Thesis submitted for the degree of Doctor of Philosophy
at the University of Oxford

EVOLUTION AND FUNCTION OF LONG NONCODING RNAs IN *DROSOPHILA*

Robert Young

BALLIOL COLLEGE
OXFORD

TRINITY TERM
2011

EVOLUTION AND FUNCTION OF LONG NONCODING RNAs IN *DROSOPHILA*

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Robert Young, Balliol College, Trinity Term 2011

ABSTRACT

Not all transcribed DNA encodes protein, and some of these noncoding RNAs (ncRNAs), such as *roX1* and *roX2*, may play important roles in the cell. The functional roles of the majority of these, however, remain largely unknown. In this thesis, I first used EST and mRNA evidence to define 2,788 lincRNA loci within the *Drosophila melanogaster* genome. I suggest that up to 1,652 of these are functional, as 1,411 show evidence for significant evolutionary constraint while 241 fast-evolving loci are enriched in short RNA species. A distinct set of 1,119 lincRNA loci were defined by RNA-seq, the vast majority of which show clear primary sequence constraint. Their expression profiles and enrichment in particular chromatin domains indicate that these lincRNAs are likely involved in developmental regulation. I also identified 42 potential analogous lincRNAs with shared genomic locations between *Drosophila* and mouse. Constrained, non-embryonic lincRNAs defined by ESTs are transcribed preferentially in the vicinity of protein-coding genes encoding transcription factors and I demonstrated that one of these, which I name *dEvf-2*, positively regulates the expression of its genomically adjacent transcription factor, *Dll*, in cell culture. Finally, I used a reverse genetics approach to search for lincRNA promoter mutations and examined the effect of these on lincRNA expression. My findings suggest that many, previously unknown, functional lincRNAs exist within the *Drosophila* genome and are worthy of further in-depth experimental investigation.

TABLE OF CONTENTS

DECLARATION

The work presented in this thesis has been conducted by the author Robert Young in the Department of Physiology, Anatomy and Genetics at the University of Oxford. Except where acknowledgement is made, all the work reported in this thesis is my own. Specifically, all datasets described in the Materials sections of **Chapters 2-6** were generated by the referenced authors and brief descriptions of their methods are supplied purely for ease of understanding. The 5' RACE sequences for *dEvf-2* described in **Chapter 5.5.5** were produced, as referenced in the text, by Charlotte Tibbit. I hereby declare that this thesis has not been submitted either in the same or different form, to this or any other university for a degree.

ACKNOWLEDGEMENTS

I would like to sincerely thank Prof. Chris P. Ponting for his enthusiasm, guidance, and patience in supervising this project. I would like to thank Dr. Ji-Long Liu for co-supervising my work, and mentoring the wet-lab aspects of this project. I am also incredibly grateful to Dr. Ana Marques for her help throughout all stages of this project, and for inspiring me to keep on trying.

I would like to acknowledge all members of the Ponting, Liu, and DARCGENs groups for useful discussions, helping complete the daily crossword, and creating an enjoyable atmosphere in which to work. In particular, I thank Steve Meader for his academic and personal support, and Siân, Hannah and Charlotte for their friendship and invaluable help with experiments.

Outside the lab, I am indebted to my parents and sisters for their wholehearted support and to Graham Baker, for his valued friendship. Finally, I would like to thank Heather for her constant encouragement and bearing the brunt during the more difficult times.

I would also like to thank the UK Medical Research Council and Balliol College for financial support.

ABBREVIATIONS

BAP	Bacterial Alkaline Phosphatase
BDGP	Berkeley *Drosophila* Genome Project
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
bp	Base Pair
BWT	Burrows-Wheeler Transform
BX-C	Bithorax Complex
CAGE	Cap Analysis of Gene Expression
CAP	Calf Intestine Alkaline Phosphatase
cDNA	Complementary DNA
ChIP	Chromatin Immuno-Precipitation
CNS	Central Nervous System
CPC	Coding Potential Calculator
CUT	Cryptic Unstable Transcript
DAPI	4',6-diamidino-2-phenylindole
d_N	Non-synonymous Substitution Rate
DNA	Deoxyribonucleic Acid
DPE	Downstream Promoter Element
DRE	Downstream Replication Element
DRSC	*Drosophila* RNAi Screening Centre
d_S	Synonymous Substitution Rate
dsDNA	Double-stranded DNA
dsRNA	Double-Stranded RNA
ENCODE	Encyclopaedia of DNA Elements
eRNA	Enhancer RNA
esiRNA	Endogenous Short Interfering RNA
EST	Expressed Sequence Tag
FACS	Fluorescence-activated Cell Sorting
FDR	False Discovery Rate
FISH	Fluorescent *in situ* hybridisation
FPKM	Fragments per Kilobase of Exon Model per Million Mapped Fragments
FPR	False Positive Rate
FWHM	Full Width at Half Maximum

GO	Gene Ontology
HMM	Hidden Markov Model
Hsr	Heat-Shock Response
IGS	Inter Gap Segment
Inr	Initiator Motif
IPS	Indel Purified Segment
ISM	*In situ* Mix
IUM	Initially Unmappable
kb	Kilobase
lincRNA	Long Intergenic noncoding RNA
lncRNA	Long noncoding RNA
Mb	Megabase
MCS	Multi-Species Conserved Segment
MFE	Minimum Free Energy
min	Minute
miRNA	Micro RNA
modENCODE	Model Organism Encyclopaedia of DNA Elements
mRNA	Messenger RNA
MSL	Male Specific Lethal
ncRNA	Noncoding RNA
NTR	New Transcribed Region
NSCLC	Non-Small Cell Lung Cancer
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline Solution
PCR	Polymerase Chain Reaction
Phylo-HMM	Phylogenetic Hidden Markov Model
piRNA	Piwi-Interacting RNA
Pol II	RNA Polymerase II
RACE	Rapid Amplification of cDNA Ends
RAM	Random-Access Memory
RLM-RACE	RNA Ligase-Mediated Rapid Amplification of cDNA Ends

RNA	Ribonucleic Acid
RNAi	RNA Interference
RPKM	Reads per Kilobase of Exon Model per Million Mapped Reads
rRNA	Ribosomal RNA
RT-PCR	Reverse Transcription and Polymerase Chain Reaction
S2	Schneider 2
SAGE	Serial Analysis of Gene Expression
siRNA	Small Interfering RNA
SNP	Single Nucleotide Polymorphism
snRNA	Small Nuclear RNA
snoRNA	Small Nucleolar RNA
ssRNA	Single-Stranded RNA
SVM	Support Vector Machine
TAP	Tobacco Acid Pyrophosphatase
TBP	Tata-Binding Protein
TF	Transcription Factor
Transfrag	Transcribed Fragment
tRNA	Transfer RNA
TSS	Transcriptional Start Site
TU	Transcriptional Unit
UAS	Upstream Activating Sequence
UTR	Untranslated Region
wt	Wild-type
X-gal	Bromo-chloro-indolyl-galactopyranoside
XCI	X Chromosome Inactivation

Chapter 1: INTRODUCTION

1.1 Transcription has been observed beyond annotated protein-coding genes in all species examined to date.

1.1.1 The transcriptome is complex and contains a large proportion of previously unannotated 'dark matter'.

Widespread transcription beyond the boundaries of currently annotated protein-coding gene sets, the so-called 'dark matter' (Johnson et al. 2005), of the genome has been recognised for a number of years but its biological significance remains unclear. A complex picture of transcription with many interleaving transcripts has emerged from a series of microarray experiments, such as the tiling arrays which include exonic, intronic and intergenic regions and cover most of the nonrepetitive bases in the genomes of humans (Bertone et al. 2004; Kampa et al. 2004) and the fruit fly, *Drosophila melanogaster* (Stolc et al. 2004). This widespread transcription beyond annotated gene sets has also been noted in large-scale cDNA collections (Carninci et al. 2005) and in whole-transcriptome shotgun sequencing (RNA-seq) experiments (e.g. Cloonan et al. 2008; Guttman et al. 2010). This phenomenon is not restricted to multicellular animals as, when placed in rich media, up to 85% of the genome of the budding yeast *Saccharomyces cerevisiae* is observed to be

transcribed (David et al. 2006). This previously unknown transcription has improved many gene annotations through the addition of novel 5' and 3' untranslated regions (UTRs) as well as the inclusion of novel exons. Furthermore, transcription which is intronic, antisense or even between these gene models is now thought to be commonplace in most eukaryotic genomes (Ponjavic and Ponting 2007). An example of a human locus with such complex transcription is shown in **Figure 1.1**. It has been observed that there is up to five-fold more transcribed ribonucleic acid (RNA) in the nucleus of the cell, relative to the cytoplasm (Cheng et al. 2005), which suggests that most products of transcription are retained in the nucleus. Many of these novel RNA transcripts lack the ability to encode a functional protein, and they are therefore defined as noncoding RNAs (ncRNAs).

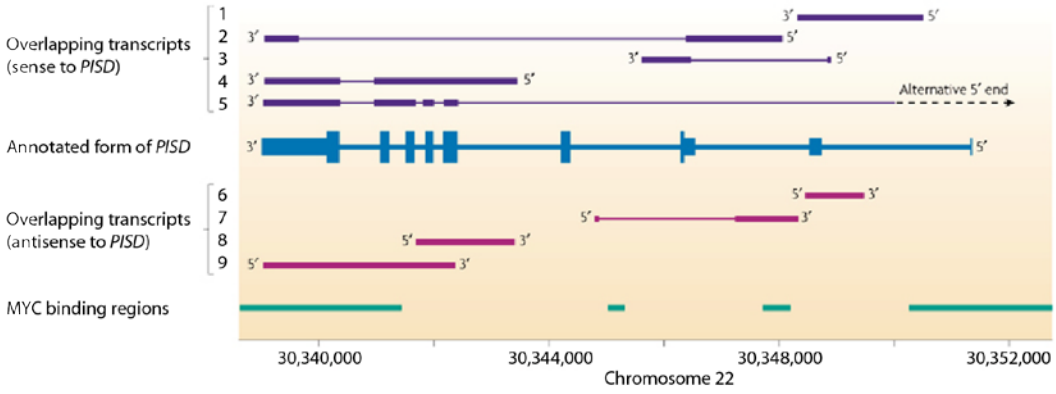

Figure 1.1 Five transcripts (in purple) that overlap the RefSeq-annotated form (in blue) of the phosphatidylserine decarboxylase gene (*PISD*) on the same strand, and four transcripts (in pink) that overlap the gene on the opposite strand are shown. Binding sites of the transcriptional factor MYC are shown in green. Taken from Kapranov et al. 2007.

This pervasive transcription is accompanied by estimates of the amount of functional material in eukaryotic genomes which far exceed that which encodes

protein-coding genes and their known associated regulatory elements. In humans, only 1.1% of the genome encodes protein-coding genes (Church et al. 2009) while at least 2.6-12% has been estimated to be biologically relevant (reviewed in Ponting and Hardison 2011). Similarly, in *D. melanogaster*, where at least 47% of the genome is thought to be functional (Meader et al. 2010), only 19% encodes coding sequence (Tweedie et al. 2009). Observations such as these suggest that many of these ncRNA molecules may have functional roles to play in the cell.

The reliability of microarray-based estimates of the amount of transcription has, however, since been questioned. It has been noted that validation rates from these experiments vary from 25-70% which suggests that a large proportion of the transcription observed may represent experimental artefacts (van Bakel et al. 2010), resulting from cross-hybridisation between probes or high levels of background fluorescence (Ponting and Belgard 2010). Transcription of annotated genes appears to be replicated in different experiments more frequently than the novel, often noncoding, transcripts (van Bakel and Hughes 2009), implying that these ncRNAs are more likely to be susceptible to experimental biases and represent false positives.

Transcripts matching unannotated exons have also been detected within large expressed sequence tag (EST) and complementary DNA (cDNA) collections which contain RNA molecules sequenced from a large variety of sources. As of January 2008, the EST database in GenBank (Benson et al. 2005) contained over 48 million sequences from approximately 2,000 species. A sizeable

proportion of these records may be unable to encode a functional peptide, which would again suggest that the noncoding proportion of the transcriptome is significantly greater than was previously thought. For example, nearly half of 21,243 human cDNAs have no clear open reading frame (ORF) with which to encode a protein (Ota et al. 2004). In mouse, 34,030 (33%) of a set of high quality, frequently complete, cDNAs produced by the FANTOM consortium also lack an ORF (Carninci et al. 2005). This technique does not produce a complete set of all expressed transcripts from a sample and, as these libraries are known to be frequently derived from brain or gonadal sources (Daines et al. 2011), it is likely that they will underestimate the full extent of this transcription. Nevertheless, it is clear that this dark matter can make an appreciable contribution to the RNA within a cell, and that a large amount of this transcription is likely to be non-coding.

The recent emergence of a second generation of technologies to sequence the transcriptome has overcome many of the difficulties of microarray- and EST-based surveys and yielded several orders of magnitude more sequence information. This type of approach, which will be discussed in more detail in **Chapter 4**, is able to generate millions of short (35 – 75 bp) sequencing reads in parallel, and has been used to study the transcriptomes of a variety of species (e.g. Mortazavi et al. 2008). These experiments have confirmed that a large proportion of the genome is transcribed into ncRNAs, although this amount is often less than that proposed by microarray experiments (van Bakel et al. 2010). Novel, non-coding transcripts are often represented by only a few

sequencing reads, which suggests that the majority are expressed at very low levels (Costa et al. 2010) and may partially explain why they have remained unidentified until now. The high-throughput and unbiased nature of this technique means that it is possible to naïvely survey the transcriptome of a variety of tissues and/or species under different conditions and approach a more accurate estimate of the total amount of transcription.

1.1.2 Several classes of short ncRNAs have been previously annotated.

These observations of rampant transcription of the genome were not the first to propose the existence of functional ncRNAs. Housekeeping RNAs, such as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) which are involved in mRNA translation, have been known for many years (Claude 1937; Holley et al. 1965). Several classes of regulatory ncRNAs have also been known and well studied for almost 20 years. Small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) are both involved in RNA processing. MicroRNAs (miRNAs), endogenous short interfering RNAs (esiRNAs) and Piwi-interacting RNAs (piRNAs) associate with various members of the Argonaute family of proteins and use shared sequence similarity to direct them to suppression expression of a target transcript, whether it is a protein-coding gene or transposable element ((Peters and Meister 2007). None of these classes of short RNAs are longer than ~30 nt long (Jung et al. 2010) and many of these form families of sequences containing similar secondary structures which are stored in the Rfam database of known RNA families (Gardner et al. 2009).

While a study of these short RNAs in *D. melanogaster* is beyond the scope of this thesis, I now describe the three main classes of such RNAs involved in regulating expression of genetic elements which have been described to date and whose genome-wide distributions I investigate in later chapters – miRNAs, esiRNAs and piRNAs.

MiRNAs were the first type of short RNAs identified, first in the worm, *Caenorhabditis elegans* (Lee et al. 1993), and then in a variety of other eukaryotic species. They have since been implicated in suppression of gene expression when loaded onto the Argonaute-1 protein (Ago1, Okamura et al. 2004), in *D. melanogaster*. They are 21-25 nt long RNA molecules (He and Hannon 2004) whose sequence is complementary to a stretch of the mRNA sequence of the target gene. Those miRNAs which perfectly match their target are most prominent in plants and are thought to form a double-stranded RNA (dsRNA) duplex with the mRNA which is then targeted for degradation by the RNA interference (RNAi) pathway (Lai 2003). In animals, where most matches are imperfect, miRNAs often target the 3' UTR and this in turn inhibits the translation of the mRNA (Perkins et al. 2005). Deeply conserved miRNAs can be detected in eukaryotic species (Wheeler et al. 2009) and they have also been detected in viral genomes (Pfeffer et al. 2004), suggesting that they are likely to be important in regulating fundamental biological processes. There are currently 237 miRNA genes in the *D. melanogaster* genome that have been annotated and placed in the miRNA repository miRBase (Kozomara and Griffiths-Jones 2011).

A separate set of short RNAs known as esiRNAs have been implicated in the suppression of transposable element expression. These are ~18 nt long and defined as being in a complex with the Argonaute-2 protein (Ago2, Czech et al. 2008). These are largely transcribed in clusters from repetitive, transposable element-derived sequence. Their biogenesis requires both *Ago2* and *Dicer-2* (Ghildiyal and Zamore 2009).

PiRNAs are similar to esiRNAs, as they are also involved in transposon suppression and are transcribed from clusters of repetitive sequence (Yin and Lin 2007). Often, the same cluster can produce both esiRNAs and piRNAs (Ghildiyal et al. 2008). However, piRNAs are slightly longer (26 – 31 nt) and are thought to be important predominantly in the germline (Aravin et al. 2003). They complex with the Piwi and Aubergine proteins and their expression can be amplified by a so-called ping-pong model which also requires the Argonaute-3 protein (Brennecke et al. 2007). Through their interaction with the Piwi protein, piRNAs have also been implicated in heterochromatin formation within the subtelomeric and pericentromeric regions from which they are transcribed (Yin and Lin 2007). PiRNAs can be maternally deposited through the germline in *D. melanogaster* (Brennecke et al. 2008).

1.2 Long noncoding RNAs, including long intergenic noncoding RNAs (lincRNAs), are the least well characterised members of novel transcription.

This novel dark matter also contains a large number of long noncoding RNA molecules whose functions, unlike the short RNAs discussed above, are mostly unknown. Long ncRNAs (lncRNAs) are defined as being $>$ 200 bp and lacking in protein-coding potential (Ponting et al. 2009). These long ncRNAs can be classified into one of five groups, based on their genomic location (Ørom and Shiekhattar 2011):

1. Antisense to a protein-coding gene.
2. Intronic to a protein-coding gene.
3. Corresponding to the 5' or 3' UTR of a protein-coding gene.
4. Independently transcribed from within a protein-coding gene locus.
5. Intergenic (lincRNAs).

Those lncRNAs that do not overlap with another annotated gene model are known as long intergenic ncRNAs (lincRNAs, Guttman et al. 2009). It is this sub-type of lncRNAs which I will focus on for most of this thesis.

Most lncRNAs that have previously been identified are thought to be transcribed by Pol II, as they are often polyadenylated (Inagaki et al. 2005).

Indeed, recent mouse and human lincRNA sets have been defined using H3K4me3 and H3K36me3 histone modifications (Guttman et al. 2009; Khalil et al. 2009) which are markers for Pol II activity (Mikkelsen et al. 2007) and lncRNAs have been found with a range of other epigenetic modifications (Amaral and Mattick 2008). These ncRNAs are frequently spliced, and are expressed in a low (Mattick 2003) and tissue-specific (Halasz et al. 2006) manner, where many are thought to remain in the nucleus (Backofen et al. 2007). Several of these lncRNAs have been investigated in detail, where most have been demonstrated to play a role in the regulation of gene expression (Ponting et al. 2009). Below, I briefly review a selection of mechanisms by which ncRNAs (and specifically lincRNAs) have been demonstrated to regulate the expression of protein-coding genes.

1.2.1 Chromatin Modification

Individual lincRNAs which aid in the regulation of chromosome-wide gene expression have been identified in both *Drosophila* and mammals. These lincRNAs are involved in a phenomenon known as dosage compensation which equalises the dosage of gene expression from the X chromosome between females with two X chromosomes and males with only one X chromosome (Lewin 2003). In *Drosophila*, this is proposed to be achieved by hypertranscription from the single X chromosome in males (Mukherjee and Beermann 1965) while, in mouse, transcription from one X chromosome is inactivated in female cells (Monk and Harper 1979).

In *Drosophila*, transcription from the male X chromosome is regulated by the male-specific lethal (MSL) complex (Hamada et al. 2005; Straub et al. 2005), a complex containing several proteins (MSL1, MSL2, MSL3, Males absent on the first (MOF), and Maleless (MLE)) and two lincRNAs (RNA on X1 (*roX1*) and RNA on X2 (*roX2*)), which are both transcribed from the X chromosome (Amrein and Axel 1997). Mutant analysis has suggested that the complex binds 30-40 'entry' sites on the X chromosome (Kelley et al. 1999) and then spreads in *cis* to coat the entire chromosome, leading to H4K16 acetylation at actively transcribed gene loci (Gelbart et al. 2009), a more diffuse chromosome morphology, and hypertranscription (Deng and Meller 2006). *RoX1* and *roX2* are functionally redundant, despite sharing little sequence similarity and displaying distinct embryonic expression profiles (Meller and Rattner 2002). These observations can be reconciled, at least in part, by experiments showing that most of the sequence of these lincRNAs is not required for normal function (Meller and Rattner 2002). While *roX1* and *roX2* are both transcribed from the X chromosome, this is not necessary as *roX1* function can be successfully rescued by reintroducing it as an autosomally-encoded gene (Meller et al. 1997). This suggests that these lincRNAs are likely to be acting through a *trans*-regulatory mechanism.

The mechanism in mice is quite different to that of *Drosophila*, despite it regulating a similar process. X chromosome inactivation (XCI) is thought to require the coating of the X chromosome by only one lincRNA – the 15 kb X-inactive specific transcript (*Xist*, Brockdorff et al. 1992). This lincRNA is

transcribed from the inactive X chromosome and spreads strictly in *cis* to coat the chromosome and prevent transcription (Penny et al. 1996). Whether this requires a complex of proteins, as in *Drosophila*, or whether the act of coating by this lincRNA is sufficient to modify the X chromosome chromatin environment, remains unknown (Storz 2002). The sequence of *Xist* is only 60% conserved across eutherian mammals (Barciszewski and Erdmann 2003) but the gene structure is conserved between mouse and human, with several short well-conserved regions (Nesterova et al. 2001). *Xist* expression is regulated by an antisense-encoded lncRNA, which has been named *Tsix* (Lee et al. 1999) and whose promoter is 13 kb downstream from *Xist. Tsix* is transcribed across the *Xist* promoter (Ogawa and Lee 2002) where it represses active chromatin modifications (Navarro et al. 2005) and, thereby, *Xist* transcription. *Tsix* transcription must continue through the *Xist* promoter as truncation of the transcript removes its regulatory ability (Ohhata et al. 2008).

LncRNAs can also regulate the expression of several adjacent genes on the chromosome, and this is seen in a number of imprinted gene clusters. Imprinted genes are expressed from only one allele in a diploid animal, and this expression depends on whether it is inherited from the maternal or paternal allele (Alberts et al. 2002). Imprinting was first identified in *Drosophila* (Crouse 1960) but it is in mammals that lncRNAs have been implicated in this process. Imprinted genes generally exist in clusters, suggesting that they are regulated as a single domain where lncRNA expression from one allele is often associated with repression of the protein-

coding gene on that allele (Prasanth and Spector 2007). One such locus is the mouse lncRNA *Airn*, whose locus overlaps the *Igf2r* gene. *Airn* is normally transcribed exclusively from the paternal allele, where it prevents expression of a gene cluster containing *Igf2r*, *Slc22a2* and *Slc22a3* from that allele, despite being antisense only to *Igf2r* (Sleutels et al. 2002). Disrupting the *Airn* promoter causes *Igf2r* to be expressed from the paternal as well as the maternal allele (Wutz et al. 1997). The mechanism causing this remains unclear, although it appears to involve H3K9me3 recruitment at the *Slc22a3* promoter (Nagano et al. 2008). The human *IGF2R* gene is not similarly regulated by an antisense lncRNA (Oudejans et al. 2001) and the *Airn* locus is poorly conserved across related species (Pang et al. 2006), suggesting that this type of lncRNA regulation may be species-specific. Not all imprinted lncRNAs are so taxonomically restricted, however, as the maternally-expressed lncRNA *H19* has been identified in both placental and marsupial mammals, where it appears to be involved in regulating the expression of the nearby protein-coding gene *IGF2* in neonatal tissues (Smits et al. 2008).

1.2.2 *Cis*-regulation

LincRNAs which regulate the expression of a single protein-coding gene have also been identified. When this protein-coding gene is found genomically adjacent to the lincRNA locus, I will refer to this regulation as *cis*-regulation. For example, the human *DHFR* locus contains two promoters, where transcription of a lncRNA through the upstream, minor promoter reduces transcription of the protein-coding transcript from the second, major promoter

(Martianov et al. 2007). It is thought that the act of transcription of the noncoding transcript through the second promoter directly prevents transcription at this promoter, a process known as transcriptional interference. In this example the general transcription factor IIB (TFIIB) forms a DNA:RNA:protein triplex with the upstream-encoded lncRNA which inhibits TFIIB binding at the major promoter, thereby preventing its transcription.

