Characterization of long non-coding RNAs in the Hox Complex of Drosophila

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2016

VICTORIA LEE COYNE

SCHOOL OF MEDICAL SCIENCES

LIST OF CONTENTS

LIST OF CONTENTS……..……..2
LIST OF FIGURES……..……..……4
LIST OF TABLES……..……..……...5
LIST OF ABBREVIATIONS……..……..7
ABSTRACT……..……..…… 8
DECLARATION……..……..…….... 9
COPYRIGHT STATEMENT……..……..…… 9
ACKNOWLEDGMENTS……..……...…….... 10
1. INTRODUCTION…….……..…....…..11
1.1. Discovery of the importance of DNA…….…….…….. 11
1.2. Regulatory non-coding DNA…….…….……12
1.3. Hox Genes and regulatory DNA…….……..…… 13
1.4. Classification of lncRNAs…….……..……... 15
1.5. Identification of functional lncRNAs…….……..……...17
1.6. Methods for investigating lncRNA functions……19
1.7. Identified functions of lncRNAs……21
1.8. Understanding Hox Genes and their conservation…… 27
1.9. Hox Gene collinearity……..29
1.10. Homeotic mutations……....30
1.11. Upstream regulation of the Hox Complex by segmentation genes…...…..33
1.12. *Cis* regulation of Hox Genes……35
1.13. Polycomb and trithorax proteins…… 38
1.14. Polycomb and trithorax complexes…… 39
1.15. Polycomb and trithorax complex recruitment to response elements…….. 46
1.16. Project summary and aims…….. 49
2. METHODS…… 51
2.1. Identification of lncRNA clusters within the *D. melanogaster* genome….. 51
2.2. GO Term analysis of protein-coding genes within *D. melanogaster* lncRNA clusters…… 51
2.3. Fly husbandry…….. 52
2.4. RNA collection, sequencing and annotation from *D. pseudoobscura* and *D. virilis* species……...52
2.5. Protein-coding gene ortholog comparison from lncRNA clusters in *D. melanogaster* and *D. virilis*.……54
2.6. Identification of lncRNAs in the Hox complex of *D. melanogaster*…… 54

2.7. Analysis of PcG, TrxG and HDAC binding to Hox lncRNAs in *D. melanogaster*……..55
2.8. Probe synthesis and imaging…….. 56
2.8.1 Genomic DNA extraction……..56
2.8.2 PCR primers and amplification……..57
2.8.3 Cloning……...57
2.8.4 Probe synthesis…….. 59
2.8.5 Embryo collection and fixation…….. 59
2.8.6 Embryo prefixation for hybridization……60
2.8.7 RNA probe hybridization……...60
2.8.8 Fluorescent detection……..60
2.9. Prediction of PREs using jPREdictor and evolutionary changes…….. 61
2.10. Gal4 driven ectopic expression of *Hox-G* and G-PRE……... 61
2.10.1. Cloning…… 61
2.10.2. Transformations and ectopic expression……..…….... 63
2.10.3. Inverse PCR…… 63
2.11. Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated proteins (Cas9) mediated transgenesis……...64
2.12. FLP manipulation of *Hox-G*…….. 67
2.13. Segmentation gene crosses for lncRNA expression investigations……. 67
3. RESULTS…… 68
3.1. Comparative analysis of lncRNA clustering and cluster conservation in Drosophilid genomes…….. 68
3.2. Identification of lncRNAs in the Hox complex of *D. melanogaster*……. 84
3.3. Expression patterns of Hox lncRNAs: Conservation and affects of segmentation gene mutations…….. 92
3.4. Regulatory protein binding at lncRNAs…… 101
3.5. Sequence based predictions of PREs and their evolutionary conservation..109
3.6. Homeotic mutations from Gal4 driven expression of *Hox-G* and G-PRE…… 121
4. DISCUSSION…… 144
4.1. Key Outcomes…… 144
4.2. Identification and conservation of lncRNA clusters…….. 144
4.3. Identification of lncRNAs in Drosophila Hox complex and transcribed PREs…… 148
4.4. Gain and loss of function of a novel lncRNA and adjacent PRE……... 155
4.5. Future Perspectives…….. 166
5. REFERENCES …….. 168

FINAL WORD COUNT: 57, 004

LIST OF FIGURES

1.1. Interactions of lncRNA *HOTAIR* to induce silencing to target loci……26

1.2. Common mechanisms of lncRNAs function……..…. 27

1.3. Expression and regulation of maternal gradient and segmentation genes in establishment of A-P segments and Hox gene activation…… 35

1.4. Hox gene expression through embryogenesis and adult flies……... 37

1.5. Hox genes, lncRNAs and PREs in the BX-C of *D. melanogaster*……49

2.1. Pipeline for identification of novel lncRNAs in stage 4-6 *D. virilis* embryos and lncRNA clusters……52

2.2. Cuffcompare class codes illustrations based on their loci relative to mRNA genes. 54

2.3. Plasmid map of *Hox-G* in pUAST in the forward orientation…… 63

2.4. InFusion summary for making changes to plasmids…… 66

2.5. Plasmid design for CRISPR/Cas9 injections……...66

3.1.1. *D. melanogaster* lncRNA clusters with 25 kb and 100 kb cutoff distances……... 70

3.1.2. PANTHER GO-Slim Term analysis of mRNAs from top 20 lncRNA clusters in *D. melanogaster*……74

3.1.3. Properties of lncRNAs identified in *D. virilis* embryos……... 78

3.1.4. Identification of lncRNA clusters in *D. virilis*……..... 80

3.1.5. Comparison of matching orthologous mRNAs found in 20 highest clusters for *D. melanogaster* and *D. virilis*……83

3.2.1. ANT-C *D. melanogaster* transcript identification during early embryogenesis using RNA-seq and CAGE……...87

3.2.2. BX-C *D. melanogaster* transcript identification during early embryogenesis using RNA-sequencing and CAGE…….. 89

3.2.3. Mapped reads from Hox complex of *D. virilis* RNA-seq showing syntenic lncRNA transcripts from embryogenesis stages 4-6……... 91

3.3.1. Nascent transcript FISH expression patterns of lncRNA and adjacent genes in WT stage 5 *D. melanogaster* embryos……... 93

3.3.2. Evolutionary syntenic conservation of lncRNA transcription and expression patterns in early developing Drosophilid embryos……... 96

3.3.3. Time series of transcript expression of *Hox-O, pri-mir-iab-4* and *iab-8* in *D. virilis*…… 98

3.3.4. Early embryonic altered expression of lncRNAs in homozygous segmentation gene mutants…… 100

3.4.1. ChIP-seq profiles of PcG proteins binding at lncRNAs…… 102

3.4.2. ChIP-seq and ChIP-ChIP profiles of TrxG proteins binding at lncRNAs…… 105
3.4.3. HDAC ChIP-seq binding profiles of 0-12hr embryos……...108
3.5.1. Prediction of PREs in the Antennapedia complex of *D. melanogaster, D. pseudoobscura* and *D. virilis* using the jPREdictor program……..110
3.5.2. Prediction of PREs in the Bithorax complex of *D. melanogaster, D. pseudoobscura* and *D.virilis* using the jPREdictor program……... 112
3.5.3. Analysis of conservation of *D. melanogaster* ANT-C sequence with PcG/TrxG protein binding and predicted PREs…… 115
3.5.4. Analysis of conservation of *D. melanogaster* BC-C sequence with PcG/TrxG protein binding and predicted PREs……118
3.5.5. PRE/TRE motif clustering of PcG/TrxG binding regions of experimental and validated PRE/TREs……120
3.6.1. Gal-4 driver and PBac(WH)1 and 2 adult fly images……..124
3.6.2. Mutations from Gal4 driven UAS-Piggy Back constructs upstream and within *Hox-G*……126
3.6.3. Homozygous fly lines generated by P-element insertion of the experimental PRE upstream of *Hox-G* (G-PRE)……...129
3.6.4. Homozygous fly lines generated by P-element insertion of full length *Hox-G* transcript……... 130
3.6.5. Phenotypes caused by Gal4 driven expression of *Hox-G* and G-PRE sequences… 131
3.6.6. CRISPR-Cas9 generated mini-white expression as reporter gene within *Hox-G*...134
3.6.7. Adult images of pTVCherry insert into *Hox-G* (G-pTVCherry) and heat shock Cre flies……135
3.6.8. Homeotic phenotypes from Cre excised G-pTVCherry…… 136
3.6.9. Adult images of PBac(WH)3, PBac(WH)2/PBac(WH)3 and Flippase flies used to generate partial duplication and deletion of *Hox-G*…… 138
3.6.10. Homeotic mutations arising from flippase mediated duplication and deletion of 3' *Hox-G* fragment……... 139
3.6.11. Necrotic and mutated legs from Gal-4-UAS and Cre experiments…… 140
4.2.1. Hox complex alignment of protein coding genes and lncRNAs in 3 different Drosophila species…… 150
4.3.1. Comparable expressions of Wg, Dpp and En/Hh in developmental primordial compartments and Ubx regulatory network of haltere…… 160

LIST OF TABLES

1.1. PcG and TrxG complexes……..…….. 42
2.1. GEO Accession numbers linked to modENCODE RNA-seq…….. 55
2.2. Accession numbers of datasets from GEO repository of PcG, TrxG and HDAC

ChIP experiments…… 56
2.3. Primer sequences for genomic DNA amplification used for probe synthesis……. 58
3.1.1. Summary of novel lncRNAs identified from RNA-seq in *D. virilis*…….. 77
3.2.1. Summary table of lncRNA transcripts throughout the Hox complex of *D. melanogaster*…….. 85
3.6.1. Gal4 drivers lines expression patterns and stages of expression…….....123
3.6.2. Penetrance table Mutations from Gal4 driven UAS-Piggy Back constructs upstream and within *Hox-G*.……127
3.6.3. Penetrance scores of Gal4 *Hox-G* transcript and G-PRE overexpression……... 132
3.6.4. Penetrance scores mutations caused by Cre excised G-pTVCherry and Flippase mediated partial duplication of *Hox-G*.……141
3.6.5. Sequenced pUAST insertion sites and associated genes…….. 142
4.1.1. Alignment of RNA-seq time points and *D. melanogaster* embryogenesis stages…. 151

LIST OF ABBREVIATIONS

A-P	Anterior-posterior
abd-A	Abdominal-A
Abd-B	Abdominal-B
ac	Acetylation
ANT-C	Antennapedia complex
Antp	Antennapedia
Ash1	Absent, small or homeotic 1
bcd	Bicoid
bx	Bithorax
BX-C	Bithorax complex
bxd	Bithoraxoid
CAGE	Cap analysis of gene expression
ChIP	Chromatin immunoprecipitation
CoREST	Co-repressor for element-1 -silencing transcription factor
CRISPR	Clustered regulatory interspaced short palindromic repeats
D-V	Dorsal-ventral
dally	Division abnormally delayed
Dfd	Deformed
Dll	Distalless
DNMT	DNA methyltransferases
Dpp	Decapentaplegic
Dsp1	Dorsal switch protein 1
E(z)	Enhancer of zeste
en	Engrailed
ENCODE	Encyclopaedia of DNA elements
esc	Extra sex combs
eve	Even-skipped
Exd	Extradenticle
Fab-7	Frontabdominal-7
fkh	Forkhead
FLC	Flowering locus control
ftz	Fushi tarazu
GEO	Gene expression omnibus
grh	Grainy head
gt	Giant
h	Hairy
H3K27	Histone 3 lysine 27
H3K4	Histone 3 lysine 4
H3K9	Histone 3 lysine 9
H4K16	Histone 4 lysine 16
hb	Hunchback
HDAC	Histone deacetylase
HMTase	Histone methyltransferase
HOTAIR	Hox transcript antisense
HOTTIP	HOXA transcript at the distal tip
Hox	Homeobox
HSPG	Heparin sulphate proteoglycan
Hth	Homothorax
iab-4	Infraabdominal-4
iab-8	Infraabdominal-8
IGV	Integrated Genome Viewer
inv	invected
Kis	Kismet
kni	Knirps
Kr	Krüppel
lab	Labial
lncRNA	Long non-coding RNA
MBT	Malignant brain tumour
me1	Monomethylation
me2	Dimethylation
me3	Trimethylation
miRNA	MicroRNA
MLL	Mixed lineage leukemia
MOF	Males absent of the first
mRNA	messenger RNA
MSL	Male specific lethal complex
ncRNA	Non-coding RNA
ntFISH	Nascent transcript fluorescent in situ hybridization
NURF	Nucleosome remodelling Factor
ORF	Open reading frame
pb	Proboscipedia
PcG	Polycomb Group
Pcl	Polycomb-like
Ph	Pleiohomeotic
PhoRC	Pho repressive complex
piRNA	Piwi-interacting RNA
PRC 1/2	Polycomb repressive complex 1/2
PRE	Polycomb response element
PS	Parasegment
PSS	Pairing sensitive silencing
Psc	Posterior sex combs
psq	Pipsqueak
PSS	Pairing sensitive silencing
rRNA	Ribosomal RNA
run	Runt
Scm	Sex combs on midleg
Scr	Sex combs reduced
siRNA	Small interfering RNA
Su(z)12	Suppressor of zeste 12
TAC1	Trithorax acetylation complex 1
TF	Transcription factor
tkv	Thick veins
TRE	Trithorax response element
Trl	Trithorax-like
tRNA	Transfer RNA
TRR	Trithorax-related
Trx	Trithorax
TrxG	Trithorax Group
TSS	Transcriptional start site
Ubx	Ultrabithorax
vg	Vestigial
wds	Will die slowly
wg	Wingless
Xist	X-inactive specific transcript
z	Zeste

ABSTRACT

The University of Manchester
Victoria Lee Coyne
Doctor of Philosophy
Characterization of long non-coding RNAs in the Hox Complex of Drosophila

Long non-coding RNAs (lncRNAs) are often defined as transcripts >200nts that have no discernable 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; Rinn and Chang, 2012) across vastly 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). Investigations into lncRNAs in wide ranging organisms, have revealed that many influence gene activity by forming ribonucleoprotein complexes that affect the conformational state of chromatin (Rinn and Chang, 2012). A genomic region that has revealed several functional lncRNAs in diverse organisms is the Hox complex (Pauli et al., 2011; Pettini, 2012; Rinn et al., 2007). The Hox complex encodes 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 bilatarian animals (Moreno et al., 2011). Misexpression or mutation of Hox genes causes morphological and pathophysiological defects (Quinonez and Innis, 2014). We investigated clustering of lncRNAs throughout the *D. melanogaster* genome using available annotations and carried out RNA-seq in *D. virilis* to expand the repertoire of lncRNAs and identify clusters of lncRNAs. We found the Hox complex to be heavily enriched with lncRNAs in both organisms, and syntenic transcripts from *D. melanogaster* could be identified in *D. pseudoobscura* and *D. virilis*. Several lncRNAs aligned with polycomb response elements (PREs); transcription of PREs has previously been linked to a switch in their activity (Herzog et al., 2014). However, we found that transcribed PREs in *D. melanogaster* move positions relative to the protein-coding genes in other drosophilids, whilst the transcriptional units remain in the same syntenic region. Conservation of syntenic transcripts without evidence of remaining a PRE suggest that the transcription is not linked to PRE function, agreeing with recent findings that transcription of PREs does not affect their function (Kassis and Muller, 2015). We investigated functions of a novel lncRNA and adjacent PRE in the Hox complex by ectopic expression and utilization of other genetic manipulation tools. Overexpression of either the lncRNA or PRE and partial duplication of the lncRNA caused phenotypes such as missing halteres and/or T3 legs, misshaped T3 legs or malformed abdominal segments. The observations that ectopic expression of this lncRNA and an adjacent regulatory element from the Hox complex causes phenotypes that can be linked to adjacent Hox gene misregulation, *Antp* and *Ubx*, suggest that they are likely to have roles in the regulation of at least one of these Hox genes.

DECLARATION

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to Dr. Matthew Ronshaugen for introducing me to the fascinating world of lncRNAs and for his patience and guidance throughout the course of my PhD. I would also like to thank my co-supervisor Prof. Sam Griffiths-Jones for his input in bioinformatic analyses that I learnt along the way. Two lab members were particularly patient and helpful when I began in the lab, Dr. Maria Ninova and Dr. Tom Pettini, both of who guided me through many techniques that were new to me and that I may never have figured out without them. Catherine Sutcliffe, the technician from the Ashe Lab, was also amazingly insightful when helping me figure out what was likely to have gone wrong during experiments and how to avoid mistakes, along with another previous Ashe Lab member, Dr. Franziska Bonath. One of my master's students, Thomas Minchington, was an excellent lab member who helped identify the enhancers, and another lab member, Tom Bleazard, was also indispensible when it came to some of the bioinformatics. I am very grateful to both of them. Several other lab members, past and present, have been a pleasure to work with, Dr. Ana Kozomara, Dr. Suzanne Hunt, Dr. Katarzyna Hooks, Steven Woods, Laila El Haddad, Dr. Antonio Marco, Elizabeth Stoops, Rachel Symons, Scott Wilcockson, Marco Pinheiro and Dr. Yamini Arthanari.

Much of this work would not have been possible without the University of Manchester's Genomic Technologies Facility, the DNA Sequencing Facility, the Fly Facility and the confocal microscope I was allowed to use from Prof. Nancy Papalopulu's lab.

The one person who has made all the tough times during my PhD bearable is my partner, Alex Robertson. He has spent a lot of time running around after me and waited patiently to pick me up and drop me off when I work late or early on countless occasions and generally put up with a lot of inconveniences during the past four years. I would also like to thank my mum, Wendy Colton, for always giving me the confidence to achieve anything I set my mind to and my brother, Danny Coyne, for always being able to cheer me up. Also, all the people I dance with regularly have always been able to put a smile on my face, especially my dance partner and friend, Alistair Ho.

Finally, I am very grateful to the Biotechnology and Biological Sciences Research Council for funding my research and making this PhD possible.

1. INTRODUCTION

1.1 Discovery of the importance of DNA

Throughout nature there is enormous variety in the organization of body plans of metazoan animals that are believed to have diversified from unicellular eukaryotic life forms over 600 million years ago (King, 2004). A key challenge for biology is to understand how genetic changes have led to the vast array of phenotypic variations within and between the species that have evolved since the transition to multicellularity. Gregor Mendel demonstrated during the 1860s that morphological traits could be inherited by offspring. Mendel carried out experiments demonstrating transmission of heritable phenotypic traits in the pea plant, *Pisum sativum*. This, in part, has led to our modern understanding that transmissible elements, now known as genes, are passed on to progeny in a simple and consistent pattern. Also, that these elements can exhibit different phenotypic consequences, as seen in recessive and dominant inheritance (Miko, 2008). Twenty years later, a German embryologist and cytologist, Oscar Hertwig, proposed that heredity was a result of chromosomes originating from the nucleus of sperm and egg cells (Hertwig, 1885), narrowing down the origin of these heritable traits. Then, in 1933 Thomas Hunt Morgan was awarded the Nobel Prize in Physiology or Medicine for discovering the role of chromosomes in heredity. Morgan's work identified spontaneous mutations in *D. melanogaster* that gave visible phenotypes, allowing him to demonstrate sex linkage and the linear arrangement of genes within chromosomes (www.nobelprize.org).

Initially, investigations into how morphological diversity evolved focused on alterations in protein-coding genes, such as nucleotide mutations, timing of expression, pace or magnitude of synthesis of mRNA into protein. It was originally thought that changes in the DNA sequence of protein-coding genes, could explain proteins adaptations giving rise to novel functions and that these changes would constitute the sole source of morphological evolution.

In 1909 Archibold Garrod proposed that one gene codes for one specific enzyme. Nobel Prize Laureates, Beadle and Tatum later corroborated Garrod's theory in experiments, in 1941 when they introduced mutations to genes that coded for metabolic enzymes. Different enzymes were then shown to be responsible for metabolism of different nutrients required by *Neurospora crassa* in order for them to grow (Beadle and Tatum, 1941). This lead to work by Avery, MacLeod and McCarty who recognized that DNA was the molecule ultimately responsible for transmissible genetic traits between different strains of *Streptococcus pneumoniae* cells (Avery et al., 1944). Austrian chemist, Erwin Chargaff, analyzed the composition of DNA from a variety of species and found that the number of guanine bases was always equal to the number of cytosine and the number of adenine molecules always equaled the number of thymine bases (Vischer and Chargaff, 1948). At the time Chargaff did not realize the implications of his work and it was later that the x-

ray diffraction photography of DNA molecules, carried out by Rosalind Franklin, that allowed Watson and Crick to figure out that the structure of DNA is a double helix (Watson and Crick, 1953). The discovery of DNA as the chemical blueprint of genes, and not proteins, initiated a much deeper understanding of the exact composition, structure (Brown, 1952) and conformation (Watson and Crick, 1953) of DNA. We also now understand that there are key principles followed by the genetic code, with few exceptions, shared throughout all species on earth (Hinegardner and Engelberg, 1963; Woese, 1964). There are 61 trinucleotide codons that are translated into 20 amino acids (Chaney and Clark, 2015), plus start and stop codons that are shared almost uniformly by every organism This degeneracy can allow for slight changes in nucleotide sequence without changing the amino acid sequence of the final protein, but the wrong change to the nucleotide sequence is also able to cause lethality to the whole system (Koonin and Novozhilov, 2009). Together these findings have led to the deep knowledge we have today on how DNA functions throughout all life forms on earth.

1.2 Regulatory non-coding DNA

The hypothesis in the mid 20th century that one gene produces one enzyme, or polypeptide, is now known to be an inaccurate description (Beadle and Tatum, 1941; Ingram, 1956, 1957). King and Wilson (1975) investigated genetic variation throughout the genomes of humans and their closest evolutionary relative, the chimpanzee, and found that protein sequences were 99% similar, whereas the non-coding sequences had more variability (King and Wilson, 1975). We now know that DNA consists of an enormous collection of functional elements, ranging from small to large non-coding RNA (ncRNA) molecules, protein-coding messenger RNAs (mRNAs) and regulatory sequences of DNA able to act on specific genes to alter their expression. It came as a great a surprise in 2001 when the International Human Genome Sequencing Consortium realized that just <1.5% of the human genome contained instructions for making proteins from mRNA (Lander et al., 2001). This has lead to an ever increasing curiosity as to if or how the other >98% of the genome functions and how much may constitute 'junk DNA', a term first coined by the geneticist Susumu Ohno (Ohno, 1972). Then, in September 2012 an international project was founded to catalogue all functional DNA called the Encyclopedia of DNA Elements (ENCODE) and they estimated that 80% of the human genome had a biochemical function (Consortium, 2012). This caused quite a bit of controversy as the 80% estimated by ENCODE was a much higher proportion than anyone had expected. The high number generated by this study was likely due to the designation used in the study for functional DNA, as this was assigned to any DNA that was bound by proteins or underwent any chemical modifications with little phenotypic functional evidence for the majority of these regions. Furthermore, much of the data generated was from pluripotent and stem cells and therefore also

not likely to be a true representation of the human *in vivo* genomic environment (Graur et al., 2013).

We now know that regulatory DNA changes are more likely to account for morphological changes between organisms, as changes in protein-coding sequence or numbers of genes alone do not alter rapidly enough to account for the vast variations in morphology (Taft et al., 2007). It was originally thought that a more complex organism arose from having more genes, but we now know that the simple roundworm, *C.elegans*, has many more protein-coding genes (~19,300) than insects (~13,500) and two-thirds as many as humans (Taft et al., 2007). We have now also established that proteins amino acid sequences change very little between closely related species. For example, humans share 29% identical protein sequences with chimpanzees and the other orthologous proteins typically undergo just 2 changes in amino acids composition (Watanabe et al., 2004). It is now much more widely accepted that the amount of non-coding DNA sequence increases with organism complexity and is likely to have a larger impact on phenotypic evolution than the number of protein coding genes (Taft et al., 2007).

An earlier large-scale ENCODE project attempted to identify how much of the human genome is transcribed by combining data across different tissues at different time points from different experiments. In this experiment they found that 93% of total bases were transcribed in at least 2 independent experiments when based on the same technology being used and this percentage was reduced to 74% if confirmed by 2 different technologies (Consortium et al., 2007). In 2010, the number of protein-coding genes in the human genome was calculated to be ~22,000 (Pertea and Salzberg, 2010) and in 2012 the number of loci containing mRNAs was 20,944 producing 111,451 transcripts (Pertea, 2012). When considering lncRNAs in the same study, 40,765 loci produced 89,981 transcripts and small RNAs just 11,366 transcripts from 11,195 loci. By 2015, after further transcriptome analysis of thousands of biological samples and cell lines, the number of lncRNA transcripts in the human genome rose to 58,648, making up 68% of the 90,013 total expressed genes (Iyer et al., 2015).

1.3 Hox Genes and regulatory DNA

The genes that are responsible for establishing regional identities along the anterior posterior axis, the Hox genes, are exceptionally conserved in virtually every bilatarian organism (Heffer and Pick, 2013). The Hox genes were discovered in *D. melanogaster* from their ability to transform segment identities in developing embryos. For example, Hox genes have the ability to transform one type of appendage to another in the adult fly (Bateson, 1894; Lewis, 1978). *D. melanogaster*, along with other insects, polychaetes, onchophorans and sea urchins, have one Hox cluster, but throughout evolution there have been various duplications leading to 4 Hox clusters in mammals and 8 in teleosts (Heffer and Pick, 2013). It has been suggested by some that this duplication of

protein-coding Hox genes has had a role in evolutionary developmental changes (Wagner et al., 2003), as the duplicated versions of Hox genes may acquire novel functions. However, Hox gene sequences do not diverge enough to account for the vast variation within vertebrate and invertebrate species. This was exemplified in a study that took the mammalian ortholog of Drosophila Hox gene *Sex combs reduced* (*Scr*), in mammals *Hoxa5,* and misexpressed it in *D. melanogaster* to produce the same phenotype as ectopic expression of *D. melanogaster's Scr*, whereby antennae are transformed into T1 legs (Zhao et al., 1993). It is now thought that changes in their regulation are more likely to account these differences.

Classic investigations have focused on enhancers, regulatory regions of DNA that have binding sites for multiple sequence-specific TFs and coactivators able to drive transcription by recruiting transcription machinery (Veitia, 2008). Enhancers can bypass adjacent genes to activate their target genes and are able to act from 100s of kilobases away from their targets (Levine and Tjian, 2003) to loop around to reach the promoter regions when activated (Levine, 2010). Within the ANT-C, there is a T1 enhancer downstream of *ftz* that loops over past *ftz* to activate the next gene upstream of *ftz, Scr* (Gindhart et al., 1995). A 450bp promoter proximal tethering element, just upstream of *Scr*, was found to be essential for mediating the T1 enhancer interaction with the *Scr* promoter and this tethering element could also induce T1 enhancer activation of *ftz* when placed 5' of the *ftz* promoter (Calhoun et al., 2002). Another regulatory DNA element known as chromosomal boundary elements, or insulators, protect genes from inappropriate activation by enhancers and if these become mutated then it is possible for enhancers to activate the genes they were protecting (Levine, 2010). Insulators are associated with binding of specific proteins to block enhancer activity and are thought to partition chromatin into regulatory domains to prevent the spread of epigenetic marks or chromatin modifying proteins (Negre et al., 2010). A combination of tethering elements, promoter specificity and insulators are therefore able to direct gene expression although it is still not clear what guides these relationships they all play major roles in genome organization (Levine, 2010). The regulatory roles these DNA elements in gene expression have been shown to be a major factor in embryonic patterning and variations in phenotypes of metazoans (Starr et al., 2011) and we now know that many of these sites are transcribed into ncRNA (Simonatto et al., 2013).

In the considerably compact genome of *D. melanogaster*, ~75% of the genome was found to be transcribed at some points during development and ~20% of the genome is estimated to encode mature mRNA and 60% is transcribed into primary mRNA transcripts (Graveley et al., 2011). FlyBase currently reports 13,907 mRNA genes, 2,470 lncRNAs and 565 small RNAs out of a total of 17,728 genes (includes rRNA and tRNA) (release 6.11). However, the *D. melanogaster* transcriptome has not been studied as extensively as the human transcriptome, with 135 RNA expression profiles and 411 gene structure analyses (http://www.modencode.org/#33), suggesting that in the future these numbers could increase with further study. The use of next generation

sequencing technology has led to the increasing discovery of thousands more transcripts in a wide range of diverse species, with an increasing number seeming to have no protein-coding potential. This has led people to question if all of these non-coding transcripts could be functional, transcriptional noise or micro-peptides that do not fit our current understanding of protein classification (Ji et al., 2015). It has now been established that as there is a direct correlation with organism complexity and proportion of the genome that is transcribed into ncRNA (Fatica and Bozzoni, 2014). This has led to the theory that variation in the non-coding portion of the genome could responsible for the diverse morphological variances seen throughout the eukaryotic kingdom, whether it be ncRNA or regulatory DNA sequences (Gaiti et al., 2015).

1.4 Classification of lncRNAs

Between 1992-2011, the total number of publications identifying lncRNAs steadily increased before a sharp rise after 2011 (Quek et al., 2015). Despite this, a very small number of lncRNAs have actually been assigned functions (Quek et al., 2015). Due to the lack of understanding of lncRNAs, they are still arbitrarily classed 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 and 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 form 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 small ncRNAs all have 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 and that have normally not been functionally characterized have generally been assigned the term lncRNAs with a limit of being at least 200nts in length to separate them from the small ncRNAs. These lncRNA transcripts have now been identified throughout massively 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 belief that these transcripts lacked any functional biological relevance, as sequence conservation is typically linked to significant and usually conserved function (Ohno, 1972; Struhl, 2007).

The small numbers of functionally characterized lncRNAs have been associated with a multitude of biological processes ranging from transcriptional activation, silencing, splicing, protein complex organization, cell cycle progression, apoptosis and response to stress (Quinn and Chang, 2016). LncRNAs play crucial roles in transcriptional gene regulation and when their function is disrupted it often leads to severe biological disorders. However, due to lack of information based on functional characteristics, an early attempt at classification of 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). Given the limited number of functionally understood lncRNAs it remains to be seen if any of these categories will be useful.

1.5 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 is important due to the time, cost and effort that need to go into understanding how a single or group of lncRNAs may function. There have been a number of different approaches to this using a variety of available technologies. To begin with, transcription must be reliably detected and be independent of protein-coding genes. This was problematic before stranded RNA-seq aided the identification of transcripts that were within or overlapping protein-coding gene transcription, as they would often be considered immature RNA and consolidated with the mRNA. Stranded RNA-seq has rapidly become the preferred method of detection due to its ever-decreasing costs and ability to rapidly and sensitively produce billions of reads that can be mapped to a genome (Mutz et al., 2013), allowing identification of *de novo* transcription and relative concentrations. Other technologies that have been used to identify lncRNAs include tiling arrays, serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE) and Chromatin immunoprecipitation (ChIP).

Once lncRNAs are annotated throughout a specific genome, the transcripts that are most likely to have biological functions need to be identified from the mass of transcription. Many lncRNAs have been contested as 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). The assertion that many lncRNA transcripts are a result of transcriptional noise is based on them frequently being found to have very low levels of transcription and being shorter and less complicated with no splicing events or fewer exons/introns than mRNAs (Ravasi et al., 2006). The arguments for the likelihood of lncRNAs being a consequence of transcriptional noise is that they are thought to arise from cryptic promoters within protein-coding gene regions, or intergenic regions that are simply devoid of histones thereby allowing access to transcription machinery that initiates at sequences that resemble promoters by chance (Struhl, 2007). Adding weight to this argument is that many lncRNAs 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 perturb their functions, which should lead 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, as 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 actually found that lncRNAs exhibit a much higher tissue specificity, regardless of levels of expression, than protein-coding genes, leading many to believe that this is a good indicator of functionally relevant lncRNAs (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 traditional central dogma of biology. These tools assess an RNA sequence based foremost on possession of a continuous stretch of DNA that begins with an initiator 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., 2008b). 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 programs now calculate minimum ORF cutoffs 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 *tarsal-less* (*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).

Other bioinformatic prediction analysis tools are also typically used to evaluate if a transcript has the potential to encode proteins by searching for similarities to any known protein family domains. This can be done relatively easily using programs such as Pfam (Finn et al., 2016), SUPERFAMILY (Gough et al., 2001) or BLASTX (Gish and States, 1993), which also consider ORF conservation, as this is considered another way to indicate coding potential. Further analysis of novel transcripts can be to investigate if they may belong to other RNA families. This can be carried out with Rfam to assess if there are any regions of consensus secondary structures and functions in a similar manner to Pfam (Nawrocki et al., 2015). RNA folding prediction tools have also been used in an attempt to identify conserved secondary structures, such as RNAz (Washietl et al., 2005) and Evofold (Pedersen et al., 2006). However, many of these use thermodynamic stability to predict canonical base interactions and there is still very little information available as to how lncRNAs are structured *in vivo* or indeed *in vitro,* to make these predictions useful. Due to the rapidly increasing amounts of transcripts being identified across a wide range of species, from different tissues and time points, bioinformatic tools have been developed combining these various aspects that need to be considered in order to accurately distinguish coding and non-coding transcription, such as the Coding Potential Calculator (CPC) (Kong et al., 2007), Coding Potential Assessment Tool (CPAT) (Wang et al., 2013) and phylogenetic Codon Substitution Frequency (phyloCSF) (Lin et al., 2011).

1.6 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 can be used to experimentally investigate functionality of lncRNAs, such as loss-of-function and gain-of-function experiments. 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 and so has been found to be more effective at targeting lncRNAs found in the same cellular compartment, such as *OIP5-AS1* and *DANCR* (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 (CRISPR-Cas9) 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, colocalisation 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 labeling of *Xist* showed that it was 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).

Arguably the most useful 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 (Fig.1.1). For example, a lncRNA from the *DHFR* loci can form a triplex with the DNA at the *DHFR* promoter and prevent Poll II transcription (Martianov et al., 2007) and lncRNAs have been found interacting with mRNAs, such as half-STAU1-binding site RNAs (½-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 analyze 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).

1.7 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. One example of a developmentally linked lncRNA that participates in imprinting, ensuring monoallelic expression of a parental gene by epigenetic mechanisms is the lncRNA, *Kcnq1ot1*. *Kcnq1ot* suppresses paternally inherited genes via interactions with G9a and polycomb repressive complex 2 (PRC2) to trimethylate H3K9 and H3K27 respectively (Pandey et al., 2008).

There are few known lncRNAs that do have high levels of sequence conservation, such as metastasis-associated long adenocarcinoma transcript 1) *MALAT1*, a functional lncRNA involved in alternative splicing and epigenetic regulation of genes that regulate cell cycle. *MALAT1* was found to be highly conserved in sequence from humans to zebrafish (Yang et al., 2011). However, there are very few examples of similar levels of conservation of other lncRNAs and instead there is better evidence that lncRNAs conserve folding and structure, allowing them to maintain interactions with the other biological molecules and therefore conserve function but not sequence (Diederichs, 2014). This has been observed for the X-inactive specific transcript (*Xist*), a mammalian lncRNA that plays a major role in the inactivation of chromosome X of females during early embryonic development (Plath et al., 2002). *Xist* was found to have conserved structural features consisting of repeat regions, one in particular that was sufficient to carry out its X-inactivation in both humans and mice via its interaction with an epigenetic silencing complex, the PRC2 (Minks et al., 2013; Wutz et al., 2002; Zhao et al., 2008).

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 a number of 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 *lgf2r* genes promoter (Latos et al., 2012). *Airn* is thought to utilize multiple unknown silencing mechanisms to mediate silencing of paternal alleles of *lgf2r, Slc22a3* and *Slc22a2,* at different developmental stages. However, the *lgf2r* 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 *lgf2r* promoter. A similar example of this mechanism is found in *S. cerevisiae*, whereby the Rap1 activator induces expression of an intergenic lncRNA, *ZRR1*, displacing Rap1 from the *ADH1* promoter leading to *ADH1* repression (Bird et al., 2006). This has 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 neighboring 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 affect gene expression by changing the nucleosome density in promoter or enhancer regions, thus facilitating or restricting TF access to DNA. This is 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). In *S. pombe*, a converse action was found for activation of the *fbp1+* locus, where chromatin is progressively remodeled into an open conformation by RNA pol II transcription of several lncRNAs through *fbp1+*, leaving the chromatin more accessible to TFs (Hirota et al., 2008). In *S. cerevisiae*, similar antisense transcription, overlapping the *PHO5* promoter, causes activation attributed to

displacement of nucleosomes that increased the rate at which chromatin could be remodeled to facilitate recruitment of RNA pol II (Uhler et al., 2007). Although this mechanism of gene regulation by nucleosome repositioning has only been supported in limited studies and for quite lengthy lncRNAs, it does reflect the general convention of lncRNA remodeling 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. This was first demonstrated in *cis* as the means by which *Xist* establishes inactivation on the X chromosome of mammalian females to account for dosage compensation between males and females (Borsani et al., 1991; Brockdorff et al., 1991; Brown et al., 1991). The *Xist* lncRNA forms part of the X-chromosome inactivation center (Xic) and is itself regulated by other lncRNAs, *Tsix* that runs antisense to *Xist* and *Jpx*. *Xist* produces a 17 kb transcript that coats the X chromosome that it is transcribed from during the initiation stages of chromosome inactivation (XCI) (Clemson et al., 1996), thereby acting in *cis*. The *Xist* transcript begins by structurally remodeling the X chromosome and positioning its target sites into the *Xist* silencing compartment (Chaumeil et al., 2006). *Xist* folds into three-dimensional structures that are able to identify target sites and move them towards the *Xist* locus, allowing *Xist* to spread further whilst remaining tethered to its original transcription site (Engreitz et al., 2013). A 1.6 kb ncRNA known as *RepA*, within the *Xist* locus, then directly recruits PRC2 via interacting with the Ezh2 subunit of the complex to carry out XCI by trimethylation of lysine 27 of histone H3 (H3K27me3) of the X (Zhao et al., 2008).

Its antisense transcript, *Tsix*, negatively regulates transcription of *Xist* in 3 different ways. In one instance it mediates interchromosomal pairing of the two X chromosomes by interacting with a chromatin insulator, CTCF, to facilitate communication between them to ensure that only one X chromosome will be inactivated (Xu et al., 2007). *Tsix* can also recruit Dnmt3a, a DNA methyltransferase that can silence *Xist* (Sun et al., 2006) and also inhibit the interaction of the RepA ncRNA, thereby preventing binding to PRC2 (Zhao et al., 2008). Positive regulation of *Xist* is mediated in *trans* by another lncRNA, *Jpx* (Tian et al., 2010). *Jpx* is upregulated when the X chromosome is inactivated and then removes the repression from a single *Xist* allele, by physically binding the CTCF protein and removing it from one of the *Xist* promoters and titrating it away (Sun et al., 2013b). Collectively, these functions of lncRNAs at the Xic, amongst other explanations of lncRNA function, mirror accounts of traditional protein regulation by multiple TFs acting in cascades or in response to certain stimuli, lending weight to the RNA world theory. This theory hypothesizes that protein and DNA came later in the evolution of life and therefore it would not be impossible to imagine that ncRNAs could reflect ancient mechanisms that have not

needed to progress into proteins, or the young versions of *de novo* proteins being formed that have developed functions that may acquire further functions if they do develop into proteins.

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. There appears to be no significant sequence similarity between *Rsx* and *Xist* besides an enrichment of 5' tandem repeats and similar motifs that may form stem-loops, but the two are not homologous (Grant et al., 2012). However, transcription of an analogous lncRNA is detected in other metatherians and clearly functions in a very similar manner, utilizing H3K27me3 for gene silencing (Grant et al., 2012). 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). The spread of MSL occupancy across the X chromosome is dependent on acetylation of H4K16 (H4K16ac), an epigenetic mark that increases accessibility and is carried out by an acetylase component of the MSL itself, *males absent on the first* (MOF) (Bell et al., 2010). The hyperacetylation of X prevents compaction of chromatin, thought to increase the access of TFs to DNA and therefore increasing gene expression. This demonstrates that chromatin reorganization via interactions of lncRNAs and methyltransferase, acetyltransferase, deacetylase or demethylase proteins 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 R-loop, a 3-stranded structure of RNA, DNA and single-stranded DNA (Sun et al., 2013a) that have also been shown to form heterochromatin in mammals. Increased expression of AS I leads 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 recruit 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.1.1) (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.

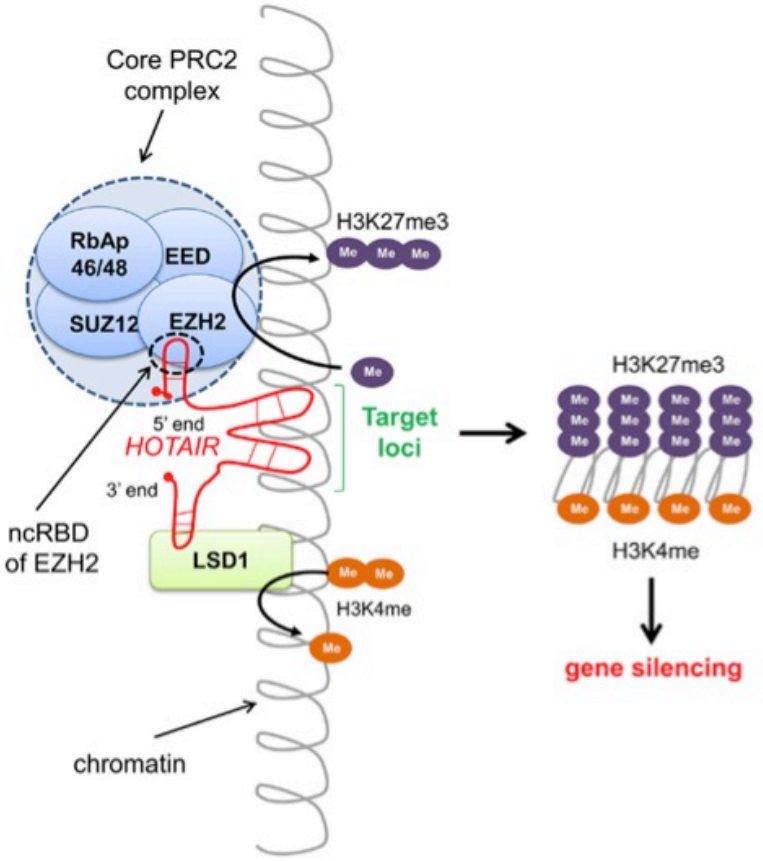


Figure 1.1. Interactions of lncRNA *HOTAIR* to induce silencing to target loci. HOTAIR interacts with LSD1 at the 3' end and the EZH2 component of the PRC2 complex at the 5' end leading to H3K27me3 and H3K4me modifications to induce silencing of target loci on a different chromosome than the *HOTAIR* 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), with the RNAcentral database now listing 91 lncRNA sequences (including transcript variants) from human Hox loci (Consortium, 2015). Other functional lncRNAs have been identified in both fruit flies and mice from within Hox clusters, such as *bxd* (Bender et al., 1983) and *iab-8* (Zhou et al., 1999) in *D. melanogaster*, and *Evx1as, Hoxb5/6as* (Dinger et al., 2008a) and *LncRNA-HIT* (Carlson et al., 2015) in mouse. The majority of these are strongly linked to Polycomb silencing and trithorax activating proteins by either a direct interaction or gain and loss of epigenetic marks associated with their transcription (www.rnacentral.org). Other chromatin modifying complexes found to be regulated by lncRNAs, such as the CoREST complex, Smcy homolog, X-linked (SMCX), G9a, growth arrest and DNA-damage-inducible, alpha (GADD45A) DNA methyltransferases (DNMTs) and LSD1 (Han and Chang, 2015) have been characterized, further suggesting that gene regulation by lncRNAs is a highly conserved mechanism, with four main models summarized in figure 1.2.

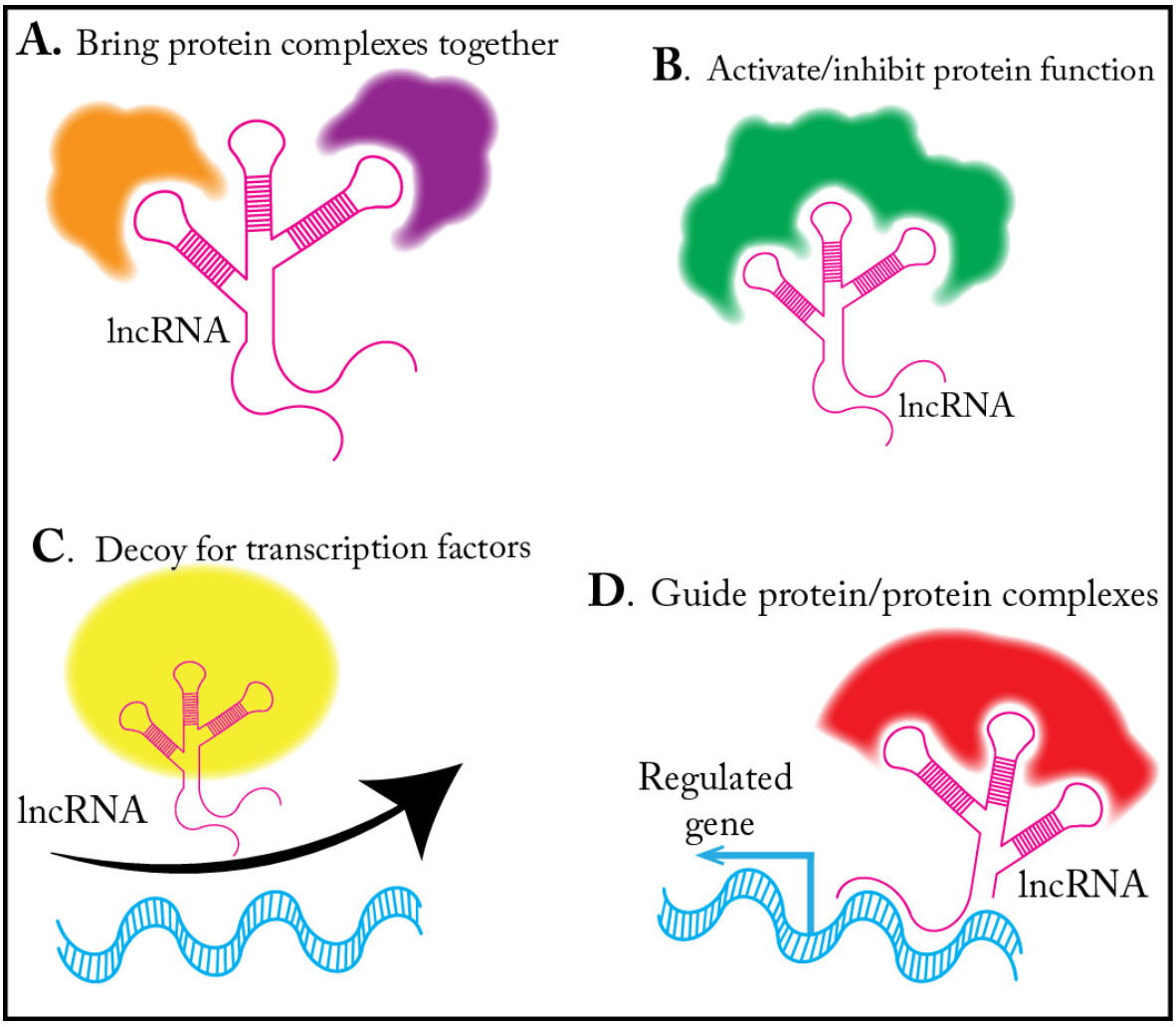


Figure 1.2. Common mechanisms of lncRNA function. A) lncRNAs can bind to proteins or protein complexes with different regions of their transcript acting as a scaffold as seen for *HOTAIR, ANRIL* and *Kcnq1ot1*. B) lncRNAs are able to allosterically alter proteins to activate or silence, examples include *ncRNA*$_{CCND1}$ and *HSR1*. C) lncRNAs can act as a decoy and remove TF's from regulatory regions to prevent activation/silencing of targets. D) lncRNAs can guide proteins or complexes to targets for regulation, includes *Gas5, NRON* and *PANDA*. Functional reviews that inspired figure are Rinn and Chang, 2012 and Geisler and Coller, 2013.

1.8 Understanding Hox Genes and their conservation

For many years the question of what the genetic differences are between the vast number of morphologically varied species on the planet, and how they are able to develop into highly organized functional organisms, has fascinated people from a variety of disciplines. Nearly 40 years ago developmental geneticists identified a set of genes, termed the bithorax complex (BX-C), responsible for the regulation of development of the Drosophila embryo. These genes were genetically characterized by Nobel Prize winner Edward B. Lewis in 1978 and shown to be responsible for segment identity, specifying structures along the anterior-posterior (A-P) axis of *D. melanogaster* (Lewis, 1978). The Hox genes have since been found to pattern the early A-P axis in all bilateral species (Garcia-Fernandez, 2005). Later, William McGinnis *et al.* discovered a

conserved sequence shared by all Hox genes in the BX-C and Antennapedia complex (ANT-C). This conserved DNA region encodes a 60 amino acid homeodomain that is the DNA binding region of these proteins (McGinnis, 1984b). The term homeodomain is derived from the term 'homeotic', which originates from observations made by William Bateson in 1894, when he wrote a book describing transformations of one body part to another, such as when antennae are transformed into legs. He coined the term homeosis and suggested that homeotic transformations could be the basis of morphological evolution (Bateson, 1894).

In animals, there are now 11 different classes of homeodomain genes and 14 classes in plants (Holland, 2013). The Hox genes belong to the largest class, ANTP, which includes Parahox, NK cluster and others, that together with Hox genes are responsible for patterning the mesoderm, nervous system and gut (Holland, 2013). The Hox genes are thought to be key to the diversification of body plans of all bilaterians and their regulated expression and molecular function remain a mystery to this day. Extra layers of regulation from non-coding RNAs that have been recently discovered further confound explanations of phenotypic evolution. In mammals, duplication events, along with gene losses and gains have produced 39 Hox genes on 4 clusters that are expressed along the A-P axis (Heffer and Pick, 2013). Different Hox genes are able to recognize similar DNA-binding sequences and replace the function of one another and their specificity *in vivo* is thought to come from their interactions with cofactors and other DNA-binding partners (Heffer and Pick, 2013).

In 1915, one of the earliest homeotic mutations was identified as a spontaneous mutation that produced a fly with a partial transformation of the third thoracic segment to the second, visible as a haltere developing into wing like structures, earning the title *bithorax* (*bx*) (Bridges, 1923). Then in 1919 a similar mutation was identified from a nearby region on the chromosome that was named *bithoraxoid* (*bxd*) by Bridges, followed by identification of the dominant mutation, designated *Ultrabithorax* (*Ubx*) (Hollander, 1937). The protein coding genes producing these mutations became known as Hox genes, found on the right arm of the third chromosome, split into two complexes in *D. melanogaster*. These eight Hox genes are split between two complexes, the BX-C, containing *Ubx, abd-A* and *Abdominal-B* (*Abd-B*) that determine parasegments 5-14, making up the posterior two-thirds of the embryo, and the ANT-C, containing *Sex combs reduced* (*Scr*), *Deformed* (*Dfd*), *proboscipedia* (*pb*) and *labial* (*lab*); establishing identity of the front third of the segments of the embryo (Pearson et al., 2005). Early studies in *D. melanogaster* revealed that Hox genes exhibit a spatially collinear relationship in expression, mirroring their arrangement on the chromosomes, 5' to 3', and in the same physical order that they affect each parasegment phenotypically (Lewis, 1978), along the A-P axis (McGinnis and Krumlauf, 1992). This collinear organization is conserved in nearly all other metazoans (McGinnis, 1984a) leading to a vast array of investigations over the years to attempt to understand the importance of this organization, along with how the Hox genes function and how they are regulated. However, to date, no unambiguous

biological explanation has been found for the link between genomic organization of Hox genes on the chromosome and the evolutionary constraint of expression in the order of the segments the genes regulate on the animals body.

1.9 Hox Gene collinearity

A Drosophila embryo is composed of 14 parasegments by stage 5 of embryogenesis that form the larvae and adult segments from the posterior half of one of these parasegments combined with the anterior of its neighbor parasegment (Fig.1.3.1) (Lempradl and Ringrose, 2008). To give identity to each embryonic parasegment, the eight Hox genes work in various highly specific combinations at different regions of the developing embryo, tightly regulated along the A-P axis using *cis*-regulatory instructions. In 1990 a model was proposed whereby a 'Hox code' determined the specific morphologies of each vertebrae after studying chick and mice, as they have different morphologies during development, but homologous regulatory genes (Kessel and Gruss, 1991). This model proposed that the sequential activation of Hox genes, from the silenced state, allows for specific combinations of Hox genes to specify each segments identity, as they are activated along the A-P segments of the animals body plan. However, a few years later another group noticed that Hox genes expressed from the posterior of the embryo/complex, starting at *Abd-B*, were capable of repressing genes that were more anterior in Drosophila. So if there was no functional *Abd-B*, the larvae would develop several A4 segments, specified by *abd-A*. Also, *Ubx* was repressed by *abd-A*, and *Antennapedia* (*Antp*) was repressed by all three BX-C genes (Duboule and Morata, 1994). This functional hierarchy of the Hox complex was termed the posterior prevalence rule where the loss-of-function of a Hox gene leads to homeotic transformations of the segments normally regulated by that gene, into more anterior segments. This posterior prevalence control has also been demonstrated in mammals (Nolte, 2015) and has also been alluded to for the non-coding transcripts within the Hox complex (Gummalla et al., 2012; Yekta et al., 2008). However, the simple model in which posterior Hox genes repress more anterior Hox genes has been found to have exceptions. For example, *abd-A* and *Abd-B* have been shown to have distinct functions in the same histoblast nest cells and larval epithelial cells during the development of abdominal epithelia. One study found that *abd-A* was required for proliferation, positively regulating *wg*, and suppression of *Ubx* and *Abd-B* was required for identity of histoblast nest cells only (Singh and Mishra, 2014).

Almost all animals conserve the ancestral Hox complex organization. Arthropods have a single cluster of 10 Hox genes and most insects with wings have 8 as two of these genes lost their homeotic functions, *Hox3,* which evolved into *zen, zen2* and *bicoid* and *fushi tarazu* (*ftz*) (Negre et al., 2005). Although the Hox genes are arranged in clusters in most species and maintenance of this and the collinear expression throughout evolution suggests that preservation of this is fundamental.

Within the Drosophila lineage however, the Hox complex has been split in a number different places. Most Hox clusters in Drosophila are split between *Antp* and *Ubx,* but in *D. virilis* (Von Allmen et al., 1996) and *D. repleta* (Ranz et al., 2001) the split occurs between *Ubx* and *abd-A*. In *D. buzzattii* there are two splits, another one occurring between *labial* (*lab*) and *proboscipedia* (*pb*) (Negre et al., 2003) that relocated *lab* near *Abd-B* and *abd*-A, therefore altering the order of the Hox genes and no longer following the same collinearity rule. These differences in arrangement do not change the expression patterns or apparently the functions of the genes in Drosophila and the non-coding regulatory regions are maintained (Negre et al., 2005), leaving questions as to why the clustering has been so well preserved throughout most metazoans.

Temporal collinearity is seen in mice where the timing of gene expression matches the physical arrangement and the anterior of the complex is expressed first moving along the posterior in both time and space (Duboule, 1992). This model posits that embryos in organisms that form in sequential bilaterally paired segments along the neural tube of mesoderm, from the anterior to posterior (somites) will sequentially express Hox genes along the body axis. The Hox genes are thought to all start in an off state, grouped in a silencing complex, and progressively activate in groups being released from the silencing complex (Maeda and Karch, 2011). However, Drosophila Hox genes are thought to activate at the same time, as their segments develop simultaneously (long-germ band) and therefore do not require fully clustered Hox genes. This is further demonstrated by a short-germ band insect, *Tribolium casteneum* that develops segments sequentially and has an intact Hox cluster, possibly maintained by this need for temporal collinearity (Shippy et al., 2008).

Many *cis*-regulatory elements control the spatial and temporal Hox transcription patterns, also arranged on the chromosome in the same order of the body segments that they affect (Karch et al., 1985; Sanchez-Herrero and Akam, 1989). An alternative explanation for preservation of Hox genes in clusters is the preservation of these *cis*-regulatory regions. Mutations in these regions can cause loss-of-function phenotypes and transform one segment towards the adjacent anterior segment, whilst also transforming further anterior segments towards the posterior segment. This was demonstrated by Ed Lewis who found that loss-of-function of the *iab-4* regulatory region could cause both A4 to A3 and A2 to A3 transformations, suggesting that the *cis*-regulatory region responsible for A3 had become activated one segment too early in A2 (Lewis, 1985).

1.10 Homeotic mutations

Early work investigating the functions of the Hox proteins was mostly carried out using Drosophila as they visibly display mutant phenotypes correlated with Hox gene mutations or misregulation. The most famous of these transformations is the *Ubx* fly that has four wings instead of two as the third thoracic segment develops the same phenotype as the second (Lewis, 1978).

Investigations over the years have revealed that ectopic expression of Hox genes would cause homeotic transformations, indicative of their function (Pick and Heffer, 2012). Each TF encoded by a Hox gene can have a number of different target genes, including themselves, and act either by itself or in conjunction with other Hox genes (Pearson et al., 2005). One example displaying the versatility of Hox genes in Drosophila is the regulation of *decapentaplegic* (*dpp*) transcription that is activated by *Ubx* and repressed by *abd-A* (Capovilla and Botas, 1998) to keep the *dpp* protein in a localized region in the gut to trigger genes that go on to change these cells shape into the correct morphology (Bienz, 1994). In another context *Ubx* acts in conjunction and redundantly with *abd-A* to repress *Distal-less* (*Dll*) in the epidermis of the abdomen, another homeodomain TF that stimulates development of appendages and therefore needs to be restricted to specific cells in order to form limbs in the correct positions (Vachon et al., 1992). The switch in *Ubx* function between repression and activation could be an effect of a number of factors, for example the different cellular environments created by different tissue layers, as the effect on *Dll* occurs in the epidermis and on *dpp* the visceral mesoderm, providing different signals from both the A/P and D/V axes. *Ubx* is best known for its role in haltere and wing development, from the original studies that produced a mutant four wing fly (Lewis, 1978) and its roles in the control of the wings and legs are now better understood, but continuing efforts are being made to understand its vast array of molecular functions in the developing embryo.

Some winged insects have four wings and no haltere/balancer organs, whereas Drosophila have two wings on the second thoracic segment and two balancer organs on the third called halteres, thought to have developed from the hindwings of four-winged ancestors (Carroll et al., 1995). *Ubx* is expressed in haltere imaginal discs but not wing discs and is thought to be the master switch between haltere and wing development, keeping the numbers of cells in the haltere much lower than wing imaginal discs throughout embryonic and larval development (Roch and Akam, 2000). If *Ubx* is mutated the halteres are transformed into wings (Lewis, 1978) and ectopic expression causes the wings to transform into halteres (White, 1985). A better understanding of the molecular biology of this came from a study that measured the effects of *Ubx* on wing and haltere size by its regulation of *dpp*, a morphogen that is responsible for cell proliferation (Rogulja and Irvine, 2005), expressed in both wing and haltere imaginal discs. From the posterior compartment of the wing disc *engrailed* (en) stimulates *hh* (Zecca et al., 1995), which in turn induces *dpp* production from the central stripe of cells, known as the AP organizer (Basler and Struhl, 1994). Dpp is secreted into the wing disc and spreads in both anterior and posterior directions, which is thought to induce incorporation of more cells into the developing wing (Capdevila and Guerrero, 1994). In halteres there is a similar, but fainter, stripe of *dpp* expression in the same domain as wing discs that does not secrete into neighboring cells due to high levels of the Dpp receptor, *thickveins* (*tkv*). Tkv is expressed evenly throughout the haltere disc, increasing signal transduction and restricting diffusion of Dpp (Lecuit and Cohen, 1998). The wing disc has

low *tkv* expression in and around the AP organizer and high levels only in lateral regions. This allows *dpp* to dissipate from the center lower levels are only detected in regions where the *tkv* expression is higher (Crickmore and Mann, 2006). *Ubx* is able to regulate the activity of *dpp* from multiple approaches; 1) it can increase levels of *tkv*, therefore reducing *dpp* diffusion (Crickmore and Mann, 2006). 2) It can repress *scribbler* (AKA *master of thick veins* (*mtv*)), a repressor of *tkv*, allowing diffusion of Dpp. 3) Ubx and En can suppress *division abnormally delayed* (*dally*), a heparin sulfate proteoglycan (HSPG) that reduces *dpp* diffusion (Crickmore and Mann, 2007; de Navas et al., 2006). These studies found their manipulations of these components via mutations, and gain and loss-of-functions would cause reductions and increases of wing and haltere sizes ranging from 30-60%, leading them to believe that additional mechanisms were involved that would explain a full transformation of one to the other.

Ubx also acts in the abdomen of Drosophila and in conjunction with *abd-A* will prevent legs from forming through the repression of *Dll* in hexapods (Vachon et al., 1992). In other arthropods, such as crustaceans and onychophora, a different *Ubx* sequence at the C-terminus produces an altered *Ubx* that does not repress *Dll* allowing limbs to develop (Gebelein et al., 2002; Ronshaugen et al., 2002). *Dll* is a conserved homeobox gene that is required for leg development (Panganiban et al., 1997) and is only active in the thoracic segments of Drosophila. The repression of *Ubx* in the abdomen is aided by two Hox cofactors *Extradenticle* (*Exd*) and *Homothorax* (*Hth*) (Gebelein et al., 2002). *Exd* is a well characterized Hox gene cofactor that can expand the binding site from ~6 to ~10 bases (Mann and Chan, 1996) and has been shown to enhance binding of *Scr* to a *forkhead* (*fkh*) regulatory element (Ryoo and Mann, 1999). *Hth* interacts with *Exd* and the trimer of a Hox protein included with these two has been shown to be necessary for efficient binding to a natural Hox gene target enhancer (Ryoo et al., 1999). Although *Ubx* and *Antp* are closely related Hox genes, *Antp* does not repress *Dll* and is not enhanced by *Exd* and *Hth*, allowing leg formation from the thorax. A 304bp cis-regulatory element was identified for Ubx binding that encoded thoracic Dll expression and abdominal Dll suppression, named the Dll304 enhancer. Within the Dll304 element, binding sites for the *Ubx-Hth-Exd* trimer were distinguished from a region termed the Distalless repression element (DllR). The DllR can repress enhancer activity leading to repression of *Dll* in the abdomen (Gebelein et al., 2002). In this study they also found that only a specific subset of *Ubx* isoforms contained the linker domain (UbxIa) required for repression, demonstrating further complexities of *Ubx*. Another experiment analyzed the *in vivo* effects of *Ubx* isoforms on *dpp* in the mesoderm during embryogenesis and found that they had different DNA binding abilities with its cofactor *Exd*, possibly as a result of altering the distance between the DNA binding homeodomain and cofactor interaction motif (Reed et al., 2010). *Ubx* is coexpressed with *Dll* in some animals that do have legs on their abdomen and it is thought that there was a divergence in *Ubx* function of specific isoforms in the hexapod lineage in the C-terminal region (Galant and Carroll, 2002; Ronshaugen et al., 2002). In the T3 leg of insects *Ubx*

and *Dll* are both expressed, but *Ubx* does not prevent leg formation. This is circumnavigated by a delay in *Ubx* expression, allowing *Dll* to activate its own autoregulatory enhancer, which prevents late *Ubx* expression form interfering with the legs development (Estella and Mann, 2008; Galindo et al., 2011). These experiments demonstrate the massive complexity and diversity exhibited by just a single Hox gene in a subset of functions within a small time period of embryogenesis. *Ubx* functions in other tissues and has altered roles during different times of development that also have to be tightly regulated for an organism to produce all the correct appendages and organs in the correct places; a principle that is extrapolated to the other Hox genes throughout the animal kingdom (Pearson et al., 2005).

1.11 Upstream regulation of the Hox Complex by segmentation genes

It is now known that Hox expression domains in Drosophila are established by earlier expression of segmentation genes (Irish et al., 1989) and interactions between Hox genes themselves (Morata and Kerridge, 1982; Struhl, 1982a). This expression is subsequently maintained by chromatin remodeling (Simon et al., 1990). Prior to Hox gene activation and before the zygotic genome is activated, polarity is established by maternally deposited RNAs partitioning into broad regions of the oocyte (Fig.1.3) (Bull, 1966; Kalthoff, 1971; Nusslein-Volhard and Wieschaus, 1980; Sander, 1975; Yajima, 1964), diffusing from the anterior and posterior tips through binding to the cytoskeleton (Koch and Spitzer, 1983; Pokrywka, 1995). Maternally deposited mRNAs are translated into proteins upon fertilization, creating gradients (Driever and Nusslein-Volhard, 1988) and therefore act as morphogens that regulate translation of themselves as well as transcription of zygotic genes (Driever and Nusslein-Volhard, 1989; Markussen et al., 1995; Rongo et al., 1995). Most of these discoveries were made in insects, primarily *D. melanogaster,* where these maternal genes regulate gap gene expression to segment the embryo into four broad domains along the A-P axis (Cohen and Jurgens, 1990; Driever et al., 1989; Finkelstein and Perrimon, 1990; Hulskamp et al., 1990).

In Drosophila, there are four main gap genes, *Krüppel (Kr), hunchback (hb), knirps (kni)* and *giant (gt).* These proteins are sequence-specific TFs that, if mutated, leave gaps in the body plan of a developing embryo in the region that is it usually expressed in (Knipple et al., 1985; Nusslein-Volhard and Wieschaus, 1980; Stanojevic et al., 1989). Gap genes regulate themselves to determine specific boundaries and go on to cooperate with products of the maternal effect genes to regulate four early pair-rule genes, *ftz, even-skipped (eve), runt (run)* & *hairy (h)* (Fig.1.3) (Carroll and Scott, 1986; Ingham et al., 1986; Nusslein-Volhard and Wieschaus, 1980). The pair-rule genes are expressed in seven, narrow, distinctive lateral stripes, and encode further TFs that control their own activity as well as the segmentation genes (Fig.1.3) (DiNardo and O'Farrell, 1987; Nusslein-Volhard and Wieschaus, 1980; Pankratz and Jackle, 1990).

Segmentation genes are expressed in 14 stripes along the A-P axis forming parasegments (Martinez-Arias and Lawrence, 1985) and together with gap and pair rule genes control the expression of Hox genes (Carroll et al., 1988) (Fig.1.3). Activation of particular combinations of Hox genes specifies the unique identity of each segment and are responsible for generating the structures that form along the A-P axis of all bilateria (Carrasco et al., 1985; Lewis, 1978; McGinnis and Krumlauf, 1992). It is the activities of these early A-P TFs, as well as those expressed concurrently dorsal-ventrally (D-V), that contributes to, and then specifies, each cells lineage (Guo et al., 2010; Lawrence and Struhl, 1996; Morata and Lawrence, 1975). Furthermore, mutations, or lack of strict spatiotemporal regulation, in any of these TF's are capable of causing loss of embryonic viability or mild to severe developmental defects (reviewed (St Johnston and Nusslein-Volhard, 1992).

Regulation of transcription can only be carried out by TFs if epigenetic modifications reorganize chromatin to allow physical accessibility of transcription machinery to genes. These epigenetic modifications maintain activity or repression of genes by keeping chromatin in open or closed states, and these signals can be retained by daughter cells throughout mitosis. In recent years, it has been discovered that lncRNAs are a requirement for the regulation of a number of genes that are also epigenetically regulated by Polycomb group (PcG) and Trithorax group (TrxG), throughout many evolutionarily divergent species (Mallo and Alonso, 2013; Morris and Mattick, 2014; Steffen and Ringrose, 2014). PcG and TrxG proteins were discovered regulating the Hox complex in *D. melanogaster (Lewis, 1978)*, a region that we now know is enriched in transcription of lncRNAs that have been shown to regulate Hox genes (Lemons and McGinnis, 2006; Mallo and Alonso, 2013). It would stand to reason that the lncRNAs of the Hox complex would also need to be spatiotemporally regulated in order to function at the correct time during development.

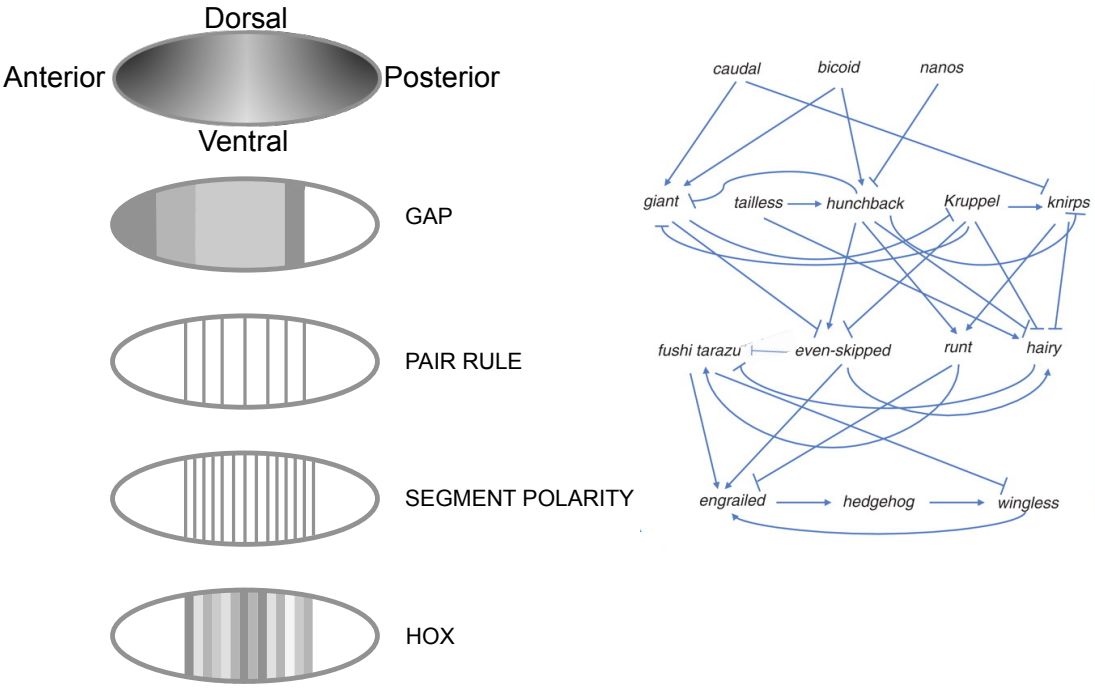


Figure 1.3. Expression and regulation of maternal gradient and segmentation genes in establishment of A-P segments and Hox gene activation. Maternal gradients establish A-P and D-V axis in an egg and regulate the expression of downstream genes, until 14 parasegments are established and Hox genes are active. Figure adapted from (Carroll et al., 2009).

1.12 *Cis* regulation of Hox Genes

The association of lncRNAs with chromatin modifying complexes has been primarily examined in embryonic development, when it is crucial that gene expression is tightly regulated to ensure correct specification of cell and tissue types (Koziol and Rinn, 2010). This is particularly apparent in the clusters of developmental genes that encode the Hox transcription factors. Two protein complexes with opposing functions in transcriptional activation and silencing are responsible for maintaining Hox genes in an on or off state, the TrxG and PcG proteins respectively (Ingham, 1980; Lewis, 1978). It is now recognized that ncRNA plays a vital role in Hox gene regulation, although the transcripts are still being uncovered and the mechanisms by which they exert their control are yet to be fully understood (Pauli et al., 2011).

Many *cis*-regulatory elements have now been identified that regulate the spatial expression of Hox genes in developing embryos, both within the Hox cluster and from distant elements on the chromosome (Akbari et al., 2006; Mallo and Alonso, 2013). The *cis*-regulatory spatial control of Hox genes in *D. melanogaster* matches the collinear arrangement on the chromosome (Maeda and Karch, 2006; Martin et al., 1995), consistent with their Hox gene arrangement. However, in species that have multiple smaller Hox clusters, many of the *cis*-regulatory elements are not within the clusters and function distantly to maintain spatially collinear expression of Hox genes in the correct segments of the developing embryo (Herault et al., 1998; Negre et al., 2005; Seo et al.,

2004). The *cis*-regulatory elements in the BX-C of Drosophila are some of the first identified and highly studied. The regions controlling *Ubx* are *abx/bx* and *bxd/pbx*, then regions termed *iab-2/3/4* control *abd-A* expression and *iab-5/6/7* control *Abd-B* (Martin et al., 1995). These are arranged into parasegment specific domains and provided an explanation as to how three Hox genes were able to provide specific identities for nine segments of the developing embryo (Fig.1.4) (Castelli-Gair and Akam, 1995; Martin et al., 1995). The TFs responsible for activation of Hox genes and dividing the embryo into 14 parasegments, the maternal and segmentation genes, bind to these regulatory elements in different specific combinations to initiate particular Hox gene activities in individual parasegments (Casares and Sanchez-Herrero, 1995; Kornberg and Tabata, 1993; Shimell et al., 1994). These early TFs are rapidly degraded and the regulatory elements are bound by PcG and TrxG proteins that maintain Hox gene expression through deposition of chemical modifications on histone tails to reorganize chromatin (Maeda and Karch, 2006).

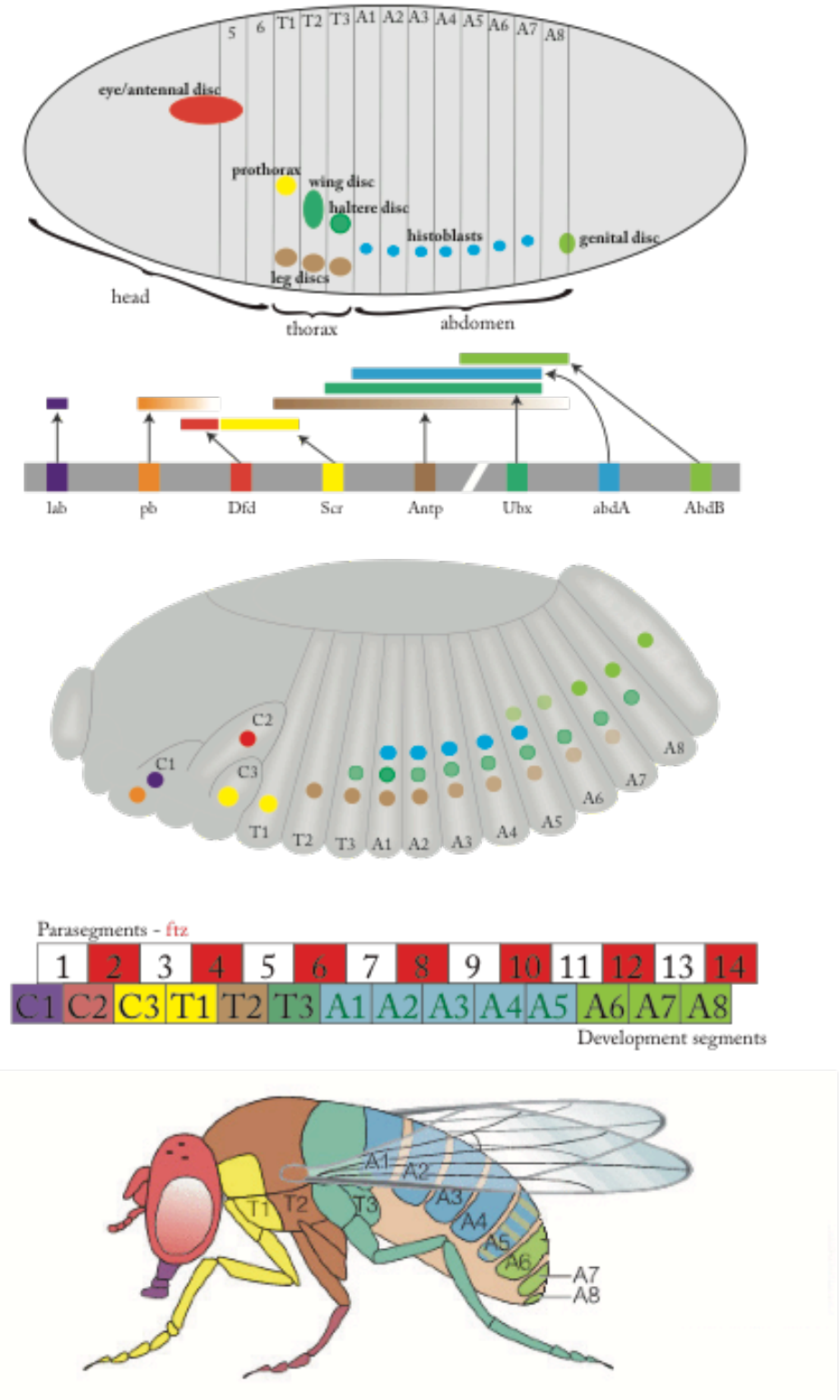


Figure 1.4. Hox gene expression through embryogenesis and adult flies. Arrangement of future discs in blastoderm embryo and Hox gene expression patterns in both early and late stage embryos (~stage 5 and 17 respectively). Parasegments 1-14, even numbers in red show where *ftz* is expressed and how this aligns to the developmental segments and finally the adult fly with matching colors for gene expression. Figure adapted from; 2008 Sinauer Associates Sadava, D. et al. Life: The Science of Biology, 8th ed; Molecular and Developmental Biology Course, Dr.Brian E. Stavely, Memorial University of Newfoundland; Atlas of Drosophila Development, Volker Hartenstein, Epidermis pgs. 24-25 (www.sdbonline.org).

1.13 Polycomb and trithorax proteins

Many of the biological functions we now understand for lncRNAs, and best understood in terms of their specific binding and mechanisms of gene regulation, is through their association with polycomb proteins, most frequently with PRC2. The polycomb proteins are known for their ability to maintain repression of target genes, whereas trithorax for its antagonistic effect maintaining gene expression, frequently during embryogenesis and growth (Geisler and Paro, 2015). In 1981, mutations in the *extra sex combs* (*esc*) gene resulted in altered Hox gene expression, mutating flies so the thorax and anterior of the abdomen transformed into more posterior segments (Struhl, 1981). In 1982, *polycomb-like* (*pcl*) was recognized for its silencing role in the maintenance of normal Drosophila segmental identities that resembled those caused by *Pc* and therefore expected to negatively regulate Hox genes (Duncan, 1982). By 1985 20 genes had been identified that regulated the spatial expression of Hox genes, with partial transformations found in single mutants, but strong homeotic phenotypes when more than 2 were mutated, suggesting the genes has some dependence upon each other (Jurgens, 1985). The group of genes that caused similar homeotic phenotypes as those first identified in *Pc* mutants that were attributed to lack of repression, they all became known as PcG (Jurgens, 1985).

At the same time PcG proteins were being investigated, another mutation causing homeotic phenotypes of the abdomen and thorax was identified that could result in flies having up to six wings and therefore became known as *Trx* (Ingham, 1980). This was later discovered to work in concert with another group of proteins termed TrxG that were able to counteract PcG silencing marks and maintain active transcription of target genes (Kennison and Tamkun, 1988). Initially the PcG and TrxG proteins were thought to instigate gene activation or repression, but it is now understood that they maintain these states through chemical modifications of histones that subsequently alter the structure of chromatin into open or compact conformations (Luger et al., 1997a; Luger et al., 1997b). Many of the homologs of the PcG and TrxG proteins have now been identified in vastly evolutionary divergent multicellular organisms from plants through to mammals, along with the PRC1 and PRC2 complexes and the epigenetics marks associated with them, suggesting that this method of transcriptional control is likely to be an ancient mechanism (Schuettengruber et al., 2007). However, the reasonably straightforward relationships that have been elucidated for the core components of the different PcG and TrxG complexes (Table.1.3.1) has become immensely complex as novel interacting partners are continually identified that earn them the classification of PcG or TrxG proteins, whilst also remaining established members other regulatory gene groups (Geisler and Paro, 2015; Schwartz and Pirrotta, 2013). It is now thought that the multifaceted composition of each complex enables specific actions and gene targeting that may be influenced by protein components interacting with a complex, or lncRNA molecules (Brockdorff, 2013; Geisler and Paro, 2015; Mondal and Kanduri, 2013). Furthermore, much of the

original and continuing investigations into PcG and TrxG proteins has been based on studies of Hox gene regulation. Recently however, whole genome binding profiles have revealed hundreds of other targets that mostly consist of developmental genes (Mendenhall and Bernstein, 2008; Ringrose, 2007).

1.14 Polycomb and trithorax complexes

Since initial studies identified PcG and TrxG proteins as regulators of Hox genes, genome-wide profiling technology has advanced considerably. This has enabled us to detect protein binding partners and DNA targets using immunoprecipitation and sequencing studies, revealing core complexes formed by both TrxG and PcG (Klymenko et al., 2006; Negre et al., 2006; Saurin et al., 2001; Schuettengruber et al., 2011; Schwartz, 2006; Slattery et al., 2014). Further biochemical analysis has characterized large multimeric complexes formed by PcG and TrxG proteins in chromatin. Polycomb proteins consist of 2 core complexes that have been conserved throughout metazoans, Polycomb repressive complex 1 (PRC1) (Franke et al., 1992; Shao et al., 1999), PRC2 (Czermin et al., 2002; Muller et al., 2002) and in Drosophila another complex, pleiohomeotic repressive complex (phoRC) (Klymenko et al., 2006). The mammalian homolog of *Pho* (*Ying-Yang*, *YY1*) is able to act as a ubiquitously expressed transcriptional repressor and activator via protein-protein and protein-DNA interactions through multiple mechanisms, including an interaction with a histone deacetylase (RPD3), but genomic binding profiles of *YY1* and PcG proteins does not shown them biding to the same regions across genome and so *YY1s* role in PcG-mediated repression is yet to be established in mammals (Mendenhall et al., 2010; Vella et al., 2012) (Table.1.3.1).

The PRC1 complex of Drosophila contains core components *polyhomeotic* (*ph*), *Pc*, *Posterior sex combs* (*Psc*) and *Sex combs extra* (*Sce*) and its function is to recognize H3K27me3 by the chromodomain of *Pc*, anchoring the complex to chromatin, and to monoubiquitinate H2AK118 to compact chromatin and potentially stall RNA pol II (Geisler and Paro, 2015). The monoubiquitination is carried out by *Sce* and its activity is enhanced by forming a heterodimer with *Psc*. *Sex combs on midleg* (*Scm*) is also considered a core component of PRC1, although its exact molecular function as a PcG protein is not understood. *Scm* is a transcriptional repressor shown to interact with *ph* through its SPM domain and colocalise at polytene chromosome sites *in vivo* (Peterson et al., 1997). It contains multiple malignant brain tumor (MBT) repeats and its a human homolog (*Scm-like2*) is implicated in malignant brain tumors (Santiveri et al., 2008). Loss-of function experiments of *Scm* has shown that it is essential for cell survival and ommatidia development (Guo and Jin, 2015) as well as being linked to death and serious homeotic transformations (Gaytan de Ayala Alonso et al., 2007). Interestingly, a component of a TrxG complex has recently been linked to the regulation of *Scm*. The gene *will die slowly* (*wds*) is a

component of the trithorax related complex (TRR Complex) that is thought to position the N-terminus of H3K4 for efficient trimethylation. Based on findings of its homologous counterpart, WDR5 in humans, it is likely to be a subunit of a TrxG complex, mixed lineage leukemia 1 and 2 (MLL1/2) (Couture et al., 2006). In mammals, the WDR5 protein was found to colocalise with MLL1 in the HOXA cluster in human fibroblast cells, specifically at the TSS of multiple Hox genes in this cluster. The investigators found that knockdown of *HOTTIP* RNA, a 3,764nt spliced and polyadenylated lncRNA, led to a broad loss of H3K4me2/3 across the HOXA locus and a reduction in MLL1 and WDR5 binding to TSS of HOXA genes, plus an increase of MLL1 and WDR5 binding to the *HOTTIP* locus. Interestingly, the *HOTTIP* locus itself has bivalent chromatin marks, H3K4me3 and H3K27me3, along with both PRC2 and MLL complexes binding to it, indicative of noncoding regulatory sequences that are poised for activation (Bernstein et al., 2006). As part of this study, *in vitro* transcribed *HOTTIP* was bound to WDR5 and immunoprecipitation of WDR5 retrieved the *HOTTIP* RNA leading the researchers to determine that *HOTTIP* directly binds WDR5 as an adapter protein to target the WDR5-MLL1 complex to HOXA genes to achieve H3K4me3 and transcription of target genes (Wang et al., 2011b).

The PRC2 complex of Drosophila is responsible for catalyzing trimethylation of H3K27me3, the repressive mark that recruits PRC1, carried out by the SET domain of its *Enhancer of zeste* (*E(z)*) component (Czermin et al., 2002). *E(z)* has very little HMTase activity without inputs from the subunits *Suppressor of zeste 12* (*Su(z)12*) (Cao and Zhang, 2004), *Chromatin assembly factor 1, p55 subunit* (*Caf1-p55*) (Nekrasov, 2005) and *esc* (Ketel et al., 2005). *Su(z)12* and *Caf1-p55* are thought to anchor *E(z)* to chromatin and *Esc* increases the HMTase activity (Nekrasov, 2005). Other PcG proteins interact with PRC2, thought to modulate enzyme activity or guide the complex to specific genomic sites, with 2 key proteins being *Jumonji, AT rich interactive domain 2* (*Jarid2*) and *Pcl*. These create distinct PRC2 complexes that are thought to enhance silencing and guide the PRC2 components to specific sites. *Pcl* has been shown to increase H3K27me3 in flies (Nekrasov et al., 2007), humans (Cao et al., 2008; Sarma et al., 2008) and mice (Walker et al., 2010) and loss-of-function leads to derepression of Hox genes (Duncan, 1982) even though H3K27me3 is still present at lower levels and PRC1 and PhoRC binding to PREs is not affected (Nekrasov et al., 2007). *Jarid2* can increase *E(z)* enzyme activity of H3K27me2/3 through its N-terminal and knockdown resulted in reduced PRC2 at target promoters, thus *Jarid2* is thought to bind DNA through it's C-terminus and recruit PRC2 to specific promoters (Li et al., 2010). *Pcl* and *Su(z)12* have also been found colocalized at PREs regulating *Ubx* and *Abd-B* and loss of *Pcl* results in a reduction in *Su(z)12* binding along with H3K27me3 indicating *Pcl* stabilizes PRC2 to specific PREs (Nekrasov et al., 2007). The *Pcl* containing PRC2 complex differs between embryonic and larval stages as embryonic *Pcl* containing complexes co-fractionate with *E(z)*, whereas in larvae *Pcl* does not associate with *E(z)* and instead forms a different complex (Savla et al., 2008). This novel Pcl-complex was investigated in wing imaginal discs using ChIP, where PcG

proteins maintain silencing of *Ubx* through components of *Pho*, PRC1 and PRC2 (including *E(z)*) binding to the *cis*-regulatory *bxd* PRE (Papp and Muller, 2006; Savla et al., 2008). One study investigated recruitment of this novel complex and which other PcG proteins were required for it to bind and function. As *E(z)* typically associates with *Pcl* they tested loss-of-function *E(z)* and found there was no difference in *Pcl* binding to the *bxd*-PRE, but H3K27me3 and *Pc* was lost, showing that the Pcl-complex was distinct from typical PRC1 and PRC2 and did not require *E(z)* to bind (Savla et al., 2008). *Pho* or *phol* were required for *Pcl bxd*-PRE binding in wing imaginal discs, demonstrating that the PhoRC complex acts as a general PcG complex recruiter to PREs (Mohd-Sarip et al., 2002). The experiments carried out in wing imaginal discs show that PhoRC is required for Pcl-complex to bind in larvae and is necessary for PRC2 and therefore PRC1 recruitment to chromosomal sites, reflecting a possible tissue and/or target site specificity and/or a role at particular developmental time periods. The need for unique complexes functioning at different times in different tissues, targeting different genes is likely to be widespread throughout different species development, becoming more complex with the complexity of the organism, requiring a much deeper understanding than is currently available.

Another distinct PcG complex that has been conserved in Drosophila and mammals can be formed with histone deacetylase 1 (HDAC1), *Caf1-55* and *esc* (Tie et al., 2001). The product of *esc* binds directly to the PcG protein *E(z)* (Jones et al., 1998) as well as colocalizing on chromosomes (Tie et al., 1998) and is essential for PcG silencing in the first 6 hours of embryogenesis, but not for later maintenance (Struhl, 1982b). *Caf1-55* is a histone binding protein that is also part of the NURF TrxG complex (Table.1.3.1) and the chromatin assembly complex 1. HDAC1 was shown to be essential for PRE/PcG mediated silencing of *Ubx*, where *E(z)* is also bound and therefore likely to act together (Tie et al., 2001). Histone deacetylation in yeast is restricted to a few nucleosomes from the site it is recruited to and HDAC1 can acetylate all four histone tails (Kurdistani and Grunstein, 2003). Consistent with this evidence is the finding that PRE-silencing disruption is followed by the activating marks of hyperacetylated H4 (Cavalli and Paro, 1999), thereby linking deacetylation by HDACs with silencing. HDAC1 in Drosophila can also interact with *Groucho* to regulate segmentation genes during early embryogenesis (Chen et al., 1999) and has been directly linked to regulation of other Hox genes, such as *Scr*, via interaction with PcG proteins (Chang et al., 2001). Also, in mammals the homolog of *pho*, *YY1*, recruits the homolog of HDAC1, (mammalian HDAC1 and HDAC2 are derived from Drosophila HDAC1) (Yang et al., 1996). More recently, the Drosophila paralog of HDAC1, HDAC3, was found to suppress apoptosis in Drosophila imaginal tissue and mutations in either caused dominant suppression of position effect variegation, linking them further to chromatin organization (Zhu et al., 2008).

In Drosophila 4 core trithorax protein complexes have been identified, Brahma associated protein complex (BAP) (Dingwall et al., 1995), nucleosome remodeling factor (NURF) complex (Badenhorst et al., 2002), trithorax acetylation complex (TAC1) (Petruk, 2006) and absent, small,

or homeotic discs 1 complex (Ash1) (Bantignies et al., 2000) (Table.1.3.1). The first mammalian homolog of *trx* was identified by its random translocations associated with human leukemias and subsequently named *mixed lineage leukemia* (*MLL*) (Tkachuk et al., 1992; Ziemin-van der Poel et al., 1991), and mammalian equivalent of TAC1 is the MLL complex. Many of the TrxG proteins have now been found in other diverse species, although in fewer cases in the same complexes. For example the ATP-dependent chromatin remodeling complex NURF can be found in mammals, *C.elegans* and plants, whereas the BAP and Ash1 complexes have so far been identified in mammals, but not *C.elegans* or plants. TAC1 is found in Drosophila alone (Schuettengruber et al., 2011). Although there is less evidence of TrxG complexes in other species, the mechanism they use is considered to be conserved due to identification of the same active epigenetic marks found associated with gene expression. Generally, H3K4 and H3K36 methylation accompany active chromatin and H3K9, H3K27 and H4K20 methylation leads to repressed chromatin (Fuchs et al., 2006) carried out by the SET domain of histone methyltransferases (HMTases). Set1 was first identified in *S. cerevisiae* and demonstrated an ability to methylate H3K4 when in complex with the COMPASS complex; it was later discovered to be the homolog of mammalian *MLL*, aiding in the understanding of *MLL* functional analysis (Gu et al., 1992; Miller et al., 2001). It was also found that Set1 alone could not catalyze methylation and the other subunits of the COMPASS complex were necessary for assembly and regulation of methylation patterns (Miller et al., 2001; Schneider et al., 2005).

The SET containing HMTase genes in Drosophila are *Ash1, CG4565, CG32732, eggless, G9a, Histone methyltransferase 4-20, Nuclear receptor binding SET domain protein, Set1, Set2, Set3, Suppressor of variegation 3-9* (*Su(var)3-9*), *E(z)*, *trithorax-related* (*trr*) and *trx* (Dillon et al., 2005; Mis et al., 2006; Schotta et al., 2004; Shilatifard, 2012). *Set1, trx* and *trr* are homologs of yeast *Set1* and are found interacting with unique complexes composed of homologs of COMPASS subunits that methylate H3K4, with mammalian orthologs identified for each (Mohan et al., 2011). The Drosophila *Set1* is a direct yeast ortholog and *trx* and *trr* are more distantly related, but loss of any one of these is lethal to Drosophila. This suggests that the loss of methylation from any one of the *Set1, trr* or *trx* COMPASS complexes has specialized functions during development (Mohan et al., 2011). Drosophila *Set1* is maternally deposited as part of a COMPASS-like complex that globally catalyzes H3K4me2/3 and has been shown to be required for the completion of later developmental stages (Ardehali et al., 2011). However, the *trr* and *trx* genes are thought to have developed more specific gene targets as the human homologs, *MLL1/2* and *MLL3/4* respectively, do not have overlapping targets (Eissenberg and Shilatifard, 2010). Humans have 6 homologs of yeast Set1, *Set1A, Set1B* (corresponding to Drosophila *Set1*), *MLL1, MLL2* (in Drosophila *trx*), *MLL3* and *MLL4* (in Drosophila *trr*) (Eissenberg and Shilatifard, 2010), all of which have been found in human equivalent COMPASS complexes responsible for H3K4me1/2/3, containing

subunits that have been conserved from yeast, indicating an ancient origin for this essential method of gene regulation (Shilatifard, 2012).

To date the exact mechanism of how TrxG inhibits PcG is not completely understood, especially as both PcG and TrxG proteins bind to PRE/TREs in both active and silent states and can be reprogrammed at specific points in development, whilst both remaining bound (Steffen and Ringrose, 2014). The main exception is the TrxG proteins, *Ash1*, which binds only to active TSSs regulated by PREs and prevents silencing marks in the promoter and coding region (Papp and Muller, 2006). It has been clearly demonstrated that the H3K36me2/3 marks appear to significantly reduce binding of the PRC2 subunit *Caf1-55* (Schmitges et al., 2011). This is now better understood by characterization of another TrxG protein, *kismet* (*kis*), a gene that was previously linked to gene activation and acts as an antagonist of PcG proteins (Dorighi and Tamkun, 2013). *Kis* is necessary for *Ash1* and *trx* recruitment (Srinivasan et al., 2008) and loss-of-function leads to increased PcG H3K27me3 and reduced H3K36me2 (Dorighi and Tamkun, 2013). However, these proteins have not been found to physically interact, so although *kis* is needed for *Ash1* recruitment and subsequent H3K36me2 to inhibit H3K27me3, there appears to be a factor not yet accounted for in this recruitment mechanism. One key finding was when *Ash1* was shown to physically interact with the lncRNA, *D4Z4 binding element-transcript* (*DBE-T*) that has been associated with facioscapulohumeral muscular dystrophy (FSHD) (Cabianca et al., 2012). The D4Z4 is a repeat that is thought to function in a similar way to a PRE/TRE as it was bound by PcG proteins and *Ash1* was recruited via its SET domain by the lncRNA, promoting target gene expression. Further work on *Ash1, kis* and lncRNAs could reveal if lncRNAs are acting as the scaffold between *kis* and *Ash1*, in a similar mechanism previously seen for the lncRNA *HOTAIR* (Tsai et al., 2010).

Table 1.1. PcG and TrxG complexes

PcG			
Complex	**Protein**	**Function**	**Ref**
PhoRC Pho repressive complex Can bind PREs and recruit PRC1, PRC2 and components of SWI/SNF	**Sfmbt** Scm-related gene containing four mbt domains	Essential for Hox repression in Drosophila, dependent on Pho binding sites. MBT repeats bind H3K9me1/2 and H4K20me1/2 but not H3K9me1 or H4K20me3 or unmethylated. The interaction with methylated histones when bound to PRE maintains repression	(Klymenko et al., 2006)
	pho pleiohomeotic	Sequence specific DNA binding protein that tethers MBT domains of Sfmbt essential for Polycomb repression	(Alfieri et al., 2013; Brown et al., 1998)
	phol pleiohomeotic-like	Sequence specific DNA binding protein able to replace pho functions binding to same sequence	(Brown et al., 2003)
PRC2 Polycomb repressive complex 2 Recruits PRC1 by catalyzing trimethylation of H3K27	**E(z)** Enhancer of zeste	SET domain methylates H3K9 and H3K27 leading to transcriptional repression. Specifically required for repression of Hox genes during first 6hrs of embryogenesis	(Czermin et al., 2002; Kuzmichev, 2002; Simon, 1995; Tie et al., 2001; Tschiersch, 1994)
	esc extra sex combs	Specifically required for repression of Hox genes during first 6hrs of embryogenesis. Interacts with E(z)	(Simon, 1995; Struhl, 1982b)
	Su(z)12 Suppressor of zeste 12	Essential for H3K27me3 in rate limiting manner and E(z) cofactor. Mutations cause strong homeotic phenotypes	(Birve et al., 2001; Chen, 2008)
	Caf1-55 Chromatin assembly factor 1, p55 subunit	Histone chaperone protein that interacts with histone H4. Essential for cell proliferation and viability. Necessary for binding PRC2 in *vitro*	(Anderson et al., 2011; Nekrasov, 2005; Roth, 1996)
	Pcl Polycomb-like	Specifically required for repression of Hox genes during first 6hrs of embryogenesis. Needed for high levels of H3K27me3	(Nekrasov et al., 2007)
	Jarid2 Jumonji, AT-rich interactive domain	Associates with all known PRC2 components and mutants affect H3K27 methylation. Required for transcriptional repression	(Herz et al., 2012)
PRC1 Polycomb repressive complex 1 Binds H3K27me3 (brought about by PRC2) and monoubiquitinates H2AK119 leading to chromatin compaction, RNA pol II stalling and transcriptional silencing	**Sce** (dRing1) Sex combs extra	E3 ubiquitin ligase mediates monoubiquitination H2AK118 – tag for transcriptional repression	(Wang et al., 2004a)
	Pc Polycomb	Contains chromodomain that recognizes and binds H3K27me3. Able to inhibit histone acetylation by CREB-binding protein	(Cao et al., 2002; Messmer, 1992; Tie et al., 2016)
	ph polyhomeotic	Interacts with SPM domain of Scm thought to mediate self-binding and be involved in an autoregulatory loop	(Fauvarque, 1995; Peterson et al., 1997)
	Psc Posterior sex combs	Inhibition of remodeling and transcription	(King, 2005)
	Scm Sex comb on midleg	Interacts with *ph* through SPM domain. Transcriptional repressor necessary for PcG silencing, important for cell survival and ommatidium development	(Guo and Jin, 2015)

Table 1.1. PcG and TrxG complexes continued

TrxG			
PBAP/BAP Polybromo-Containing / Brahma Associated Proteins Complex ATP-dependent. Binds acetylated histones via bromodomain and remodel chromatin	**brm** brahma	Able to suppress PcG-mediated homeotic transformations. Zeste dependent recruitment to TREs upon activation	(Dejardin and Cavalli, 2004; Dingwall et al., 1995; Kennison and Tamkun, 1988)
	mor moira	Binds brm, shown to regulate Hox genes	(Crosby, 1999; Kennison and Tamkun, 1988)
	osa osa	Non-specific DNA binding, recruited by zeste. Interacts with brm, shown to regulate *AntpP2*. Required for segmentation	(Kal et al., 2000; Vázquez, 1999)
	Snr1 Snf5-related 1	Physically interacts with trx. Shown as positive regulator Hox genes	(Rozenblatt-Rosen et al., 1998; Zraly, 2003)
NURF Nucleosome Remodeling Factor ATP dependent (Iswi-SNF2L) chromatin remodeling complex, facilitating transcription. that recognizes H3K4me3 mark by Nurf-301/BPTF	**Iswi** Imitation SWI	Energy transducing component for nucleosome sliding and counteraction of repression	(Tsukiyama, 1995)
	Nurf-38 Nucleosome remodeling factor 38	Catalyzes incorporation nucleotides into growing chain. Shows binding to *Trl*	(Kugler and Nagel, 2010; Xiao, 2001)
	Caf1	See above	
	E(bx) Enhancer of bithorax	Reads H3K4me3 via PHD finger. Thought to recruit complex to specific genes. Needed for efficient and accurate nucleosome sliding. Positive regulator bithorax complex	(Badenhorst et al., 2002; Xiao, 2001)
TAC1 Trithorax Acetylation Complex 1 Possesses H3K4 methyltransferase activities	**trx** trithorax	Histone methyltransferase. C-terminal SET domain methylates H3K4. N-terminal required for H3K27 acetylation by CBP	(Klymenko and Muller, 2004; Tie et al., 2014; Tie et al., 2009)
	nej nejire	Lysine acetyltransferase, established role as histone acetylase. Mutations shown to reduce expression *Ubx*.	(Petruk, 2001)
	Sbf SET domain binding factor	Found in complex, closely linked physically in Hox maintenance	(Petruk, 2001)
Ash1	**ash1** absent, small, or homeotic discs 1	SET domain methylates H3K36. Interacts with trx. Antagonizes PcG repression	(Klymenko and Muller, 2004; Rozovskaia et al., 1999)
	nej	See above	

1.15 Polycomb and trithorax complex recruitment to response elements

The chemical modifications catalyzed by PcG and TrxG for maintenance of silencing or activation can remain stable over many cell divisions, providing an epigenetic memory essential to the cell identity. They carry out this role through binding to DNA *cis*-regulatory elements, termed Polycomb Response Elements (PREs) or Trithorax Response Elements (TREs), as it was originally thought that the PcG complex binds to maintain repression or TrxG to maintain activation, or that all are repressed until TrxG counteracts the silencing (Geisler and Paro, 2015; Klymenko and Muller, 2004). However, it was subsequently found that characteristic marks for repression, H3K27me3, and activation, H3K4me3, along with TrxG and PcG proteins both colocalized at these regulatory elements regardless of transcriptional activity (Beisel et al., 2007; Enderle et al., 2011; Papp and Muller, 2006). However, these experiments have been carried out in whole embryos or heterogenic tissue so this is still not clear if this is true within a single cell at the same locus. Therefore, it is still not fully understood how PRE/TREs function and it seems likely that there could be different classes based on variable results during investigations into their mechanisms.

A popular early theory was that repression mediated by PREs could be reversed by transcription through the sequence, at certain points in development when the gene was required, allowing activation or re-activation of the target gene, maintained throughout adulthood. These models of function come from experiments investigating a PRE called *Frontabdominal-7* (*Fab-7*), located between *Abd-B* and *abd-A* in the Hox complex of *D. melanogaster*, that had been shown to have both Pc and GAGA factor (expressed from the *trl* gene) bound *in vivo* (Strutt et al., 1997) that had previously been found to regulate *Abd-B's* expression (Busturia and Bienz, 1993). Cavalli and Paro (Cavalli and Paro, 1998) transgenically cloned a regulatory fragment termed *Fab-7* that silenced a flanking UAS-mini-white reporter as observed by the prevention of Gal4 binding or activity, a method that they had previously shown to be efficient when Pc was bound (Zink and Paro, 1995). They tested the effect of a single copy *Trl* loss-of-function mutant allele and found silencing from *Fab-7* increased, measured by a decrease in eye pigmentation that was inherited by progeny, leading them to believe that wild-type *Trl* counteracts *Fab-7* silencing by PcG proteins and was stably inherited (Cavalli and Paro, 1998). The investigators had already established that PcG proteins, Pc, Ph and Psc could be displaced by activation of Gal4 transcribing through the construct, thereby activating *lacZ* (Zink and Paro, 1995) and so measured the effects of short pulses of transcription, at different stages of development, through *Fab-7* to release the PcG proteins and measure mini-white expression. They found that 70% of adult flies had red eyes, although not 'completely uniform', and quantification showed a 2.5% increase in eye pigmentation compared to controls, if Gal4 was induced in embryos, not larvae. They also found the silencing effects of the PRE to be temperature dependent and concluded that transcription through *Fab-7* derepressed

silencing, likely through displacement of PcG proteins, and that this active chromatin state induced by *Fab-7* was heritably maintained through cell divisions (Cavalli and Paro, 1998). Paro later tested if this process would work the same using other PREs in the Hox complex and one controlling *hedgehog* (*hh*) in imaginal wing disc development. These gave similar results suggesting that this could be a widely used developmental mechanism (Maurange, 2002; Rank et al., 2002), further corroborated by findings that several of these were transcribed during development (Cumberledge et al., 1990; Lipshitz et al., 1987; Sanchez-Herrero and Akam, 1989).

The model proposed by Cavalli and Paro (1998), that transcription through PREs, such as *Fab-7,* can derepress silencing, has since been under scrutiny as Cavalli later reported that this line has a duplication of the *Fab-7* transgene (Bantignies et al., 2003). This was shown whilst demonstrating that *Fab-7* transgenes can pair in long-range interactions with each other and the endogenous *Fab-7*, enhancing silencing (Bantignies et al., 2003). Further investigations into the *Fab-7* fragment that had been used revealed that the sequence also contained other regulatory elements including an insulator adjacent *iab-6* regulatory region that if removed would not perpetuate the derepression of mini-white silencing after Gal4 induced transcription, and instead full silencing would return (Rank et al., 2002). The affects of transcribing through PREs have been further tested by Erokhin *et al* (2015), who used ChIP to investigate if PcG proteins are displaced when transcription is initiated. This study has demonstrated that transcription through a PRE does not displace PcG proteins or remove the repressive epigenetic marks, even if persistently transcribed at high levels throughout development and suggest other adjacent regulatory elements are responsible for the switch in PRE state from silencing to activation (Erokhin et al., 2015). Another study tested transcription of the lncRNA *bxd* at its endogenous loci by mutating the promoter of transcription. The lncRNA*, bxd*, contains a transcribed PRE in the sense direction of the first intron and previously, it was shown that deletion of the *bxd-PRE* does not prevent *Ubx* expression, but leads to misexpression (Sipos et al., 2007). When the promoter of *bxd* was mutated transcription could not be detected of either the *bxd* transcript or PRE. They also only detected slight changes in the Hox gene regulated by *bxd*, *Ubx*, when the native *bxd* and *bxd-PRE* transcript could no longer be made, as *Ubx* expression advanced more rapidly to match later stages of its expression pattern during embryogenesis, but the flies developed normally leading them to believe there was no function of the noncoding transcript itself and therefore transcription (Pease et al., 2013).

A more recent example of possible PRE/TRE function was investigated at the *vestigial* (*vg*) locus and has shown a strand specific switch that seems able to confer activation or silencing of a PRE/TRE using orientation specific non-coding transcription (Herzog et al., 2014). In this study transcription of the reverse strand *in vivo* would displace PRC2 and inhibit methyltransferase activity of E(z) to induce activation of *vg*. However, when tested *in vitro*, transcription of both strands would prevent PRC2 silencing suggesting differential regulation when *in vivo*, whereas the

repressive effects of forward transcription seemed to correlate with long-range pairing of PREs. Another study demonstrated the similarities between PREs from two similar homeodomain genes, *en* and *invected* (*inv*) that are adjacent to each other on the chromosome and regulate segmentation during embryogenesis and posterior compartment formation in imaginal discs (Gustavson et al., 1996). There are four PREs adjacent to these genes (Cunningham et al., 2010), two of which are both required for maintenance of *en* stripes (DeVido et al., 2008). These PREs were able to replace the *bxd*-PRE when tested for pairing sensitive silencing (PSS) of the mini-white gene in a reporter construct and restrict *Ubx* expression in a *lacZ* reporter construct (Americo et al., 2002). The *bxd*-PRE could also maintain perfect *en* stripe expression using a *lacZ* reporter construct, but the *inv* was slightly misexpressed between stripes (Cunningham et al., 2010) demonstrating some, but not all PREs can be interchanged.

Taken together these variable accounts of PRE/TRE actions may be due to differences in PcG/TrxG proteins and distinct complexes bound and therefore different classes of PRE/TREs that function differently, but with so few investigated, especially outside of the Hox complex, it would be difficult to substantiate this against the argument that they are behaving different due to being scrutinized in an environment outside their native site of action and at different time periods to their wild type expression. However, one experiment observed that Gal4 binding to PREs alone was able to derepress the silencing and Gal4 binding alone increased TrxG binding regardless of whether it was transcribed or not, suggesting that Gal4 binding somehow reduces PcG protein binding, although how this comes about is not at all understood yet (Erokhin et al., 2015). An interesting feature of PREs that has been revealed more recently, using chromosome conformation capture technology, is that they can work together to organize chromatin (Delest et al., 2012), explaining the formation of concentrated regions of PcG proteins seen using fluorescent labeling within a nucleus (Pirrotta and Li, 2012). It has also been demonstrated in several instances that PcG targets can have multiple PREs regulating the target genes and PRE/TREs are frequently found within gene clusters, with the best characterized being the bithorax complex of Hox genes (Maeda and Karch, 2011).

One aspect of PRE/TREs that has proven somewhat useful in Drosophila is their prediction based on enrichment of consensus sequence motifs of DNA binding PcG and TrxG proteins. These include *pho* (Brown et al., 1998; Mihaly et al., 1998), *Trl* (Strutt et al., 1997), *Dorsal switch protein 1* (*Dsp1*) (Dejardin et al., 2005), *zeste* (*z*) (Saurin et al., 2001), *grainy head* (*grh*) (Blastyak et al., 2006), *pipsqueak* (*psq*) (Lehmann et al., 1998) and *Sp1* (Brown et al., 2005). There are also now many studies (available from Gene Expression Omnibus (GEO)) that have carried out ChIP-ChIP and ChIP-seq throughout different stages of development and in specific tissues for various components of both PcG and TrxG complexes that can be used in conjunction with predictions. As well as the consensus sequences for the DNA binding proteins that are used to identify PRE/TREs, another sequence, GTGT, is enriched in PREs that when deleted in the *vg*

PRE, reduced the silencing capabilities (Okulski et al., 2011). The jPREdictor (Fiedler, 2006) program has been widely used for PRE prediction and is based on clustered consensus binding motifs and assigning scores to regions based on the weighted sum of the occurrence of motif pairs. This algorithm is useful for identifying sites that are highly similar to the consensus, but cannot identify two or more neighboring weak binding sites. We also do not know if other DNA-binding proteins or alternative factors can recruit PcG and TrxG complexes, which seems likely as ChIP studies reveal many more sites bound by these proteins than are identified by the jPREdictor program (Schuettengruber et al., 2009; Tolhuis et al., 2006) and PREs not predicted by the program have been verified as functional PREs (Cunningham et al., 2010). Therefore, it is essential to combine information from different aspects to identify a PRE/TRE before experimentally validating its function.

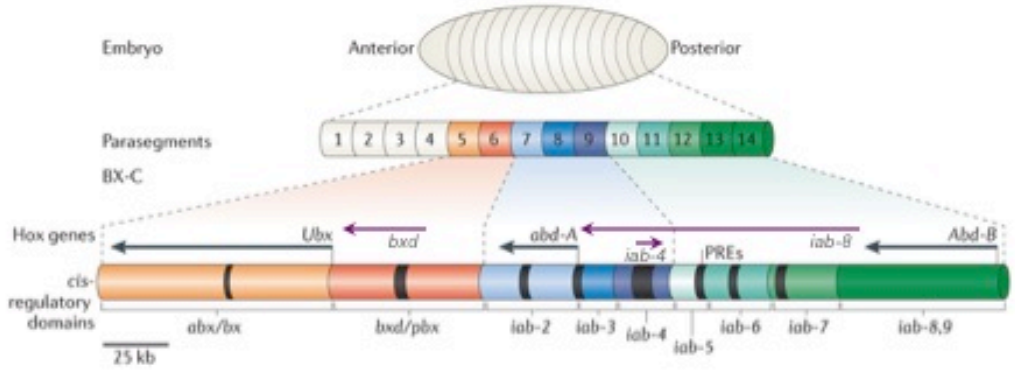


Figure 1.5. Hox genes, lncRNAs and PREs in the BX-C of *D. melanogaster*. Protein coding Hox genes, *Ultrabithorax* (*Ubx*), *abdominal-A* (*abd-A*) and *Abdominal-B* (*Abd-B*) depicted in relation to the well characterized lncRNAs, *bithoraxoid* (*bxd*), infraabdominal-8 (*iab-8*) and *infraabdominal-4* (*iab-4*) and their expression domains of a developing embryos parasegments. The identified polycomb response elements (PREs) are shown as black bars throughout the BX-C (Steffen and Ringrose, 2014).

1.16 Project Summary and Aims

The aim of this project was to better understand the regulation of the Hox genes by the transcription of regulatory regions of DNA. Hox genes have various roles throughout development and their specific expression patterns have long been known to be modulated by *cis*-regulatory elements (Mallo and Alonso, 2013). We now know that many of these regions are transcribed, but still lack an understanding of the functional roles of this transcription and therefore seek to understand its relevance (Starr et al., 2011). There is uncertainty that all transcripts that fall into the current classification of lncRNAs are functional and not transcriptional noise and in certain cases the act of transcription itself seems sufficient to carry out regulation, without the RNA having an apparent function (Starr et al., 2011). We therefore sought to identify lncRNAs that were most likely to have biological roles in the Hox complex of *D. melanogaster* and investigate their functions.

As yet there is no standard methods for identifying functional lncRNAs, so we investigated RNA expression patterns, syntenic conservation, ChIP-seq signatures and motif predictions to identify transcribed loci that cold be functionally relevant.

We also investigated the enrichment of clusters of lncRNAs throughout the *D. melanogaster* genome and if this enrichment was conserved in *D. virilis*, based on protein coding genes within these regions. This revealed that the Hox complex was enriched for lncRNAs when compared to the rest of the genome and this enrichment was also conserved. A subset of lncRNAs were selected that were most likely to be functional within the Hox complex. We then also reasoned that there would be evolutionary conservation of the regions of a developing embryo that the lncRNAs were transcribed in, in both *D. melanogaster* and *D. virilis*, so carried out ntFISH and found those that were conserved in expression patterns. RNA-seq was also carried out in *D. pseudoobscura* to compare syntenically conserved lncRNAs in the Hox complex and available ChIP datasets were used to investigate regulatory proteins that bind to our subset of lncRNAs.

To better understand the roles of lncRNAs in the Hox complex, we selected one that had not been previously studied that we had identified as a potentially good candidate for investigating. Our aim was to find out if the lncRNA was functional and identify genes that it may be involved in regulating or what the consequences were of perturbing its usual expression. The experiments ectopically expressing the lncRNA and partially duplicating its second exon produced a variety of homeotic mutations on the adult flies that could linked to the lncRNAs endogenous expression or the adjacent Hox genes. Interestingly, ectopically expressing the regulatory DNA region just upstream of the lncRNA produced matching phenotypes, although the CRISPR-Cas9 experiments using a mini-white reporter indicates it has a silencing action on the lncRNA. Whilst carrying out the partial duplication of the lncRNA, we should have seen the affects of a partial deletion of the same region, however, flies expressing the mini-white marker to indicate this did not hatch, therefore, leading us to believe that the deletion was lethal. Altogether, these experiments have led us to believe that we have successfully identified a functional lncRNA in the Hox complex of *D. melanogaster* that is conserved over ~63 million years (Tamura et al., 2004) as we identified a syntenic transcript that was also expressed in a similar pattern in *D. virilis* embryos.

Further analysis of lncRNAs included investigating alterations in their expression patterns in mutant backgrounds of segmentation genes that have been shown to regulate Hox gene expression, to test if they also have an effect on the expression of lncRNAs. This demonstrated that segmentation gene mutant backgrounds would lead to altered expression patterns of lncRNAs and are therefore likely to be regulating their expression. Finally, we characterized the spatiotemporal expression of 3 lncRNAs relative to each other, 2 of which are transcribed antisense, from the same locus, as the other large lncRNA. This time series demonstrates that these lncRNAs follow both spatial and temporal collinearity demonstrated by the Hox genes and are expressed in very specific patterns on the developing embryo.

2. METHODS

2.1 Identification of lncRNA clusters within the *D. melanogaster* genome

Files containing coordinates of all annotated lncRNAs for *D. melanogaster* (FlyBase release 6.08) were downloaded from FlyBase in General Transfer Format (GTF) and the base-pair distance distribution between adjacent lncRNAs plotted in a histogram in order to determine an appropriate threshold to identify clusters. The number of lncRNA pairs (y-axis) was plotted against distance separating lncRNAs (x-axis) and this showed that the majority of lncRNAs were clustered <4 kb, with a sharp decrease in the number of lncRNAs separated by over 4 kb, suggesting that at this distance we would find more distinct clusters. However, based on previous lab work on lncRNAs in the Hox complex and the distance typically found between protein-coding genes, we knew this distance was very small and would separate regions of known gene clusters into small sections and possibly only identify a few very dense lncRNA clusters. We also realized that protein-coding genes could separate lncRNAs by great distances but could actually link them into the same cluster, so we removed these distances from our analysis and instead used a trial and error approach testing out a range of distances until the most dense clusters matched visibly enriched clusters of lncRNAs plotted out across the chromosome.

2.2 GO Term analysis of protein-coding genes within *D. melanogaster* lncRNA clusters

The clusters with the 20 highest numbers of lncRNAs were analyzed by the Panther classification system (http://pantherdb.org) to identify the overrepresented GO-Slim terms found in each list of protein-coding genes from the enriched lncRNA clusters. The GO-Slim terms are a subset of the 7024 GO terms, developed to give 218 broader terms, giving an overview of the Biological Processes that are represented in a list of genes. We used PANTHER's statistical overrepresentation test to find any general associations between the protein-coding genes within our top 20 clusters and reported as GO-Slim terms found to be significantly increased (P=<0.05) against the background set of *D. melanogaster* GO-Slim terms. Overrepresented indicates the terms have appeared more frequently than found by chance if testing the same number of genes randomly in the organism and the fold enrichment is a measurement of how many more times this term is enriched compared to the background frequency.

2.3 Fly husbandry

Flies were maintained in large communal 25°C incubators in the Fly Facility in the Michael Smith Building of the University of Manchester unless otherwise stated, feeding on standard fly media.

2.4 RNA collection, sequencing and annotation from *D. pseudoobscura* and *D. virilis* species

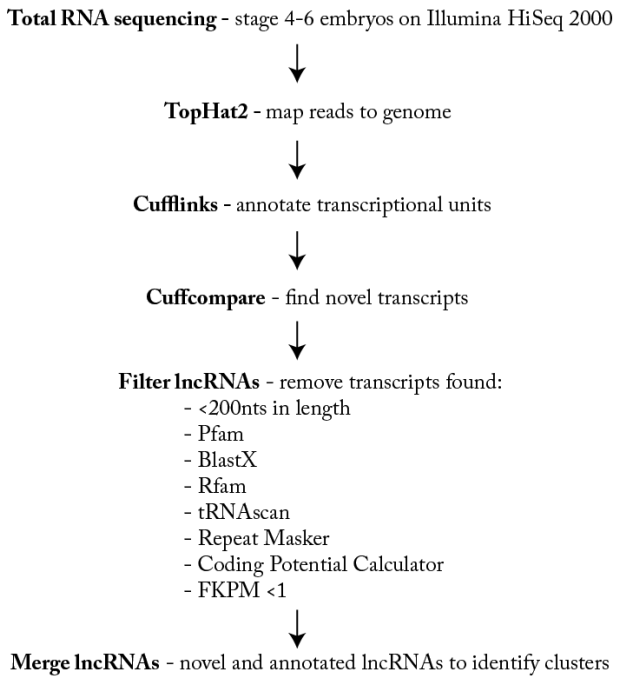


Figure 2.1. Pipeline for identification of novel lncRNAs in stage 4-6 *D. virilis* embryos and lncRNA clusters. Transcripts identified as positive by the different filters were removed in once sequences were analyzed with each of the tools revealing the number and location of all novel lncRNAs from the RNA-seq. These were then merged with the currently identified lncRNAs from FlyBase to identify clusters.

Figure 2.1 shows a summary of the pipeline used to identify novel lncRNAs. Adult flies were left to lay eggs on apple juice agar supplemented with yeast for 1 hour in a 25°C incubator. The *D. virilis* eggs were then left to develop at 25°C for a further 9hrs, whilst *D. pseudoobscura* developed for 5hrs at 25°C in order to obtain approximate stages 4-6 in both species (Campos-Ortega and Hartenstein, 1997), as *D. virilis* develops ~1.5x lower than *D. melanogaster* (Ninova et al., 2014). The embryos were collected from the apple juice agar plates and a *mir*Vana™ miRNA Isolation Kit (Life Technologies Cat #AM1560) was used to isolate and collect total RNA following the manufacturers procedure for total RNA isolation (https://tools.thermofisher.com/content/sfs/manuals/cms_055423.pdf). The RNA was poly(A)+ selected and libraries were prepared using Illumina's HiSeq Stranded mRNA Library Prep Kit (Illumina, Cat #RS-122-2101) and tested for quality and quantity using a TapeStation, before being sequenced in the Genomic Technologies Core Facility at the University of Manchester using

an Illumina HiSeq 2000 System. The 101bp paired-end RNA-seq reads were mapped to the *D. virilis* genome (FlyBase release1.03) with TopHat2 (Kim et al., 2013) as it aligns spliced reads, allowing no more than 2 mismatches for each 101nt read. The options used for TopHat2 were: -p 8 -a 5 -m 2 -i 28 --library-type fr-firststrand. Transcriptome assembly was carried out using Cufflinks version 2.2.1 with the options: cufflinks --library-type fr-firststrand --multi-read-correct --overhang-tolerance 2 --min-intron-length 8 -p 4. Cuffcompare was then used to detect novel transfrags annotated by the Cufflinks run against the *D. virilis* genome FlyBase release1.06. The class codes that most reliably indicate a novel transcript, that is not likely to be a novel isoform or alternate UTR of a previously annotated gene, are class codes 'u' (unknown intergenic), 'x' (exonic overlap with reference on opposite strand) and 'i' (a transfrag falling entirely within a reference intron) (Fig.2.2). The u, x and i class codes from the Cuffcompare output were then strictly filtered for reliability of falling into the classification of lncRNAs. Any lncRNAs <200nts in length or with an FKPM of <1 were removed and all transfrag types were separated based on their class code from Cuffcompare that reports their position relative to current gene annotations. These were all investigated for functional protein domains using the Pfam (version 29.0) database (Finn et al., 2014; Finn et al., 2016) and BlastX for amino acid sequence similarity to known proteins (Altschul et al., 1990; Gish and States, 1993). Rfam (version 12.0) was used to identify any characteristics matching known RNAs (Nawrocki et al., 2015), as well as tRNAscan-SE to ensure none of the novel transcripts were tRNAs (Schattner et al., 2005), both with default parameters using the online servers. Further tests were run to identify any repetitive DNA elements with RepeatMasker (version 4.0.6) within these sequences, using the *D. melanogaster* library and RMBlast (version 2.2.28) as the search engine, with default sensitivity with the online server (Smit, 2013-2015). Also, to check that they were not in fact protein-coding genes, each sequence was investigated using the coding potential calculator (CPC) with default parameters (Kong et al., 2007), as this investigates six sequence features to distinguish between coding and non-coding RNAs. Once the novel lncRNA list was complete the histograms containing the properties of the total lncRNAs were constructed using an online histogram statistics software calculator (http://www.wessa.net/histo.wasp) based on R code from Modern Applied Statistics with S (Venables and Ripley, 2002).

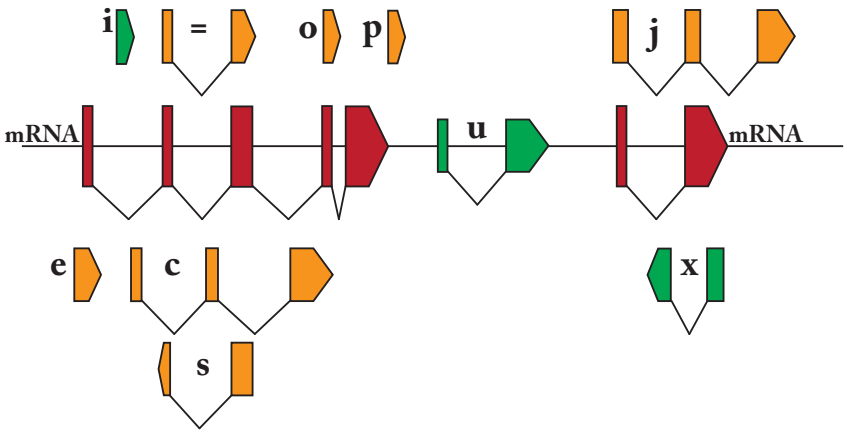


Figure 2.2. Cuffcompare class codes illustrations based on their loci relative to mRNA genes. The class codes descriptions from Cuffcompare are as follows: **=** - Complete match of intron chain, **c** – contained, **j** – potentially novel isoform (at least one splice junction is shared with reference transcript, **e** - Single exon transfrag overlapping a reference exon and at least 10bp of a reference intron, indicating a possible pre-mRNA fragment, **o** – generic exonic overlap with a reference transcript, **p** – possible polymerase run-on fragment (within 2 kb of reference transcript), **u** – unknown intergenic transcript, x – exonic overlap with reference on the opposite strand, **s** – an intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors). Green transfrags represent the transfrags retained for further analysis as most likely to be novel, putative lncRNAs and not part of a reference transcript and orange are those that could be matched in some way to a reference transcript.

2.5 Protein-coding gene ortholog comparison from lncRNA clusters in *D. melanogaster* and *D. virilis*

Once the top 20 most highly enriched lncRNA clusters were identified in *D. virilis*, the genomic co-ordinates were used to collect the protein coding genes from each cluster. These were extracted from the track named 'Dmel Orthologs' that had been downloaded from FlyBase as a GFF3 file and curated by OrthoDB (http://www.orthodb.org). The protein coding genes from the top 20 lncRNA clusters of both *D. melanogaster* and *D. virilis* were then compared to find matching genes in both sets and highlighted using excel. Gene clusters that had been identified were then mapped out onto corresponding chromosome arms or scaffolds to allow visualization of similarities and splits or movements of clusters.

2.6 Identification of lncRNAs in the Hox complex of *D.melanogaster*

RNA sequencing tracks, taken from developmental intervals every two hours during embryogenesis up to 24hrs, were downloaded from the modENCODE *D. melanogaster* browser (modencode.org). The GEO accession numbers found associated with the MultiMapper option

from modENCODE are listed in table 2.1. This consisted of poly(A)+ purified RNA samples from different experiments that used the Illumina GAII sequencing platform. The Illumina GAII platform did not include strand information and was mapped with TopHat v1.0.10. The mapped RNA was then visualized using the Integrative Genome Viewer (IGV) (https://www.broadinstitute.org/igv/) to align RNA fragments with the current annotated genome (dm3) and distinguish between known genes and regions that were transcribed and not annotated, and therefore, likely to be lncRNAs.

In order to annotate the start positions of transcription, 5' cap analysis gene expression (5' CAGE) data was included with the RNA-seq profiles that had been carried out in 2hr windows from egg laying through to 8hrs AEL. The CAGE data was downloaded in SAM alignment format from modENCODE that had been generated as part of the Berkeley Drosophila Genome Project (BDGP), and visualized in IGV along with RNA-seq. The peaks were manually curated as stacks of reads that were significantly higher in one loci compared to low-level background noise. A GFF3 file was created to document estimated start and end positions of transcripts identified as possible novel lncRNAs with starts and ends annotated using a combination of 5' CAGE and RNA-seq reads to approximate the ends.

Table 2.1. GEO Accession numbers linked to modENCODE RNA-seq

Time after egg laying	GEO Accessions
0-2	SRX008271, SRX008258, SRX008238, SRX008227, SRX008180
2-4	SRX008193, SRX008190, SRX008270, SRX008179, SRX008015
4-6	SRX008250, SRX008217, SRX008265, SRX008181, SRX008027
6-8	SRX008257, SRX008212, SRX008210, SRX008175, SRX008025
8-10	SRX008274, SRX008273, SRX008252, SRX008249, SRX008010
10-12	SRX008243, SRX008274, SRX008247, SRX008208, SRX008198
12-14	SRX008277, SRX008235, SRX008225, SRX008177, SRX008018
14-16	SRX008262, SRX008237, SRX008233, SRX008196, SRX008007
16-18	SRX008278, SRX008242, SRX008205, SRX008187, SRX008006
18-20	SRX008259, SRX008222, SRX008215, SRX008213, SRX008020
20-22	SRX008256, SRX008241, SRX008221, SRX008214, SRX008011
22-24	SRX008266, SRX008251, SRX008234, SRX008167, SRX008019

2.7 Analysis of PcG, TrxG and HDAC binding to Hox lncRNAs in *D. melanogaster*

Files were downloaded from the Gene Expression Omnibus (GEO) database containing mapped or unmapped ChIP-ChIP or ChIP-seq reads from experiments investigating PcG, TrxG or HDAC binding in *D. melanogaster* or *D. pseudoobscura* (see Table 2.2). If not available as

mapped files, then SAM files were downloaded and mapped in the same way as previously mentioned for RNA-seq experiments. Once mapped the files were loaded into IGV to investigate peaks of stacked reads at loci of interest.

Table 2.2. Accession numbers of datasets from GEO repository of PcG, TrxG and HDAC ChIP experiments.

SPECIES	PROTEIN	TISSUE	SRA/BED accession #
D. mel	Pc	4-12hr embryos	SRX681771
	Pc	S2 cells	SRX027823
	Pcl	0-8hr embryo	SRX025472
	Ph	4-12hr embryos	SRX681770
	Dsp1	4-12hr embryos	SRX681772
	Pho	4-12hr embryos	SRX681813
	Psc	S2 cells	SRX027827
	Suz(12)	5-13hr embryos	SRX671953
	Mdg4	0-12hr embryos	GSM409072
	Brm	pupae	GSM400395
	Fs(1)h	Kc167 cells	SRX202999
	Trl	8-16hr embryos	SRX495313
	Trl	16-24hr embryos	SRX025479
	Trr	S2 cells	SRX193314
	Utx	S2 cells	SRX193315
	Trx	S2	SRX027830
	HDAC	0-12hr embryos	SRP001789
D. pse	Trl	4-12hr embryos	SRX032424
	Pc	4-12hr embryos	SRX681788
	Dsp1	4-12hr embryos	SRX681790
	Pho	4-12hr embryos	SRX681816

2.8 Probe synthesis and imaging

2.8.1 Genomic DNA extraction

D.melanogaster w^{1118} flies and *D.virilis* were maintained on standard medium at 25˚C. 25 flies were frozen for 15mins and homogenized using a sterile pestle and mortar containing 250µl of Buffer A (0.1M Tris-HCl pH 9.0, 0.1M EDTA pH8.0, 1% (w/v) SDS). The homogenate was transferred to 1.5ml eppendorf tubes and after a 30min incubation at 70˚C, rapidly agitated in 35µl

8M KAc, before a 30min incubation on ice. Sample was then spun for 5mins at 13,000 rpm and the aqueous layer agitated with 1 volume Phenol:Chloroform and spun for 5mins at 13,000 rpm. This step was repeated with another volume of Phenol:Chloroform, before transferring the supernatant to a new tube, and rapid agitation of supernatant with 150µl Propan-2-ol. The sample was spun for another 5mins at 13,000 rpm, and the supernatant removed, leaving the DNA pellet. The pellet was washed with 1ml 70% EtOH, spun for a further 5mins at 13,000 rpm, and dried for 10mins. The pellet was then resuspended in ddH_2O. Throughout all DNA isolation procedures, quantity and quality of DNA was measured using a Nanodrop spectrophotometer.

2.8.2 PCR primers and amplification

Primers were designed using Primer3web (version 4.0.0) (Table.2.3) and ordered lyophilized from Integrated DNA Technologies®, then used at 20µM concentrations in PCR reactions. PCR reaction volumes were 50µl. Myfi™ polymerase (Bioline, Cat #BIO21117) was used with 330-490ng genomic DNA template and colony PCR with M13 primers. Recommended PCR conditions were used as instructed by manufacturers guidelines for MyFi™ polymerase and varied depending on annealing temperatures of primers and length of products. All products were checked for efficiency of amplification and size on 1% agarose gels using Hyperladder I (Bioline, Cat #BIO-33053) with 1ppm ethidium bromide and 0.5M TAE running buffer.

2.8.3 Cloning

PCR products were cloned using the TOPO® TA Cloning® Kit, Dual Promoter, with One Shot® TOP10 chemically competent *E. coli* cells (Invitrogen, Cat #K460040). The cloning reaction consisted of 1µl PCR product, 0.5µl salt solution, 1.5µl dH_2O, and 0.5µl TOPO vector, mixed and incubated for 30mins at room temperature of which 2µl was then transformed into TOP10 chemically competent cells on ice for 30mins. The cells were then heat-shocked in a 42˚C water bath for exactly 30secs and transferred back to ice. 250µl of room temperature super optimal broth with catabolite repression (S.O.C) (provided in the TOPO® TA Cloning® Kit) was added and the tube shook at 225rpm, 37˚C for 1 hour. 50µl and 25µl of the S.O.C containing transformed TOP10F' cells was spread on prewarmed (37˚C) selective agar plates (1% Tryptone (Melford, Cat #T1332), 0.5% yeast extract (Melford, Cat #GY1333), 1.8M NaCl (Fisher Scientific, Cat #10112640), 2% agar (Melford, Cat #GM1002), 100µg/ml ampicillin (Sigma, Cat #A0166), 20mM IPTG (Bioline, Cat #BIO-37036), 80µg/ml X-gal (Bioline, Cat #BIO-37035)) and incubated overnight at 37˚C. 8 white colonies from each plate were restreaked on a fresh selective plate and grown for another 8 hours at 37˚C, before colony PCR was carried out to screen for those with the inserts by checking the size on 1% agarose gel.

Table 2.3. Primer sequences for genomic DNA amplification used for probe synthesis. Sequences are for *D. melanogaster* unless otherwise stated. The *Hox-G* probes start from the promoter, numbered 1, and work across towards the 3' end and further and were used together for signal.

Name	Primers
AntpP1	5'-AGACTTTCTCCCATTTGTTCC-3' 5'-AAGTTCACACTCATGGCAAAG-3'
AntpP2	5'-GCACTAACAACAAGCAACTGC-3' 5'-GAGCAAACAATTCGGAGACAG-3'
Scr	5'-CCCGTCCAATTGTATCTGCGAGT-3' 5'-AAACTGCACTGTGGTGTGGAGGA-3'
ftz	5'-TTGCAAAGACTCGAAACGCA-3' 5'-GTTTTGGGCTTGTGTTTGGC-3'
Ubx - promoter	5'-TTTCTCCTTTGTTTTAGCACCAA-3' 5'-TCGCCACTCAGTTGAAGGAA-3'
abd-A - 5' end	5'-ACGGCTGGAACTGTGGATAC-3' 5'-AATACAACGCAACCCGAGAC-3'
Abd-B – 5' end	5'-ATGAGGAGGAGGTCCGAGAT-3' 5'-GGGAAGGGGTGAACACTACA-3'
TipX - whole	5'-GCTCTAGATGGAAGCTTAAGTTTAAGTTAAG-3' 5'-GCTCTAGAGCGGACCTGTGCAGTTCCTCC-3'
lincX - 5' end	5'-TTAAAGACAGAGCCCAACGATGC-3' 5'-ACCGATCAGCCAACACAATCAAC-3'
D.virilis lincX 5' end	5'-GGATTTAAGGTGCGTCGTGT-3' 5'-CCCTCTGTCAAACCACAGGT-3'
Hox-O - 5' end	5'-TGCGGAAAACAGGAATACAA-3' 5'-GTTTCAGCGTGACCCTTGTT-3'
D.virilis Hox-O – 5' end	5'-CTATGTTTGCCAACGGTGTG-3' 5'-ACGCGTTCTCTTCTTGCATT-3'
iab-4 – 5' end	5'-TCCCCATTAAATCGCATCGC-3' 5'-CGGGTGGAATGTGCAAATGA-3'
D.virilis iab-4 – 5' end	5'-AGAAACCCCGTTTACGCTTT-3' 5'-TCAAATGTCAGCCGTCAGAG-3'
iab-8 – mid exon	5'-CAGCACCATAATTCAGGGCC-3' 5'-CCTTCCCACTTTTGCCCTTC-3'
D.virilis iab-8 - 5' end	5'-ATCTGTCACAACCACCGTCA-3' 5'-CTTTACAGCCTCGATGCACA-3'
iab-7 PRE - whole	5'-TGGTTTCCAACTCTAGCGGT-3' 5'-TTGGGTTCGGTAAGAGGTCT-3'
Bxd - 5' end	5'-AAGCGGATGGGATGTAGATG-3' 5'-ACTGCCTCCGCTAACAAAGA-3'
D.virilis Bxd - 5' end	5'-GGCACACGGATCCATAAGAA-3' 5'-CGCACAACCAACTCAAAAGA-3'
Tre2 - whole	5'-CCAAGTATCGAGGCGCTAAG-3' 5'-ATGGCCTCATAATCGTTTGC-3'
Hox-G – 1	5'-GGAATATAGGGCCACCGACT-3' 5'-ATTGTGTACGTTCGCTGCAA-3'
Hox-G – 2	5'-CCACCTTTTGGGCTAACAAA-3' 5'-GACCACAAGATGGCTGGAAT-3'
Hox-G – 3	5'-AACCGGCTACCTGGCTAAAT-3' 5'-AAGAAAGCGGCGAACTGTAA-3'
Hox-G - 4	5'-ACGAGAGACTTCCTGCCAAA-3' 5'-TAATCCGACGCCAATCCTAC-3'
Hox-G – 5	5'-CAATTTTTGACCACGCCTTT-3' 5'-ACTTGAAACGGCCAAAAATG-3'
Hox-G - 6	5'-AGGCATTTATCATCGGCAAG-3' 5'-TTAATGGCTTTTCGCAGCTT-3'
D.virilis Hox-G -1	5'-GACTGCGCTCGTAATTCTCC-3' 5'-AGGTGTCACAGGCTCACAGA-3'
D.virilis Hox-G -2	5'-GACTGCGCTCGTAATTCTCC-3' 5'-AGGTGTCACAGGCTCACAGA-3'
D.virilis Hox-G -3	5'-CCGATACTGAAGGGCGAATA-3' 5'-CTGGGCAAATTGCTTTGTTT-3'
D.virilis Hox-G -4	5'-AGCCGATGCCTCAGACTAAA-3' 5'-AGGAACTCGAAACAGCAGGA-3'
M13 primers	5'-GTAAAACGACGGCCAG-3' 5'-CAGGAAACAGCTATGAC-3'

2.8.4 Probe synthesis

Colonies containing the inserts were then grown in 3mls S.O.C medium with 0.1% Ampicillin overnight at 37°C, 225rpm in round bottom snap cap Falcon tubes (BD Falcon, Cat #352051). The following morning, the S.O.C. medium containing the amplified colonies was purified with a Purelink® Quick Plasmid Miniprep kit (Invitrogen, Cat #K210010) and sent for sequencing with M13 primers to the in house sequencing facility. Once the correct inserts were confirmed and checked for orientation, they were amplified in 100µl PCR reactions using M13 primers with the same concentrations of reagents and conditions as previously described. The entire product was run on a 1.5% agarose gel with EtBr using large wells and the bands containing the correct size insert cut out with sterile heavy duty single edge carbon steel blades (Agar Scientific, Cat #AGT5016). The cut bands in the agarose gel were weighed for use with a Purelink® Quick Gel Extraction kit (Invitrogen, Cat #K210012), followed by purification of the gel extracted DNA using a Purelink® Quick PCR Purification kit (Invitrogen, Cat #K310001).

All RNA probes for nascent transcript fluorescent in situ hybridization (ntFISH) were made using digoxigenin (DIG) (Roche, Cat #11277073910), fluorescein (FITC) (Roche, Cat #11685619910), biotin (BIO) (Roche, Cat #11685597910 or labeled nucleotides. Synthesis with the T7 polymerase (Promega, Cat #P2075) was carried out in a reaction containing 4µl (x5) supplied buffer, 2µl DTT (Promega, Cat #V3151), 2µl DIG labeling mix (Roche, Cat #11277073910), 40U RNasin (Promega, Cat #N2111), 80U T7 polymerase, 350ng template, then made up to a 20µl total reaction volume with dH_2O. The T7 RNA probe reaction was placed in a 37°C incubator for 2½ hours. For synthesis of the opposite strand, SP6 polymerase (NEB, Cat #M0207) was used in a reaction containing 4µl (x10) supplied buffer, 4µl labeling mix, 2µl RNasin, 40U SP6 polymerase, 2µg template and made to a final volume of 40µl with dH_2O. The SP6 RNA probe reaction was placed in a water bath set to 40°C for one hour. 6µl (T7) or 12µl (SP6) dH_2O was added once the incubation periods were finished, and 1µl taken from each to visualize on a 1.5% agarose gel. The labeled RNA was then precipitated with either, 2.5µl (T7) or 5µl (SP6) 4M LiCl (Fisher Scientific, Cat #L121), and 75µl (T7) or 150µl (SP6) 100% ethanol, then spun at 4°C for 30mins at 16,602xg. The pellet was then washed with 70% ethanol and air dried before being resuspended in 100µl hybridization solution ((50% formamide (Sigma, Cat #F9037), 5x saline sodium citrate (SSC) (Sigma, Cat #S6639), 0.1% Tween (Sigma, Cat #P1379, 1% fragmented salmon sperm DNA, 0.05% (10mg/ml stock) heparin, 22.5% dH_2O)].

2.8.5 Embryo collection and fixation

Drosophila embryos were collected, after laying for 7 hours (*D. melanogaster*), or 10.5 hours (*D. virilis*) from apple juice plates, supplied with yeast paste (Sigma, Cat #51475) to facilitate

laying, and placed in 500ml polypropylene breeding cages. Embryos were collected using a small paint brush and wash buffer (1M NaCl, 4mM Triton x-100 (Sigma, Cat #X100)) and prepared by dechorionating with a 2.5% sodium hypochlorite solution (Fisher Scientific, Cat #SS290-1) for 1½mins and then fixed by incubation in a fixation solution (500µl PBS (Sigma, Cat #P5493), 500µl dH_2O, 5ml heptane (Sigma, Cat #34873) 4ml 10% formaldehyde (Polysciences, Cat #04018-1)) in scintillation vials placed on an orbital shaker for 45mins at 220rpm. After removal of the aqueous phase, the embryos' vitelline was removed by addition of 8ml methanol (Fisher, Cat #A452-1) to the scintillation vials followed by rapid agitation. After removal of the upper phase of heptane, devitellinised embryos were stored under methanol at -20°C.

2.8.6 Embryo prefixation for hybridization

After being transitioned to ethanol, the embryos were cleared of cholesterol by rocking for 1hour in xylenes and washed with ethanol and methanol before post fixation for 25mins in a solution of 50% methanol and 50% PBT (1x PBS, 0.1% Tween) with 5% formaldehyde. The embryos were then washed with PBT before being transitioned into hybridization solution with a 10min incubation with rocking at room temperature in 50% PBT and 50% hybridization solution, then incubated in 100% hybridization solution for ~2hrs in a water bath set to 55°C, changing the hybridization solution every 30mins.

2.8.7 RNA probe hybridization

Approximately 50µl of prefixed embryos were hybridized with each probe. For hybridization the probes were diluted in hybridization solution (1:100), heated to 83°C for 2½mins and quickly transferred to ice to denature any secondary structure that could have formed. The probe mixture was added to the embryos and incubated in a water bath set to 55°C for 22hrs with some occasional gentle agitation. The embryos were then washed 3 times with prewarmed hybridization solution, once for 5mins, and then twice for 30mins.