In *Drosophila*, the most widely studied example of *cis*–regulatory ncRNAs is found at the bithorax complex (BX-C) shown in **Figure 1.2**. A number of different ncRNAs are transcribed across the various regulatory elements, but the results produced from different groups are inconsistent (Lempradl and Ringrose 2008). For example, *bxd* ncRNAs (shown in pink in **Figure 1.2**) were originally thought to positively regulate *Ubx* transcription as their expression is coincident in larval imaginal discs (Sanchez-Elsner et al. 2006). However, higher resolution *in situ* images have revealed that these ncRNAs are rarely expressed in the same cell as *Ubx* (Petruk et al. 2007).

One functional study has shown that ectopic expression in the embryo of an antisense ncRNA transcript spanning the *iab-7/8* region leads to misregulation of *Abd-B* – it becomes repressed in embryo parasegments PS10 and PS11, but ectopically activated in adult abdominal segments (Hogga and Karch 2002). Therefore, the behaviour of this ncRNA may be life-cycle stage-dependent. A similar transcribed element *Fab-7* has previously been shown to be able to act in this manner, but specifically during embryogenesis (Cavalli and Paro 1998). The *iab-7/8* region is, however, clearly involved in *cis*-regulation of the type

discussed above for *Airn* and *DHFR*, which confirms that this mechanism can occur in *D. melanogaster*.

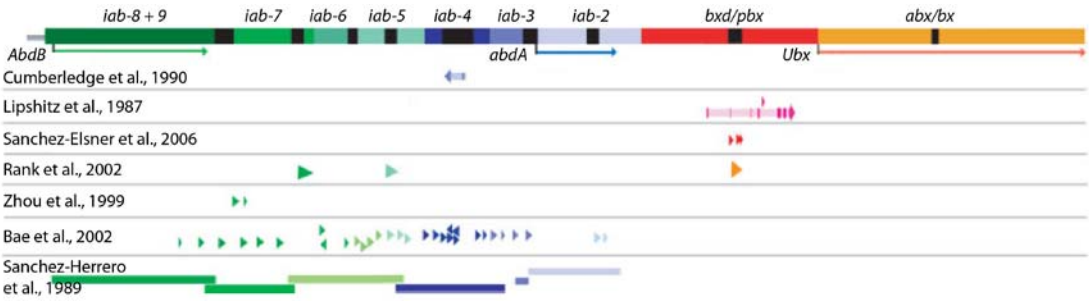

Figure 1.2 Noncoding transcription in the BX-C locus. The direction of transcription as detected by strand-specific probes is indicated by arrowheads. Where information on exons is available, these are indicated as darker bars within the transcript. Taken from Lempradl and Ringrose 2008.

Attempts have been made to identify examples of potential *cis*-regulation at a genomic scale. An analysis of cDNA collections identified several thousand *cis*-encoded antisense RNAs – RNA transcribed from the same locus as the target gene, but from the opposite strand – in several species, including 2,377 in *D. melanogaster* (Numata et al. 2007). Due to their antisense relationship, it was suggested by the authors that these ncRNAs suppress target gene expression by an RNAi-type mechanism but this has not been tested experimentally. These antisense RNAs in *Drosophila* frequently target genes which are involved in transcriptional regulation and genes whose products are found in the cell nucleus (Numata et al. 2007).

Mouse lincRNAs are similarly found near to genes annotated with specific GO terms and, in particular, to genes also involved in the regulation of transcription (Guttman et al. 2009; Ponjavic et al. 2009). Several of these

lincRNA-protein coding gene pairs have been subsequently investigated experimentally by *in situ* hybridisation (Ponjavic et al. 2009). They were shown to have similar expression patterns in the developing mouse brain, which suggests that the lincRNA may positively regulate the expression of its adjacent protein-coding gene.

Whether this type of *cis*-regulation by ncRNAs is a common phenomenon regulating many genes in *Drosophila*, or whether the BX-C example represents an unusual locus in this regard, remains to be comprehensively tested. I investigate this question in **Chapter 5**.

1.2.3 Enhancer RNAs

Similar *cis*-encoded RNAs have also been identified which are transcribed through gene enhancers and which may function in a positive manner to regulate target gene expression. An enhancer is defined as a DNA sequence which can increase a gene's transcription in *cis*, independently of its genomic orientation relative to that gene's promoter (Bulger and Groudine 2010).

A large set of ~2,200 ncRNA loci known as enhancer RNAs (eRNAs) has been defined in macrophage cells which overlap a set of enhancers and whose transcription appears to be associated with macrophage activation by endotoxin (De Santa et al. 2010). Potential eRNAs have been identified in *Drosophila* (Kharchenko et al. 2011) where, as in mouse, they are transcribed bidirectionally and need not be polyadenylated. The mechanism causing this phenomenon and whether the ncRNA is a cause or consequence of increased

transcription is unclear (Kim et al. 2010). Recent experiments on a set of human eRNAs have suggested that the mature RNA molecule is essential. SiRNA-mediated knockdown of seven (of 12 tested) putative eRNAs resulted in a significant disruption of transcription of a genomically adjacent protein-coding gene (Ørom et al. 2010). This behaviour seems to rely on the presence of the promoter of the nearby gene as, when this is removed at the *Arc* locus, the transcription of the related eRNA is disrupted (Ørom and Shiekhattar 2011). Much more comprehensive studies of eRNAs are required to reveal how many play such a functional role, and by what mechanism(s) this role is achieved.

1.2.4 Transcription Elongation

A lincRNA has also been defined which negatively regulates elongation of transcription beyond the gene promoter, rather than regulating the initiation of transcription as described above. *7SK*, a 330 bp (Zieve and Penman 1976) mammalian lincRNA, associates with the general elongation factor P-TEFb (Nguyen et al. 2001; Yang et al. 2001) to suppress the kinase activity of its Cdk9 subunit. Upon cellular stress, such as UV irradiation, P-TEFb kinase activity increases, which is thought to be due to the release of *7SK* from the complex (Blencowe 2002). The downstream genes which are affected by this remain to be identified.

1.2.5 Post-transcriptional Regulation

LncRNAs have also been discovered which are involved in gene regulation beyond the initial act of transcription. Here, I describe the involvement of ncRNAs in regulating gene expression through their interaction with RNA-processing factors in response to cellular stress, in both humans and *Drosophila*. I also briefly mention two examples of regulation of individual protein-coding genes, at the level of mRNA splicing and protein localisation within the cell.

In *D. melanogaster*, three noncoding heat-shock response (hsr-ω) transcripts are induced from the 93D puff by heat shock, CO_2 exposure and after the release of ecdysone in third instar larvae (Lakhotia and Sharma 1996). This 10-20 kb locus is functionally conserved in all *Drosophilid* species. One short transcript is cytoplasmic, while the other two remain at the locus from which they are transcribed and within nuclear 'omega' speckles which are thought to be storage sites for RNA-processing proteins (Prasanth et al. 2000). Although they are upregulated in response to stress, these transcripts must also play a housekeeping role in the developing animal as only 20-25% of trans-heterozygote mutant embryos hatched (Mohler and Pardue 1982).

In human cells, non-coding transcription from the sat III repeat sequences is observed upon heat shock (Rizzi et al. 2004). These transcripts also remain associated with the locus after transcription, but are not expressed constitutively. They also associate with several RNA-processing factors (Jolly

and Lakhotia 2006), some of which are the same as in *Drosophila*, e.g. heat-shock factor 1 (HSF1).

In both this example, and that of the dosage compensation ncRNAs (**Chapter 1.2.1**), the ncRNAs function by similar mechanisms: the *roX1* and *roX2* loci in *Drosophila* and *Xist* in eutherian mammals coat the X chromosome to aid in equalling the level of transcription from this chromosome in both sexes, while both the hsr-ω and sat III ncRNAs are upregulated in response to heat shock and remain at the locus from which they are transcribed. These ncRNAs may share a common evolutionary origin. Alternatively, they may instead represent convergent evolution events as suggested by their lack of sequence similarity. Due to this ambiguous evolutionary history, I term such ncRNAs not as orthologues, but as analogues. I discuss further this concept of analogy in the context of lincRNAs in **Chapters 4** and **5**.

A lncRNA which is encoded in the antisense orientation to the transcription factor *Zeb2* causes upregulation of *Zeb2* when ectopically expressed in human epithelial cells (Beltran et al. 2008). This is thought to be achieved by binding of the lncRNA to a splice site within the long 5' UTR of the *Zeb2* mRNA, which prevents splicing at this site and preserves a necessary internal ribosome entry site for translation (Beltran et al. 2008).

The behaviour of proteins themselves can also be regulated by lincRNAs, as exemplified by the ncRNA repressor of the nuclear factor of activated T-cells (*NRON*). This lincRNA contains three exons and is alternatively spliced to

produce transcripts varying in length from 800 bp to 3.7 kb (Willingham et al. 2005). Through its interaction with the nuclear import factor KPNB1, *NRON* contributes to preventing the protein product of nuclear factor of activated T-cells (*NFAT*) from entering the nucleus which, in turn, prevents NFAT from promoting transcription of its target genes (Willingham et al. 2005).

1.3 Several lncRNAs have been implicated in different forms of cancer.

LncRNAs are now known to be involved in various cellular mechanisms (discussed in **Chapter 1.2**). This suggests that, when their loci are disrupted, they may give rise to an abnormal phenotype. In humans, this might be embodied through an involvement in disease. In fact, lncRNAs may be involved in various nervous system diseases through their association with protein-coding genes which are important in diseases such as Fragile X syndrome and Alzheimer's disease (Qureshi et al. 2010). Several lncRNAs have also been implicated in cancer which might be expected as their principal role appears to be in regulating gene expression, a dysregulation of which is a hallmark of tumourigenesis (Alberts et al. 2002). In the following section, I discuss three specific loci as examples of disease-associated lincRNAs – *ANRIL*, *MALAT-1*, and *PTENP1* – which are now recognized as being involved in the pathogenesis of different, but overlapping, types of cancer.

1.3.1 ANRIL

ANRIL is transcribed into a 3,834 bp RNA molecule made up of 19 exons (Pasmant et al. 2007) and is found within the 9p21.3 locus. This locus contains over 50 variants at single positions, known as single nucleotide polymorphisms (SNPs), which are associated with susceptibility to a variety of complex diseases, including coronary artery disease, ischemic stroke, aortic aneurysm, type II diabetes, glioma, several carcinomas, malignant melanoma, and acute lymphoblastic leukaemia (summarised in Cunnington et al. 2010). The 9p21.3 locus also contains two cyclin-dependent kinase inhibitor genes known as *CDKN2A* and *CDKN2B*. Recombinant mice lacking a 70 kb region orthologous to 9p21.3, which includes the 3' end of *ANRIL*, show an increased mortality during development and as adults (Visel et al. 2010).

Those SNPs which have been associated with disease are more strongly associated with *ANRIL* than *CDKN2A* or *CDKN2B* expression (Cunnington et al. 2010). The expression of all three genes in the locus is positively correlated which suggests they may be co-regulated. Certain SNPs appear to have opposing effects on *ANRIL* and *CDKN2B* expression (Cunnington et al. 2010) which may be because *ANRIL* overlaps *CDKN2B*, but is transcribed from the antisense strand. Antisense transcription from *CDKN2B* has previously been shown to downregulate it (Yu et al. 2008). *ANRIL* and *CDKN2A* share some promoter sequence, including a common CTCF-binding site (Rodriguez et al. 2010), which may account for the correlated effects of SNPs on their expression. It has been speculated that this regulation of other genes by

ANRIL may contribute to cellular aging and, thereby, its involvement in several, seemingly unrelated diseases (Pasmant et al. 2011). *ANRIL* may be regulated by a Polycomb-mediated mechanism as Chromobox 7 (CBX7), which is part of the Polycomb repressive complex 1 (PRC1), binds directly to the *ANRIL* RNA (Yap et al. 2010).

1.3.2 MALAT-1

MALAT-1 (metastasis associated in lung adenocarcinoma transcript-1) expression is found in a wide range of tissues but its expression is significantly greater in metastatic non-small cell lung cancer (NSCLC) than in tumours which did not metastasise (Ji et al. 2003). High expression of *MALAT-1* is also related to a reduced survival rate of patients with stage I NSCLC (Ji et al. 2003). Ji et al. identified two alternative isoforms of *MALAT-1* whose transcripts are ~8 kb long and ~70% identical when aligned between human and mouse. Since this study, *MALAT-1* has been found to be similarly upregulated in other carcinomas, including breast, pancreas and colon cancer (Lin et al. 2007), which suggests that *MALAT-1* may be generally important in carcinogenesis. *MALAT-1* is retained in the nucleus, where it influences alternative splicing of hundreds of transcripts by regulating the localisation of multiple pre-mRNA splicing factors such as the serine/arginine splicing factor SF1 to nuclear speckles (Tripathi et al. 2010).

1.3.3 PTENP1

PTENP1, a non-coding processed pseudogene of *PTEN*, has been identified as a tumour suppressor gene, in that it appears to prevent cancer formation (Alberts et al. 2002) Through its widespread homology to *PTEN* and conservation of miRNA-binding sites in the 3' UTR region, this lncRNA is able to positively regulate *PTEN* expression, perhaps by acting as a sponge for negative regulatory miRNAs (Poliseno et al. 2010). Furthermore, several independent patients with sporadic colon cancer have been identified with DNA losses at the *PTENP1* locus which are related to dysregulation of *PTEN*, demonstrating that *PTENP1* is clinically relevant (Poliseno et al. 2010). This mechanism of gene regulation by non-coding pseudogenes may be more widespread, as five other pairs of genes and pseudogenes were identified with similarly conserved miRNA binding sites in this study and thousands of pseudogenes have been found throughout the human genome (Zhang et al. 2003).

1.4 Genome-wide analyses of lncRNAs

In recent years attempts have been made to move away from individual case studies of lncRNAs and take a more comprehensive view of the importance of lncRNAs. While all genomes studied to date contain a considerable number of lncRNA loci, the number of these which are biologically functional has proven to be particularly controversial. Research in this field has taken two distinct

approaches – computational and experimental analyses – which I describe in more detail below.

1.4.1 Computational analyses of lncRNA function

LncRNAs may form three-dimensional secondary structures which could be required for the mature RNA molecules to exert their function. As such, various lncRNA sets have been screened for sequence signatures to suggest they form such structures (the methods required to do this are described in **Chapter 2.1.7**). It has been predicted that there are up to 10,000 conserved structural RNA loci in *Drosophila* (Rose et al. 2007) although these in total cover only 5 Mb – a relatively small portion of the widespread transcription observed in the genome (Manak et al. 2006). The ENCODE pilot study, which studied 1% of the human genome (Birney et al. 2007), predicted between 1,500 and 1,800 structured regions. However, there was little overlap in the structure predictions produced using different methods (Washietl et al. 2007) which suggests that they contain a large number of false positive predictions.

An alternative, and complementary, approach to discriminate between functional and non-functional lncRNAs is to analyse their evolutionary signals when comparing their sequences between related species. This approach assumes that, if a locus is functional, deleterious mutations which disrupt this function will be preferentially purged from the population. Such sequences will therefore be better conserved relative to neutral sequence for species in which they are functional. While some of the most well studied lncRNAs are known

to be poorly conserved between species (Pang et al. 2006), it is still true that those lncRNAs whose primary sequence is important for their biological function should display evolutionary conservation.

Much work has focused on lincRNAs transcribed from the mouse genome, with several studies using cDNA evidence produced by the FANTOM consortium. An initial study of 60,770 cDNA clones predicted 4,280 strong lincRNA candidates (Okazaki et al. 2002). Of these, 454 (10.6%) contained sequences which mapped to the human genome and it was suggested that these conserved loci, if not more, were likely to encode genuine functional lincRNAs. It was noted that these loci shared similar frequencies of substitution and indel mutations with intergenic sequences (Wang et al. 2004), which are assumed to contain a low density of functional elements. However, when accounting for local substitution rates and the large-scale variation in G+C content observed within the mouse genome (Eyre-Walker and Hurst 2001), a similar set of lincRNAs was shown to be evolutionarily conserved when compared to neighbouring presumed neutral sequence (Ponjavic et al. 2007). LincRNA promoters are particularly highly constrained, suggesting that it could be the act of transcription itself which is more often important for these lincRNAs functionality (Carninci et al. 2005; Ponjavic et al. 2007).

A second set of 1,675 mouse lincRNAs was defined using chromatin markers for active promoters (H3K4me3) and actively transcribed exonic sequence (H3K36me3) (Guttman et al. 2009). When analysed in a similar manner, this set was shown to be similarly constrained to those lincRNAs defined above

using the FANTOM sequences (Marques and Ponting 2009), suggesting that they contain a similar proportion of functional lincRNA loci. However, only 11% of these loci were also found in the previous set. This difference may be because the two sets were defined using different tissue and cell culture samples and it may be that both have only sampled a small proportion of a much larger mouse lincRNA repertoire that remains to be fully defined (Marques and Ponting 2009).

Attempts to similarly describe the lincRNA complement of *D. melanogaster* have suffered from a number of limitations. One of the first (Tupy et al. 2005) discovered 69 lincRNAs (although they call them mRNA-like ncRNAs) from a set of fewer than 8,000 cDNA sequences by considering only intergenic sequences over 200 bp from the nearest gene and with no significant ORF $>$ 100 amino acids. From this, it was predicted that there were no more that 50-100 lincRNA genes in the *D. melanogaster* genome. A more recent study (Hiller et al. 2009) identified conserved introns between Drosophilid species and used these to define novel transcripts. 129 introns were discovered within novel lincRNAs, of which 29 are conserved beyond the Sophophora group of species (see **Chapter 1.5.1**). However, this approach only provides a partial structure where the full extent of the transcript remains unknown, and the expression across only seven of these introns (58% of the 12 which were tested) was verified by RT-PCR. A screen of intergenic expression by RT-PCR suggested that there may be up to 8,350 ncRNAs in the *D. melanogaster* genome (Li et al. 2009). They found many short ($<$ 120 nt) transcripts that

were not found by tiling microarray experiments (Manak et al. 2006). The contribution of these ncRNAs to lincRNA (> 200 nt) transcription thus remains unclear.

In *Drosophila*, there has been little research into evolutionary conservation amongst lincRNAs, beyond looking for presence/absence of individual sequences in related species. One study of 136 lincRNAs (Inagaki et al. 2005), also defined using cDNA evidence, compared lincRNAs to protein-coding genes and found that fewer lincRNAs (94, 69%) than protein-coding genes (8,881, 85%) were conserved between *D. melanogaster* and *D. pseudoobscura*. These two species are quite distant relatives in the *Drosophila* clade (see **Chapter 1.5.1**), having diverged approximately 25 million years ago. The substitution rate at synonymous sites (d_S) within orthologous genes between these species is estimated to have a median value greater than one (Heger and Ponting 2007) which implies that it may be difficult to align noncoding bases outwith of protein-coding regions. The evolutionary constraint of lincRNAs in *Drosophila* therefore remains very much an open question.

These types of large-scale cDNA and EST collections are thought to remain one of the better ways to identify novel lincRNAs (Xue and Li 2008), particularly as these loci often show relatively low and restricted expression profiles, and may therefore be missed by other technologies, such as genome-wide tiling arrays, which struggle to capture very lowly expressed transcripts (van Bakel et al. 2010). I use this type of data to define a novel set of lincRNA loci in *D. melanogaster* in **Chapter 3**.

1.4.2 Experimental analyses of lncRNA function

The majority of high-throughput experiments that test the biological roles of lncRNAs have assumed that their regulated expression is an indicator of functionality. A study of 1,602 mouse lncRNAs identified 178 which are differentially expressed between tissues and 70 which respond to lipopolysaccharide stimulation in macrophage cells (Ravasi et al. 2006). A similar study found 174 lncRNAs whose expression changed significantly during mouse embryoid body differentiation (Dinger et al. 2008). These data also suggested that intronic and bidirectional (those transcribed in a head-head orientation within 1 kb from a neighbouring protein-coding gene) lncRNA expression tends to be correlated with that of the associated protein-coding gene, although there was no relationship between protein-coding genes and *cis*-encoded antisense lncRNAs. A survey of *in situ* hybridisation images from the adult mouse brain collected by the Allen Mouse Brain Atlas (Lein et al. 2007) revealed 849 lincRNAs that are expressed in the brain, 513 of which showed distinct regional patterns of expression (Mercer et al. 2008). Unlike the embryoid body study, a variety of patterns was observed for bidirectional lncRNAs and protein-coding genes, which implies that any related expression is not simply a consequence of open chromatin, but could be functionally important.

As for the computational approach (**Chapter 1.4.1**), efforts have been made to identify lncRNAs which are evolutionarily conserved between species and therefore potentially functional. Detailed analysis of four mouse lincRNAs

revealed that their brain expression patterns can be conserved between diverse vertebrates (Chodroff et al. 2010). A custom microarray has been used to identify 6,923 lncRNAs in mouse, and eight of these were validated by northern blotting (Babak et al. 2005). Five of these eight were also found to be expressed in rat, but none were found in any of the human tissues or cell lines tested. Over a much smaller evolutionary distance, between human and chimp, a similar degree of conservation of intergenic and exonic transcription was observed using tiling arrays designed to target 1% of the human genome (Khaitovich et al. 2006). The least divergence in expression was seen for brain samples, with the greatest divergence detected in testis. That similar patterns were seen for exonic and intergenic transcription, most of which is thought to consist of ncRNAs, suggested that the majority of these ncRNAs are likely to be functional, but this functionality may not be conserved over the longer evolutionary timescales studied by Babak et al.

1.5 *Drosophila* as a model organism

In this thesis, I have chosen to use *Drosophila* as a model organism to study lincRNAs and their potential functionality. Much of the work discussed above, particularly on a genome-wide scale, has been done on mammalian species, including mouse and human, which has led to the suggestion that lincRNAs may be specific to species such as these with a complex genome organisation (Prasanth and Spector 2007). My research tests this hypothesis by searching for lincRNAs and then investigating their functionality in a species distantly

related to mammals – the fruit fly *D. melanogaster*, which diverged from the mouse lineage approximately 700 million years (Ponting 2008). Throughout this thesis, I will refer to *D. melanogaster* when I wish to describe this species explicitly, but I will also use the term *Drosophila* to refer to features common to all Drosophilid species. *Drosophila* was chosen as the study organism for both the computational and experimental work which is described in this thesis for a number of reasons.

1.5.1 Computational advantages of Drosophila

The *Drosophila* genome is relatively small and tractable (Celniker and Rubin 2003), and a large proportion (at least 47%, Meader et al. 2010) of it is thought to be functional. The sequenced regions of the *D. melanogaster* genome, which contain both the euchromatic and the majority of the largely repetitive heterochromatic regions, comprise only 120 Mb (Tweedie et al. 2009). Unlike most mammalian genomes (Lander et al. 2001), *Drosophila* carries a small amount of repetitive elements, where less than 6% of the *D. melanogaster* genome is thought to derive from repetitive transposable-element derived sequence (Bergman et al. 2006). These regions are difficult to assemble into a genome (Phillippy et al. 2008) and are thought unlikely to harbour many functional elements.

Much of the non-coding DNA in the *Drosophila* genome shows signs of functionality, suggesting that any lincRNAs within these regions may also be functional. 22-26% of non-coding sequences are highly conserved between *D.*

melanogaster and *Drosophila virilis* (Bergman and Kreitman 2001) while studies on the more closely-related species *D. melanogaster* and *D. simulans* have suggested that approximately 50% of non-coding DNA may be constrained and therefore functional in *D. melanogaster* (Andolfatto 2005; Meader et al. 2010). Furthermore, up to 20% of differences in intergenic sequence between *D. melanogaster* and *D. simulans* may have been driven by positive natural selection (Andolfatto 2005) which is consistent with the sequences in which they reside also being functional.

The *Drosophila* genome has been extensively studied and, as such, there are a wide range of computational datasets that can aid with both identification and analysis of potential lincRNA functionality. The *Drosophila* bioinformatics web service FlyBase (Tweedie et al. 2009) publishes genome-wide annotations and regular updates of these, as well as other genomics resources for *Drosophila*, such as the locations of all mapped transposable element insertions. Recently, the modENCODE collaboration (Celniker et al. 2009) has begun publishing a range of genome-wide datasets including characterisation of regulatory elements, chromosome proteins, small RNAs and the origins of DNA replication. Specifically, they have also produced a large amount of RNA-seq data describing the *D. melanogaster* transcriptome, which I analyse in detail in **Chapter 4**.