2.8.8 Fluorescent detection

The embryos were transitioned into PBT, by 10mins of rocking in 50% PBT and 50% hybridization solution, and then 4x 5min washes in 100% PBT and 2x 30min washes in PBT containing 2x Western Blocking Reagent (WBR) (Roche, Cat #11921673001). 1µl of primary antibodies, combinations of sheep anti-DIG (Roche, Cat #1333089), rabbit anti-FITC (Invitrogen, Cat #A-889) and mouse anti-BIO (Invitrogen, Cat #03-3700), were then diluted in 400µl PBT with 2x WBR and added to each tube of embryos. These were incubated on an orbital

shaker overnight at 4˚C, before being washed 4x in PBT for 10mins each, followed by 30mins shaking in PBT with 2x WBR. The secondary antibodies, combinations of anti-sheep Alexa Fluor®555 (Invitrogen, Cat #A21436), anti-rabbit Alexa Fluor®647 (Invitrogen, Cat #A21573) and anti-mouse Alexa Fluor®488 (Invitrogen, Cat #A21202), were diluted 1:400 in PBT and 2x WBR, then incubated with the embryos in the dark, on an orbital shaker at room temperature for 1½hrs. Another 5 washes were carried out with PBT, the first for 1min, the other 4 for 15mins each, and then the embryos were mounted on glass slides with ProLong Gold Antifade reagent with DAPI (Life Technologies, Cat #P36930) and a cover slip. These were left to dry in the dark for 24hrs before being stored at -20˚C and visualized with an Olympus Fluoview fv1000 confocal microscope. The images were then processed using Fiji software (Schindelin et al., 2012) and Adobe® Photoshop®.

2.9 Prediction of PREs using jPREdictor and evolutionary changes

The jPREdictor program scores sequences by using a previously validated motif set (Ringrose *et al* 2003) to apply a weight to each motif. The jPREdictor program uses randomization and sampling to calculate a significant cutoff score to identify the threshold at which a PRE regulatory element is confidently identified (Fiedler, 2006). Each motif in the sequence being examined is weighted by the numbers of the motif found in the model and divided by the numbers of motifs found in the background set, then normalized to the length of the sequence to generate a cutoff score. We ran the full DNA sequence of the ANT-C and BX-C and their equivalent regions in *D. melanogaster*, *D. pseudoobscura* and *D. virilis*. The summary file containing coordinates of all potential PRE/TREs and their scores that were above the threshold was used to align each predicted PRE to the Hox genes and lncRNAs detected using the SnapGene software (www.snapgene.com). SnapGene was also used to identify all motifs for the sequences that were known or possible PREs, along with the control.

2.10 Gal4 driven ectopic expression of *Hox-G* and G-PRE

2.10.1 Cloning

The *Hox-G* and G-PRE sequences were PCR amplified from W1118 *D. melanogaster* genomic DNA with PrimeSTAR GXL DNA polymerase (Clontech, Cat #R050A) for high-fidelity and blunt ends, following the manufacturers standard protocol. The PCR primers were designed to amplify and add 15bp overhangs for cloning into the pUAST vector. The primers were: G-PRE forward 5'- CGAGGGTACCTCTAGTGGTCACGATCGTGATCGTGGT-3' and reverse 5'– ACAAAGATCCTCTAGTCCTCGAAAAGTAAACGCCCATAAACAATG–3'; *Hox-G*

forward 5'–CGAGGGTACCTCTAGCATTCGAGTGCATTTTTCACTCAACAC–3'; reverse 5'-ACAAAGATCCTCTAGGCTCCTTTCAAATGAAAATATATATAAATGAATTTTAAG-3'. The primers were designed using SnapGene software for Infusion cloning. The PCR products were cleaned using the Purelink® Quick PCR Purification kit and validated on agarose gels and by Nanodrop as before. The pUAST vector was linearized at the multiple cloning site using XbaI to cut (NEB, Cat #R0145S) following manufacturers guidelines and after heat inactivation was also cleaned using the Purelink® Quick PCR Purification kit. The Infusion HD Liquid Cloning kit (Clontech, Cat #638909) was then used to insert the PCR products following manufacturers guidelines and transformed into Top10 cells for amplification on ampicillin agar plates. Colony PCR was used to identify colonies with inserts based on size using the same primers from original genomic DNA PCR amplification for cloning. Positive colonies were then grown in Lysogeny broth (LB) (Bertani, 1951) overnight and plasmids were collected in the morning using the Purelink® Quick Plasmid Miniprep Kit (Invitrogen, Cat #K210010) and sequenced in house for verification. Figure 2.3 shows a plasmid map of the *Hox-G* transcript in pUAST.

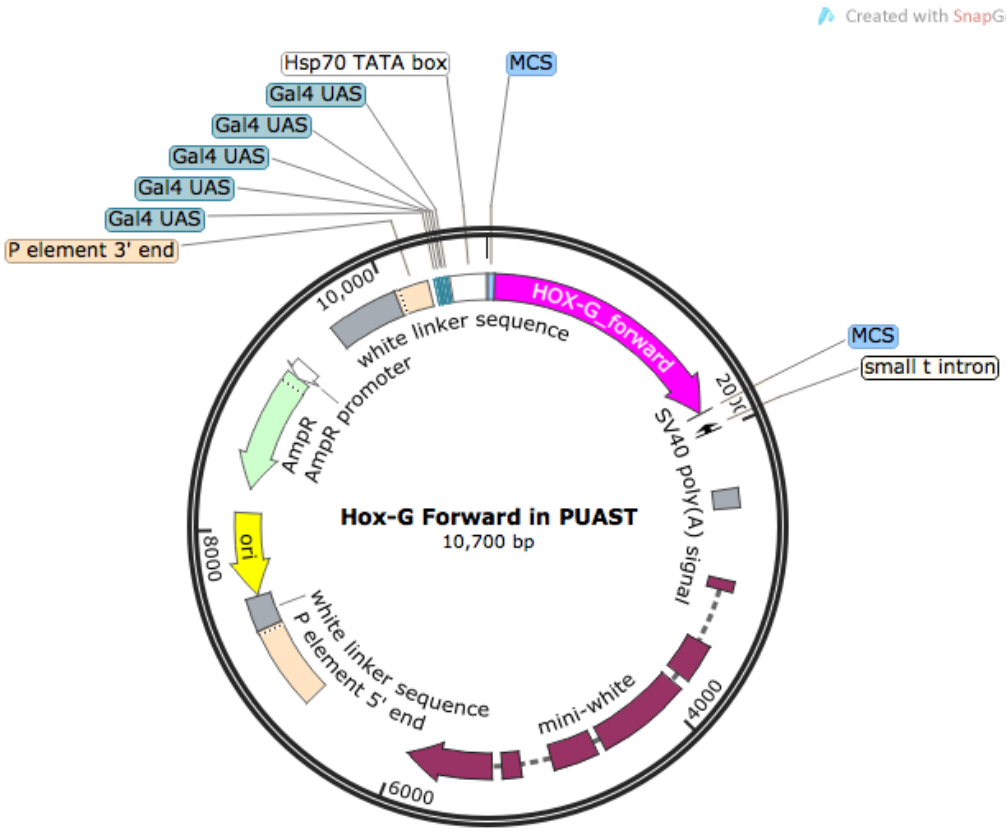


Figure 2.3. Plasmid map of *Hox-G* in pUAST in the forward orientation. Map was created using SnapGene software and was a result of the output from designing the InFusion cloning of the gene into the vector.

2.10.2 Transformations and ectopic expression

The pUAST vectors containing the *Hox-G* and G-PRE sequences in the same forward orientation that *Hox-G* is transcribed from were sent to BestGene Inc for microinjection into w^{1118} flies expressing transposase for random insertion into the genome. Flies expressing the mini-white marker were sent back and we homozygozed 8 lines of each insert by crossing single male and female virgins and waiting for two generations without white eyes. Different Gal4 drivers were crossed to the homozygozed pUAST lines along with PBac(WH) (Thibault et al., 2004) lines that had been identified in or flanking *Hox-G* that also had UAS binding sites and allow for Gal4 driven expression of adjacent genes. The Gal4 driver lines and PBac(WH) lines used were:

Bloomington 1553 - w[*]; wg[Sp-1]/CyO; P{w[+mW.hs]=GAL4-dpp.blk1}40C.6/TM6B, Tb[1]
Bloomington 1774 - w[*]; P{w[+mW.hs]=GawB}69B
Bloomington 30564 - y[1] w[*]; P{w[+mW.hs]=en2.4-GAL4}e16E
Bloomington 7062 - w[*]; P{w[+mC]=matalpha4-GAL-VP16}V2H
Harvard Exelixis – f00519 – referred to as PBac(WH)1
Harvard Exelixis – f01872 – referred to as PBac(WH)2
Harvard Exelixis – f02656 – referred to as PBac(WH)3

Although the Harvard stocks do not list any balancers on either their website or FlyBase, it was quickly discovered that PBac(WH)2 flies contained a balancer with Humeral and ebony, likely TM6B, as when crossed to a line containing a P-element with mini-white/TM3, flies with white eyes and black bodies were produced. PBac(WH)1 and PBac(WH)3 seemed to be homozygous when tested the same way. The parents were removed prior to offspring hatching and the majority were expected to be heterozygous for the Gal4 driver and UAS binding sites and if not then the markers were used to select the flies that contained both. Adults were visually inspected under light microscopes for any visible mutations and counted to assess penetrance. F1 flies only were counted by 3 people carrying out the crosses and the types and numbers of fly phenotypes counted and recorded in shared tables. Flies with visible mutations were frozen in -20˚C and suspended in PBS for imaging to prevent drying out and to aid in positioning of fly for imaging. Flies that had specific leg phenotypes were frozen and legs were removed and mounted on slides using CMCP-10 High Viscosity Mountant (Polysciences, Cat #16300-250).

2.10.3 Inverse PCR

Inverse PCR (iPCR) was carried out on flies that were positive for mini-white and had therefore had the pUAST vector integrated into their genome, in order to identify the insertion sites. This work was carried out by two masters students, Philippa Jackson and Margrete

Langmyhr, using the methodology from the Berkeley Drosophila Genome Project. The only alterations being that they used GXL polymerase and sequenced in house.

2.11 Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated proteins (Cas9) mediated transgenesis

An integration site within the second exon of *Hox-G* was identified that had a suitable motif for insertion of a pTVCherry vector (Baena-Lopez et al., 2013). This motif was a unique site that the guide RNA recognizes when associated with the Cas9 enzyme to guide the enzyme to the DNA to make a double strand cut. The sequence identified in *Hox-G* for the guide RNA was: 5'-GAGTGGGAGTTGGGGGG//CGTGGG-3', with the double forward slash indicating the where the double strand break occurs. This was inserted into a modified vector that we designed, to combine elements of two other plasmids, vasaFUS (Baena-Lopez et al., 2013) and pCFD3-dU6:3gRNA (Port et al., 2014). This modified plasmid was designed to reduce the number of plasmids carrying different elements. The pTVCherry plasmid is to be linearized with the SceI (NEB, Cat #R0694) in order to recombine into the double strand break under the vasa promoter. Also the CRISPR RNA sequence that associates with the Cas9 enzyme (Cong et al., 2013) with the unique guide RNA sequence included is now on in the same plasmid under the U6:3 promoter to reduce the number plasmids needing to be injected down to two. For this we used PCR to take two sections from vasaFUS, a section containing the SceI gene and the vasa 3' UTR sequence using the primers: For = 5'-GGATGGGATCAAGATCG-3' and Rev = 5'-ATGATGGACCAGATGGGTG-3'. For the PCR of vasa promoter including the 5' UTR section the primers were: For = 5'-CCTGCAGCTGGTTGTAGGTG-3' and Rev = 5'-CACCACACTGGACTAGTAG-3'. Both these fragments then had overhangs added for cloning into pCFD3, side by side, by PCR using primers: SceI + vasa 3' UTR For = 5'-TGATCCACTAGAAGGCCTGCAGCTGGTTGTAGGTG-3' and Rev = 5'-GTGTACCGAATTAGGCACCACACTGGACTAGTAGGTACC-3'; vasa 3' UTR For = 5'-AAAAAAAATATCAATGGATGGGATCAAGATCGCCAAAAAAG-3' Rev = 5'-TGGACTAGTAGGTACATGATGGACCAGATGGGTGAGG-3'. The pCFD3 vector was then linearized using SpeI (NEB, Cat #R0133) and the two vasa/SpeI fragments cloned into it at the same time using the Infusion HD Liquid Cloning Kit (Clontech) as previously described. Colonies were screened by PCR for insert size as before and finally sequenced to check orientation and specificity of cloning before being sent for injections. The pCFD3 backbone with the SceI gene added with the vasa promoter and UTRs was ready to have guide RNA added for the specific site for targeting of Cas9. This was done by PCR linearizing the plasmid at the point where the bases were to be added using primers that contained 15bp overhangs to allow recircularization of the plasmid. The primers used to linearize and add guide RNA for *Hox-G* were: 5'-

GCTTAACTTAAACTTACAGTGTTTTAGAGCTAGAAATAGC-3' and 5'-ACTGTAAGTTTAAGTTAAGCAGGTCTTCTCGAAGACCCCG-3'. Once PCR linearized and cleaned with the Purelink® Quick PCR Purification kit, the plasmid was then recircularized with InFusion, transformed, miniprepped and sequenced as previously described (Fig.2.5).

The pTVCherry vector had 1.5 kb homology arms for both the 3' and 5' directions spanning outwards from the double strand break. The primers for amplification of these arms from w^{1118} *D.melanogaster* were: 3' arm For = 5'-CGTGGGGCTAAAGAAATGTC-3'; 3' arm Rev = 5'-CAGTTGTGCACTGAGCAACC-3'; 5' arm For = 5'-CCCCCCAACTCCCACTCCGC-3'; 5' arm Rev = 5'-AGGGTGAAATGTAGCTCCGC-3'. pTVCherry was linearized at each multiple cloning site (MCS) separately to insert the 5' homology arm first and then the re-linearized at the other MCS to insert the 3' homology arm. NotI-HF (NEB, Cat #R3189) was used to cut the 5' homology arm MCS and SpeI was used to linearized pTVCherry at the MCS site where the 3' homology arm was inserted. InFusion cloning was carried out as before to insert homology arms and recircularize the pTVCherry with overhangs added to each homology arm to introduce the necessary 15bp overhangs with GXL polymerase. The primers used to add the overhangs to the homology arms were: 3' arm + overhangs For = 5'-CGAAGTTATCACTA GCGTGGGGCTAAAGAAATGTCGT-3'; 3' arm Rev = 5'-GGAGATCTTTACTAGC AGTTGTGCACTGAGCAACCA-3'; 5' arm For = 5'-CCCGCGGTAGCGGCC CCCCCCAACTCCCACTCC-3'; 5' arm Rev = 5'-GCATGCAATGCGG CCAGGGTGAAATGTAGCTCCGC-3'. The plasmids were transformed into Top10 cells as before and checked for correct size on agarose gel after colony PCR. A correct size colony was grown in LB broth overnight, extracted with the Purelink® Quick Plasmid Miniprep and sent for sequencing in house as before (Fig.2.4). The plasmids were sent to BestGene Inc for injection into flies expressing Cas9 under the vasa promoter (BDSC #55821 - y[1], M{vas-Cas9.RFP}ZH-2A, w[1118]), using 250g/ul of modified pCFD3-dU6:3gRNA and 500g/ul of pTVCherry with homology arms. 300 embryos were injected, with 80 surviving and 4 transformants were produced. Flies expressing the mini-white marker were sent back and homozygozed by single male and virgin female crosses and waiting for two generations without white eyes. The homozygous flies were sequenced to check that the pTVCherry plasmid inserted to the correct place by PCR with GXL polymerase and ethanol precipitated before being sent for sequencing in house. The homozygous flies were raised at different temperatures and frozen before having their heads imaged. Figure 2.4 from Clontech summarizes this. The flies that were homozygous for the pTVCherry insert were crossed to Cre flies - y[1] w[67c23] P{y[+mDint2]=Crey}1b; D[*]/TM3, Sb[1] (Bloomington #851) to test the effects of partial or imprecise excision of the insert, by screening for flies that had lost mini-white and therefore had white eyes.

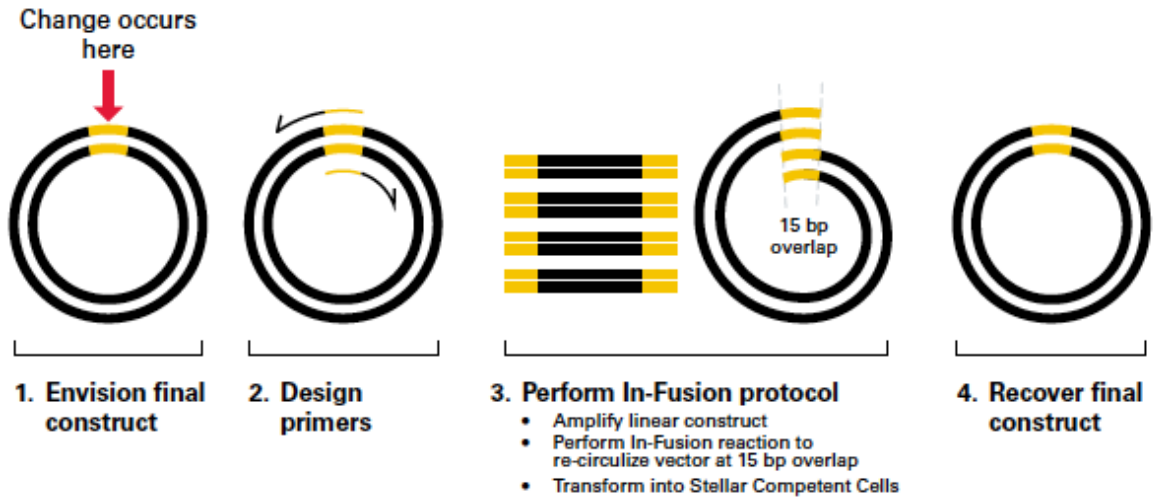


Figure 2.4. InFusion summary for making changes to plasmids. We modified this to insert 20nts (5'-GAGTGGGAGTTGGGGGGCGT-3') to the modified pCFD3 plasmid. Image taken from InFusion application notes (http://info.clontech.com/Mutagenesis-Tech-Note-Sign-up-2014.html).

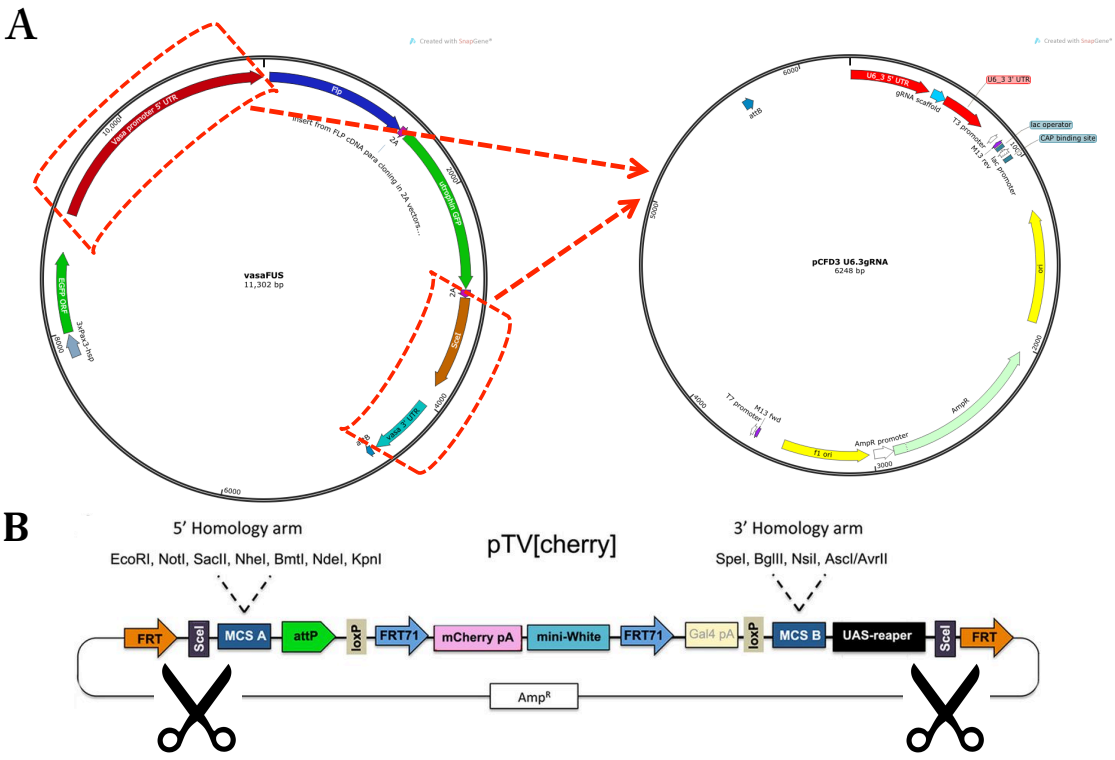


Figure 2.5. Plasmid design for CRISPR/Cas9 injections. A) The two red dashed boxes show the vasa promoter and 5' and 3' UTRs that were taken from vasaFUS and placed into the pCFD3 vector. B) the pTVCherry vector with scissors to indicate SceI cut sites for linearization.

2.12 FLP manipulation of *Hox-G*

The PBac(WH)2 and PBac(WH)3 flies were crossed to bring one insertion over the other, using the *e-* and *Hu-* on the balancer as negative selection markers. The flies were then expanded slightly and several generations checked for continuity of mini-white expression and no white-eyed flies, to ensure the insertions were stable on sister chromatids. These flies were then crossed to Flippase (FLP) - y^1 w^{67c23}; MKRS, P{hsFLP}86E/TM6B, P{Crew}DH2, Tb^1 (Bloomington #1501) in order to induce unequal homologous recombination. Due to the orientation of the PBac elements and the position of the FRT sites, the flies were predicted to have either four copies of min-white and a partial deletion of the second exon, or no mini-white, just *rosy* from the FLP insert and therefore a partial duplication of the second exon. The only flies that hatched had *rosy* eyes and no mini-white could be detected, so we took that to mean that the deletion was lethal and the duplication was the causation of the mutations observed (Fig.3.6.10).

2.13 Segmentation gene crosses for lncRNA expression investigations

Flies used with mutations on segmentation genes were: Kr^1 (Bloomington #3494), Kr^{17} (Kyoto #101324), eve^1 (Bloomington #5344), eve^3 (Bloomington #299), h^{25} (Bloomington #1781). All flies were raised on standard cornmeal medium at 25°C with 12h light and dark cycles. Flies lay for 7hrs in polypropylene cages on apple juice agar supplemented with yeast paste to facilitate laying and were collected, fixed, stained and imaged as previously described. These alleles are all homozygous lethal in ¼ of laid eggs, identified by *ftz* expression pattern. Slides were screened for mutations and imaged and lay at different temperatures if temperature sensitive alleles had been reported.

3. RESULTS

3.1 Comparative analysis of lncRNA clustering and cluster conservation in drosophilid genomes

In order to identify regions in drosophilid genomes that may have a conserved enrichement for lncRNAs, we searched for lncRNA clustering in both *D. melanogaster* and *D. virilis*. These two drosophilids are ~63 million years divergent (Tamura et al., 2004) and although the sequence of lncRNAs may not be well conserved, we used the orthologous protein-coding genes within clusters to identify syntenic and potentially conserved orthologous lncRNAs. By identifying the regions most highly enriched for lncRNAs in both species that also retain clusters of orthologous genes, we are hoping to identify clusters of conserved functional lncRNAs.

We began by identifying clusters from the 2,470 currently annotated lncRNAs in the *D. melanogaster* genome (FlyBase release 6.08). We first attempted to determine the distance between adjacent lncRNAs to empirically estimate cluster size distribution. We mapped the length distribution of distances between all adjacent lncRNAs and plotted a histogram of numbers of lncRNAs across each distance. This indicated that the distance between two adjacent lncRNAs, termed the cutoff distance, should be 4 kb as above 4 kb the frequency of enriched clusters of lncRNAs rose significantly, indicating this should be the distance to separate one cluster from the adjacent cluster. However, this does not reflect the variation in intervening protein coding gene size. For example, the distances between lncRNAs within established gene clusters, such as Hox clusters where average protein coding gene sizes are large, showed that distances of >20 kb between lncRNAs was more appropriate. As lncRNAs can be separated by large protein-coding genes, but still belong to a cluster; we removed the protein-coding genes length when calculating the distances between lncRNAs. Several cutoff distances were tested until reported lncRNA clusters closely matched visibly enriched regions of lncRNAs when examining the distributions in a genome browser and seeing denser regions of clustering, as well as falling within the confines of known gene clusters, such as Hox or enhancer of split (E[spl]-C), without encroaching on neighboring regions. We have reported the numbers of lncRNAs in the 20 most enriched (clusters with the highest numbers of lncRNAs before a 25 kb space separates two adjacent lncRNAs) clusters when using both a 25 kb cutoff distance (Fig.3.1.1.A). A 100 kb cutoff distance is shown for comparison of enriched regions, as this was the distance later used for the *D. virilis* analysis (Fig.3.1.1.A). A cutoff of 25 kb was found to be the most suitable distance for resolution of visibly enriched clusters and also identified some known protein-coding gene clusters, without excessively overlapping adjacent regions (Fig.3.1.1.B-F). A gene cluster consists functionally related genes that are adjacently arranged physically closer than would be expected on the chromosome. The number of genes required to form a cluster can range from 2-100s (Lee and Sonnhammer, 2003), so we therefore reported all clusters with a distance of <25 kb between them. The top 20 out of 594 were later investigated for possible conserved clustering of syntenic mRNAs.

Figure 3.1.1 shows the 20 clusters containing the largest numbers of lncRNAs when both the 100 kb and 25 kb cluster cutoff was applied. The cluster IDs are comprised of the cluster cutoff size and the numbers of lncRNAs in the cluster (Fig.3.1.1.A). At the 100 kb cutoff most of the genome falls within large lncRNA clusters (Fig.3.1.1.B-F). However, the lncRNA distribution in these large clusters is very inhomogeneous and more discreet regions of clustering can be clearly seen if we examine the lncRNAs found in these cluster (blue dashes Fig.3.1.1.B-F). Therefore, we incrementally reduced the cutoff distance by 5 kb intervals until the clusters reported smaller regions that resembled the visibly denser regions. The regions found to contain the 20 highest amounts of lncRNAs when a cutoff of 25 kb was applied are shown as blue blocks across each chromosome arm (Fig.3.1.1.B-D). Chromosome arm 2R and X did not have any of the top 20 clusters with the 25 kb cutoff, although they did have multiple clusters at a 100 kb cutoff. The majority of the clusters with a 25 kb cutoff are found across chromosome arm 2L. Interestingly, this chromosome arm is also considerably smaller than any of chromosome 3's arms. We considered if the increase in clusters on chromosome 2R could be due to increased density of all genes, maybe as a result of less duplication events or more deletion events than chromosome 3. However, this seems unlikely as it's size matches chromosome X and this does not have any of the top 20 clusters with the 25 kb cutoff, leading us to believe the clustering is not merely a result of compaction. Remarkably, a few of the enriched clusters from the 25 kb cutoff are next to each other on chromosome arms 2L and 3R. In some cases this is likely to be due to the reduction in distance set between clusters, however, 25 kb cluster IDs 25-10, 25-11 and 25-12 were still separated when the 100 kb cutoff limit was applied (100-3, 100-4, 100-5), indicating a definite separation across this region.

We are confident that this method has identified small regions of enriched lncRNAs, however, are aware that there could be improvements or more empirical ways of testing this. We can see that 100 kb is not suitable and 25 kb does seem appropriate as this indicates that some of the most highly enriched lncRNA regions are within known protein coding gene clusters, such as Hox and (E[spl]-C). These gene clusters are regulated by PcG and TrxG proteins, which have been shown to have strong links to lncRNA and so it seems possible that these regions could be rich in lncRNAs that may be involved in the regulation of these complexes. This demonstrates that some known regions regulated by PcG and TrxG proteins are enriched for lncRNAs and could therefore be a good indication that these lncRNAs are functional. Our algorithm has been designed to be able to run with any cutoff window size to be used in different organisms that may have bigger genomes and therefore gene density, although could be improved with further testing to find the most appropriate cutoff distances between clusters.

A

100kb ID	100-1	100-8	100-3	100-27	100-13	100-0	100-34	100-20	100-33	100-28	100-10	100-12	100-36	100-38	100-5	100-23	100-11	100-19	100-56	100-39
lncRNAs	231	173	163	148	135	134	134	119	94	76	75	68	64	60	53	51	50	48	45	45

25kb ID	25-5	25-4	25-19	25-14	25-2	25-12	25-15	25-0	25-8	25-6	25-16	25-11	25-7	25-20	25-10	25-17	25-9	25-18	25-13	25-3
lncRNAs	25	25	23	22	22	20	19	18	18	18	17	17	17	16	16	15	15	14	14	14

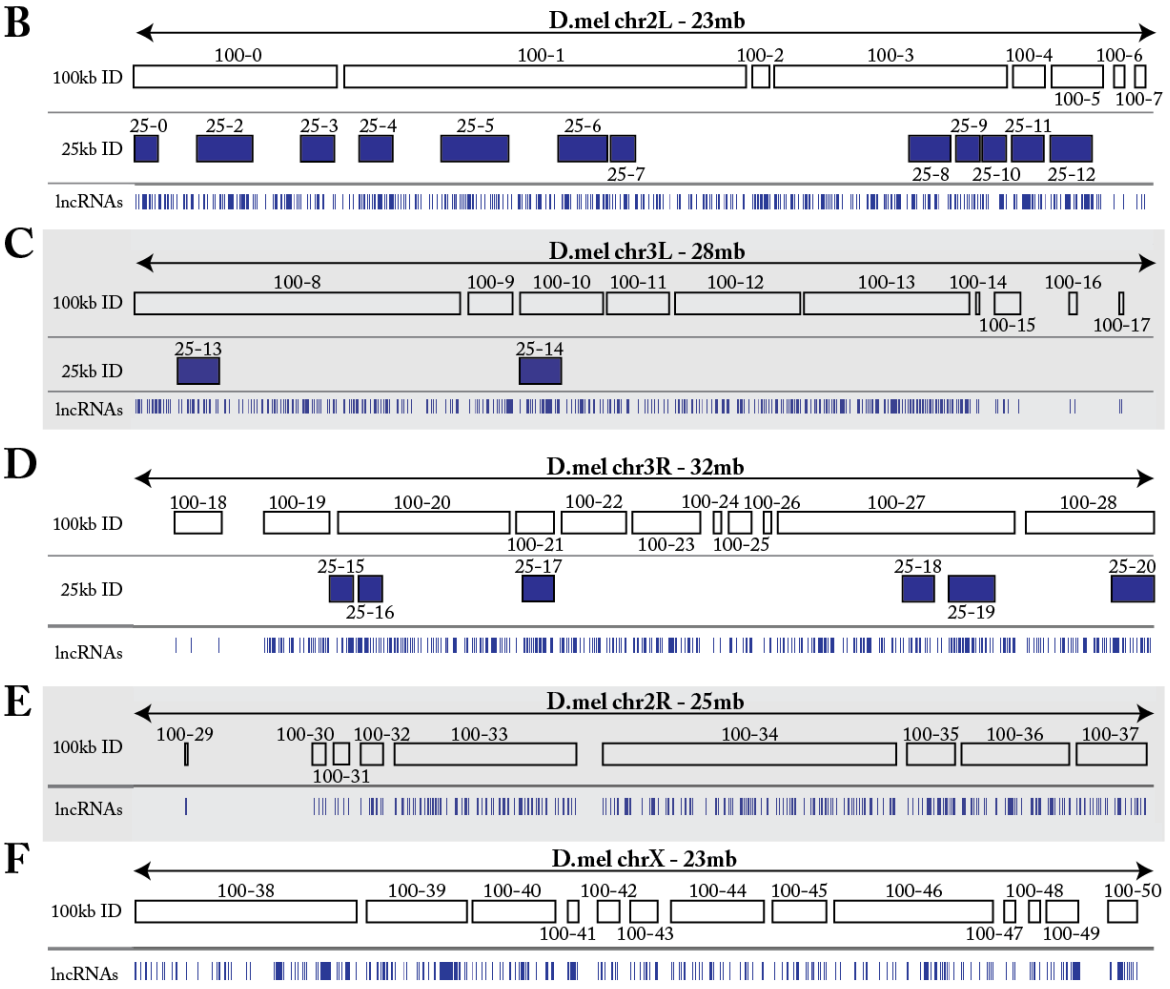


Figure 3.1.1. *D. melanogaster* lncRNA clusters with 25 kb and 100 kb cutoff distances. Chromosome X, along with chromosome arms 2L, 2R, 3L and 3R are depicted with all clusters identified using both a 100 kb cutoff to break each cluster, or 25 kb cutoff to increase resolution. The 20 clusters with the highest amount of lncRNAs are listed using cluster ID's for reference (arbitrarily numbered in order along each chromosome) and the numbers of lncRNAs found in each cluster for each 100 kb and 25 kb cutoff limits (A). Empty blocks on each chromosome show the cluster with the highest numbers of lncRNAs on each chromosome using a cutoff of 100 kb. Dark blue blocks with numbers are the top 20 clusters when a 25 kb cutoff limit was imposed between each cluster (B-D). The blue dashes show dashes for each lncRNA contained within the 100 kb clusters.

To investigate if there was an association of other clustering or enrichment of gene functions with the clusters having the highest numbers of lncRNAs, we investigated the protein-coding genes found within the top 20 highest. We did this by using PANTHER (Protein ANalysis Through Evolutionary Relationships) (Mi et al., 2016), an up to date gene classification system containing a large database of curated genes and functionally related families as part of the Gene Ontology Reference Genome Project (Reference Genome Group of the Gene Ontology, 2009). Using PANTHER, we evaluated the protein-coding genes from each clusters for functionally related biological processes. Gene ontology (GO) is a set of defined terms that provide descriptions of genes properties divided into 3 main broad descriptions; molecular function, biological process, or cellular component. These properties can be reported on a number of different levels giving either a broad overview or detailed descriptions. For example, a level below biological process to give a broad description could be 'developmental process' that could then be further subcategorized into many more specific descriptions, for instance 'pattern specification process', that can further be specified into more terms, until a very specific description, such as 'head segmentation' is reached.

Figure 3.1.2 is a graphical display of the fold enrichment of each significantly shared GO-Slim terms for each of the protein-coding gene lists, indicating shared functional characteristics. The majority of the GO-Slim terms found overrepresented in each list of protein-coding genes are unique and those that do have some overlap are those that also have a lower enrichment. The GO-Slim terms that have the highest fold (>50) of enrichment are:

- Female Gamete Generation
- Defense Response to Bacterium
- Digestive Tract Mesoderm Development
- Embryo Development
- Spermatogenesis
- Segment Specification
- Pattern Specification Process
- Regulation of Sequence-specific DNA binding
- Transcription Factor Activity
- Cellular Glucose Homeostasis
- Muscle Organ Development
- Regulation of Liquid Surface Tension
- Chromatin Remodeling

The GO-Slim terms that are found to be most highly associated with genes within the enriched lncRNA clusters have several things in common, mainly that all occur during embryogenesis, some exclusively. The terms limited to embryogenesis are female gamete generation, digestive tract mesoderm generation, embryo development, spermatogenesis, segment specification, pattern specification process and muscle organ development as these terms reflect various stages of generation of an embryo (gamete generation and spermatogenesis) (Beck, 2002; Beuchle et al., 2001; Chen et al., 2005b; Enriquez et al., 2010; Foronda et al., 2012; Maitre and Heisenberg, 2013; Moazed and O'Farrell, 1992; Tixier et al., 2013). TF activity, regulation of sequence-specific DNA binding and chromatin remodeling have been discussed in the context of Hox gene regulation and lncRNAs during embryogenesis and link to the pattern specification process, but these processes also occur in adulthood to maintain homeostasis. The regulation of liquid surface tension is needed for oocyte positioning and mesoderm/endoderm cell internalization in Drosophila (Maitre and Heisenberg, 2013) and later in life for processes such as wound healing. The defense to bacterium is part of the Toll-signaling pathway, also recognized for its role in dorsoventral patterning, but also defends against fungal and gram-positive infections during development (Hetru et al., 2003). Finally, cellular glucose homeostasis is regulated during muscle development and then to maintain normal glucose concentrations in adult flies (Tixier et al., 2013). It may not be surprising that the lncRNAs clusters are linked to development as ~44% of the total protein-coding genes known for *D. melanogaster* are expressed during embryogenesis (Tomancak et al., 2007) and many lncRNAs have been shown to be involved in developmental processes in a diverse range of species (Fatica and Bozzoni, 2014). This could well be linked to the large amount of chromatin reorganization that must occur throughout development in order to activate and silence various genetic pathways. Furthermore, it is not surprising that there is evidence of clustering linked to development, as it is likely to be more energy efficient to modify a large section of chromatin into open and closed conformations for rapid gene regulation.

Interestingly, one of the highly enriched lncRNA clusters, 25-15 (25 kb cutoff), is a well known gene cluster containing the developmentally important and deeply conserved Hox gene cluster (Garcia-Fernandez, 2005). The Hox cluster contains a notable 19 lncRNAs, but has the least protein-coding genes in the top 20 lncRNA clusters (Fig.3.1.5.A). Previous studies have shown that Hox gene expression initiates in response to the early expression of gap genes that are associated with segment specification (as well as female gamete generation), but before those necessary for digestive tract mesoderm development and muscle organ development (Gebelein et al., 2004). Furthermore, it is widely established that TF products from segmentation specification genes regulate Hox genes and Hox gene TFs are responsible for regulation of genes involved in digestive track and muscle organ development (Beck, 2002; Carroll et al., 1988; Enriquez et al., 2010), placing Hox genes at the center of the highest enriched biological processes from the GO-Slim term analysis.

PANTHER's GO term analysis indicates that a large portion of lncRNAs clustered transcription can be linked to embryonic development the *D. melanogaster*. We decided to explore if we could identify similar lncRNA clusters in the distantly evolutionary related species of *D. virilis*. There is no current RNA seq data available from stages of *D. virilis* embryogenesis and there are currently just 565 annotated lncRNAs in FlyBase for this organism. Using the data from *D. melanogaster* as a guide, we sequenced total RNA from embryonic developmental times where we anticipate identification of many more lncRNAs to combine to the current annotation. We will then compare the protein-coding genes found in *D. virilis* lncRNA clusters obtained from a new larger dataset. Current evidence suggests that initial Hox gene expression is detected between stages 4-6 of embryogenesis in *D. melanogaster*, with expression of many segmentation, muscle, organ, and digestive tract genes also coinciding with these stages (Tomancak et al., 2002). We therefore carried out total RNA-sequencing of the corresponding stages of *D. virilis* development, as currently the only total RNA-sequencing datasets available for *D. virilis* were from adult flies.

We extracted total RNA from *D. virilis* and *D. pseudoobscura* embryos that were aged to stages 4-6 (Campos-Ortega and Hartenstein, 1997). The RNA was poly(A)+ selected and stranded libraries were prepared and tested for quality and quantity, before being sequenced in the Genomic Technologies Core Facility at the University of Manchester (see methods section 2.4 for details). The total RNA was sequenced and the reads mapped to the *D. virilis* genome with TopHat2 (Kim et al., 2013). Cufflinks (Trapnell et al., 2010) was then used to annotate all potential transcripts throughout the genome from the reads that had been mapped with TopHat2. Cufflinks reports overlapping stacks of reads as a transfrag and annotates splice junctions, allowing identification of introns and exons. Cufflinks also allows reporting of low abundance reads facilitating detection and filtering based on FKPM values. To extract the novel lncRNAs from *D. virilis*, we first utilized the Cuffcompare tool, from the Cufflinks package (Trapnell et al., 2010). This tool compares the currently annotated lncRNAs from the *D. virilis* genome downloaded from FlyBase to the Cufflinks annotated embryonic RNA-seq.

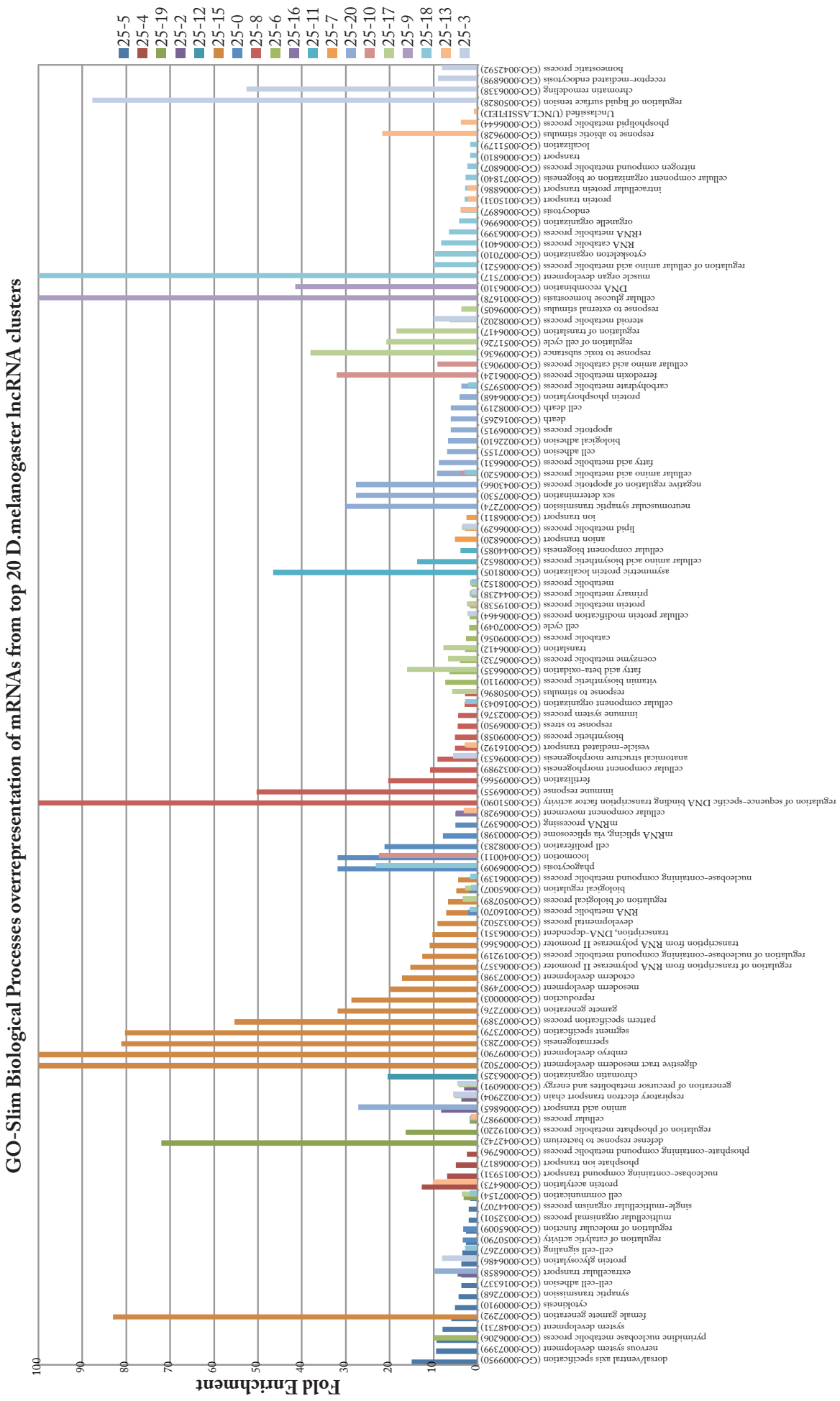


Figure 3.1.2. Legend next page

Figure 3.1.2. PANTHER GO-Slim Term analysis of mRNAs from top 20 lncRNA clusters in *D. melanogaster*. The protein-coding genes from the clusters found to have the highest numbers of lncRNAs were analyzed for overrepresentation of biological processes using PANTHER GO-Slim terms (www.pantherdb.org). The graph displays significantly (P=<0.05) overrepresented GO-Slim terms of the protein-coding genes found in each of the top 20 lncRNA cluster compared to the background frequency of annotated GO-Slim terms for all genes of the *D. melanogaster* genome. The fold enrichment of GO-Slim terms compared to the background set is shown on the y axis, with those displayed as 100, actually being >100. The GO-Slim terms and their identifiers are displayed on the x-axis. The key indicates the *D. melanogaster* lncRNA cluster ID that matches the 25 kb ID from Fig 3.1.1. The clusters that share GO-Slim terms are shown bars that overlap by 90% and all have <25-fold enrichment.

Each transfrag from the Cuffcompare report is given a 'class code', which indicates its context relative to transcripts from the current annotation (see Fig.2.2). The class codes that most reliably indicate a novel transcript, that is not likely to be a novel isoform or alternate UTR of a previously annotated gene, are class codes 'u' (unknown intergenic), 'x' (exonic overlap with reference on opposite strand) and 'i' (a transfrag falling entirely within a reference intron). We therefore retained only those transcript annotations that have 'u', 'x' and 'I' class codes.

We removed all transfrags that had an FKPM <1. This cutoff has been found to be robust and conservative for transcript detection of low-level mRNAs (Mortazavi et al., 2008). The transfrags that that have FKPMs >1 were then scrutinized for additional evidence that they were lncRNAs. We also removed all transfrags that were <200nts, as this will generally distinguish lncRNAs from other well-known smaller ncRNAs, such as microRNAs, PIWI-associated RNAs and siRNAs (Dinger et al., 2008b). We tested the protein-coding potential of each transfrag using the coding potential calculator (CPC) and removed any transfrags that the program reported as coding (Kong et al., 2007).

The transfrags were also searched against Pfam for the potential to code for protein domains conserved in protein families (Finn et al., 2016) and BlastX for amino acid sequence similarity to known proteins (Altschul et al., 1990; Gish and States, 1993). Rfam was used to investigate the RNA sequences of each transfrag for matches to known consensus RNA secondary structures or sequence similarity to multiple sequence alignments (MSAs) of RNA families (Nawrocki et al., 2015). Along with Rfam, tRNAscan was also used to confirm that none of the transfrags left could be classed as a tRNA (Schattner et al., 2005). The RepeatMasker program was used to screen for repeats or low complexity DNA sequences (Smit, 2013-2015). Any transfrags with significant matches according to any of these programs were removed.

The initial number of transfrags that were identified from RNA-seq with Cufflinks was 82761. 3032 of these did not correspond to currently annotated genes and were investigated for the possibility of being categorized as lncRNAs. After filtering, 542 novel lncRNAs were identified from the transcriptome of *D. virilis* stage 4-6 embryos (Campos-Ortega and Hartenstein, 1997) (Table.3.1.A). These were further categorized based on location of transcripts relative to annotated genes, either as intergenic (class code u) or antisense (class code x). No transfrags marked class code i remained after filtering. The newly identified transcripts were also split into single exon and multi-exon and compared to mRNAs for exon numbers and size distribution in (Fig.3.1.3).

RNA-seq reads	124607186
Cufflinks Transfrags	82761
FlyBase annotated lncRNAs (all multi-exon)	565
Cuffcompare novel transfrags	3032
Classed as lncRNA after filters	542
Novel intergenic single exon	352
Novel intergenic multi-exon	69
Novel antisense single exon	91
Novel antisense multi-exon	30
Total lncRNAs = novel + annotated	1107

Table 3.1.1. Summary of novel lncRNAs identified from RNA-seq in *D. virilis*. Numbers of lncRNAs identified using RNA-seq from stage 4-6 *D. virilis* embryos (Campos-Ortega and Hartenstein, 1997). Cuffcompare identified transfrags that were compared to the current lncRNA annotation with Cuffcompare. These were then filtered for quality and any indication they may belong to another class of RNA.

Human lncRNAs have been found to predominantly consist of 2 exons (46%), compared to just 6% of protein-coding genes having 2 exons (Derrien et al., 2012). A preference for 2 exon lncRNA transcripts was also found in filamentous fungus *Neurospora crassa* (Arthanari et al., 2014) indicating a strong conservation of this tendency across highly divergent species. Our data shows that single exon lncRNAs account for the largest fraction of total lncRNAs, however, the *D. virilis* genome is still poorly assembled with many regions of unknown sequence and many unlinked contigs. This leads to reads that cannot be mapped correctly as they are split between contigs and therefore increases the numbers of single exon transfrags (Table.3.1.A). However, if instead we just consider the multi-exon transfrags, then we can see an overall preference for 2 exon lncRNAs that still clearly contrasts with the distribution of exon numbers in mRNAs that mostly ranges between 10-20 exons (Fig.3.1.3.A-B).

The length distributions of the total lncRNAs is shown in figure 3.1.3 C-F and indicates similar lengths for single exons from both lncRNAs and mRNAs, although the majority of single exon mRNAs fall below 250nt (Fig.3.1.3.E-F). However, the cutoff for lncRNAs is 200nts, so although there could be lncRNA transcripts that are <200nts, they would not be included in this analysis as this is not how they are currently classified (Fig.3.1.3.E-F). It is unsurprising that multi-exon mRNAs are considerably longer than multi-exon lncRNAs (Fig 3.1.3.C-D) as in our data they may be split across unlinked contigs and the average size of mRNAs are longer than lncRNAs, for example, 2880nts for *D. melanogaster* mRNA transcripts vs. 994nts for lncRNA transcripts (FlyBase release notes r6.12).

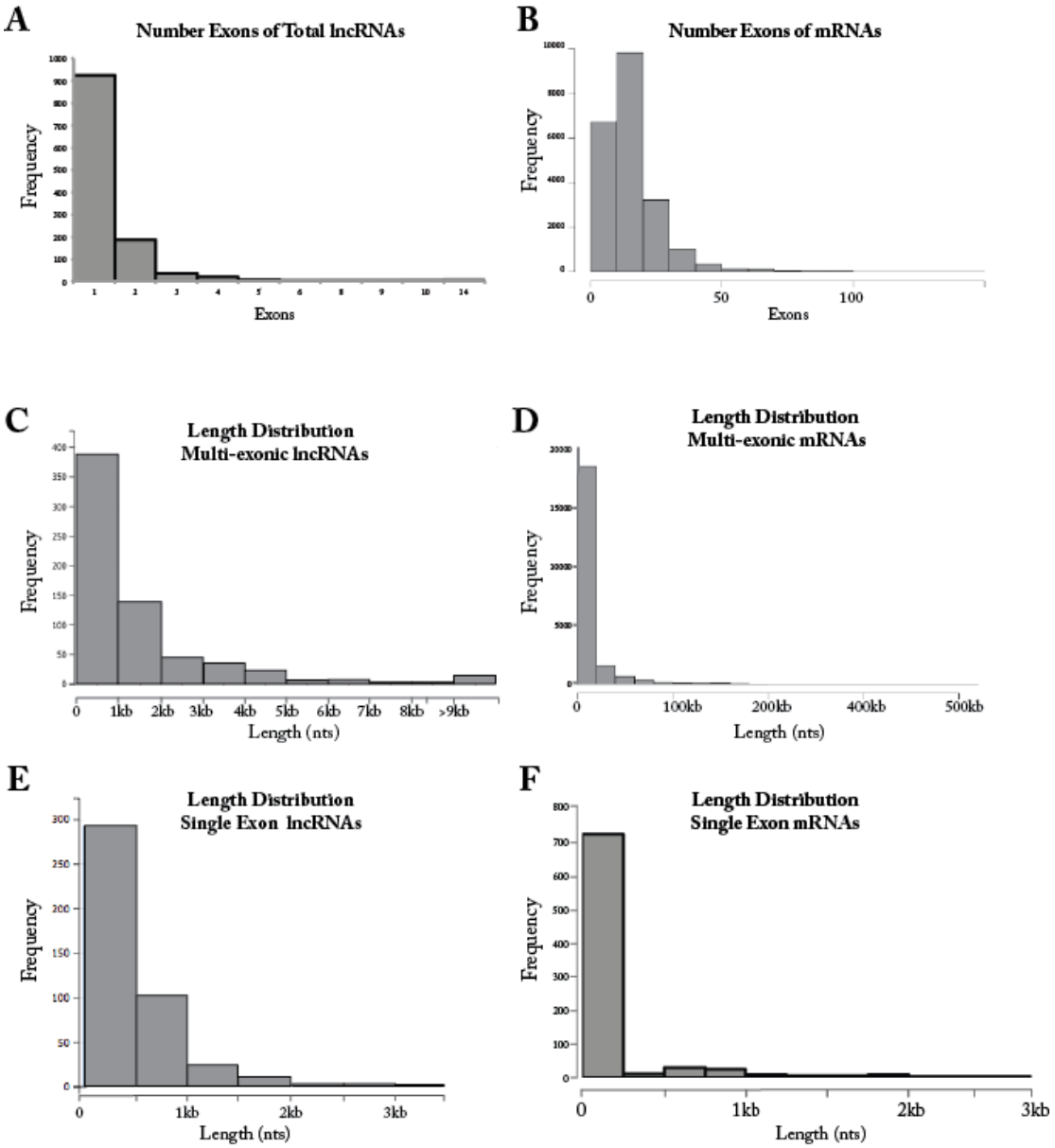


Figure 3.1.3. Properties of lncRNAs identified in *D. virilis* embryos. Numbers of exons of total lncRNAs (novel and previously annotated by FlyBase) were compared to the numbers of exons found in mRNAs (A-B). The length distributions of multi-exonic lncRNAs and multi-exonic mRNAs are shown (C-D) along with the length distributions of single exon lncRNAs and mRNAs (E-F).

Using the novel lncRNAs identified from our RNA-seq combined with those that were already annotated on FlyBase, we identified 186 lncRNA clusters within the *D. virilis* genome when imposing a 100 kb cutoff (Fig.3.1.4). The larger cutoff is empirically determined and needed in part due to the *D. virilis* genome being approximately double the size of *D. melanogaster's* genome (male=339mb, female=307mb vs. 170mb and 175mb respectively), due to the incomplete assembly of the genome, and because there are ~1/3 of the number of lncRNAs identified in these early developmental stages. However, each genome has a similar number of protein-coding genes identified, 17,674 and 17,728 respectively and therefore the *D. virilis* genes are likely to be more

dispersed. A cutoff of 100 kb separating lncRNAs was found to generate clusters that had similar numbers in each cluster to those identified in *D. melanogaster* when the 25 kb cutoff was used (Fig.3.1.1.A). Some of the *D. virilis* scaffolds shown in figure 3.1.4 B-E are particularly good for comparison as they are well assembled and similar in size to chromosome arms from *D. melanogaster.* These allow identification of clusters that are much less likely to have been split over smaller scaffolds (Fig3.1.4.F-J).

As for *D. melanogaster*, the 20 *D. virilis* clusters with the highest numbers of lncRNAs were used for further analysis. The mRNAs found in *D. virilis* clusters were used to identify the *D. melanogaster* orthologs from FlyBase as curated by OrthoDB (http://www.orthodb.org). The protein-coding gene orthologs from *D. virilis* were then compared to the protein-coding genes that had been identified from the lncRNA clusters in *D. melanogaster* to investigate the conservation of orthologs in lncRNA-enriched clusters. Comparing the synteny of genes found in these regions of *D. virilis* with *D. melanogaster* may reveal a split in clusters of genes across ends of scaffolds and this could allow tracking of movement and separation of gene clusters between the two species (Fig.3.1.5.A). Many genes have yet to be accurately identified in the *D. virilis* genome, as this species has not been examined as intensively as *D. melanogaster*. However, there are clear regions of micro-synteny that have been identified in both species as being rich in lncRNAs (Fig.3.1.5.A).

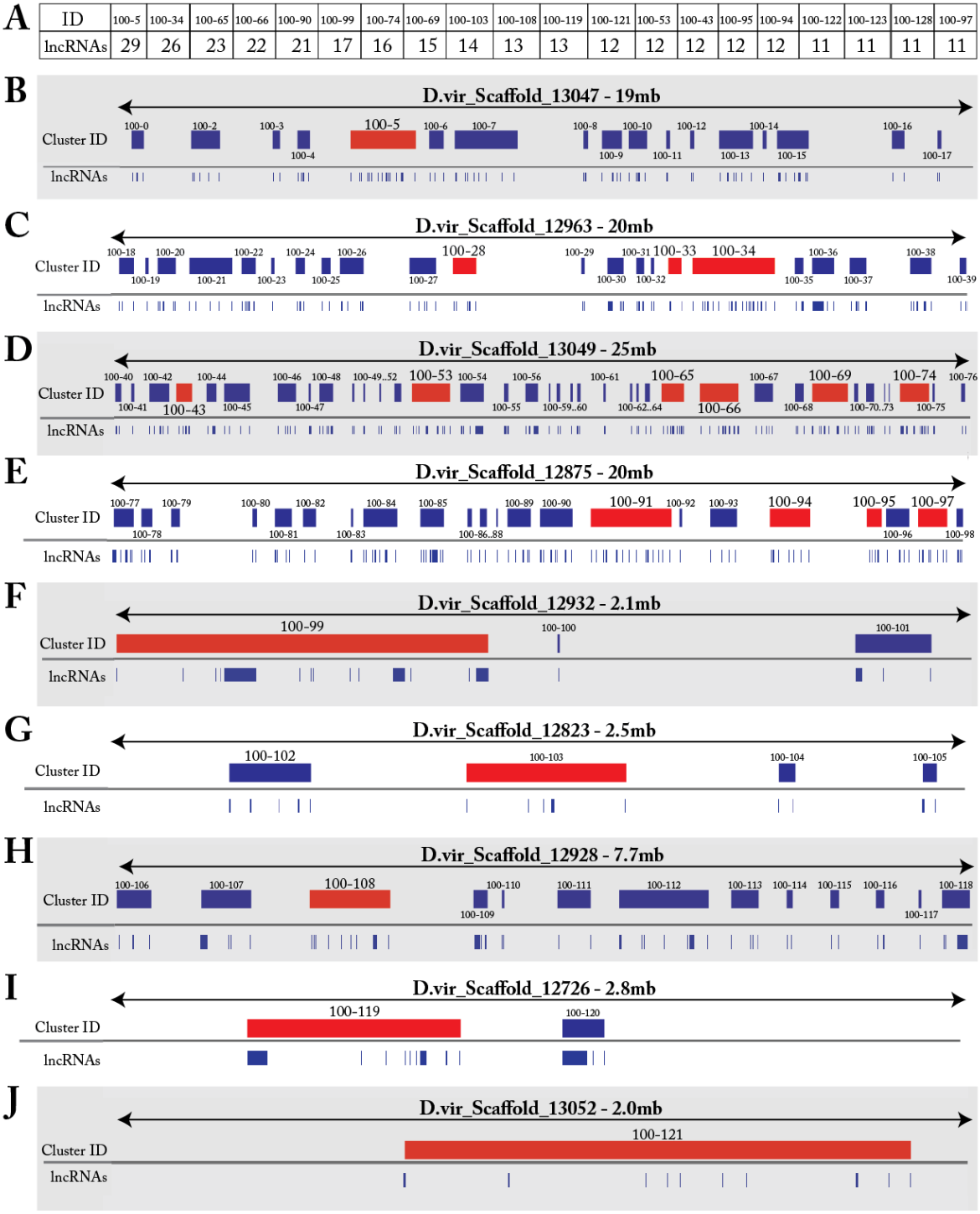


Figure 3.1.4. Identification of lncRNA clusters in *D. virilis*. The same clustering algorithm used to discover lncRNA clusters in *D. melanogaster* was applied to the novel identified lncRNAs and currently annotated lncRNAs from *D. virilis*. A cutoff length of 100 kb was used to separate each cluster to identify similar numbers of lncRNAs per cluster as those found in *D. melanogaster* (Fig 3.1.1. A), when using 25 kbs as a cutoff length. Table A lists the 20 cluster IDs with the highest numbers of lncRNAs and the highest 17 clusters are depicted on the scaffolds that they are found on (B-J). The red boxes are the clusters that contained the 20 highest numbers of lncRNAs and the blue boxes are all other clusters. Each lncRNA matched to a cluster is shown by blue dashes underneath along with their cluster ID number.

D. melanogaster cluster 25-19 and *D. virilis* cluster 100-5 are good examples of regions with some of the highest number of lncRNAs that also contain clusters of orthologous protein-coding genes (highlighted in light green in Fig.3.1.5.A). This cluster is also found to be the most significantly overrepresented by GO-Slim analysis for the term 'defense response to bacterium' (GO:0042742) (Fig.3.1.2). In *D. melanogaster*, cluster 25-19 is on chromosome arm 3R along with the 7th most enriched lncRNA cluster, 25-15. These 2 clusters are separated by ~20mb on chromosome 3R in *D. melanogaster,* but in *D. virilis* are almost directly adjacent to each other on scaffold_13047 and are found within the most highly enriched lncRNA cluster, 100-5 (Fig.3.1.5.B). The protein-coding genes found in *D. melanogaster* cluster 25-15 include four belonging to the ANT-C of Hox genes. Not all of the Hox genes in this cluster were detected in *D. melanogaster*, but they were in the most highly enriched lncRNA cluster, 100-5, in *D. virilis*. The cluster 25-19 that lies next to 25-15 in *D. virilis* (100-5) includes genes from the (E[spl]-C), a deeply conserved cluster of developmental genes involved in neurogenesis (Wurmbach et al., 1999).

We took the cluster with the highest number of lncRNAs in the *D. melanogaster* genome (25-5) and identified the protein coding genes in this cluster to compare to all of the protein coding genes in the top 20 most lncRNA enriched lncRNA clusters in *D. virilis*. The protein coding genes found in *D. melanogaster* 25-5 does not seem to be conserved as a cluster in the *D. virilis* genome as the matching protein-coding genes identified in both datasets are quite scattered in the *D. virilis* clusters (highlighted in purple Fig.3.1.5). The genes highlighted in orange, cluster 25-2 in *D. melanogaster,* have been split in *D. virilis* (clusters 100-34 and 100-28) but still seem to have several genes retained together. However, there is also no clear reason based on the genes known functions, but many of these genes have not been investigated yet and there could be unknown roles that could link their functions. Corresponding clusters 25-14 and 100-74 have just a few shared protein-coding genes (Fig.3.1.5.A - yellow) and this cluster was not found to have any overrepresented GO-Slim terms. However, there is no obvious link between these genes. The relative proximity of *D. melanogaster* cluster 25-13 to cluster 25-15 (turquoise and yellow) on chromosome arm 3L corresponds to the *D. virilis* cluster 100-43 and 100-74 on scaffold_13049, although further separated. *D. melanogaster's* cluster 25-2 on chromosome 2L seems to have been split over *D. virilis* scaffold_12963, with the majority identified in cluster 100-34 and the remaining in 100-28. The fold enrichment for the GO-Slim terms for this cluster was <10 (Fig.3.1.2), which would suggest these genes do not have any known shared functions. Other notable clusters are *D. melanogaster* 25-12 and *D. virilis* 100-133. These contain a large cluster of histone proteins, overrepresented by the GO-Slim term 'chromatin organization' (GO:0006325) (Fig.3.1.2). This is particularly interesting given that literature has established one of lncRNAs key functions to be associations with histone modifying complexes as a method of gene regulation (Quinn and Chang, 2016; Tsai et al., 2010).

Using the GO term analysis on the protein-coding genes from *D. melanogaster* lncRNA enriched clusters, we identified that stages 4-6 would allow us to identify lncRNAs enriched clusters throughout the genome of *D. virilis*. The lncRNA clusters should correspond to some of the same clusters that had been identified in *D. melanogaster* (Fig.3.1.1). This also aligns with expression of the Hox complex and therefore allows identification of lncRNAs within the Hox complex to compare to *D. melanogaster*. As the *D. virilis* genome assembly still contains thousands of small scaffolds, many lncRNAs could be missed or incorrectly annotated if split between two scaffolds. However, the larger scaffolds where lncRNA clusters were identified are similar sizes to *D. melanogaster* chromosome arms and contain many of the protein-coding genes identified in *D. melanogaster* lncRNA clusters and therefore can still be used for comparison. As the *D. virilis* genome is bigger than *D. melanogaster*, the cutoff between lncRNA clusters was determined to be most accurate at 100 kb, as this also yielded lncRNA clusters with similar numbers of lncRNAs within (Fig.3.1.4). When comparing the protein-coding genes from each of the 20 most highly enriched lncRNA clusters from both species, many of the same regions have been detected, even though we also included the previously annotated lncRNAs from *D. virilis* (Fig.3.1.5). One of the most interesting outcomes from this was that the Hox complex and (E[spl]-C) are directly adjacent in the *D. virilis* genome as these two complexes are deeply conserved (Heffer and Pick, 2013; Wurmbach et al., 1999) and regulated by PcG and TrxG (Delest et al., 2012; Schaaf et al., 2013), therefore implying that lncRNAs within these regions could be linked to regulation of these genes based on previous knowledge of interactions between PcG and lncRNAs. The comparison of lncRNA enriched regions between the two species also demonstrates that several of the same regions remain enriched for lncRNAs over ~63 My of evolution (Tamura et al., 2004).

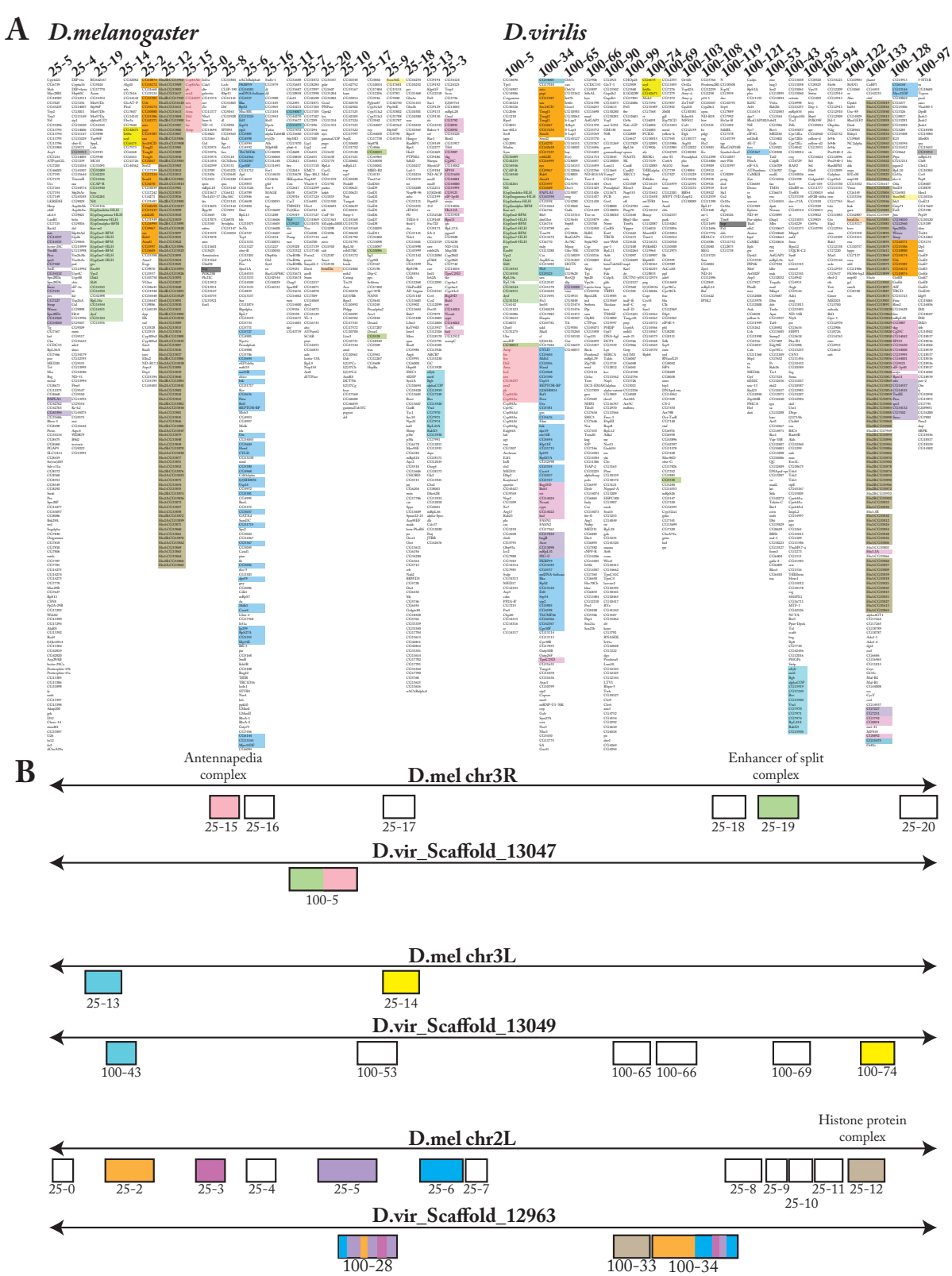


Figure 3.1.5. Comparison of matching orthologous mRNAs found in 20 highest clusters for *D. melanogaster* and *D. virilis*. A) Orthologs of protein-coding genes found in *D. virilis* top 20 lncRNA clusters were compared to the protein-coding genes from *D. melanogaster* top 20 lncRNA clusters and highlighted if found in both lists. Each cluster is colour coded to *D. melanogaster* for tracking into *D. virilis*. The clusters with the highest lncRNAs is furthest left. B) The colour codes are used to depict the cluster on *D. melanogaster* chromosome arms to the corresponding cluster on *D. virilis* scaffolds and how some have been split and distributed. Columns 100-34 and 100-90 of the *D. virilis* section A have 71 and 95 genes respectively not shown as no significant matches to *D. melanogaster* genes.

3.2 Identification of lncRNAs in the Hox complex of *D. melanogaster*

Based on the notable enrichment of lncRNAs found in the ANT-C in both *D. melanogaster* and *D. virilis*, along with literature establishing functions for lncRNAs in the Hox complex of several species including *H. sapiens* and *D. melanogaster* (Mallo and Alonso, 2013), we decided to investigate in detail the lncRNAs in the Hox complex. Total RNA of *D. melanogaster* had been previously sequenced as part of a large-scale project to investigate the transcriptome across 30 developmental stages, from 0-2hr embryos to adult flies (Graveley et al., 2011). We investigated the Hox complex for evidence of lncRNA transcription throughout each of the 2hr embryogenesis time windows (0-24hr) by visualizing the mapped reads with a genome data visualization tool, the Integrative Genome Viewer (IGV) (Fig.3.2.1-A). We originally identified 13 lncRNAs that had not been previously annotated in 2012 (Gindhardt et al. 2012) (Table.3.2.1) and updated the table to reflect the changes made in 2014 by the FlyBase genome annotators to include these transcripts as lncRNAs in FlyBase after verification of the work carried out by Young *et al* (Young et al., 2012).

We found the majority of lncRNA candidates demonstrated peak expression within the 4-6hr window of transcription (Table.3.2.1, Fig.3.2.1.B-C and Fig.3.2.2.B-C). Introns can be detected within some of the lncRNAs as split reads across two adjacent sites joined by blue lines, allowing gene models to be built (Fig.3.2.1-B). All potential lncRNAs were tested for coding potential using CPC, as previously described (Kong et al., 2007), followed by assessments for protein domains (Pfam) (Finn et al., 2016), consensus RNA structures (Rfam) (Nawrocki et al., 2015) and repetitive elements (Repeat Masker) (Smit, 2013-2015). The current list of lncRNAs found within the Hox complex includes the original annotations made using the 0-24hr RNA-seq profiles, denoted as 'Hox- (Letter)' (Table.3.2.1) and updated with 'CR' codes from the FlyBase curators updates.

We focused on novel RNA transcripts expressed during embryogenesis and only if there was a clearly visible transcript within a two-hour window, rather than at very low levels across several time points. We omitted lncRNAs if they were adjacent to genes other than Hox genes for further analysis, as they could be involved in the regulation of a different gene in the Hox complex. For example, *Amalgam* sits between *bcd* and *Dfd* and codes for an immunoglobin that has an antisense lncRNA (CR45593), a lncRNA directly adjacent to the 3' end (CR44930), followed by the primary transcript for mir-993 that also contains another antisense transcript (CR43435). LncRNAs have previously been linked to the regulation of immunoglobulins (Yu et al., 2015) and miRNAs (Du et al., 2016) and although there was no evidence that these lncRNAs were associated with *Amalgam*, we chose to avoid them in favor of identifying lncRNAs more likely to be associated with Hox genes based on proximity.