The genomes of 12 *Drosophila* species have been sequenced (Clark et al. 2007) which, when coupled with their well-established phylogeny (**Figure 1.3**, (Powell 1997), has made a range of comparative evolutionary studies in

Drosophila possible. These genome sequences have been used for, amongst other things, protein-coding gene orthology assignment (Heger and Ponting 2007), identification of functional elements (Stark et al. 2007), and miRNA prediction (Stark et al. 2007). *Drosophila* has also been used for population genomics studies, in *D. melanogaster* (Sackton et al. 2009) and *D. simulans* (Begun et al. 2007).

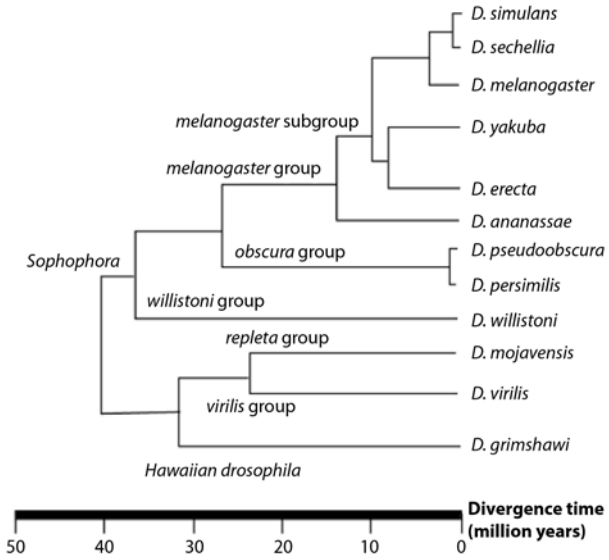

Figure 1.3 Phylogeny of 12 sequenced Drosophilid genomes (taken from the *Drosophila* Assembly/Alignment/Annotation website; http://rana.lbl.gov/drosophila/graphics/tree.gif).

1.5.2 Experimental advantages of *Drosophila*

Drosophila also possesses a number of features which makes it a useful model organism for experimental genetic investigation. It is small, can be cultured on simple media and is highly fertile, with a short life cycle (**Figure 1.4**) that takes only nine days to go from the egg to adult stage (Alberts et al. 2002). *Drosophila* is a reasonably complex organism with a functioning nervous

system, which makes it ideal to study the function of genes such as lincRNAs in complex tissues.

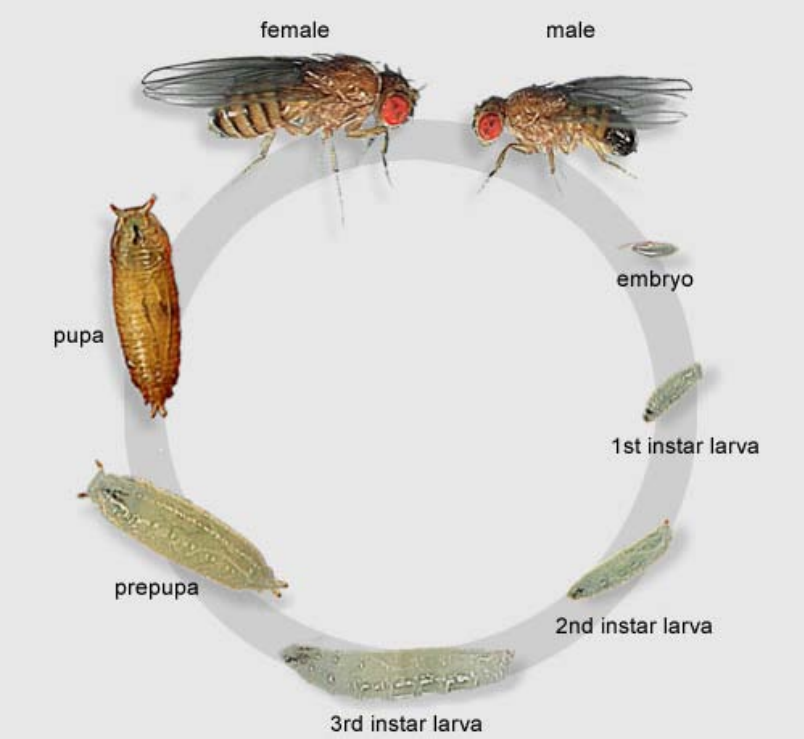

Figure 1.4 *Drosophila* life cycle. Taken from Wolpert et al. 2006.

These advantages have meant that *Drosophila* is the model genetic organism which has been in use for the longest period of time (Alberts et al. 2002) and this has led to a number of experimental tools being developed. For example, various sets of molecular deletions have been generated which are estimated to cover approximately 80% of the *D. melanogaster* genome (Venken and Bellen 2005) while the Berkeley *Drosophila* Genome Project (BDGP) is aiming to generate a single transposable element mutation in every *D. melanogaster* transcriptional unit (Spradling et al. 1999). The database resulting from this project is queried in **Chapter 6** to identify putative mutations in lincRNA promoters. More complex genetic tools also exist, such as the frequently used

UAS/GAL4 system, which allows one to drive gene expression in a developmental- or tissue-specific manner (Brand and Perrimon 1993).

1.6 Project Aims and Thesis Structure

The aims of this thesis are to search for lincRNA loci in the *D. melanogaster* genome and to propose to what extent these loci are functional, using both computational and experimental approaches. Cursory inspection of the FlyBase genome browser shows that transcription takes place outside annotated gene models in even a well-studied model organism such as *D. melanogaster* but the frequency and biological relevance of this has remained largely unclear.

In **Chapter 3**, I define a set of 2,788 lincRNA loci using EST, mRNA and cDNA evidence. Of these, 1,411 show evidence of being evolutionarily constrained between *D. melanogaster*, *D. simulans* and *D. yakuba*. These loci are also enriched for several other indicators of functionality and predicted RNA secondary structures which suggests that they may function at the level of the mature RNA molecule. A second set of 241 lincRNA loci are identified as being fast-evolvers which may frequently represent the primary transcripts of esiRNA and piRNA sequences.

RNA-seq data are used in **Chapter 4** to identify a second distinct set of 1,119 lincRNA loci. The vast majority of these are multi-exonic and evolutionarily constrained, which again suggests that they may be functional. This constraint increases with increasing specificity of expression and many of these loci are

found within chromatin domains containing specifically expressed, developmentally regulated genes which suggests that these lincRNAs may be important regulators of development. A large number (151) of lincRNAs are male-specific but, unlike their protein-coding gene counterparts, they do not appear to be experiencing sexual selection pressures. Finally, I identified a significantly greater number than expected (42) of putative analogous lincRNAs whose genomic positions are orthologous in *Drosophila* and mouse. These pairs of lincRNAs may share a common function in the two species.

In **Chapter 5**, I investigate the possible function of the lincRNAs identified in **Chapter 3** in more detail. These lincRNAs are shown to be enriched in the genomic neighbourhoods of transcription factor genes, which suggests that lincRNAs may play a general role in regulation of this type of gene when transcribed from adjacent loci. Four lincRNAs share a ubiquitous expression profile with their neighbouring transcription factor. I validate a functional relationship for one pair – the transcription factor *Distal-less* (*Dll*) and a lincRNA which I name *dEvf-2* – using RNAi to show that the lincRNA positively regulates expression of the mRNA of the transcription factor. I also define the structure of this locus by RACE.

I use a reverse genetics approach in **Chapter 6** to identify transposable element mutations in the putative promoters of lincRNAs defined in **Chapter 3**. TSSs containing canonical promoter motifs near to lincRNA loci are defined and cross-referenced with the mutants stored in the Bloomington Stock Centre. Five mutations are identified but none of these significantly reduce the

expression of their associated lincRNA; only one has any effect and this is actually to up-regulate lincRNA transcription.

Finally, in **Chapter 7** I discuss the significance of the work undertaken here and the potential for future experiments based on the findings presented in this thesis.

Chapter 2: MATERIALS AND METHODS

In this thesis I utilise a variety of datasets, both computationally and experimentally derived, which I describe in **Chapter 2.1**. The various methods that I use to analyse these data, again both computational and experimental, are described in **Chapter 2.2**. Individual data and methods which are used in only one project are described in their relevant chapter.

2.1 Datasets

2.1.1 *D. melanogaster* Transcriptome Data

2.1.1.1 mRNA Sequences

21,863 mRNA sequences available for *D. melanogaster* were obtained from GenBank (Benson et al. 2005) via the UCSC Genome Informatics FTP server on 4 January 2008. These had been produced from RNA extractions (**Chapter 2.2.6.1**) of a variety of tissues, which were then reverse transcribed into double-stranded cDNA. The cDNA libraries were then cloned into a bacterial vector where the sequencing reaction proceeded based on established primer sequence found near to the cloning site in the vector (Adams et al. 1993).

2.1.1.2 Expressed Sequence Tags

All 542,688 expressed sequence tag (EST) records for *D. melanogaster* were also obtained from GenBank on 4 January 2008. These had been generated in the same way as mRNA sequence, but are the products of only one-pass of

sequencing using universal primers (Putney et al. 1983) and so are generally of lower quality. Most EST sequences are 500-800 bp long, but need not necessarily represent the full-length, or either of the ends, of an RNA molecule.

2.1.1.3 **Full-length cDNA Clones**

A set of 11,040 full-length cDNA clones was obtained from the BDGP (Stapleton et al. 2002). These clones had also been generated in a similar manner to the mRNA sequences deposited in GenBank and then sequenced from both ends. However, as the transcript can be longer than the length of the two sequences combined, these records need not contain comprehensive information regarding the structure of these transcripts.

2.1.1.4 **RNA-seq dataset**

RNA-seq reads from 30 developmental time points were acquired from the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra?term=srp001065). Each sequencing run was available as a single FASTQ file, or as two linked files for paired-end reads.

Briefly, these reads had been generated using the following protocol. RNA was extracted from the relevant tissue as in **Chapter 2.2.6.1** and split into ~200 bp fragments by centrifugation. These fragments were reverse transcribed into cDNA and PCR primer adaptors ligated onto both ends of the fragment as shown in **Figure 2.1**. The cDNA was attached to a flow cell densely coated in primers complementary to the primers ligated onto the fragments such that

individual fragments bound to isolated regions of the cell. This was followed by several rounds of DNA amplification to increase the DNA concentration on the flow cell. As shown in **Figure 2.1**, both ends of newly amplified DNA fragments bound to primer sequences on the flow cell. Finally, the double-stranded DNA (dsDNA) was denatured to produce single-stranded DNA and attached to the slide by only one adaptor ready for the sequencing reaction.

The cDNA samples produced here were sequenced using the Illumina Genome Analyser IIx which uses a cyclic reversible termination reaction. The sample was washed with four fluorescent modified nucleotides, where only one nucleotide can be incorporated onto each template. The remaining nucleotides were washed off, and the resulting image captured to record the identity of the incorporated nucleotides at each position. A cleavage step followed which removed the terminating group from the added nucleotide and allowed a second nucleotide to be added in subsequent steps. This cycle was repeated a number of times (n = 75, to generate the 75 bp reads analysed here) to record the sequence at each position along the DNA fragment.

For paired-end reads, the DNA fragments were then attached back to the slide using the other adaptor, and denatured from the first adaptor. Sequencing then proceeded as above from the opposite end of the fragment.

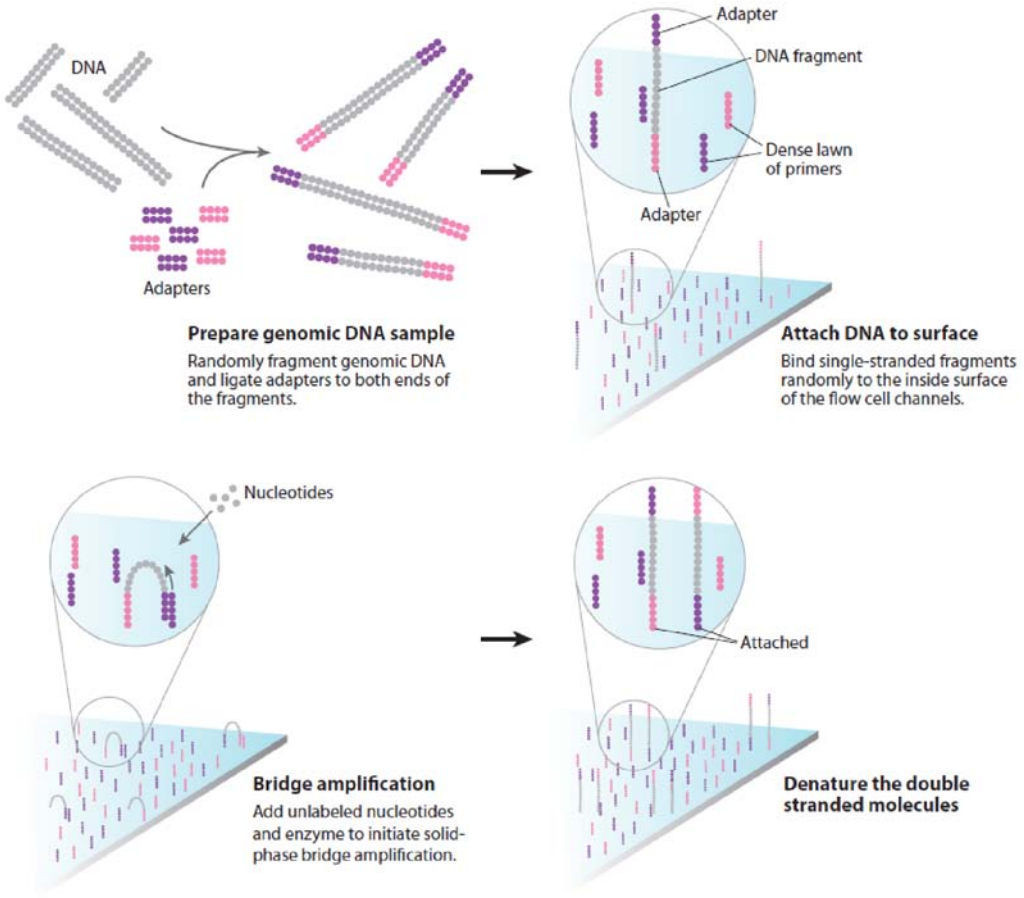

Figure 2.1 Preparation of DNA samples for Illumina GAIIx sequencing. Taken from Mardis 2008.

The data are presented in a FASTQ file which consists of the recorded DNA sequences and their associated quality scores which represent the probability that each base has been called incorrectly, based on the recorded images.

2.1.2 *Drosophila* Genome Sequence and Annotation

D. melanogaster was the first animal with a nervous system to have its genome sequenced (Myers et al. 2000), and the first to be sequenced using whole-genome shotgun technology (Venter et al. 1998). This process required shearing genomic DNA into randomly sized fragments before cloning and

sequencing. Computer algorithms were then used to recognise shared sequence between fragments which arise from them overlapping the same regions of the genome. By extending this process, the entire sequence along each chromosome arm was assembled. Subsequent releases have increased the sequence quality, removed gaps from the assembly and increased the coverage of heterochromatic regions (Celniker and Rubin 2003). This work exclusively uses the third build of the *D. melanogaster* genome from April 2006, which is known asBDGP5.

The genomes of *D. yakuba and D. simulans* were sequenced using the same whole-genome shotgun approach, as part of the *Drosophila* 12 Genomes Consortium (Clark et al. 2007).

Unless stated elsewhere, *Drosophilid* genome sequences and all other annotations of the *D. melanogaster* genome were obtained from the *Drosophila* bioinformatics database FlyBase, Release 5.4 (Tweedie et al. 2009). These include, for example, the structures of previously annotated gene models, their functions, and the locations of mapped transposable element insertions. The information provided by these annotations comes from a number of sources – experimental, manual curation, or automatic curation. I do not discriminate between these, instead treating all sources as equally valid.

2.1.3 Whole-genome alignments – BLASTZ, Chaining and Netting

Most research into the evolutionary processes which have shaped DNA sequences is inherently comparative, in that it involves comparing two related sequences to determine the changes that have occurred since they diverged from their common ancestor. A first step to performing analyses at a genome-wide scale must be the identification of which sequences in the genomes under consideration are truly related, i.e. are homologous.

In **Chapters 3** and **4** I make use of whole-genome alignments between D. *melanogaster* and *D. simulans* or *D. yakuba* which are available as axt alignment files from the UCSC Genome Informatics FTP server (http://hgdownload.cse.ucsc.edu/downloads.html#fruitfly). Species more distantly related to *D. melanogaster* are not considered since alignments of noncoding DNA are less reliable due to recurrent and back mutations over these greater evolutionary distances (Pollard et al. 2004).

These alignments were originally produced using the BLASTZ algorithm (Schwartz et al. 2003) which uses a modification of the gapped BLAST approach (Altschul et al. 1997). This involves finding near-exact short sequence matches and then extending them in both directions, first without gaps, and then allowing gaps in the alignment.

Specifically, BLASTZ records all 12-mers that are identical between the two sequences, while allowing for one transition mutation. These are then extended

in each direction while prohibiting gaps from being introduced into the alignment. The alignment is scored at each stage of this extension, where identical positions are given positive scores and non-identical positions receive negative scores. The value of these scores depends on the individual nucleotides, e.g. an AC alignment position corresponds to -114 while AG scores -31. The total alignment score is adjusted for a measure of sequence complexity (Chiaromonte et al. 2002) to prevent this score being artificially inflated by regions of severely biased nucleotide content. Extension continues until this summed score for the alignment decreases below a certain threshold.

If this score is above a second threshold (2,200 for alignments between *D. melanogaster* and *D. simulans* or *D. yakuba*) then the extension process from the original almost identical 12-mer is repeated, but allowing for gaps in one or more of the sequences being aligned. Like nucleotide substitutions, gaps decrease the overall alignment score: a gap of length k decreases the alignment score by $400 + 30k$. The minimum score threshold for retaining an extended gapped alignment is generally higher than that for an ungapped alignment, being 4,000 when aligning *D. melanogaster* with *D. simulans* or *D. yakuba sequences*.

BLASTZ next attempts to join adjacent alignments by repeating this process of finding short matches in the intervening sequence and then extending them, but with slightly altered parameters. Requiring the match to extend for only 7 positions, rather than 12, increases the sensitivity of finding almost exact matches. The thresholds for retaining an extended alignment are also reduced.

This process produces a set of alignment 'islands' which are fragmented across the genome and need to be connected to generate a complete whole-genome alignment. This problem can be solved by the application of the AXTCHAIN program (Kent et al. 2002). This chains alignments together by allowing them to be separated by gaps in the sequences. Only genuine gaps are allowed by the requirement that the order of these alignments in both species is conserved across the gaps.

The CHAINNET program (Kent et al. 2002) is then used to resolve situations where the same sequence in one genome is aligned to multiple positions in another genome. The chains generated above are grouped into longer sections of presumed conserved synteny, with chains retained where they give the highest alignment score. Gaps which occur between high quality (i.e. high scoring) chains can be filled in with lower quality (i.e. lower scoring) chains and in this way a hierarchy of alignments across the genome is produced.

2.1.4 Short Intron Alignments

When attempting to identify functional sequences in the genome, it is important to compare their characteristics to sequences which are assumed to be free of functional constraint.

Ancestral repeats are one such type of sequence which is frequently used to represent neutrally evolving and assumed functionless DNA (Lunter et al. 2006). These are repeat elements (also known as transposable element-derived sequence) found in orthologous locations in all species under consideration and

which are thought to have been present in the common ancestor of these species. However, in *Drosophila*, transposable elements have been identified that show evidence of adaptive, rather than neutral, evolution (González et al. 2008). They are therefore an unreliable source of neutral sequence in *Drosophila* genomes.

Instead, short introns are used as the best proxy for neutral sequence. These introns (≤ 86 bp) exhibit a similar substitution rate to synonymous sites (Haddrill et al. 2005), but are not expected to experience the same, albeit relatively weak, selection for translational efficiency (Akashi 1994). 128 alignments of such short introns in *D. melanogaster* with their orthologous sequence in *D. simulans* were obtained directly from Haddrill et al. I determined the orthologous intron sequence in *D. yakuba* using the same protocol as Haddrill et al. by mapping the *D. melanogaster* sequences onto the *D. yakuba* genome with BLAT (**Chapter 2.2.1.1**) and then retaining only the 97 uniquely mapping sequences. The six bases at the 5' end and the 16 bases at the 3' end of each intron were removed since they are required for correct splicing of the intron and hence do not evolve neutrally. The short length of these alignments makes it impossible to accurately estimate their evolutionary parameters using baseml. Consequently, I concatenated them separately for each species to generate a single neutrally-evolving alignment against *D. melanogaster* of ~5 kb and ~4 kb for *D. simulans* and *D. yakuba*, respectively.

2.1.5 Indel-Purified Segments (IPSs)

One way in which DNA sequences can evolve is through the acquisition or loss of particular regions, a process which results in gaps in the alignment when two related sequences are compared. These gaps are referred to as 'indels' or insertion-deletion mutations because, when considering only two sequences, it is not possible to infer whether there was an insertion mutation in one lineage, or a deletion in the other. Like other mutations, it is generally assumed that these mutations will tend to be deleterious to the sequences which contain them and will be preferentially purged from functional sequence.

This work looks at long inter-gap segments (IGSs), known as Indel-Purified Segments (IPSs), which do not tolerate such indel mutations. These had previously been defined using genome-wide alignments between *D. melanogaster* and *D. simulans*. Here, the length of IGSs which do not deviate from the predicted neutral geometric distribution has been estimated to be 15-55 bp (Meader et al. 2010). By applying a false discovery rate of 10%, Meader et al. showed that 84 Mb of the *D. melanogaster* genome is covered by 335,675 IPSs. I consider the distribution of their IPSs within different sequence classes in **Chapters 3** and **4**.

2.1.6 Multi-Species Conserved Segments (MCSs)

Sequence evolution can also be investigated by examining multiple sequences simultaneously in the same alignment. This contains more information than simple pairwise alignments of the type described above (Dubchak and Frazer

2003), since it covers a greater evolutionary time for changes between sequences to have occurred. Also, if the phylogeny of the species is known, it is possible to infer the ancestral state and therefore which changes have since arisen in which lineages (e.g. Feng-Chi Chen et al. 2007).

These advantages are exploited by the PhastCons program (Siepel et al. 2005) which can be used to define deeply conserved regions known as Multiple-species Conserved Segments (MCSs) across a multiple genome alignment. In this work, I have examined a published set of 1,080,980 MCSs which cover 49 Mb (41%) of the *D. melanogaster* genome. These are available from the UCSC Genome Informatics FTP server and were based on a newly available 15-way genome alignment of 12 *Drosophila* genomes (Clark et al. 2007), and genomes of the malarial mosquito *A. gambiae* (Holt et al. 2002), the honey-bee *Apis mellifera* (Honeybee Genome Sequencing Consortium 2006), and the red flour beetle *Tribolium castaneum* (Richards et al. 2008).

2.1.7 Predicted RNA Secondary Structures

One line of evidence that an RNA molecule is functional would be its ability to fold up in three-dimensional space into what is known as a secondary structure. This would suggest that it is the mature lincRNA molecule, rather than the act of transcription at its locus, which is required for its biological activity. This structure may be required for the correct functioning of the molecule, e.g. by binding with a protein partner. Here, I discuss two

computationally derived sets of RNA secondary structure predictions which I compare to my lincRNA annotations in **Chapters 3** and **4**.

2.1.7.1 **EvoFold**

The EvoFold predictions were made using the EvoFold algorithm (Pedersen et al. 2006) which infers secondary structures based on the entire space of all possible structures within a set of MCSs, described in **Chapter 2.1.6**Maximum likelihood analysis is used to determine if substitutions in the multiple alignment are compensatory, i.e. they maintain the predicted structure of the folded RNA molecule. EvoFold has high sensitivity, but relatively low specificity (Pedersen et al. 2006).

In the *Drosophila* clade, the EvoFold algorithm was applied to the same 15-way alignment used to generate the MCSs. The conserved elements to which the EvoFold algorithm was applied contained 98% of known ncRNAs. Stark et al. predicted 19,672 RNA secondary structures in the *D. melanogaster* genome and I obtained these also from the UCSC Genome Informatics FTP server.

2.1.7.2 **RNAz**

A second, complementary, method to predict RNA secondary structures is based on detecting conserved thermodynamic stability in these structures across multiple species alignments. The RNAz program compares the average stability of each of the sequences in an alignment to the stability of the consensus sequence – these will be similar if the structure is conserved across

all sequences, but the stability of the consensus will be much lower than the average if there are different structures present in the different sequences (Washietl et al. 2005).

RNAz predictions have been made in *D. melanogaster* (Rose et al. 2007) using the 12-way *Drosophila* species alignment (Clark et al. 2007). As for EvoFold, these have a high sensitivity (they recover 96% of all known miRNAs) but a low specificity (at a *p*-value threshold of 0.9, 45% of structure predictions are estimated to be false positives). The set of 42,482 secondary structure predictions published by Rose et al. cover 5.1 Mb.

2.1.8 Short RNA Species

One possible explanation for lincRNA function is that they represent the unprocessed primary transcripts of one of several types of short RNA species that have been reported to date – miRNAs, esiRNAs, and piRNAs (**Chapter 1.1.2**). The preparation of samples of each of these types of sequence is described in detail below. I consider their distributions across my lincRNA annotations in **Chapters 3** and **4**.

2.1.8.1 **miRNA candidate Sequences**

A recent study has examined miRNA expression in a variety of tissues across the *D. melanogaster* life cycle (Ruby et al. 2007). I obtained all 954,575 unique sequences deposited by Ruby et al. in the NCBI Gene Expression Omnibus (GSE7448). 954,525 of these mapped to 108,584 positions in the *D.*

melanogaster genome using Bowtie (**Chapter 2.2.1.2**). All parameters were kept constant as only the sequence information and not any related quality scores was available from their FASTA files. Overlapping miRNA candidates were clustered to produce 53,348 regions in the six main chromosome arms ranging in size from 13-2,023 nt long, and which cover 1.3 Mb of the genome.

2.1.8.2 **Endogenous short interfering RNA (esiRNA) candidate Sequences**

EsiRNA candidate sequences were obtained through a combination of immunoprecipitation with Ago2, the binding partner for esiRNAs (**Chapter 1.1.2**), and selection of sequences 18-29 nt long from bulk RNA extractions (Czech et al. 2008). I obtained 8,263,257 sequences from this study via the Gene Expression Omnibus (GSE11086). As in the original study, I mapped these to the *D. melanogaster* genome using BLAT (**Chapter 2.2.1.1**) and only those 89,059 sequences mapping with 100% identity were retained. Overlapping esiRNA candidate sequences found in the euchromatic regions of the genome were merged into clusters and those overlapping known piRNA regions (described in below) were removed. This resulted in 21,262 clusters which covered 874 kb.

2.1.8.3 **piRNA Sequences**

Immunoprecipitation with the Piwi protein was used to extract potential piRNA sequences from *wt* ovaries (Yin and Lin 2007). It was noted by these authors that most of these piRNA sequences were transcribed from 369

genomic clusters, which are thought to represent the primary piRNA transcripts. Many of these are in the heterochromatic regions of the genome or in the unplaced chromosome U sequence. Yin and Lin identified only 151 clusters in the euchromatic region of the six main chromosome arms, and I consider only these in this work.

2.1.9 Chromatin Domains

The chromatin architecture of the *Drosophila* genome has been investigated by a genome-wide analysis of the distributions of a large set of chromatin and histone modification proteins (Filion et al. 2010). The positions of 53 proteins, taken from most known chromatin protein complexes, and four histone modification proteins were probed in the *Drosophila* embryonic cell line Kc167 using the DamID protocol (van Steensel and Henikoff 2000). These different protein profiles were split into five domains of combinatorial protein presence/absence (Filion et al. 2010).

These domains, and their coverage of the *D. melanogaster* genome are as follows: novel heterochromatin (1,711 regions covering 56 Mb); Polycomb-protein marked heterochromatin (2,114 regions covering 23 Mb); HP1-marked heterochromatin (423 regions covering 5.9 Mb); euchromatin containing narrowly-expressed genes (2,036 regions covering 11 Mb); and euchromatin containing broadly-expressed genes (2,114 regions covering 21 Mb) (Filion et al. 2010). The coverage of these regions across different gene classes is discussed in **Chapters 3** and **4**.

2.2 Methods

2.2.1 Transcriptome Mapping

The first stage in analysing transcriptome data is to identify the loci from which these RNA molecules were transcribed. This is done by mapping the RNA sequences to their corresponding DNA sequence in the genome. I use two programs for this mapping – BLAT for analysing long EST-era sequences, and Bowtie for short next-generation sequencing data.

2.2.1.1 **BLAT**

BLAT (Kent 2002), which is an acronym for BLAST-Like Alignment Tool, was developed as an improvement, appropriate for the genomic age, to the frequently used program BLAST (Altschul et al. 1990). Like BLAST, it initially recognises short regions ('hits') of local identity between the query sequence and the genome before extending these into a complete alignment. However, by indexing the genomic database and scanning through the query sequence, rather than indexing the query and scanning through the database, BLAT is several orders of magnitude quicker than BLAST. This is critical when analysing thousands of sequences, such as those of ESTs deposited in GenBank (**Chapter 2.1.1.2**). BLAT can also trigger extensions based on any number of nearby hits, as opposed to BLAST which uses only one or two. Finally, BLAT is able to recover the intron/exon structure of a spliced RNA molecule by generating a single alignment from multiple exons with correctly

placed splice sites. A minimum sequence similarity of 90% between query and genome is recommended, which comfortably covers sequencing error, but suggests that BLAT is not best suited for mapping sequences across more diverged species.

The first, search, stage of BLAT requires an index of non-overlapping sequences, of length K and covering the complete genome of interest, to be constructed. I use the default K for nucleotide searches of the genome throughout this work, which is 11. This index does not include K-mers that occur too frequently (by default, I consider 1,024 instances to be too frequent), such as those found within repetitive transposable-element derived sequences, or that contain ambiguous sequence. Next, each overlapping K-mer in the query sequence, e.g. an EST sequence, is compared against all K-mers in the index and hits with up to one mismatch are retained. Multiple hits in different K-mers but the same query sequence are also allowed, if they are near each other and the distance between them in the query sequence and the genomic index is comparable. Small indel events are allowed between these hits, but only if the hits themselves are perfectly matched. Hits are then placed into buckets of 64 kb based on their position in the indexed database. These buckets known as clumps are then extended by 500 bp and merged if there is less than 300 bp between them.

The second stage generates the alignment of the query sequence to the genome. If an individual K-mer in the query matches more than once each hit is extended until one becomes unique or the length exceeds a given threshold.

Each hit within a sequence is then extended as far as possible with no mismatches, and hits are merged into an alignment. Any gaps in this alignment are filled recursively, until the number of hits stops increasing or all gaps are < 5 bp. Finally, other extensions are allowed such as one or two mismatches or indels in a hit, if the mismatched position is bordered by multiple matches. If there are many equally likely positions for a large gap, then the gap is moved to find the best hit to the consensus GT/AG splice site. Different alignments can be stitched together after this final stage, but this is not implemented in the gfServer version of BLAT used exclusively in this work.

I use BLAT to map all mRNA, EST, and cDNA sequences described in **Chapter 2.1.1** in **Chapter 3**. I also use it to map RACE sequences in **Chapter 5**, and genomic DNA sequences in **Chapter 6**.

2.2.1.2 **Bowtie**

The next-generation sequencing data which I analyse in this work are all mapped to the *D. melanogaster* reference genome using Bowtie (Langmead et al. 2009). This algorithm is designed specifically for the short sequences generated using this technology and is faster by several orders of magnitude than traditional mapping programs, such as BLAT. This is critical when working with these data since they contain several orders of magnitude more information than EST data sets. Bowtie is currently the fastest program available. Its novel indexing of the search genome, which I describe below,

produces relatively small indexes which can be pre-computed and downloaded directly from the Bowtie website (the *D. melanogaster* index is only 150 MB). Bowtie is also a part of a suite of programs designed specifically for mapping and reconstructing RNA-seq information, which I use in **Chapter 4**.

Like BLAT, Bowtie relies on indexing the reference genome although this is done in a different way to BLAT. The Burrows-Wheeler Transform (Burrows and Wheeler 1994) modifies a set of K-mers so that they can be searched efficiently. This transformation involves appending a value $ to the K-mer, where $ is a novel character not found in the K-mer and found alphabetically before all characters in the K-mer. A matrix is then formed of all cyclic rotations of the modified K-mer and these are sorted alphabetically, as shown in **Figure 2.2a**. The right-most column then becomes the BWT(K-mer), and has the same length as the original K-mer. This matrix has the useful property of 'last-first mapping', which means that the i^{th} occurrence of character X in the last column corresponds to the same character as the i^{th} occurrence of X in the first column.

The original text can be recovered using the UNPERMUTE algorithm (Burrows and Wheeler 1994) as shown in **Figure 2.2b**, but it is the search by EXACTMATCH (Ferragina and Manzini 2000) in **Figure 2.2c** which shows the strength of this approach. The search begins with the 3' end of the query sequence (here, the query sequence is acaacg) and progressively adds one position 5' to define an increasingly small number of rows (and therefore genomic positions) which match this sequence. This stage of the search ends

when only one row remains from which the UNPERMUTE algorithm can recreate the query sequence, i.e. a unique position has been found, or no matching rows remain.

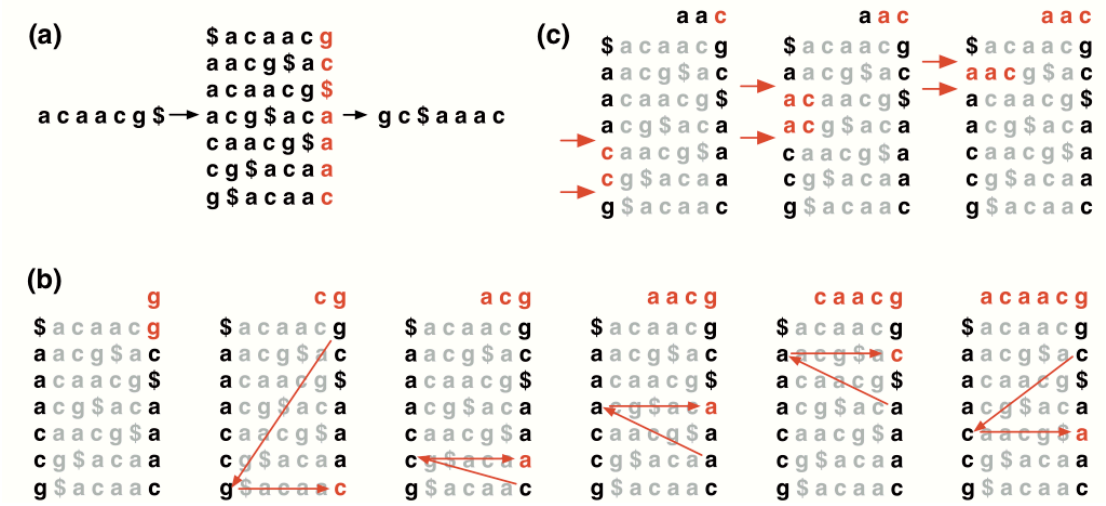

Figure 2.2. Burrows-Wheeler transform. (a) The Burrows-Wheeler matrix and transformation for 'acaacg'. (b) UNPERMUTE repeatedly applies the last first mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column). (c) Steps taken by EXACTMATCH to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'. Taken from Langmead et al. 2009.

If no matching rows are found, Bowtie implements a backtracking stage where an already matched position is substituted for a mismatch. This position is chosen to minimise the quality scores of the mismatched positions and the search described above resumes to see if a position can be found for the query. The first valid alignment found need not necessarily be the best possible alignment, but here I force Bowtie to keep looking for this using the --best option. The number of mismatches which are allowed within the high-quality 5' end of a sequencing read and the total quality score which is allowed to be covered by mismatches can be set by the user. I leave these values at their defaults (2 and 70). In order to prevent excessive backtracking to the 3' end of

the query, Bowtie uses two Burrows-Wheeler transformed indexes – one for the regular genome sequence, and one for the reversed (known as mirror) sequence – which can be searched sequentially. In the first phase, the forward sequence is searched but mismatches are not allowed in the 3' half of the query sequence. The reverse sequence is searched in the second phase where mismatches are not allowed in the 3' half of the reversed query, which corresponds to the 5' end of the original sequence. The number of backtracking events should thus be kept to a minimum but an absolute limit to the number of backtracks for each sequence is also imposed. I leave this value at its default which is 800 when using the --best option.

Bowtie searches are made up of three phases which use, in turn, the mirror index, the forward index, and the mirror index again. The first identifies hits with no mismatches in the 5' left hand side of the query sequence, while the second identifies hits with no mismatches in the other half of the sequence. The third stage phase finds hits with one mismatch in the 5' side of the sequence but none in the other half, and hits with one mismatch in both halves.

Bowtie is much faster than comparable next-generation sequence mappers, and has a similar sensitivity (Langmead et al. 2009). For the same set of 36 bp reads, Bowtie is both 65 times faster and maps 13% more reads than SOAP (Li et al. 2008) and is 36.7 times faster while mapping only 4% fewer reads than Maq (Li et al. 2008). For 76 bp reads, Bowtie is 14.9 times faster than Maq and maps 44.5% of reads, compared to 44.9% which are mapped by Maq

(SOAP is unable to map such long reads). It is noted that the number of reads mapped may not be the only measure of a mapper's behaviour and that other parameters, such as the amount of random-access memory (RAM) occupied while mapping and the accuracy of reads mapped (Bao et al. 2011), could also have been considered.

In **Chapter 4**, I use Bowtie to map the 75 bp reads produced by the modENCODE consortium. In **Chapter 6**, I map a set of 26 bp reads derived from 5' cDNA sequences.

2.2.2 Coding Potential Calculator

Once a set of novel transcripts has been assembled, it is essential to determine which of these are truly noncoding and can thus contribute to lincRNA gene models, and which are actually coding and represent previously unannotated protein-coding transcripts. Proving a negative, that a transcript never encodes a functional protein, is not possible while experimental techniques to test this, such as mutating the start codon of a putative ORF, are also not very amenable to high-throughput analysis of the type required here.

Several statistical tools are available to analyse the primary sequence of a transcript and to determine its ability to encode a functional protein. In this work, I use the Coding Potential Calculator (CPC) (Kong et al. 2007) which these authors report to be as accurate as other techniques, such as the 'Coding or Non-coding' algorithm (Liu et al. 2006), but works much faster and is therefore more suitable for the genome-wide analyses I conduct here.

CPC first considers features of the longest possible ORF encoded by a transcript, using the *framefinder* software (Slater 2000). The quality of this ORF is quantified by the log-odds score reported by this software. CPC scores how much of the transcript is covered by this ORF, which is also related to its quality. Whether the ORF begins with a start codon and ends with an in-frame stop codon is also recorded.

CPC also tests the similarity of this putative encoded protein to other proteins in the UniProt database (The UniProt Consortium 2010). The number of hits found with BLASTX with an *E*-value $< 10^{-10}$ is recorded. The quality of these hits is also recorded as the hit score which is defined in **Equation 2.1**. It is thought that a genuine-protein coding transcript, as opposed to a non-coding transcript, is likely to match several previously reported proteins and that these matches are likely to be of high quality, resulting in a higher hit score. Finally, CPC also calculates the frame score which penalises BLASTX hits in different frames in the putative protein – again, if the peptide is genuine, any similarities with other proteins should match the one true frame.

$$\text{HIT SCORE} = \frac{\sum_{i=0}^{2} S_i}{3} \text{ where } S_i = \text{mean}\left\{-\log_{10} E_{ij}\right\}$$

Equation 2.1 Calculation of HIT SCORE by CPC, where E_{ij} is the *E*-value of the j^{th} hit in frame i.

These six measures are combined using a Support Vector Machine (SVM) to produce a single coding-potential score. In general, the more positive this score is, the more likely a transcript is to encode a functional protein and the more

negative it is, the more likely a transcript is to be noncoding. The authors of CPC suggest that a score between 0 and 1 be recorded as 'weak protein-coding', while anything above 1 is 'protein-coding'. In the same way, a score between 0 and -1 should be recorded as 'weak non-coding' and anything below -1 is 'non-coding'. I benchmark these scores in **Chapter 3.5.2** and confirm that, in *D. melanogaster*, any transcript scored below 0 can be accurately annotated as non-coding, while anything above 0 is likely to be protein-coding.

I use CPC to assess the protein-coding potential of EST-based transcripts in **Chapter 3** and of RNA-seq-based transcripts in **Chapter 4**.

2.2.3 Nucleotide Substitution Rates

When DNA sequences diverge from a common ancestor, they accumulate changes as a result of mutation events and the resulting natural selection or genetic drift processes acting on them. These related sequences are known as homologues, which can be further split into two categories – orthologues and paralogues. Orthologous sequences are those which have arisen through divergence of the genomes in which they are carried, e.g. due to a speciation event, while paralogous sequences arise through duplication within the same genome. In this work, I consider only orthologous sequences as defined by genome-wide BlastZ analysis (see **Chapter 2.1.3**). I also remove any gaps in orthologous alignments, and the nucleotides immediately bordering these gaps, due to the previously reported unreliability of these alignment positions (Lunter et al. 2008). Specifically, these neighbouring bases are biased to be

identical due to the high score penalty caused by introducing a gap in the alignment.

This divergence between related sequences can be quantified in a number of ways, with the most straightforward method being to simply record the total number of differences between the sequences. The *p*-distance (Salemi and Vandamme 2003) is calculated like this to be the number of nucleotide substitutions per site. However, this is an oversimplification which underestimates any divergence, particularly when multiple substitutions may have taken place at the same site between the sequences since they diverged. For example, back substitutions, where a nucleotide mutates from one state to another, and then back to the original state (e.g. C → A → C), and parallel substitutions (e.g. C → A in all species) will not be counted. Similarly, multiple substitutions at a single site (e.g. C → A → G) will only be counted as a single mutation event (C → G).

A better estimate of divergence is to infer it using a formal substitution model and maximum likelihood statistics. By modelling substitutions as random stochastic events, the substitution process can be considered to be a homogeneous Markov process which is summarised by a rate matrix *Q*. This matrix has a number of assumptions: (i) a substitution event at an individual nucleotide position is independent of all other substitutions, and is only dependent on the current state at that position (the Markov property); (ii) nucleotide frequencies have remained constant since the DNA sequences

diverged (the homogeneity property); and (iii) the underlying substitution rate has also remained constant over this time period.

The simplest model using such an approach is the Jukes-Cantor Model, which assumes a single substitution rate and equal proportions of all four nucleotides across the sequence (Jukes and Cantor 1969). This was subsequently modified to allow different rates of transition substitutions (between purines or pyrimidines) and transversion substitutions (between a purine and a pyrimidine) (the Kimura model, Kimura 1980), and different nucleotide frequencies (the F81 model, Felsenstein 1981). The HKY85 model (Hasegawa et al. 1985) incorporates both of these features, but is not as parameter-rich as models such as the general time-reversible model (Yang 1994), which allows all possible nine parameters in Q (three nucleotide frequencies, and five modifications to the underlying substitution rates) to be estimated separately. This type of model is only appropriate when studying long evolutionary distances, as a large number of substitution events are required to reliably estimate the several different estimates of substitution rates. It is noted that all of these models are reversible, in that mutations in either direction (e.g. C → A and A → C) are equally likely.

The HKY85 model was selected as it allows most of the sequences to be analysed while giving similar results to more complicated models. When applying the GTR model, a proportion of the sequences are unable to be studied as their divergence is insufficient. The Q matrix for HKY85 model is defined overleaf:

$$
Q = \begin{array}{cccccc}
 & & A & C & G & T \\
 & A & * & \pi C\alpha_1 & \pi G & \pi T\alpha_1 \\
 & C & \pi A\alpha_1 & * & \pi G\alpha_1 & \pi T \\
 & G & \pi A & \pi C\alpha_1 & * & \pi T\alpha_1 \\
 & T & \pi A\alpha_1 & \pi C & \pi G\alpha_1 & *
\end{array}
$$

where the rows and columns represent the possible states for two sequences at an individual alignment column. The symbols πA, πC, πG, and πT represent the nucleotide frequencies of the four bases and α_1 describes the ratio of transition-to-transversion mutations. The diagonal elements (denoted by *) are scored so that the sum of each row equals zero. This matrix therefore represents the pattern of all possible substitution events, with the estimated branch length corresponding to the 'amount' of evolution which has taken place since the sequences diverged from their common ancestor.

Substitution patterns and other model parameters were estimated by maximum likelihood using baseml, which is part of the PAML package of programs (Yang 2007). Baseml is run using a control file, where the user can add a number of modifications to the model being fitted. The most important parameter is the choice of substitution model and here I always select the HKY85 model defined above. The 'clock' option can also be changed to allow

different substitution rates across different branches of the phylogenetic tree – I leave this at 'no clock' since I use only pairwise alignments and it is not possible to determine on which of the two branches individual substitution events took place. The user can also supply a phylogenetic tree, or ask baseml to estimate this as well. As I have only used pairwise alignments in this work, there is only one possible tree so this choice is not relevant. The maximum likelihood estimation procedure is used to optimise the likelihood of the observed data (the given sequence alignment) by modifying the values of the model parameters, branch lengths and tree topology.

I have used the estimated total branch length as a measure of the substitution rate between two aligned sequences. This is done for a variety of sequence types in both **Chapters 3** and **4**.

2.2.4 Genome-wide Association

Examining the intersection of two genomic features can be informative as to their biological roles. For example, the enrichment of a particular chromatin domain within a gene set may suggest that the expression of members of this set is regulated in a similar fashion. It is important to determine the significance of any such bias, but most tools available for this, such as EASE (Hosack et al. 2003), consider genes as single entities and are therefore unable to take different gene lengths into account. This makes them biased towards longer genes, which may be a particular issue here if, as expected, lincRNAs show a different length distribution to previously defined protein-coding gene

models (Ponjavic et al. 2007). GONOME (Stanley et al. 2006) corrects for this by testing genome-wide association at the nucleotide level, but it is mainly used to determine biases in GO terms and it is not generally applicable to different types of genomic features. Instead, I have used the Annotator program (described in Ponjavic et al. 2007) to examine the representation of many different features within the gene sets I define in this work.

The Annotator takes a set of sequences of interest (known as *segments*) and determines their over- or under-representation within a second sequence set (the *annotations*), relative to the null hypothesis that the segments are randomly distributed within a particular *workspace*. The *workspace* is the background against which any association will be tested and often represents the assembled portions of the genome although it can be modified in other ways, for example to consider only intronic or non-genic sequences. The *segments* represent the feature of interest, which is usually the gene type, e.g. lincRNAs. The *annotations* are then the genomic feature whose prevalence within these *segments* is being investigated.

The first stage is to calculate the proportion of nucleotides within the *segments* and the *workspace* which also intersect the *annotation*. This number is referred to as the 'observed proportion'. Next, *segments* are sampled randomly with replacement, while still keeping their length distribution constant and these are given a random location within the *workspace*. This process continues until the number of sampled nucleotides within the *workspace* equals the initial number of nucleotides in the *segments* found within the *workspace*. The

'sampled proportion' can then be calculated as the proportion of sampled nucleotides within the *workspace* which intersect the *annotation*.

This sampling process is repeated many times, typically 10,000, and a record is kept of how many times the sampled proportion exceeds the observed. A *p*-value can be estimated from this and, if this is lower or greater than a given threshold, usually $p < 0.025$ or $p > 0.975$, then the over- or under-representation, respectively, of the *annotation* within the *segments* is reported as being significant. When multiple *annotations* are tested, e.g. when analysing a set of GO terms, possible false discovery *annotations* must be controlled for when assigning significance to any results obtained. Those rare *annotations* which have an expected density of segments of less than 1% are first removed, to reduce the number of tests. The rate of type I errors (the number of truly non-significant *annotations* incorrectly called as significant) is then estimated as the false discovery rate (Benjamini and Hochberg 1995), and a *p*-value threshold is selected which keeps the number of false positives at an acceptable level, which is generally less than one.

I have used the Annotator to examine the distribution of *segments*, such as lincRNA loci within gene neighbourhoods which, in this thesis, I will call territories. A gene territory covers all the nucleotides that are closer to that gene than to either preceding or subsequent genes on the chromosome, as shown in **Figure 2.3**.