Table 3.2.1. Summary table of lncRNA transcripts throughout the Hox Complex of *D. melanogaster*.

lncRNA transcript coordinates (dm3)	nt #	Current FlyBase annotations	Embryonic expression (hrs)	Description
pncr:3R:2,525,282-2,525,886	605	pncr:3R/ CR33938	4-14	Intergenic in cuticle proteins
CR45593:2,587,235-2,589,200	1966	CR45593	4-24	Antisense *Ama*
CR44930:2,591,207-2,592,431	1225	CR44930	4-12	Downstream *Ama*
CR42721:2,592,772-2,605,450	12679	CR42721	10-20	pri-mir-993
CR43435:2,603,839-2,604,674	836	CR43435	2-22	Antisense CR42721
CR45915:2,636,541-2,636,968	428	CR45915	-	Antisense CR42651
CR42651:2,634,518-2,642,101	7584	CR42651	4-24	pri-mir-10
CR45901:2,659,085-2,659,969	885	CR45901	-	Antisense *Scr*
CR45902:2,660,065-2,660,689	625	CR45902	-	Intronic sense strand *Scr*
CR45903:2,661,102- 2,661,412	311	CR45903	-	Intronic sense strand *Scr*
CR45904:2,661,591-2,662,080	490	CR45904	-	Antisense *Scr*
CR45905:2,662,388-2,663,330	943	CR45905	-	Antisense *Scr*
CR45900:2,684,561-2,684,914	353	CR45900	-	Intergenic *Scr-ftz*
lincX:2,703,400-2,711,000	7600	CR44931	4-6, 22-24	Intergenic *ftz-Antp*
TipX:2,718,647-2,719,191	544	CR45559	2-24	Downstream *Antp*
Hox-A:2,720,819-2,721,871	1052	CR44932	2-24	Downstream *Antp*
Hox-B:2,729,183-2,731,371	1,210	CR43252	4-10	Intronic sense strand *Antp*
Hox-F:2,826,191-2,827,144	420	CR45899	4-6	Divergent 5' *Antp*
Hox-G:2,863,724-2,865,357	1633	CR44933	6-14	Intergenic *Antp-Sodh1*
bxd:12,567,847-12,598,911	31064	bxd	4-10	Intergenic *Ubx-Glut3*
Tre1:12,591,129-12,592,078	950	(removed)	4-6	Intronic sense strand *bxd*
Tre2:12,589,406-12,590,514	1109	(removed)	4-14	Intronic sense strand *bxd*
Tre3:12,589,091-12,589,441	351	(removed)	4-6	Intronic sense strand *bxd*
Hox-L:12,608,818-12,610,372	1554	CR44945	6-14	Intergenic *bxd-Glut3*
CR45750:12,626,322-12,626,925	603	CR45750	-	Intergenic *Glut3-abd-A*
CR45751:12,627,021-12,627,599	578	CR45751	-	Intergenic *Glut3-abd-A*
iab-8:12,657,493-12,750,579	93086	iab-8	2-14	Intergenic *abd-A-Abd-B*
iab-4:12,675,726-12,681,913	6187	iab-4	4-16	pri-mir-iab4
Hox-O:12,718,215-12,723,595	5380	CR43167	4-6	Antisense *iab-8*
iab-7 PRE:12,725,342-12,725,811	470	iab-8	4-24	Antisense exon *iab-8*
Hox-P:12,739,371-12,740,294	924	CG10349 (removed)	4-6	Previously annotated mRNA
Hox-Q:12,778,101-12,779,459	1,359	CR46267	4-6	Antisense *Abd-B*

Mapped RNA-seq reads were investigated throughout embryogenesis for evidence of lncRNA transcription throughout the Hox complex. Some had already been annotated and putative novel lncRNA were assigned as Hox-A-R. FlyBase updated their annotations 2 years later and recognized the majority of those that we had detected, now assigned CR codes. The lncRNAs that have not been annotated by FlyBase were also removed from our later investigations as being unreliable. The '–' in the embryonic expression section indicates there was no detectable RNA-seq expression within 0-24hrs of embryogenesis.

In order to better annotate lncRNA genes structures we compared the RNA-seq findings with reads from 5' Cap Analysis of Gene Expression (CAGE), available up to 8 hours, to identify the transcription start sites (TSSs) used during the same 2hr periods after egg laying. CAGE reads for many lncRNAs also showed that the highest levels of transcription of the candidates was during the 4-6hr period, similar to the RNA-sequencing data (Fig.3.2.1-D and Fig.3.2.2-D). However, in the ANT-C, CAGE reads indicate that transcription during the 2-4hr window almost matches the 4-6hr window for the 4 lncRNAs shown. This suggests their transcription could begin earlier but

not be detected by the RNA-seq that utilizes poly(A) selection at the 3' end of transcription, possibly due to polymerase pausing.

There is evidence from CAGE that one of these lncRNAs, *Hox-G*, is transcribed to almost its full levels less than 2 hours after egg laying, suggesting considerably earlier activation of this gene (Fig.3.2.1-D). *Hox-G* consists of two exons and higher levels of transcription of the second, larger exon, something also seen for the transcription of *lincX* (Fig.3.2.1-B). It is 38,650bps from the nearest Hox gene, *Antp*, so it was not clear if it should be considered a part of the ANT-C. However, there are no other coding or noncoding genes between this lncRNA and the Antp promoter and we later establish using ntFISH that the lncRNA is expressed in the same cells as *Antp's* second promoter (Fig.3.3.1-B). This coupled with the corresponding timing of transcription with ANT-C Hox genes would suggest that it is indeed associated with the Hox complex.

TIPX and *lincX* have been previously characterized in our lab (Pettini, 2012) and were used for comparison to novel lncRNAs. *TIPX* is a single exon lncRNA that is highly transcribed in the 4-6hr window and maintains high levels of expression for much longer than the other lncRNAs identified (Fig.3.2.1.B-C). *Hox-F* is also a single exon transcript that is very similar in length to *TIPX* at 1 kb. *Hox-F* diverges from *Antp*, transcribed 1264nts away, from the opposite strand at much lower levels than the other lncRNAs found within the ANT-C (Fig.3.2.1.B-C). The 4 lncRNAs shown in the ANT-C (Fig.3.2.1) were chosen for further study over other lncRNAs from the original list (Table.3.2.1) as others had either: very low level transcription in any single 2 hour embryonic RNA-seq window, no/very low CAGE reads mapping near to the TSS, or because of their context relative to other genes; for example, a sense strand in a UTR region, intron, or adjacent to a non-Hox protein-coding gene.

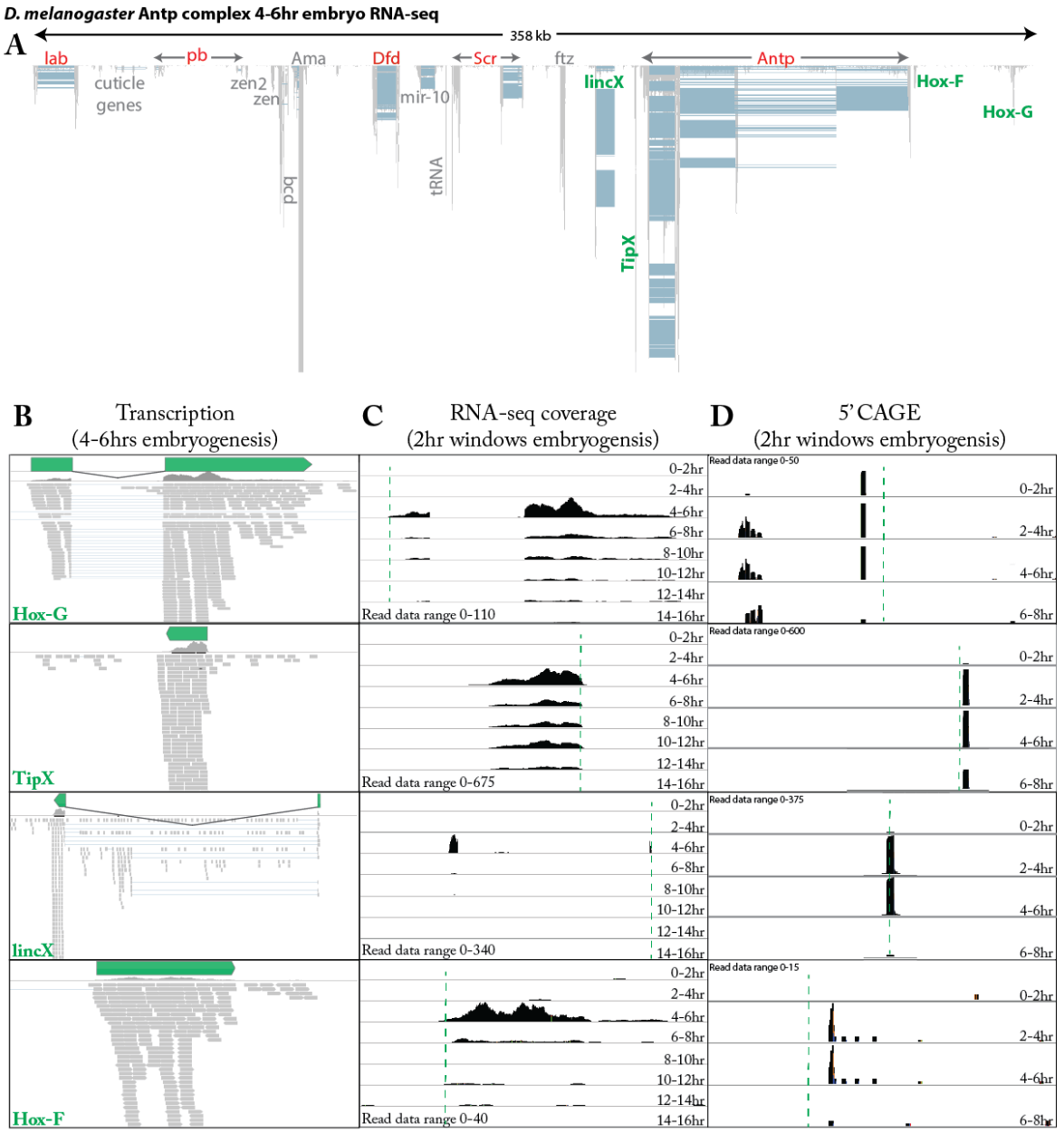


Figure 3.2.1. ANT-C *D. melanogaster* transcript identification during early embryogenesis using RNA-sequencing and CAGE. RNA-sequencing reads throughout the ANT-C were analyzed for possible lncRNAs, as indicated in green writing (A). These reads were used to determine transcript models using split reads that map to adjacent loci to find exons separated by introns (B) and 5' CAGE sequence tags to find the TSS of the 5' end (D). Coverage tracks show abundance of transcription across each locus and can be used to demonstrate relative abundance of transcription at each locus during the 2-hour time periods of embryogenesis that samples were taken from (C). The CAGE experiments were also carried out in 2-hour windows during early embryogenesis and can be used to measure relative abundance of transcription at each of these time points (D). The green dotted lines and arrows indicate where the RNA-seq reads begin (C).

The BX-C of *D. melanogaster* has been previously shown to contain two very large non-coding transcripts, *bxd* and *iab-8*. *Bxd* has been studied for over 50 years (Hannah-Alava, 1964) and *iab-8* for nearly 20 years (Zhou et al., 1999) and it has been shown that they have important roles in mediating the activity of the Hox genes *Ubx* and *abd-A* respectively. Based on the read coverage in tracks showing the levels of transcription across 0-16hrs of embryogenesis, the 4-6hr window contains the majority of lncRNA transcription (Fig.3.2.2-C), similar to the RNA-seq coverage from the ANT-C lncRNAs (Fig.3.2.1-C). However, the 5' CAGE sequencing indicates that lncRNA transcription in the BX-C transpires within the 4-6hr window, at least 2 hours after CAGE sequencing for the lncRNAs in the ANT-C (Fig.3.2.1-B).

Bxd consists of three small exons in the largest transcript, and 3 other shorter isoforms, all expressed at this time period, with high levels of transcription throughout the second intron and more restricted transcription in a small region in the first intron. The transcription within the first intron matches three TREs, once known simply as Tre1, 2 and 3 (Sanchez-Elsner et al., 2006). However, note that this journal article has since been retracted based on manipulation of gel images that they claimed showed protein recruitment to this site and therefore the name has been removed from FlyBase. We will continue to use the names Tre1, 2 and 3 for simplicity as not all databases at this time have updated to remove them and there is still clear transcription of this site. Furthermore, the ChIP-seq data indicates that both PcG and TrxG proteins bind to this exact site, implicating the loci as a PRE/TRE (Fig.3.4.1 and Fig.3.4.2)

Within an intron and antisense to *iab-8* is another previously characterized transcript, *iab-4,* that is processed to produce a microRNA, *mir-iab-4* (Ronshaugen et al., 2005). Another transcript was identified in our original screen that was designated *Hox-O,* found within another intron of *iab-8* and also transcribed from the opposite strand. This transcript consists of two exons and is similar to *Hox-G* and *lincX*, in that the first exon is smaller and transcribed at much lower levels than the second, larger exon (Fig.3.2.2-B). When viewing the stranded RNA-seq, we also noticed that the second exon of *iab-8* was transcribed in both directions. This region has previously been shown to have silencer functions and been annotated as a PRE, specifically iab-7 PRE (Hagstrom et al., 1997).

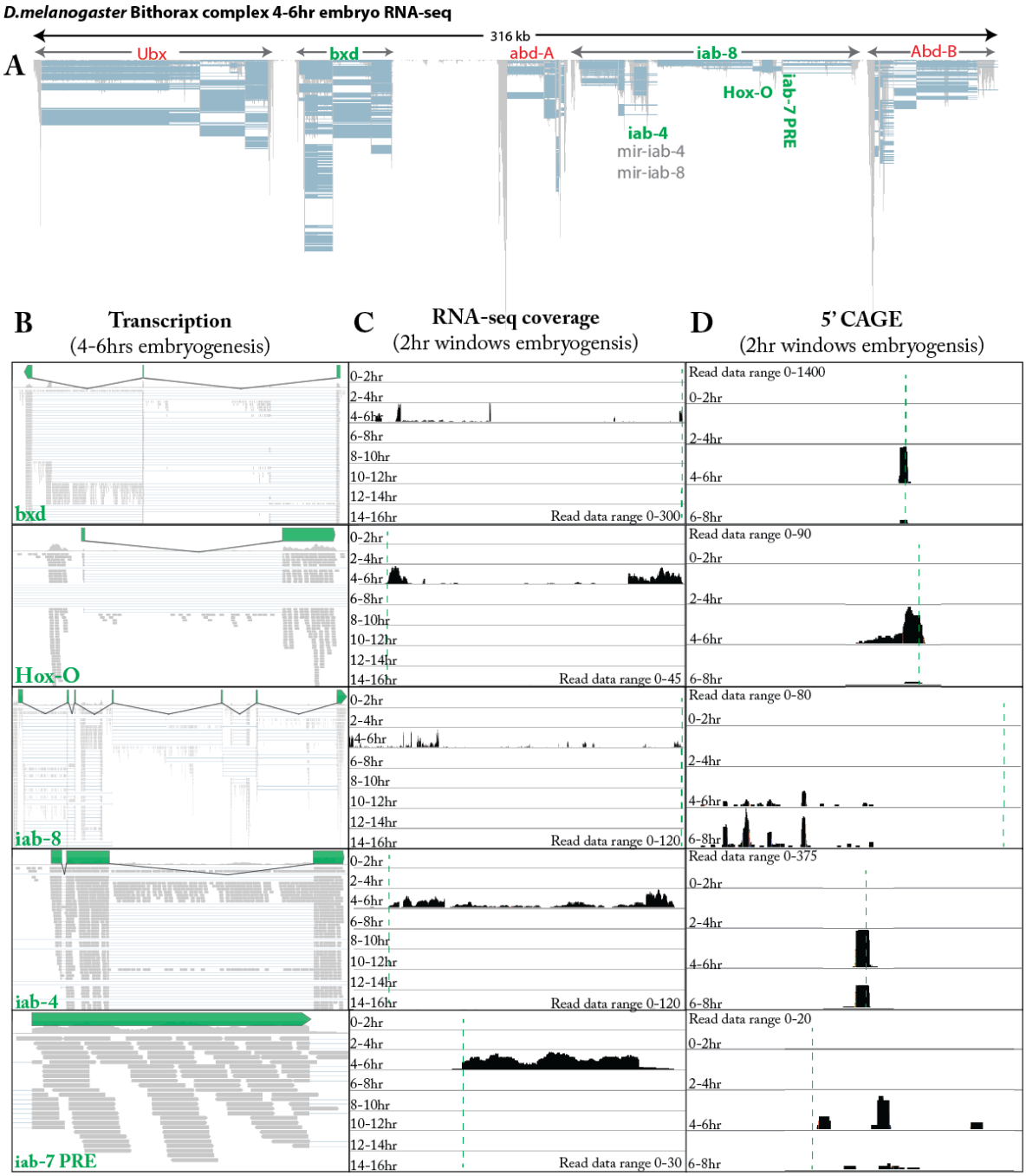


Figure 3.2.2. BX-C *D. melanogaster* transcript identification during early embryogenesis using RNA-sequencing and CAGE. RNA-sequencing reads throughout the BX-C were analyzed for possible lncRNAs, as indicated in green writing (A). These reads were used to determine transcript models using split reads that map to adjacent loci to find exons separated by introns (B) and 5' CAGE sequence tags to find the TSS of the 5' end (D). Coverage tracks show abundance of transcription across each locus and can be used to demonstrate relative abundance of transcription at each locus during the 2-hour time periods of embryogenesis that samples were taken from (C). The CAGE experiments were also carried out in 2-hour windows during early embryogenesis and can be used to measure relative abundance of transcription of the TSS at each of these time points (D). The green dotted lines and arrows indicate where the RNA-seq reads begin and direction of transcription (C).

In order to identify lncRNA transcripts that are most likely to be functionally conserved, we further investigated the Hox complex lncRNAs in *D. melanogaster* for evidence of syntenic conservation in *D. virilis*. We used the previously generated RNA-seq data as that matched the time period of most lncRNA expression from *D. melanogaster*. The mapped reads were again manually curated using IGV to identify regions of transcription to ensure accurate annotation of each transcriptional unit (Fig.3.2.3). There was evidence for eight lncRNAs from *D. melanogaster* that were syntenic with lncRNAs in the Hox complex of *D. virilis.*

The 2 lncRNAs, *lincX* and *TIPX*, downstream of *Antp* have previously been detected by ntFISH, after prediction by a BLAST based approach (Pettini, 2012). This study provided the first evidence of their transcription by RNA-seq, demonstrating conserved exon-intron structure for each (Fig.3.2.3-B). *Hox-F* was likely to have been previously overlooked due to its location, directly upstream of *Antp* and therefore considered part of the UTR. However, the stranded RNA-seq shows that *Hox-F* is clearly transcribed from the other DNA strand in both *D. melanogaster* and *D. virilis*, resembling divergent transcripts found to regulate essential developmental regulatory genes in mammalian studies through epigenetic manipulation of chromatin (Lepoivre et al., 2013; Luo et al., 2016). Further upstream of *Antp*, a transcript was identified that is syntenic to *Hox-G* and also found to have two exons (Fig.3.2.3-B).

Unlike *D. melanogaster* and most other sequenced drosophilids, the split between the ANT-C and BX-C occurs between *bxd* and *abd-A* in *D. virilis*, rather than between *Hox-G* and *Ubx.* Despite the break separating the 2 complexes at different loci, *bxd* remains adjacent to *Ubx* in *D. virilis*, further affirming an established role for this lncRNA in the regulation of Ubx (Petruk, 2006; Sanchez-Elsner et al., 2006). In the *abd-A-Abd-B* interval of the Hox complex in *D. virilis*, *iab-8* similarly spans most of the DNA between *abd-A* and *Abd-B*. The syntenic transcript of *iab-8* is 110 kb, similar to *D. melanogaster's iab-8* at 90 kb. It also consists of at least 8 exons mirroring *D. melanogaster's iab-8* transcript very closely (Figs.3.2.3-B, Fig.3.2.2-B). Within the syntenic *iab-8*, the lncRNA encoding the microRNA *iab-4* can be detected. However, the primary transcript does not show spliced exons as seen in *D. melanogaster* and instead looks like a large single exon transcript. Furthermore, a transcript that is seemingly syntenic to *Hox-O*, being antisense to *iab-8,* can also be detected, but it is spliced about halfway into the second exon of the syntenic *iab-8* lncRNA.

To further investigate evolutionary conservation of the lncRNAs identified in the Hox complex of *D. melanogaster* and *D. virilis*, RNA-seq was also carried out in *D. pseudoobscura*. *D. pseudoobscura* RNA was collected from embryos that had aged between 4-6hrs as they develop at approximately the same rate as *D. melanogaster*. The RNA was sequenced using the same method as previously described for *D. virilis* and reads mapped to the *D. pseudoobscura* genome (r3.03). Our sequencing depth was not as good as for the other two drosophilids and *Hox-F* could not be detected, but those that could were strikingly similar in exon-intron structure and synteny.

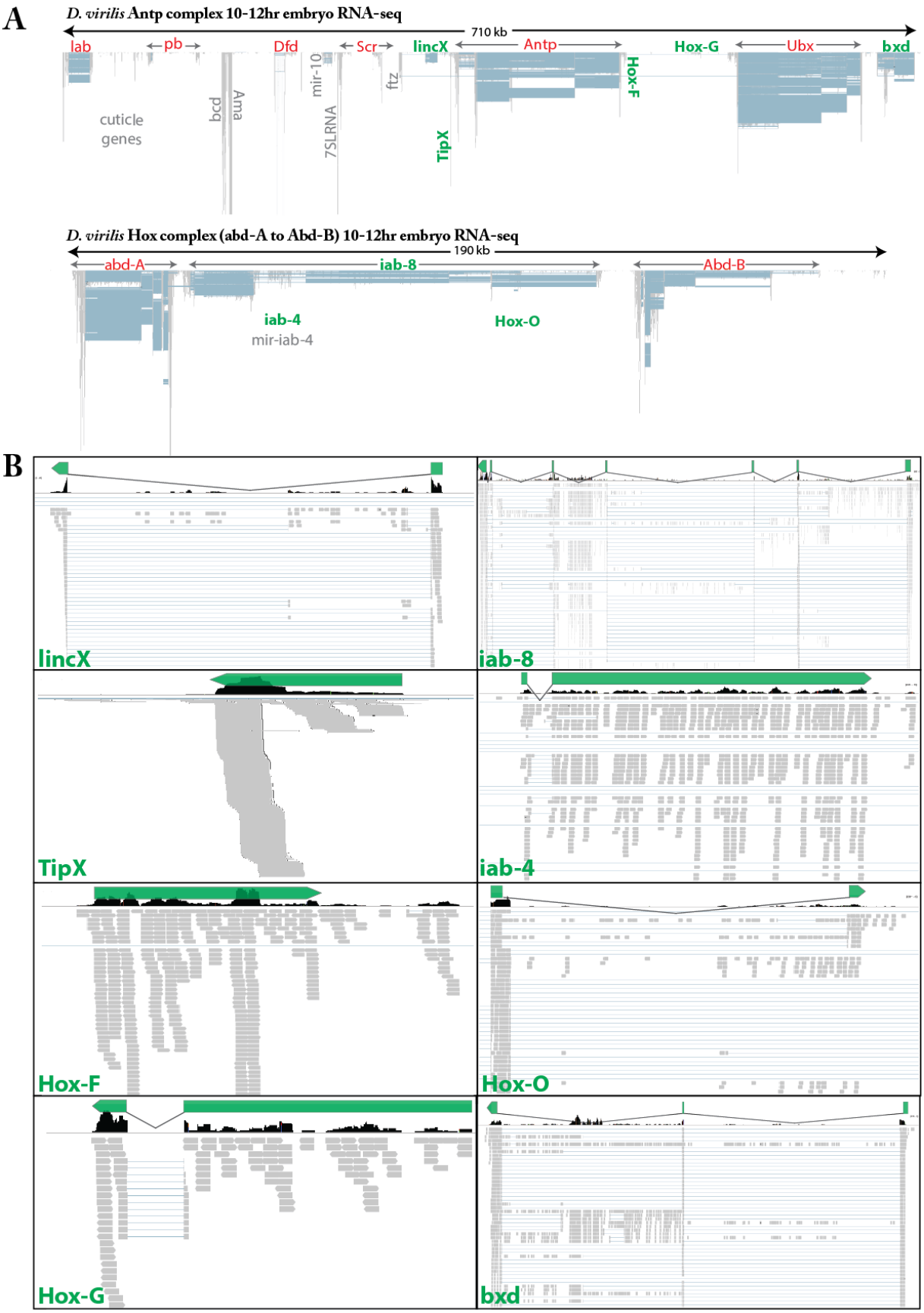


Figure 3.2.3. Mapped reads from Hox complex of *D. virilis* RNA-seq showing syntenic lncRNA transcripts from embryogenesis stages 4-6 (Campos-Ortega and Hartenstein, 1997). A) Overview of RNA-seq of Hox complex of *D. virilis*. Red writing denotes Hox protein-coding genes, green writing is the lncRNAs and grey is the others genes within the complex. B) Each of the lncRNAs are shown with the coverage track directly above in black peaks to show regions of highest transcription. Grey bars are single reads that can be split between genomic loci, indicated by blue lines between reads.

3.3 Expression patterns of Hox lncRNAs: Conservation and affects of segmentation gene mutations

The RNA-seq profile of *D. melanogaster* was used to depict the relative loci and gene structures of lncRNAs and Hox transcripts, along with the pair-rule gene, *ftz,* found between *Scr* and *Antp* (Fig.3.3.1-A). Previous investigations into lncRNA regulation in the Hox complex of *D. melanogaster* indicate that they are involved in the regulation of adjacent protein coding genes and this is reflected in their position of expression in developing embryos (Bender et al., 1983; Gummalla et al., 2012; Petruk, 2006; Pettini, 2012). We therefore wanted to investigate how the novel lncRNAs were expressed in developing embryos relative to adjacent genes and previously annotated lncRNAs. We engineered RNA probes that allowed us to visualize the expression of these lncRNAs and the neighboring genes using fluorescently labeled antibodies. The lncRNAs and nearby genes were then imaged with confocal microscopy, each giving overlapping but often distinct regions on the developing embryo. The lncRNAs can often be seen expressed in many of the same cells as the adjacent genes in a variety of both broad and restricted patterns (Fig.3.3.1-B).

The lncRNAs *lincX* and *TIPX* have been shown to be involved primarily in the regulation of the Hox gene *Scr*. They are found adjacent and upstream of *Scr* on the chromosome (Pettini, 2012). The expression of the entire Hox complex, the lncRNAs and Hox genes, can generally be seen progressing from the anterior to the posterior of the embryo in the same spatiotemporal manner observed previously for Hox genes. This begins at *lab* in ANT-C and moves through to *Abd-B* in the BX-C. This collinear expression is similar to the original findings of Ed Lewis for the collinear arrangement on the chromosome matching the order they are expressed on the developing embryo (Lewis, 1978) (Fig.3.3.1-B). This is exemplified when considering *iab-8, iab-4, Hox-O, TIPX* and *lincX;* however, the expression pattern of *Hox-G* seems to skip the domain of *AntpP1's* expression to match that of *AntpP2.*

LncRNAs have been shown to commonly regulate local gene expression in *cis* in the Hox complex and this has been demonstrated in *bxd's* regulation of the adjacent *Ubx* and *iab-8s* regulation of *abd-A* (Quinn and Chang, 2016). It is therefore not surprising when lncRNAs are expressed in similar regions as the genes they are regulating, such as *Bxd's* RNA expression on stage 5 embryos (Campos-Ortega and Hartenstein, 1997), which overlaps most of the domain of *Ubx* expression, with *Ubx* also appearing in a wide layer of cells anterior to *Bxd's* expression.

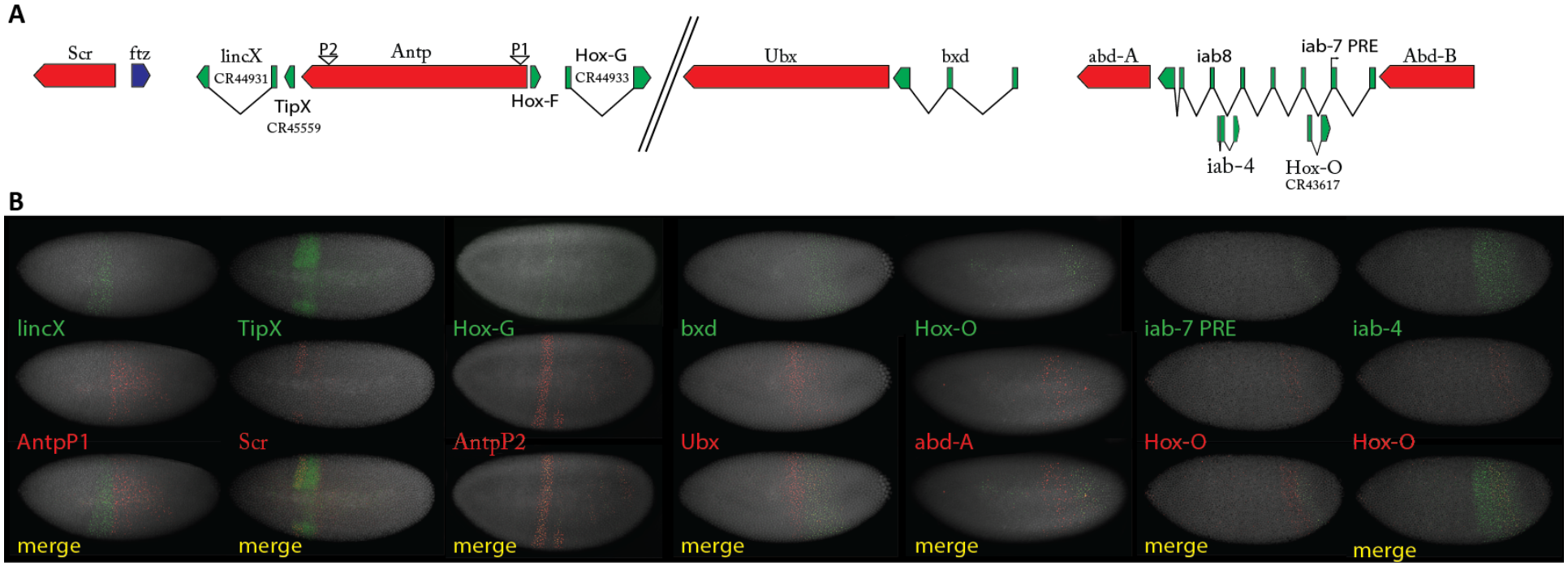


Figure 3.3.1. Nascent Transcript FISH Expression Patterns of lncRNA and Adjacent Genes in WT Stage 5 *D. melanogaster* embryos

Nascent transcript in situ hybridization was on wild type (W1118) embryos and shows lateral bands of transcript expression domains. A) Relative location of lncRNA genes to neighboring Hox genes and *ftz*, including the exon and intron structures of the lncRNAs. The location of the probes to detect the first promoter (P1) and the second promoter (P2) of *Antp* is shown by the empty triangles as these each have their own expression domains on the developing embryo, as demonstrated in panel B. *LincX*, *TipX*, *bxd* and *iab-8* have been previously characterized and were used to compare differences in domains of expression in an embryo to any novel lncRNA expression. (B) The lncRNA, *Hox-G*, matches the expression of the second promoter of Antp (P2), but with fainter expression. *Bxd* regulates *Ubx* in a number of studies and is found expressed more to the posterior of the embryo. *Hox-O* is a transcript found antisense to a single intron of *iab-8* and has a unique expression pattern in a lateral stripe posterior to *abd-A* in the embryo. RNA probes detected that the second exon of *iab-8* is transcribed in both directions (*iab-7 PRE*), expressing in the same domain in either orientation, and also matching the pattern of *iab-8*. *Abd-B* not shown as it does not express until later embryonic stages.

The very long, multiexonic lncRNA, *iab-8* is expressed in a pattern matching *ftz* stripe 7 and extends anteriorly about halfway into the space between stripes 6 and 7, matching segment A8-PS14 (Fig.3.3.2-B). *iab-8* has two antisense transcripts, *Hox-O* and *iab-4* and produces a miRNA, *mir-iab-8* (Tyler et al., 2008). The *iab-4* transcript has been investigated and produces a microRNA, *mir-iab-4,* that has been shown to be involved in the regulation of *Ubx*, through its ability to transform the halteres into wings (Ronshaugen et al., 2005). *Hox-O* lncRNA transcript has yet to be investigated. Its transcription can be seen within *iab-4's* expression domain, but with a more restricted pattern that begins further to the posterior of an early stage 5 embryo. *Hox-O's* expression then moves towards the anterior of the embryo, once higher expression can be detected in later stage 5 embryos (Campos-Ortega and Hartenstein, 1997) (Fig.3.3.1-B).

A regulatory region has been identified as *iab-7* and shown to be an important component for appropriate activation of *Abd-B* (Celniker et al., 1990; Sanchez-Herrero, 1991). The *iab-7* PRE is important for polycomb dependent maintenance of silencing and this silencing has been shown to be disrupted by transcribing through it. The effect of this disruption is thought to be orientation dependent due to the position of the promoter (Hogga and Karch, 2002), indicating this PRE could function as a bidirectional switch. This is something that has also been seen for other PREs (Herzog et al., 2014). *iab-8* is expressed in the same pattern as the *iab-7* PRE, a single exon that is transcribed in both directions from the second exon of *iab-8*. There is very little evidence suggesting that antisense transcription of *iab-7* PRE at its endogenous loci has been recognized. There is no antisense transcript reported in FlyBase for the second exon of *iab-8*, but two other groups have some evidence. One group annotated coding sequence (CDS) from their whole transcriptome microarray for a transcript matching this exon in the antisense orientation referred to as BK002593.1 in their data, that overlapped transcription about halfway into the exon (Hild et al., 2003). However, they have used low stringency settings in their gene predictions with the Fgenesh software (Salamov and Solovyev, 2000) and the majority of their transcripts are not validated by other methods that do use high stringency analysis. The antisense transcript at the *iab-7* PRE locus is expressed at a very low level and the exon is just 462nts in length so it is likely to be discarded based on expression. Another group has annotated peaks of promoters across the genome using paired end 5'-complete cDNA sequencing with an analysis termed RAMPAGE and identified a TSS at the 3' end of the second *iab-8* exon (TSS_RAMPAGE_019667) that indicated there is transcription in the antisense orientation (Batut et al., 2013). Although this other evidence is not conclusive, our strand-specific RNA-seq, along with strand-specific ntFISH has corroborated that this exon is transcribed in both directions.

Hox-G's expression pattern is unusual in that it perfectly matches the expression of the second promoter of *Antp,* (*AntpP2*), something not seen for any of the other lncRNAs that all have their own unique expression in the embryo (Fig.3.3.1-B). This second promoter produces a distinct isoform of *Antp* and is ~105 kb from *Hox-G's* transcript on the chromosome. *Hox-G* produces

much fainter band than *AntpP2*, likely due to the very low expression levels (Fig.3.3.1-B). However, it can be seen in each of the three regions matching *AntpP2*, in half a lateral stripe near the center of the embryo, a full stripe anteriorly just below the half stripe and then also a very faded stripe towards the posterior of the embryo.

For the syntenically conserved lncRNAs, RNA probes were constructed to carry out ntFISH in *D. virilis* to test if the lncRNAs were expressed in similar patterns as *D. melanogaster* embryos. Figure 3.3.2-B shows those that were detected in both species, in stage 5 embryos (Campos-Ortega and Hartenstein, 1997), with *ftz* used in both organisms to denote segments. For clarification, the *ftz* stripes are designated stripes 1-7 from the anterior (left) to posterior (right) of the embryo. The *lincX* lncRNA transcript expression aligns between *ftz* stripes 1 and 2, matching the anterior of PS2 to the first half of the anterior of PS4. The *lincX* transcript therefore corresponds to the anterior of the T1 boundary (PS2-T1) in both drosophilids in panel B of figure 3.3.2. *Hox-G* expression is much less visible, but can be seen matching the second *ftz* stripe in both Drosophilids, indicting PS2 and then a small part of stripe 3 (PS6) (Fig.3.3.2-B). The posterior *Hox-G* and *AntpP2* stripe, in *D. melanogaster,* matches the position of *ftz* stripe 7 (PS14), but is not visible in the *D. virilis* embryo.

Bxd overlaps *ftz* stripes 4, 5 and 6 in both species, encompassing PS 8-12, with a few cells and extends anteriorly from *ftz* stripe 4 in both *D. melanogaster* and *D. virilis* embryos, suggesting it could extend into segment A2, with a uniform distribution (Fig.3.3.2-B). The expression pattern of *Hox-O* is slightly different in each species. In *D. melanogaster, Hox-O* covers *ftz* stripe 6 and extends outwards, both posteriorly and anteriorly, about halfway towards both stripe 5 and 7, suggesting its expression pattern matches segments A4-A5. However, in *D. virilis*, the syntenic *Hox-O* transcript appears to start over *ftz* stripe 6 and extends posteriorly to cover stripe 7, aligning to PS 12-14. This is particularly interesting as the *Hox-O* transcript itself is positioned in the second intron of *iab-8* in *D. melanogaster*, but the syntenic transcript in *D. virilis* extends into the first intron of *iab-8*, and therefore mirroring its physical location further along the Hox complex on the chromosome.

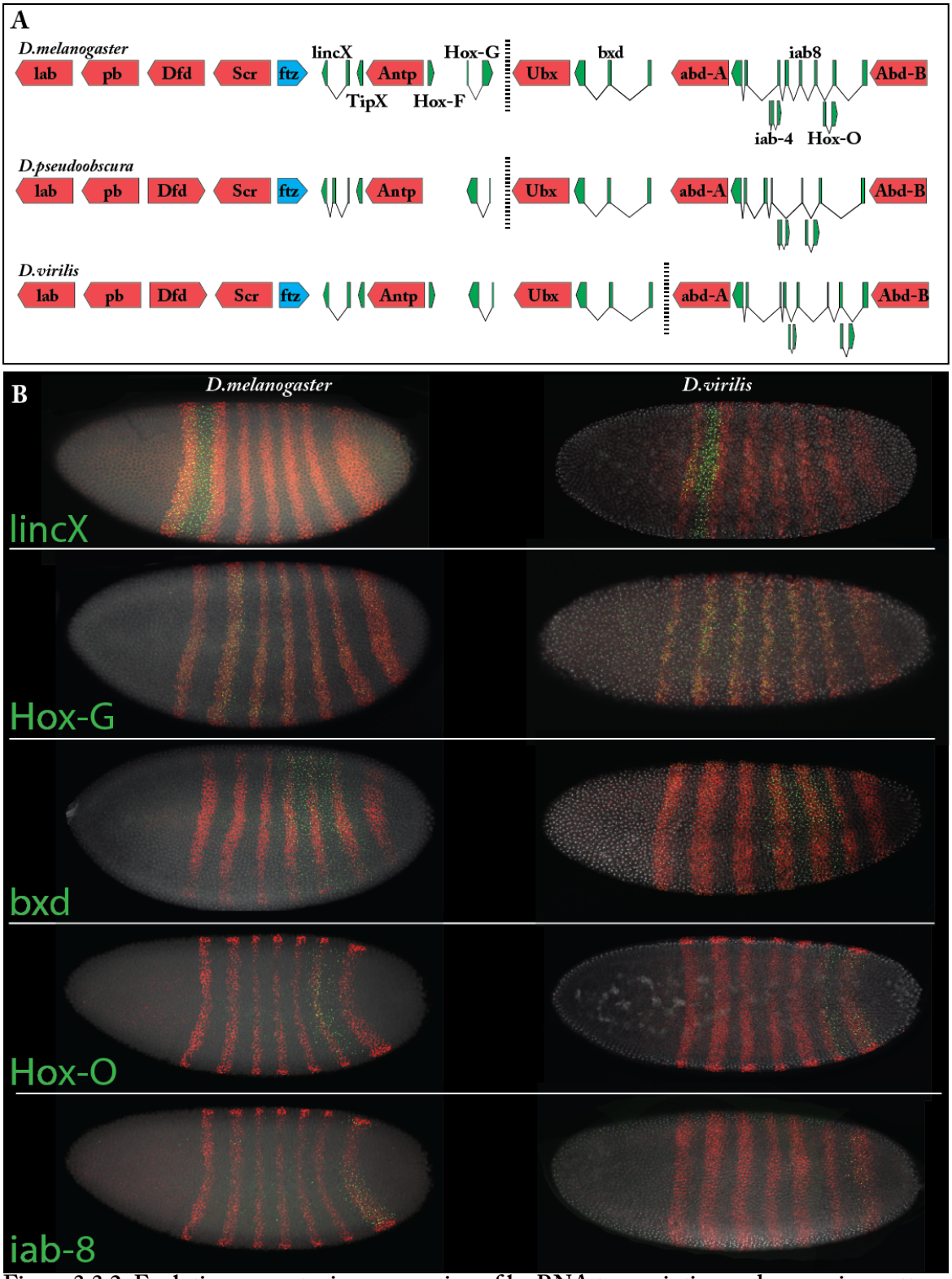


Figure 3.3.2. Evolutionary syntenic conservation of lncRNA transcription and expression patterns in early developing Drosophilid embryos. Analysis of RNA-seq in developing embryos of *D. pseudoobscura* and *D. virilis* identified syntenic transcript conservation throughout 60 million years of evolutionary divergence. Syntenically conserved transcripts can be seen in similar positions (green) relative to Hox genes (red) in both other species from the lncRNAs identified in *D. melanogaster*, further supported by analogous exon and intron gene structures (A). The dotted black line for each organism and D. pseudoobscura shows the break points of the ANT-C and BX-C and *D. virilis* have been reversed for comparison to *D. melanogaster*. The lncRNAs can been seen expressed in the same domains of Stage 5 embryos (Campos-Ortega and Hartenstein, 1997) for both *D. melanogaster* and *D. virilis* (B) when using *ftz* (red stripes) to mark segments. Embryos are oriented anterior to the left and posterior to the right. Note the image of the late stage 5 embryo is rotated to show the dorsal view face down.

When *D. melanogaster* and *D. virilis* were investigated for conserved expression of *Hox-O*, we noticed that the expression in *D. virilis* was much more prominent, giving 2 clear lateral stripes near the posterior of the embryo. We decided to take advantage of this to compare its expression pattern to the adjacent lncRNAs expressed in the same region and investigate the timing that expression could be detected throughout embryogenesis. We used the RNA probes already constructed at the 5' end of each gene, *iab-8, Hox-O* and *pri-mir-iab4* (Fig.3.3.3-A) to explore the relative spatiotemporal patterns.

Figure 3.3.3-B shows that *pri-mir-iab-4* is initially expressed in stage 4 embryos (Campos-Ortega and Hartenstein, 1997) in a wide band near the posterior of the embryo before *Hox-O* or *iab-8* can be seen. By early stage 5, *Hox-O* can be detected in a narrower band in a very similar position as the posterior of *pri-mir-iab-4* and *iab-8* is still not visible until late stage 5 when it can be seen appearing in a narrow band at the posterior of the embryo. At late stage 5 *Hox-O* separates into 2 distinct lateral stripes, expressed either side of *iab-8* suggesting a regulatory role for either *Hox-O* in *iab-8's* expression or vice versa. By early stage 6 and into late stage 6 of embryogenesis, *pri-mir-iab4* is barely visible and *Hox-O* is fading slightly, with the more posterior stripe almost gone, but *iab-8* is still clear. At stage 7, there is still a faint signal for both *pri-mir-iab4* and *Hox-O* that is almost completely gone in stage 8 embryos, whereas the expression of *iab-8* is maintained until stage 15, when it finally fades and can no longer be detected past this stage.

These results indicate specific spatiotemporal regulation of each of these lncRNAs and agree with the theory of collinearity of Hox genes, as they are initiated from *iab-8*, along the chromosome to *lincX* (shown left to right Fig.3.3.1-A). The tightly restricted patterns of lncRNA expression seen on the developing embryo align with other studies that have demonstrated lncRNAs have highly specific expression profiles in comparison to mRNAs {Quinn, 2016 #1074}. Interestingly, the majority of the lncRNA expression patterns are unique and do not match adjacent Hox genes, but overlap some slightly, possibly suggesting they are independently regulated. Several of the lncRNAs that we can identify as syntenically conserved in *D. virilis* are transcribed in similar regions and in similar stages of development in both species, suggesting that they are conserved orthologs (Fig.3.3.2). This is further exemplified by the similarities in intron-exon structural arrangements and therefore this level of conservation is indicative of function {Diederichs, 2014 #853}. The expression patterns of *iab-8, Hox-O* and *pri-mir-iab-4* shows just how dramatically these transcripts differ in where and when they are expressed, even though they are overlapping or adjacent to each other (Fig.3.3.3). This indicates that they are being individually regulated and therefore may have specific roles in the respective cells they are expressed in.

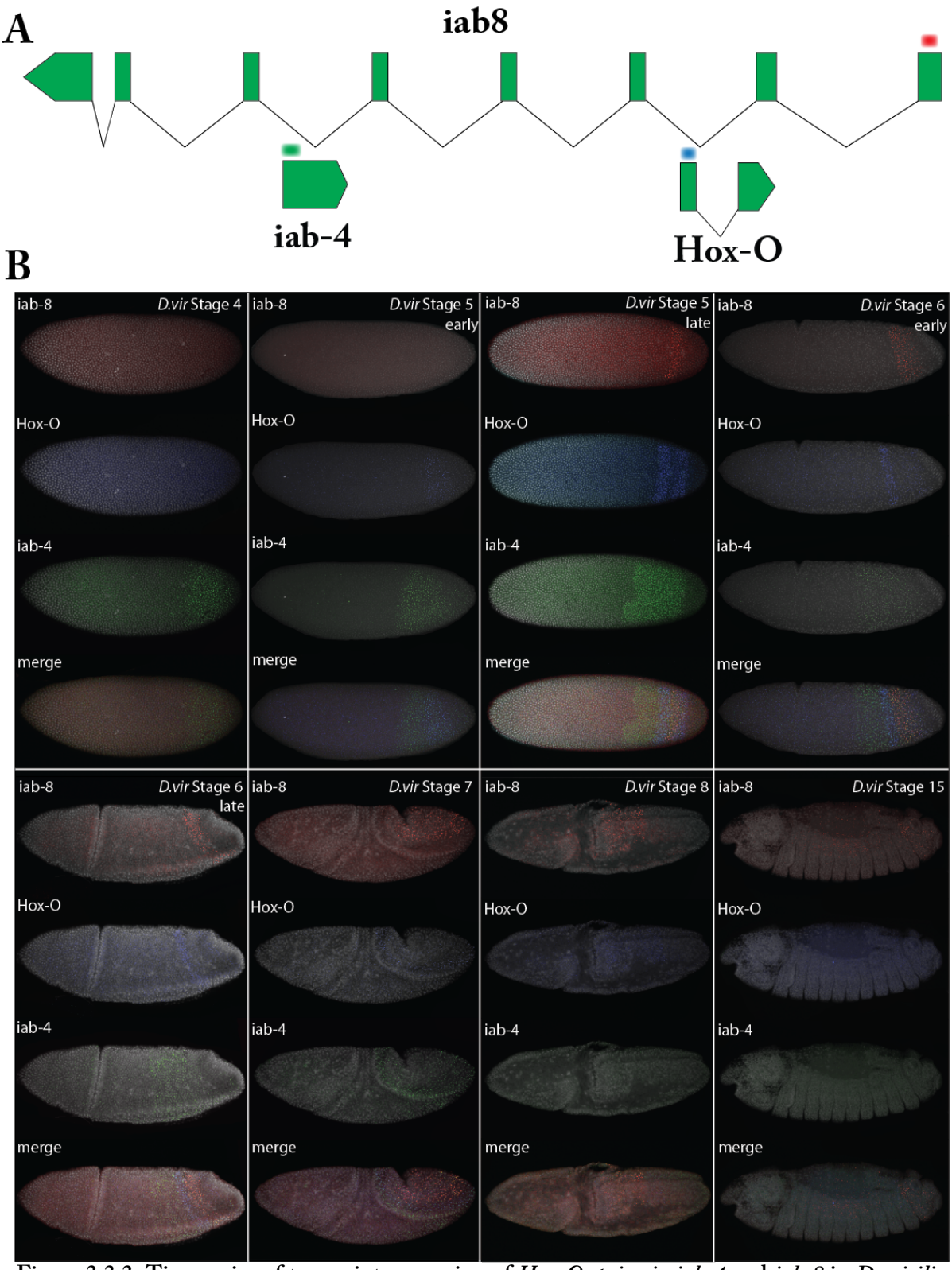


Figure 3.3.3. Time series of transcript expression of *Hox-O*, *pri-mir-iab-4* and *iab-8* in *D. virilis*. Developing embryos were fluorescently stained for antisense transcripts to *iab-8*, *Hox-O* and *pri-mir-iab4* and show independent patterns of spatiotemporal expression of the different lncRNAs. Red = *iab-8*, blue = *Hox-O*, green = *pri-mir-iab4* shown as individual images and merged in each panel with DAPI staining shown in grey.

Based on the observation that lncRNAs in the Hox complex are mostly expressed in a collinear pattern relative to Hox genes, we investigated if the same segmentation genes as those that regulate Hox genes could also regulate the lncRNAs. To investigate this we used *D. melanogaster* flies that had mutations in the segmentation genes, *Kr, h* and *eve* that were lethal when homozygous but would account for ~25% of eggs laid in a cage, allowing us to identify and analyze expression patterns in them. This allowed us to identify embryonic lethal mutations using ntFISH in early developing embryos. The segmentation gene *ftz*, is expressed in 7 evenly distributed stripes in wild type embryos, but will fail to produce these stripes in homozygous mutant embryos of segmentation genes (Fig.3.3.4). This allowed us to screen for the homozygous mutants and find out if lncRNAs expression was altered and therefore also regulated by the same segmentation genes responsible for Hox gene regulation.

Homozygous *Kr-* mutant embryos (1 in 4) could be identified by the expression of just four stripes of *ftz*. This is due to a loss of central segments of the embryo, causing a general deletion of 3 of the middle *ftz* stripes. Kr^{1} has been reported as missing T3, A1 and A2 (Bullock et al., 2004) and Kr^{17} as missing T1 through to A4 (Preiss et al., 1985). These mutants also displayed altered expression of lncRNAs, *bxd, lincX* and *TIPX* as they can be seen in much more expanded regions than their wild type expression (Fig.3.3.4). *bxd* expands towards the anterior of the embryo and *lincX* and *TIPX* are expressed more posterior in the embryo, indicating that the mechanism of negative regulation has failed, but the TF that instigates they expression is still active in these regions.

Homozygous *eve-* embryos can be identified by missing *ftz* expression in stripe 1 (Fig.3.3.4). In eve^{1} mutants, *lincX* appears to be half the width, missing the most anterior half of its expression. However, in the eve^{3} mutant *lincX* appears to be the normal width, but missing a patch in the same region that *ftz* stripe is no longer expressed (Fig.3.3.4). The missing region of expression for *lincX* and *ftz* seems to align well with *snail*, a D-V gene expressed at this time (data not shown), suggesting D-V genes could also play roles in the regulation of both of these genes. The eve^{1} allele is temperature sensitive, so we also tested embryos from flies raised at 29°C. The homozygous eve^{1} mutants can be identified by having just 6 *ftz* stripes, as stripe 1 is completely missing, and *lincX* and *TIPX* expression is seen as a narrow, faint band just posterior to *ftz* stripe 2, suggesting they have either been negatively regulated and silenced or failed to activate in these cells.

Homozygous eve^{1} and h^{25} demonstrated clear alterations in the expression of *Hox-G*. The homozygous h^{25} embryos could be identified by the *ftz* stripes merging and covering most of the embryo (Fig.3.3.4). *Hox-G* was stained along with *AntpP2* to determine if *AntpP2's* expression altered in the same way as *Hox-G,* as they are usually expressed in the same cells in wild type embryos and we thought they might have been regulated by the same mechanisms. Interestingly, although both *Hox-G* and *AntpP2* expression changed, it was in very different respects. The

anterior stripe of *Hox-G* could not be detected in the *eve*^{1} mutants and became very faint in the *h*^{25} mutants, but in both the posterior stripe became much more prominent than could be detected in wild type embryos, suggesting it's expression levels had increased significantly. However, the expression of *AntpP2* is barely altered, with only the gap between the 2 anterior stripes disappearing and filling with its expression and the posterior stripe remaining faint and seemingly in the same cells. This would imply that at least for the lncRNAs tested, the segmentation genes are involved in their regulation.

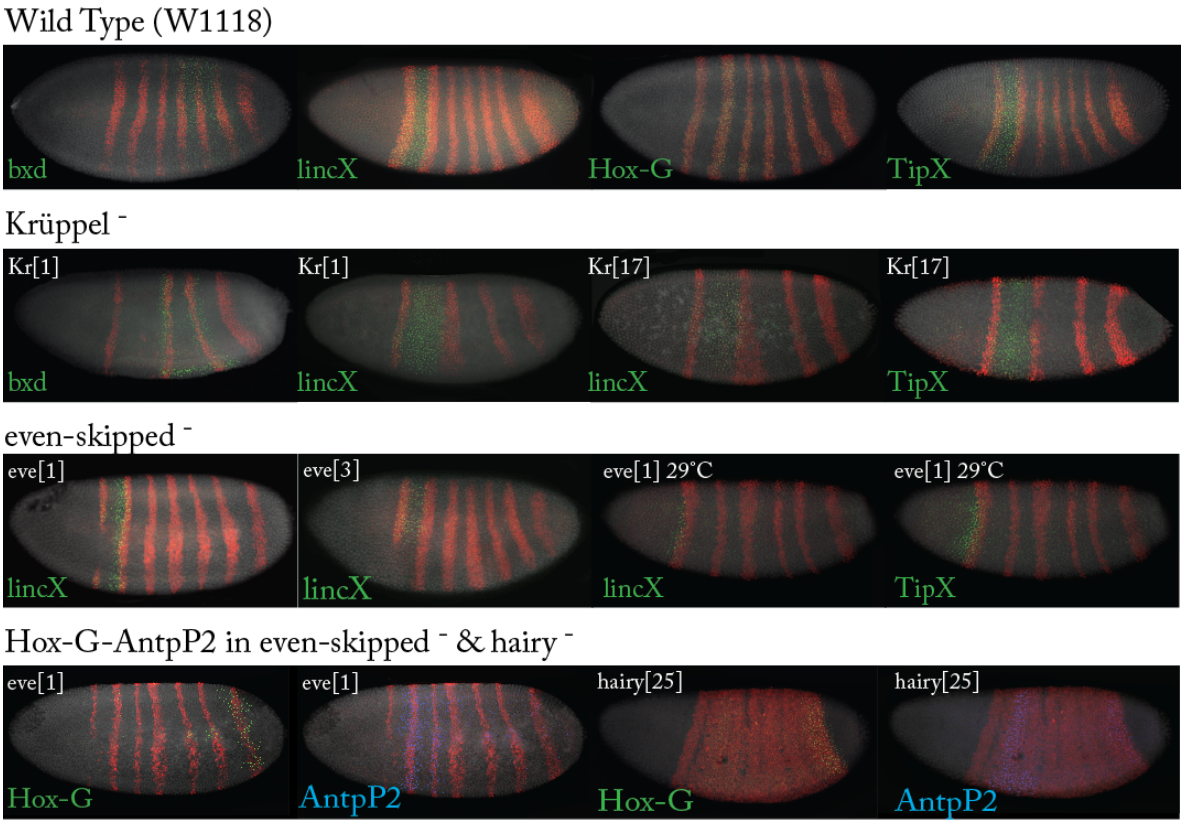


Figure 3.3.4. Early embryonic altered expression of lncRNAs in homozygous segmentation gene mutants. Nascent transcript FISH (ntFISH) was carried out on embryos of various segmentation gene mutants to test if they also played a role in the regulation of lncRNAs in the Hox complex. Wild type (W1118) expression of lncRNAs is shown in the top panel in green alongside *ftz* in red to demarcate the segments and as a guide to measure any lncRNA changes. *ftz* was also used to identify the homozygous mutant embryos that could not form appropriate segments. *Bxd, lincX* and *TipX* have altered expression in *Krüppel* mutants. *Hox-G, lincX* and *TipX* have altered expression in *eve* mutants, *eve[1]* is a temperature sensitive allele and gave more dramatic results when the embryos were laid at 29°C. *Hox-G* had a similar change in its expression in *hairy* mutants.

3.4 Regulatory protein binding at lncRNAs

PcG and TrxG proteins make epigenetic modifications to histones that lead to restructuring of chromatin and altering gene expression. Many PcG and TrxG proteins were identified as when mutated they fail to carry out this role to maintain on or off states of Hox genes in later development and adult flies (Cavalli, 2002; Lewis, 1978). Furthermore, it is now well established that lncRNAs have roles in directing the action of chromatin modifying complexes (Bohmdorfer and Wierzbicki, 2015). We therefore investigated available datasets for evidence of PcG or TrxG proteins binding the lncRNA loci we had identified in the Hox complex. Whole genome investigations into various subunits of the PcG and TrxG complexes, have utilized ChIP-ChIP and ChIP-seq to identify loci throughout the *D. melanogaster* genome that these proteins bind to that could indicate PRE/TREs. These experiments were carried out in a variety of developmental stages, tissues and cell types, including whole embryos, imaginal discs, pupae and the embryonic cell lines, Kc167 (originates from 8-12hr embryos), or Schneider 2 (S2) cells (harvested from 20-24hr embryos). The ChIP data shown in Figure 3.4.1 and 3.4.2 were obtained from the GEO repository (see methods for details).

Figure 3.4.1 and 3.4.2 show PcG and TrxG protein ChIP-seq profiles at lncRNAs in the Hox complex that give well-defined peaks. Above the single exon transcripts, *TIPX, Hox-F* and *Tre2* distinct PcG peaks can be seen across the entirety of their transcribed region. These transcribed regions are significantly enriched as the region can be seen with visibly increased numbers of reads compared to the surrounding, background noise, seen as a large black pyramid structure that stands out above the transcripts. There is very little binding of either PcG or TrxG proteins to *lincX*, except minor peaks of Pc, Pcl and Ph binding within the intron (Fig.3.4.1). The lncRNA *Hox-G* has distinct binding peaks for Pc, Psc, Su(z)12 and Ph and a smaller peak for Pcl, just upstream of the TSS, indicating this region could be a PRE and regulate *Hox-G* or another gene (Fig.3.4.1). The *bxd* transcript does not appear to have any other binding of these PcG proteins, outside of the *Tre2* transcript. *iab-8* has many binding sites, overlapping both exons and introns, including the *iab-7* PRE. However, the antisense transcripts, *Hox-O* and *iab-4,* do not appear to have any PcG peaks, even when amplified and auto scaled to facilitate detection of low level binding (Fig.3.4.1).

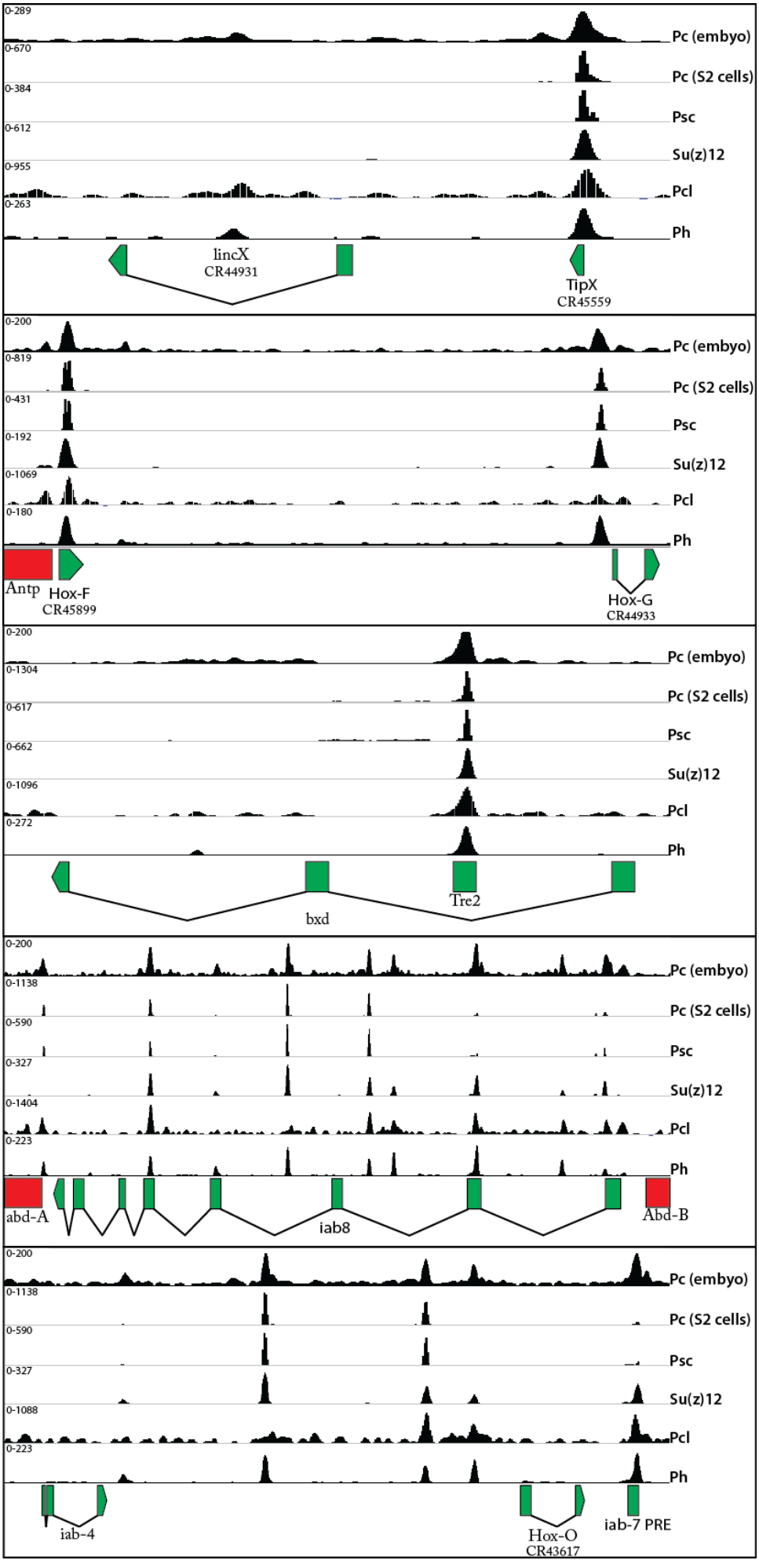


Fig 3.4.1– Legend on next page

Figure 3.4.1. ChIP-seq profiles of PcG proteins binding at lncRNAs. PcG protein binding peaks from either S2 cells or whole embryos. Embryonic samples are collected from either 0-8hrs or 5-13hrs post egg laying, depending on experiment. ChIP-seq tracks were visualized using IGV and auto scaled showing peaks of reads that stack up at particular sites. PcG binding peaks are included for Polycomb (Pc), Posterior sex combs (Psc), Su(z)12, Polycomblike (Pcl) and polyhomeotic (ph). The position and structure of the lncRNAs are indicated in green and the protein coding Hox genes in red. The transcription unit Tre2 is indicated as a green block within the intron of *bxd*.

Figure 3.4.2 shows that the binding profiles of many TrxG proteins overlap the binding profiles of many PcG proteins. The single exon lncRNAs are bound by numerous components of these complexes, along with the region just upstream of *Hox-G*. Similar to PcG proteins, *lincX* seems to have no clear peaks of TrxG protein binding, suggesting it may not be directly regulated by, recruit, or participate in chromatin modifications via PcG and TrxG proteins. *TIPX* has distinct binding peaks for many members of the TrxG proteins shown, except brm and Utx. *Hox-G* and *Hox-F* also have well-defined peaks indicating TrxG protein binding, with the exception of mod(mdg4) and indistinct peaks for Lid and Utx. In the BX-C, *Tre2* is bound by most members of the TrxG proteins shown, except for mod(mdg4) (Fig 3.4.2). A second region of the *bxd* transcript, within the second exon, is also clearly bound by mod(mdg4), fs(1)h and trr. The whole of *bxd's* second intron is quite heavily transcribed (Fig.3.2.3-B) making it hard to determine if it contains any specific transcripts that overlap the protein bound region and could be classed as lncRNAs. There also the possibility that it could be processed into a stable intronic sequence RNA (Pek et al., 2015) in order to function. Alternatively, transcription may not be necessary for this DNA region to carry out any functions it may have and the heavy transcription could be a result of poor pre-mRNA splicing efficiency (Guilgur et al., 2014).

The *iab-8* transcript has several regions that bind different combinations of TrxG proteins. mod(mdg4) binds all regions that are also bound by at least one other TrxG protein (Fig.3.4.2). The three sites of *iab-8* bound by Trl, the first intron, the second exon (*iab-7 PRE*) and fourth intron, appear to lack binding of the other TrxG proteins shown, except for mod(mdg4). However, all other regions of *iab-8* that indicate regulation by TrxG proteins are clearly lacking peaks of Trl. The antisense transcripts within *iab-8, iab-4* and *Hox-O* appear not to overlap any of the TrxG peaks and instead are flanked by previously identified regulatory regions (Gummalla et al., 2012).

The ChIP profiles show that some of the single exon lncRNA loci are clearly bound by several members of the PcG and TrxG proteins and therefore are likely to be classed as transcribed PRE/TREs (Fig.3.4.1 & 3.4.2). The majority of multi-exon lncRNAs do not show indications of acting as PRE/TREs as they do not have distinct binding of the PcG and TrxG proteins at their loci, with the exception of *iab-8*, but this transcript is ~90 kb in length and the binding sites appear random so are likely to be coincidental with the region of transcription (Fig.3.4.1 & 3.4.2). The other interesting peak of PcG and TrxG binding is just upstream of *Hox-G* as it seems to align with the region the promoter would be expected to be found (Fig.3.4.1 & 3.4.2) and so it could be interesting to find out if the PRE/TRE is linked to the possible functions of *Hox-G*. Interestingly, there are a few examples in this data where there is a PRE/TRE just upstream of a two exon lncRNA, as can be seen for *lincX*, whereby the transcribed *TipX* is just upstream and *Hox-O* also has a peak just upstream (Fig.3.4.1 & 3.4.2). However, given the frequency of the distribution of PRE/TREs throughout the Hox complex, this could be coincidental and this theory would require testing throughout the genome.

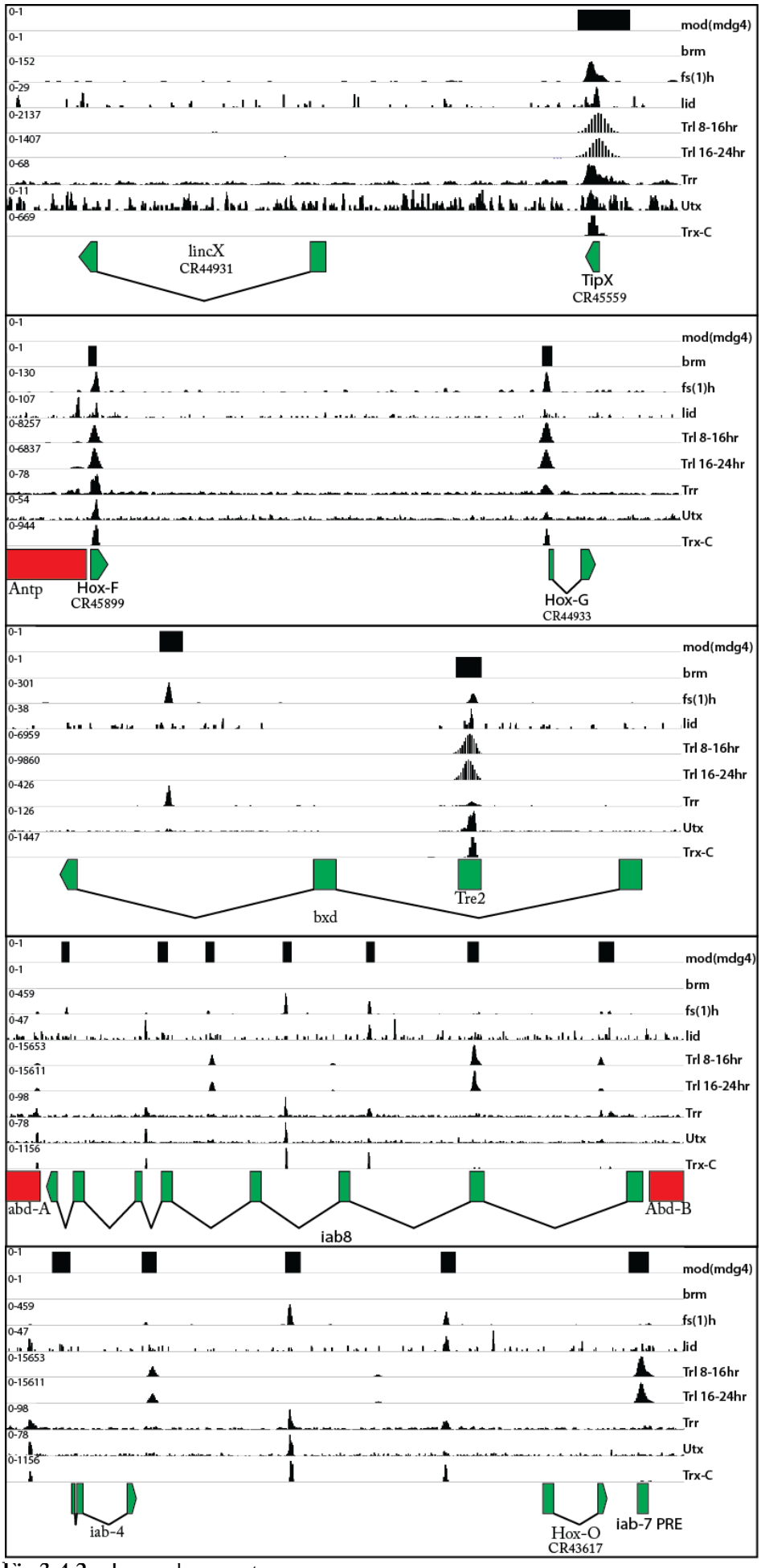


Fig 3.4.2 – Legend on next page

Figure 3.4.2. ChIP-seq and ChIP-ChIP profiles of TrxG proteins binding at lncRNAs. TrxG protein binding peaks from S2 cells and whole embryos. Embryonic samples are collected from 0-12hrs, 8-16hrs or 16-24hrs post egg laying, depending on experiment. Tracks were visualized using IGV and either auto scaled showing relative signal enrichment over control (input DNA), sequencing read coverage, or in the case of mod(mdg4) and brm, the presence of peaks from studies that had already carried out peak calling experiments. TrxG binding peaks are included for modifier of mdg4 (mod(mdg4)), brahma (brm), female sterile (1) homeotic (fs(1)h), little imaginal discs (lid), Trithorax-like (Trl), trithorax-related (Trr), Utx histone demethylase (Utx) and the C-terminal of the trithorax (trx) protein. The position and structure of the lncRNAs are indicated in green and the protein coding Hox genes in red. The transcription unit *Tre2* is indicated as a green block within the intron of *bxd*.

Drosophila have 5 HDACs that are enriched at active promoters, with enrichment correlating to gene expression level (Negre et al., 2011). In particular HDAC4 and HDAC1 mark PREs and are frequently correlated with H3K27me3 and Pho bound regions, while HDAC3 is correlated with H3K36me3 transcribed exons. Furthermore, HDACs have been shown to be recruited by lncRNAs to repress target genes (Kim et al., 2012) and one in particular, HDAC1, has been found to be necessary for PcG silencing at PREs (Tie et al., 2001) and for homeotic gene regulation in Drosophila (Chang et al., 2001). HDACs are also frequently linked to the regulation of lncRNAs (Castelnuovo and Stutz, 2015). In order to further investigate the roles of these lncRNA transcripts and to gain an insight into how they are regulated, HDAC ChIP-seq was investigated for binding to the lncRNA loci. HDACs remove acetyl groups from histones thereby remodeling chromatin into a transcriptionally repressed state. Figure 3.4.3 shows that the same regions that are bound by PcG and TrxG proteins are those that HDAC proteins bind in 0-12hr embryos. *Hox-G* is the only lncRNA within the Hox complex that has distinct peaks for all HDACs shown and the single exon transcripts of the ANT-C seem to be bound by HDAC4 and HDAC6. Interestingly, *lincX* has a clear binding peak for HDAC11 within its intron and a small peak for HDAC1, implicating these proteins in the regulation of *lincX*. There are small peaks of PcG proteins binding to the same region of *lincX* as HDAC11 that could suggest a shared role in the regulation of *lincX* or recruitment of these proteins by the lncRNA; however, this would require thorough investigations to confirm.