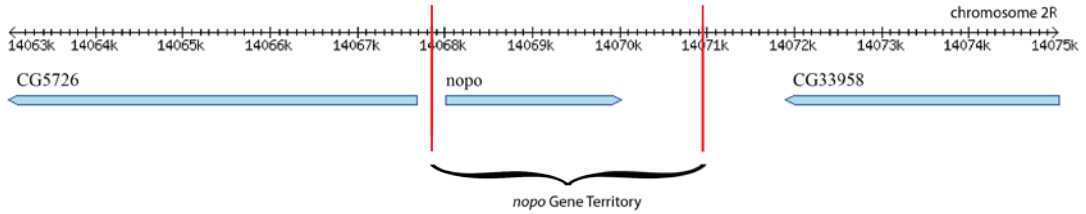

Figure 2.3. Example gene territory for *nopo* (FBgn0034314).

In **Chapter 5**, gene territories are defined only for protein-coding genes, and these cover the complete genome. The territories for overlapping genes were defined by considering them as a single gene stretching from their 5' to 3' maximal borders. LincRNAs are therefore found in one or two (if they straddle the territory border) gene territories. In **Chapter 6**, territories are defined for both protein-coding genes and lincRNA loci. Due to the incompleteness of lincRNA loci defined by EST evidence, I consider these territories to be the maximum possible extent of the gene model or lincRNA locus.

2.2.5 Fly-Handling

Fly stocks are kept at 25°C on standard medium comprising of 80 g/l maize, 18 g/l dried yeast, 10 g/l soya flour, 80 g/l malt extract, 40 g/l molasses, 8 g/l agar, 6.6 ml acid mix (containing 50% propioninc acid and 3.2% phosphoric acid)/l. Flies are anaesthetised using CO_2 for all experiments. *y, w* is used as the control wild-type (*wt*) strain (Bloomington Stock Centre).

2.2.6 RT-PCR

2.2.6.1 **RNA Extraction**

RNA is extracted from S2 cells and a variety of tissues using the Qiagen miRNeasy kit. This kit recovers the total RNA complement from a sample, as well as short RNA species $<$ 200 nt long, although specific extraction of short RNA was not done during this work. The miRNeasy kit can extract up to 100 µg of RNA from up to 50 mg tissue, which is approximately six individuals (whether at the larva, pupa, or adult fly stages), 200 adult heads or 10^6 S2 cells. The cells are removed from culture as a pellet by centrifuging them for 5 min at 300 x g and removing the supernatant.

The extraction protocol broadly followed the manufacturer's instructions. Samples are homogenised with a sterile pestle (or by vortexing in the case of S2 cells) in the presence of 400 µl lysis buffer. This stage breaks the cell membranes to release the RNA molecules while simultaneously breaking apart high molecular weight DNA and cellular organelles. Samples are then homogenised by centrifugation for 2 min at 12,000 x g through a QIAshredder homogeniser to increase the RNA yield. 80 µl of chloroform is added (to aid in later phase separation) and the samples were centrifuged for 15 min at 12,000 x g and 4°C. The aqueous phase is removed and added to 1.5 volumes of 100% ethanol to promote later binding of the RNA. The sample is then placed in a spin column and washed as instructed with the buffers RWT and RPE, before being eluted into 40 µl of RNase-free H_2O.

The concentration of the extracted RNA was measured using a Nanodrop ND-1000 spectrophotometer. This machine exposes a sample to UV light, which is absorbed by RNA at 260 nm, and this absorption is linearly related to the RNA concentration of the sample. Proteins absorb nearer 280 nm, and therefore the ratio of 260:280 nm absorption is used as an indicator of RNA purity. A value of ~2 indicates pure RNA; samples which deviate significantly from this are discarded from subsequent experiments.

2.2.6.2 **Reverse transcription**

cDNA libraries are created from RNA extractions using the QuantiTect Reverse Transcription Kit (Qiagen). This kit contains a mix of primers to initiate the transcription reaction across all regions of extracted RNA molecules, including 5' UTRs. The reverse transcriptase enzyme used here has two distinct activities: an RNA dependent DNA polymerase (to generate the initial cDNA strand) and a hybrid-dependent exoribonuclease (to degrade RNA:DNA hybrids made by the first stage).

The protocol followed is the one recommended by the manufacturer, with three changes described below. 1 μg of RNA is used for every reaction and this is incubated with the genomic DNA wipeout buffer at 42°C for eight, rather than two mins, to ensure complete degradation of genomic DNA. The reverse transcription reaction is then allowed to proceed for 30 mins to maximise the yield of cDNA molecules > 200 bp, before being stopped by incubation at 95°C for 3 mins. Each reaction is carried out twice – once with the reverse

transcriptase enzyme and once with the enzyme replaced with H_2O, to act as a negative control that tests for the absence of genomic DNA contamination in the initial RNA extraction.

2.2.6.3 **Polymerase Chain Reaction**

The Polymerase Chain Reaction (PCR) is used to test for the presence or absence of a particular sequence in a DNA sample. The sample is incubated with forward and reverse primers, which are specific to the sequence of interest. I make 25 µl reaction mixes containing the following components: 1x NH_4 Reaction Buffer, 2 mM $MgCl_2$, 0.4 mM dNTPs, 1 unit *Taq* DNA polymerase (Bioline), 1 µM forward and reverse primer, and ~10 ng DNA sample. The DNA samples can be either cDNA (created as described in **Chapter 2.2.6.2**) or genomic DNA (created as described in **Chapter 6.4.2**). Primers targeting the constitutively-expressed TATA-binding protein (TBP) mRNA are used as a positive control for PCR reactions using a cDNA template.

PCR reactions are carried out in a G-Storm GS4 thermal cycler. The PCR reaction programme consists of an initial denaturation step of 2 mins at 94°C, before the following steps are repeated for up to 40 cycles:

1. Denaturation of the double-stranded DNA molecules at 95°C for 30 s.

2. The primers anneal to the DNA at 60°C for 30 s.

3. The new DNA strands are created by extension at 72°C for 30 s.

The final extension step is allowed to proceed for 5 mins, still at 72°C. Finally, the reaction mix is held at 4°C.

PCR products are examined by separating them on a 1% agarose gel (1x TAE buffer, containing 0.04 M Tris-acetate and 1 mM EDTA) containing ethidium bromide (1 μg/ml) run for approximately 1 hour at 100 V. Approximately 6 μl of sample is loaded in the presence of 1 μl of 1x loading buffer (2.5% bromophenol blue and 30% glycerol, New England Biolabs). The size of these products is determined by running them alongside a 100 bp or 1 kb ladder of DNA molecules of known size (New England Biolabs). These gels are then photographed under UV light (Alpha Innotech AlphaImager 3400).

2.2.7 Real-time PCR

Regular PCR as described above yields essentially a binary result, which is either the presence or absence of the sequence of interest. Real-time PCR is an extension of this protocol which monitors the PCR reaction as it proceeds to make a quantitative measure of the amount of sequence present in the original sample. This technique has three advantages over traditional PCR as it is more sensitive to the presence of rare transcripts, such as lincRNAs; it has an increased range of quantities over which a sequence can be detected; and there is less post-processing (no need for agarose gel separation). The amount of DNA produced during the exponential phase of the PCR reaction, which is proportional to the starting amount, is recorded as the threshold cycle C_T – the PCR cycle number at which the sample reaches a fluorescent intensity

about a given threshold. In my experiments, I allow this threshold to be set automatically by the real-time PCR software (Applied Biosystems).

The fluorescence is generated during the PCR reaction process by the presence of double-stranded DNA (dsDNA), using the SYBR green technology (Zipper et al. 2004) in the Platinum SYBR Green qPCR SuperMix-UDG (Invitrogen). This dye binds all dsDNA and, when it binds, it fluoresces. This is an advantage if investigating the expression of several genes, as I do, because there is no need for any gene-specific optimisation as would be necessary for the probes required by the alternative Taqman protocol (Holland et al. 1991). However, it also increases the risk of false-positives, as any non-specific amplification will be recorded. This is controlled by using non-template controls (PCR reactions where H_2O is added rather than the DNA template) and then examining the melting curves at the end of the reaction. These curves describe the temperature at which the dsDNA in the sample melts, and is related to its sequence and length. The non-template control curves should be very different to the experimental samples, as any melting will only represent the presence of non-specific dimerisation of the primer pairs, while the experimental curves should appear similar if they are amplifying the same sequence. Manual inspection is used to remove outliers until the real-time software is satisfied that the standard error between technical replicates is at an acceptable level. This is done before looking at the C_T values to reduce experimenter bias.

Real-time PCR reactions are prepared according to the manufacturer's instructions. I carry out four technical replicates of 10 µl reactions for each sample. These contain 1x real-time PCR mix (containing the *Taq* polymerase, required buffers, nucleotides, and SYBR Green dye), ROX reference dye (50 nM) forward and reverse primers (1x primer mix, or 500 nM forward and reverse primer), ~25 ng cDNA sample, and are made up to 10 µl with RNase-free H_2O. Note that all real-time PCR reactions are carried out using cDNA, prepared as described in **Chapter 2.2.6.2**. The reactions take place in the Applied Biosystems 7500 Fast Real-Time PCR machine.

2.2.7.1 **Absolute Quantification**

The quantity of DNA in each sample is calculated using the absolute method. First, a standard curve is plotted to associate a known quantity of DNA with a particular C_T value for each pair of primers. A serial dilution of the initial cDNA library is created and the average C_T for each dilution is plotted against the $\log_{10}$(quantity), as shown in the example in **Figure 2.4**. For ease of calculation, the original library is arbitrarily assigned a quantity of 100 for each primer set. The y-intercept (c) and the gradient (m) of the linear relationship between these two quantities are recorded. Subsequent experimental C_T values can then be converted into DNA quantity using **Equation 2.2**.

$$Quantity = 10^{\left(\frac{C_T - c}{m}\right)}$$

Equation 2.2 Calculation of absolute DNA quantity from a C_T value.

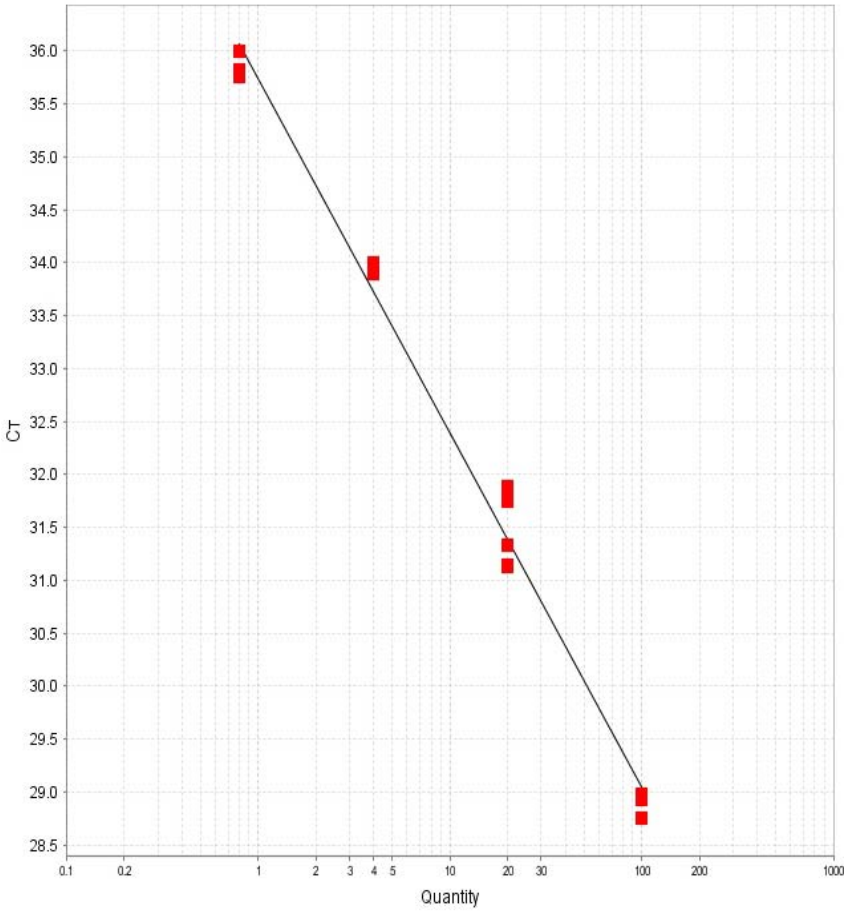

Figure 2.4 Example standard curve for lincRNA EC069024.

2.2.7.2 **Comparative C_T**

When comparing the quantities of a particular sequence between two treatments (e.g. *wt* and mutant), these should be normalised against a sequence assumed to present at the same quantity in both samples. This is particularly important when analysing cDNA libraries, as this controls for variation between individual RNA extractions or reverse transcription

reactions. I use Glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*), an enzyme required for glycolysis (Alberts et al. 2002), as such a control. This constitutively expressed gene is a traditional 'house-keeping' gene required at a similar level in all cells. It is frequently used as a loading control in western blots or as a control in RT-PCR experiments of the type I perform here.

For each experimental sample, what I call the relative expression of each gene is calculated using the following process. The quantity of each replicate for all genes is calculated from the C_T as described in **Chapter 2.2.7.1** and each quantity is then divided by the mean *Gapdh* quantity for that biological sample to calculate the normalised quantity. Relative expression values are then calculated for each gene by dividing the normalised quantities by the grand mean of the control treatment. The mean and standard errors of these expression values can then be plotted, where the mean for the control is automatically one.

2.2.8 Statistical methods

2.2.8.1 Empirical Cumulative Distribution Function

I frequently display a range of numbers, such as the substitution rates of all lincRNAs, as their cumulative distribution. This requires N data points to be ordered by either increasing or decreasing magnitude as X_1, X_2,...,X_N. The empirical cumulative distribution function can then be calculated as in **Equation 2.3**. This function is graphed for a number of data sets.

$$E(n) = \frac{n(i)}{N}$$

Equation 2.3 Empirical Cumulative Distribution Function. n(i) is the number of data points with values less than or equal to X_i.

The average of a distribution is usually summarised by the median, which is the mid-point of the empirical cumulative distribution, i.e. X_i when E(n) = 0.5. This is useful when the data are not normally distributed because, unlike the mean, it is not so affected by very extreme values.

2.2.8.2 **Unpaired Student's T-Test**

The Student's T-test is used to determine whether two normally distributed samples are drawn from the same population distribution and therefore if the samples are significantly different from each other. It is assumed that the variance of the two samples is equal and the mean values of the samples are compared as shown in **Equation 2.4**, for unpaired samples of potentially different size. The T-statistic is expected to be normally distributed, with the number of degrees of freedom equal to (n_1+n_2)-2 where n_1 and n_2 correspond to the sample size of groups 1 and 2, respectively.

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{X1X2} \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } S_{X1X2} = \sqrt{\frac{(n_1 - 1)S_{X1}^2 + (n_2 - 1)S_{X2}^2}{n_1 + n_2 - 2}}$$

Equation 2.4 Calculation of the t-test statistic for samples 1 and 2. $\overline{X_n}$ corresponds to the sample mean while S^2_n is the sample variance and n_n is the sample size of group n.

2.2.8.3 **Wilcoxon or Mann-Whitney U Test**

The Wilcoxon or Mann-Whitney U test is also used to determine whether two samples are drawn from the same population distribution. The median values of the samples are compared and therefore the distribution need not be normal. This test can be considered to be the non-parametric equivalent of the unpaired Student's T-test.

2.2.8.4 **Chi-squared Test**

This non-parametric test examines whether there is an association between two categorical variables. The expected frequencies of the combinations of these two variables are calculated under the null hypothesis of no association as the product of the number of occurrences of the two variables divided by the total number of observations. The chi-squared statistic can then be calculated as in **Equation 2.5**, whose distribution and associated *p*-value is well known.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Equation 2.5 Chi-squared calculation. O_i and E_i denote the observed and expected frequency of observation i, respectively.

2.2.8.5 **Conducting Statistical Analyses**

All statistical analyses described in this thesis were conducted using either the R statistical package (http://www.r-project.org), or Microsoft Excel.

Chapter 3: lincRNAs SHOW SEVERAL SIGNATURES OF FUNCTIONALITY

3.1 Abstract

Little evidence exists about the presence of lincRNAs and their functionality, or lack thereof, in the *D. melanogaster* genome. I have comprehensively investigated the noncoding transcriptome of *D. melanogaster*, using EST and mRNA evidence to define a set of 2,788 lincRNA loci.

1,411 of these were annotated as evolutionarily constrained between *D. melanogaster* and two sister species, *D. simulans* and *D. yakuba*. These lincRNAs are depleted in indel mutations, and also contain an excess of sequences conserved across the entire *Drosophilid* clade, suggesting that they may be functional in each of these species. Constrained lincRNAs are enriched in predicted RNA secondary structures, implying that it may be the mature RNA sequences which are required for function. These functions may be important for developmental regulation of the fly since these loci are also frequently found within Polycomb-group protein-associated heterochromatin.

Using the same approach, 241 lincRNAs were identified as possessing a significantly increased substitution rate in these three species. These sequences often represent the primary transcripts of short RNA species. Their rapid evolution may be driven by the positive selection experienced by some of these

short sequences to maintain their ability to suppress harmful transposable elements in the genome.

Taken together, this work shows that lincRNAs have been previously overlooked by previous *D. melanogaster* annotations but are now deserving of further experimental scrutiny.

3.2 Introduction

The genome of *D. melanogaster* is one of the best-studied animal genomes to date but there likely remain many functional elements, such as lincRNAs, to be discovered within it. *D. melanogaster* was the first animal to have its genome sequenced using whole-genome shotgun technology (Myers et al. 2000), and regular updates to its extensive annotations are released by the FlyBase service (Tweedie et al. 2009). Still, it is clear that transcription exists beyond these annotated genes as intergenic ESTs can be observed using the FlyBase genome browser. Microarray experiments (Manak et al. 2006) have even suggested that virtually all of the *D. melanogaster* genome may be stably transcribed at some point during development. This previously unrecognised transcription may yield a number of lincRNAs which further extend the characterisation of the *Drosophila* genome.

LincRNAs have previously been defined using EST evidence in a variety of species, including mouse (1,987 loci, Carninci et al. 2005; Ponjavic et al. 2007; A.C. Marques, unpublished), pig (1,004 loci, Seemann et al. 2007), and Arabidopsis (19 loci, MacIntosh et al. 2001). EST sequences may indeed be an

optimal method for defining novel lincRNAs (Xue and Li 2008). Several databases of lincRNAs defined by such evidence now exist, including RNAdb, a comprehensive database of mammalian ncRNAs (Pang et al. 2007).

In this chapter, I take a comprehensive approach to defining functional lincRNAs in *D. melanogaster.* A set of 2,788 lincRNA loci is defined using over 575,000 expressed sequences. The majority of these loci are likely to be functional, where the 1,411 constrained lincRNAs (79.4% of the 1,776 which could be reliably assigned between all three *Drosophilid* species by my simulation procedure) contain a variety of other genomic features suggestive of functionality and may be important for developmental regulation. 241 (13.6%) loci are annotated as fast evolving and appear to function primarily as the precursors of short RNA species: their increased substitution rates may be caused by the positive selective pressure experienced by the short RNA sequences they contain. These lincRNAs should be investigated experimentally in greater detail and I describe my preliminary work on this in **Chapters 5** and **6**.

3.3 Materials

3.3.1 Synonymous substitution rates

The distribution of substitution rates of lincRNAs is compared to that of synonymous sites within protein-coding genes. The sequence evolution of these genes has been analysed using codeml, which is part of the PAML package of

programs (Yang 2007). This is similar to baseml, which is described in **Chapter 2.2.3**, but it allows the substitution rates at non-synonymous sites (known as d_N) and d_S within protein-coding sequences to be estimated separately. The ratio of these is often used as a measure of the selection acting on the sequence (Yang 1994), where synonymous sites are assumed to be evolving under little or no selective pressure. Although certain synonymous codons are known to be experiencing relatively weak selection for translation efficiency in *Drosophila* (Akashi 1994), I use them here as a conservative proxy for neutral sequence. I have recorded all published d_S values for predicted *D. melanogaster*-*D. simulans* and *D. melanogaster*-*D. yakuba* orthologues (Heger and Ponting 2007).

3.4 Methods

3.4.1 Sampling neutral substitution rates

The significance of the substitution rate observed for each lincRNA, relative to that seen in putatively neutral sequence, is determined by sampling from a concatenated alignment of short introns, created as described in **Chapter 2.1.4**. Separate alignments were created for *D. melanogaster* against *D. simulans* or *D. yakuba*.

A neutral alignment is generated for each lincRNA by sampling with replacement from a corresponding concatenated intron alignment. For each base in the lincRNA, an identical nucleotide is randomly selected from the *D.*

melanogaster intron alignment and the corresponding aligned base in the other species recorded, as shown in **Figure 3.1**. 1,000 such neutral alignments are generated for each lincRNA and their substitution rates estimated using baseml (**Chapter 2.2.3**). The probability of the true substitution rate being equal to, or greater than, the observed rate can then be calculated. LincRNAs with $p < 0.025$ are then labelled as 'constrained', whereas those with $p > 0.975$ are labelled as 'fast-evolving'. All other lincRNAs are labelled as 'neutral'.

The false positive rate (FPR) for these classifications is estimated by splitting the observed p-values into 40 bins which each cover 2.5% of the probability distribution, and then calculating the mean number of entries in the neutral bins ($p \geq 0.025$ and ≤ 0.975). The ratio of this to the height of, for example, the bin containing the constrained lincRNAs is then the estimated FPR for these constrained lincRNAs.

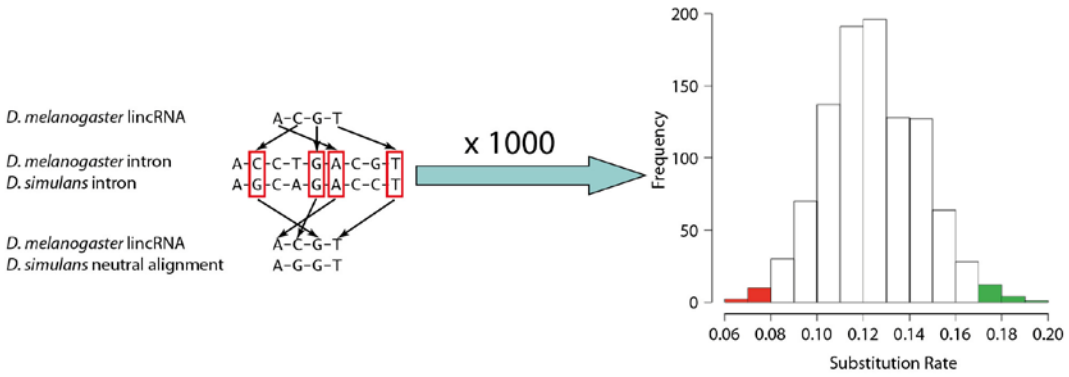

Figure 3.1 Sampling protocol to estimate the distribution of putatively neutral substitution rates. LincRNAs whose observed substitution rates fall into the red portion of the distribution are annotated as 'constrained', while those in the green portion are annotated as 'fast-evolving'. All other lincRNAs are labelled as 'neutral'.

3.5 Results

3.5.1 Definition of transcriptional units, and their overlap with **known gene models**

The EST, mRNA and cDNA sequences describing the *D. melanogaster* transcriptome (**Chapter 2.1.1**) were mapped to the *D. melanogaster* genome by BLAT, with all parameters kept at their default values. The best-scoring position as calculated in **Equation 3.1** was retained for those sequences which mapped to multiple positions in the genome. In total, 499,059 (86.7%) of these sequences were placed uniquely onto one of the six main chromosomal arms, avoiding the heterochromatic and unassembled regions of the genome.

$$\text{Alignment score} = \text{Number of matches} - \text{Number of mismatches}$$

Equation 3.1 Alignment scoring by BLAT

These sequences were merged to produce transcriptional units (TUs) (Carninci et al. 2005), which I define as the boundaries of a cluster of one or more sequences connected through shared exonic or intronic bases. Not all pairs of sequences in a cluster need overlap, and this overlap can be on opposing DNA strands (the strandedness of mapped sequences is not actually considered here). 14,897 TUs were defined, of which 3,122 are intergenic, in that they have no overlap with FlyBase-defined genes. Two typical TUs – one intergenic (BI609486) and one covering a protein-coding gene (FBgn0031424, *VGlut*) – are shown in **Figure 3.2**. Intergenic TUs are named after the 5' most

transcribed sequence in the cluster. Therefore, although they are named after GenBank accession numbers, there is not a direct correspondence between a TU and its related GenBank entry (unless the TU is supported by only one sequence, in which case the TU is simply the position of that one sequence). Multiple, overlapping genes (Celniker and Rubin 2003) are merged into a single TU, which are then referred to as gene models, by this procedure.