In the BX-C it is also noticeable that HDAC4 and HDAC6 tend to bind to the same regions that align with the binding of PcG and TrxG proteins (Fig.3.4.1 and 3.4.2). At the single exon lncRNA, *Tre2,* HDAC4 and HDAC6 bind along with HDAC1. Within the rest of *bxd*, there are no other sites that give prominent peaks of HDAC binding. HDAC4 and HDAC6 seem to bind independently of the other HDAC proteins throughout *iab-8*, with the possible exception of the *iab-7* PRE locus, which has a tentative peak for HDAC1. There are then some possible independent peaks that show HDAC1 binding in other regions of *iab-8*. Similar to PcG and TrxG ChIP-seq, no HDAC protein binding aligns with the antisense transcripts, *iab-4* or *Hox-O* (Fig.3.4.3).

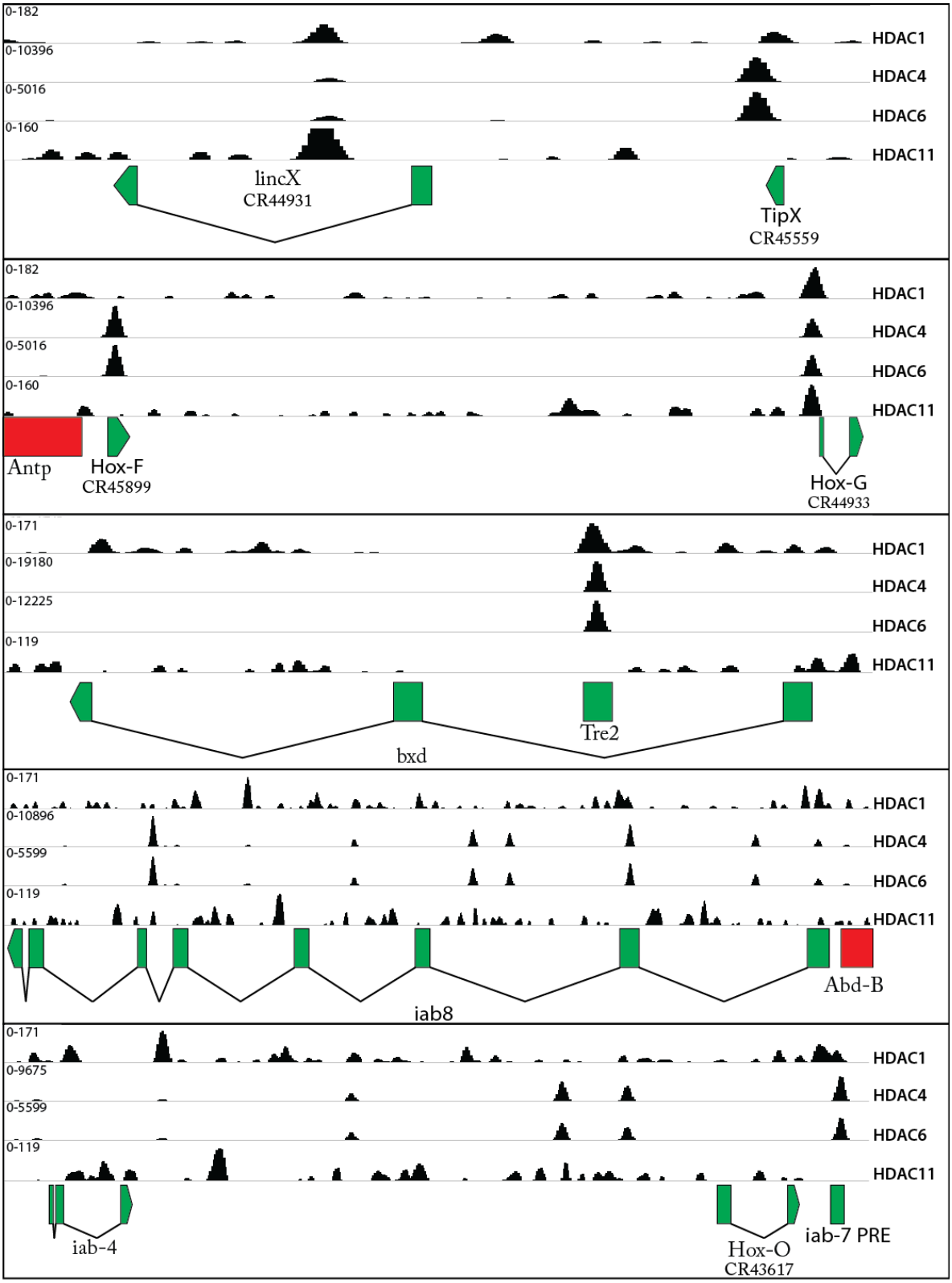


Figure 3.4.3. HDAC ChIP-seq binding profiles of 0-12hr embryos. Whole embryos were collected when aged between 0-12hrs post egg laying and ChIP-seq was carried out to determine where on the genome the HDACs were bound. The data was viewed using IGV and left to auto scale, with read numbers shown to the left, in each window in order to detect possible low-level binding. The read counts are shown starting from zero to the left of each track. Datasets are from GEO Series GSE20000.

3.5 Sequence based predictions of PREs and their evolutionary conservation

Motifs have been identified for PcG and TrxG components that bind DNA and these have been used to predict potential PRE/TREs. We utilized the jPREdictor program, which uses positive and negative training sets to score DNA sequences for enrichment of motif clustering whilst taking into account distances between individual motifs (Fiedler, 2006). We calculated weighted motif scores across the full-length sequences of the ANT-C and BX-C for *D. melanogaster, D. pseudoobscura* and *D. virilis* and used a graphical output to visualize all predicted PRE/TREs above the cutoff limit (Fig.3.5.1 and 3.5.2). Each predicted PRE/TRE was aligned with reference to Hox genes and other protein-coding genes or miRNAs (if aligned to a predicted PRE/TRE) and lncRNAs to identify their relative location to transcription. This allowed us to ascertain lncRNA candidates predicted to be associated with PRE/TREs. Furthermore, by using this analysis tool to predict PRE/TREs across the 3 different species we can investigate if predicted PRE/TREs are maintained throughout evolution in relation to lncRNAs.

Figure 3.5.1 graphically displays the results of the jPREdictor program across the ANT-C of *D. melanogaster, D. pseudoobscura* and *D. virilis* comparing the scores of each region above the calculated cutoff (red line) across the whole DNA sequence. There are some clearly defined high scoring regions that indicate strong candidate PRE/TREs within the ANT-C of each species and some that also appear to have been conserved throughout ~60My of evolution. The highest scoring peak in *D. melanogaster's* ANT-C is number 2, just upstream of Hox gene *Dfd*, which corresponds to peak 28 in *D. pseudoobscura.* This peak becomes less clear in *D. virilis* where there is a much smaller peak in the comparable region. However, there is still a good score for a peak within the *Dfd* locus (#32), which could indicate a shift in this regulatory region or be a consequence of motif turnover as some PRE/TREs have been shown to alter their genomic position and motif composition quite rapidly through evolution (Hauenschild et al., 2008).

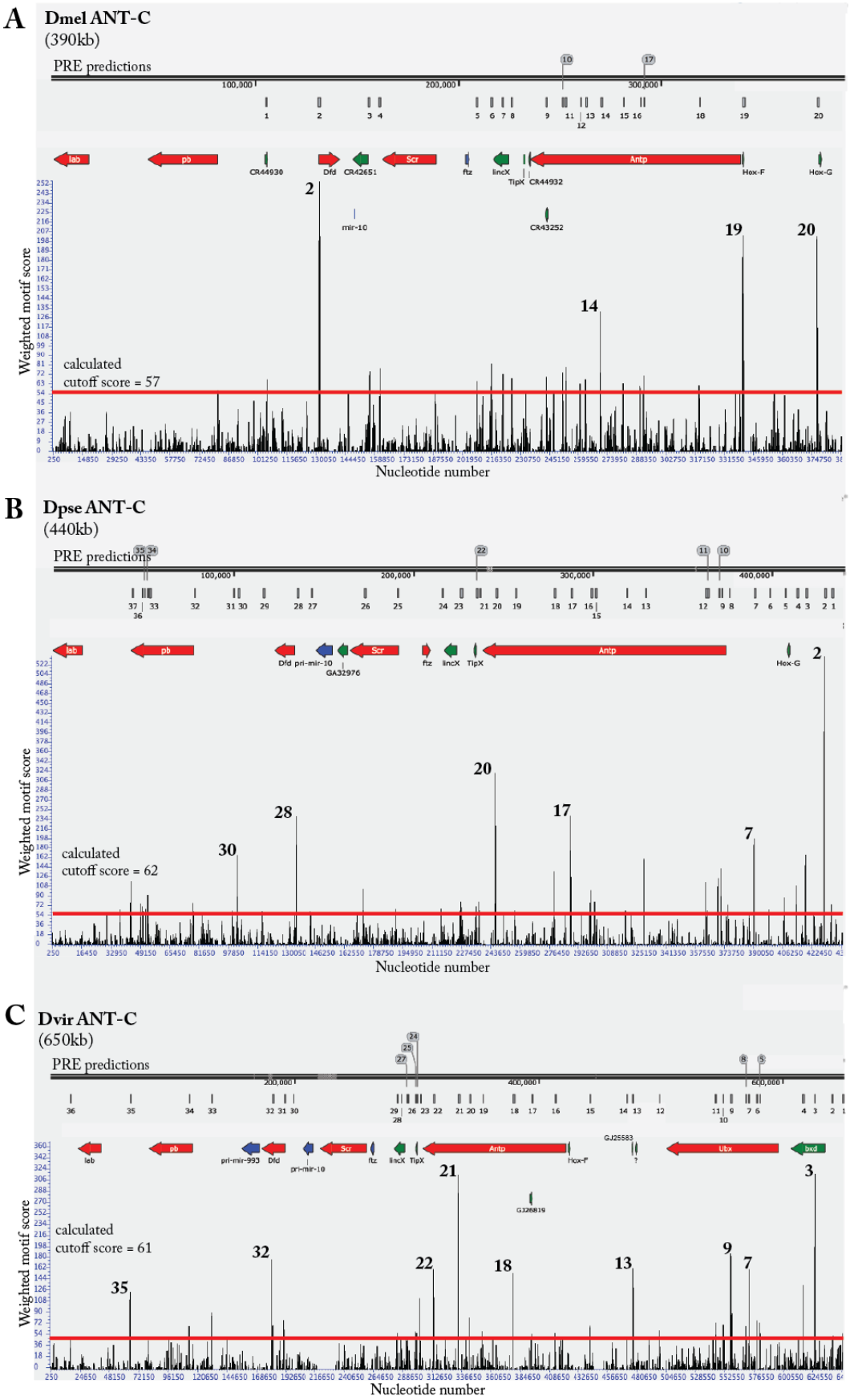


Figure 3.5.1. Prediction of PREs in the Antennapedia complex of *D. melanogaster*, *D. pseudoobscura* and *D. virilis* using the jPREdictor program. Graphical view of PRE/TRE predictions in each species with cutoff (red line). All PRE/TRE predictions above the threshold are depicted above the protein-coding genes (red), miRNAs (blue) and lncRNAs (green) and arbitrarily numbered. The potential PRE/TREs with the highest scores are numbered on the graph.

In *D. melanogaster's* ANT-C there are 2 significant peaks of PRE/TRE prediction upstream of the Hox gene *Antp,* numbered #19 and #20 (Fig.3.5.1-A). Peak #20 corresponds to a region just upstream of the lncRNA transcript *Hox-G* that does not show evidence of transcription and peak #19 perfectly aligns with lncRNA *Hox-F* that is clearly transcribed (Fig.3.2.2-A). Peak #19 from *D. melanogaster* aligns well with peak 7 in *D. pseudoobscura* (Fig.3.5.1-B), although it diminishes considerably, and in *D. virilis* there is no detectable peak corresponding to *Hox-F*, even though there is an apparently syntenic transcript (Fig.3.5.1-C). *D. melanogaster's* peak #20 appears to be detected in both *D. pseudoobscura* (#2) and *D. virilis* (#13) although the distance from this predicted PRE/TRE and the syntenic transcript of *Hox-G* becomes more distant. Interestingly, by using BLAST to search for *D. melanogaster Hox-G* sequence in *D. pseudoobscura* and *D. virilis* genomes, showed that the best alignment is between the same relative positions with respect to the corresponding peaks in each genome, suggesting the annotated syntenic transcripts may not be the orthologs of *Hox-G*.

The other notable peak within the ANT-C of *D. melanogaster* is #14, within the Hox gene *Antp* (Fig.3.5.1-A). This peak also appears to be conserved in *D. pseudoobscura* and *D. virilis*, aligning with peaks #17 and #21 respectively (Fig.3.5.1.B-C). The *D. pseudoobscura* and *D. virilis* ANT-C also share a high scoring PRE/TRE prediction close to the 3' end of *Antp* that is very low scoring in *D. melanogaster* (peak #9), demonstrating evolutionary changes of these response elements concurrent with other investigations (Hauenschild et al., 2008).

A prediction of PRE/TREs in the BX-C of the 3 Drosophila species also demonstrates the remarkable positional conservation for some of the predicted PREs. The highest scoring region in *D. melanogaster* aligns with the lncRNA *Tre2* (Fig.3.5.2.A-peak #14) and although a syntenic transcript could not be identified in either *D. pseudoobscura* or *D. virilis* RNA-seq, this region within *bxd* is still predicted to have a high confidence PRE/TRE (Fig.3.5.2.B-C, peaks #12 and #3 respectively).

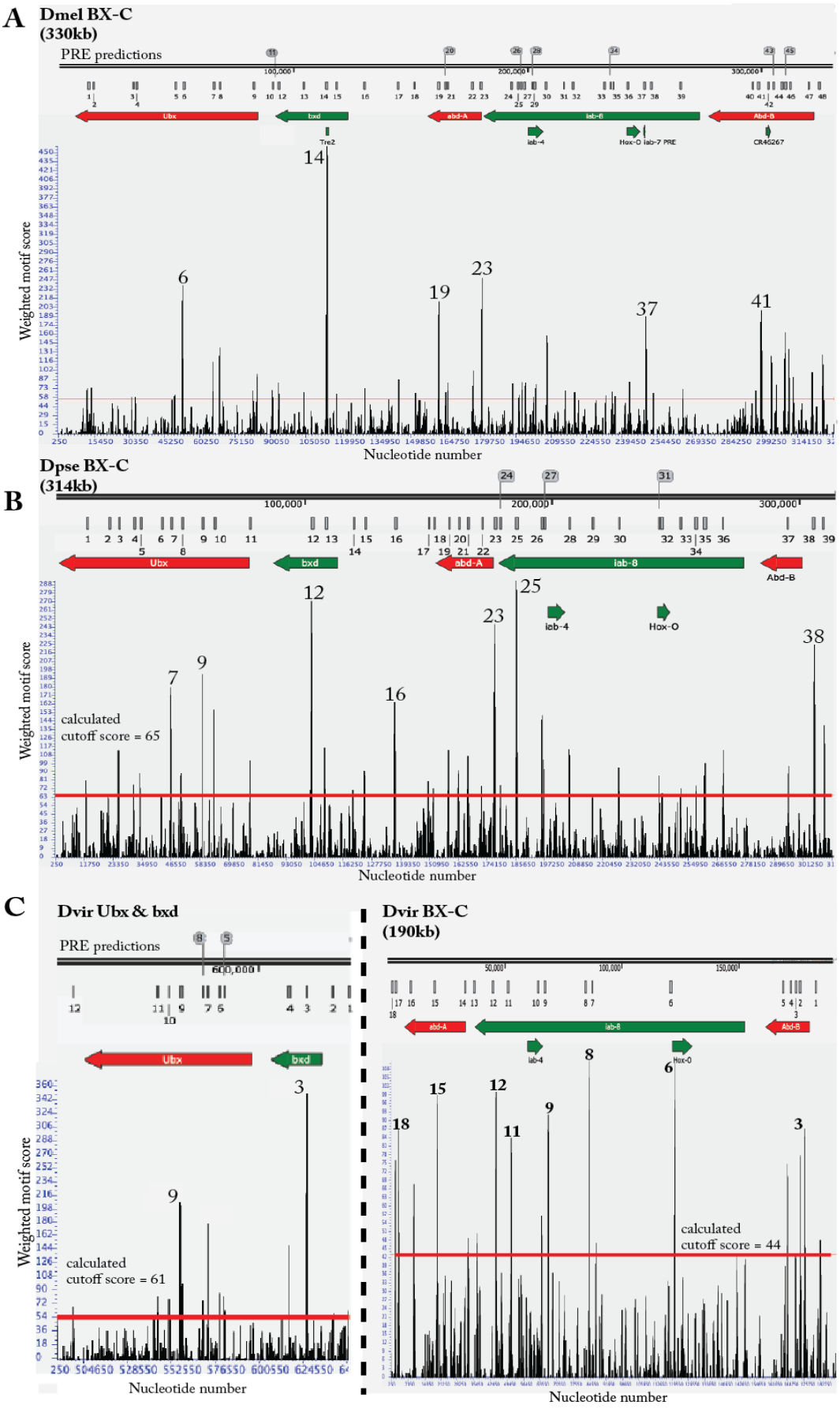


Figure 3.5.2. Prediction of PREs in the Bithorax complex of *D. melanogaster*, *D. pseudoobscura* and *D. virilis* using the jPREdictor program. Graphical view of PRE/TRE predictions in each species with cutoff (red line). All PRE/TRE predictions above the threshold are depicted above the protein-coding genes (red), miRNAs (blue) and lncRNAs (green) and arbitrarily numbered. The potential PRE/TREs with the highest scores are numbered on the graph.

The conserved prediction of the *Tre2* peak is particularly notable in the *D. virilis* genome where the Hox gene *Ubx,* and lncRNA *bxd,* are part of the ANT-C. We therefore analyzed a much larger region of 650 kb using jPREdictor (Fig.3.5.1-C). However, the program still found this region to be particularly high scoring for a PRE/TRE. This TRE has been known for nearly 30 years to be transcribed (Lipshitz et al., 1987) and identified as a functional TRE affecting *Ubx* for over 15 years (Rozovskaia et al., 1999). Another prominent and conserved PRE/TRE prediction peak in the BX-C is #6 (*D. melanogaster*), corresponding to numbers 7 and 9 in *D. pseudoobscura* and *D. virilis* respectively (Fig.3.5.2) found within *Ubx*.

Further upstream of the highest peak within *Ubx* is another positionally conserved high scoring prediction in *D. melanogaster* (peak #6), *D. pseudoobscura* (peak #9) and *D. virilis* (peak #7). This sequence has been identified as a predicted silencer PRE region due to HDAC1/HDAC4 binding overlapping H3K27me3, whilst lacking H3K27me3 (Negre et al., 2011). An origin recognition complex (ORC) protein-binding site is also almost perfectly aligned with the sequence of predicted PRE peak #6 in *D. melanogaster* (Eaton et al., 2011). There is also a TSS that aligns to the same PRE peak #6, identified by RAMPAGE, a combination of template switching and cap trapping (Batut et al., 2013; Batut and Gingeras, 2013). The identification of a TSS suggests that this PRE is likely to be transcribed, although this is not clear from RNA-seq that there is a single exon due to a high amount of transcription throughout this intron (Fig.3.2.3-A).

The beginning of *abd-A* also has a high scoring region predicted with jPREdictor in *D. melanogaster* and *D. pseudoobscura* (both peak #23) along with a lower scoring peak in *D. virilis* (peak #14), still above the threshold (Fig.3.5.2). This is likely to represent a PRE/TRE overlapping *abd-A's* promoter, a trait observed for several other PREs such as *en, hh* and *pho* (Muller and Kassis, 2006). Near the 3' end of *iab-8* there is a region that scores high in both *D. virilis* and *D. pseudoobscura* (peaks #12 and #25 respectively), whereas this specific region in *D. melanogaster* seems not to have a PRE/TRE, although overall the *D. melanogaster's iab-8* has several more predicted PRE/TREs than the other 2 species.

There is a small peak just downstream of *iab-4* in both *D. melanogaster* and *D. pseudoobscura* (peaks 30 and 28 respectively) that appear to match *D. virilis* peak #9 and *D. melanogaster's iab-4* has some low scoring peaks throughout that are just over the threshold, whereas there are no PRE/TRE predictions for the syntenic *iab-4* of *D. pseudoobscura* or *D. virilis* (Fig.3.5.2). The *iab-7* PRE, identified in *D. melanogaster*, is another example confirming that the jPREdictor tool is a reliable prediction method of PRE/TREs as this site has another high scoring predicted peak (#37). Using RNA-seq, a syntenic transcript could not be found in either *D. pseudoobscura* or *D. virilis* and there is no obvious matching high scoring peak in these 2 other species. Finally, peak#48 in *D. melanogaster*, peak #3 in *D. virilis* and peak #38 in *D. pseudoobscura* seem to be a match relative to the promoter region of *Abd-B* and then within *Abd-B, D.*

melanogaster and *D. virilis* have 7 and 4 predicted PRE/TREs respectively, but just 1 was found in *D. pseudoobscura's Abd-B*.

The jPREdictor was designed to predict PRE/TREs in *D. melanogaster*, so to test the reliability of its predictions in other species we used pre-existing ChIP-seq data to investigate binding of PcG and TrxG proteins to these sites that had been carried out in both *D. melanogaster* and *D. pseudoobscura*. The sequencing reads for each ChIP-seq experiment were mapped to either *D. melanogaster* r6.08 or *D. pseudoobscura* r3.03 and visualized with IGV to align the phastCons file, downloaded from UCSC, with gene annotations and PRE/TRE predictions (Fig.3.5.3). The phastCons program measures evolutionary conservation using multiple sequence alignments across the whole genomes of 27 insect species to estimate the probability of each nucleotide belonging to a conserved element, taking into account flanking sequences and the process of DNA substitution (Siepel, 2005). PhastCons relies on whole-genome alignments carried out on these insects that are regularly updated as new genome releases come out. However, not all of these insects' genomes have been reliably assembled and the failures in alignments can be reflected by phastCons analysis. Therefore, the program will miss some conserved elements, but the majority of the positively conserved elements in the Hox complex do match known conserved regions. This is evident in protein-coding Hox genes that are known to be conserved by their 180bp homeobox sequence (Heffer and Pick, 2013). However, *Dfd*, a Hox protein that is transcribed in the opposite orientation relative to other Hox genes in *D. pseudoobscura*, is not found to be positively conserved within the phastCons analysis, suggesting this can cause problems for multiple sequence alignments and therefore conservation scoring.

It is interesting to compare the prediction tools with the ChIP-seq data sets and syntenic lncRNAs found by RNA-seq. The *Hox-G* region has a peak indicating sequence conservation throughout 27 species with the phastCons analysis (Fig.3.5.3-A), along with PcG and TrxG ChIP-seq binding and a predicted PRE/TRE. This is the only site in *D. melanogaster's* ANT-C that has positive results in all 3 criteria. In *D. pseudoobscura* the syntenic transcript does not align with protein binding or PRE/TRE prediction in that region, but instead seems shifted. This could indicate that this transcript is not *Hox-G's* ortholog, or that the PRE/TRE has been separated by a greater distance in *D. pseudoobscura*. *Hox-F* has clear protein binding matching the PRE/TRE predictions in *D. melanogaster*, but no indication of sequence conservation, along with very little evidence of protein binding and no PRE/TRE prediction (Fig.3.5.3.A-C). At the *TIPX* loci in *D. melanogaster*, there is a phastCons peak showing sequence conservation and clear PcG/TrxG protein binding peaks, but no PRE/TRE predicted and in *D. pseudoobscura*, the transcript no longer seems to be bound by these proteins.

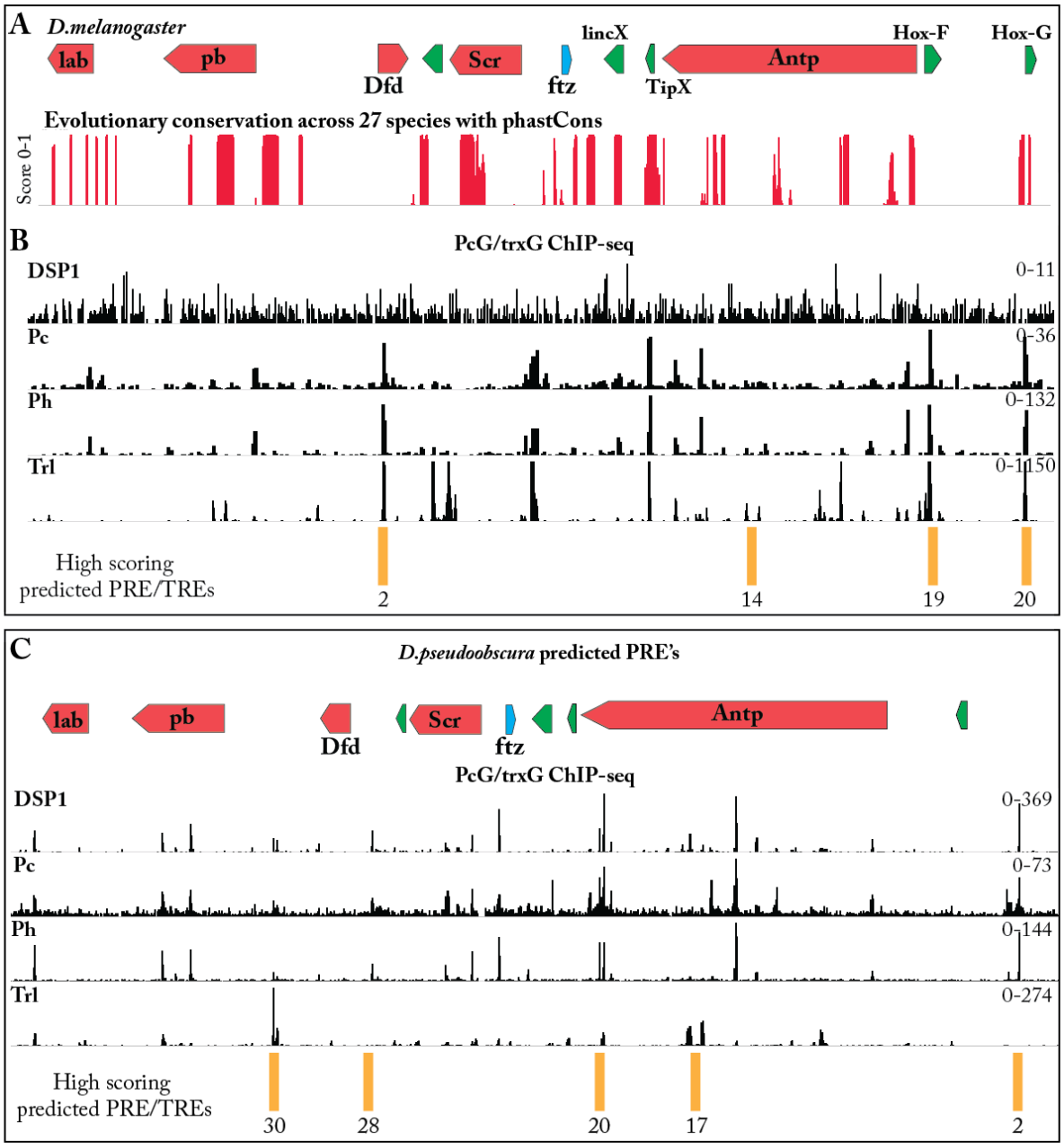


Figure 3.5.3. Analysis of conservation of *D. melanogaster* ANT-C sequence with PcG/TrxG protein binding and predicted PREs. The phastCons multiple sequence alignment across 27 insect species is aligned to Hox genes and lncRNAs in the ANT-C of *D. melanogaster*, showing the regions with the highest conservation as red peaks (A). PcG and TrxG protein ChIP-seq from the same study carried out in *D. melanogaster* and *D. pseudoobscura* (GEO series GSE60428) was also aligned to each species ANT-C genes and PRE/TRE high scoring predictions (yellow bars) using jPREdictor (B-C).

The *Dfd* promoter region is bound by PcG and TrxG proteins in both *D. melanogaster* and *D. pseudoobscura*, aligning with PRE/TRE predictions in each species. This region is likely to have been missed by the phastCons analysis due to the change in orientation (Fig.3.5.3.A-C). Overall, throughout the ANT-C, it does not seem that either PRE/TREs (based on predictions) are any more conserved than regions that are bound by PcG or TrxG proteins, but *lincX, TIPX* and *Hox-G* loci do seem to have an increase in evolutionary conservation throughout the 27 species analyzed by the phastCons study (Fig.3.5.3-A).

The phastCons profile in the BX-C does not provide very clear indications of specific regions of conservation, as there is no clear increase in conservation aligning to any of the regulatory or protein-coding regions that we would expect and instead gives messy and unclear signal throughout (Fig.3.5.4-A). We can however, see indications of protein binding peaks at all predicted PRE/TREs in *D. melanogaster's* BX-C. In *D. pseudoobscura* we can see ChIP-seq peaks aligning to 6 of 8 predicted PRE/TREs, for numbers 7, 9, 12, 23, 26 and 38, suggesting that these regulatory regions can be reliably predicted in other species. The *iab-7* PRE, just downstream of *Hox-O*, aligning to predicted PRE/TRE #37, was not detected in *D. pseudoobscura*. However, there are distinct peaks of Trl, Ph and Pc binding in *D. pseudoobscura* in a very similar position relative to the syntenic *Hox-O* transcript, suggesting this could be the syntenic region of *iab-7* PRE (Fig.3.5.4.A-B).

The ChIP-seq peak within *Ubx* seems well conserved in both species matching predicted PRE/TREs in each, peak #6 in *D. melanogaster* and peak #7 in *D. pseudoobscura*. In *D. pseudoobscura* another predicted PRE/TRE within *Ubx* (#9) has some evidence of protein binding in the ChIP-seq profiles and aligns with a similar small peak in *D. melanogaster* that could match a low scoring predicted PRE/TRE, either 7 or 8 (Fig.3.5.4.A-B). The *Tre2* transcript has clear binding of PcG and TrxG proteins along with a high scoring PRE/TRE prediction in *D. melanogaster* and although the syntenic transcript could not be identified in the RNA-seq due to noisy transcription throughout *bxd's* introns in other species (Fig.3.2.4-B), there is a clear syntenic region based on the ChIP-seq and PRE/TRE predictions (Fig.3.5.4-B).

Using the prediction tool combined with the ChIP datasets allows us to have a good assessment of where PRE/TREs are in both the *D. melanogaster* and *D. pseudoobscura* genomes. We can use the Hox genes, which remain well conserved in relative positions and intron-exon structure, to inform us of syntenic location of PRE/TREs between these two species. When comparing the positions of the PRE/TREs between the two species we can see that there are very few that appear to remain in the same syntenic position using either prediction alone (Figs.3.5.1 & 3.5.2) or including ChIP data (Figs.3.5.3 & 3.5.4). This agrees with a previous study that found PRE/TREs evolved rapidly, dramatically changing in numbers and positions {Hauenschild, 2008 #695}. However, this study did not consider transcription of PRE/TREs and we can see within the

Hox complex that syntenic lncRNA transcripts that align with PRE/TREs in *D. melanogaster*, do not continue to have evidence of PRE/TRE function from either prediction methods or ChIP data in *D. pseudoobscura* (Figs.3.5.3 & 3.5.4). In some cases the PRE/TRE region (transcribed or not) has evidence of evolutionary sequence conservation throughout 27 insect species (Fig.3.5.3) and in some cases this region remains a PRE/TRE (upstream *Hox-G*), but in others loses any indications that it is a PRE/TRE (*TipX*) (Fig.3.5.3).

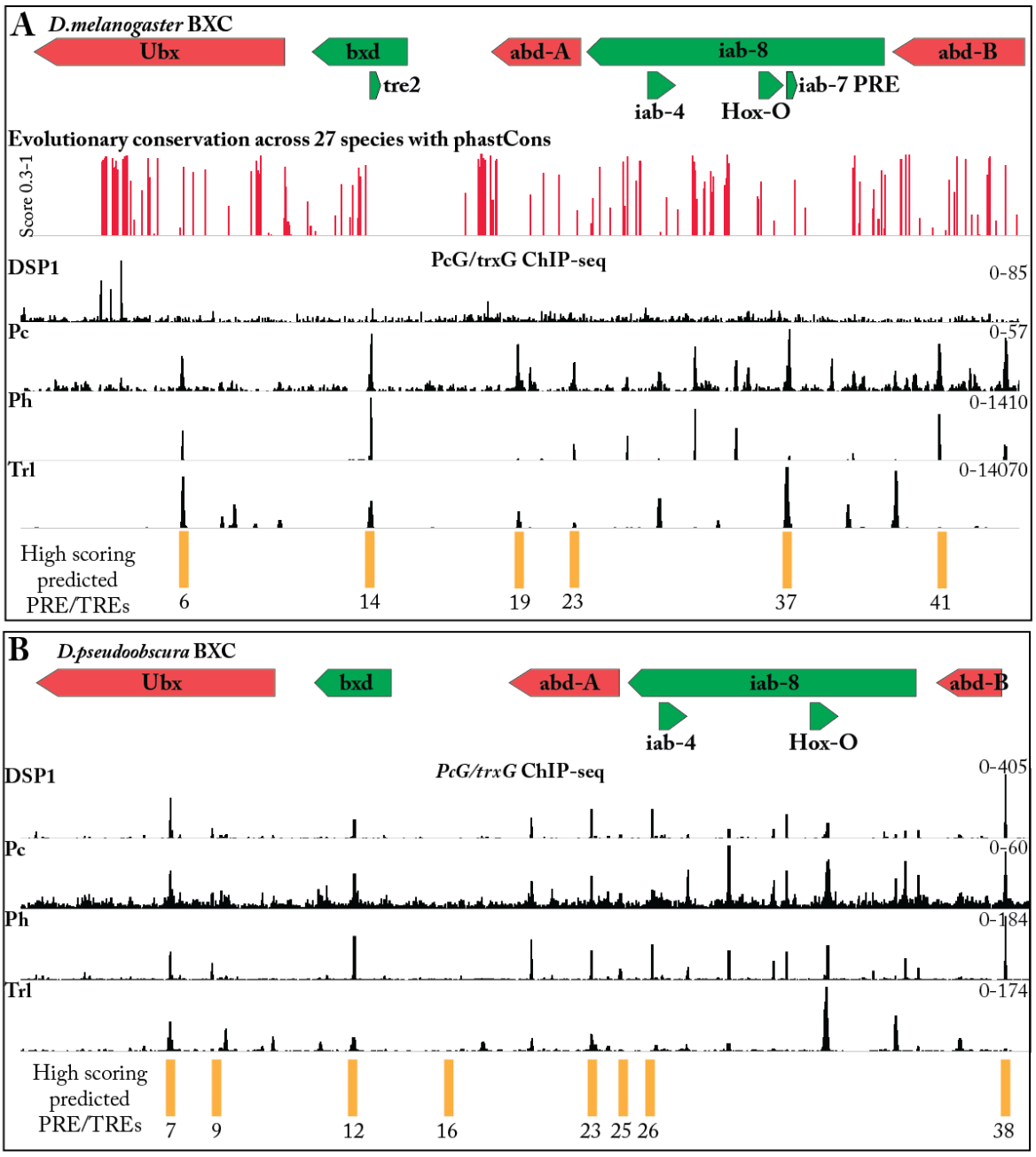
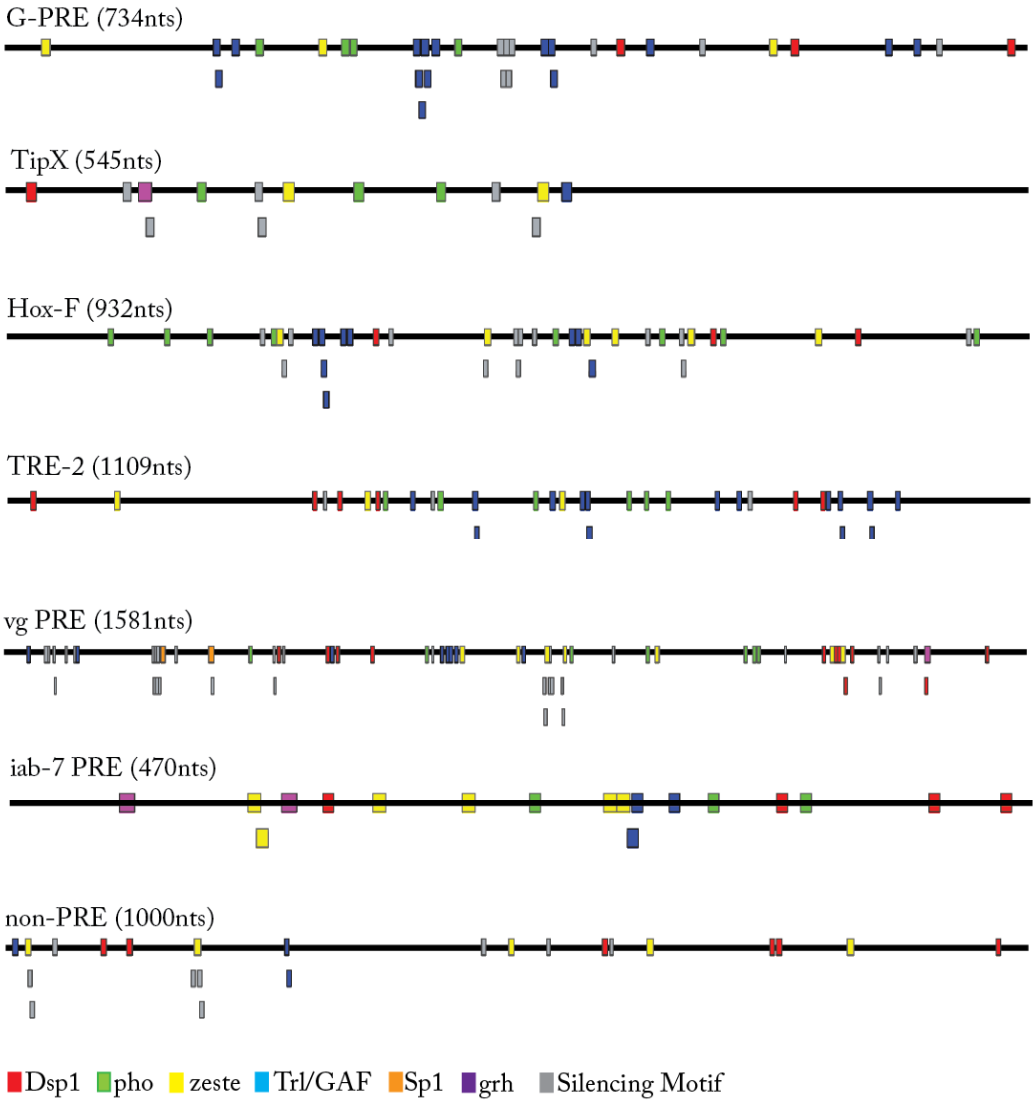


Figure 3.5.4. Analysis of conservation of *D. melanogaster* BC-C sequence with PcG/TrxG protein binding and predicted PREs. The phastCons multiple sequence alignment across 27 insect species is aligned to Hox genes and lncRNAs in the BX-C of *D. melanogaster*, showing the regions with the highest conservation as red peaks (A). PcG and TrxG protein ChIP-seq from the same study carried out in *D. melanogaster* and *D. pseudoobscura* (GEO series GSE60428) was also aligned to each species ANT-C genes and PRE/TRE high scoring predictions (yellow bars) using jPREdictor (B-C).

The jPREdictor scores the PRE/TRE predictions based on clustering of motifs that are from both PcG and TrxG proteins, but does not attempt to discern if a site is more likely to be either a PRE or TRE or can act as both. In an attempt to characterize the transcribed PRE/TREs that would therefore also be classed as lncRNAs due to their transcription, we annotated the DNA binding motifs of both PcG and TrxG proteins, along with a silencing motif that is essential for PRE silencing functions (Okulski et al., 2011) (Fig.3.5.5). We also annotated the PRE identified upstream of *Hox-G* as this seems to be a region that has been well conserved throughout insects and has very high scoring PRE/TRE predictions clear peak demonstrating PcG and TrxG protein binding (Fig.3.4.1 and 3.4.2), along with 4 members of the HDAC family (Fig.3.4.3) suggesting this is a particularly key regulatory element during development. To compare the motif distribution and frequency of motifs of the putative PRE/TREs, we also analyzed a random intergenic DNA sequence that has no evidence of being a PRE or TRE or having any regulatory functions, along with the *vg* PRE that has previously been characterized.

The motifs of DNA binding proteins belonging to or involved in the recruitment or function of PcG members are Dsp1, Pho, Sp1 and Grh and those that belong to TrxG are Trl/Gaf and Zeste. The silencing motif, GTGT, was identified in the *vg* PRE as necessary for the silencing capabilities of this PRE (Okulski et al., 2011) that has recently been linked to the sequence specific DNA binding of *Combgap* (*cg*) (Ray et al., 2016). The highest number of motifs found in the G-PRE is for TRL protein with 15. This number seems significant when comparing to the *vg* PRE that is almost twice the length and only contains 8 Trl motifs (Fig.3.5.5). Furthermore, the Trl motif can be seen to cluster in the G-PRE sequence by the blue rectangles in 3 specific regions, suggesting these are the specific sites that Trl could be recruited to. There are low numbers of other motifs when comparing to other PRE/TREs, particularly the non-PRE sequence.

TIPX is interesting as it is transcribed and has been shown to clearly have PcG and TrxG proteins bind to its sequence in ChIP-seq datasets. However, it is not predicted to be a PRE/TRE with the jPREdictor program and very few DNA binding motifs are found to be present in its sequence. This suggests it could have a unique method of action compared to the established PREs. *Hox-F* has similarities to G-PRE in that it has clusters of TRL motifs, although also has 8 PHO motifs and 13 silencing motifs. The lncRNA *Tre2* also has many TRL motifs (15) in clusters but very few silencing motifs. Given the length of *Tre2*, the other motifs identified are unlikely to be significant if comparing to the non-PRE, suggesting that this is truly likely to be a TRE. The *vg* PRE has several clusters of the silencing motif that have been shown to be essential for its silencing function and also contains 11 DSP1 motifs, although it is close to the number that would be found by chance in such a large region. The *iab-7* PRE has no silencing motifs but has been shown to require PHO and TRL/GAF to carry out its silencing activity, but the silencing motif is not well understood and has only been investigated in the *vg* PRE and is unlikely to be required for silencing by all PREs (Okulski et al., 2011).



Number of motifs

	Dsp1	pho	zeste	Trl/GAF	Sp1	grh	Silencing Motif
G-PRE	3	4	3	15	0	0	8
TipX	1	3	2	1	0	1	6
Hox-F	3	8	6	9	0	0	13
TRE-2	6	6	3	15	0	0	3
vg PRE	11	7	7	8	2	1	31
iab-7 PRE	4	3	6	3	0	2	0
non-PRE	6	0	5	3	0	0	9

Figure 3.5.5. PRE/TRE motif clustering of PcG/TrxG binding regions of experimental and validated PRE/TREs. DNA sequence motifs bound by members of PcG/TrxG proteins are depicted in genomic regions of the Hox complex that had peaks from ChIP experiments (Fig 2.4 & 2.5). An experimentally validated PRE at the vestigial (vg) locus (Herzog et al., 2014) and an unbound region are also shown. The G-PRE (upstream *Hox-G*) and non-PRE have no evidence of transcription. The *vg* PRE and iab-7 PRE is transcribed in both directions and *TipX, Hox-F* and *TRE-2* appear to be transcribed in one direction. A summary table of the number of motifs found for each PcG/TrxG protein is shown along with the number of silencing motifs, demonstrated to be essential for PRE silencing (Okulski et al., 2011).

3.6 Homeotic mutations from Gal4 driven expression of *Hox-G* and G-PRE

Over-expression or ectopic expression is a common technique used for the study of gene function that has been particularly successful in investigation of Hox genes, as when ectopically expressed they often caused homeotic transformations that could be easily seen and interpreted in 1st instar larvae and in the adult fly. The development of the Gal4-UAS system has been made these types of experiments very simple genetically and very precise with respect to temporal and spatial expression of the gene of interest (Rorth, 1996). We chose to explore the functions of the *Hox-G* lncRNA using this method to find out if it possessed any diffusible or trans function. As the lncRNAs in the Hox complexes of diverse animals regulate the Hox genes themselves we wanted to examine if adult homeotic phenotypes would be generated by ectopic expression that might indicate either that *Hox-G* was involved in regulating Hox genes and possibly which Hox gene(s) it could be. We began by identifying available transposable elements in the region of *Hox-G* that may contain sequences that would allow them to be manipulated to study *Hox-G's* function. We identified 3 PBac(WH) elements inserted upstream, within the second exon and downstream of the *Hox-G* transcript. These containing FRT sites that enable flippase mediated recombination and terminal UAS sites that allow Gal4 to drive expression from neighboring promoters (Fig.3.6.2-A) (Rorth, 1996; Thibault et al., 2004). Table 3.6.1 summarizes the expression patterns and developmental stages of the Gal4 drivers used for overexpression of the lncRNA and PRE.

Additionally we also cloned both the *Hox-G* transcript and the G-PRE into a p(UAST) expression vector (Brand and Perrimon, 1993), to allow us to test the effects of driving expression from a variety of tissues and at different developmental times to test the effects of ectopic *Hox-G* transcription. The putative *Hox-G* PRE was similarly cloned to allow investigation of its PRE activity in a test for PSS and to act as a control for expression of the *Hox-G* lncRNA. These were randomly inserted by P-element transformation into *D. melanogaster's* genome by microinjection into w^{1118} flies (BestGene Inc) for screening of the mini-white marker to show insertion. From these injections, 9 transformants were identified by orange eye color for each the G-PRE and lncRNA constructs. All lines were viable with no apparent phenotype when homozygous suggesting that the insertions did not disrupt the functions of any genes required for viability. The homozygous lines were then crossed to different embryonic Gal4 drivers.

The Gal4 lines subsequently used to drive ectopic RNA expression were imaged (Fig.3.6.1) along with the PBac(WH) transgenic fly lines (Fig.3.6.1) and p(UAST) transformants (Fig.3.6.3 and 3.6.4) to determine the specificity of any phenotypes observed when Gal4 lines were crossed with UAS lines. We refer to the PBac(WH) lines as 1, 2 and 3 for simplicity, with PBac(WH)1 being upstream of *Hox*-G, PBac(WH)2 being in the second exon and PBac(WH)3 being downstream. We observed a wide variety of phenotypes resulting from crossing PBac(WH)1 and PBac(WH)2 to early embryonic Gal4 drivers (Fig.3.6.2.B-J). These phenotypes were quite strong

and generally included a combination of missing appendages, necrotic/black marks in the thorax and abdomen and abdominal cuticle malformations. One striking phenotype that was frequently observed when the maternal Gal4 driver, α-tub Gal4-2, was crossed to the PBac(WH)1 element was that adult flies would develop an abdomen that would collapse ~3days after hatching. This was also seen when PBac(WH)1 was driven by 69B-Gal4 (Fig.3.6.2-I). This normally led to early death, potentially due to starvation or dehydration. Along with this phenotype, these flies often had black tissue growing on their heads, occasionally in combination with abdominal malformations, but more often independent of each other (Fig.3.6.2-B). A more frequent phenotype in α-tub Gal4-2 driving PBac(WH)1 was a missing T3 or the T3 becoming misshapen and twisted and looked like it may be overgrown, as it would have regions of an enlarged leg width, or in other cases the T3 would look like it was possibly transforming into an antennae based on bristle patterns and round, antennae-like shape replacing the tarsal segments (Fig.3.6.2-B and E and Fig.3.6.11). Adult flies with a missing T3 leg would frequently have a black mark in the abdomen near the region closest to the leg primordial from the thorax. When these flies were opened to investigate the black mark, it was found that a partially formed T3 leg was growing from the thorax, compressed into the abdomen, where it grew inside the fly (Fig.3.6.2-E). Another common phenotype was abnormal patterning of the dorsal abdominal segments where they were found to merge into each other when PBac(WH)1 was crossed to α-tub Gal4-2 (Fig.3.6.2-D) and PBac(WH)2 was crossed to 69B-Gal4 and *en*-Gal4 (Fig.3.6.2.H-J). A rare phenotype was for the whole abdomen to twist slightly (~40°) in the PBac(WH)2: α-tub-Gal4-2 cross (Fig.3.6.2-C). This twisted abdomen phenotype has been seen in perturbations of the evolutionary related genes *rotated abdomen* and *twisted*, the Drosophila orthologs of human *O-mannosyltransferase-1* and *2* respectively, genes that are linked to brain, eye and muscle development (Ichimiya et al., 2004; Lyalin et al., 2006). Both *twisted* and *rotated abdomen* are expressed in embryonic stage 10 (Ichimiya et al., 2004) and *rotated abdomen* is also maternally deposited (Lyalin et al., 2006) although the regulation of these genes is yet to be established and so it is not yet clear if these genes are linked to our phenotypes or not.

When 69B-Gal4 was crossed to PBac(WH)2 we frequently observed that along with abnormal abdominal segmentation, the most posterior abdominal segment, containing the genitalia, has become enlarged (Fig.3.6.2-F and H). Another rare phenotype was for the wings not to have unfolded in adult flies. This was found only in crosses of 69B-Gal4 x PBac(WH)1 and in the fly shown was combined with a collapsed abdomen possibly suggesting overall poor development (Fig.3.6.2-H). A frequently observed phenotype in all Gal4 driven crosses was missing T3 or tarsal leg segments or legs that looked overgrown and were bigger than wild-type, becoming twisted, particularly T3 (Fig.3.6.11). Also common was necrotic patches of black tissue at various positions along the legs (Fig.3.6.11). The underdeveloped legs would often have a stump where it appeared the most distal tarsal regions of the leg did not finish developing (Fig.3.6.2-J,

3.6.10-D and 3.6.11). A summary of penetrance of each of these phenotypes is recorded in Table 3.6.1. The table groups each of the Gal4 drivers together with red columns to match PBac(WH)1 crosses. This way we can assess if specific phenotypes occur more frequently in certain driver lines or different PBac(WH) insertion sites. The main difference between phenotypes that were observed when using different Gal4 driver lines to drive *Hox-G* or G-PRE was when using α-tub-Gal4-2 that is maternally deposited and has a ubiquitous strong expression throughout development and continues to be strongly expressed in adults (Kalfayan and Wensink, 1982). The α-tub-Gal4-2 driver produced abdominal defects and may have been weaker in the T3 leg, as the ingrown leg was only seen when using this line also. Besides the α-tub-Gal4-2, there was no clear link between using certain Gal4 driver lines and the types of phenotypes produced as all gave similar or the same phenotypes. Furthermore, the same phenotypes were generated from driving ectopic expression of *Hox-G,* G-PRE or Pbac(WH) constructs.

Table 3.6.1. Gal4 drivers lines expression patterns and stages of expression

Gal4 line	Stages of expression	Pattern of expression	References
α-tub-Gal4-2	• Maternally deposited • Throughout development • Adulthood	• Ubiquitous	(Kalfayan and Wensink, 1982; Matthews et al., 1989; Natzle and McCarthy, 1984)
en-Gal4	• Embryonic stages 4-16 • 3rd instar larvae	• Native *en* • Segment polarity expression • Fat body, cuticle, imaginal disc, digestive system	(Harrison et al., 1995; Tomancak et al., 2002; Tomancak et al., 2007; Weiss et al., 2001)
69B-Gal4	• Embryonic stages 9-17 • 3rd instar larvae	• Ectoderm • Imaginal discs	Brand, A, 1997 (personal communication to FlyBase)(Staehling-Hampton et al., 1994a)
dpp-Gal4	• larval stage • 3rd instar larvae	• Imaginal discs • Morphogenetic furrow • Midgut	(Cherbas et al., 2003; Mukherjee et al., 2000; Staehling-Hampton et al., 1994a)

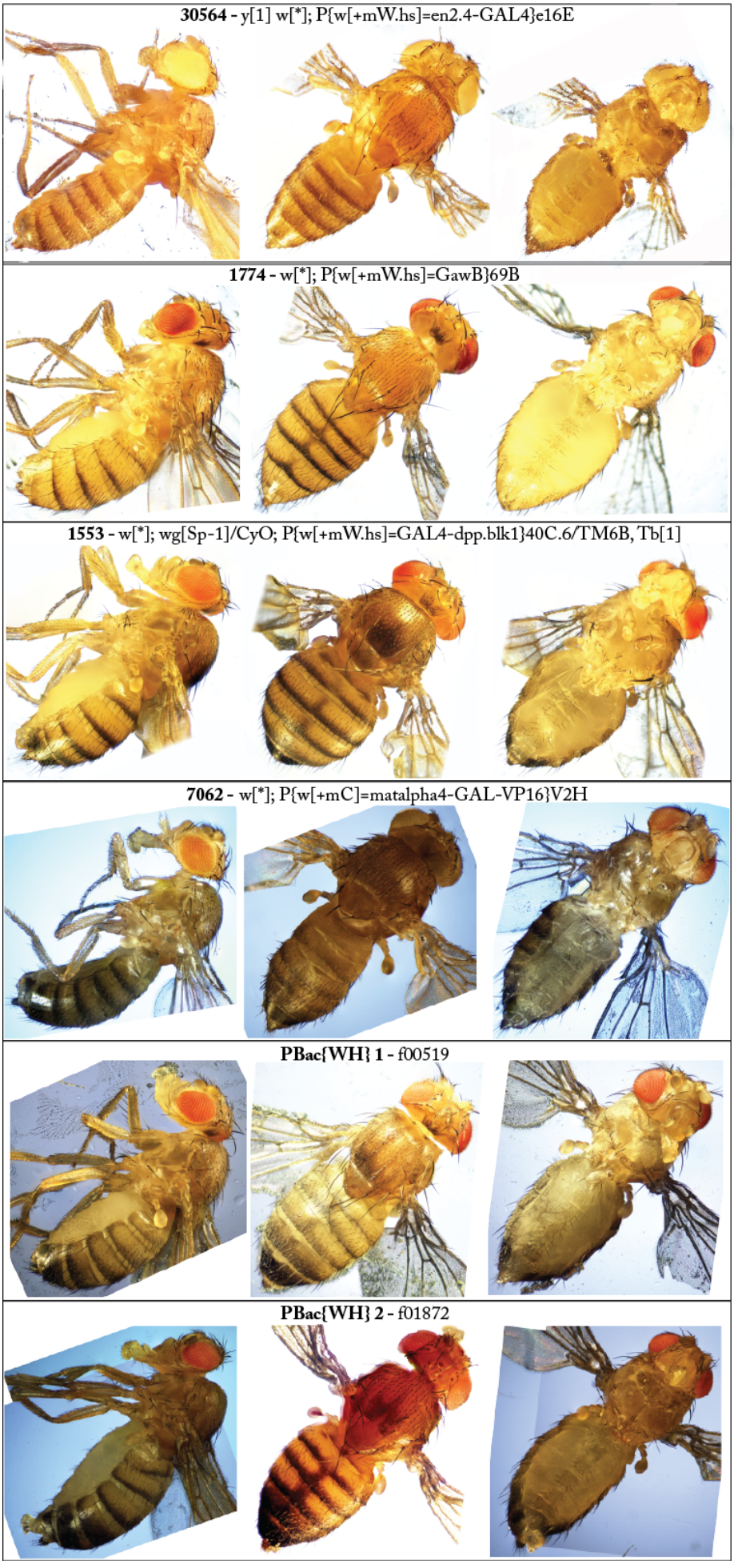


Fig 3.6.1 – Legend on next page

Figure 3.6.1. Gal-4 driver and PBac(WH)1 and 2 adult fly images. Side (left), dorsal (middle) and ventral (right) views of adult flies used to drive expression of *Hox-G*. Numbers in bold of Gal4 driver flies relate to the Bloomington Stock center ID. The numbers beginning 'f' of PBac(WH) flies are the Harvard Exelixis identifiers.

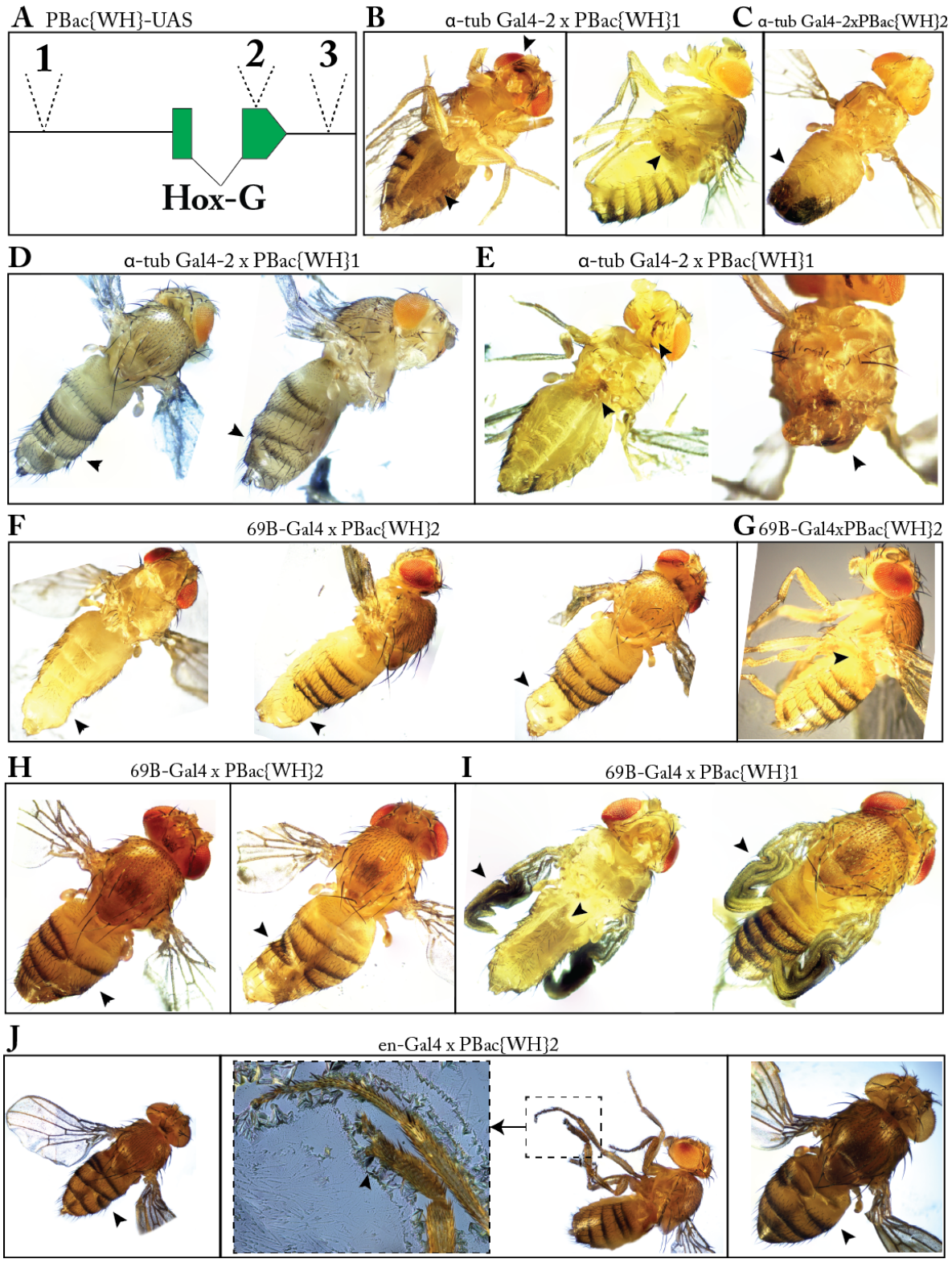


Figure 3.6.2. Mutations from Gal4 driven UAS-Piggy Back constructs upstream and within *Hox-G*. Flies containing UAS binding sites within the second exon of *Hox-G* were used from the Harvard Exelixis Collection (https://drosophila.med.harvard.edu) (f01872) to drive expression of the remaining portion of the second exon, downstream of the insertion site. Flies generated had a number of homeotic phenotypes, such as abdominal stripes that had not properly formed (A and D), a missing T3 leg (C) and leg sections missing from T3 legs (B and C). B is a magnification of C showing the shortened T3 leg, missing tibia sections t1-5, whilst still forming the tarsal claw. - Also note the regional expression of mini-white within the fly's eye (C) and the slightly smaller wing (A).

Observed Mutations	Fig	control PBac(WH)1 n=100	control PBac(WH)2 n=100	control en-Gal4 n=100	control 69B-Gal4 n=100	control dpp-Gal4 n=100	control α-tub Gal4-2 n=100	α-tub Gal4-2 xPBac(WH)1	α-tub Gal4-2 xPBac(WH)2	dpp-Gal4 xPBac(WH)1	dpp-Gal4 xPBac(WH)2	en-Gal4 x PBac(WH)1	en-Gal4 x PBac(WH)2	69B-Gal4 xPbac(WH)1	69B-Gal4 xPbac(WH)2
Missing 1 haltere	**3.6.5-G**	0	0	0	0	0	0	0	0	5% n=65	0	5% n=80	8% n=114	0	8% n=121
Reduced haltere	**3.6.5-E**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Missing both halteres	**3.6.5-C**	0	0	0	0	0	0	0	0	0	3% n=131	0	6% n=83	0	0
Missing 1 x T3	**3.6.5-D**	0	0	0	0	0	0	10% n=104	0	0	6% n=101	0	0	0	0
Missing 2 x T3	**3.6.5-D**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Overgrown T3	**3.6.11**	0	0	0	0	0	0	11% n=104	0	0	0	7% n=80	23% n=70	0	0
Necrotic legs	**3.6.11**	0	0	0	0	0	0	0	15% n=77	17% n=90	12% n=101	0	0	3% n=98	8% n=121
Abnormal abdominal stripes	**3.6.5-C &D**	0	2%	0	0	0	1%	10% n=104	0	0	0	16% n=80	20% n=61	7% n=98	19% n=121
Black growth head	**3.6.8-D**	0	0	0	0	0	0	6% n=104	0	NE	0	0	NE	0	NE
Twisted Abdomen	**3.6.2-C**	0	0	0	0	0	0	0	3% n=77	NE	NE	NE	NE	0	NE
Unfolded wings	**3.6.2-I**	0	0	0	0	0	0	NE	NE	NE	NE	NE	NE	1% n=98	NE
Collapsed abdomen <3days from hatching	**3.6.10-D**	0	0	0	0	0	0	8% n=104	NE	NE	NE	NE	NE	9% n=98	NE

Table 3.6.2. Penetrance table mutations from Gal4 driven UAS-Piggy Back constructs upstream and within *Hox-G*. Controls of mutations found in PBac(WH)1, PBac(WH)2 and Gal4 driver lines. Penetrance of each mutation shown as a percentage with numbers counted. The Gal4 driver lines have been grouped and the PBac line (WH1) highlighted in pink for easier comparison. Figure references displayed for each phenotype. NE = not evaluated.

Lines carrying the p(UAST) vector inserted carrying either the G-PRE sequence or the *Hox-G* sequence were homozygozed and crossed to the same set of Gal4 driver lines. Representative images of adult flies with homozygous p(UAST) insertions that produced phenotypes are shown in Figures 3.6.3 and 3.6.4 to show they were healthy and to compare to the mutant phenotypes. The phenotypes from Gal4 driven expression of *Hox-G* or G-PRE generate very similar phenotypes to those observed in the PBac(WH) fly crosses with Gal4, although at varying degrees of penetrance. One of the most noticeable and frequent phenotypes was missing halteres. This was found in the majority of the crosses, with either 1 haltere missing or both. None of the flies from the PBac(WH) experiments had produced flies missing both halteres suggesting that the p(UAST) ectopic expression increases the penetrance of the phenotype (Fig.3.6.5.B and D). Furthermore, flies missing 1 or 2 halteres also often were missing at least 1 T3 leg, but in this case no black inclusions were observed in the abdomen and no rudimentary leg inclusions could be found in the abdomen, and when opened up there was no signs of T3 formation. This may suggest an early action or trigger in the leg disc to prevent initiation of leg development (Fig.3.6.5.B, E and F). Furthermore, the abnormal segmentation phenotype observed as disruption of the pigmented cuticle stripes seemed more dramatic in the flies with some abdominal segments completely

missing and causing some flies abdominal segments to form irregularly and lose abdominal symmetry through fusion of tergites (Fig.3.6.5.B, F and G). One phenotype that was specific to the α-tub Gal4-2 driver was the misshapen ventral abdominal hairs that would also frequently be seen with abnormal dorsal stripes (Fig.3.6.5.G).

Table 3.6.3 summarizes the penetrance of each of the phenotypes from ectopic expression of *Hox-G* and G-PRE, individually separating mutations to show the penetrance of each, as many of these phenotypes would frequently be seen together in a variety of combinations. We can see that many of the phenotypes affect the T3 segment of the adult fly, as one of the most common and striking phenotypes is the missing halteres and T3 legs (Tables. 3.6.2 & 3.6.3). Particularly interesting is that ectopic overexpression of both *Hox-G* and G-PRE produces very similar phenotypes, possibly linking them to the regulation of each other or the same gene. Furthermore, ectopic overexpression using the PBac constructs, both adjacent and within *Hox-G* has produced adult flies with similar phenotypes as the ectopic overexpression from the pUAST experiments, suggesting that expression from both the endogenous locus, as well as other loci, has the same effect. Another frequent phenotype is necrosis of the legs from all experiments and there are a range of other phenotypes that recur in different experiments overexpressing *Hox-G* and G-PRE, giving no direct link to a single Hox genes function for either of these sequences.

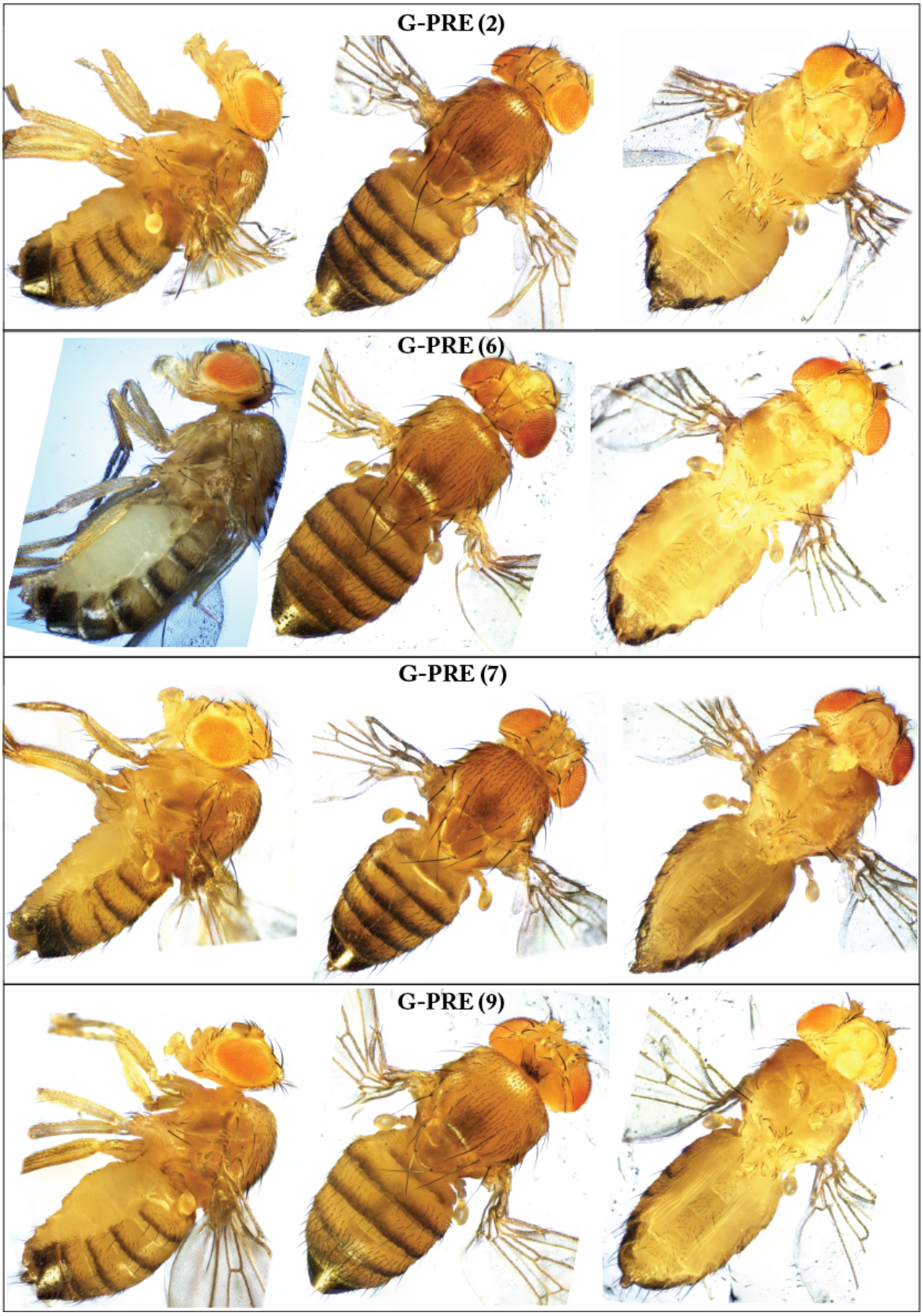


Figure 3.6.3. Homozygous fly lines generated by P-element insertion of the experimental PRE upstream of *Hox-G* (G-PRE). Side (left), dorsal (middle) and ventral (right) views of adult flies containing the homozygozed p(UAST) vector carrying the G-PRE sequence with line numbers in brackets.

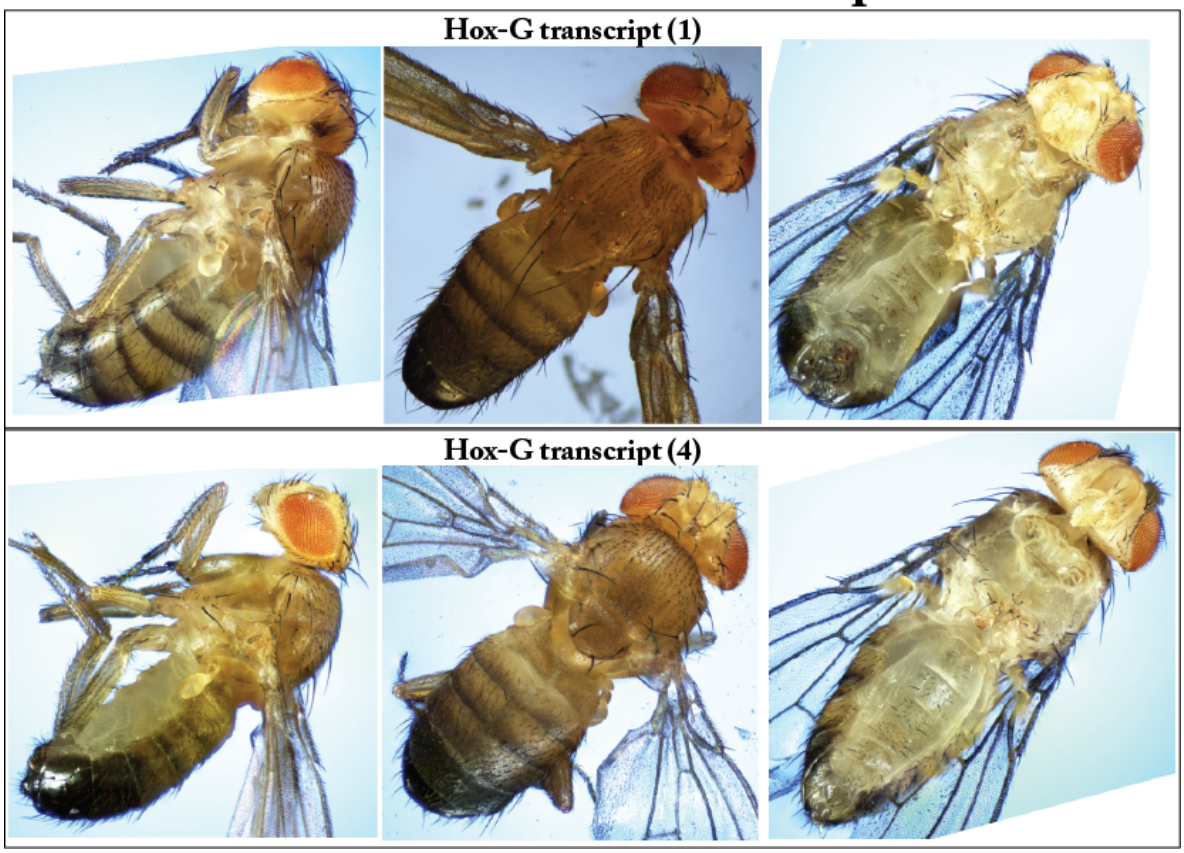


Figure 3.6.4. Homozygous fly lines generated by P-element insertion of full length *Hox-G* transcript. Two of the positive transformant lines that were injected with p(UAST) vector carrying the full-length transcript of *Hox-G* are shown. Transformant line 1, Hox-G-transcript(1), and line 4, Hox-G transcript(4), are the two lines that demonstrated various mutations when *Hox-G* was overexpressed in them.