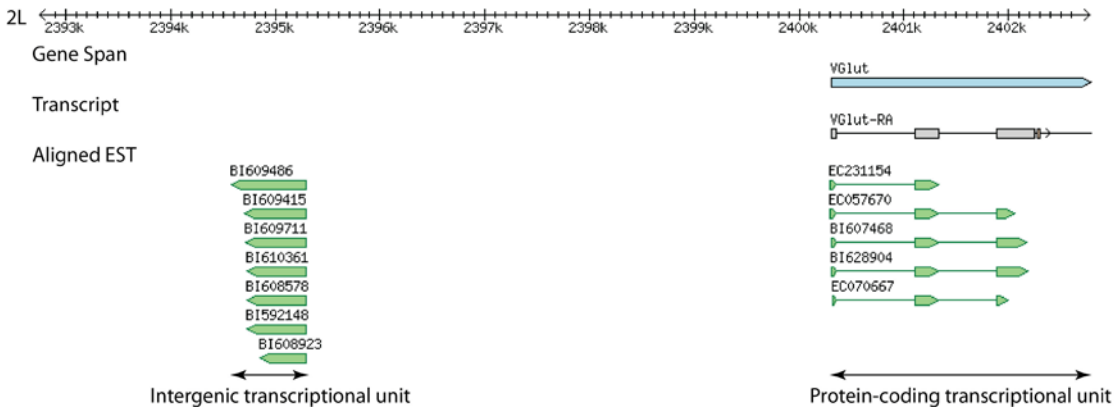

Figure 3.2 Example intergenic (BI609486) and protein-coding gene (FBgn0031424, *VGlut*) TUs.

The coverage of the *D. melanogaster* transcriptome was estimated by measuring the coverage of previously defined genes by this TU set. Of the 14,652 genes annotated in FlyBase, 13,470 (91.9%) overlapped a TU by at least one base. This significant coverage implies that the remaining intergenic TUs should cover a large proportion of any intergenic transcription in *D. melanogaster*.

3.5.2 Benchmarking of CPC

Despite the high quality of the *D. melanogaster* genome annotation (Celniker and Rubin 2003), it was unclear how many of these intergenic TUs are genuine

lincRNAs and how many represent previously unannotated protein-coding genes. The protein-coding potential of TUs is predicted by CPC, as described in **Chapter 2.2.2**, which reports a score reflecting the ability of the TU to encode a protein (Kong et al. 2007).

I decided to benchmark the behaviour of CPC against a set of known transcripts in *D. melanogaster*, both protein-coding and noncoding. I first considered a score of 0 as a cut-off where transcripts scoring below this were annotated as noncoding, while transcripts scoring above this were annotated as coding. I also analysed transcripts in both strands, as this is how I later tested the TUs defined in **Chapter 3.5.1**. Only 130 (0.62%) out of 20,822 coding transcripts in FlyBase were annotated as noncoding, while 49 (7.5%) out of 652 noncoding sequences (annotated noncoding RNAs, miRNAs, and miscellaneous RNAs) were annotated as coding. This results in a specificity of 99.4% to detect a true protein-coding transcript, and a sensitivity of 92.5% to recover a noncoding transcript. I therefore consider any sequence with a CPC score below 0 on both strands to be truly noncoding, and a sequence with a score above 0 to be protein-coding.

3.5.3 Annotating lincRNAs

334 of the 3,122 intergenic TUs were predicted by CPC to have protein-coding potential on at least one DNA strand. This is significantly greater than the 234 predicted using a false-negative rate of 7.5% estimated in **Chapter 3.5.2** (two-tailed chi-squared test, $p = 0.02$), and suggests that this set likely

contains up to 100 previously unrecognised protein-coding exons or genes. As it is not possible to recognise which of these 234 TUs with protein-coding potential are the 100 truly protein-coding loci, all 234 were discarded. The remaining 2,788 TUs are considered *bona fide* lincRNA loci, while the remaining genomic regions between all TUs are annotated as intergenic space. These sequence types, plus FlyBase-defined gene models, are summarised in **Table 3.1**.

Sequence Type	**Number of loci**	**Median number of supporting ESTs per bp**	**Median length (bp)**	**Total length (bp)**
Gene models	11,775	6.6×10^{-3}	2,280	73,032,174
lincRNA loci	2,788	2.6×10^{-3}	532	1,706,601
Intergenic	14,902	0	715.5	45,030,001

Table 3.1 Summary of three sequence types.

LincRNA loci are generally shorter than gene models and the intergenic regions surrounding them (Mann-Whitney, $p < 2.2 \times 10^{-16}$ for both comparisons). In total, this set of lincRNA loci represents 1.7 Mb (1.4%) of the *D. melanogaster* genome, while the gene models cover 73 Mb (61.0% of the genome).

It was also noted that lincRNAs are supported by fewer transcribed sequences than gene models (Mann-Whitney, $p < 2.2 \times 10^{-16}$). This suggests that they are transcribed at a lower level, or show a more restricted expression profile than protein-coding genes, which agrees with the observations of lincRNAs defined by RNA-seq data described in **Chapter 4**.

3.5.4 Substitution rate of gene models, lincRNA loci and intergenic regions

It is expected that functional regions of the genome, such as those which encode (protein-coding) genes, may show increased sequence conservation between species, as the vast majority of mutations within them may have a detrimental effect on function (Li 1997). The gene model, lincRNA locus and intergenic region sequences defined in **Chapter 3.5.3** were aligned to their orthologous sequences in *D. simulans* or *D. yakuba* and the substitution rates between these species were estimated using baseml.

By assuming previously annotated genes to be functional, and intergenic regions to be much less so, it is shown in **Figure 3.3** that lincRNA substitution rates more closely resemble those of functional gene models, than the supposedly more evolutionarily neutral intergenic regions. The median substitution rates for gene models and lincRNAs are very similar and, in alignments with *D. simulans*, the median lincRNA substitution rate (0.052) is actually slightly lower than that for the gene models (0.053). All comparisons of substitution rates (gene models *versus* lincRNA loci, gene models *versus* intergenic regions, lincRNA loci *versus* intergenic regions) within the two sets of alignments were significant (Mann-Whitney, all $p < 0.01$).

The lincRNA substitution rate distributions are also skewed to the right, where there is an excess of lincRNAs with a higher substitution rate. This suggests that not all lincRNAs may be experiencing purifying selection of the

same type as protein-coding genes. This is investigated in further detail in **Chapter 3.5.5**.

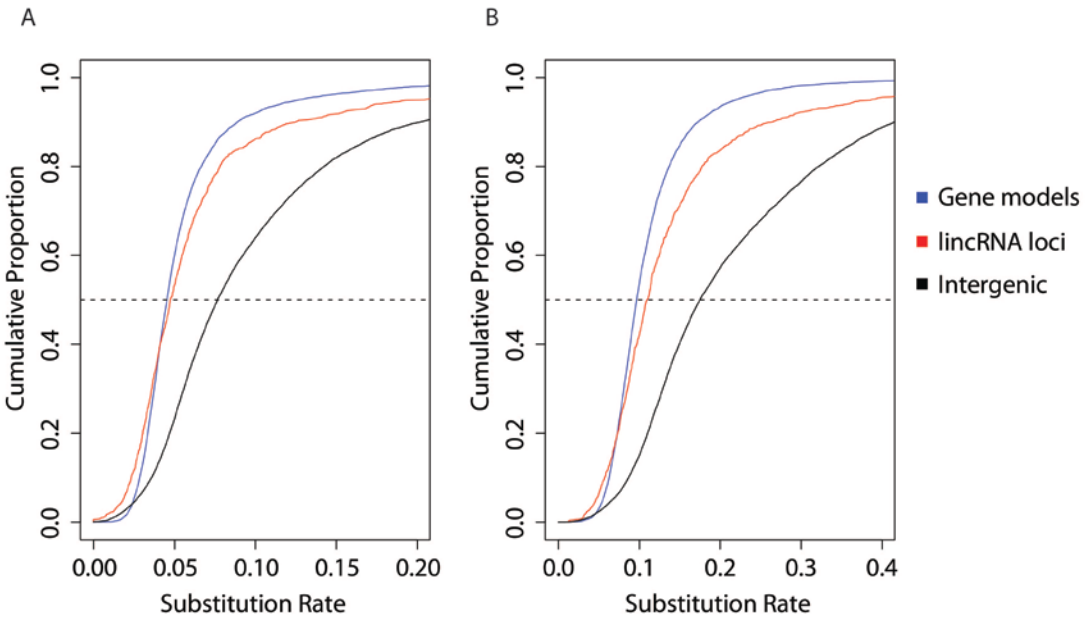

Figure 3.3 Cumulative frequency distributions of nucleotide substitution rates of gene models (blue), lincRNA loci (red) and intergenic regions (black). The dashed line indicates the 50th percentile. A. Alignments between *D. melanogaster* and *D. simulans.* B. Alignments between *D. melanogaster* and *D. yakuba*.

3.5.5 Evolutionary classification of lincRNAs

Individual lincRNA loci were classified as being significantly constrained, neutral, or fast-evolving using the sampling procedure outlined in **Chapter 3.4.1**. The results of this procedure are summarised in **Table 3.2**, with a FPR less than 6%. As suggested by the substitution rate distributions in **Figure 3.3**, the majority of loci are significantly constrained, although a sizeable number also show an increased rate of sequence evolution relative to that of short introns (**Table 3.2**).

To increase the reliability of this classification procedure, I discarded lincRNA loci from further analyses if they were not placed into the same class by their alignments with both *D. simulans* and *D. yakuba*. 1,012 loci were removed, which resulted in a set of 1,411 constrained lincRNA loci, 124 which are neutrally evolving, and 241 fast evolving loci. As in the individual species analyses, the neutrally evolving lincRNA loci (median length 247.5 bp) tend to be shorter than either the constrained (median length 596 bp) or the fast evolving set (median length 500 bp) (Mann-Whitney, $p < 2\text{x}10^{-9}$ for all comparisons). This reduced length of neutrally evolving loci may limit the accuracy of the substitution rate estimation. An increased variation in the sampled distribution for these lincRNAs would then make it more unlikely that the observed rate would significantly deviate from this wider distribution.

	***D. simulans* alignments**				***D. yakuba* alignments**			
	Number of lincRNAs	**Median length (bp)**	**Median substitution rate**	**FPR (%)**	**Number of lincRNAs**	**Median length (bp)**	**Median substitution rate**	**FPR (%)**
Constrained	1,858	570	0.036	0.7	2,109	559	0.085	0.4
Neutral	239	280	0.121	N/A	132	285	0.298	N/A
Fast	497	566	0.458	5.5	295	575	0.589	5.9

Table 3.2 Summary of lincRNAs defined by substitution rates using alignments with *D. simulans* or *D. yakuba*.

3.5.6 Consistency of neutral lincRNA substitution rate with that of synonymous substitutions in protein-coding genes

To further test the performance of my sampling procedure, I compared the substitution rate distributions of the three lincRNA classes to that of synonymous sites estimated using the same species alignments. As noted in **Chapter 3.3.1**, these sites are a conservative proxy for neutral sequence but, as shown in **Figure 3.4**, they very closely resemble the distribution of the neutrally evolving lincRNA loci. Substitution rates between these two sequence classes are similar (Mann-Whitney, $p = 0.84$ and 0.05 for alignments with *D. simulans* and *D. yakuba*, respectively). The substitution rate distributions of both constrained and fast-evolving lincRNA loci are highly significantly different from that for these synonymous sites (Mann-Whitney, $p < 2.2\text{x}10^{-16}$ for all comparisons).

From this, I conclude that my protocol is appropriate for classifying lincRNAs as evolutionarily constrained, fast-evolving or neutral. The remainder of this chapter focuses on analysing the differences observed between these three classes.

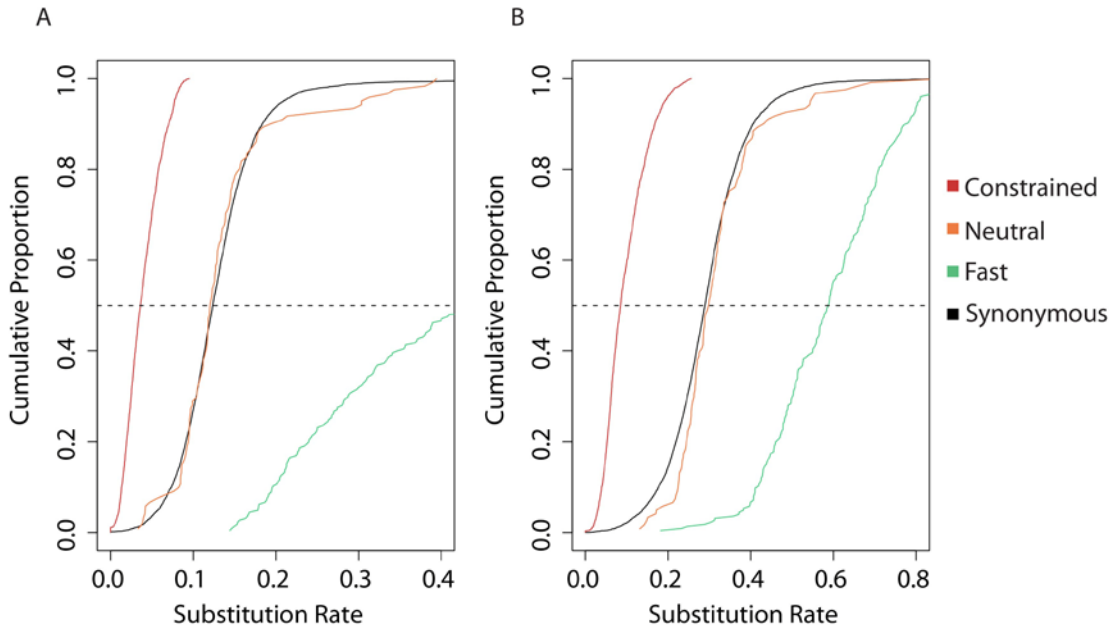

Figure 3.4 Cumulative distributions of substitution rates for different lincRNA classes (as defined in Table 3.2) and synonymous sites in *D. melanogaster*. The dashed line indicates the 50th percentile A. Alignments with *D. simulans*. B. Alignments with *D. yakuba*.

3.5.7 Specific enrichment of functional genomic features in constrained lincRNAs

I next considered how the results of this primary sequence analysis compared to other indicators of functionality in the *D. melanogaster* genome. The Annotator (**Chapter 2.2.4**) was used to examine the coincidence of a variety of genomic features within the different lincRNA sequence classes. In all of these analyses, the *workspace* selected was made up of all regions outside annotated gene models, as these were the regions from which lincRNAs were defined. When considering the EvoFold predictions of RNA secondary structure (**Chapter 2.1.7.1**), the *workspace* is made up of MCSs not covered by gene models as it was these regions from which EvoFold predictions were derived.

Beyond showing a suppressed substitution rate in their primary sequence, the constrained lincRNAs are significantly and specifically enriched in two measures of evolutionary conservation (**Figure 3.5**), which provides further evidence for their functionality. They exhibit an 8.3% ($p < 0.001$) enrichment in indel-purified segments (IPSs), while a deficit of these is seen for both neutral and fast-evolving lincRNA loci (-22.1% and -39.9%, respectively with $p < 0.001$). This is perhaps not unsurprising, as it is known that substitution and indel mutation events are correlated across the genome (Hardison et al. 2003). It does confirm, however, that lincRNA loci respond similarly to both of these different types of mutation events – constrained lincRNAs tend not to tolerate these events, while they are both relatively frequent within fast-evolving lincRNA loci.

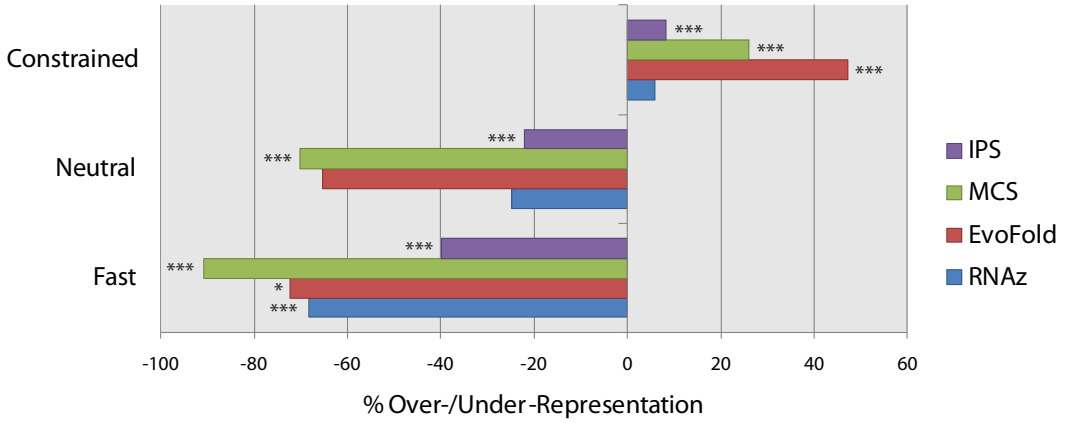

Figure 3.5 Enrichments or deficits of IPSs, MCSs, EvoFold predictions and RNAz predictions within different sequence types relative to genome-wide random expectations (* indicates $p < 0.05$ and *** indicates $p < 0.001$).

Constrained lincRNAs are also enriched (26.0%, $p < 0.001$) in MCSs, which are regions of deep conservation across 15 insect species as defined in **Chapter**

2.1.6. Conversely, both neutral and fast-evolving lincRNAs are depleted in MCSs (-70.3% and -91.0%, respectively with $p < 0.001$) in these regions. Almost all (1404, 99.5%) constrained lincRNA loci overlap an MCS by at least one base. Their function may therefore be conserved across all 15 species examined here. The patterns of lincRNA evolution observed in three relatively closely related species thus appear to be replicated across these other species, which implies that these lincRNA loci may be functional in *Drosophila* and related insect species.

The enrichment of predicted secondary structures within constrained lincRNA loci suggests that it may be the folding of the mature RNA molecule which is transcribed from these loci, rather than simply the act of transcription across them, which may be responsible for any biological function. As shown in **Figure 3.5** this set of lincRNAs is specifically enriched in both EvoFold and RNAz predictions of RNA secondary structure (47.3% and 5.84%, respectively), although only the enrichment in EvoFold predictions is statistically significant ($p < 0.001$). The lack of significance for the RNAz predictions may in part be caused by the reported high false-positive rate (45%) in making these predictions. This cannot, however, be the only reason as this rate is lower than the false-positive rate of 62% reported by EvoFold. The fast-evolving lincRNAs are significantly depleted in both prediction sets (-72.4%, $p = 0.04$ for EvoFold predictions and -68.4%, $p < 0.001$ for RNAz), while the neutral lincRNA loci show no deviation from the expected frequency of these structure predictions.

The distribution of these lincRNA loci across a set of genome-wide chromatin domains is also informative as to their functional roles (**Figure 3.6**). The constrained lincRNAs are specifically enriched in both types of euchromatin, although their concentration in regions showing a more regulated expression profile is greater and more significant (85.8% enrichment, $p < 0.001$ against 2.37% enrichment, $p = 0.04$). They are also frequently found in Polycomb-protein marked heterochromatin (15.5% enrichment, $p = 0.04$).

In contrast, HP1-associated heterochromatin is specifically enriched within fast-evolving lincRNA loci (261.3%, $p < 0.001$). The novel heterochromatic regions identified by this study are under-represented in all three lincRNA classes (data not shown).

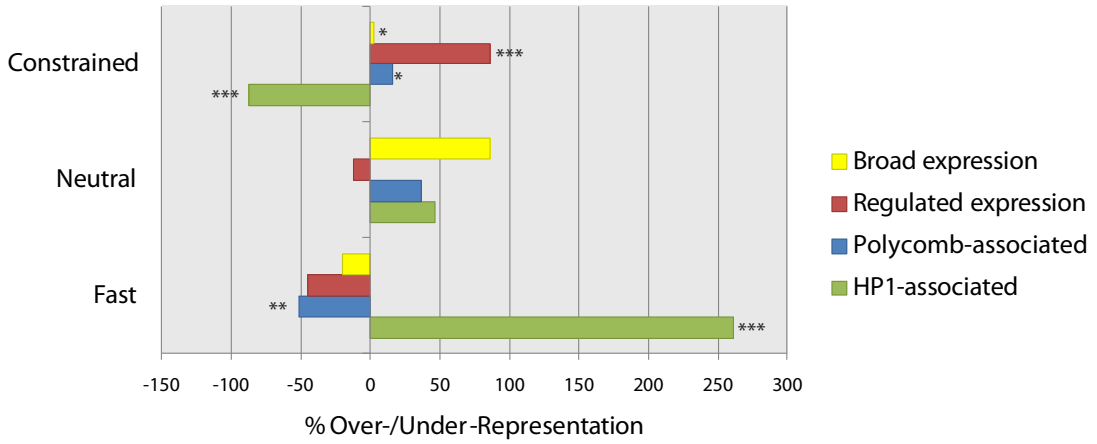

Figure 3.6 Enrichments or deficits of chromatin states within the three lincRNA classes relative to genome-wide random expectations (* indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$).

3.5.8 Enrichment of short RNA species in fast-evolving lincRNAs

One possible explanation of lincRNA functionality could be that they represent the primary transcripts of previously identified short RNA species. I tested this by calculating the genomic overlaps of three different types of short RNAs (miRNAs, esiRNAs and piRNAs, defined in **Chapter 2.1.8**) within these lincRNA loci, as shown in **Figure 3.7**. Interestingly, all three are specifically enriched within fast-evolving lincRNAs (146%, 113% and 321% for miRNAs, esiRNAs and piRNAs, respectively, all $p < 0.002$). 47 fast-evolving lincRNA loci overlap a previous miRNA annotation by at least one base, while 98 similarly overlap an esiRNA, and 15 are found within the piRNA clusters. In total, 126 (52.3%) fast-evolving lincRNA loci may represent the primary transcript of at least one short RNA molecule.

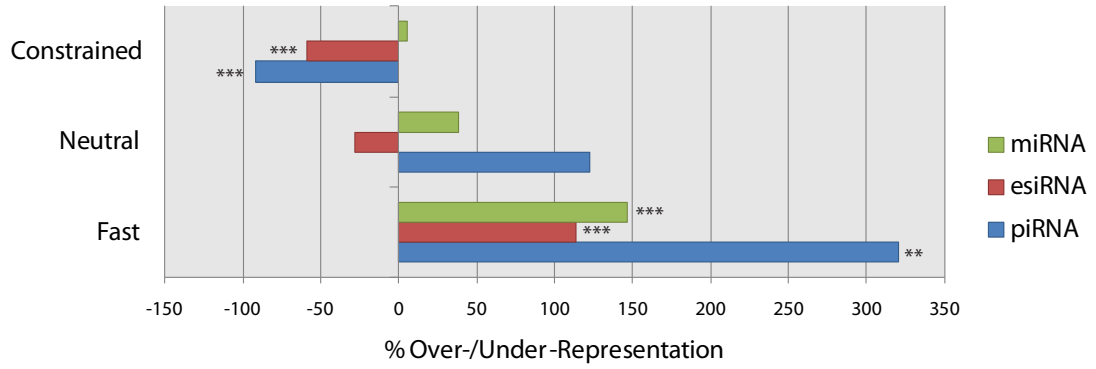

Figure 3.7 Enrichments or deficits of short RNA species within the three different lincRNA classes relative to genome-wide random expectations (** indicates $p < 0.01$, *** indicates $p < 0.001$)

3.6 Discussion

The results presented here confirm that even recent gene catalogues in FlyBase (Tweedie et al. 2009) remain incomplete. I identified 3,122 transcribed but intergenic loci of which 2,788 are likely to encode true lincRNAs. By considering multiple overlapping genes to be part of the same TU, this corresponds to a 23.7% increase in the number of annotated loci in the *D. melanogaster* genome. These lincRNAs are, however, generally shorter than protein-coding genes and cover only 1.4 Mb, an increase of only 2.4% in the number of annotated bases. The short length, lack of recognizable protein-coding potential and relatively low or restricted expression (as implied by a reduced number of supporting transcribed sequences) may explain why most lincRNA loci have been missed by previous annotation attempts.

The evolutionary signatures exhibited by this lincRNA set suggest that the vast majority of them are likely to be functional in multiple *Drosophila* species. Their substitution rates are suppressed relative to untranscribed, intergenic regions (**Figure 3.3**). Intergenic regions contain functional DNA elements, such as enhancers or transcription factor binding sites, which are expected to evolve under purifying selection (Andolfatto 2005; Meader et al. 2010). The true neutral substitution rate is therefore likely to be even higher than the rate recorded for these intergenic sequences and therefore the extent of substitution rate suppression observed due to putative lincRNA functionality is likely to have been underestimated. In a similar manner, it is

unlikely that all the bases within a protein-coding gene as defined in **Chapter 3.5.1** are functional. Not all sites are contained within critical functional domains while many of them will be found within introns, which evolve more rapidly than protein-coding exons (Haddrill et al. 2005). The substitution rate relative to such sequences experiencing purifying selection is therefore likely to be even lower than that recorded for these gene models.

Nevertheless, 1,411 lincRNA loci (79.4% of those which could be tested) individually showed evidence of significant sequence constraint relative to short intron sequence between the three species tested here. I interpret this constraint to mean that these loci are likely to be functional: mutations within these sequences are likely to be detrimental to this function, and are therefore preferentially purged. Other evidence supporting the functionality of this set of lincRNAs is discussed below. A further 241 (13.6%) of lincRNAs were individually annotated as evolving significantly faster than the short intron sequences. Mutation events within this set of sequences are assumed to occur at the same rate as within the short introns, but they are fixed at a higher frequency than expected for neutrally evolving sequence. This may be because such changes are beneficial to the organism, which implies that the sequence confers a functional role. These analyses of constraint and fast evolution have greatest power to detect constrained sequence when the aligned sequences are functional. Consequently, they have less power to predict the functionality of lincRNAs that have experienced lineage-specific selective pressures. For example, a functional lincRNA which is constrained between *D. melanogaster*

and *D. simulans*, but which is now evolving adaptively in the *D. yakuba* lineage would be discarded by my analyses. My total estimate of 1,652 functional lincRNA loci in the *D. melanogaster* should therefore be considered to be a lower bound of the number of functional loci whose evolutionary pressures have been preserved over the last 10 million years of *Drosophilid* evolution.

The constrained lincRNA loci are enriched in a number of other predictors of functionality. In a similar manner to substitutions, they tend not to tolerate indel mutations (shown by their enrichment in IPSs) and this pattern appears to be conserved across all 15 insects from which the MCSs were predicted. This function may well be mediated by the mature RNA structure, as these lincRNAs are enriched in EvoFold-predicted RNA secondary structures. While they are not significantly enriched in RNAz-based predictions, they still contain an increased density of these structures relative to neutrally and fast evolving lincRNAs, both of which are significantly depleted. The function(s) of constrained lincRNAs may be developmentally regulated, as shown by their significant and specific enrichment in Polycomb-protein marked heterochromatin. Polycomb target genes are frequently involved in important developmental processes (Sparmann and van Lohuizen 2006), and it may be that the 352 (24.9%) lincRNAs transcribed within these regions are similarly regulated. Alternatively, as it is known that Polycomb-protein marked heterochromatin can be present in wide domains over the target gene and it frequently targets transcription factors (Schuettengruber et al. 2007), this

signal may arise simply from the genomic tendency of constrained lincRNAs and transcription factors to coincide as described in **Chapter 5**. This hypothesis could be tested by examining the expression profiles of lincRNAs transcribed from within these domains, either through real-time RT-PCR or fluorescent *in situ* hybridisation (FISH). Any effects on these profiles by mutations at the Polycomb locus (e.g. Classen et al. 2009) would then confirm that lincRNA expression was regulated by the action of this heterochromatin type.

On the other hand, fast evolving lincRNAs appear to function frequently (52.3%) as the precursors of short RNA species. Both esiRNA and piRNA, whose candidate sequences are enriched in these lincRNA loci, are involved in transposable element suppression. EsiRNAs are thought to function largely in somatic tissues (Ghildiyal et al. 2008). Components of the piRNA biogenesis pathway are important in the germline (Aravin et al. 2007), which suggests that piRNAs might be similarly important for the development of this tissue, although somatic piRNAs have also been reported (Li et al. 2009). Many loci produce both esiRNAs and piRNAs, even though these short RNAs are produced by distinct biogenesis pathways (Ghildiyal and Zamore 2009). When I removed piRNA-overlapping esiRNA candidates, the significant enrichment remained, hence this can be considered to be an esiRNA-specific signal. The enrichment of piRNAs within these sequences is consistent with a similar enrichment of HP1-associated heterochromatin in fast evolving lincRNA loci. Like aubergine (Reiss et al. 2004), a key component of the piRNA biogenesis

pathway (Brennecke et al. 2007), HP1 has been implicated in trans-silencing of transposable elements (Ronsseray et al. 1996). HP1-bound heterochromatin is found within pericentromeric and subtelomeric heterochromatin (Eissenberg and Elgin 2000), which are also frequent sources of piRNA transcription (Aravin et al. 2007).

That fast evolving lincRNAs represent the primary transcripts of transposon suppressors may explain why they appear to have evolved rapidly. Both esiRNAs and piRNAs can be transcribed from repeat elements within the genome (Yin and Lin 2007; Kawamura et al. 2008) and these regions are known to be particularly difficult to place when assembling a genome (Phillippy et al. 2008). This makes orthology assignment difficult and, consequently, alignments in these regions are less reliable. Inaccurate alignments between non-orthologous sequences will have an increased number of differences relative to what is expected from neutrality, and this would be interpreted here as an apparent increased rate of evolution.

Alternatively, these lincRNAs may have evolved quickly because they have adapted to past changes in their target transposon sequences. Different transposable elements are known to evolve at different rates, at least in mouse (Gaffney and Keightley 2006), and, where their suppression is based on sequence complementarity, this should be matched by a comparable evolutionary rate in suppressor sequences. The Red Queen Hypothesis (van Valen 1973) states that "it takes all the running you can do to keep in the same place" (Carroll 1865). In this context, natural selection would preserve

mutations in the short RNA sequences so that they 'run' just to maintain their complementarity and ability to suppress their target transposon. Distinguishing between these two hypotheses would require the identification of the target(s) of individual fast evolving lincRNAs through experimentation, such as investigating whether mutation of the lincRNA locus leads to an increase in transposable element activity. This could be followed by a study of the paired lincRNA/transposon sequences in a variety of related species and/or strains. If the lincRNA sequence is indeed evolving adaptively, then it is predicted that the two sequences should contain compensatory mutations, ensuring that their complementarity is maintained as the lincRNA 'runs' after the transposon sequence. If, instead, the mutations are apparently randomly placed between the two sequences, then it is likely that the lincRNA was originally annotated as fast evolving due to it being misaligned and that there is no functional relationship between the lincRNA and the identified transposable element.

That fast evolving lincRNAs are also enriched in miRNA candidate sequences might be slightly more surprising. MiRNA families have been identified which are deeply conserved across all metazoan species (Wheeler et al. 2009) and evolutionary conservation is often used to aid miRNA predictions, such as those based on the 12 *Drosophila* genome sequences (Stark et al. 2007). Newly arisen miRNAs have, however, been identified in *Drosophila* that show signatures of adaptive evolution (Lu et al. 2008) and miRNAs appear to arise at a faster rate in *Drosophila*, as compared to other insects (Marco et al.

2010). Subsequent curation of the miRNA candidate sequence-overlapping lincRNAs is required to confirm the presence of genuine mature miRNAs within these loci (Kozomara and Griffiths-Jones 2011) but it may be that those miRNAs which map within the fast evolving lincRNAs are recent acquisitions to the *D. melanogaster* genome. This could explain why they were overlooked by previous annotation attempts. More experimental evidence, such as observing expression of these lincRNAs only in species closely related to *D. melanogaster*, would be required to confirm or dispute this hypothesis.

The work presented here confirms the view that interrogating EST databases can still reveal novel and functionally interesting lincRNAs. Using this approach I identified 2,788 lincRNA loci in the *D. melanogaster* genome and predict that at least 1,652 of these are likely to be functional. I investigate the functions of a subset of these lincRNAs in **Chapters 5** and **6**, but the overwhelming majority of these loci remain to be experimentally characterised. By adding these loci to the list of annotated genes in *D. melanogaster*, it will be possible to reach a more comprehensive understanding of the genome of this important model genetic organism.

Chapter 4: LincRNAs LIKELY FUNCTION AS CONSERVED DEVELOPMENTAL REGULATORS

4.1 Abstract

I used a large whole transcriptome shotgun sequencing (RNA-seq) data set to define a set of 1,119, mainly novel, lincRNAs of which only 14% share any overlap with the lincRNAs defined in **Chapter 3**. These lincRNAs are generally shorter and encode fewer alternative transcripts than similarly defined gene models. Significant evolutionary constraint suggests that almost all of these lincRNA loci are functional. Such functions may be related to the regulation of developmental processes, as these lincRNAs are predominantly expressed in early developmental stages and they are frequently found in chromatin domains associated with genes with regulated expression profiles or developmental functions. Specifically, lincRNAs appear to be important in male-specific organs, such as the testes. I also identified, for the first time, a set of lincRNA loci whose locations may be conserved over 700 million of years of evolution, between *Drosophila* and mouse. This study demonstrates that large numbers of functional lincRNAs are a general feature of animal genomes.

4.2 Introduction

Although lincRNA loci have traditionally been identified using EST sequences of the type described in **Chapter 3**, a second generation of sequencing technologies has emerged recently which has a number of advantages over previous approaches (Marguerat and Bähler 2010). These technologies have already been used in a wide range of applications, from genome resequencing (Metzker 2010) to sequencing of DNA associated with particular chromatin proteins using chromatin immunoprecipitation (ChIP) by a technique known as ChIP-seq (Mardis 2007). In this chapter, I focus on analysing an RNA-seq dataset, which represents a very large number of short (75 bp) sequence reads from various cDNA libraries, which were produced as described in **Chapter 2.1.1.4**.

The major advantage of RNA-seq, when compared to Sanger sequencing of ESTs, is its ability to generate greater amounts of sequence data by several orders of magnitude in a rapid and cost-effective manner (Marguerat and Bähler 2010). This allows researchers to investigate several different tissues, and under different physiological conditions. The biases for brain and gonadal tissues seen in EST libraries (Daines et al. 2011) can thus be avoided when attempting a more comprehensive and objective description of an organism's transcriptome.

The form of RNA-seq data is also well suited to the definition of novel, lincRNA transcripts which are expressed at a low level (Costa et al. 2010). It

makes no assumptions about transcript structure. This has been used to build a catalogue of over 100,000 putative ncRNAs in *D. melanogaster* using overlapping sequence tags from a variety of short RNA sequence studies (Jung et al. 2010). Here, it is thought that these clustered tags represent the primary transcript of the short RNA. This lack of reliance on previously defined loci allows protein-coding and lincRNA transcripts to be defined using identical criteria, making direct comparisons between them possible for the first time.

RNA-seq data can also reveal the structure of the novel transcripts to high resolution, which is difficult using either EST or microarray evidence. It does not rely on previous annotations (Wilhelm et al. 2010) and the transcripts based on RNA-seq have single base pair resolution (Marguerat and Bähler 2010). Gene models and splice junctions are called solely based on the recorded sequences, while microarray studies require potential splice sites to be defined *a priori* so that probes matching them can be placed on the slide (Mortazavi et al. 2008). As an aside, RNA-seq data also avoid the problems of cross-hybridisation between probes (Ponting and Belgard 2010) and G+C biases in hybridisation (Halasz et al. 2006) observed in microarray studies. RNA-seq transcripts models are thought to be more complete than those built using other approaches, as a recent study of the *D. melanogaster* transcriptome revealed that exonic sequence could be added to 25% of currently annotated gene models and $>$ 8% of genes could be extended to include a previously unrecognised 5' UTR (Daines et al. 2011). These authors also identified novel

alternative isoforms within their transcript models which could not have been readily detected by other means.

RNA-seq experiments show a high degree of power in detecting differential gene expression. They contain little background noise (Wilhelm et al. 2010) and are able to detect changes in expression over a greater dynamic range than is possible with current microarray platforms (Marguerat and Bähler 2010). It is possible to control the total amount of transcription which has been assayed (Marguerat and Bähler 2010) by selecting only reads which meet predefined conditions, such as those mapping uniquely to genomic sequence. Gene expression values are counted digitally, as a normalised measure of the number of sequencing reads which map to an individual gene, which makes it straightforward to statistically test for differences between different samples. This can even be used to detect allele-specific differences in a heterozygote, as was recently carried out to quantify and describe divergent expression between *D. melanogaster* and a related sister species, *Drosophila sechellia* (McManus et al. 2010).

The raw data presented here were produced by the modENCODE consortium. Like the ENCODE project (Birney et al. 2007), this is a collaboration between various research groups which aims to describe all the functional elements in the genome of two model organisms – *D. melanogaster* and *C. elegans* (Celniker et al. 2009). The collaboration is studying transcript structure, transcription-factor binding sites, chromatin marks and the sites of DNA replication, amongst others, at a genome-wide level in these two species. I

make use of a set of over four billion reads produced by poly(A)$^+$-selected RNA-seq (**Chapter 2.1.1.4**) which is, to my knowledge, the greatest amount of sequence data available for the *D. melanogaster* transcriptome. These data are supplied across 30 developmental time points which allows the investigation of the temporal expression profiles of lincRNAs throughout development. This was not previously possible with conventional EST collections such as those analysed in **Chapter 3**.

In this chapter, I describe the identification of 1,119 *Drosophila* lincRNAs using the modENCODE RNA-seq data and present several lines of evidence to support their functionality. These lincRNAs do not frequently tolerate either substitution or indel mutations, which is consistent with their having biological roles in *D. melanogaster*. From their temporal expression patterns and distribution across chromatin domains, lincRNAs appear important for developmental regulation at specific stages within the life cycle. I also observe that lincRNAs are frequently expressed in only one sex, and that this may be related to the development of sex-specific tissues as they do not appear to be involved in sexual selection. Finally, I identify 42 analogous lincRNAs shared between *Drosophila* and mouse but separated by 700 million years of evolution.

4.3 Materials

4.3.1 InParanoid Database

Orthologous protein-coding genes shared between *D. melanogaster* and *M. musculus* were obtained from the InParanoid7 database of eukaryotic orthologues (Ostlund et al. 2010). This is a general-purpose orthology tool, which has been reported to be the most accurate of all available orthology databases (Hulsen et al. 2006; Altenhoff and Dessimoz 2009). It reports pairwise orthologous genes, known as inparalogues, which may have undergone duplication in one or both species since the divergence from their last common ancestor. Genes which duplicated before this divergence (outparalogues) are not considered.

There are 5,557 orthologous clusters currently reported between *D. melanogaster* and *M. musculus* by InParanoid7. The mean number of proteins in each cluster is 1.2 for *D. melanogaster* and 1.8 for *M. musculus*. 9,391 genes in *M. musculus* represented by proteins contained within these clusters are orthologous to only one *D. melanogaster* gene while 273 are orthologous to multiple genes. 4,507 *D. melanogaster* genes are orthologous to only one *M. musculus* gene and 1,644 are orthologous to multiple genes. I have currently defined 11,775 protein-coding loci in *D. melanogaster* and, as 6,151 of these appear to have orthologous sequences in *M. musculus*, this corresponds to 52.2% of *Dros*ophila loci being conserved in mouse.

I use these orthology relationships to identify protein-coding gene territories (described in **Chapter 2.2.4**) which contain orthologous regions in *M. musculus*. The territories associated with the 6,151 protein-coding loci represent a total of 39.5 Mb (32.9%) of the *D. melanogaster* genome.

4.3.2 Mouse lincRNAs and gene territories

I investigate the relationship between these lincRNAs and a set of lincRNAs identified using the mouse FANTOM3 cDNA collection (Ponjavic et al. 2007). These loci were defined similarly to the lincRNAs described in **Chapter 3**. As of the current release of mouse genome annotations (Ensembl v59), 1,987 of these 3,122 lincRNAs remain intergenic (A.C. Marques, unpublished) and therefore only these 1,987 loci are analysed here.

A set of mouse protein-coding gene territories was constructed using the current Ensembl annotations. The method for partitioning intergenic space is described in **Chapter 2.2.4**.

4.4 Methods

4.4.1 TopHat

Mapping RNA-seq reads to a reference genome requires one to take account of the intron/exon structure of the transcripts being analysed and I do this here using the TopHat program (Trapnell et al. 2009). TopHat is able to identify novel splice sites between adjacent exons of the same transcript, which is

important when investigating previously unannotated transcripts such as lincRNAs. It is also optimised to detect splice junctions in transcripts, like lincRNAs, which are expressed at a very low level. TopHat was selected over QPALMA (De Bona et al. 2008), which uses machine learning to identify novel splice junctions, because the latter's vastly reduced efficiency made it impractical for the amount of data analysed here.

The TopHat pipeline, which uses three lines of evidence to annotate a novel splice junction, is summarised in **Figure 4.1**. Reads are split into 25 bp segments and mapped independently with Bowtie (described in **Chapter 2.2.1.2**). Those which do not map to the genome are collected separately as a set of possible intron-spanning, 'initially unmappable' (IUM) reads. Those which do map are assembled into consensus islands using the Maq assembly module (Li et al. 2008) and all possible canonical splice sites between neighbouring islands are recorded. With the 75 bp reads used here, these canonical sites can take the form of GT-AG, GC-AG and AT-AC. The neighbouring islands need not be directly adjacent, but there is a maximum genomic distance between them which can be set by the user. I supplied the minimum and maximum possible intron size as 52 bp and 11929 bp, respectively, as these lengths covered 95% of the intron length distribution in *D. melanogaster*, using the updated FlyBase Release Version 5.27 (Tweedie et al. 2009). The transcripts which span introns shorter than this minimum distance are merged into a single exon. Introns which are within this length

distribution and which are supported by one of the canonical models mentioned above are retained.

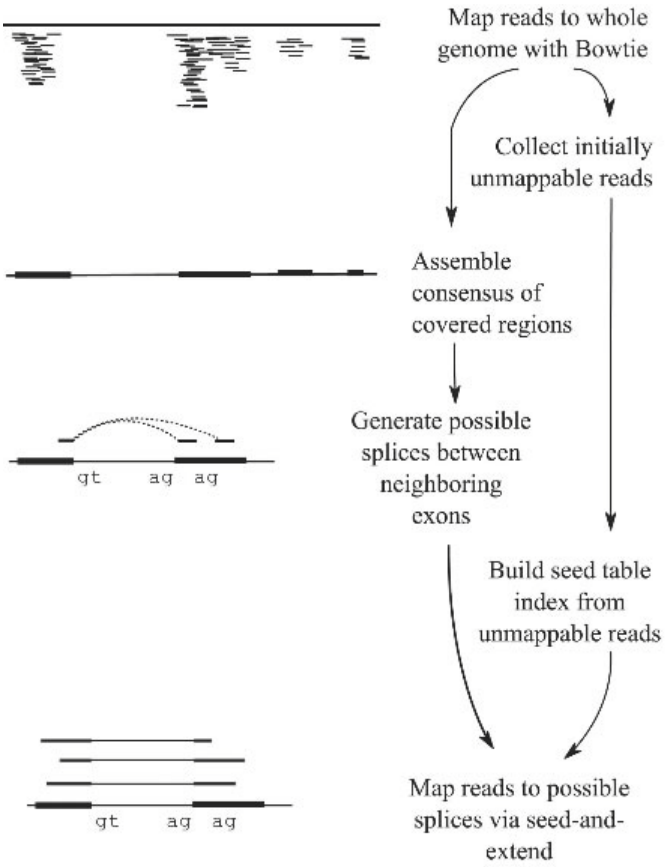

Figure 4.1 The TopHat pipeline. Taken from Trapnell et al. 2009.

Secondly, paired-end reads can also support a splice junction, where one mate pair maps to one island and the other maps to a different, neighbouring one. The user needs to supply TopHat with the expected (average) and standard deviation of the internal distance between mate-pairs. I calculate this separately for each sequencing run where the paired-end reads are mapped individually using Bowtie. The standard deviation is estimated by first calculating the full-width at half maximum (FWHM) of these distances. The

FHWM is defined as the width of the internal distance frequency distribution at half the maximum frequency observed. By assuming these distances to be normally distributed, the standard deviation is then calculated as shown in **Equation 4.1**

$$\text{Std. Deviation} = \frac{FWHM}{2.35482}$$

Equation 4.1 Calculation of standard deviation from FWHM.

The third situation which can support a splice junction takes place when an individual read spans the junction: such situations are identified by a 'seed-and-extend' strategy. All canonical splice junctions within the high-quality 5' end (the 5' most 25 bp, by default) of the IUM reads are indexed by 10-mers made up of 5 bp upstream and 5 bp downstream of the putative splice site. The 10-mers surrounding the possible splice junctions identified between neighbouring islands are then searched sequentially for identical hits in the IUM reads. If an identical hit is found, the alignment is extended left and right from the splice junction and retained only if the total number of mismatches is below 2 bp (the default setting).

A splice junction from any of the three lines of evidence is only retained if it is supported by sufficient read coverage. This threshold is 15% of the more deeply sequenced exon adjacent to the splice site, as a previous RNA-seq study suggested that 86% of minor isoforms are supported by at least 15% of the coverage of the major isoform (Wang et al. 2008).

TopHat was provided with *D. melanogaster* gene models from FlyBase release 5.27 gene annotations (Tweedie et al. 2009) and from the set of lincRNAs defined in **Chapter 3**. To exclude putative intergenic transcripts that represent unannotated exons of proximal protein-coding genes, I defined raw junctions (the optional option j for TopHat) as the adjacent end-points of neighbouring EST-defined lincRNA loci and FlyBase genes. This directs TopHat to seek reads that span 5' and 3' positions of previously unannotated splice junctions. All other options were left at their default values.

4.4.2 Transcript assembly

4.4.2.1 **Cufflinks**

Reads as mapped by TopHat (**Chapter 4.4.1**) are assembled into transcripts by the Cufflinks program (Trapnell et al. 2010), as summarised in **Figure 4.2**. Briefly, the mapped reads, known as fragments, are clustered as in **Figure 4.2a** into overlapping bundles where each bundle can be analysed separately. Mutually incompatible fragments that must have originated from different transcripts are identified, as shown in **Figure 4.2b.** The resulting overlap graph is used to draw paths through these fragments, connecting possible compatible fragments. The total number of paths equals the total number of incompatible fragments and the minimum number of transcripts. Fragments which are compatible with more than one transcript are associated with all compatible transcripts in **Figure 4.2c.**

The abundance of each of these transcripts is estimated using a statistical model described in **Figure 4.2d-e**. Fragments are assigned to individual transcripts using a maximum likelihood function which takes account of the length distribution of fragments. For example, it is more probable that the spliced violet fragment which spans the red and blue transcripts was originally part of the blue transcript as, otherwise, it would have to be much longer than the other fragments in the sample. The maximum number of iterations for maximum likelihood estimation was increased here from the default 5,000 to 25,000 to improve the accuracy of assigning fragments to individual

transcripts. The relative abundance of each fragment can then be used to estimate the relative abundance of each transcript as the number of fragments per kilobase of exon model per million mapped fragments (FPKM).

4.4.2.2 **Cuffcompare**

Each developmental stage was analysed in isolation due to constraints on the total amount of computer memory available and, as such, the reads mapped from each developmental stage produced a specific set of transcripts. These were collated into a single set of consensus transcripts using Cuffcompare, a program supplied along with Cufflinks (Trapnell et al. 2010).

4.4.2.3 **Cuffdiff**

Expression differences were tested using the Cuffdiff program, which is available along with Cufflinks (Trapnell et al. 2010). All high-confidence transcripts can be tested for significant differences between two samples using the variance estimates produced by the maximum likelihood function described in **Chapter 4.4.2**. A modified one-sided T-test is applied to the transcript's FPKM values, with a Benjamini-Hochberg correction made for multiple testing.