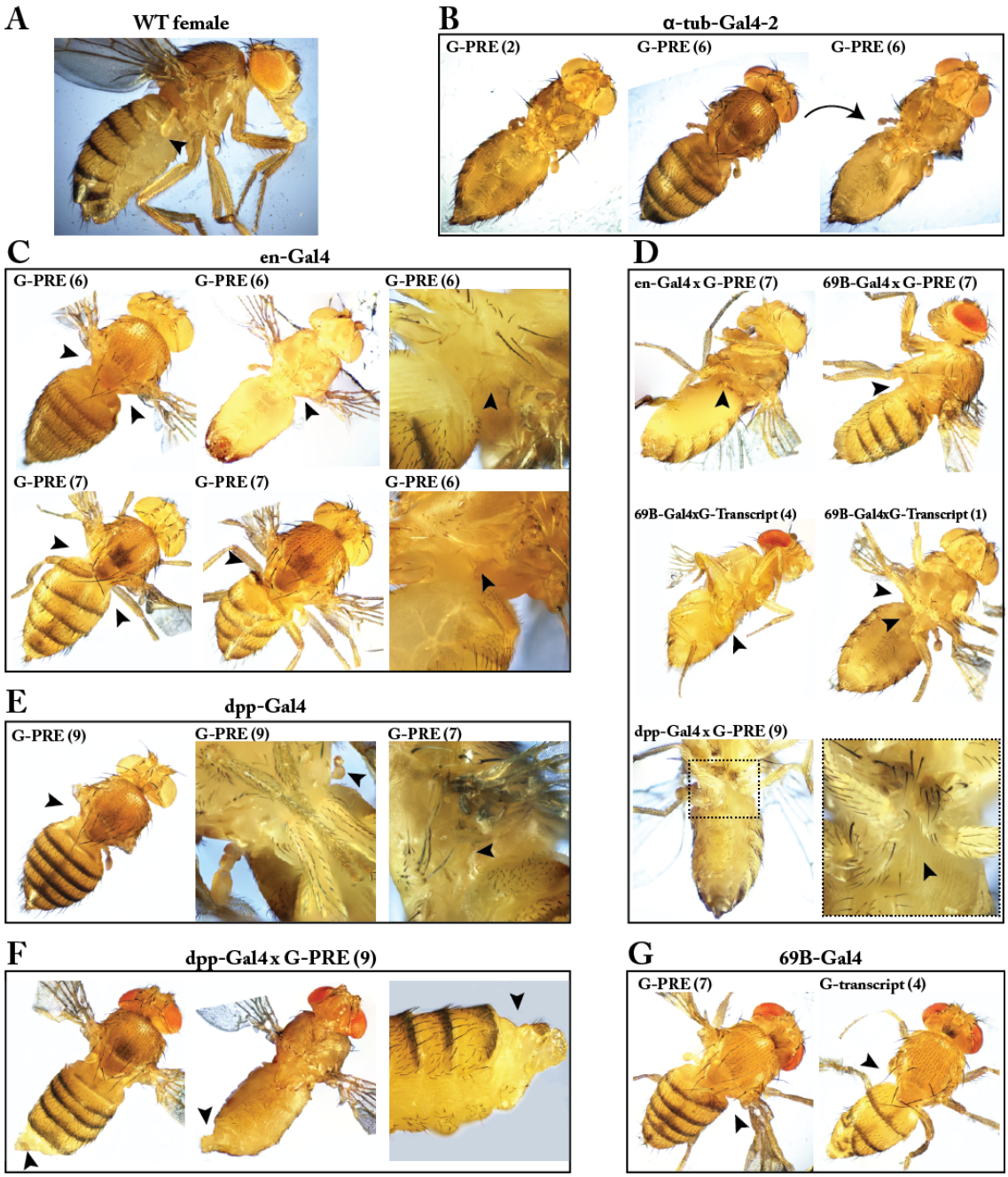


Figure 3.6.5. Phenotypes caused by Gal4 driven expression of *Hox-G* and G-PRE sequences. The *Hox-G* transcript and upstream PRE were overexpressed during embryogenesis using the Gal4-UAS system. A) WT female (Oregon R) side view to show T1, T2, T3 leg arrangements and halteres on the thorax and regular pigmented abdominal stripes. B) Ectopic expression of α-tub-Gal4-2 in two G-PRE lines causes sternal bristles to grow irregularly (G-PRE-2) and loss of abdominal segments combined with irregular sternal bristles abd blackening within the abdomen (G-PRE-6). C) Flies overexpressing G-PRE(7) from the *engrailed*-Gal4 promoter have lost one or both halteres (black arrows) and have either a normally formed abdomen, or show a partial or full loss of abdominal segment 1 (A1). D) Demonstrates various phenotypes under different Gal4 drivers, occurring in overexpression of both G-PRE and *Hox-G*. Frequently missing one T3 on the same side as a missing haltere (69B-Gal4 x G-transcript-1) or occasionally both T3, both halteres and A1 stripe (69B-Gal4 x G-PRE-7). E) Flies have partial or full loss of halteres under the *decapentaplegic*-Gal4 driven expression of G-PRE. F) Same fly rotated shows loss of posterior stripe, possibly A6, and the genitalia has angled to the fly's left (black arrows). G) Flies have dramatically altered abdomens under the 69B-Gal4 driven expression, missing A1 (G-PRE-7) or have mutated abdominal segments (G-transcript-4). These flies are also missing halteres and T3 legs.

Observed Mutations / Fig	control G-PRE(2) n=100	control G-PRE(6) n=100	control G-PRE(7) n=100	control G-PRE(9) n=100	control G-transcript(1) n=100	control G-transcript(4) n=100	en-Gal4 x G-PRE(6)	en-Gal4 x G-PRE(7)	en-Gal4 G-transcript(1)	69B-Gal4 xG-PRE(2)	69B-Gal4 x G-PRE(7)	69B-Gal4 x G-PRE(9)	69B-Gal4 G-transcript(4)	69B-Gal4 G-transcript(1)	dpp-Gal4 x G-PRE(9)	dpp-Gal4 x G-PRE(7)	α-tub Gal4-2 x PRE(2)
Missing 1 haltere - 3.6.5-G	0	0	0	0	0	0	9% n=187	12% n=139	0	0	9% n=69	0	11% n=94	16% n=102	3% n=64	5% n=65	1% n=185
Reduced haltere - 3.6.5-E	0	0	0	0	0	0	0	0	0	0	0	3% n=64	0	0	0	0	0
Missing both halteres - 3.6.5-C	0	0	0	0	0	0	6% n=187	0	0	0	4% n=75	0	0	0	11% n=64	3% n=65	0
Missing 1 x T3 - 3.6.5-D	0	0	0	0	0	0	0	0	0	0	0	0	9% n=94	8% n=87	4% n=71	0	2% n=52
Missing 2 x T3 - 3.6.5-D	0	0	0	0	0	0	0	1% n=139	0	0	0	0	0	0	0	0	0
Overgrown T3 - 3.6.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9% n=93	0	11% n=150
Necrotic legs - 3.6.11	0	0	0	0	0	0	31% n=81	0	13% n=62	11% n=66	0	0	0	20% n=102	0	14% n=65	4% n=111
Abnormal abdominal stripes - 3.6.5-C &D	3%	2%	0	0	0	0	8% n=142	0	0	0	7% n=71	0	19% n=94	0	0	0	5% n=185
Abnormal sternal bristles - 3.6.5-B	0	0	0	0	0	0	NE	NE	NE	0	0	0	0	0	NE	0	11% n=266
Collapsed abdomen <3days from hatching - 3.6.10-D	0	0	0	0	0	0	NE	NE	NE	0	0	0	0	0	NE	NE	26% n=90
Black growths - 3.6.8-E	0	0	0	0	0	0	NE	NE	NE	0	0	0	0	0	NE	NE	NE
Black growth head - 3.6.8-D	0	0	0	0	0	0	NE	NE	NE	0	0	0	0	0	NE	NE	NE

Table 3.6.3. Penetrance scores of Gal4 *Hox-G* transcript and G-PRE overexpression. Controls show mutations found in 100 homozygous p(UAST) transformants. G-PRE lines are highlighted in pink and *Hox-G* transcripts are left white. The separate Gal4 drivers are shown together in groups. The percent of penetration for individual mutations is calculated with number of flies recorded underneath. The Gal4 driver lines have been grouped and the G-PRE lines highlighted in pink and *Hox-G* transcript left white for easier comparison. Figure references displayed for each phenotype. NE = not evaluated.

Given the striking and specific nature of the phenotypes been seen for Gal4 driven expression of *Hox-G* and for the associated PRE, we decided to investigate functions of *Hox-G* further using the CRISPR/Cas9 system. The CRISPR/Cas9 system allows vector integration or manipulation of a target locus at its endogenous loci, therefore ruling out any positional effects from random insertion or manipulation within the genome. This allowed us to study *Hox-G* by making a double strand break and screening for homologous recombination of a donor plasmid using mini white as a reporter. The donor plasmid had many features such as loxP, attP and FRT sites that would allow us to further investigate *Hox-G's* DNA and RNA. We identified an integration site in the second exon that had a suitable targetable motif to design guide RNA necessary for specific cutting of the double stranded DNA by the Cas9 enzyme (Fig.3.6.6-A). A second plasmid, pTVCherry (Baena-Lopez et al., 2013) was modified to contain 1.5 kb homology arms extending in both 3' and 5' directions from the Cas9 cut site to facilitate homologous recombination of this donor plasmid. Flies expressing the Cas9 enzyme were then injected with a combination of Cas9 specific interacting RNA containing guide RNA sequence and SceI enzyme for linearization of the other plasmid to be integrated, pTVCherry (Baena-Lopez et al., 2013). The flies were screened for mini-white and homozygozed. The pTVCherry containing the mini-white

gene was now inserted within *Hox-G's* second exon (termed G-pTVCherry) with no apparent alteration to the fly's morphology. However, at 25°C the flies eyes had slight variable patches of darker and lighter pigmentation from the mini-white reporter, termed variegation (Kassis, 2002). PREs are frequently linked to variegation of mini-white and another form of silencing called pairing sensitive silencing (PSS) that are dependent on genomic position and the regulatory DNA in the surrounding environment (Kassis, 2002). Variegation has also been shown to be affected by changes in temperature, particularly the PcG gene *E(z)*(Chan, 1994), therefore, we tested development at 18°C and 29°C to find out if this would affect levels of variegation. Interestingly, variegation is found to typically occur in heterozygotic transgenes and flies showing PSS have lighter eyes in homozygotes than heterozygotes (Chan, 1994). Therefore, we balanced the G-pTVCherry over the TM3 balancer to investigate the effects of heterozygosity.

Figure 3.6.6-A shows an overview of the CRISPR/Cas9 strategy used to integrate the pTVCherry plasmid into the *Hox-G* locus, generating an allele we termed G-pTVCherry. We imaged the w^{1118} flies used for injection that expressed the Cas9 enzyme under control of the *vasa* promoter, to allow comparison to the homozygous and heterozygous G-pTVCherry eye colors produced by flies raised at different temperatures. Homozygous G-pTVCherry flies raised at 29°C show much stronger effects of variegation and PSS in 100% of the flies (Fig.3.6.6.B). Flies raised at 18°C have much darker red eyes that still show a mosaic of very dark red and slightly lighter red, again in 100% of the flies, suggesting there is still some variegation at the lower temperature but PSS can not be detected. When the G-pTVCherry allele is moved over a balancer (TM3), the variegation is lost and instead the flies have a uniform light orange eye color at all temperatures that is not noticeably darker than the homozygotes (25°C is shown in Fig.3.6.6-B). This would suggest that the mini-white is able to act as a reporter for variegation and PSS at the endogenous *Hox-G* locus and strongly supports that the *Hox-G* locus is at or near a temperature sensitive PRE.

We then decided to utilize other components of the pTVCherry plasmid to further investigate effects of altering the wild-type state of the *Hox-G* transcript. We began by cutting the loxP sites by introducing the *Cre* protein. The effect of the *Cre* enzyme on loxP sites is orientation specific. The pTVCherry loxP sites are both in the same relative orientation, therefore causing the DNA between the 2 sites to be excised as a circular loop. The *Cre* enzyme makes 2 double strand breaks and rejoins the DNA, removing mini-white and allowing for screening of successful excision based on eye color. When repairing the DSB between the 2 loxP sites, the 3' end is degraded whilst the adjacent 5' end is extended, using the sister chromatid as a template, usually creating an intermediate Holliday junction before homologous recombination occurs (Voziyanov et al., 1999). However, it is not clear exactly what happens in the Cre-loxP system when the sister chromatid is a balancer chromosome and therefore suppresses homologous recombination, although there is some evidence of rare TM3 crossing over (Crown et al., 2014).

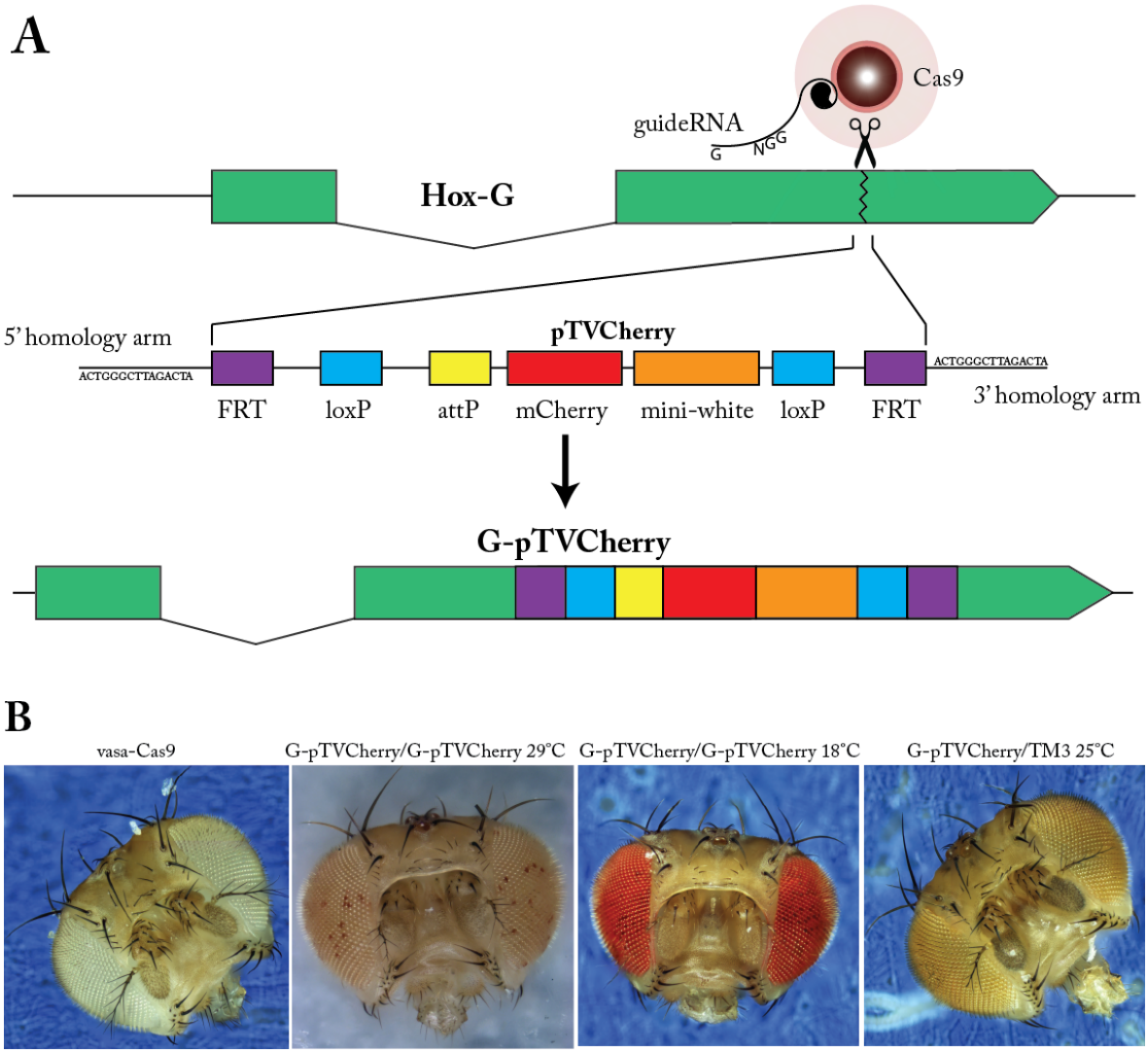


Figure 3.6.6. CRISPR-Cas9 generated mini-white expression as reporter gene within *Hox-G*. Flies were generated using the CRISPR-Cas9 system to insert pTVCherry (Baena-Lopez et al., 2013) into the second exon of *Hox-G* (A). White eyed flies expressing the Cas9 enzyme (B) were injected and the mini-white gene was used as a marker to screen for orange eye transformants and homozygozed. Flies developed at 29°C show a combination of highly variegated and PSS expression of mini-white in 100% of offspring and those developed at 18°C have dark red eyes, showing higher expression of mini-white and no silencing affects (B). When the G-pTVCherry was moved over a balancer chromosome, the variegation was lost and mini-white was expressed in what would be considered a normal expression of a transgene carrying mini-white (B).

Figure 3.6.7. Adult images of pTVCherry insert into *Hox-G* (G-pTVCherry) and heat shock Cre flies. Adult flies imaged from the side (left), dorsal (middle) and ventral (right – legs removed). Homozygous and balanced G-pTVCherry flies, generated by CRISPR/Cas9 system, are shown, along with flies used in the Cre experiment, Bloomington stock number 851.

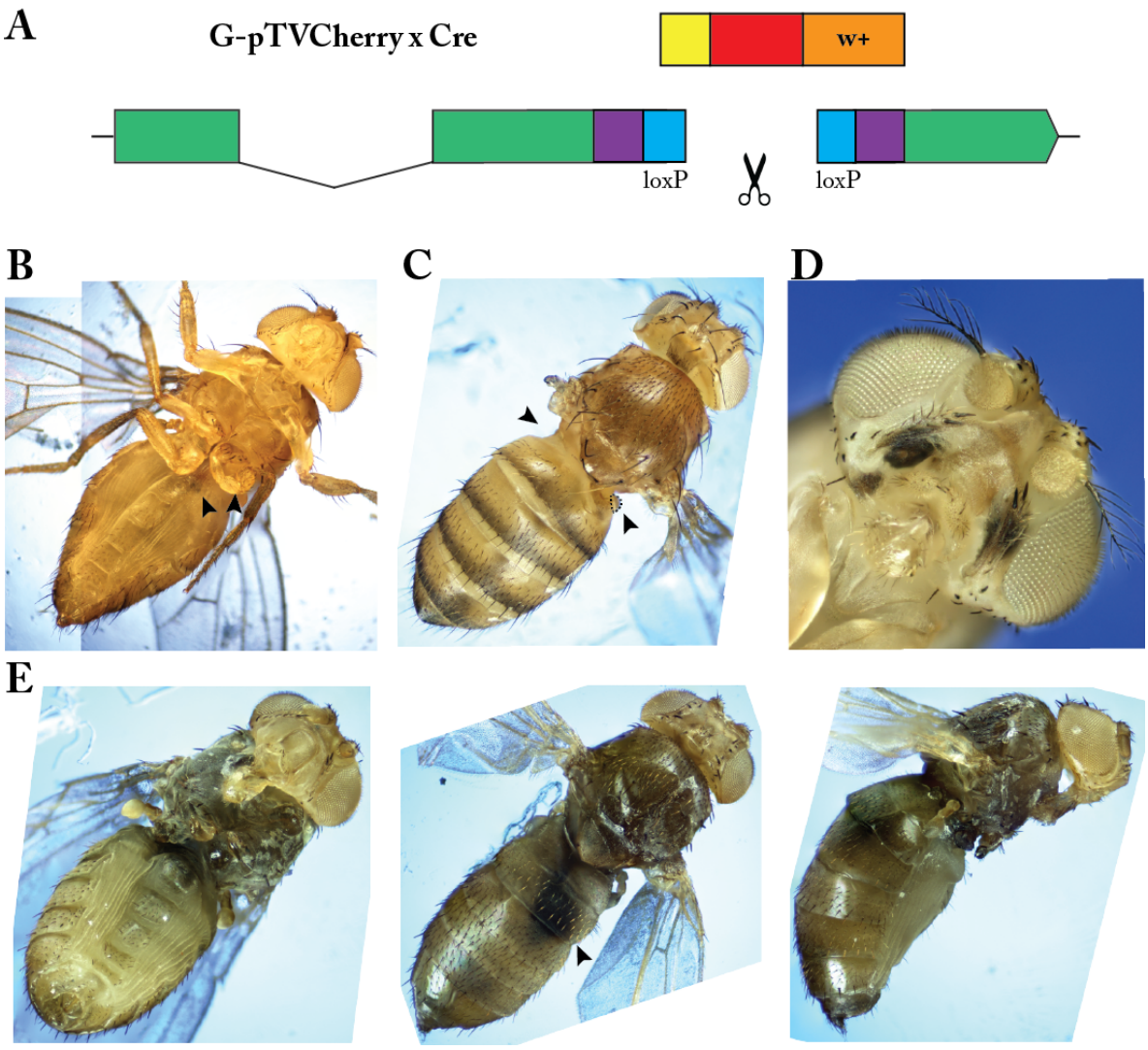


Figure 3.6.8. Homeotic phenotypes from Cre excised G-pTVCherry. The pTVCherry insert into the second exon of *Hox-G* (G-pTVCherry) contained loxP sites that were used to excise most of the vector from the genomic DNA (A). Cre excised offspring were selectable by mini-white removal (white eyes) leaving 2x34bp FRT sites in the genome. Similar phenotypes as seen in the Gal4 driven UAS-G-PRE and UAS-*Hox-G*-transcript were found in the offspring of the Cre excised G-pTVCherry. B shows a female missing a T3 leg with a misshapen T2 leg on the left side with a black lump formed in the abdomen in place of the leg (black arrows). C shows a female with a left missing haltere and the right haltere (black dotted lines) formed from the dorsal of the abdomen, rather than the thorax. D shows a fly head with black forming either side of the mouth.

We then took advantage of the loxP sites that had been integrated with the pTVCherry vector to see if excision could cause any visible phenotypes through imprecise DNA repair at the cut sites. We found that excision of the DNA between the 2 loxP sites did indeed cause multiple mutant phenotypes in *w-* offspring, corresponding with phenotypes seen in the Gal4 driven overexpression experiments (Fig.3.6.8.B-E) and speculate that this was due to adverse DNA changes at the *Hox-G* locus as the mini-white reporter was removed. These specific similarities included missing halteres, black/necrotic patches on the head (Fig.3.6.8-C), abnormal abdominal stripes (Table.3.6.4), necrotic legs (Fig.3.6.11) and missing T3 legs, many with black growths identified within their abdomen that resembled underdeveloped legs (Fig.3.6.8-B). New phenotypes that also arose were the supernumerary tarsal segments growing on T3 legs (Fig.3.6.11) and a rare large black growth found in the abdomen (Fig.3.6.8-E).

We then utilized the FRT sites available in the PBac(WH)2 and PBac(WH)3 lines to duplicate or remove part of the second exon of *Hox-G* (Fig.3.6.10-B). In this scheme the with PBac(WH)2 line was crossed to PBac(WH)3 line in order have one of each on sister chromatids (Fig.3.6.10-A), before crossing them to a fly expressing the FLP enzyme. This can cause recombination between the 2 sister chromatids and when the DNA is replicated and cells are divided, can lead to cells having either a duplication or deletion of the second half of the second exon of *Hox-G* (Fig.3.6.10-A). The offspring will then inherit one of the recombined chromosomes and this can be tracked based on eye color variations, as all copies of mini-white in this case would segregate with the deletion allele and none with the allele that carried the duplication. None of the offspring were found to have orange eyes, indicating the deletion was lethal. However, pale *ry+* (from the FLP construct) hatched and had similar mutation phenotypes previously identified in Gal4 and Cre experiments (Fig.3.6.10.D-G). These included underdeveloped and missing T3 legs, overdeveloped T2 legs and abdomens collapsing in less than 3 days after hatching, with penetrance summarized along with the Cre-loxP flies (Fig.3.6.10.D-G and Table.3.6.4). Examples of various leg phenotypes are shown in Figure 3.6.11 to compare similarities and differences throughout different experiments altering the expression of *Hox-G*.

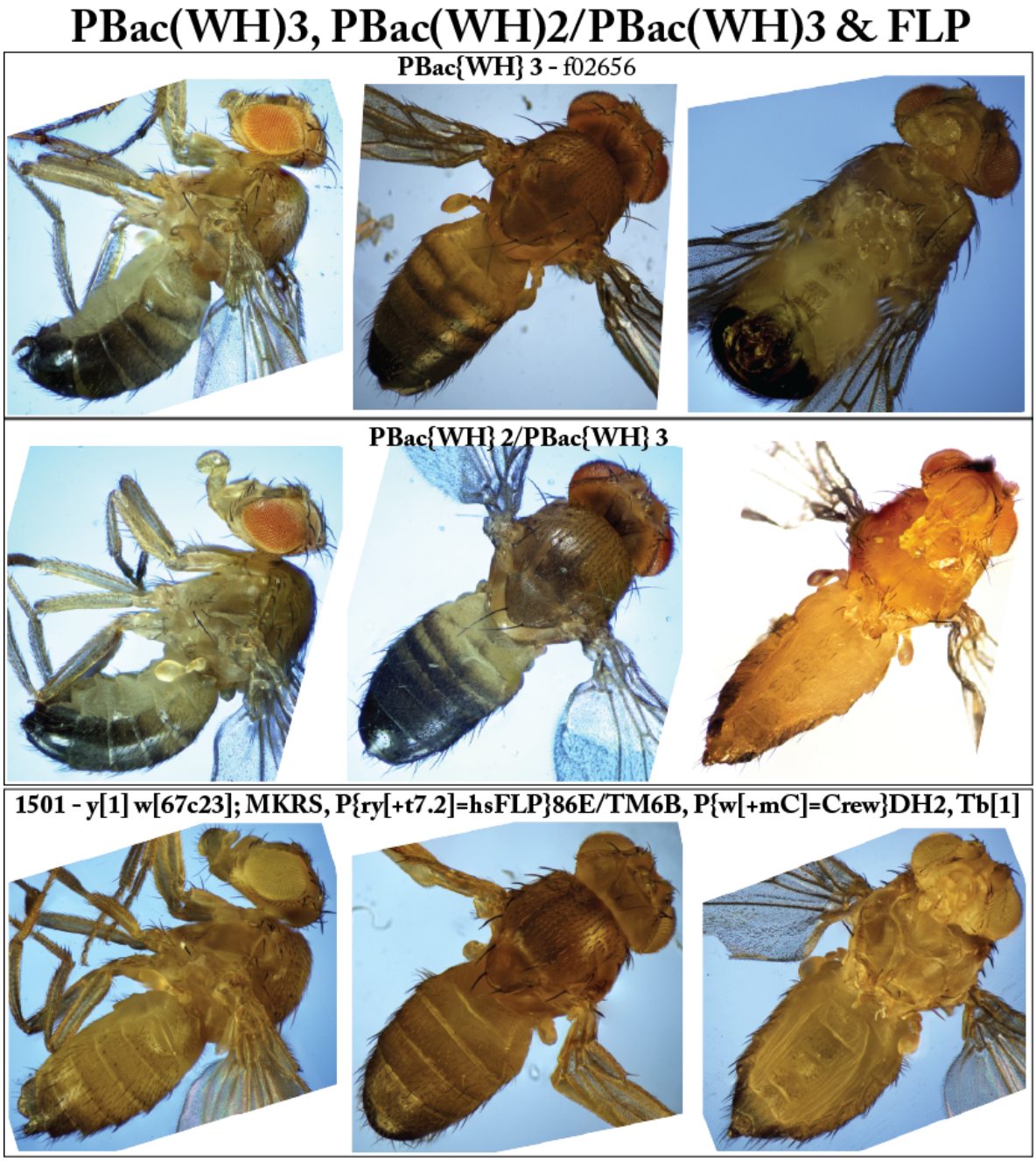


Figure 3.6.9. Adult images of PBac(WH)3, PBac(WH)2/PBac(WH)3 and Flippase flies used to generate partial duplication and deletion of *Hox-G*. Adult flies used in the FLP experiments to partially duplicate or delete the second exon of *Hox-G* are shown from the side (left), dorsal (middle) and ventral (right – legs removed). The Pbac(WH)2 flies can be seen in Figure 3.6.1

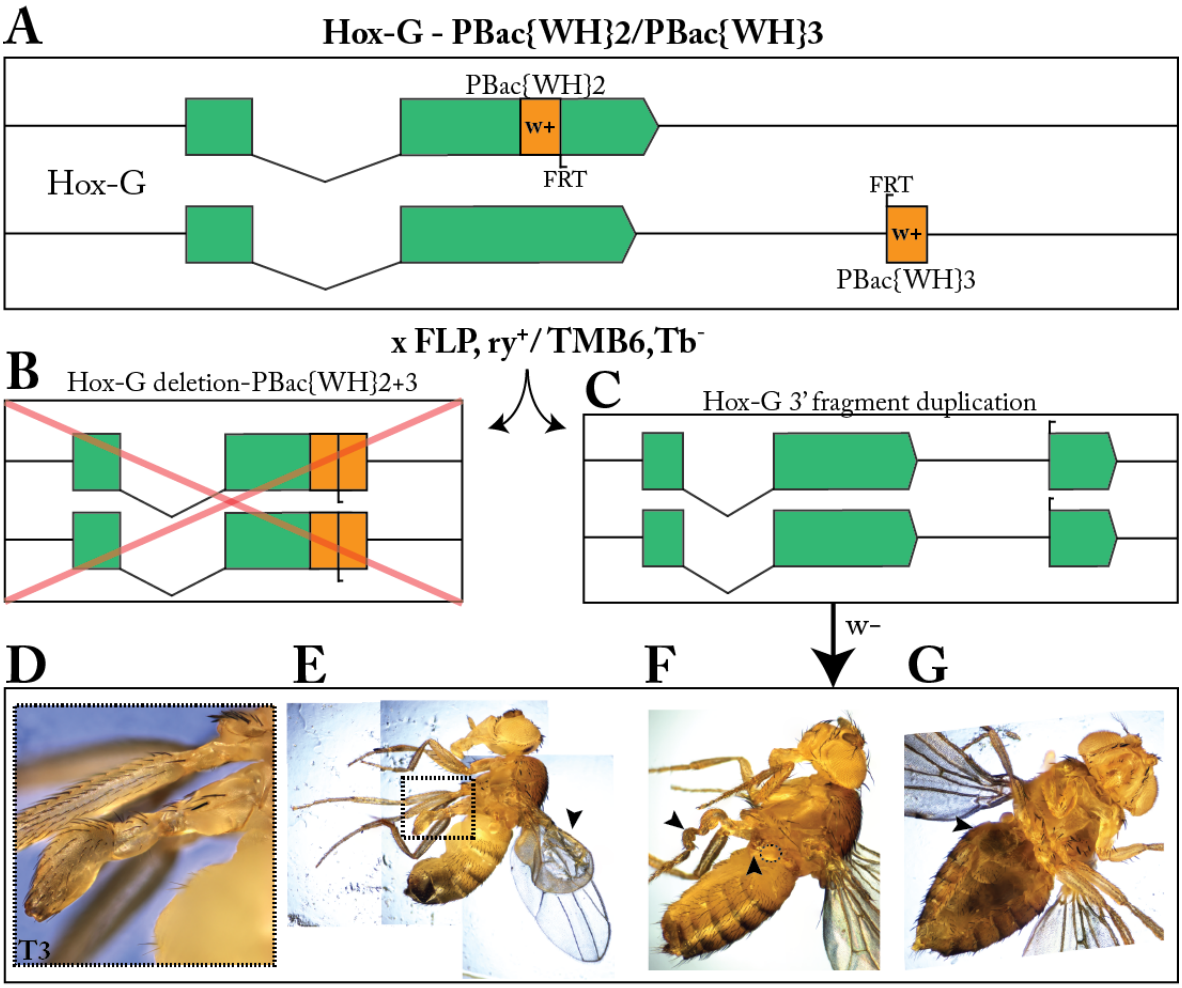


Figure 3.6.10. Homeotic mutations arising from flippase mediated duplication and deletion of 3' *Hox-G* fragment. PBac insertions on sister chromatids carrying FRT sites (A) were used to generate a partial deletion or partial duplication of *Hox-G,* via Flippase mediated uneven homologous recombination (B and C). T3 leg of a male fly is malformed (D), along with the wing of the same fly (E-black arrows). F) Female missing a T3 leg and T2 is malformed (black arrows). The haltere is circled with a dotted line. G) Female missing T3 with a black mass in the abdomen (black arrow). Also, the abdomen of G is collapsed, something noticed in ~50% of the flies <3 days old.

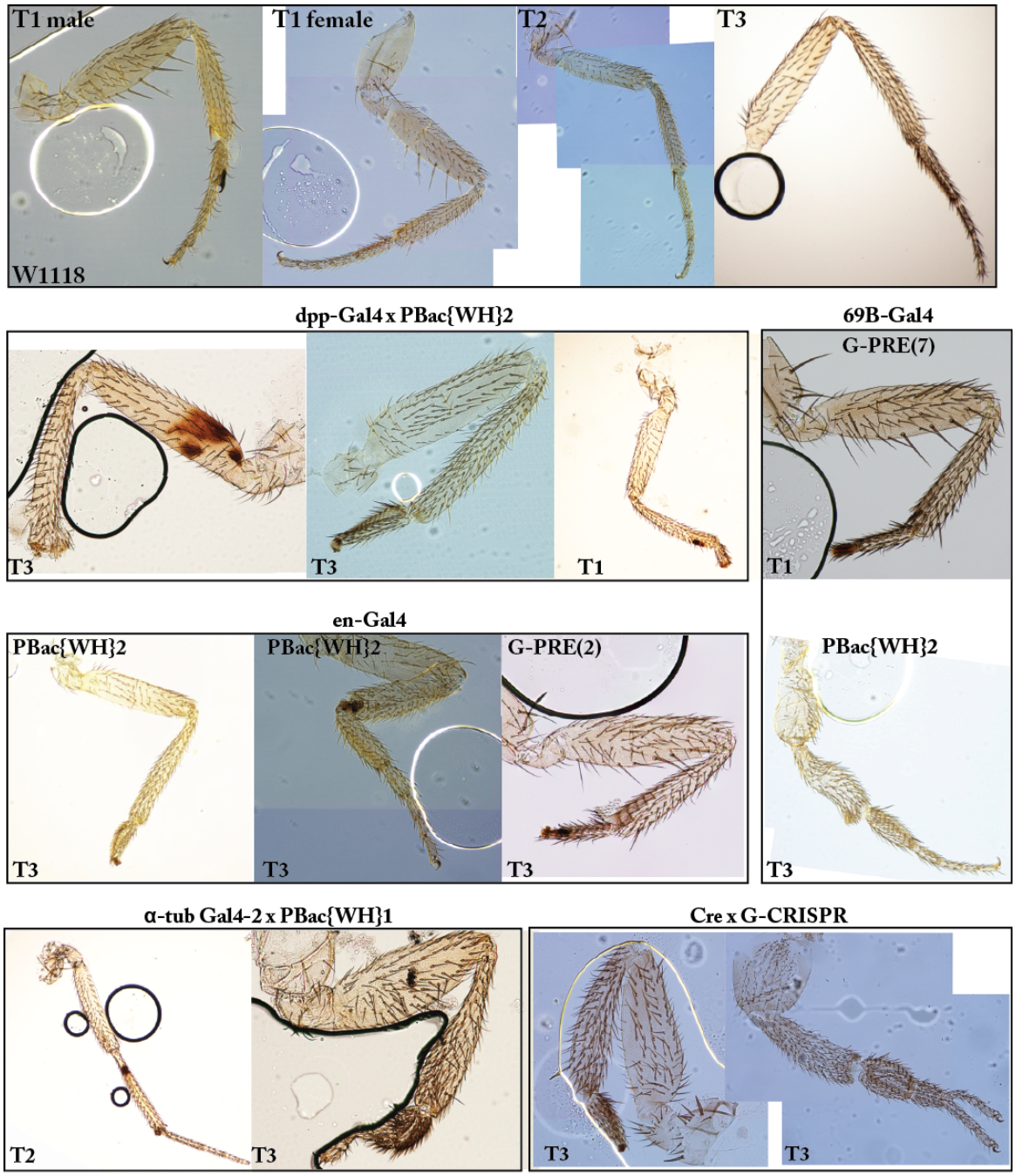


Figure 3.6.11. Necrotic and mutated legs from Gal-4-UAS and Cre experiments. Various leg phenotypes from different experiments. WT T1 from both male (sex comb) and female are shown with healthy T2 and T3 legs. Various examples of necrotic black marks, under and over developed legs, along with possible transformation into antennae and supernumerary formation.

Table 3.6.4. Penetrance scores mutations caused by Cre excised G-pTVCherry and Flippase mediated partial duplication of *Hox-G*. All mutations found in previous experiments were assessed in offspring of Cre removal of loxP sites from the G-pTVCherry construct and FLP induced duplication of the second half of the second exon of *Hox-G*. The percent of penetrance was calculated for each individual mutation and numbers of offspring recorded indicated below.

Observed Mutations / Fig	Missing 1 haltere - 3.6.5-G	Reduced haltere - 3.6.5-E	Missing both halteres - 3.6.5-C	Missing 1 x T3 - 3.6.5-D	Missing 2 x T3 - 3.6.5-D	Overgrown T3 - 3.6.11	Necrotic legs - 3.6.11	Abnormal abdominal stripes - 3.6.5-C &D	Abnormal sternal bristles - 3.6.5-B	Collapsed abdomen <3days from hatching - 3.6.10-D	Black growths - 3.6.8-E	Black growth head - 3.6.8-D	Supernumerary leg- 3.6.11
Cre x G-pTVCherry	15% n-74	0	0	7% n-74	0	0	3% n-185	18% n-74	0	0	0	5% n-74	3% n-185
FLP x Pbac [WH]2-3	0	0	0	6% n-119	0	0	18% n-119	0	0	48% n-119	11% n-27	0	0

Table 3.6.5. Sequenced pUAST insertion sites and associated genes. iPCR was used to identify coordinates of inserted vectors carrying UAS binding sites for Gal4 driven expression of *Hox-G* or G-PRE. If a single gene is stated then the insertion is within that gene, if two genes are separated by dashes, then the insertion is intergenic between the two.

Fly Line	Position *D. mel* r6.12	Closest or associated gene(s)	Description of associated gene functions
Hox-G (3)	chrX:19,851,125	*pico*	Pico is an intracellular adapter protein belonging to the MRL family of proteins, which transduce signals from growth factor receptors to changes in the actin cytoskeleton. Pico roles include the regulation of growth and cell migration (FlyBase)
Hox-G (4)	chr3R:27,581,145	*widerborst*	Required for planar cell polarization for wing hair orientation (Hannus et al., 2002)
Hox-G (5)	chr2R:7,489,940	*wech --- Coop*	*wech* - Plays a role in tumor formation. Crucial component for the physical link between integrins and the cytoskeleton in the epidermal muscle attachment sites. (Uniprot) *Coop* - corepressor of *Pangolin* and antagonizes *Wg* signaling. (Song et al., 2010)
G-PRE (6)	chr2R:6,170,835	*Ecdysone receptor—Cyp6w1*	*Ecdysone Receptor* - Receptor for ecdysone. Binds to ecdysone response elements (ECRES) following ecdysone binding, and recruitment of a complex containing the histone methyltransferase trr, leads to activate transcription of target genes. (UniProt) *Cyp6w1* – May be involved in the metabolism of insect hormones and in the breakdown of synthetic insecticides. (UniProt)
G-PRE (7)	chr3L:22,739,722	*lethal(3)04053*	Gene function is unknown
G-PRE (9)	chr3L:2,417,777	*CG45186*	Predicted to organize cytoskeleton (FlyBase curators 2004), no phenotypes reported

Insertion sites were determined for some of the lines generated by P-element transformation of the pUAST vector carrying UAS binding sites to drive expression of *Hox-G* and G-PRE (Table.3.6.5). The genomic coordinates were identified and we have reported the gene that the vector either inserted into or adjacent genes if intergenic. For the genes that are adjacent or have had the vector inserted within, we have given a description of the genes function. The description of the protein-coding genes function is to ascertain if the phenotypes may be attributed to the vector insertion or if driving expression of *Hox-G* or G-PRE may have caused disruption to the genes function. Interestingly, when identifying any regulatory elements identified at any of the insertion sites, the 3 *Hox-G* sites were directly in a TF binding site of Trl based on ChIP-ChIP

(Negre et al., 2011) and predictions (mod et al., 2010). This could be a coincidence, but is noted as P-elements containing PRE and enhancer/promoter sequences for *en* (Cheng et al., 2012; Hama et al., 1990; Kassis et al., 1992) and the BX-C (Bender and Hudson, 2000). The gene *pico* has been linked to wing disc mutations, leading to larger wings when overexpressed in the wing pouch (Lyulcheva et al., 2008). A different stock of *en*-Gal4 that has the same insertion site and expression to the stock we used (Neufeld et al., 1998; Weiss et al., 2001), was used by Lyulcheva *et al* (2008) and they found a larger posterior compartment growth of wing pouch and overall body growth. However, Lyulcheva *et al* (2008) made no mention of halteres, legs or abdomen mutations and in that case we do not believe our phenotypes are connected to disruption of that gene. The *widerborst* insertion site had no features mapped to this particular locus (Feature mapper in FlyBase) and there was no reported evidence for leg or haltere phenotypes when searching FlyBase or using Google to search. The *widerborst* gene could only be found linked to wing development in genetic screens (Molnar et al., 2012; Molnar et al., 2006) and therefore any perturbations to *widerborst* seems unlikely to be responsible for the phenotypes generated by our investigations. *Cyp6w1* belongs to the cytochrome P450 family and is expressed in appendages, highest in antennal segment 3 than legs and low levels in head and body (Wang et al., 1999). We also cannot find any reports to suggest it affects the halteres, legs or abdomen using FlyBase and Google searches and therefore have no reason to believe the inserted vector has caused any *Cyp6w1* disruptions that would explain our results. *Ecdysone receptor* is linked to the regulation of development of many tissues and organs throughout embryogenesis, larval and pupal stages and has been linked to mutations of several organs and tissues, including wings (Nijhout et al., 2014). Therefore, disruption of the *Ecdysone receptor* could theoretically be linked to some of the mutant phenotypes we have generated. However, this seems unlikely as the mutations produced from the vector inserted adjacent to the *Ecdysone receptor,* G-PRE(6), produced mutant phenotypes affecting halteres, legs and abdomen when ectopically expressed (Table.3.6.3). The mutations generated from ectopic expression of G-PRE(6) were similar to those seen for ectopic expression of a number of other insertion sites, whereas mutations currently reported for the *Ecdysone receptor* include mutations we did not generate.

4. DISCUSSION

4.1 Key outcomes

The aim of this project was to better understand the regulation of Hox genes by the transcription of regulatory DNA and to investigate if the transcription was functional. To do this we identified a novel regulatory region consisting of a multi-exon lncRNA with a previously unidentified PRE/TRE in the adjacent upstream region. We obtained evidence that the PRE/TRE had a silencing effect on the lncRNA in specific cells, at the endogenous loci, by using CRISPR to introduce a mini white reporter. We investigated ectopic overexpression of both the lncRNA and the PRE/TRE in order to distinguish if the lncRNA transcript or RNA from the PRE/TRE were functional by introducing many copies to the whole developing embryo and assessing the adults for visible phenotypes. This led to phenotypes that could be linked to misregulation of Hox genes, such as missing halteres, missing T3 legs, supernumerary growths on legs and abnormal abdominal stripes. Therefore, this indicates that the RNA transcript has a function in Hox gene regulation and based on the phenotypes, the Hox gene being affected seems to be *Ubx* or *Antp*. Understanding the regulation of these genes will aid in our understanding of how they are able to control key developmental activities that can lead to severe developmental defects if not properly regulated. This lncRNA and the adjacent PRE/TRE had strong affects on the development of *D. melanogaster* embryos indicating this region and the transcribed lncRNA is critical for healthy development. This knowledge will further aid in our understanding of the fine tuning of Hox gene regulation and how lncRNAs function, along with their importance during development.

4.2 Identification of lncRNA enriched clusters

There has been significant controversy over the numbers of functional lncRNAs predicted from RNA sequencing data, as many transcripts are believed to be transcriptional noise (Struhl, 2007). Low sequence conservation and transcription of lncRNA transcripts can be used to support an argument suggesting that many observed transcripts have no function (Mattick and Makunin, 2006; Wang et al., 2004b; Young et al., 2012). However, they are highly abundant with the number of mRNA loci in humans calculated at 20,944 and lncRNA loci at 40,765 (Pertea, 2012) and many now have identified functions. Many lncRNAs also demonstrate specific tissue and subcellular localizations (Dinger et al., 2008a), particular temporal expression (Carninci et al., 2005), have conserved promoters, are alternatively spliced and demonstrate and open chromatin structure at their promoters for transcription and regulations by TFs (Rinn and Chang, 2012). The

various regulatory processes involved in lncRNA transcription are used to argue that many lncRNAs are likely to have functional roles (Mercer et al., 2009; Morris and Mattick, 2014).

The balance of evidence would suggest that the noncoding transcriptome of most higher eukaryotes is likely to be composed of a number of functional lncRNAs present in a much larger population of nonfunctional or spurious RNA transcription. Therefore, to identify lncRNAs that were most likely to be functional, we used features such as conservation, clustering and features that indicate precise regulation. Evidence from many animal models suggests that these features are a useful way to narrow down to functional lncRNAs (Amoutzias and Van de Peer, 2008; Kung et al., 2013; Sproul et al., 2005; Spurlock et al., 2015; Wang et al., 2011a). A typical definition of a cluster is physical clustering, where a group of two or more genes that have similar function are proximal on a chromosome (Medema et al., 2015). Gene distribution can also alter in organisms, depending on how compact the genome is, therefore altering the density of genes and the numbers of nucleotides used to separate one cluster from the next (Hurst et al., 2004). This makes cluster identification usually quite specific to each study, with some arguing there is no way to have a single definition of a cluster (Jain, 1988). The intercluster distance, also known as the linkage function, is one of the main differences between different studies (D'Haeseleer, 2005) and can be defined in a number of ways, but usually requires prior knowledge of clustering in the genome.

Using a bespoke algorithm we identified lncRNA clusters in *D. melanogaster* by first determining intercluster distances between lncRNAs ranging from 100 kb to 10 kb. We used 5 kb intervals to test different intercluster distances until the clusters that were identified matched visibly compact stretches of lncRNAs with spaces between them. Figure 3.1.1 shows the clusters when a 100 kb cutoff is used to separate them and this grouped most of the genome into a small number of large clusters. Also shown is a 25 kb cutoff that was empirically determined to be the most appropriate intercluster distance where a larger number of discrete clusters are apparent. We investigated the top 20 most highly enriched lncRNA clusters, meaning the cluster contained the largest number of lncRNAs. As conservation of lncRNAs is limited we explored the use of conservation of syntenic lncRNAs in clusters as a feature to identify conservation. We performed the same cluster analysis on *D. virilis* to determine if there was evidence of conservation of regions that are the most enriched for lncRNAs. However, lncRNAs in *D. virilis* have not been well annotated so we first identified an appropriate developmental stage to expand and annotate the repertoire of lncRNAs for comparison using knowledge from the *D. melanogaster* clustering. Using GO term analysis on the protein-coding genes found in the 20 clusters from *D. melanogaster*, we identified stages 4-6 as likely to be enriched in lncRNAs overall. Many question the validity of a GO-term analysis as they have been found to have redundant terms describing the same thing and the descriptions given are not always meaningful (Gillis and Pavlidis, 2013), or may be incomplete or biased depending on the research carried out (Thomas et al., 2012). Nevertheless, GO analysis as a general guide is still a commonly used tool for annotating functions of lists of genes, and there

are few obvious alternatives available. The Gene Ontology Consortium has daily updates that reflect up-to-date literature. However, many GO-term analysis tools are not updated as frequently with the current gene annotations, including the widely used DAVID, which had not been updated for 5 years at the time of this analysis (Huang da et al., 2009). We therefore chose to use PANTHER as it is updated monthly with current GO-terms and has demonstrated accuracy and comprehensiveness on the *D. melanogaster* genome (Mi et al., 2016; Mi et al., 2003).

Our results show that the regions containing the highest numbers of clustered lncRNAs have a tendency to be those containing protein-coding genes linked to development. This is not particularly surprising as most functionally characterized lncRNAs have roles in development and their misexpression is often linked to cell proliferation that can lead to tumor progression in cancers (Fatima et al., 2015). However, analysis of the GO terms and the clusters allowed us to narrow down the stages of embryogenesis that were likely to have the most actively transcribed lncRNAs. When investigating the most enriched GO-Slim terms, 'pattern specification process' and 'segment specification' stands out, assigned to the 7th highest cluster (25-15), which covers most of the ANT-C. This gained our attention as Hox genes have been strongly linked to lncRNAs previously in flies and mammals (Mallo and Alonso, 2013). Other GO-Slim terms with >50 fold enrichment within this cluster included 'digestive tract mesoderm development', 'embryo development', 'spermatogenesis' and 'female gamete generation', demonstrating the wide range of the few protein-coding genes in this complex. LncRNAs have previously been suggested to play critical roles in coordinating the wide range of specific regulatory functions carried out by Hox genes (Dasen, 2013). Therefore, our findings that the lncRNA cluster at the ANT-C also contains the highest ratio of lncRNAs (19) (Fig.3.1.1) to protein-coding genes (12) (Fig.3.1.5) with the highest number of significantly overrepresented GO-Slim terms (19 in total) (Fig.3.1.2) would suggest that at least a proportion of these transcripts should be functional.

Cluster 25-18 also gained our attention as it contains several GO-terms linked to development, most notably >100 fold enrichment for 'muscle organ development'. This also directed us to Bownes stage 6 embryogenesis (Campos-Ortega and Hartenstein, 1997) as this is when the initiation of muscles begins, derived from mesoderm progenitor cells (Furlong et al., 2001). Another two well-known complexes were identified in the top 20 lncRNA clusters of both *D. melanogaster* and *D. virilis*, the (E[spl]-C) (25-19) and the Histone complex (25-12). The E[spl]-C are all TFs that most likely evolved by duplication and have gene inhibitory roles in neurogenesis in the same genetic pathways as *Notch* (Lai et al., 2000). E[spl]-C gene expression is detected earlier in the bearded family members (BFM), from stage 4-6, than the HLH members that are detected from stage 7-8 (Knust et al., 1987; Wech et al., 1999; Weiszmann et al., 2009). Interestingly, the E[spl]-C is also known to have regulatory input from the PcG proteins and therefore there is a possibility that some of the lncRNAs within the complex could be functioning with PcG or TrxG complexes to direct chromatin states (Schaaf et al., 2013). The large number of

lncRNAs in the Histone complex may be related to the extreme levels of recombination and pseudogenation that has been described there, which potentially generates lncRNAs from the pseudogenes of decaying Histone genes (Hurles, 2004; Sisu et al., 2014). This is fascinating as pseudogenes once also bore the label of 'junk DNA' along with lncRNAs, but have since demonstrated potential in neighboring gene regulation through interference and miRNA decoy with strong links to cancer progression, mirroring much of what is becoming known about lncRNAs (Pink et al., 2011). Therefore, these lncRNAs could harbor some interesting roles, such as that seen by the histone H2A/K pseudogene in humans that has been linked to cell proliferation (Guo et al., 2016). There are now ongoing investigations into possible functions of lncRNAs that have been derived from pseudogenes throughout the genome (Milligan and Lipovich, 2014).

Another GO-Slim term that was >100 fold enriched was 'regulation of sequence-specific DNA binding transcription factor activity' from the cluster 25-8, a fairly non-specific term as there are many DNA-binding TFs in the *D. melanogaster* genome. However, this cluster also has 'immune response' as a GO-Slim term with >50 fold enrichment and amongst other GO-Slim terms is 'immune system processes'. Therefore, it seems likely that a subset of the TF genes may be linked to immunity and share common functions. This could be during stage 5 of embryogenesis, as *D. melanogaster's* systemic immunity comes from specialized haemocytes that undergo the first phase of haematopoiesis during this stage. Furthermore, the second phase involves chromatin remodeling, a process now commonly linked to lncRNAs and interestingly the progenitors are maintained by *Collier, Serrate, Antp, and hh* (Crozatier and Meister, 2007). The cluster 25-3 also has two high scoring GO-Slim terms, 'regulation of liquid surface tension' (nearly 90 fold enrichment) and 'chromatin remodeling' (>50 fold enrichment). Both of these terms have strong links to development, as liquid surface tension is involved in cell fate determination, mechanical control of tissue and organ morphogenesis and patterning during development (Lecuit and Lenne, 2007; Mammoto and Ingber, 2010).

Many of the most confidently overrepresented GO-terms from *D. melanogaster* are linked to embryogenesis and the known protein-coding clusters, Hox, E[spl]-C and histone, which were identified are expressed from egg laying to stage 7. Also, the cluster with the strongest links to lncRNAs and PcG regulation is the Hox complex, which contains genes that all give distinct expression patterns between stages 4-6 (Weiszmann et al., 2009). We therefore determined that sequencing RNA from stage 4-6 *D. virilis* embryos would allow us to identify many novel lncRNA transcripts. Doing this allowed us to identify 542 novel lncRNAs to add to the 565 previously annotated lncRNAs (Table.3.1), doubling the number of annotated lncRNAs. The small number of previously annotated lncRNAs in *D. virilis* were mainly identified using prediction algorithms such as Gnomon by the FlyBase Consortium (personal communications), which mostly rely on sequence similarity and only multiexonic lncRNA transcripts were previously included. Therefore, we separated lncRNAs into multi-exon and single exon and intergenic or antisense as it was useful

for quality control as the genome of *D. virilis* is fragmented in ~13,500 scaffolds with a lot of low quality sequence (Carvalho and Clark, 2013). This fragmented assembly can lead to identification of two genes at ends of scaffolds that are in fact one and so the most reliable lncRNAs are those well within the scaffold boundaries. We also noticed that the Hox gene, *Scr*, has a massive run of N's in the DNA sequence leaving the *Scr* gene in FlyBase annotated as just 595nts, whereas its *D. melanogaster* counterpart is nearly 27 kb. This difference in size is highly unlikely to be real given how well conserved Hox genes are and that other homologous Hox genes in *D. virilis* are quite similar in size and exon-intron structure. These poor quality regions could potentially generate several lncRNAs as reads that belong to *Scr* may be fragmented in to smaller transfrags that do not have the characteristics of protein coding genes. Nevertheless, although identification of some clusters will be disrupted due to poor genome assembly, there are several large scaffolds that should allow us to identify some lncRNA enriched clusters to compare to *D. melanogaster*. With this in mind, we carefully, manually annotated the Hox complex for further analysis, allowing us to identify likely 'false' lncRNAs that would otherwise be missed by automated identification.

The 100 kb intercluster distance used in *D. virilis* lncRNA cluster analysis was empirically determined to approximate the number and size of the clusters identified in *D. melanogaster*. We were satisfied that the 100 kb cutoff in *D. virilis* was a suitable match to the 25 kb cutoff used for *D. melanogaster* as numbers of lncRNAs in the top 20 highest clusters were within a range that overlapped in both species (Fig.3.1.4). Most of the top 20 clusters occurred on the largest *D. virilis* scaffolds, which is unsurprising as the smaller scaffolds are more prone to breaks and low quality sequence hindering cluster identification. The orthologs of *D. melanogaster* protein-coding genes were then matched to the protein-coding genes found in *D. virilis* lncRNA clusters and compared to each other for any matching single genes, color coded by *D. melanogaster* clusters. This revealed that the E[spl]-C and ANT-C are directly adjacent in the *D. virilis* genome and are part of the cluster with the highest number of lncRNAs (Fig.3.1.5) (highlighted in pink and green). This method of identifying lncRNA clusters in Drosophila genomes has revealed some interesting conserved arrangements of regions of protein-coding genes that have high numbers of lncRNAs throughout 40-60My evolutionary divergence, revealing a novel method of identifying possible homologs of lncRNAs worthy of further investigation.

4.3 Identification of lncRNAs in Drosophila Hox complex and transcribed PREs

Of particular interest are the E[spl]-C and ANT-C clusters, which are both enriched for lncRNAs and are associated with regulation by PcG proteins. These clusters provide a possible connection between the many studies finding that lncRNAs are often involved in chromatin regulation. Furthermore, PcG regulated genes are often clustered, a process that is conserved from plants to humans (Bantignies and Cavalli, 2011; Rosa et al., 2013; Shen et al., 2012). Based on

previous evidence of functional lncRNAs in the Hox complex, we manually annotated all lncRNAs in both the ANT-C and BX-C in *D. melanogaster* noting the 2hr windows they were expressed and their locations relative to other genes in the Hox complex (Table.3.2.1). We saw a wide variety of temporal expression of lncRNAs, with the majority of the expression beginning in the 4-6hr window after egg laying (AEL).

The candidate list was narrowed down to lncRNAs that were most likely to regulate Hox genes based on position in the complex, as those that act in *cis* to affect Hox genes are likely to be in physically adjacent regions on the chromosomes based on previous evidence (Bertani et al., 2011; Guil and Esteller, 2012; Quinn and Chang, 2016; Wang et al., 2011b). The lncRNAs that are closer to other protein coding genes in the Hox complex, such as *Amalgam*, were excluded from further analysis. In the ANT-C, our lab has previously investigated *lincX* and *TipX* (Pettini, 2012) and from our screen two more, upstream of *Antp,* were retained for further analysis as they were expressed at reasonable levels in the 4-6hr window and had not previously been identified as miRNA primary transcripts (Fig.3.2.2). To further characterize the lncRNA genomic structure we used CAGE datasets carried out in 2hr windows AEL to identify TSSs. This led to an interesting observation that there were large differences in timing of the accumulation of reads from the CAGE analysis and observed production of transcripts seen in the NGS transcriptome analysis for all of the lncRNAs in the ANT-C, but not the BX-C. Previously, TSSs have been seen transcribed prior to a genes transcription in cell lines in response to stimuli. This study investigated immediate-early genes (IEGs) and found ncRNA transcription was initiated before the IEGs (Aitken et al., 2015). These IEGs have properties frequently associated with lncRNA dysregulation, namely that they are involved in differentiation and proliferation and often become constitutively expressed in cancers due (Aitken et al., 2015; Quinn and Chang, 2016). Furthermore, chromatin architecture is also thought to play a key role in IEG expression and mature mRNAs of IEGs have been seen to peak 3hrs later than the pre-mRNA (Tullai et al., 2007). A GO-term analysis in the study by Aitken *et al* (2015) of the IEGs revealed those that the overrepresented GO-terms included: regulation of gene expression, regulation of transcription from RNA polymerase II promoter, regulation of RNA metabolic processes and regulation of metabolic processes, terms all found associated with the ANT-C (Fig.3.1.2). This study also investigated if gene length contributed to the 3hr delay in transcription from the TSS through to the rest of the gene and found that short genes underwent the same delay before reaching their transcriptional peak. Furthermore, the same study found that IEGs were associated with promoter-proximal pausing and that this included lncRNAs such as *NEAT1* and *MALAT1*, along with other ncRNAs. It is therefore likely that the lncRNAs in the ANT-C are transcribed very early, particularly *Hox*-G, which could indicate that they are functioning for up to 4hrs before being polyadenylated and therefore detected by these particular RNA-seq experiments.

In the BX-C of *D. melanogaster* there are three well-documented lncRNAs, *bxd, iab-8* and *iab-4*. We also identified *Hox-O*, another antisense transcript to *iab-8* and *iab-7* PRE, transcribed antisense to the second exon of *iab-8*. There are two known miRNAs transcribed from the same region of *iab-8* on opposite strands, one from the *iab-8* transcript, *mir-iab-8* and the other from the 3' end of *iab-4, mir-iab-4* (Tyler et al., 2008) (Fig.3.2.3). Three other lncRNAs have been identified from RNA-seq in the intergenic region between *bxd* and *abd-A*, all of which show low levels of expression, (small grey patches in Fig.3.2.2-A). The locus encompassing the second exon of *iab-8* that is transcribed from both strands has been previously identified as a PRE, *iab-7* PRE, recognized for maintenance of silencing of *iab-7* the regulatory domain in parasegments that are anterior to PS12 (Hagstrom et al., 1997). Its bidirectional transcription has not been previously reported and this could help explain a number of previous observations from a study that indicated transcription through the *iab-7* PRE could interfere with PcG repression in an orientation dependent manner (Hogga and Karch, 2002). It would be interesting to investigate if the *iab-7* PRE was acting as a bidirectional switch or if the transcript has another function. *Hox-O* is also particularly interesting as it is transcribed antisense to *iab-8*, just adjacent to *iab-7* PRE, is spliced into 2 exons and is expressed in the 4-6hr window of transcription in both RNA-seq and CAGE data.

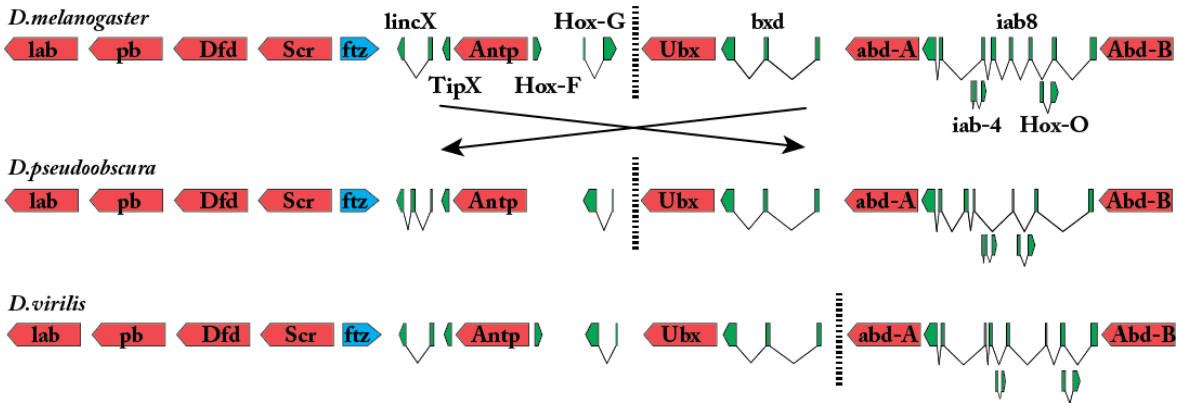


Figure 4.2.1. Hox complex alignment of protein coding genes and lncRNAs in 3 different Drosophila species. Hox genes are shown as red block arrows and the segmental gene *ftz* as a blue arrow. The lncRNAs are in green with the intron-exon structure depicted. The black arrows indicate opposite orientation on the chromosome and black dashed lines show the different break points of the complex in each species.

The syntenic conservation of lncRNAs was investigated in *D. virilis* and *D. pseudoobscura*. Several from the list of lncRNAs in *D. melanogaster* could be identified in both other species based on position, with very similar intron-exon structures for the majority (Fig.4.2.1). This was a similar case for those that we could detect in *D. pseudoobscura* RNA-seq, however, many could not be detected due to low sequencing depth from these runs (data not shown). The RNA-seq evidence shows that these lncRNAs have been conserved for over ~63 million years of evolution that would typically indicate that they may possess conserved functions. To further examine conserved and

divergent aspects of the lncRNAs we used a ntFISH approach to visualize the lncRNAs in developing embryos. We visualized the lncRNAs transcript expression and adjacent protein coding or lncRNA genes, in order to investigate differences or similarities in expression. Interestingly, all Hox genes and lncRNAs are visible in a stage 5 embryo, a stage that matches 130-180 minutes of development and therefore aligns to the 2-4hr RNA-seq window (Table 4.1). This matches the findings from the CAGE data for the ANT-C lncRNAs. However, in the BX-C both RNA-seq and CAGE demonstrate peaks of expression in the 4-6hr window and no signs of transcription can be seen prior to this in these datasets. This would suggest that in fact all of these lncRNAs we can detect with ntFISH are transcribed from around 2hrs into embryogenesis, although an explanation of why this is not demonstrated in the CAGE analysis of BX-C remains elusive.

Bownes Stage #	1	2	3	4	5	6	7	8	9
Time period (mins)	0-15	15-70	70-90	90-130	130-180	180-195	195-200	200-230	230-260
Events	CC 1 - 2 pronuclear fusion	CC 3-8 early cell division	CC 9 pole bud formation	CC 10-13 syncytial blastoderm	cellularization of blastoderm	- gastrulation forms mesoderm and endoderm - pole cells shift dorsally	-gastrulation completes -pole cells form pocket -germband elongation	-rapid germ band elongation -mesodermal parasegmentation	-slow germ band elongation -cephalic furrow formation
Sequencing windows	0-2hr				2-4hr				

Bownes Stage #	10	11	12	13	14	15	16	17
Time period (mins)	260-320	320-440	440-580	580-620	620-680	680-800	800-900	900-hatch 1st instar larva
Events	-features of head form -stomodeum invaginates	-formation of segmentation and tracheal pits -midgut reaches posterior pole	-germband shortening -anterior & posterior midgut fusion -ventral closure -optic lobe invaginates	-end germband retraction -CNS & PNS differentiation	-dorsal closure midgut & epidermis -head involution begins	-dorsal closure finishes -dorsal epidermal segmentation -discs invaginate	-advanced denticles are visible -shortening of ventral nerve cord	-organs formed -trachael tree fills with air -ventral cord retracts
Sequencing windows	4-6hr	6-8hr	8-10hr		10-12hr	12-14hr	14-16hr	16-18hr 18-20hr

Table 4.1.1 Alignment of RNA-seq time points and *D. melanogaster* embryogenesis stages. The 2 hour sequencing windows are shown with respect to embryogenesis stages, the timing of stages and the events occurring in each stage. CC= Cleavage Cycle. Created using information from www.sdbonline.org and Bownes stages from (Campos-Ortega and Hartenstein, 1997).

The lncRNAs that are directly adjacent to each other, *lincX* and *TipX*, are expressed in distinct patterns, overlapping some domains of *Scr* and bordering *Antp*, which also corresponds to the collinear arrangement of all genes along the chromosome (Fig.3.3.1). This is similar for the expression patterns of *iab-7* PRE, *iab-4*, *Hox-O* and *bxd*, which are all expressed further posterior on the embryo in the order they are arranged on the chromosome. The most interesting expression pattern is that of *Hox-G* as it is expressed in the same cells as the second promoter of *Antp* (*AntpP2*), rather than a unique pattern that is shown by other lncRNAs of the Hox complex. The *Hox-G* transcript is oriented in the opposite direction as *AntpP2,* is 105 kb away, and has one of the earliest peaks of CAGE reads in the 0-2hr window (Fig.3.2.2-D). This evidence would seem to suggest *Hox-G* could be involved in the *cis*-regulation of *AntpP2*. *Hox-G* may also be adjacent to or

overlap the *cis*-regulatory element controlling *AntpP2s* expression. There is also a possibility that *Hox-G* can interact with proteins or protein complexes, acting hundreds of kilobases away from its target whilst still tethered to chromatin, reflecting our knowledge of other lncRNAs such as *Xist/RepA* (Kung and Lee, 2013). Furthermore, this highlights the complexity of identifying what would seem like fairly simple temporal characteristics, and that several lines of evidence should be considered that may help build a picture of the kinetics of transcription. We also tested if these expression patterns were conserved in *D. virilis* and aligned the RNA-seq syntenic transcripts from both *D. pseudoobscura* and *D. virilis* to compare the orientation and intron-exon structures. This shows remarkable syntenic conservation of the arrangement and structures of these lncRNAs within the Hox complex, with just minor adjustments in orientation (*Hox-G*), exon number (*lincX* and *iab-8*) and distance relative to adjacent genes (*Hox-O* and *iab-4*). Furthermore, some of the syntenic lncRNAs could be detected in *D. virilis* and are expressed in almost identical patterns. The only exception was a slight anterior shift for *Hox-O*, which could reflect the move of the transcript position towards a more 5' location of the syntenic *iab-8* transcript and Hox complex overall. There could also be a change in the association of the *Hox-O* promoter with more posteriorly activating or anteriorly repressive regulatory elements in *D. virilis*. It is particularly interesting that the break in the complex is between Hox genes *Antp* and *Ubx* in *D. melanogaster* and *D. pseudoobscura*. However, when the split is between different genes in *D. virilis*, *Ubx* and *abd-A, bxd* still remains adjacent to *Ubx*, suggesting this is necessary for its regulation of *Ubx*.

Next, we investigated if PcG and/or TrxG proteins were binding to or near the lncRNAs in the Hox complex, as we already know that they maintain Hox genes expression, and may also be regulating lncRNAs (Mallo and Alonso, 2013). Furthermore, there is a substantial amount of evidence that lncRNAs are involved in gene regulation by associating with PcG and TrxG complexes (Mallo and Alonso, 2013), although binding to the DNA does not necessarily indicate this. The ChIP-seq profiles of both PcG (Fig.3.4.1) and TrxG (Fig.3.4.2) demonstrate interesting differences at each lncRNA locus. Single exon transcripts *TipX*, *Hox-F* and *Tre2* appear to be bound by nearly all of these proteins, indicating that these 3 are all PRE/TREs and it is likely that their transcription has a role in how they function as they are expressed in specific domains of the developing embryo. The multi-exon lncRNAs *lincX, Hox-G, iab-4* and *Hox-O* have no discernable peaks, suggesting they are not acting as PRE/TREs and may not in fact associate with these proteins or do so with the RNA transcript itself. Furthermore, HDAC proteins do not show clear binding to these multi-exon transcripts to indicate PREs (Fig.3.4.3) and PRE predictions using the jPREdictor program also failed to identify high scoring PREs within any of these transcripts (Fig.3.5.1 and 3.5.2).

Of particular interest is the lncRNA *bxd*. This lncRNA has been reported to have no function, as preventing transcription of *bxd* in early embryos did not lead to any phenotypic affects

(Pease et al., 2013). This led to the conclusion that the DNA underlying the lncRNA carries out the regulatory effects previously observed for *bxd* (Pease et al., 2013). In this study, the transcription of *Tre2* was also prevented without any phenotypic effects to the fly. This aligned with findings by Erokhin *et al* (2015), who found that transcription through this PRE/TRE does not displace PcG proteins or act as a switch to maintain gene expression in repressed states to an active state (Erokhin et al., 2015). Interestingly, our results show that syntenic transcription is often conserved, but in these regions PcG/TrxG no longer binds and PRE prediction does not score these loci high. This would suggest that the transcript may not be related to the PcG/TrxG function, although a syntenic and similarly expressed lncRNA has been maintained, which usually suggests some conserved function. *TipX* is a good example that demonstrates high sequence conservation throughout 27 insect species and a locus that ChIP-seq clearly demonstrates PcG and TrxG protein binding in *D. melanogaster*. However, *TipX* was not predicted to be a PRE by jPREdictor in *D. melanogaster*, but *TipX* in *D. pseudoobscura* gets a very low scoring PRE prediction (barely above the cutoff score) and then in *D. virilis* scores even higher (Fig.3.5.1). However, *D. pseudoobscura* ChIP-seq does not show PcG or TrxG proteins binding to the syntenic *TipX* transcript for those tested (Fig.3.5.3), which seems particularly strange as it scores higher with the jPREdictor prediction tool. It is a different case for the lncRNA *Hox-F*, that although was not detected by RNA-seq in *D. pseudoobscura*, a syntenic transcript could be identified in *D. virilis*, suggesting that it may exist in *D. pseudoobscura*, but we just did not detect it.

There are ChIP-seq peaks of PcG/TrxG protein binding at *Hox-F* in *D. melanogaster* and this corroborated by a peak at this locus with the jPREdictor. Then there is no evidence of PcG/TrxG proteins binding in this region in *D. pseudoobscura* and the PRE predictions no longer identify this region as a PRE, suggesting it does not function as a PRE in this species. Other putative PREs do show conservation in sequence, genomic position and PcG/TrxG protein binding, particularly at the PRE/TRE at the promoter regions of *Hox-G* and *Dfd*, *Tre2* and one within *Ubx*. Also, the *iab-7* PRE shows conserved protein binding at the same position in *D. pseudoobscura*. These observations reflect those of Hauenschild *et al* (2008), who found that PREs could be separated into two classes, those that have evolutionary constrained positions and those that do not. They found generally that PREs could evolve very rapidly through Drosophila genomes in motif composition, numbers of PREs and genomic positions. This study did not take into account the whether or not the PREs were transcribed, which could further subdivide the categories of PREs. However, this may not be very useful if the transcription is merely coincidental so would need further investigations into the relevance of transcription of PREs. This is particularly evident in our data as the syntenic transcripts can still be found in the same position relative to Hox genes whilst the PRE is no longer detected either computationally or experimentally. This could be an indication that the transcripts that originate from PREs may not function as part of the PRE as many remain transcribed from the same positions when there is no

detectable PRE in divergent species. This could also mean that those that are conserved in position and continue to be transcribed in different species could be a specific subset that may have use for the lncRNA in the context of the PRE. This is something our data does not support in the Hox complex. Further analysis throughout evolutionary divergent Drosophila genomes to identify PREs both computationally and experimentally combined with further RNA-seq throughout development could be used to investigate if the transcribed PREs are transcribed by chance due to the fairly frequent distribution of PRE/TREs across specific regions. This would help clarify the frequency of transcribed PREs that are syntenically conserved and if they are also transcribed throughout evolution or if most of the transcripts remain syntenically conserved whilst the DNA no longer shows signs of being classified as a PRE. An alternative scenario could be that PREs evolve so rapidly that requirement for transcription also changes and the transcription of the RNA when the site is not a PRE in other species is transcriptional noise.

By analyzing motifs within a region predicted to be a PRE, the core region, known as the minimal PRE, can be identified by a high-density region of motifs and is usually sufficient to carry out silencing (Dejardin and Cavalli, 2004; Okulski et al., 2011). The GTGT motif has been found in repeats in PREs and deletion of just one of these repeats has been shown to dramatically reduce silencing, but its role in silencing is not understood. However, this GTGT motif deletion has currently only been tested in the *vg* PRE by Okulski *et al* (2011) and we identified 31 GTGT repeats in at least three clusters of this 1581nt PRE. One GTGT cluster has multiple overlapping motifs surrounded by *Trl* and *zeste* binding sites around the core of the PRE sequence (blue and yellow bars Fig.3.5.5). *Trl* encodes GAF, which seems to function mostly in activation by promoting open chromatin conformation (Benjajati, 1997). However, GAF has also been identified as a transcriptional repressor (Mishra, 2003) and Zeste has been shown to bind to DNA and stimulate transcription from nearby promoters and long-range interactions able to bypass insulators (Kostyuchenko et al., 2009). Therefore, it seems possible that other factors, not yet identified, are involved in guiding the activity of PcG or TrxG proteins bound to PRE/TREs. Interesting similarities can be seen between the motif organization of these PREs, as *Hox-G* associated PRE has a region of GTGT repeats near the center, flanked by several Trl motifs, as has *Hox-F* and the *vg* PRE has Trl motifs adjacent on one side of the central GTGT repeat. The *iab-7* PRE has no GTGT repeats, but is relatively enriched in Zeste motifs for its size and *Hox-F* is enriched for both GTGT and Trl motifs when compared to the other PRE/TREs. However, *TipX* is quite devoid of known motifs for any of the proteins, although it clearly binds many members of PcG and TrxG proteins based on ChIP-seq profiles (Fig.3.4.1 and 3.4.2). This demonstrates that there are massive gaps in knowledge about PRE/TREs as the *iab-7* PRE has demonstrated silencing functions (Mishra et al., 2001), but appears not to require the silencing motif and is enriched for Zeste motifs, typically indicative of activation. There are some possible parallels between the motif structure within PREs, but these are vague and require better knowledge of

PRE/TRE functions to ascertain if there are specific subsets of PRE/TREs that function similarly based on motif compositions. Unfortunately just a handful of PREs have been characterized and recent work testing *in vivo* function has begun to highlight deficits in previous methods of investigation into their functions. Many have been investigated using random P-element insertion, but there have been studies that have shown that they can be quite sensitive to their genomic position and so this may not be reflecting their true function (Okulski et al., 2011). Also, the majority of the information about PRE/TREs comes from ChIP-seq or ChIP-ChIP experiments in a variety of specific cell types or in whole organisms, often over quite substantial time periods where the PRE could be differentially bound. A similar problem exists with much of the associated RNA-seq analysis. This results in situations where we do not know if the transcribed regions are in the same cells or stages as those bound by PcG/TrxG proteins. For example, we cannot differentiate if a PRE/TRE locus is bound by specific proteins during transcription or if this is just the case in the cells where that locus is silent, or vice versa. We are also currently likely to be missing information on how dynamic the changes are over shorter periods of time as much of the ChIP-seq is carried out in blocks of several hours when we know several developmental stages have occurred. We have provided evidence that there is no consistent pattern of conservation for lncRNA transcription being associated with PRE/TREs in the Hox complex. Thus a direct connection between PRE/TRE function and lncRNA transcription is not apparent, as lncRNAs are syntenically conserved in evolution whilst PRE/TRE positions move relative to the lncRNA. A deeper understanding of this will come from much more precise investigations, preferably at endogenous loci and including many more than the handful that have currently been examined.

4.4 Gain and loss of function of a novel lncRNA and adjacent PRE

We chose the lncRNA *Hox-G* from the Hox complex to investigate further and the PRE from its promoter region, which we termed G-PRE. We used P-element insertion of each sequence to instigate ectopic transcription using the Gal4-UAS system individually in developing embryos. We also used the CRISPR-Cas9 system to insert a mini-white reporter within the endogenous *Hox-G* loci and found quite dramatic variegation. We used the FLP-FRT and Cre-loxP technologies to manipulate the sequence at the endogenous *Hox-G* loci to test the effects. First we identified transgenic flies where PBac elements had been inserted upstream (PBac(WH)1), within (PBac(WH)2) and downstream (PBac(WH)3) of the *Hox-G* loci. These Pbac(WH) elements also included UAS binding sites for Gal4 driven expression and FRT sites for further manipulation. Ectopically overexpressing PBac(WH)1 and PBac(WH)2 produced flies that had homeotic phenotypes, as well as other mutations that were all linked to disrupted imaginal discs or abdominal development (Fig.3.6.2). Many of these phenotypes were recapitulated when both *Hox-G* and G-PRE were randomly inserted and expressed ubiquitously, resulting in slightly more extreme T3 segment phenotypes. For example, when the PBac elements were driven by α-tub

Gal4-2, the T3 leg was not everted but could be seen as a dark shadow that had formed within the thorax (Fig.3.6.2.B&E). However, there was no evidence of leg formation in those missing one or both T3 legs in the experiments that ectopically overexpressed *Hox-G* or G-PRE from the pUAST vector (Fig.3.6.5). Particularly interesting is the observation that the same phenotypes could be recovered when driving transcription of either the G-PRE or the *Hox-G* transcript, which could be an indication that one is regulating the other or that they both have very similar functions. Furthermore, the main differences in mutant phenotypes between different Gal4 drivers was a result of the maternal driver, α-tub Gal4-2, which caused stronger disruption of abdominal development than the other drivers (Fig.3.6.1 & Table 3.6.3). This could be due to it's strong ubiquitous expression or from being maternally deposited through to stage 16 of embryogenesis (Weiszmann et al., 2009). We also believe we duplicated a segment of the second exon of *Hox-G* based on eye color and literature on FLP recombination and the same phenotypes were observed (Fig.3.6.10 & Table 3.6.4). We also supervised 2 masters students, Margrete Langmyhr and Philippa Jackson, whilst they carried out FISH on Gal4 driven *Hox-G* and G-PRE crosses. They used the Gal-4 drivers that we had seen the most striking phenotypes with, to investigate if Hox genes expression domains were altered. Based on the chromosomal position and phenotypes they focused on Antp and Ubx expression patterns whilst *Hox-G* and G-PRE was being ectopically overexpressed, but did not detect a change in expression of the embryos they imaged (personal communication). Although these results were preliminary, there was no indication that the Hox genes tested were expressed outside their endogenous domains.

The most striking phenotype observed was missing halteres, frequently combined with missing T3 leg(s). This phenotype is exceedingly rare but has been observed in three other studies that we can find. Interestingly, one such study involved disruption of a TRE/PRE in the *Ubx* lncRNA *bxd*. Deletion of the *Tre2* fragment in a *bxd* intron and mutated binding sites for Trl and Pho caused halteres and T3 legs to no longer form (Kozma et al., 2008). They focused on a 185bp core PRE fragment containing a cluster of motifs for Pho and Trl and found deletion or mutating both motifs was necessary for higher penetrant mutations (27 and 24% respectively). They also replaced this PRE with others that had similar clustered motifs for Pho and Trl, including the *iab-7* PRE, and found that this did not cause mutant phenotypes. This PRE replacement suggested that PREs are interchangeable and therefore the order and numbers of motifs were not specifically structured. However, when they replaced *Tre2* with the closest human sequence containing both Pho and Trl motifs, it was not able to replace the *Tre2* function and therefore there was some information missing in the human fragment. This study also investigated Ubx expression in both embryo and larval tissue and found that it was not detectably misexpressed, but did note that the levels of Ubx seemed subtly increased. This core cluster of Pho and Trl binding motifs can be seen in Figure 3.5.5 on the center region of *Tre2* by blue and green blocks, also identifiable in *iab-7* PRE, *vg* PRE, *Hox-F* and G-PRE, somewhat near the center of these identified PREs.

The second study overexpressed *bifid* (*bi*) (previously *optomotor blind*) in the dorsal compartment of the haltere disc, which led to reduction or loss of halteres and mutations of *bi* resulted in overgrowth (Simon and Guerrero, 2015). This gene responds to different levels of Ubx, although is not a direct target of Ubx, but is upregulated by Dpp and Wg (Grimm and Pflugfelder, 1996). The *dpp* gene is directly down regulated by Ubx in the halteres and reduces the diffusion by repression of *sbb* and *dally* and increasing expression of *tkv*, the *dpp* receptor, affecting haltere size (de Navas et al., 2006). In the wing and haltere, *dpp* is activated by Hh and together their signaling targets are linked to cell proliferation and survival, with the difference in size between the two structures attributed to Ubx repression of many of their targets that otherwise lead to cell growth in the wing (Simon and Guerrero, 2015). Simon and Guerrero found *bi* was expressed in both wing and haltere discs, but was not repressed by Ubx. Instead, *bi's* function in the wing is thought to be the prevention of apoptosis, but in the haltere functions differently and limits growth by repressing the targets of Dpp and Hh involved in growth, such as *Dorsocross2* (*Doc2*), *knot* (*kn*) (AKA *collier*), *spalt major* (*salm*), *dally* and *dally-like* (*dlp*). This is similar to the mechanisms of Ubx, but Bi does not interfere with the functions of Ubx. Interestingly, a recent study investigated the *in vivo* interactome of Hox proteins and found Doc2 associated with Abd-B, Scr, Antp, Abd-A and Ubx; Kn associated with Scr, Ubx, Abd-A, and Antp; and Salm associated with Antp, Abd-B, Scr and Abd-A (Baeza et al., 2015). It is not yet clear exactly how these different complexes function, but indicates that there could be many regulatory networks that are not yet understood in appendage formation.

The third study that showed loss of halteres used the same *en*-Gal4 driver used in our investigations to drive expression of *Socs36E* and found this suppressed activities of the Janus Activated Kinase/Signal Transducers and Activators of Transcription (JAK/STAT) and Epidermal growth factor receptor (EGFR) signaling pathways (Callus and Mathey-Prevot, 2002). *Egfr* is well established in haltere development and if not down regulated in the haltere leads to haltere to wing transformations (Pallavi et al., 2006). The ligands that bind to activate the Egfr/Ras pathway are Vein, Spitz, Gurken, and Keren (Shilo, 2005). *Egfr* and *Vein* were identified as directly down regulated by Ubx by Pallavi *et al* (2006). These three studies demonstrate alternative methods of generating a fly that is missing halteres, with the main link between them being the gene *Ubx*, either by a direct regulation of the gene itself from the PRE or indirect interference with its interacting partners. For example, if the levels of *Ubx* were altered, as indicated by Kozma *et al* (2008), then this could lead to the downstream effects that have been seen for the genes regulated by Ubx, as they may respond differently to different levels of Ubx. Given the location of *Hox-G* and G-PRE in the Hox complex, and taking in to account that the majority of the phenotypes manifest in the T3 segment of the fly, it is possible that their overexpression could have altered expression levels of *Ubx*. However, endogenous *Hox-G* expression in stage 5 embryos matches *AntpP2* expression both spatially and temporally, when considering both RNA-seq and ntFISH data.

Although, all three genes demonstrate transcriptional activity when using ntFISH to image the transcription in stage 5 embryos (Fig.3.3.1) and intersect in the T3 segment. This segment corresponds to the anterior PS6 (Fig.1.3.1), so there is a possibility that either *Antp* or *Ubx* are being affected in these cells.

PS6 is marked by the 3rd stripe of *ftz* and overlaps the posterior half of T3 and the anterior half of A1 (Fig.1.3.1). This corresponds to the segments where we see the majority of the Gal4 driven *Hox-G* and G-PRE phenotypes. This is particularly remarkable considering that this is still the case when *Hox-G* or G-PRE is ubiquitously expressed by a maternal driver, the alphaTub67C promoter (Hacker and Perrimon, 1998), or the *engrailed* promoter, which aligns to the anterior half of each PS. Therefore, the Gal4 driven expression in PS6 would be expressed in T3, not A1 in blastoderm embryos (Weiss et al., 2001), matching the region of the majority of the phenotypes. Other Gal4 promoters used have a slightly more restricted expression: the Gal4-69B promoter shows expression in the ectoderm (outmost layer of embryo) in stages 9-17 (Staehling-Hampton et al., 1994a) and the haltere discs, wing discs, ventral thoracic disc and eye-antennal disc in 3rd instar larvae (Brand, 1997 – personal communication to FlyBase); and the Gal4-dpp expression is found in larval eye-antennal disc (Kim et al., 1996), morphogenetic furrow (Mukherjee et al., 2000), genital discs, salivary gland, midgut (Gorfinkiel et al., 1999), and the dorsal-ventral (Marquez et al., 2001) and anterior-posterior (Tomoyasu et al., 1998) compartment boundaries of the wing disc.

Interestingly, there are only one or two examples in our data that show mutations of the wing, whereas the majority seems unaffected when homeotic mutations can be seen. However, when the haltere was missing we also noticed the metathoracic spiracle, Sp2, was almost always also missing, leaving a smooth surface on the thorax where it would normally be (attempts to image this failed). Also, the T3 legs were frequently missing along with the haltere and in a few severe case, both halteres and T3 legs were gone. Alternatively to the legs missing, T3 and T2 legs could often look overgrown and twisted, have necrotic/weak black patches on the legs that would often cause breaks or have missing structures towards the distal leg where tibia and tarsal segments should have formed. In rare cases, supernumerary legs and a possible antennal like growth could be seen (Fig.3.6.11). Supernumerary legs, legs lacking distal features and overgrown/twisted legs have been seen when *wg* has been ectopically expressed (Diaz-Benjumea and Cohen, 1994; Wilder and Perrimon, 1995) and gain of function *hh* has led to overgrowth of several imaginal discs, including haltere duplications and supernumerary legs (Felsenfeld and Kennison, 1995), this genes targets include *wg* and *dpp*. Inactivation of *buttonhead* (*btd*) and *Sp1* reduces size of legs through reduction of *wg* and *dpp* transcription and ectopic expression of *btd* in the wing, eye or haltere discs (dorsal discs) leads to transformation into legs and antennae (ventral discs) by altering the expression of *en, wg* and *dpp* (Estella et al., 2003).

The other noticeable phenotype that was found both when the haltere was missing or on its own, was malformations of the abdominal stripes, predominantly affecting the A1 segment in *en*-Gal4 and 69B-Gal4 overexpression of G-PRE and *Hox-G* (Fig.3.6.2 and 3.6.5). Rather than a transformation of A1 to other body segments, A1 seems to simply have been deleted or not have formed at all; when only one side is affected the rest of the abdomen appears to collapse into the missing region. Occasionally other abdominal segments were affected in a similar manner, with crossing over of the tergites that could leave the abdomen bent in a similar manner where the abdominal segment seems partially deleted. This was most prominent in the Gal4 driven expression of the Pbac elements in or near *Hox-G* (Fig.3.6.2). This phenotype would seem to indicate either some disruption of histoblast or spiracle development (Bownes, 1976). However, as misexpression with multiple early and ubiquitous drivers show almost solely late phenotypic effects disrupted segmentation is a not likely. The other interesting abdominal phenotype shows the dorsal abdomen collapsing in on itself within 3 days of hatching, similar to old flies near death. We cannot find any similar genetic phenotypes in the literature, however, desiccation does result in a similar appearance. Overall this could be due to a lack of understanding the underlying cause and we would speculate that maybe the gut or pleura has not formed properly as these flies died within 3-4 days of hatching. This phenotype was seen in nearly half of the flies that we think had a putative partial duplication of the second exon of *Hox-G* (based on eye color) and in 26% of a G-PRE line driven by the maternal alpha-tubulin driver and then a lower percentage (<10%) of the alpha-tubulin and 69B driven PBac constructs (Tables 3.6.2, 3.6.3, 3.6.4 and Fig.3.6.10).

Together these mutations can link imaginal primordia, the discs that eventually form the wings, legs, halteres, thorax, labial, genitals and eyes and the histoblast nests that form the abdominal epidermis, spiracles, gut and salivary glands (Beira and Paro, 2016; Kopp et al., 1999). Most imaginal disc primordia are specified as a cluster of cells at specific positions in a blastoderm embryo. They then invaginate as an epithelial layer from the ectoderm, or in the case of the genital disc, will recruit cells from the mesoderm (Beira and Paro, 2016). These founder cells were known as polyclones as they consist of cells from multiple origins, rather than a single cell (Crick and Lawrence, 1975; Wieschaus and Gehring, 1976) but do not have a determined lineage in blastoderm embryos (Vincent and O'Farrell, 1992). Early gene networks sequentially pattern the A-P body plan of early embryos, instigated by maternally deposited genes that lead to segmentation and Hox gene activation (Peel et al., 2005). This results in unique transcriptional regulation of common targets such as *wg, hh,* and *dpp* in each segment (Fig.4.3.1) (Morata, 2001; Scott and Carroll, 1987), which in turn leads to imaginal disc patterning as they require many of the same input genes for A-P specification and formation of compartments for appendage development (Martinez-Arias and Lawrence, 1985).

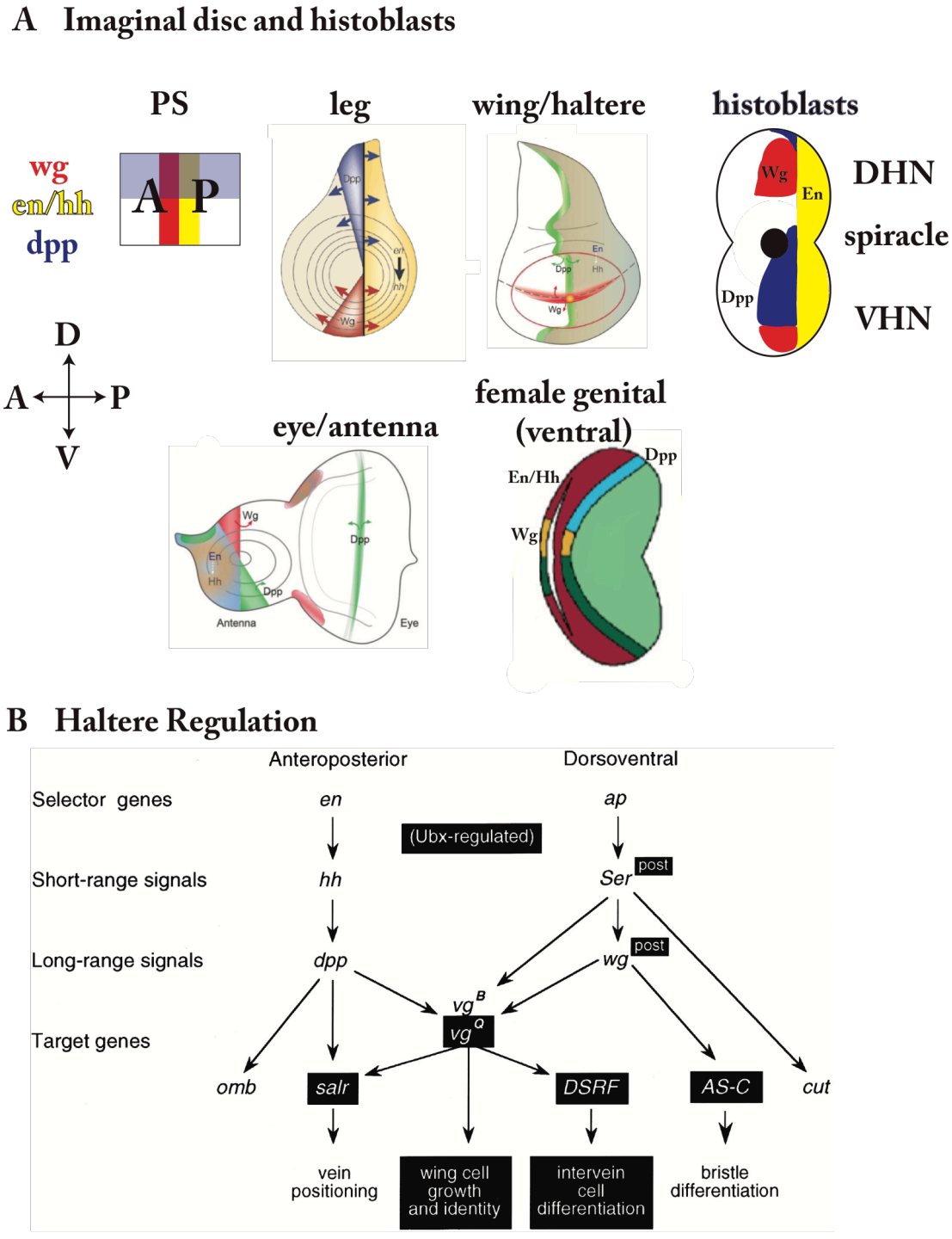


Figure 4.3.1. Comparable expressions of Wg, Dpp and En/Hh in developmental primordial compartments and Ubx regulatory network of haltere. A)Expression domains of Wg, Dpp and En/Hh are organized in similar manners in the parasegments, legs, wing (and haltere are very similar) and histoblast nests upon fusion of dorsal and ventral histoblast nests (DHN and VHN respectively). B) Ubx suppresses genes responsible for cell growth and differentiation (black boxes) in the haltere, which is responsible for differences in size between wing and halteres (as no Ubx expression in wing). vg^{Q} = quadrant enhancer (regulated by Ubx), vg^{B} = boundary enhancer, not Ubx regulated. *wg = wingless, en = engrailed, hh = hedgehog, dpp = decapentaplegic, ap = apterous, omb = optomotor blind (AKA bifid), Ser = Serrate, salr = spalt-related, DSRF = Drosophila serum Response Factor (AKA blistered), vg = vestigial, AS-C* = unknown. Leg disc image (Morata, 2001), wing/haltere/eye/antenna disc image (Beira and Paro, 2016), histoblast image (Kopp et al., 1999), female genital disc image (Chen and Baker, 1997) Ubx regulatory network (Weatherbee et al., 1998).

The progenitor cells of wing and haltere discs (dorsal thoracic) can be detected as 24 (wing) and 12 (haltere) cells by stage 13-14 embryos (Bate and Arias, 1991) but seem to originate associated with the leg discs in stage 12 (Cohen et al., 1993). By stage 15 the leg discs have relocated dorsally to invaginate into the embryo from the epithelium becoming increasingly partitioned into unique domains, proliferating and evaginating during larval and pupal stages, and finally forming bristles, trichomes and adult cuticle (Fristrom, 1993). *Ubx* has been implicated in regulation of growth and patterning of T2 and T3 leg discs (Rozowski and Akam, 2002; Stern, 2003) and is considered the master regulator of the size and differentiation between wing and haltere discs able to act at several stages of development (Roch and Akam, 2000; Weatherbee et al., 1998). *Antp* also plays an important role in leg development (Casares and Mann, 1998) and *Hox-G* is expressed in the same domain as the second promoter of *Antp* in stage 5 embryos (Fig.3.3.1). *Antp* functions in leg discs by repressing *hth*, an antennal determining gene along with *exd* (Casares and Mann, 1998). Antenna and legs are two homologous structures that are influenced by *Antp* expression, as *Antp* promotes leg development through the repression of *hth* and *exd* (Casares and Mann, 1998). If *Antp* is ectopically expressed in the head, then the antennae are transformed into legs by the same mechanisms that lead to conventional leg formation (Casares and Mann, 1998). Interestingly, legs develop in *hth* or *exd* mutants without *Antp, Ubx* (T3 patterning) or *Scr* (T1 patterning) expression, suggesting the ground state is to form legs from the disc primordial. This provides an explanation as to why other Hox genes can induce antennae to leg transformations (Casares and Mann, 1998; Morata and Sanchez-Herrero, 1998). In particular, *Scr, Antp, Ubx* and *abd-A* can also repress *hth* preventing Exd nuclear localization (Yao et al., 1999), which demonstrates shared functions of genes from both ANT-C and BX-C. Antp also blocks the eye selector gene, *eyeless* (*ey*), by protein interactions between the DNA-binding domains of Antp and Ey, leading to inhibition and loss-of-function of both proteins in different tissues (Plaza et al., 2001). This again held true for other Hox proteins, *Scr, Ubx, Abd-B* and *abd-A* when expressed using the same dpp^{blink}-Gal4 line (Staehling-Hampton et al., 1994b) used in our investigations, which for them led to eyes being significantly under developed (Plaza et al., 2001). As we did not see defects in eye development when overexpressing either *Hox-G* or G-PRE, but did see leg and haltere mutations (Table.3.6.2 and 3.6.3), it seems unlikely that *Hox-* or G-PRE is able to upregulate *Scr, Ubx, Abd-B* or *abd-A* ubiquitously, or in all the regions that the different Gal4 driver are expressing the transgenes. If *either Hox*-G or G-PRE were upregulating *Antp* in all regions in which the transgenes were ectopically overexpressed then we would expect see eye mutations consistent with its protein binding to Ey demonstrated by the Plaza (2001) study. Therefore, it would seem that if *Hox-G* or G-PRE is regulating a Hox gene then it must need a specific cellular environment to do so that does not match that found in the eye discs.

The genital imaginal disc primordia establishes cell lineages in a blastoderm embryo (Dübendorfer, 1982) and the Hox gene *Abd-B* is primarily responsible for this specification

(Estrada and Sanchez-Herrero, 2001). Lack of *Abd-B* induces leg or antenna formation in place of genitalia due to ectopic *Dll* expression, which is activated by Wg and Dpp, but usually repressed by Abd-B in wild-type situations (Estrada and Sanchez-Herrero, 2001). Homeotic genes *abd-A, Abd-B* and *caudal* are used to specify the three lineages of the genital primordial precursor cells from a 22 cell cluster of the ventral epidermis, into female or male genitalia and anal primordia. This occurs during mid-embryogenesis and by the third instar larval stage the compartments are organized by common imaginal disc genes, *en, hh, wg, ptc,* and *dpp* (Sanchez, 1997). *Egfr* is required for the initial development of all three genital disc precursor cells (Chen et al., 2005a) and apoptosis during genital disc development is regulated by JNK in cells that also expressed *en* and a balance of anti- and pro-apoptotic factors (Benitez et al., 2010). The eye-antenna disc has not been found to require information from Hox genes found in either the ANT-C or BX-C and instead 7 other master selector genes are thought to be responsible for eye and antenna development (Kumar and Moses, 2001a). These master selector genes are regulated upstream by EGFR and Notch signaling, which have homeotic functions, and loss of EGFR function led to deletion of both eyes and antenna, with eye and antenna specification from the master selector genes occurring in the second larval stage (Kumar and Moses, 2001a, b). Wg and Hh are responsible for size and shape (Kumar and Moses, 2001a). Lack of Dpp and Wg overlap in the eye disc prevents the leg and antenna specific gene *Dll* from being expressed, where it functions to specify proximodistal axis (Duong et al., 2008). Also, Wg morphogen, at high levels, is responsible for inducing cell death of peripheral ommatidia necessary for proper eye development (Kumar et al., 2015). We did not see any mutant phenotypes of the eye or genital discs when investigating effects of the partial duplication of *Hox-G* or overexpression studies, indicating that Hox genes have not been affected in these regions. If *Hox-G* or G-PRE is regulating a Hox gene, then it has other requirements to do so.

The mutations in our investigations do not seem to affect eye-antennal or genital disc development but seem to correspond to the regions of the embryo that we can detect *Hox-G* expression, segments T1-T2, T3-A1 and A8. This seems to suggest there is other factors involved *Hox-G's* mechanism of regulation that are specific to these regions. *Ubx* and *AntpP2* seem to be the most likely candidates of *Hox-G* regulation as it is expressed in the same pattern as *AntpP2* in stage 5 embryos but the mutations seem more consistent with the literature on *Ubx* functions. However, it does not appear to be a clear case of *Hox-G* driving or suppressing expression of either *Ubx* or *AntpP2* as the Gal4 driver lines are ubiquitous or expressed in all imaginal discs. However, we have only seen mutations that affect the discs that develop in the regions that seem to correspond to *Hox-G's* native expression. Furthermore, the fluorescent protein expression on the larval epidermis appears to correspond to the larval dorsal trichomes, small fine hair-like structures that are also patterned differently on T2 and T3 legs in response to different levels of *Ubx* (Davis et al., 2007). Trichomes are also found on other appendages, such as wings and eyes, linking Hox genes *Ubx* and

abd-A to genes *wg* and *hh* and the Notch and Egfr signaling pathways (Arif et al., 2015). Another possible theory for how G-PRE or *Hox-G* functions could be linked to the physical interaction observed between the ANT-C and BX-C chromatin; as seen particularly during development in tissues where both complexes are repressed (Bantignies et al., 2011). This interaction between ANT-C and BX-C is dependent on PcG proteins and theoretically the transcription of *Hox-G* or G-PRE could lead to a disruption of these interactions. However, very little is known about how these interactions are communicated. Other possible explanations of *Hox-G* or G-PRE's actions could include isoform specific effects on *Antp* or *Ubx* as both have multiple isoforms and little is known about the transcript from the second *Antp* promoter. The isoform that is either expressed or regulated in the cells that *Hox-G* is expressed in is likely to form very specific complexes with a variety of interacting partners in order to regulate genes at specific times in development. There is still a lot to be learnt about different interactions with Hox proteins, as demonstrated by a study that found Hox proteins interact with a large number of different TFs in developing embryos through conserved short linear amino acid motifs that can alter Hox proteins binding partners and potentially be a factor in tissue specific differences in binding activity (Baeza et al., 2015). It was demonstrated eight years ago that Hox proteins require additional cofactors to carry out their multiple actions on targets that is dependent on cellular environments (Berger et al., 2008; Noyes et al., 2008) and just recently the mediator complex subunit 19 (MED19) was identified as directly binding the homeodomain in order to access the RNA pol II machinery (Boube et al., 2014).

The mediator complex consists of highly conserved proteins found throughout eukaryotes and it is able to form a number of complexes with a variety of different conformations and subunit compositions that directly affect its interactions with TFs (Allen and Taatjes, 2015). Mediator is essential for transcriptional activation as it has the ability to convey signals from enhancer or promoter bound TFs to RNA pol II for transcription, can reorganize chromatin and regulate elongation, promoter pausing and release, and has been linked to a number of developmental diseases and cancers (Allen and Taatjes, 2015). We have strong indications from the CAGE and RNA-seq profiles that *AntpP2* and *Hox-G* are subject to promoter pausing, A previous study immunodepleted mediator and found that transcription was lost at the *AntpP2* and *en* promoters and is therefore necessary for their transcription (Park et al., 2001), although exactly how the Mediator complex achieves this is still a mystery (Allen and Taatjes, 2015). One theory is that it may be blocking nucleosome assembly to allow assembly of a pre-initiation complex, demonstrated in yeast and Drosophila. Furthermore, Mediator interacts with SWI/SNF chromatin remodeling complex, a complex that is also recruited by the PcG PhoRC complex (Table.1.4.1). Mediator, along with other proteins, is required for DNA looping of linearly separated sequences that interact to regulate transcription (Allen and Taatjes, 2015). One class of lncRNAs interacts with Mediator complex, termed activator RNA (aRNA), as they increase the levels of transcription of adjacent genes via a gene looping mechanism that is poorly understood. So far the interaction of aRNAs

and Mediator has been investigated in mammals and the aRNAs are thought to interact with Mediator subunit 12 (MED12) (Lai et al., 2013), which when mutated caused developmental defects in both mammals and Drosophila. Interestingly, in *D. melanogaster*, mutating MED12 or MED13 caused defects in wing and eye development; mutations in other Mediator subunits were linked to wing, eye and anteroposterior mutations leading to the Mediator complex being considered a master regulator of cell fate determination (Yin and Wang, 2014). It is not too surprising that imaginal disc development is affected in *D. melanogaster* as MED12 and MED13 are required to stimulate Wnt signaling in metazoans (Allen and Taatjes, 2015) and *wg* encodes a ligand of the Wnt signaling pathway (Swarup and Verheyen, 2012).

The requirement of Mediator for *AntpP2* transcription and the demonstrated roles of the class of lncRNAs, aRNAs, in interacting with Mediator to increase transcriptional activity, create a possible scenario for the actions of *Hox-G*. This could also help to explain why the mutant phenotypes found in our investigations seem to be limited to the endogenous primordial discs or cells that *Hox-G* could be expressed in. For example, if *Hox-G* were interacting with a Mediator complex, it is likely to only be able to interact with certain subunit(s) that may only be available in specific cells. Then other proteins required for transcription of *Hox-G's* targets could also only be available or able to function at certain loci, creating a unique environment containing specific Mediator proteins, specific TF's and RNA pol II transcription machinery that may not be found in other cells. If overexpression of *Hox-G* were to lead to increased levels of *Antp*, then it could be that the mutations that would normally be linked to *Ubx* could have been carried out by *Antp* as several Hox proteins seem able to replace each others functions, as previously mentioned. It is hard to imagine how each of the experiments testing *Hox-G's* function have produced similar phenotypes, particularly how overexpressing G-PRE led to identical phenotypes. However, much of our understanding of lncRNAs is in it's infancy and there are very few *in vivo* or *in vitro* investigations into the functions of PREs. Based on the limited knowledge we have of PRE/TREs, the transcribed TREs have been mostly associated with maintaining active transcription, possibly acting as a decoy to PcG proteins to displace them from the PRE (Davidovich et al., 2013; Herzog et al., 2014). If this were true for G-PRE, and this PRE was responsible for silencing of *Hox-G*, then this could provide one possible explanation as to how overexpressing G-PRE ubiquitously could lead to higher levels of *Hox-G*. Alternatively, G-PRE could also be regulating a Hox gene and not *Hox-G* and the matching phenotypes could be caused by one Hox gene substituting the function of another in a different tissue, as seen in a number of cases in different organisms (Foronda et al., 2009). Furthermore, the key imaginal disc and histoblast regulators, *en, hh* and *wg* are all also regulated by PcG and TrxG proteins (Beira and Paro, 2016) and G-PRE or may be affecting their regulation in the correct environment as PREs are also interchangeable as previously discussed.

Therefore, there are a few plausible scenarios that could be used to explain how *Hox-G* or G-PRE function, based on our results and literature: 1) There is another factor expressed in *Hox-G's* endogenous regions that is also necessary for its function. This could be specific PcG/TrxG or Mediator complexes. 2) G-PRE may be able to suppress *Hox-G* in the cells that don't normally express *Hox-G* and therefore *Hox-G* is only overexpressed in its endogenous domains. 3) G-PRE could be linked to the communication of ANT-C and BX-C, through interactions with other PREs and *Hox-G* could be involved in this, but this communication could be lost when ectopically overexpressed. 4) *Hox-G* or G-PRE could be responsible for both *Antp* and *Ubx* regulation or just one of these Hox genes that has gone on to replace the function of the other when dysregulated by the overexpression of *Hox-G* or G-PRE. 5) *Hox-G* is bound by proteins and loops around to bring those proteins into proximity of *AntpP2's* promoter, thus affecting transcription, possibly by forming a triplex to stabilize the interaction. Unfortunately, there is still much to be learnt about PRE/TRE functions and how or if their transcription is relevant, along with a better understanding of lncRNAs interactions with PcG/TrxG/Mediator complexes and how these interact with TFs and basal transcription machinery. Given our knowledge so far, it seems likely that cellular environments are very unique along both the A-P and D-V axes and also temporally very dynamic during development. Therefore, although many proteins function in similar ways in different cells, they could also have unknown functions in a subset of cells that could be easily missed by experiments on whole embryos spanning many hours of development. The different functions carried out by different proteins are likely to be largely affected by their environment and availability of interacting partners that they can form complexes with. When considering just how complicated gene regulation is, it seems we are still only just beginning to unravel the many factors involved. We still do not fully understand the large variety of protein complexes that exist, or how many of them function, but these complexes can regulate a gene or gene complex at promoters, enhancers or by affecting the surrounding chromatin. Protein complexes can also bind to regulatory regions of DNA, such as PRE/TREs, to alter the chromatin state, either locally or via looping mechanisms to their targets. However, how PRE/TREs function is still not well understood and whether or not the transcription found at a subset of PRE/TRE loci is relevant is still questionable. We have given evidence supporting the claim that transcription may not be linked to the PRE/TRE function. However, we have also shown that some of these transcripts are evolutionarily conserved and could therefore still be functional, but the function may not be associated with the underlying DNAs function as a PRE/TRE. We also now know that transcripts themselves are further regulated by different ncRNAs in a number of ways, such as splicing, localization, degradation and elongation (Quinn and Chang, 2016). Furthermore, lncRNAs have been shown to associate with a number of different protein complexes, particularly those that modify chromatin. Therefore, even though our understanding of functional lncRNAs is still in the very early stages, there are steadily growing lists of different gene regulation mechanisms linked to lncRNAs.

However, our lack of understanding of how protein complexes function or even a knowledge of all of those that exist in each cell at different stages of development, will need further investigation for us to understand particular lncRNAs roles within these complexes. Our investigations have led to identification of regulatory DNA, which is likely is be a PRE or TRE and an adjacent lncRNA that seems to function by regulating one or both of the adjacent Hox genes, *Ubx* and *Antp*. Literature seems to indicate that a protein complex could be involved, but this will require much more work to fully understand the mechanisms utilized by this lncRNA.

4.5 Future perspectives

This work focuses on gaining a better understanding of the long studied Hox complex of *D. melanogaster* and aims to better understand the regulation of these key developmental TFs by regulatory DNA that we now can detect to be transcribed. We have demonstrated that the lncRNA, *Hox-G*, has functional RNA and the adjacent PRE/TRE is able to selectively silence the *Hox-G* loci. It would now be useful to further understand the exact mechanisms employed by the *Hox-G* transcript in order for it to carry out its functions, along with the exact genes that it is being targeted to. A key to understanding the functions of the lncRNA is to find out the proteins, DNA or/and RNA that the lncRNA interacts with. Current methods for this have been largely designed for mammalian cell culture studies and require a very large amount of cells to carry this out. However, we wanted to design a methodology that would allow us to investigate these interactions *in vivo* and therefore utilized the CRIPSR/Cas9 system to integrate donor DNA into the second exon of *Hox-G,* carrying components that we could use for further analysis. These components included loxP and attP sites and the mini-white gene. The mini-white gene allowed us to screen for insertion and homozygosity, as well as acting as a reporter for PRE/TRE activity in our investigations. To further understand the functions of *Hox-G*, the next steps would be to use a Cre enzyme to remove the donor DNA between the two loxP sites (Sauer and Henderson, 1988), leaving just the attP site (Fig.2.5) in the second exon of *Hox-G*. This would remove mini-white and allow for highly efficient integration of other donor DNA using the PhiC31-attP-attB integration system (Keravala and Calos, 2008).

One way to continue the investigations could be to introduce MS2 stem loops (Peabody, 1993). This would allow live imaging of the RNA transcript as the MS2 stem loops should be transcribed along with *Hox-G*, then the flies would be expressing the MS2 coat protein (MCP) that strongly binds to the MS2 loops. The MCP coat protein can be conjugated to a fluorescent protein and used to carry out live imaging of RNA transcripts (Garcia et al., 2013) and ribonucleoprotein complexes can be investigated by cross-linking and immunoprecipitation of the fluorescent protein and MCP (Yoon et al., 2012). Further experiments utilizing the CRISPR/Cas9 technology could be carried out in order to block transcription of the lncRNA. This would involve using a 'dead' Cas9 enzyme that can bind DNA in multiple positions at the promoter region and

therefore block transcription machinery (Larson et al., 2013). These experiments would give a real insight into the proteins and DNA targets, along with enabling identification of any other RNA associated with the *Hox-G* transcript. To further understand just how the RNA may be associating with proteins, analysis can be carried out on RNA secondary and tertiary structure and mathematical modeling to link this to RNA binding domains of the proteins (Weeks, 2010). If Hox genes do not show altered expression patterns when lncRNAs that are thought to modulate them are perturbed, then qPCR could be used to test if levels of expression are being affected. These experiments could be used on many of the lncRNAs throughout the Hox complex with very readily available tools to find out how this novel class of molecules affects their functions.

5. REFERENCES

Aitken, S., Magi, S., Alhendi, A.M., Itoh, M., Kawaji, H., Lassmann, T., Daub, C.O., Arner, E., Carninci, P., Forrest, A.R., Hayashizaki, Y., Consortium, F., Khachigian, L.M., Okada-Hatakeyama, M., Semple, C.A., 2015. Transcriptional dynamics reveal critical roles for non-coding RNAs in the immediate-early response. PLoS Comput Biol 11, e1004217.

Akbari, O.S., Bousum, A., Bae, E., Drewell, R.A., 2006. Unraveling cis-regulatory mechanisms at the abdominal-A and Abdominal-B genes in the Drosophila bithorax complex. Dev Biol 293, 294-304.

Alfieri, C., Gambetta, M.C., Matos, R., Glatt, S., Sehr, P., Fraterman, S., Wilm, M., Muller, J., Muller, C.W., 2013. Structural basis for targeting the chromatin repressor Sfmbt to Polycomb response elements. Genes Dev 27, 2367-2379.

Allen, B.L., Taatjes, D.J., 2015. The Mediator complex: a central integrator of transcription. Nature reviews. Molecular cell biology 16, 155-166.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Americo, J., Whiteley, M., Brown, J.L., Fujioka, M., Jaynes, J.B., Kassis, J.A., 2002. A complex array of DNA-binding proteins required for pairing-sensitive silencing by a polycomb group response element from the Drosophila engrailed gene. Genetics 160, 1561-1571.

Amoutzias, G., Van de Peer, Y., 2008. Together we stand: genes cluster to coordinate regulation. Dev Cell 14, 640-642.

Anderson, A.E., Karandikar, U.C., Pepple, K.L., Chen, Z., Bergmann, A., Mardon, G., 2011. The enhancer of trithorax and polycomb gene Caf1/p55 is essential for cell survival and patterning in Drosophila development. Development 138, 1957-1966.

Ardehali, M.B., Mei, A., Zobeck, K.L., Caron, M., Lis, J.T., Kusch, T., 2011. Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. Embo J 30, 2817-2828.

Arif, S., Kittelmann, S., McGregor, A.P., 2015. From shavenbaby to the naked valley: trichome formation as a model for evolutionary developmental biology. Evol Dev 17, 120-126.

Arthanari, Y., Heintzen, C., Griffiths-Jones, S., Crosthwaite, S.K., 2014. Natural antisense transcripts and long non-coding RNA in Neurospora crassa. PloS one 9, e91353.

Avery, O.T., Macleod, C.M., McCarty, M., 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J Exp Med 79, 137-158.

Badenhorst, P., Voas, M., Rebay, I., Wu, C., 2002. Biological functions of the ISWI chromatin remodeling complex NURF. Genes Dev 16, 3186-3198.

Baena-Lopez, L.A., Alexandre, C., Mitchell, A., Pasakarnis, L., Vincent, J.P., 2013. Accelerated homologous recombination and subsequent genome modification in Drosophila. Development 140, 4818-4825.

Baeza, M., Viala, S., Heim, M., Dard, A., Hudry, B., Duffraisse, M., Rogulja-Ortmann, A., Brun, C., Merabet, S., 2015. Inhibitory activities of short linear motifs underlie Hox interactome specificity in vivo. Elife 4.

Baker, K.E., Parker, R., 2004. Nonsense-mediated mRNA decay: terminating erroneous gene expression. Curr Opin Cell Biol 16, 293-299.

Bantignies, F., Cavalli, G., 2011. Polycomb group proteins: repression in 3D. Trends Genet 27, 454-464.

Bantignies, F., Goodman, R.H., Smolik, S.M., 2000. Functional interaction between the coactivator Drosophila CREB-binding protein and ASH1, a member of the trithorax group of chromatin modifiers. Molecular and cellular biology 20, 9317-9330.

Bantignies, F., Grimaud, C., Lavrov, S., Gabut, M., Cavalli, G., 2003. Inheritance of Polycomb-dependent chromosomal interactions in Drosophila. Genes & development 17, 2406-2420.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., Cavalli, G., 2011. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell 144, 214-226.

Basler, K., Struhl, G., 1994. Compartment boundaries and the control of Drosophila limb pattern by hedgehog protein. Nature 368, 208-214.

Bate, M., Arias, A.M., 1991. The embryonic origin of imaginal discs in Drosophila. Development 112, 755-761.

Bateson, W., 1894. Materials for the study of variation treated with especial regards to discontinuity in the origin of species Macmillan, London.

Batut, P., Dobin, A., Plessy, C., Carninci, P., Gingeras, T.R., 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome research 23, 169-180.

Batut, P., Gingeras, T.R., 2013. RAMPAGE: promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. Curr Protoc Mol Biol 104, Unit 25B 11.

Beadle, G.W., Tatum, E.L., 1941. Genetic Control of Biochemical Reactions in Neurospora. Proceedings of the National Academy of Sciences of the United States of America 27, 499-506.

Beck, F., 2002. Homeobox genes in gut development. Gut 51, 450-454.

Beira, J.V., Paro, R., 2016. The legacy of Drosophila imaginal discs. Chromosoma.

Beisel, C., Buness, A., Roustan-Espinosa, I.M., Koch, B., Schmitt, S., Haas, S.A., Hild, M., Katsuyama, T., Paro, R., 2007. Comparing active and repressed expression states of genes

controlled by the Polycomb/Trithorax group proteins. Proceedings of the National Academy of Sciences of the United States of America 104, 16615-16620.

Bell, O., Schwaiger, M., Oakeley, E.J., Lienert, F., Beisel, C., Stadler, M.B., Schubeler, D., 2010. Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. Nat Struct Mol Biol 17, 894-900.

Bender, W., Akam, M., Karch, F., Beachy, P.A., Peifer, M., Spierer, P., Lewis, E.B., Hogness, D.S., 1983. Molecular Genetics of the Bithorax Complex in Drosophila melanogaster. Science 221, 23-29.

Bender, W., Hudson, A., 2000. P element homing to the Drosophila bithorax complex. Development 127, 3981-3992.

Benitez, S., Sosa, C., Tomasini, N., Macias, A., 2010. Both JNK and apoptosis pathways regulate growth and terminalia rotation during Drosophila genital disc development. Int J Dev Biol 54, 643-653.

Benjajati, C., Mueller, L., Xu, N., Pappano, M., Gao, J., Mosammaparast, M., Conklin, D., Granok, H., Craig, C., Elgin, S.C.R., 1997. Multiple isoforms of GAGA factor, a critical component of chromatin structure. Nucleic Acid Res. 25, 3345-3353.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., Khalid, F., Zhang, W., Newburger, D., Jaeger, S.A., Morris, Q.D., Bulyk, M.L., Hughes, T.R., 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133, 1266-1276.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S.L., Lander, E.S., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315-326.

Bertani, G., 1951. Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. J Bacteriol 62, 293-300.

Bertani, S., Sauer, S., Bolotin, E., Sauer, F., 2011. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. Mol Cell 43, 1040-1046.

Beuchle, D., Struhl, G., Muller, J., 2001. Polycomb group proteins and heritable silencing of Drosophila Hox genes. Development 128, 993-1004.

Bienz, M., 1994. Homeotic genes and positional signalling in the Drosophila viscera. Trends Genet 10, 22-26.

Bird, A.J., Gordon, M., Eide, D.J., Winge, D.R., 2006. Repression of ADH1 and ADH3 during zinc deficiency by Zap1-induced intergenic RNA transcripts. Embo J 25, 5726-5734.

Birve, A., Sengupta, A.K., Beuchle, D., Larsson, J., Kennison, J.A., Rasmuson-Lestander, A., Muller, J., 2001. Su(z)12, a novel Drosophila Polycomb group gene that is conserved in vertebrates and plants. Development 128, 3371-3379.

Blastyak, A., Mishra, R.K., Karch, F., Gyurkovics, H., 2006. Efficient and specific targeting of Polycomb group proteins requires cooperative interaction between Grainyhead and Pleiohomeotic. Molecular and cellular biology 26, 1434-1444.

Bohmdorfer, G., Wierzbicki, A.T., 2015. Control of Chromatin Structure by Long Noncoding RNA. Trends Cell Biol 25, 623-632.

Bond, A.M., Vangompel, M.J., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., Kohtz, J.D., 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat Neurosci 12, 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., Ballabio, A., 1991. Characterization of a murine gene expressed from the inactive X chromosome. Nature 351, 325-329.

Boube, M., Hudry, B., Immarigeon, C., Carrier, Y., Bernat-Fabre, S., Merabet, S., Graba, Y., Bourbon, H.M., Cribbs, D.L., 2014. Drosophila melanogaster Hox transcription factors access the RNA polymerase II machinery through direct homeodomain binding to a conserved motif of mediator subunit Med19. PLoS Genet 10, e1004303.

Bownes, M., 1976. Larval and adult abdominal defects resulting from microcautery of blastoderm staged Drosophila embryos. J Exp Zool 195, 369-392.

Brand, A.H., Perrimon, N., 1993. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development 118, 401-415.

Bridges, C.B.a.M., T. H., 1923. The third-chromosome group of mutant characters of Drosophila melanogaster.

Brockdorff, N., 2013. Noncoding RNA and Polycomb recruitment. Rna 19, 429-442.

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

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., 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, 515-526.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., 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, 38-44.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J., 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, 527-542.

Brown, D.M.a.T., A.R., 1952. Nucleotides Part X: Some observations on the structure and chemical behaviour of the nucleic acids. J. Chem. Soc, 52-58.

Brown, J.L., Fritsch, C., Mueller, J., Kassis, J.A., 2003. The *Drosophila* pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. Development 130, 285-294.

Brown, J.L., Grau, D.J., DeVido, S.K., Kassis, J.A., 2005. An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene. Nucleic acids research 33, 5181-5189.

Brown, J.L., Mucci, D., Whiteley, M., Dirksen, M.L., Kassis, J.A., 1998. The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. Mol Cell 1, 1057-1064.

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., West, R.B., 2012. Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol 13, R75.

Bull, A.L., 1966. *Bicaudal* a genetic factor which affects the polarity of the embryo of *Drosophila melanogaster*. J Exp Zool 161, 221-242.

Bullock, S.L., Stauber, M., Prell, A., Hughes, J.R., Ish-Horowicz, D., Schmidt-Ott, U., 2004. Differential cytoplasmic mRNA localisation adjusts pair-rule transcription factor activity to cytoarchitecture in dipteran evolution. Development 131, 4251-4261.

Busturia, A., Bienz, M., 1993. Silencers in abdominal-B, a homeotic Drosophila gene. Embo J 12, 1415-1425.

Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., Gabellini, D., 2012. A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell 149, 819-831.

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

Calhoun, V.C., Stathopoulos, A., Levine, M., 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proceedings of the National Academy of Sciences of the United States of America 99, 9243-9247.

Callus, B.A., Mathey-Prevot, B., 2002. SOCS36E, a novel Drosophila SOCS protein, suppresses JAK/STAT and EGF-R signalling in the imaginal wing disc. Oncogene 21, 4812-4821.

Campos-Ortega, J.A., Hartenstein, V., 1997. The embryonic development of Drosophila melanogaster.

Cao, R., Wang, H., He, J., Erdjument-Bromage, H., Tempst, P., Zhang, Y., 2008. Role of hPHF1 in H3K27 methylation and Hox gene silencing. Molecular and cellular biology 28, 1862-1872.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., Zhang, Y., 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 298, 1039-1043.

Cao, R., Zhang, Y., 2004. SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. Mol Cell 15, 57-67.

Capdevila, J., Guerrero, I., 1994. Targeted expression of the signaling molecule decapentaplegic induces pattern duplications and growth alterations in Drosophila wings. Embo J 13, 4459-4468.

Capovilla, M., Botas, J., 1998. Functional dominance among Hox genes: repression dominates activation in the regulation of Dpp. Development 125, 4949-4957.

Carlson, H.L., Quinn, J.J., Yang, Y.W., Thornburg, C.K., Chang, H.Y., Stadler, H.S., 2015. LncRNA-HIT Functions as an Epigenetic Regulator of Chondrogenesis through Its Recruitment of p100/CBP Complexes. PLoS Genet 11, e1005680.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schonbach, C., Sekiguchi, K., Semple, C.A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559-1563.

Carrasco, M.S., Gomez Armenta, F., de Ory, M.J., Sanchez, G., Marin, A., Gil, C., Andreu, L., Canos, J., Bolinches, R., 1985. [Genetic study of plasma cholinesterases in the adult patients of the Cadiz region]. Rev Esp Anestesiol Reanim 32, 156-158.

Carroll, S.B., DiNardo, S., O'Farrell, P.H., White, R.A., Scott, M.P., 1988. Temporal and spatial relationships between segmentation and homeotic gene expression in Drosophila embryos: distributions of the fushi tarazu, engrailed, Sex combs reduced, Antennapedia, and Ultrabithorax proteins. Genes Dev 2, 350-360.

Carroll, S.B., Grenier, J., Weatherbee, S., 2009. From DNA to Diversity : Molecular Genetics and the Evolution of Animal Design.

Carroll, S.B., Scott, M.P., 1986. Zygotically active genes that affect the spatial expression of the fushi tarazu segmentation gene during early Drosophila embryogenesis. Cell 45, 113-126.

Carroll, S.B., Weatherbee, S.D., Langeland, J.A., 1995. Homeotic genes and the regulation and evolution of insect wing number. Nature 375, 58-61.

Carvalho, A.B., Clark, A.G., 2013. Efficient identification of Y chromosome sequences in the human and Drosophila genomes. Genome research 23, 1894-1907.

Casares, F., Mann, R.S., 1998. Control of antennal versus leg development in Drosophila. Nature 392, 723-726.

Casares, F., Sanchez-Herrero, E., 1995. Regulation of the infraabdominal regions of the bithorax complex of Drosophila by gap genes. Development 121, 1855-1866.

Castelli-Gair, J., Akam, M., 1995. How the Hox gene Ultrabithorax specifies two different segments: the significance of spatial and temporal regulation within metameres. Development 121, 2973-2982.

Castelnuovo, M., Stutz, F., 2015. Role of chromatin, environmental changes and single cell heterogeneity in non-coding transcription and gene regulation. Curr Opin Cell Biol 34, 16-22.

Cavalli, G., 2002. Chromatin as a eukaryotic template of genetic information. Curr Opin Cell Biol 14, 269-278.

Cavalli, G., Paro, R., 1998. The *Drosophila* Fab-7 chromosomal element conveys epigenetic inheritance during mitosis and meiosis. Cell 93, 505-518.

Cavalli, G., Paro, R., 1999. Epigenetic inheritance of active chromatin after removal of the main transactivator. Science 286, 955-958.

Celniker, S.E., Sharma, S., Keelan, D.J., Lewis, E.B., 1990. The molecular genetics of the bithorax complex of Drosophila: cis-regulation in the Abdominal-B domain. Embo J 9, 4277-4286.

Chan, C.S., Rastelli, L. and Pirrotta, V., 1994. A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. EMBO J. 13, 2553-2564.

Chaney, J.L., Clark, P.L., 2015. Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys 44, 143-166.

Chang, Y.L., Peng, Y.H., Pan, I.C., Sun, D.S., King, B., Huang, D.H., 2001. Essential role of Drosophila Hdac1 in homeotic gene silencing. Proceedings of the National Academy of Sciences of the United States of America 98, 9730-9735.

Chaumeil, J., Le Baccon, P., Wutz, A., Heard, E., 2006. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. Genes Dev 20, 2223-2237.

Chen, E.H., Baker, B.S., 1997. Compartmental organization of the Drosophila genital imaginal discs. Development 124, 205-218.

Chen, E.H., Christiansen, A.E., Baker, B.S., 2005a. Allocation and specification of the genital disc precursor cells in Drosophila. Dev Biol 281, 270-285.

Chen, G., Fernandez, J., Mische, S., Courey, A.J., 1999. A functional interaction between the histone deacetylase Rpd3 and the corepressor groucho in Drosophila development. Genes Dev 13, 2218-2230.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q., 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research 41, D983-986.

Chen, S., Birve, A., and Rasmuson-Lestander, A., 2008. In vivo analysis of Drosophila Su(z)12 function. Mol. Genet. Genomics 279, 159-170.

Chen, X., Hiller, M., Sancak, Y., Fuller, M.T., 2005b. Tissue-specific TAFs counteract Polycomb to turn on terminal differentiation. Science 310, 869-872.

Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B., Jaenisch, R., 2013. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res 23, 1163-1171.

Cheng, Y., Kwon, D.Y., Arai, A.L., Mucci, D., Kassis, J.A., 2012. P-element homing is facilitated by engrailed polycomb-group response elements in Drosophila melanogaster. PloS one 7, e30437.

Cherbas, L., Hu, X., Zhimulev, I., Belyaeva, E., Cherbas, P., 2003. EcR isoforms in Drosophila: testing tissue-specific requirements by targeted blockade and rescue. Development 130, 271-284.

Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnar, Z., Ponting, C.P., 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome Biol 11, R72.

Chu, C., Quinn, J., Chang, H.Y., 2012. Chromatin isolation by RNA purification (ChIRP). J Vis Exp.

Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., Lawrence, J.B., 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 33, 717-726.

Clemson, C.M., McNeil, J.A., Willard, H.F., Lawrence, J.B., 1996. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J Cell Biol 132, 259-275.

Cohen, B., Simcox, A.A., Cohen, S.M., 1993. Allocation of the thoracic imaginal primordia in the Drosophila embryo. Development 117, 597-608.

Cohen, S.M., Jurgens, G., 1990. Mediation of Drosophila head development by gap-like segmentation genes. Nature 346, 482-485.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., 2013. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823.

Consortium, E.P., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.K., Ooi, H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaoz, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Loytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Program, N.C.S., Baylor College of Medicine Human Genome Sequencing, C., Washington University Genome Sequencing, C., Broad, I., Children's Hospital Oakland Research, I., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N., Yu, Y., Ruan, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Armengol, L., Bird, C.P., de Bakker, P.I., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrimsdottir, I.B., Huppert, J., Zody, M.C., Abecasis, G.R., Estivill, X., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., de Jong, P.J., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.

Consortium, R.N., 2015. RNAcentral: an international database of ncRNA sequences. Nucleic acids research 43, D123-129.

Couture, J.F., Collazo, E., Trievel, R.C., 2006. Molecular recognition of histone H3 by the WD40 protein WDR5. Nat Struct Mol Biol 13, 698-703.

Crick, F.H., Lawrence, P.A., 1975. Compartments and polyclones in insect development. Science 189, 340-347.

Crickmore, M.A., Mann, R.S., 2006. Hox control of organ size by regulation of morphogen production and mobility. Science 313, 63-68.

Crickmore, M.A., Mann, R.S., 2007. Hox control of morphogen mobility and organ development through regulation of glypican expression. Development 134, 327-334.

Crosby, M.A., Miller, C., Alon, T., Watson, K. L., Verrijzer, C. P., Goldman-Levi, R., and Zak, N. B., 1999. The trithorax group gene moira encodes a brahma-associated putative chromatin-remodeling factor in Drosophila melanogaster. Mol. Cell. Biol. 19, 1159-1170.

Crown, K.N., McMahan, S., Sekelsky, J., 2014. Eliminating both canonical and short-patch mismatch repair in Drosophila melanogaster suggests a new meiotic recombination model. PLoS Genet 10, e1004583.

Crozatier, M., Meister, M., 2007. Drosophila haematopoiesis. Cell Microbiol 9, 1117-1126.

Cumberledge, S., Zaratzian, A., Sakonju, S., 1990. Characterization of two RNAs transcribed from the cis-regulatory region of the abd-A domain within the Drosophila bithorax complex. Proceedings of the National Academy of Sciences of the United States of America 87, 3259-3263.

Cunningham, M.D., Brown, J.L., Kassis, J.A., 2010. Characterization of the polycomb group response elements of the Drosophila melanogaster invected Locus. Molecular and cellular biology 30, 820-828.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., Pirrotta, V., 2002. *Drosophila* enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell 111, 185-196.

D'Haeseleer, P., 2005. How does gene expression clustering work? Nat Biotechnol 23, 1499-1501.

Darnell, R., 2012. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. Cold Spring Harb Protoc 2012, 1146-1160.

Dasen, J.S., 2013. Long noncoding RNAs in development: solidifying the Lncs to Hox gene regulation. Cell Rep 5, 1-2.

Davidovich, C., Zheng, L., Goodrich, K.J., Cech, T.R., 2013. Promiscuous RNA binding by Polycomb repressive complex 2. Nat Struct Mol Biol 20, 1250-1257.

Davis, G.K., Srinivasan, D.G., Wittkopp, P.J., Stern, D.L., 2007. The function and regulation of Ultrabithorax in the legs of Drosophila melanogaster. Dev Biol 308, 621-631.

de Navas, L.F., Garaulet, D.L., Sanchez-Herrero, E., 2006. The ultrabithorax Hox gene of Drosophila controls haltere size by regulating the Dpp pathway. Development 133, 4495-4506.

Dejardin, J., Cavalli, G., 2004. Chromatin inheritance upon Zeste-mediated Brahma recruitment at a minimal cellular memory module. Embo J 23, 857-868.

Dejardin, J., Rappailles, A., Cuvier, O., Grimaud, C., Decoville, M., Locker, D., Cavalli, G., 2005. Recruitment of Drosophila Polycomb group proteins to chromatin by DSP1. Nature 434, 533-538.

Dekker, J., Marti-Renom, M.A., Mirny, L.A., 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390-403.

Delest, A., Sexton, T., Cavalli, G., 2012. Polycomb: a paradigm for genome organization from one to three dimensions. Curr Opin Cell Biol 24, 405-414.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., Guigo, R., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research 22, 1775-1789.

DeVido, S.K., Kwon, D., Brown, J.L., Kassis, J.A., 2008. The role of Polycomb-group response elements in regulation of engrailed transcription in Drosophila. Development 135, 669-676.

Di Gesualdo, F., Capaccioli, S., Lulli, M., 2014. A pathophysiological view of the long non-coding RNA world. Oncotarget 5, 10976-10996.

Diaz-Benjumea, F.J., Cohen, S.M., 1994. wingless acts through the shaggy/zeste-white 3 kinase to direct dorsal-ventral axis formation in the Drosophila leg. Development 120, 1661-1670.

Dieci, G., Preti, M., Montanini, B., 2009. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics 94, 83-88.

Diederichs, S., 2014. The four dimensions of noncoding RNA conservation. Trends Genet 30, 121-123.

Dillon, S.C., Zhang, X., Trievel, R.C., Cheng, X., 2005. The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biol 6, 227.

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., Jacks, T., 2014. LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol Cell 54, 777-790.

DiNardo, S., O'Farrell, P.H., 1987. Establishment and refinement of segmental pattern in the Drosophila embryo: spatial control of engrailed expression by pair-rule genes. Genes Dev 1, 1212-1225.

Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E., Ru, K., Solda, G., Simons, C., Sunkin, S.M., Crowe, M.L., Grimmond, S.M., Perkins, A.C., Mattick, J.S., 2008a. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome research 18, 1433-1445.

Dinger, M.E., Pang, K.C., Mercer, T.R., Mattick, J.S., 2008b. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 4, e1000176.

Dingwall, A.K., Beek, S.J., McCallum, C.M., Tamkun, J.W., Kalpana, G.V., Goff, S.P., Scott, M.P., 1995. The Drosophila snr1 and brm proteins are related to yeast SWI/SNF proteins and are components of a large protein complex. Mol Biol Cell 6, 777-791.

Dorighi, K.M., Tamkun, J.W., 2013. The trithorax group proteins Kismet and ASH1 promote H3K36 dimethylation to counteract Polycomb group repression in Drosophila. Development 140, 4182-4192.

Driever, W., Nusslein-Volhard, C., 1988. A gradient of bicoid protein in Drosophila embryos. Cell 54, 83-93.

Driever, W., Nusslein-Volhard, C., 1989. The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo. Nature 337, 138-143.

Driever, W., Thoma, G., Nusslein-Volhard, C., 1989. Determination of spatial domains of zygotic gene expression in the Drosophila embryo by the affinity of binding sites for the bicoid morphogen. Nature 340, 363-367.

Du, Z., Sun, T., Hacisuleyman, E., Fei, T., Wang, X., Brown, M., Rinn, J.L., Lee, M.G., Chen, Y., Kantoff, P.W., Liu, X.S., 2016. Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. Nat Commun 7, 10982.

Dübendorfer, K.a.N., R., 1982. A clonal analysis of cell lineage and growth in the male and female genital discs of *Drosophila melanogaster*. Dev. Biol. 191, 42-45.

Duboule, D., 1992. The vertebrate limb: A model system to study the Hox/HOM gene network during development and evolution. Bioessays 14, 375-384.

Duboule, D., Morata, G., 1994. Colinearity and functional hierarchy among genes of the homeotic complexes. Trends in genetics : TIG 10, 358-364.

Duncan, I.M., 1982. Polycomblike: a gene that appears to be required for the normal expression of the bithorax and antennapedia gene complexes of Drosophila melanogaster. Genetics 102, 49-70.

Duong, H.A., Wang, C.W., Sun, Y.H., Courey, A.J., 2008. Transformation of eye to antenna by misexpression of a single gene. Mech Dev 125, 130-141.

Eaton, M.L., Prinz, J.A., MacAlpine, H.K., Tretyakov, G., Kharchenko, P.V., MacAlpine, D.M., 2011. Chromatin signatures of the Drosophila replication program. Genome research 21, 164-174.

Eissenberg, J.C., Shilatifard, A., 2010. Histone H3 lysine 4 (H3K4) methylation in development and differentiation. Dev Biol 339, 240-249.

Enderle, D., Beisel, C., Stadler, M.B., Gerstung, M., Athri, P., Paro, R., 2011. Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. Genome research 21, 216-226.

Engreitz, J., Lander, E.S., Guttman, M., 2015. RNA antisense purification (RAP) for mapping RNA interactions with chromatin. Methods Mol Biol 1262, 183-197.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., Plath, K., Guttman, M., 2013. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341, 1237973.

Enriquez, J., Boukhatmi, H., Dubois, L., Philippakis, A.A., Bulyk, M.L., Michelson, A.M., Crozatier, M., Vincent, A., 2010. Multi-step control of muscle diversity by Hox proteins in the Drosophila embryo. Development 137, 457-466.

Erokhin, M., Elizar'ev, P., Parshikov, A., Schedl, P., Georgiev, P., Chetverina, D., 2015. Transcriptional read-through is not sufficient to induce an epigenetic switch in the silencing activity of Polycomb response elements. Proceedings of the National Academy of Sciences of the United States of America 112, 14930-14935.

Estella, C., Mann, R.S., 2008. Logic of Wg and Dpp induction of distal and medial fates in the Drosophila leg. Development 135, 627-636.

Estella, C., Rieckhof, G., Calleja, M., Morata, G., 2003. The role of buttonhead and Sp1 in the development of the ventral imaginal discs of Drosophila. Development 130, 5929-5941.

Estrada, B., Sanchez-Herrero, E., 2001. The Hox gene Abdominal-B antagonizes appendage development in the genital disc of Drosophila. Development 128, 331-339.

Falaleeva, M., Stamm, S., 2013. Processing of snoRNAs as a new source of regulatory non-coding RNAs: snoRNA fragments form a new class of functional RNAs. Bioessays 35, 46-54.

Fatica, A., Bozzoni, I., 2014. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15, 7-21.

Fatima, R., Akhade, V.S., Pal, D., Rao, S.M., 2015. Long noncoding RNAs in development and cancer: potential biomarkers and therapeutic targets. Mol Cell Ther 3, 5.

Fauvarque, M.O., Zuber, V and Dura, J-M., 1995. Regulation of polyhomeotic transcription may involve local changes in chromatin activity in Drosophila. Mech. Dev. 52, 343-355.

Felsenfeld, A.L., Kennison, J.A., 1995. Positional signaling by hedgehog in Drosophila imaginal disc development. Development 121, 1-10.

Feng, J., Bi, C., Clark, B.S., Mady, R., Shah, P., 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 Dev 20, 1470-1484.

Fiedler, T., and Rehmsmeier, M., 2006. jPREdictor: a versatile tool for the prediction of cis-regulatory elements. Nucl. Acids Res. 34, 546-550.

Finkelstein, R., Perrimon, N., 1990. The orthodenticle gene is regulated by bicoid and torso and specifies Drosophila head development. Nature 346, 485-488.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic acids research 42, D222-230.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic acids research 44, D279-285.

Foronda, D., de Navas, L.F., Garaulet, D.L., Sanchez-Herrero, E., 2009. Function and specificity of Hox genes. Int J Dev Biol 53, 1404-1419.

Foronda, D., Martin, P., Sanchez-Herrero, E., 2012. Drosophila Hox and sex-determination genes control segment elimination through EGFR and extramacrochetae activity. PLoS Genet 8, e1002874.

Franke, A., DeCamillis, M., Zink, D., Cheng, N., Brock, H.W., Paro, R., 1992. Polycomb and polyhomeotic are constituents of a multimeric protein complex in chromatin of Drosophila melanogaster. Embo J 11, 2941-2950.

Fristrom, D.a.F., J. W., 1993. The metamorphic development of the adult epidermis. Cold Spring Harbor Laboratory Press, NY.

Fuchs, J., Demidov, D., Houben, A., Schubert, I., 2006. Chromosomal histone modification patterns--from conservation to diversity. Trends Plant Sci 11, 199-208.

Furlong, E.E., Andersen, E.C., Null, B., White, K.P., Scott, M.P., 2001. Patterns of gene expression during Drosophila mesoderm development. Science 293, 1629-1633.

Gaiti, F., Fernandez-Valverde, S.L., Nakanishi, N., Calcino, A.D., Yanai, I., Tanurdzic, M., Degnan, B.M., 2015. Dynamic and Widespread lncRNA Expression in a Sponge and the Origin of Animal Complexity. Mol Biol Evol 32, 2367-2382.

Galant, R., Carroll, S.B., 2002. Evolution of a transcriptional repression domain in an insect Hox protein. Nature 415, 910-913.

Galindo, M.I., Fernandez-Garza, D., Phillips, R., Couso, J.P., 2011. Control of Distal-less expression in the Drosophila appendages by functional 3' enhancers. Dev Biol 353, 396-410.

Galindo, M.I., Pueyo, J.I., Fouix, S., Bishop, S.A., Couso, J.P., 2007. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol 5, e106.

Garcia, H.G., Tikhonov, M., Lin, A., Gregor, T., 2013. Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning. Curr Biol 23, 2140-2145.

Garcia-Fernandez, J., 2005. The genesis and evolution of Homeobox gene clusters. Nature Reviews Genetics 6, 881-892.

Gaytan de Ayala Alonso, A., Gutierrez, L., Fritsch, C., Papp, B., Beuchle, D., Muller, J., 2007. A genetic screen identifies novel polycomb group genes in Drosophila. Genetics 176, 2099-2108.

Gebelein, B., Culi, J., Ryoo, H.D., Zhang, W., Mann, R.S., 2002. Specificity of Distalless repression and limb primordia development by abdominal Hox proteins. Dev Cell 3, 487-498.

Gebelein, B., McKay, D.J., Mann, R.S., 2004. Direct integration of Hox and segmentation gene inputs during Drosophila development. Nature 431, 653-659.

Geisler, S.J., Paro, R., 2015. Trithorax and Polycomb group-dependent regulation: a tale of opposing activities. Development 142, 2876-2887.

Gilles, A.F., Averof, M., 2014. Functional genetics for all: engineered nucleases, CRISPR and the gene editing revolution. Evodevo 5, 43.

Gillis, J., Pavlidis, P., 2013. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. Bioinformatics 29, 476-482.

Gindhart, J.G., Jr., King, A.N., Kaufman, T.C., 1995. Characterization of the cis-regulatory region of the Drosophila homeotic gene Sex combs reduced. Genetics 139, 781-795.

Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nat Genet 3, 266-272.

Gong, C., Maquat, L.E., 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature 470, 284-288.

Gorfinkiel, N., Sanchez, L., Guerrero, I., 1999. Drosophila terminalia as an appendage-like structure. Mech Dev 86, 113-123.

Gough, J., Karplus, K., Hughey, R., Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313, 903-919.

Grant, J., Mahadevaiah, S.K., Khil, P., Sangrithi, M.N., Royo, H., Duckworth, J., McCarrey, J.R., VandeBerg, J.L., Renfree, M.B., Taylor, W., Elgar, G., Camerini-Otero, R.D., Gilchrist, M.J., Turner, J.M., 2012. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature 487, 254-258.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B., Zufall, R.A., Elhaik, E., 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome biology and evolution 5, 578-590.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., Brown, J.B., Cherbas, L., Davis, C.A., Dobin, A., Li, R., Lin, W., Malone, J.H., Mattiuzzo, N.R., Miller, D., Sturgill, D., Tuch, B.B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R.E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J.E., Wan, K.H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P.J., Brenner, S.E., Brent, M.R., Cherbas, P., Gingeras, T.R.,

Hoskins, R.A., Kaufman, T.C., Oliver, B., Celniker, S.E., 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471, 473-479.

Grimm, S., Pflugfelder, G.O., 1996. Control of the gene optomotor-blind in Drosophila wing development by decapentaplegic and wingless. Science 271, 1601-1604.

Grosshans, H., Filipowicz, W., 2008. Molecular biology: the expanding world of small RNAs. Nature 451, 414-416.

Gu, Y., Nakamura, T., Alder, H., Prasad, R., Canaani, O., Cimino, G., Croce, C.M., Canaani, E., 1992. The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. Cell 71, 701-708.

Guil, S., Esteller, M., 2012. Cis-acting noncoding RNAs: friends and foes. Nat Struct Mol Biol 19, 1068-1075.

Guilgur, L.G., Prudencio, P., Sobral, D., Liszekova, D., Rosa, A., Martinho, R.G., 2014. Requirement for highly efficient pre-mRNA splicing during Drosophila early embryonic development. Elife 3, e02181.

Gummalla, M., Maeda, R.K., Castro Alvarez, J.J., Gyurkovics, H., Singari, S., Edwards, K.A., Karch, F., Bender, W., 2012. abd-A regulation by the iab-8 noncoding RNA. PLoS Genet 8, e1002720.

Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., Robson, P., 2010. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell 18, 675-685.

Guo, J., Jin, D., 2015. A genetic screen in Drosophila implicates Sex comb on midleg (Scm) in tissue overgrowth and mechanisms of Scm degradation by Wds. Mech Dev 136, 1-7.

Guo, J., Jing, R., Lv, X., Wang, X., Li, J., Li, L., Li, C., Wang, D., Bi, B., Chen, X., Yang, J.H., 2016. H2A/K pseudogene mutation may promote cell proliferation. Mutat Res 787, 32-42.

Gustavson, E., Goldsborough, A.S., Ali, Z., Kornberg, T.B., 1996. The Drosophila engrailed and invected genes: partners in regulation, expression and function. Genetics 142, 893-906.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., Lander, E.S., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223-227.

Hacker, U., Perrimon, N., 1998. DRhoGEF2 encodes a member of the Dbl family of oncogenes and controls cell shape changes during gastrulation in Drosophila. Genes Dev 12, 274-284.

Hagstrom, K., Muller, M., Schedl, P., 1997. A Polycomb and GAGA dependent silencer adjoins the Fab-7 boundary in the Drosophila bithorax complex. Genetics 146, 1365-1380.

Hainer, S.J., Pruneski, J.A., Mitchell, R.D., Monteverde, R.M., Martens, J.A., 2011. Intergenic transcription causes repression by directing nucleosome assembly. Genes Dev 25, 29-40.

Hama, C., Ali, Z., Kornberg, T.B., 1990. Region-specific recombination and expression are directed by portions of the Drosophila engrailed promoter. Genes Dev 4, 1079-1093.

Han, P., Chang, C.P., 2015. Long non-coding RNA and chromatin remodeling. RNA Biol 12, 1094-1098.

Hannah-Alava, A., 1964. Interaction of Non-Allelic Loci in Expression of the Extra-Sexcomb Phenotype in Drosophila Melanogaster. Z Vererbungsl 95, 1-9.

Hannus, M., Feiguin, F., Heisenberg, C.P., Eaton, S., 2002. Planar cell polarization requires Widerborst, a B' regulatory subunit of protein phosphatase 2A. Development 129, 3493-3503.

Harrison, D.A., Binari, R., Nahreini, T.S., Gilman, M., Perrimon, N., 1995. Activation of a Drosophila Janus kinase (JAK) causes hematopoietic neoplasia and developmental defects. Embo J 14, 2857-2865.

Hauenschild, A., Ringrose, L., Altmutter, C., Paro, R., Rehmsmeier, M., 2008. Evolutionary plasticity of polycomb/trithorax response elements in *Drosophila* species. PLoS biology 6, e261.

Heffer, A., Pick, L., 2013. Conservation and variation in Hox genes: how insect models pioneered the evo-devo field. Annu Rev Entomol 58, 161-179.

Heo, J.B., Sung, S., 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331, 76-79.

Herault, Y., Beckers, J., Kondo, T., Fraudeau, N., Duboule, D., 1998. Genetic analysis of a Hoxd-12 regulatory element reveals global versus local modes of controls in the HoxD complex. Development 125, 1669-1677.

Hertwig, O., 1885. Welchen Einfluss übt die Schwerkraft auf die Theilung der Zellen? Jena. Z. Naturwiss 18, 175-205.

Herz, H.M., Mohan, M., Garrett, A.S., Miller, C., Casto, D., Zhang, Y., Seidel, C., Haug, J.S., Florens, L., Washburn, M.P., Yamaguchi, M., Shiekhattar, R., Shilatifard, A., 2012. Polycomb repressive complex 2-dependent and -independent functions of Jarid2 in transcriptional regulation in Drosophila. Molecular and cellular biology 32, 1683-1693.

Herzog, V.A., Lempradl, A., Trupke, J., Okulski, H., Altmutter, C., Ruge, F., Boidol, B., Kubicek, S., Schmauss, G., Aumayr, K., Ruf, M., Pospisilik, A., Dimond, A., Senergin, H.B., Vargas, M.L., Simon, J.A., Ringrose, L., 2014. A strand-specific switch in noncoding transcription switches the function of a Polycomb/Trithorax response element. Nat Genet 46, 973-981.

Hetru, C., Troxler, L., Hoffmann, J.A., 2003. Drosophila melanogaster antimicrobial defense. J Infect Dis 187 Suppl 2, S327-334.

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

Hild, M., Beckmann, B., Haas, S.A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., Hoheisel, J.D., Paro, R., 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. Genome Biol 5, R3.

Hinegardner, R.T., Engelberg, J., 1963. Rationale for a Universal Genetic Code. Science 142, 1083-1085.

Hirota, K., Miyoshi, T., Kugou, K., Hoffman, C.S., Shibata, T., Ohta, K., 2008. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 456, 130-134.

Hogga, I., Karch, F., 2002. Transcription through the iab-7 cis-regulatory domain of the bithorax complex interferes with maintenance of Polycomb-mediated silencing. Development 129, 4915-4922.

Holland, P.W., 2013. Evolution of homeobox genes. Wiley Interdiscip Rev Dev Biol 2, 31-45.

Hollander, W.F., 1937. Bithorax Alleles Dros. Inf. Serv. 8, 77.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57.

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., Rinn, J.L., 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142, 409-419.

Hulskamp, M., Pfeifle, C., Tautz, D., 1990. A morphogenetic gradient of hunchback protein organizes the expression of the gap genes Kruppel and knirps in the early Drosophila embryo. Nature 346, 577-580.

Hurles, M., 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol 2, E206.

Hurst, L.D., Pal, C., Lercher, M.J., 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5, 299-310.

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

Ichimiya, T., Manya, H., Ohmae, Y., Yoshida, H., Takahashi, K., Ueda, R., Endo, T., Nishihara, S., 2004. The twisted abdomen phenotype of Drosophila POMT1 and POMT2 mutants coincides with their heterophilic protein O-mannosyltransferase activity. J Biol Chem 279, 42638-42647.

Ingham, P.W., Ish-Horowicz, D., Howard, K.R., 1986. Correlative changes in homoeotic and segmentation gene expression in Kruppel mutant embryos of Drosophila. Embo J 5, 1659-1665.

Ingham, P.W.a.W., R., 1980. *Trithorax:* A new homeotic mutation of *Drosophila melanogaster* causing transformations of abdominal and thoracic imaginal segments. Molec. gen. Genet 179, 607-614.

Ingram, V.M., 1956. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 178, 792-794.

Ingram, V.M., 1957. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180, 326-328.

Irish, V.F., Martinez-Arias, A., Akam, M., 1989. Spatial regulation of the Antennapedia and Ultrabithorax homeotic genes during Drosophila early development. Embo J 8, 1527-1537.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S.M., Wu, Y.M., Robinson, D.R., Beer, D.G., Feng, F.Y., Iyer, H.K., Chinnaiyan, A.M., 2015. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 47, 199-208.

Jain, A.K.a.D., R. C., 1988. Algorithms for Clustering Data. Prentice Hall, Inc., New Jersey.

Ji, Z., Song, R., Regev, A., Struhl, K., 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife 4, e08890.

Jones, C.A., Ng, J., Peterson, A.J., Morgan, K., Simon, J., Jones, R.S., 1998. The Drosophila esc and E(z) proteins are direct partners in polycomb group-mediated repression. Molecular and cellular biology 18, 2825-2834.

Jurgens, G., 1985. A group of genes controlling the spatial expression of the bithorax complex in *Drosophila*. Letters to Nature 316.

Kal, A.J., Mahmoudi, T., Zak, N.B., Verrijzer, C.P., 2000. The Drosophila brahma complex is an essential coactivator for the trithorax group protein zeste. Genes & development 14, 1058-1071.

Kalfayan, L., Wensink, P.C., 1982. Developmental regulation of Drosophila alpha-tubulin genes. Cell 29, 91-98.

Kalthoff, K., 1971. Photoreversion of UV induction of the malformation "double abdomen" in the egg of Smittia spec. (Diptera, Chironomidae). Dev Biol 25, 119-132.

Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M., Lewis, E.B., 1985. The abdominal region of the bithorax complex. Cell 43, 81-96.

Kassis, J.A., 2002. Pairing-sensitive silencing, polycomb group response elements, and transposon homing in Drosophila. Advances in genetics 46, 421-438.

Kassis, J.A., Muller, J., 2015. Transcription through Polycomb response elements does not induce a switch from repression to activation. Proceedings of the National Academy of Sciences of the United States of America 112, 14755-14756.

Kassis, J.A., Noll, E., VanSickle, E.P., Odenwald, W.F., Perrimon, N., 1992. Altering the insertional specificity of a Drosophila transposable element. Proceedings of the National Academy of Sciences of the United States of America 89, 1919-1923.

Kelley, R.L., Lee, O.K., Shim, Y.K., 2008. Transcription rate of noncoding roX1 RNA controls local spreading of the Drosophila MSL chromatin remodeling complex. Mech Dev 125, 1009-1019.

Kennison, J.A., Tamkun, J.W., 1988. Dosage-dependent modifiers of polycomb and antennapedia mutations in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 85, 8136-8140.

Keravala, A., Calos, M.P., 2008. Site-specific chromosomal integration mediated by phiC31 integrase. Methods Mol Biol 435, 165-173.

Kessel, M., Gruss, P., 1991. Homeotic transformations of murine vertebrae and concomitant alteration of Hox codes induced by retinoic acid. Cell 67, 89-104.

Ketel, C.S., Andersen, E.F., Vargas, M.L., Suh, J., Strome, S., Simon, J.A., 2005. Subunit contributions to histone methyltransferase activities of fly and worm polycomb group complexes. Molecular and cellular biology 25, 6857-6868.

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., 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, 11667-11672.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Kim, D.H., Doyle, M.R., Sung, S., Amasino, R.M., 2009. Vernalization: winter and the timing of flowering in plants. Annu Rev Cell Dev Biol 25, 277-299.

Kim, J., Sebring, A., Esch, J.J., Kraus, M.E., Vorwerk, K., Magee, J., Carroll, S.B., 1996. Integration of positional signals and regulation of wing formation and identity by Drosophila vestigial gene. Nature 382, 133-138.

Kim, T., Xu, Z., Clauder-Munster, S., Steinmetz, L.M., Buratowski, S., 2012. Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. Cell 150, 1158-1169.

King, I.F.K., Emmons, R. B., Francis, N. J., Wild, B., Müller, J., Kingston, R. E., and Wu, C., 2005. Analysis of a Polycomb Group Protein Defines Regions That Link Repressive Activity on Nucleosomal Templates to In Vivo Function. Mol. Cell. Biol. 25, 6578-6591.

King, M.C., Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. Science 188, 107-116.

King, N., 2004. The unicellular ancestry of animal development. Developmental Cell 7, 313-325.

Klymenko, T., Muller, J., 2004. The histone methyltransferases Trithorax and Ash1 prevent transcriptional silencing by Polycomb group proteins. EMBO Rep 5, 373-377.

Klymenko, T., Papp, B., Fischle, W., Kocher, T., Schelder, M., Fritsch, C., Wild, B., Wilm, M., Muller, J., 2006. A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. Genes & Development 20, 1110-1122.

Knipple, D.C., Seifert, E., Rosenberg, U.B., Preiss, A., Jackle, H., 1985. Spatial and temporal patterns of Kruppel gene expression in early Drosophila embryos. Nature 317, 40-44.

Knust, E., Tietze, K., Campos-Ortega, J.A., 1987. Molecular analysis of the neurogenic locus Enhancer of split of Drosophila melanogaster. Embo J 6, 4113-4123.

Koch, E.A., Spitzer, R.H., 1983. Multiple effects of colchicine on oogenesis in Drosophila: induced sterility and switch of potential oocyte to nurse-cell developmental pathway. Cell Tissue Res 228, 21-32.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., Gao, G., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic acids research 35, W345-349.

Koonin, E.V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61, 99-111.

Kopp, A., Blackman, R.K., Duncan, I., 1999. Wingless, decapentaplegic and EGF receptor signaling pathways interact to specify dorso-ventral pattern in the adult abdomen of Drosophila. Development 126, 3495-3507.

Kornberg, T.B., Tabata, T., 1993. Segmentation of the Drosophila embryo. Curr Opin Genet Dev 3, 585-594.

Kornienko, A.E., Guenzl, P.M., Barlow, D.P., Pauler, F.M., 2013. Gene regulation by the act of long non-coding RNA transcription. BMC Biol 11, 59.

Kostyuchenko, M., Savitskaya, E., Koryagina, E., Melnikova, L., Karakozova, M., Georgiev, P., 2009. Zeste can facilitate long-range enhancer-promoter communication and insulator bypass in Drosophila melanogaster. Chromosoma 118, 665-674.

Koziol, M.J., Rinn, J.L., 2010. RNA traffic control of chromatin complexes. Curr Opin Genet Dev 20, 142-148.

Kozma, G., Bender, W., Sipos, L., 2008. Replacement of a Drosophila Polycomb response element core, and in situ analysis of its DNA motifs. Mol Genet Genomics 279, 595-603.

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., Khavari, P.A., 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature 493, 231-235.

Kugler, S.J., Nagel, A.C., 2010. A novel Pzg-NURF complex regulates Notch target gene activity. Mol Biol Cell 21, 3443-3448.

Kumar, J.P., Moses, K., 2001a. EGF receptor and Notch signaling act upstream of Eyeless/Pax6 to control eye specification. Cell 104, 687-697.

Kumar, J.P., Moses, K., 2001b. The EGF receptor and notch signaling pathways control the initiation of the morphogenetic furrow during Drosophila eye development. Development 128, 2689-2697.

Kumar, S.R., Patel, H., Tomlinson, A., 2015. Wingless mediated apoptosis: How cone cells direct the death of peripheral ommatidia in the developing Drosophila eye. Dev Biol 407, 183-194.

Kung, J.T., Colognori, D., Lee, J.T., 2013. Long noncoding RNAs: past, present, and future. Genetics 193, 651-669.

Kung, J.T., Lee, J.T., 2013. RNA in the loop. Dev Cell 24, 565-567.

Kurdistani, S.K., Grunstein, M., 2003. Histone acetylation and deacetylation in yeast. Nature reviews. Molecular cell biology 4, 276-284.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D., 2002. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev. 16, 2893-2905.

Lai, E.C., Bodner, R., Posakony, J.W., 2000. The enhancer of split complex of Drosophila includes four Notch-regulated members of the bearded gene family. Development 127, 3441-3455.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., Shiekhattar, R., 2013. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature 494, 497-501.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A.,

Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing, C., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Larson, M.H., Gilbert, L.A., Wang, X., Lim, W.A., Weissman, J.S., Qi, L.S., 2013. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc 8, 2180-2196.

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., Barlow, D.P., 2012. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science 338, 1469-1472.

Lawrence, P.A., Struhl, G., 1996. Morphogens, compartments, and pattern: lessons from drosophila? Cell 85, 951-961.

Lecuit, T., Cohen, S.M., 1998. Dpp receptor levels contribute to shaping the Dpp morphogen gradient in the Drosophila wing imaginal disc. Development 125, 4901-4907.

Lecuit, T., Lenne, P.F., 2007. Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis. Nature reviews. Molecular cell biology 8, 633-644.

Lee, J.M., Sonnhammer, E.L., 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome research 13, 875-882.

Lehmann, M., Siegmund, T., Lintermann, K.G., Korge, G., 1998. The pipsqueak protein of Drosophila melanogaster binds to GAGA sequences through a novel DNA-binding domain. J Biol Chem 273, 28504-28509.

Lemons, D., McGinnis, W., 2006. Genomic evolution of Hox gene clusters. Science 313, 1918-1922.

Lempradl, A., Ringrose, L., 2008. How does noncoding transcription regulate Hox genes? BioEssays : news and reviews in molecular, cellular and developmental biology 30, 110-121.

Lennox, K.A., Behlke, M.A., 2016. Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. Nucleic acids research 44, 863-877.

Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-Cabeza, J., Garibal, M.A., Koch, F., Maqbool, M.A., Fenouil, R., Loriod, B., Holota, H., Gut, M., Gut, I., Imbert, J., Andrau, J.C., Puthier, D., Spicuglia, S., 2013. Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics 14, 914.

Levine, M., 2010. Transcriptional enhancers in animal development and evolution. Curr Biol 20, R754-763.

Levine, M., Tjian, R., 2003. Transcription regulation and animal diversity. Nature 424, 147-151.

Lewis, E.B., 1978. A gene complex controlling segmentation in *Drosophila*. Nature 276, 565-570.

Lewis, E.B., 1985. Regulation of the genes of the bithorax complex in Drosophila. Cold Spring Harb Symp Quant Biol 50, 155-164.

Li, G., Margueron, R., Ku, M., Chambon, P., Bernstein, B.E., Reinberg, D., 2010. Jarid2 and PRC2, partners in regulating gene expression. Genes Dev 24, 368-380.

Lin, M.F., Jungreis, I., Kellis, M., 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27, i275-282.

Lipshitz, H.D., Peattie, D.A., Hogness, D.S., 1987. Novel transcripts from the Ultrabithorax domain of the bithorax complex. Genes Dev 1, 307-322.

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

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., Richmond, T.J., 1997a. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389, 251-260.

Luger, K., Rechsteiner, T.J., Flaus, A.J., Waye, M.M., Richmond, T.J., 1997b. Characterization of nucleosome core particles containing histone proteins made in bacteria. J Mol Biol 272, 301-311.

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., Shen, X., 2016. Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. Cell Stem Cell 18, 637-652.

Lyalin, D., Koles, K., Roosendaal, S.D., Repnikova, E., Van Wechel, L., Panin, V.M., 2006. The twisted gene encodes Drosophila protein O-mannosyltransferase 2 and genetically interacts with the rotated abdomen gene encoding Drosophila protein O-mannosyltransferase 1. Genetics 172, 343-353.

Lyulcheva, E., Taylor, E., Michael, M., Vehlow, A., Tan, S., Fletcher, A., Krause, M., Bennett, D., 2008. Drosophila pico and its mammalian ortholog lamellipodin activate serum response factor and promote cell proliferation. Dev Cell 15, 680-690.

Ma, L., Bajic, V.B., Zhang, Z., 2013. On the classification of long non-coding RNAs. RNA Biol 10, 925-933.

Maeda, R.K., Karch, F., 2006. The ABC of the BX-C: the bithorax complex explained. Development 133, 1413-1422.

Maeda, R.K., Karch, F., 2011. Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. Curr Opin Genet Dev 21, 187-193.

Maitre, J.L., Heisenberg, C.P., 2013. Three functions of cadherins in cell adhesion. Curr Biol 23, R626-633.

Mallo, M., Alonso, C.R., 2013. The regulation of Hox gene expression during animal development. Development 140, 3951-3963.

Mammoto, T., Ingber, D.E., 2010. Mechanical control of tissue and organ development. Development 137, 1407-1420.

Mann, R.S., Chan, S.K., 1996. Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. Trends Genet 12, 258-262.

Marin-Bejar, O.a.H., M., 2015. Long noncoding RNAs: from identification to functions and mechanisms. Advances in Genomics and Genetics 2015:5, 257-274.

Markussen, F.H., Michon, A.M., Breitwieser, W., Ephrussi, A., 1995. Translational control of oskar generates short OSK, the isoform that induces pole plasma assembly. Development 121, 3723-3732.

Marquez, R.M., Singer, M.A., Takaesu, N.T., Waldrip, W.R., Kraytsberg, Y., Newfeld, S.J., 2001. Transgenic analysis of the Smad family of TGF-beta signal transducers in Drosophila melanogaster suggests new roles and new interactions between family members. Genetics 157, 1639-1648.

Marsh, D.J., Shah, J.S., Cole, A.J., 2014. Histones and their modifications in ovarian cancer - drivers of disease and therapeutic targets. Front Oncol 4, 144.

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., Akoulitchev, A., 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445, 666-670.

Martin, C.H., Mayeda, C.A., Davis, C.A., Ericsson, C.L., Knafels, J.D., Mathog, D.R., Celniker, S.E., Lewis, E.B., Palazzolo, M.J., 1995. Complete sequence of the bithorax complex of Drosophila. Proceedings of the National Academy of Sciences of the United States of America 92, 8398-8402.

Martinez-Arias, A., Lawrence, P.A., 1985. Parasegments and compartments in the Drosophila embryo. Nature 313, 639-642.

Matthews, K.A., Miller, D.F., Kaufman, T.C., 1989. Developmental distribution of RNA and protein products of the Drosophila alpha-tubulin gene family. Dev Biol 132, 45-61.

Mattick, J.S., Makunin, I.V., 2006. Non-coding RNA. Hum Mol Genet 15 Spec No 1, R17-29.

Maurange, C., and Paro, R., 2002. A cellular memory module conveys epigenetic inheritance of hedgehog expression during Drosophila wing imaginal disc development. Genes. Dev. 16, 2672-2683.

McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A., and Gehring, W. J., 1984a. A homologous protein-coding sequence in *Drosophila* homoetic genes and its conservation in other metazoans. Cell 37, 403-408.

McGinnis, W., Krumlauf, R., 1992. Homeobox genes and axial patterning. Cell 68, 283-302.

McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A., and Gehring, W. J. , 1984b. A conserved DNA sequence in homeotic genes of the *Drosophila* Antennapedia and bithorax complexes. Nature Genetics 308, 428-433.

Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., de Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C., Cruz-Morales, P., Duddela, S., Dusterhus, S., Edwards, D.J., Fewer, D.P., Garg, N., Geiger, C., Gomez-Escribano, J.P., Greule, A., Hadjithomas, M., Haines, A.S., Helfrich, E.J., Hillwig, M.L., Ishida, K., Jones, A.C., Jones, C.S., Jungmann, K., Kegler, C., Kim, H.U., Kotter, P., Krug, D., Masschelein, J., Melnik, A.V., Mantovani, S.M., Monroe, E.A., Moore, M., Moss, N., Nutzmann, H.W., Pan, G., Pati, A., Petras, D., Reen, F.J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N.J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A.K., Balibar, C.J., Balskus, E.P., Barona-Gomez, F., Bechthold, A., Bode, H.B., Borriss, R., Brady, S.F., Brakhage, A.A., Caffrey, P., Cheng, Y.Q., Clardy, J., Cox, R.J., De Mot, R., Donadio, S., Donia, M.S., van der Donk, W.A., Dorrestein, P.C., Doyle, S., Driessen, A.J., Ehling-Schulz, M., Entian, K.D., Fischbach, M.A., Gerwick, L., Gerwick, W.H., Gross, H., Gust, B., Hertweck, C., Hofte, M., Jensen, S.E., Ju, J., Katz, L., Kaysser, L., Klassen, J.L., Keller, N.P., Kormanec, J., Kuipers, O.P., Kuzuyama, T., Kyrpides, N.C., Kwon, H.J., Lautru, S., Lavigne, R., Lee, C.Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Mendez, C., Metsa-Ketela, M., Micklefield, J., Mitchell, D.A., Moore, B.S., Moreira, L.M., Muller, R., Neilan, B.A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M.S., Ostash, B., Payne, S.M., Pernodet, J.L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J.M., Salas, J.A., Schmitt, E.K., Scott, B., Seipke, R.F., Shen, B., Sherman, D.H., Sivonen, K., Smanski, M.J., Sosio, M., Stegmann, E., Sussmuth, R.D., Tahlan, K., Thomas, C.M., Tang, Y., Truman, A.W., Viaud, M., Walton, J.D., Walsh, C.T., Weber, T., van Wezel, G.P., Wilkinson, B., Willey, J.M., Wohlleben, W., Wright, G.D., Ziemert, N., Zhang, C., Zotchev, S.B., Breitling, R., Takano, E., Glockner, F.O., 2015. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 11, 625-631.

Meller, V.H., Gordadze, P.R., Park, Y., Chu, X., Stuckenholz, C., Kelley, R.L., Kuroda, M.I., 2000. Ordered assembly of roX RNAs into MSL complexes on the dosage-compensated X chromosome in Drosophila. Curr Biol 10, 136-143.

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

Mendenhall, E.M., Bernstein, B.E., 2008. Chromatin state maps: new technologies, new insights. Curr Opin Genet Dev 18, 109-115.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., Bernstein, B.E., 2010. GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet 6, e1001244.

Mercer, T.R., Dinger, M.E., Mattick, J.S., 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet 10, 155-159.

Messmer, S., Franke, A., and Paro, R., 1992. Analysis of the functional role of the Polycomb chromo domain in Drosophila melanogaster. Genes Dev. 6, 1241-1254.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic acids research 44, D336-342.

Mi, H., Vandergriff, J., Campbell, M., Narechania, A., Majoros, W., Lewis, S., Thomas, P.D., Ashburner, M., 2003. Assessment of genome-wide protein function classification for Drosophila melanogaster. Genome research 13, 2118-2128.

Mihaly, J., Mishra, R.K., Karch, F., 1998. A conserved sequence motif in Polycomb-response elements. Mol Cell 1, 1065-1066.

Miko, I., 2008. Gregor Mendel and the Principles of Inheritance. Nature Education 1, 134.

Miller, T., Krogan, N.J., Dover, J., Erdjument-Bromage, H., Tempst, P., Johnston, M., Greenblatt, J.F., Shilatifard, A., 2001. COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. Proceedings of the National Academy of Sciences of the United States of America 98, 12902-12907.

Milligan, M.J., Lipovich, L., 2014. Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet 5, 476.

Minks, J., Baldry, S.E., Yang, C., Cotton, A.M., Brown, C.J., 2013. XIST-induced silencing of flanking genes is achieved by additive action of repeat a monomers in human somatic cells. Epigenetics Chromatin 6, 23.

Mis, J., Ner, S.S., Grigliatti, T.A., 2006. Identification of three histone methyltransferases in Drosophila: dG9a is a suppressor of PEV and is required for gene silencing. Mol Genet Genomics 275, 513-526.

Mishra, K., Chopra, V.S., Srinivasan, A., Mishra, R.K., 2003. Trl-GAGA directly interacts with lola like and both are part of the repressive complex of Polycomb group of genes. Mech. Dev. 120, 681-689.

Mishra, R.K., Mihaly, J., Barges, S., Spierer, A., Karch, F., Hagstrom, K., Schweinsberg, S.E., Schedl, P., 2001. The iab-7 polycomb response element maps to a nucleosome-free region of chromatin and requires both GAGA and pleiohomeotic for silencing activity. Molecular and cellular biology 21, 1311-1318.

Moazed, D., O'Farrell, P.H., 1992. Maintenance of the engrailed expression pattern by Polycomb group genes in Drosophila. Development 116, 805-810.

mod, E.C., Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., Washietl, S., Arshinoff, B.I., Ay, F., Meyer, P.E., Robine, N., Washington, N.L., Di Stefano, L., Berezikov, E., Brown, C.D., Candeias, R., Carlson, J.W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M.Y., Will, S., Alekseyenko, A.A., Artieri, C., Booth, B.W., Brooks, A.N., Dai, Q., Davis, C.A., Duff, M.O., Feng, X., Gorchakov, A.A., Gu, T., Henikoff, J.G., Kapranov, P., Li, R., MacAlpine, H.K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S.K., Riddle, N.C., Sakai, A., Samsonova, A., Sandler, J.E., Schwartz, Y.B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K.H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S.E., Brent, M.R., Cherbas, L., Elgin, S.C., Gingeras, T.R., Grossman, R., Hoskins, R.A., Kaufman, T.C., Kent, W., Kuroda, M.I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J.W., Ren, B., Russell, S., Cherbas, P., Graveley, B.R., Lewis, S., Micklem, G., Oliver, B., Park, P.J., Celniker, S.E.,

Henikoff, S., Karpen, G.H., Lai, E.C., MacAlpine, D.M., Stein, L.D., White, K.P., Kellis, M., 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787-1797.

Mohan, M., Herz, H.M., Smith, E.R., Zhang, Y., Jackson, J., Washburn, M.P., Florens, L., Eissenberg, J.C., Shilatifard, A., 2011. The COMPASS family of H3K4 methylases in Drosophila. Molecular and cellular biology 31, 4310-4318.

Mohd-Sarip, A., Venturini, F., Chalkley, G.E., Verrijzer, C.P., 2002. Pleiohomeotic can link polycomb to DNA and mediate transcriptional repression. Molecular and cellular biology 22, 7473-7483.

Molnar, C., Casado, M., Lopez-Varea, A., Cruz, C., de Celis, J.F., 2012. Genetic annotation of gain-of-function screens using RNA interference and in situ hybridization of candidate genes in the Drosophila wing. Genetics 192, 741-752.

Molnar, C., Lopez-Varea, A., Hernandez, R., de Celis, J.F., 2006. A gain-of-function screen identifying genes required for vein formation in the Drosophila melanogaster wing. Genetics 174, 1635-1659.

Mondal, T., Kanduri, C., 2013. Maintenance of epigenetic information: a noncoding RNA perspective. Chromosome Res 21, 615-625.

Morata, G., 2001. How Drosophila appendages develop. Nature reviews. Molecular cell biology 2, 89-97.

Morata, G., Kerridge, S., 1982. The role of position in determining homoeotic gene function in Drosophila. Nature 300, 191-192.

Morata, G., Lawrence, P.A., 1975. Control of compartment development by the engrailed gene in Drosophila. Nature 255, 614-617.

Morata, G., Sanchez-Herrero, E., 1998. Developmental biology. Pulling the fly's leg. Nature 392, 657-658.

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

Morris, K.V., Mattick, J.S., 2014. The rise of regulatory RNA. Nat Rev Genet 15, 423-437.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628.

Mukherjee, A., Shan, X., Mutsuddi, M., Ma, Y., Nambu, J.R., 2000. The Drosophila sox gene, fish-hook, is required for postembryonic development. Dev Biol 217, 91-106.

Muller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., Simon, J.A., 2002. Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell 111, 197-208.

Muller, J., Kassis, J.A., 2006. Polycomb response elements and targeting of Polycomb group proteins in *Drosophila*. Curr Opin Genet Dev 16, 476-484.

Mutz, K.O., Heilkenbrinker, A., Lonne, M., Walter, J.G., Stahl, F., 2013. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol 24, 22-30.

Natzle, J.E., McCarthy, B.J., 1984. Regulation of Drosophila alpha- and beta-tubulin genes during development. Dev Biol 104, 187-198.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., Finn, R.D., 2015. Rfam 12.0: updates to the RNA families database. Nucleic acids research 43, D130-137.

Negre, B., Casillas, S., Suzanne, M., Sanchez-Herrero, E., Akam, M., Nefedov, M., Barbadilla, A., de Jong, P., Ruiz, A., 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. Genome research 15, 692-700.

Negre, B., Ranz, J.M., Casals, F., Caceres, M., Ruiz, A., 2003. A new split of the Hox gene complex in Drosophila: relocation and evolution of the gene labial. Mol Biol Evol 20, 2042-2054.

Negre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R.F., Chen, J., Hwang, L., Cheng, C., Auburn, R.P., Davis, M.B., Domanus, M., Shah, P.K., Morrison, C.A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N.A., Grundstad, A.J., Hanley, D., MacAlpine, D.M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R.L., Ren, B., Posakony, J.W., Kellis, M., White, K.P., 2011. A cis-regulatory map of the Drosophila genome. Nature 471, 527-531.

Negre, N., Brown, C.D., Shah, P.K., Kheradpour, P., Morrison, C.A., Henikoff, J.G., Feng, X., Ahmad, K., Russell, S., White, R.A., Stein, L., Henikoff, S., Kellis, M., White, K.P., 2010. A comprehensive map of insulator elements for the Drosophila genome. PLoS Genet 6, e1000814.

Negre, N., Hennetin, J., Sun, L.V., Lavrov, S., Bellis, M., White, K.P., Cavalli, G., 2006. Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biol 4, e170.

Nekrasov, M., Klymenko, T., Fraterman, S., Papp, B., Oktaba, K., Kocher, T., Cohen, A., Stunnenberg, H.G., Wilm, M., Muller, J., 2007. Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. Embo J 26, 4078-4088.

Nekrasov, M., Wild, B., and Mülller, J., 2005. Nucleosome binding and histone methyltransferase activity of Drosophila PRC2. EMBO 6, 348-353.

Neufeld, T.P., de la Cruz, A.F., Johnston, L.A., Edgar, B.A., 1998. Coordination of growth and cell division in the Drosophila wing. Cell 93, 1183-1193.

Nijhout, H.F., Riddiford, L.M., Mirth, C., Shingleton, A.W., Suzuki, Y., Callier, V., 2014. The developmental control of size in insects. Wiley Interdiscip Rev Dev Biol 3, 113-134.

Ninova, M., Ronshaugen, M., Griffiths-Jones, S., 2014. Conserved temporal patterns of microRNA expression in Drosophila support a developmental hourglass model. Genome biology and evolution 6, 2459-2467.

Nolte, C., Alexander, T. B., and Krumlauf, R., 2015. Mammalian Embryo: Hox Genes. John Wiley & Sons, Ltd.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., Wolfe, S.A., 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133, 1277-1289.

Nusslein-Volhard, C., Wieschaus, E., 1980. Mutations affecting segment number and polarity in Drosophila. Nature 287, 795-801.

Ohno, S., 1972. So much "junk" DNA in our genome. Brookhaven Symp Biol 23, 366-370.

Okulski, H., Druck, B., Bhalerao, S., Ringrose, L., 2011. Quantitative analysis of polycomb response elements (PREs) at identical genomic locations distinguishes contributions of PRE sequence and genomic environment. Epigenetics Chromatin 4, 4.

Pallavi, S.K., Kannan, R., Shashidhara, L.S., 2006. Negative regulation of Egfr/Ras pathway by Ultrabithorax during haltere development in Drosophila. Dev Biol 296, 340-352.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., Kanduri, C., 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32, 232-246.

Panganiban, G., Irvine, S.M., Lowe, C., Roehl, H., Corley, L.S., Sherbon, B., Grenier, J.K., Fallon, J.F., Kimble, J., Walker, M., Wray, G.A., Swalla, B.J., Martindale, M.Q., Carroll, S.B., 1997. The origin and evolution of animal appendages. Proceedings of the National Academy of Sciences of the United States of America 94, 5162-5166.

Pankratz, M.J., Jackle, H., 1990. Making stripes in the Drosophila embryo. Trends Genet 6, 287-292.

Papp, B., Muller, J., 2006. Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. Genes & Development 20, 2041-2054.

Park, J.M., Gim, B.S., Kim, J.M., Yoon, J.H., Kim, H.S., Kang, J.G., Kim, Y.J., 2001. Drosophila Mediator complex is broadly utilized by diverse gene-specific transcription factors at different types of core promoters. Molecular and cellular biology 21, 2312-2323.

Pauli, A., Rinn, J.L., Schier, A.F., 2011. Non-coding RNAs as regulators of embryogenesis. Nat Rev Genet 12, 136-149.

Peabody, D.S., 1993. The RNA binding site of bacteriophage MS2 coat protein. Embo J 12, 595-600.

Pearson, J.C., Lemons, D., Mcginnis, W., 2005. Modulating Hox gene functions during animal body patterning. Nature Reviews Genetics 6, 893-904.

Pease, B., Borges, A.C., Bender, W., 2013. Noncoding RNAs of the Ultrabithorax domain of the Drosophila bithorax complex. Genetics 195, 1253-1264.

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., Haussler, D., 2006. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2, e33.

Peel, A.D., Chipman, A.D., Akam, M., 2005. Arthropod segmentation: beyond the Drosophila paradigm. Nat Rev Genet 6, 905-916.

Pek, J.W., Osman, I., Tay, M.L., Zheng, R.T., 2015. Stable intronic sequence RNAs have possible regulatory roles in Drosophila melanogaster. J Cell Biol 211, 243-251.

Pertea, M., 2012. The human transcriptome: an unfinished story. Genes (Basel) 3, 344-360.

Pertea, M., Salzberg, S.L., 2010. Between a chicken and a grape: estimating the number of human genes. Genome Biol 11, 206.

Peterson, A.J., Kyba, M., Bornemann, D., Morgan, K., Brock, H.W., Simon, J., 1997. A domain shared by the Polycomb group proteins Scm and ph mediates heterotypic and homotypic interactions. Molecular and cellular biology 17, 6683-6692.

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 noncod- ing RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. Cell 127, 1209-1221.

Petruk, S., Sedkov, Y., Smith, S., Tillib, S., Kraevski, V., Nakamura, T., Canaani, E., Croce, C. M., and Mazo, A., 2001. Trithorax and dCBP acting in a complex to maintain expression of a homeotic gene. Science 294, 1331-1334.

Pettini, T., 2012. The role of novel long non-coding RNAs in Hox gene regulation, FLS. University of Manchester, Manchester.

Pick, L., Heffer, A., 2012. Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. Annals of the New York Academy of Sciences 1256, 15-32.

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

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., Carter, D.R., 2011. Pseudogenes: pseudo-functional or key regulators in health and disease? Rna 17, 792-798.

Pirrotta, V., Li, H.B., 2012. A view of nuclear Polycomb bodies. Curr Opin Genet Dev 22, 101-109.

Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., Panning, B., 2002. Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 36, 233-278.

Plaza, S., Prince, F., Jaeger, J., Kloter, U., Flister, S., Benassayag, C., Cribbs, D., Gehring, W.J., 2001. Molecular basis for the inhibition of Drosophila eye development by Antennapedia. Embo J 20, 802-811.

Pokrywka, N.J., 1995. RNA localization and the cytoskeleton in Drosophila oocytes. Curr Top Dev Biol 31, 139-166.

Port, F., Chen, H.M., Lee, T., Bullock, S.L., 2014. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 111, E2967-2976.

Prasanth, K.V., Spector, D.L., 2007. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes Dev 21, 11-42.

Preiss, A., Rosenberg, U.B., Kienlin, A., Seifert, E., Jackle, H., 1985. Molecular genetics of Kruppel, a gene required for segmentation of the Drosophila embryo. Nature 313, 27-32.

Quek, X.C., Thomson, D.W., Maag, J.L., Bartonicek, N., Signal, B., Clark, M.B., Gloss, B.S., Dinger, M.E., 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic acids research 43, D168-173.

Quinn, J.J., Chang, H.Y., 2016. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet 17, 47-62.

Quinonez, S.C., Innis, J.W., 2014. Human HOX gene disorders. Mol Genet Metab 111, 4-15.

Rank, G., Prestel, M., Paro, R., 2002. Transcription through intergenic chromosomal memory elements of the *Drosophila* bithorax complex correlates with an epigenetic switch. Molecular and cellular biology 22, 8026-8034.

Ranz, J.M., Casals, F., Ruiz, A., 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome research 11, 230-239.

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., Mattick, J.S., 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome research 16, 11-19.

Ray, P., De, S., Mitra, A., Bezstarosti, K., Demmers, J.A., Pfeifer, K., Kassis, J.A., 2016. Combgap contributes to recruitment of Polycomb group proteins in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 113, 3826-3831.

Reed, H.C., Hoare, T., Thomsen, S., Weaver, T.A., White, R.A., Akam, M., Alonso, C.R., 2010. Alternative splicing modulates Ubx protein function in Drosophila melanogaster. Genetics 184, 745-758.

Reference Genome Group of the Gene Ontology, C., 2009. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput Biol 5, e1000431.

Ringrose, L., 2007. Polycomb comes of age: genome-wide profiling of target sites. Curr Opin Cell Biol 19, 290-297.

Rinn, J.L., Chang, H.Y., 2012. Genome regulation by long noncoding RNAs. Annu Rev Biochem 81, 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., Chang, H.Y., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311-1323.

Roch, F., Akam, M., 2000. Ultrabithorax and the control of cell morphology in Drosophila halteres. Development 127, 97-107.

Rogulja, D., Irvine, K.D., 2005. Regulation of cell proliferation by a morphogen gradient. Cell 123, 449-461.

Rongo, C., Gavis, E.R., Lehmann, R., 1995. Localization of oskar RNA regulates oskar translation and requires Oskar protein. Development 121, 2737-2746.

Ronshaugen, M., Biemar, F., Piel, J., Levine, M., Lai, E.C., 2005. The *Drosophila* microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. Genes & Development 19, 2947-2952.

Ronshaugen, M., McGinnis, N., McGinnis, W., 2002. Hox protein mutation and macroevolution of the insect body plan. Nature 415, 914-917.

Rorth, P., 1996. A modular misexpression screen in Drosophila detecting tissue-specific phenotypes. Proceedings of the National Academy of Sciences of the United States of America 93, 12418-12422.

Rosa, S., De Lucia, F., Mylne, J.S., Zhu, D., Ohmido, N., Pendle, A., Kato, N., Shaw, P., Dean, C., 2013. Physical clustering of FLC alleles during Polycomb-mediated epigenetic silencing in vernalization. Genes Dev 27, 1845-1850.

Roth, S.Y.a.A., C. D. , 1996. Histone acetylation and chromatin assembly: a single escort, multiple dances? Cell 87, 5-8.

Rozenblatt-Rosen, O., Rozovskaia, T., Burakov, D., Sedkov, Y., Tillib, S., Blechman, J., Nakamura, T., Croce, C.M., Mazo, A., Canaani, E., 1998. The C-terminal SET domains of ALL-1 and TRITHORAX interact with the INI1 and SNR1 proteins, components of the SWI/SNF complex. Proceedings of the National Academy of Sciences of the United States of America 95, 4152-4157.

Rozovskaia, T., Tillib, S., Smith, S., Sedkov, Y., Rozenblatt-Rosen, O., Petruk, S., Yano, T., Nakamura, T., Ben-Simchon, L., Gildea, J., Croce, C.M., Shearn, A., Canaani, E., Mazo, A., 1999. Trithorax and ASH1 interact directly and associate with the trithorax group-responsive bxd region of the Ultrabithorax promoter. Molecular and cellular biology 19, 6441-6447.

Rozowski, M., Akam, M., 2002. Hox gene control of segment-specific bristle patterns in Drosophila. Genes Dev 16, 1150-1162.

Ryoo, H.D., Mann, R.S., 1999. The control of trunk Hox specificity and activity by Extradenticle. Genes Dev 13, 1704-1716.

Ryoo, H.D., Marty, T., Casares, F., Affolter, M., Mann, R.S., 1999. Regulation of Hox target genes by a DNA bound Homothorax/Hox/Extradenticle complex. Development 126, 5137-5148.

Salamov, A.A., Solovyev, V.V., 2000. Ab initio gene finding in Drosophila genomic DNA. Genome research 10, 516-522.

Sanchez, L., Casares, N., Gorfinkiel, I., Guerrero, 1997. The genital disc of *Drosophila melanogaster:* 2. Role of the genes *hedgehog, decapentaplegic,* and *wingless*. Dev. Genes Evol 207, 229-241.

Sanchez-Elsner, T., Gou, D., Kremmer, E., Sauer, F., 2006. Noncoding RNAs of trithorax response elements recruit *Drosophila* Ash1 to Ultrabithorax. Science 311, 1118-1123.

Sanchez-Herrero, E., 1991. Control of the expression of the bithorax complex genes abdominal-A and abdominal-B by cis-regulatory regions in Drosophila embryos. Development 111, 437-449.

Sanchez-Herrero, E., Akam, M., 1989. Spatially ordered transcription of regulatory DNA in the bithorax complex of Drosophila. Development 107, 321-329.

Sander, K., 1975. Pattern specification in the insect embryo. Ciba Found Symp 0, 241-263.

Santiveri, C.M., Lechtenberg, B.C., Allen, M.D., Sathyamurthy, A., Jaulent, A.M., Freund, S.M., Bycroft, M., 2008. The malignant brain tumor repeats of human SCML2 bind to peptides containing monomethylated lysine. J Mol Biol 382, 1107-1112.

Sarma, K., Margueron, R., Ivanov, A., Pirrotta, V., Reinberg, D., 2008. Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. Molecular and cellular biology 28, 2718-2731.

Sauer, B., Henderson, N., 1988. Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. Proceedings of the National Academy of Sciences of the United States of America 85, 5166-5170.

Saurin, A.J., Shao, Z., Erdjument-Bromage, H., Tempst, P., Kingston, R.E., 2001. A Drosophila Polycomb group complex includes Zeste and dTAFII proteins. Nature 412, 655-660.

Savla, U., Benes, J., Zhang, J., Jones, R.S., 2008. Recruitment of Drosophila Polycomb-group proteins by Polycomblike, a component of a novel protein complex in larvae. Development 135, 813-817.

Schaaf, C.A., Misulovin, Z., Gause, M., Koenig, A., Dorsett, D., 2013. The Drosophila enhancer of split gene complex: architecture and coordinate regulation by notch, cohesin, and polycomb group proteins. G3 (Bethesda) 3, 1785-1794.

Schattner, P., Brooks, A.N., Lowe, T.M., 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic acids research 33, W686-689.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat Methods 9, 676-682.

Schmitges, F.W., Prusty, A.B., Faty, M., Stutzer, A., Lingaraju, G.M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., Bunker, R.D., Wirth, U., Bouwmeester, T., Bauer, A., Ly-Hartig, N., Zhao, K.,

Chan, H., Gu, J., Gut, H., Fischle, W., Muller, J., Thoma, N.H., 2011. Histone methylation by PRC2 is inhibited by active chromatin marks. Mol Cell 42, 330-341.

Schneider, J., Wood, A., Lee, J.S., Schuster, R., Dueker, J., Maguire, C., Swanson, S.K., Florens, L., Washburn, M.P., Shilatifard, A., 2005. Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. Mol Cell 19, 849-856.

Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D., Jenuwein, T., 2004. A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev 18, 1251-1262.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., Cavalli, G., 2007. Genome regulation by polycomb and trithorax proteins. Cell 128, 735-745.

Schuettengruber, B., Ganapathi, M., Leblanc, B., Portoso, M., Jaschek, R., Tolhuis, B., van Lohuizen, M., Tanay, A., Cavalli, G., 2009. Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol 7, e13.

Schuettengruber, B., Martinez, A.M., Iovino, N., Cavalli, G., 2011. Trithorax group proteins: switching genes on and keeping them active. Nature reviews. Molecular cell biology 12, 799-814.

Schwartz, Y.B., Kahn, T. G., Nix, D. A., Li, X., Bourgon, R., Biggin, M., and Pirrotta, V. , 2006. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nature Genetics 38, 700-705.

Schwartz, Y.B., Pirrotta, V., 2013. A new world of Polycombs: unexpected partnerships and emerging functions. Nat Rev Genet 14, 853-864.

Scott, M.P., Carroll, S.B., 1987. The segmentation and homeotic gene network in early Drosophila development. Cell 51, 689-698.

Seo, H.C., Edvardsen, R.B., Maeland, A.D., Bjordal, M., Jensen, M.F., Hansen, A., Flaat, M., Weissenbach, J., Lehrach, H., Wincker, P., Reinhardt, R., Chourrout, D., 2004. Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. Nature 431, 67-71.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C.T., Bender, W., Kingston, R.E., 1999. Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell 98, 37-46.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., Ren, B., 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116-120.

Shilatifard, A., 2012. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annu Rev Biochem 81, 65-95.

Shilo, B.Z., 2005. Regulating the dynamics of EGF receptor signaling in space and time. Development 132, 4017-4027.

Shimell, M.J., Simon, J., Bender, W., O'Connor, M.B., 1994. Enhancer point mutation results in a homeotic transformation in Drosophila. Science 264, 968-971.

Shippy, T.D., Ronshaugen, M., Cande, J., He, J., Beeman, R.W., Levine, M., Brown, S.J., Denell, R.E., 2008. Analysis of the Tribolium homeotic complex: insights into mechanisms constraining insect Hox clusters. Dev Genes Evol 218, 127-139.

Siepel, A.a.H., D., 2005. Phylogenetic Hidden Markov Models, in: Nielsen, R. (Ed.), Statistical Methods in Molecular Evolution. Springer, New York, pp. 325-351.

Simon, E., Guerrero, I., 2015. The transcription factor optomotor-blind antagonizes Drosophila haltere growth by repressing decapentaplegic and hedgehog targets. PloS one 10, e0121239.

Simon, J., Bornemann, D., Lunde, K., and Schwartz, C., 1995. The extra sex combs product contains WD40 repeats and its time of action implies a role distinct from other Polycomb group products. Mech. Dev. 53, 197-208.

Simon, J., Peifer, M., Bender, W., O'Connor, M., 1990. Regulatory elements of the bithorax complex that control expression along the anterior-posterior axis. Embo J 9, 3945-3956.

Simonatto, M., Barozzi, I., Natoli, G., 2013. Non-coding transcription at cis-regulatory elements: computational and experimental approaches. Methods 63, 66-75.

Singh, N.P., Mishra, R.K., 2014. Role of abd-A and Abd-B in development of abdominal epithelia breaks posterior prevalence rule. PLoS Genet 10, e1004717.

Sipos, L., Kozma, G., Molnar, E., Bender, W., 2007. In situ dissection of a Polycomb response element in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America 104, 12416-12421.

Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., Balasubramanian, S., Harte, R., Wang, D., Rutenberg-Schoenberg, M., Clark, W., Diekhans, M., Rozowsky, J., Hubbard, T., Harrow, J., Gerstein, M.B., 2014. Comparative analysis of pseudogenes across three phyla. Proceedings of the National Academy of Sciences of the United States of America 111, 13361-13366.

Slattery, M., Ma, L., Spokony, R.F., Arthur, R.K., Kheradpour, P., Kundaje, A., Negre, N., Crofts, A., Ptashkin, R., Zieba, J., Ostapenko, A., Suchy, S., Victorsen, A., Jameel, N., Grundstad, A.J., Gao, W., Moran, J.R., Rehm, E.J., Grossman, R.L., Kellis, M., White, K.P., 2014. Diverse patterns of genomic targeting by transcriptional regulators in Drosophila melanogaster. Genome research 24, 1224-1235.

Smit, A.F., Hubley, R. and Green, P., 2013-2015. RepeatMasker Open-4.0.

Song, H., Goetze, S., Bischof, J., Spichiger-Haeusermann, C., Kuster, M., Brunner, E., Basler, K., 2010. Coop functions as a corepressor of Pangolin and antagonizes Wingless signaling. Genes Dev 24, 881-886.

Sproul, D., Gilbert, N., Bickmore, W.A., 2005. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6, 775-781.

Spurlock, C.F., 3rd, Tossberg, J.T., Guo, Y., Collier, S.P., Crooke, P.S., 3rd, Aune, T.M., 2015. Expression and functions of long noncoding RNAs during human T helper cell differentiation. Nat Commun 6, 6932.

Srinivasan, S., Dorighi, K.M., Tamkun, J.W., 2008. Drosophila Kismet regulates histone H3 lysine 27 methylation and early elongation by RNA polymerase II. PLoS Genet 4, e1000217.

St Johnston, D., Nusslein-Volhard, C., 1992. The origin of pattern and polarity in the Drosophila embryo. Cell 68, 201-219.

St Laurent, G., Wahlestedt, C., Kapranov, P., 2015. The Landscape of long noncoding RNA classification. Trends Genet 31, 239-251.

Staehling-Hampton, K., Hoffmann, F.M., Baylies, M.K., Rushton, E., Bate, M., 1994a. dpp induces mesodermal gene expression in Drosophila. Nature 372, 783-786.

Staehling-Hampton, K., Jackson, P.D., Clark, M.J., Brand, A.H., Hoffmann, F.M., 1994b. Specificity of bone morphogenetic protein-related factors: cell fate and gene expression changes in Drosophila embryos induced by decapentaplegic but not 60A. Cell Growth Differ 5, 585-593.

Stanojevic, D., Hoey, T., Levine, M., 1989. Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in Drosophila. Nature 341, 331-335.

Starr, M.O., Ho, M.C., Gunther, E.J., Tu, Y.K., Shur, A.S., Goetz, S.E., Borok, M.J., Kang, V., Drewell, R.A., 2011. Molecular dissection of cis-regulatory modules at the Drosophila bithorax complex reveals critical transcription factor signature motifs. Dev Biol 359, 290-302.

Steffen, P.A., Ringrose, L., 2014. What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. Nature reviews. Molecular cell biology 15, 340-356.

Stern, D.L., 2003. The Hox gene Ultrabithorax modulates the shape and size of the third leg of Drosophila by influencing diverse mechanisms. Dev Biol 256, 355-366.

Struhl, G., 1981. A gene product required for correct initiation of segmental determination in Drosophila. Nature 293, 36-41.

Struhl, G., 1982a. Genes controlling segmental specification in the Drosophila thorax. Proceedings of the National Academy of Sciences of the United States of America 79, 7380-7384.

Struhl, G., and Brower, D., 1982b. Early role of the esc+ gene product in the determination of segments in Drosophila. Cell 31, 285-292.

Struhl, K., 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14, 103-105.

Strutt, H., Cavalli, G., Paro, R., 1997. Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. Embo J 16, 3621-3632.

Sun, B.K., Deaton, A.M., Lee, J.T., 2006. A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. Mol Cell 21, 617-628.

Sun, Q., Csorba, T., Skourti-Stathaki, K., Proudfoot, N.J., Dean, C., 2013a. R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. Science 340, 619-621.

Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y., Lee, J.T., 2013b. Jpx RNA activates Xist by evicting CTCF. Cell 153, 1537-1551.

Swarup, S., Verheyen, E.M., 2012. Wnt/Wingless signaling in Drosophila. Cold Spring Harb Perspect Biol 4.

Swiezewski, S., Liu, F., Magusin, A., Dean, C., 2009. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature 462, 799-802.

Taft, R.J., Pheasant, M., Mattick, J.S., 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays : news and reviews in molecular, cellular and developmental biology 29, 288-299.

Tamura, K., Subramanian, S., Kumar, S., 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21, 36-44.

Thebault, P., Boutin, G., Bhat, W., Rufiange, A., Martens, J., 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, 1288-1300.

Thibault, S.T., Singer, M.A., Miyazaki, W.Y., Milash, B., Dompe, N.A., Singh, C.M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H.L., Ryner, L., Cheung, L.M., Chong, A., Erickson, C., Fisher, W.W., Greer, K., Hartouni, S.R., Howie, E., Jakkula, L., Joo, D., Killpack, K., Laufer, A., Mazzotta, J., Smith, R.D., Stevens, L.M., Stuber, C., Tan, L.R., Ventura, R., Woo, A., Zakrajsek, I., Zhao, L., Chen, F., Swimmer, C., Kopczynski, C., Duyk, G., Winberg, M.L., Margolis, J., 2004. A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. Nat Genet 36, 283-287.

Thomas, P.D., Wood, V., Mungall, C.J., Lewis, S.E., Blake, J.A., Gene Ontology, C., 2012. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. PLoS Comput Biol 8, e1002386.

Tian, D., Sun, S., Lee, J.T., 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell 143, 390-403.

Tie, F., Banerjee, R., Fu, C., Stratton, C.A., Fang, M., Harte, P.J., 2016. Polycomb inhibits histone acetylation by CBP by binding directly to its catalytic domain. Proceedings of the National Academy of Sciences of the United States of America 113, E744-753.

Tie, F., Banerjee, R., Saiakhova, A.R., Howard, B., Monteith, K.E., Scacheri, P.C., Cosgrove, M.S., Harte, P.J., 2014. Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. Development 141, 1129-1139.

Tie, F., Banerjee, R., Stratton, C.A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M.O., Scacheri, P.C., Harte, P.J., 2009. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. Development 136, 3131-3141.

Tie, F., Furuyama, T., Harte, P.J., 1998. The Drosophila Polycomb Group proteins ESC and E(Z) bind directly to each other and co-localize at multiple chromosomal sites. Development 125, 3483-3496.

Tie, F., Furuyama, T., Prasad-Sinha, J., Jane, E., Harte, P.J., 2001. The Drosophila Polycomb Group proteins ESC and E(Z) are present in a complex containing the histone-binding protein p55 and the histone deacetylase RPD3. Development 128, 275-286.

Tixier, V., Bataille, L., Etard, C., Jagla, T., Weger, M., Daponte, J.P., Strahle, U., Dickmeis, T., Jagla, K., 2013. Glycolysis supports embryonic muscle growth by promoting myoblast fusion. Proceedings of the National Academy of Sciences of the United States of America 110, 18982-18987.

Tkachuk, D.C., Kohler, S., Cleary, M.L., 1992. Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias. Cell 71, 691-700.

Tolhuis, B., de Wit, E., Muijrers, I., Teunissen, H., Talhout, W., van Steensel, B., van Lohuizen, M., 2006. Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. Nature Genetics 38, 694-699.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., Rubin, G.M., 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3, RESEARCH0088.

Tomancak, P., Berman, B.P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S.E., Rubin, G.M., 2007. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 8, R145.

Tomoyasu, Y., Nakamura, M., Ueno, N., 1998. Role of dpp signalling in prepattern formation of the dorsocentral mechanosensory organ in Drosophila melanogaster. Development 125, 4215-4224.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511-515.

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., Chang, H.Y., 2010. Long noncoding RNA as modular scaffold of histone modification complexes. Science 329, 689-693.

Tschiersch, B., Hofman, A., Krauss, V., Dorn, R., Korge, G., and Reuter, G., 1994. The protein encoded by the *Drosophila* position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. EMBO 13, 3822-3831.

Tsukiyama, T., Daniel, C., Tamkun, J., and Wu, C., 1995. ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. Cell 83, 1021-1026.

Tullai, J.W., Schaffer, M.E., Mullenbrock, S., Sholder, G., Kasif, S., Cooper, G.M., 2007. Immediate-early and delayed primary response genes are distinct in function and genomic architecture. J Biol Chem 282, 23981-23995.

Tyler, D.M., Okamura, K., Chung, W.J., Hagen, J.W., Berezikov, E., Hannon, G.J., Lai, E.C., 2008. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. Genes & development 22, 26-36.

Uhler, J.P., Hertel, C., Svejstrup, J.Q., 2007. A role for noncoding transcription in activation of the yeast PHO5 gene. Proceedings of the National Academy of Sciences of the United States of America 104, 8011-8016.

Ulitsky, I., Bartel, D.P., 2013. lincRNAs: genomics, evolution, and mechanisms. Cell 154, 26-46.

Vachon, G., Cohen, B., Pfeifle, C., McGuffin, M.E., Botas, J., Cohen, S.M., 1992. Homeotic genes of the Bithorax complex repress limb development in the abdomen of the Drosophila embryo through the target gene Distal-less. Cell 71, 437-450.

Vázquez, M., Moore, L., and Kennison, J. A., 1999. The trithorax group gene osa encodes an ARID-domain protein that genetically interacts with the Brahma chromatin-remodeling factor to regulate transcription. Development 126, 733-742.

Veitia, R.A., 2008. One thousand and one ways of making functionally similar transcriptional enhancers. Bioessays 30, 1052-1057.

Vella, P., Barozzi, I., Cuomo, A., Bonaldi, T., Pasini, D., 2012. Yin Yang 1 extends the Myc-related transcription factors network in embryonic stem cells. Nucleic acids research 40, 3403-3418.

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S.

Vincent, J.P., O'Farrell, P.H., 1992. The state of engrailed expression is not clonally transmitted during early Drosophila development. Cell 68, 923-931.

Vischer, E., Chargaff, E., 1948. The separation and quantitative estimation of purines and pyrimidines in minute amounts. J Biol Chem 176, 703-714.

Von Allmen, G., Hogga, I., Spierer, A., Karch, F., Bender, W., Gyurkovics, H., Lewis, E., 1996. Splits in fruitfly Hox gene complexes. Nature 380, 116.

Voziyanov, Y., Pathania, S., Jayaram, M., 1999. A general model for site-specific recombination by the integrase family recombinases. Nucleic acids research 27, 930-941.

Wagner, G.P., Amemiya, C., Ruddle, F., 2003. Hox cluster duplications and the opportunity for evolutionary novelties. Proceedings of the National Academy of Sciences of the United States of America 100, 14603-14606.

Walker, E., Chang, W.Y., Hunkapiller, J., Cagney, G., Garcha, K., Torchia, J., Krogan, N.J., Reiter, J.F., Stanford, W.L., 2010. Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. Cell Stem Cell 6, 153-166.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., Zhang, Y., 2004a. Role of histone H2A ubiquitination in Polycomb silencing. Nature 431, 873-878.

Wang, J., Haubrock, M., Cao, K.M., Hua, X., Zhang, C.Y., Wingender, E., Li, J., 2011a. Regulatory coordination of clustered microRNAs based on microRNA-transcription factor regulatory network. BMC Syst Biol 5, 199.

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J., Wong, G.K., 2004b. Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature 431, 1 p following 757; discussion following 757.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., Wysocka, J., Lei, M., Dekker, J., Helms, J.A., Chang, H.Y., 2011b. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120-124.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., Li, W., 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research 41, e74.

Wang, Q., Hasan, G., Pikielny, C.W., 1999. Preferential expression of biotransformation enzymes in the olfactory organs of Drosophila melanogaster, the antennae. J Biol Chem 274, 10309-10315.

Washietl, S., Hofacker, I.L., Stadler, P.F., 2005. Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences of the United States of America 102, 2454-2459.

Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T.D., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., Kube, M., Taenzer, S., Galgoczy, P., Platzer, M., Scharfe, M., Nordsiek, G., Blocker, H., Hellmann, I., Khaitovich, P., Paabo, S., Reinhardt, R., Zheng, H.J., Zhang, X.L., Zhu, G.F., Wang, B.F., Fu, G., Ren, S.X., Zhao, G.P., Chen, Z., Lee, Y.S., Cheong, J.E., Choi, S.H., Wu, K.M., Liu, T.T., Hsiao, K.J., Tsai, S.F., Kim, C.G., S, O.O., Kitano, T., Kohara, Y., Saitou, N., Park, H.S., Wang, S.Y., Yaspo, M.L., Sakaki, Y., 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature 429, 382-388.

Watson, J.D., Crick, F.H., 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737-738.

Weatherbee, S.D., Halder, G., Kim, J., Hudson, A., Carroll, S., 1998. Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the Drosophila haltere. Genes Dev 12, 1474-1482.

Wech, I., Bray, S., Delidakis, C., Preiss, A., 1999. Distinct expression patterns of different enhancer of split bHLH genes during embryogenesis of Drosophila melanogaster. Dev Genes Evol 209, 370-375.

Weeks, K.M., 2010. Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol 20, 295-304.

Weiss, J.B., Suyama, K.L., Lee, H.H., Scott, M.P., 2001. Jelly belly: a Drosophila LDL receptor repeat-containing signal required for mesoderm migration and differentiation. Cell 107, 387-398.

Weiszmann, R., Hammonds, A.S., Celniker, S.E., 2009. Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. Nat Protoc 4, 605-618.

White, R.A.H.a.A., M. E., 1985. *Contrabithorax* mutations cause inappropriate expression of *Ultrabithorax* products in *Drosophila*. Nature 318, 567-569.

Wieschaus, E., Gehring, W., 1976. Clonal analysis of primordial disc cells in the early embryo of Drosophila melanogaster. Dev Biol 50, 249-263.

Wilder, E.L., Perrimon, N., 1995. Dual functions of wingless in the Drosophila leg imaginal disc. Development 121, 477-488.

Woese, C.R., 1964. Universality in the Genetic Code. Science 144, 1030-1031.

Wurmbach, E., Wech, I., Preiss, A., 1999. The Enhancer of split complex of Drosophila melanogaster harbors three classes of Notch responsive genes. Mech Dev 80, 171-180.

Wutz, A., Rasmussen, T.P., Jaenisch, R., 2002. Chromosomal silencing and localization are mediated by different domains of Xist RNA. Nat Genet 30, 167-174.

Xiao, H., Sandalzopoulos, R., Wang, H. M., Hamiche, A., Ranallo, R., Lee, K. M., Fu, D., and Wu, C., 2001. Dual functions of largest NURF subunit NURF301 in nucleosome sliding and transcription factor interactions. Mol. Cell 8, 531-543.

Xu, N., Donohoe, M.E., Silva, S.S., Lee, J.T., 2007. Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. Nat Genet 39, 1390-1396.

Yajima, H., 1964. Studies on Embryonic Determination of the Harlequin-Fly, Chironomus Dorsalis. Ii. Effects of Partial Irradiation of the Egg by Ultra-Violet Light. J Embryol Exp Morphol 12, 89-100.

Yang, L., Lin, C., Liu, W., Zhang, J., Ohgi, K.A., Grinstein, J.D., Dorrestein, P.C., Rosenfeld, M.G., 2011. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell 147, 773-788.

Yang, W.M., Inouye, C., Zeng, Y., Bearss, D., Seto, E., 1996. Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. Proceedings of the National Academy of Sciences of the United States of America 93, 12845-12850.

Yang, Y., Wen, L., Zhu, H., 2015. Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein. Cell Biosci 5, 59.

Yao, L.C., Liaw, G.J., Pai, C.Y., Sun, Y.H., 1999. A common mechanism for antenna-to-Leg transformation in Drosophila: suppression of homothorax transcription by four HOM-C genes. Dev Biol 211, 268-276.

Yekta, S., Tabin, C.J., Bartel, D.P., 2008. MicroRNAs in the Hox network: an apparent link to posterior prevalence. Nat Rev Genet 9, 789-796.

Yin, J.W., Wang, G., 2014. The Mediator complex: a master coordinator of transcription and cell lineage development. Development 141, 977-987.

Yoon, J.H., Srikantan, S., Gorospe, M., 2012. MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs. Methods 58, 81-87.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., Ponting, C.P., 2012. Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. Genome biology and evolution 4, 427-442.

Yu, A.D., Wang, Z., Morris, K.V., 2015. Long noncoding RNAs: a potent source of regulation in immunity and disease. Immunology and cell biology 93, 277-283.

Zecca, M., Basler, K., Struhl, G., 1995. Sequential organizing activities of engrailed, hedgehog and decapentaplegic in the Drosophila wing. Development 121, 2265-2278.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., Lee, J.T., 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322, 750-756.

Zhao, J.J., Lazzarini, R.A., Pick, L., 1993. The mouse Hox-1.3 gene is functionally equivalent to the Drosophila Sex combs reduced gene. Genes Dev 7, 343-354.

Zhou, J., Ashe, H., Burks, C., Levine, M., 1999. Characterization of the transvection mediating region of the abdominal-B locus in Drosophila. Development 126, 3057-3065.

Zhu, C.C., Bornemann, D.J., Zhitomirsky, D., Miller, E.L., O'Connor, M.B., Simon, J.A., 2008. Drosophila histone deacetylase-3 controls imaginal disc size through suppression of apoptosis. PLoS Genet 4, e1000009.

Ziemin-van der Poel, S., McCabe, N.R., Gill, H.J., Espinosa, R., 3rd, Patel, Y., Harden, A., Rubinelli, P., Smith, S.D., LeBeau, M.M., Rowley, J.D., et al., 1991. Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. Proceedings of the National Academy of Sciences of the United States of America 88, 10735-10739.

Zink, D., Paro, R., 1995. Drosophila Polycomb-group regulated chromatin inhibits the accessibility of a trans-activator to its target DNA. Embo J 14, 5660-5671.

Zraly, C.B., Marenda, D. R., Nanchal, R., Cavalli, G., Muchardt, C., and Dingwall, A. K., 2003. SNR1 is an essential subunit in a subset of *Drosophila* brm complexes, targeting specific functions during development. Dev. Biol. 253, 291-308.