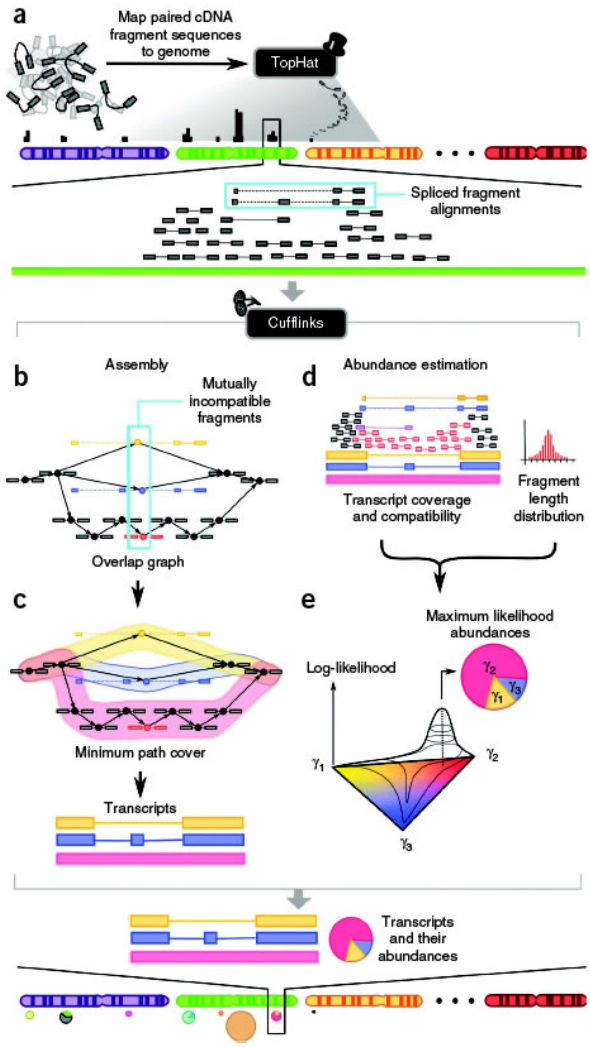

Figure 4.2 Cufflinks. A. Input mapped sequences. B. Construction of mutually incompatible fragments. C. Construction of minimum paths. D. Transcript abundance estimation. E. Log-likelihood function. Taken from Trapnell et al. 2010.

4.5 Results

4.5.1 Short-read assembly pipeline

The RNA sequences described in **Chapter 2.1.1.4** were mapped onto the *D. melanogaster* genome separately for each developmental time point data set as shown in **Figure 4.3** over a period of approximately four months.

Both pairs of each paired-end read were mapped separately using Bowtie (**Chapter 2.2.1.2**) so that the mean and standard deviation of the insert size for paired-end reads could be calculated for each sequencing run.

The 5' and 3' positions of splice junctions were mapped for each sequencing run (whether single- or paired-ended) separately using TopHat (**Chapter 4.4.1**). Reads Per Kilobase of exon model per Million reads mapped (RPKM) values were calculated for each FlyBase-defined gene model for each sequencing run. This was done by dividing the number of reads mapping to a particular gene by the length of the gene, and the total number of reads mapped in that run. As shown in **Figure 4.4a**, the RPKM values obtained for different sequencing runs are highly reproducible. Newly-called splice junctions from one or more sequencing runs but the same cDNA library were collated and appended to the previous raw junctions before reads were remapped using TopHat (with all other parameters held constant). This allowed TopHat to map reads in one sequencing run which supported a splice junction found in a separate run, but which previously had insufficient reads to be called. A single RPKM value was then calculated for each FlyBase gene model using reads from all sequencing runs for that cDNA library. The splice junctions produced by each cDNA library for each individual developmental time point were collected together, and added to the raw junctions defined by neighbouring FlyBase genes and EST-defined lincRNA loci. All reads from this time point were then mapped for a third and final time using TopHat. This allowed reads in one cDNA library to now support a splice junction found in a separate

library. Again, the results of this mapping are reproducible as shown by the example correlation in **Figure 4.4b**. They were assembled into a set of time point-specific transcripts using the Cufflinks program (**Chapter 4.4.2**). Here, the mean mate-pair insert size and standard deviation supplied to the program were calculated from all paired-end reads mapped from all replicates of the same cDNA library. The final collection of mapped reads and their associated transcript numbers is summarised in **Appendix A**.

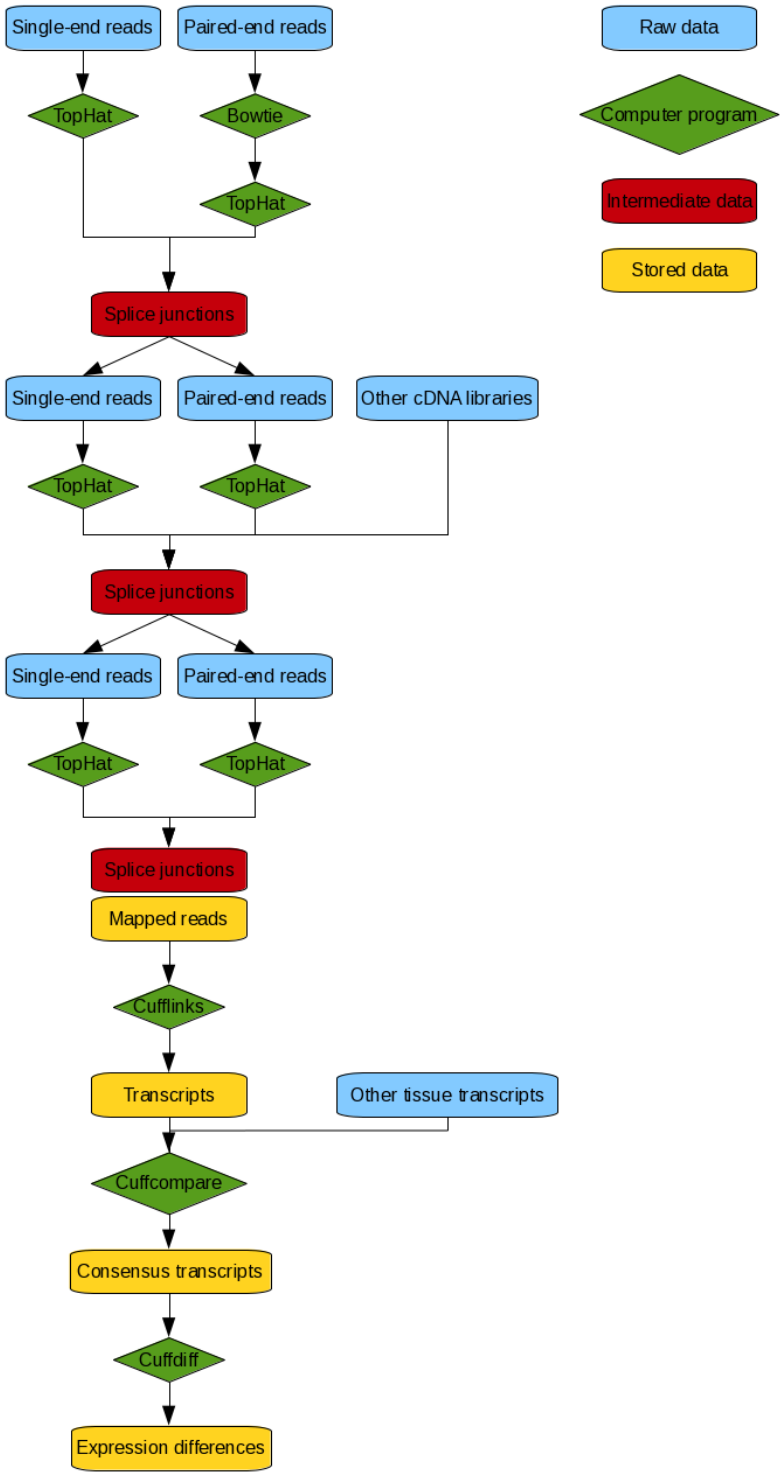

Figure 4.3 Flow diagram showing mapping pipeline used to define transcripts from RNA-seq data.

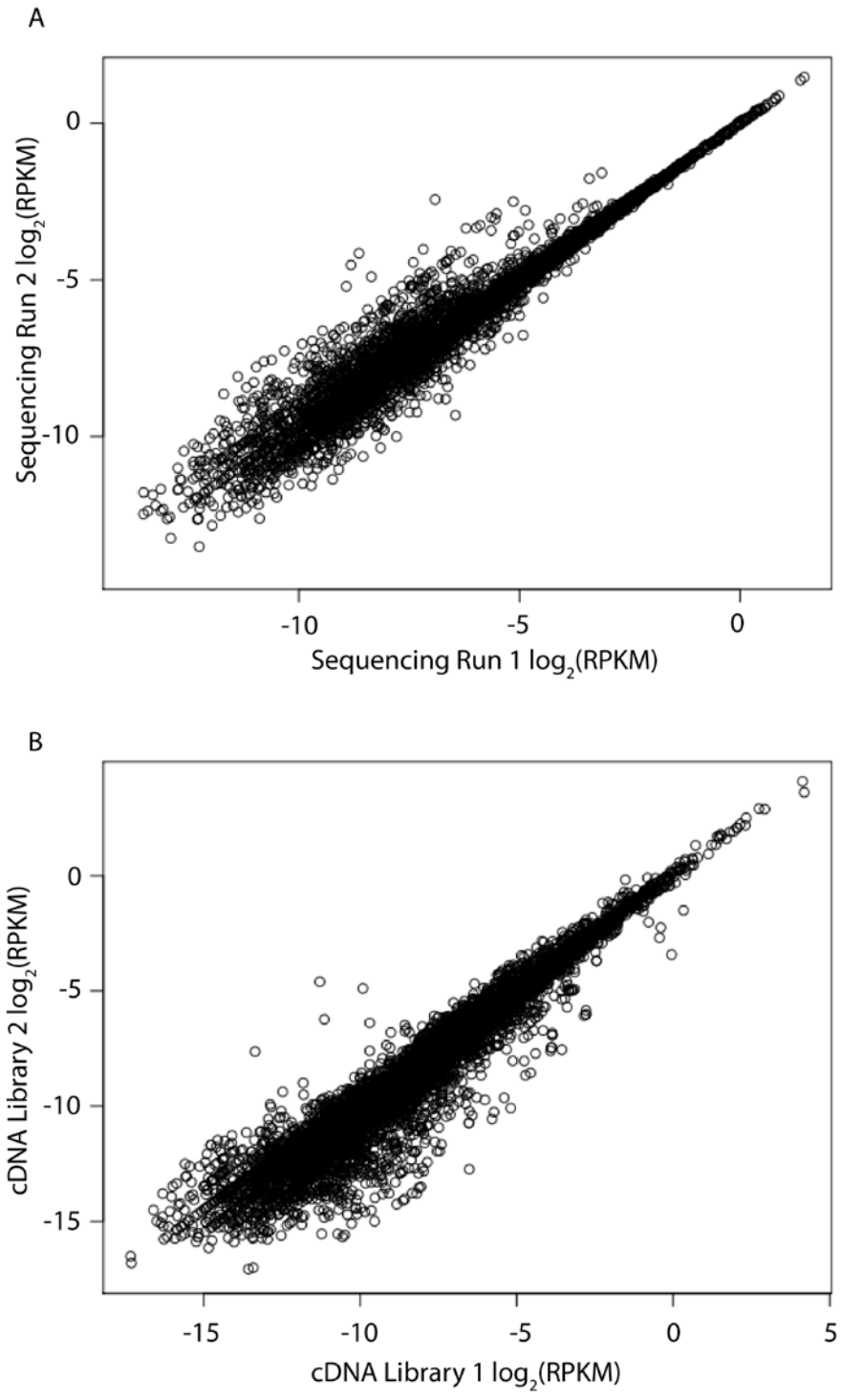

Figure 4.4 Consistency of mapping protocol. A. Example correlation of FlyBase RPKM values from sequencing runs of the same cDNA library. B. Example correlation of FlyBase RPKM values from cDNA libraries of the same tissue.

4.5.2 Comparative Transcriptomics

Cuffcompare (**Chapter 4.4.2.2**) was used to build a consensus transcript set using the transcript models for all 30 developmental time points. The mate-pair insert size and standard deviation were calculated from all paired-end reads mapped across all stages. Differential expression of these transcripts across time points was then estimated using Cuffdiff (**Chapter 4.4.2.3**). As Cuffdiff allows only pairwise comparisons, developmental time points were analysed sequentially and then separately for males and females when appropriate. Also, differences between age-matched male and female samples were investigated, with the parameters set as above. Here, instead of using the RPKM values as above, individual transcript expression levels were quantified using FPKM values (Fragments Per Kilobase of exon per Million fragments mapped) as reported by Cufflinks. The use of this quantity is more suited to paired-end reads, as it reports on the simultaneous mapping of the two read ends of the cDNA fragment, rather than on the mapping of individual reads. This also allows overlapping transcripts to be quantified separately, depending on to which transcript individual fragments had been assigned. When considering stage-specific expression (embryo, larva, pupa, adult), a gene was considered to be expressed in a stage if it was associated with an FPKM > 1 (Mortazavi et al. 2008) in at least one of the time points contained within that stage. Male- and female-specific gene models were defined if they were expressed with an FPKM > 1 in at least one stage in one sex, but absent from all stages in the other sex. These FPKM values were $\log_2$-transformed to

produce an approximately normal distribution from which standard statistical analysis could be applied.

4.5.3 Transcript and Gene Annotation

To ensure that these transcript models did not represent genomic DNA contamination in the cDNA libraries I only considered transcripts longer than 200 bp that were either multi-exonic, or unspliced and expressed in multiple tissue samples, where the transcript contained sufficient reads for Cuffdiff to test for differential expression in at least one comparison. All other transcript models were discarded.

I defined a gene model as a cluster of one or more transcripts that are connected through shared exonic or intronic bases, as shown in **Figure 4.5**. Note that not all pairs of transcripts in a gene need overlap. Models overlapping a known FlyBase gene by at least one base on either strand were associated with that gene. Those transcript models that lay in the intergenic regions represent putative intergenic non-coding RNA loci.

I calculated the coding potential of all intergenic gene loci using the Coding Potential Calculator (Kong et al. 2007) (**Chapter 2.2.2**). The exonic bases for each transcript in a model were analysed separately, and in both orientations (forward and reverse strands). A transcript was deemed to be non-coding if the coding potentials of both strands scored less than zero. If all transcripts within a model are non-coding, then it is labelled as a lincRNA locus.

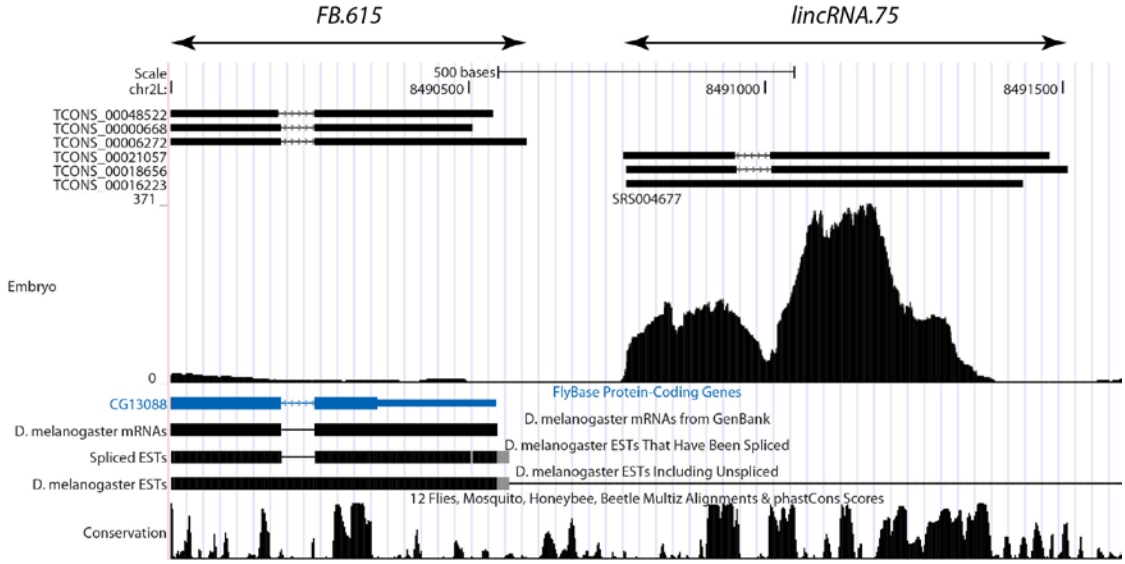

Figure 4.5 Definition of genomically adjacent protein-coding (*FB.615*) and lincRNA locus (*lincRNA.75*) gene models. The black boxes denote exons called by Cufflinks, with arrowed lines representing introns separating exons within the same transcript. A histogram of read counts that support these models' sequences is shown below (from embryonic tissues, 18-20 hours after egg-laying). Note that the FlyBase-defined gene model (*FB.615*) is expressed at a very low level in this tissue.

4.5.4 Annotation of 1,119 lincRNA loci in *D. melanogaster*

In total, I recorded 7,540 gene models (covering a total of 88.9 Mb) and 1,119 lincRNA loci (covering 2.0 Mb) in the *D. melanogaster* genome using this pipeline. Transcriptional evidence was available for 13,463 (92.5%) FlyBase genes, in that they had at least one base overlap with my gene models. Of these I identified transcription from 13,022 protein-coding loci and 441 non-protein-coding genes. I call fewer gene models than FlyBase because I consider all overlapping transcripts on either strand to be part of the same gene model, and so do not call overlapping genes as separate entities. Certain genes which are expressed in narrow spatial or temporal patterns will also be absent from this data set. Those intergenic regions between gene models and lincRNA loci, for which there is no evidence of transcription, were annotated as 'intergenic

sequences'. The numbers of each type of gene, their lengths and their expression profiles are summarised in **Table 4.1**.

At the same time as this was work was being carried out, the modENCODE consortium themselves identified 1,938 new transcribed regions (NTRs) which are not associated with previously annotated gene models (Graveley et al. 2011). Up to one-third of these may be protein-coding, and in contrast to this work they were not required to be truly intergenic (some NTRs are intronically-encoded). There is little overlap between these lincRNA models and the transcripts called by the modENCODE consortium (**Figure 4.6**). While 30.8 Mb (74.2%) of exonic gene models are also identified by modENCODE, only 200 kb (13.3%) of my lincRNA exons within 333 loci were identified by modENCODE. My use of three rounds of mapping appears to have contributed significantly to this increased sensitivity, and I have mapped over 200 million more reads than reported by Graveley et al. 1.3 Mb of the novel 3.7 Mb of transcribed sequence identified by modENCODE is intronic to my gene models and lincRNAs, while the remainder was likely identified from their total RNA samples and microarray studies. Graveley et al. report that only 84% of their NTRs were identified by the poly(A)$^+$ sequence analysed here, which would be consistent with my detection of 89% of their transcribed nucleotides.

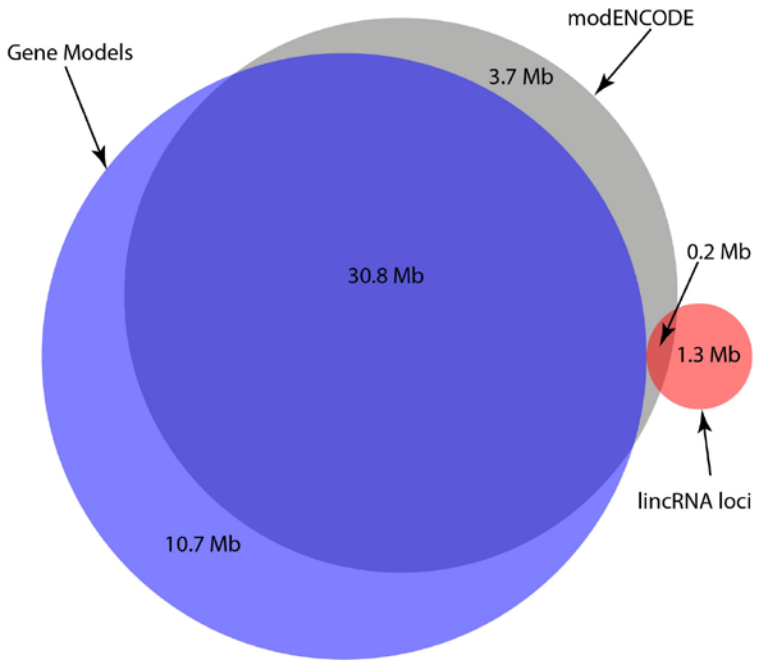

Figure 4.6 Exonic overlap between gene models predicted by the modENCODE consortium and my gene models and lincRNA loci.

LincRNA loci tend to be relatively simpler compared to gene models. As shown in **Table 4.1**, they are shorter and have fewer transcripts in each model. There are also fewer exons in the multi-exonic lincRNA transcripts, relative to the gene models (there is a median of four exons in a multi-exonic gene-defined transcript compared to two in the lincRNA transcripts). It was also noted that only 156 (13.9%) of the lincRNA loci defined here shared at least one base overlap with the lincRNA loci defined by EST evidence in **Chapter 3**. While gene models show an excess frequency of RNA secondary structures, lincRNA loci show no such bias (data not shown).

Sequence type	**Structure/ Expression**	**Number of gene loci**	**Median gene length (bp)**	**Median number of alternative transcripts**	**Median number of tissues in which expressed**	**Median $\log_2$(FPKM)**	**Standard Error $\log_2$(FPKM)**
Gene model	Multi-exonic	7,414	1,700	2	30	1.93	1.61
	Single exon, expressed in multiple tissues	126	873	1	18.5	N/A	N/A
lincRNA loci	Multi-exonic	1,049	443	1	11	-1.52	1.54
	Single exon, expressed in multiple tissues	70	235	1	2	0.30	N/A
Intergenic	N/A	8,669	669	N/A	N/A	N/A	N/A

Table 4.1 Characteristics of gene models, lincRNA loci, and untranscribed intergenic sequences.

There is no evidence that the lincRNA loci identified here represent alternative transcripts of genomically adjacent protein-coding genes. By definition, no transcripts were called that link neighbouring protein-coding genes. By looking at the raw reads which support the individual lincRNA loci models (e.g. **Figure 4.5**) it appears that there is a clear division between adjacent gene models with intervening regions showing very little, if any, evidence of transcription. In fact, only one lincRNA could be joined by its neighbouring gene with reads mapped to a density greater than 1 FPKM, the commonly used threshold for reliable expression using this type of data (Mortazavi et al. 2008). The RNA-seq data used here represent, by far, the greatest amount of information on the *Drosophila* transcriptome. If a lincRNA is part of an adjacent transcript, then these data have the best chance of detecting a read which spans the two gene models. Furthermore, I looked at the distribution of intergenic distances between the gene models and lincRNA loci. If lincRNAs do represent unannotated exons, it would be expected that they would be found close to their neighbouring genes. Actually, the lincRNA loci tend to be further away from gene models than these gene models are from each other (median 452 bp for gene-gene intervals and 2,269 bp for gene-lincRNA intervals, Mann-Whitney p-value $< 2.2\text{x}10^{-16}$).

These lincRNAs are rarely, if ever, the precursors of previously identified short RNA species. As shown in **Figure 4.7**, lincRNA loci are significantly depleted in miRNA (-13.1%, $p = 2.4\text{x}10^{-3}$), piRNA (-98.6%, $p = 1.0\text{x}10^{-4}$), and esiRNA (-56.7%, $p = 1.0\text{x}10^{-4}$) sequences relative to their overall genomic

representations. MiRNA sequences are significantly enriched in the gene models (17.4%, $p = 1.0\text{x}10^{-4}$) which would be consistent with many of them being transcribed from within protein-coding genes (Rodriguez et al. 2004; Marco et al. 2010). Conversely, esiRNAs and piRNAs are significantly enriched in the intergenic regions of the genome (3.7%, $p = 1.0\text{x}10^{-4}$ and 6.5%, $p = 1.0\text{x}10^{-4}$, respectively) – the sequences for which there is no evidence of transcription. This is likely to be because the RNA-seq data derive from poly(A)$^+$-selected libraries while esiRNAs and piRNAs do not possess polyA$^+$ tails, being transcribed by RNA Polymerase III (Miyoshi et al. 2010).

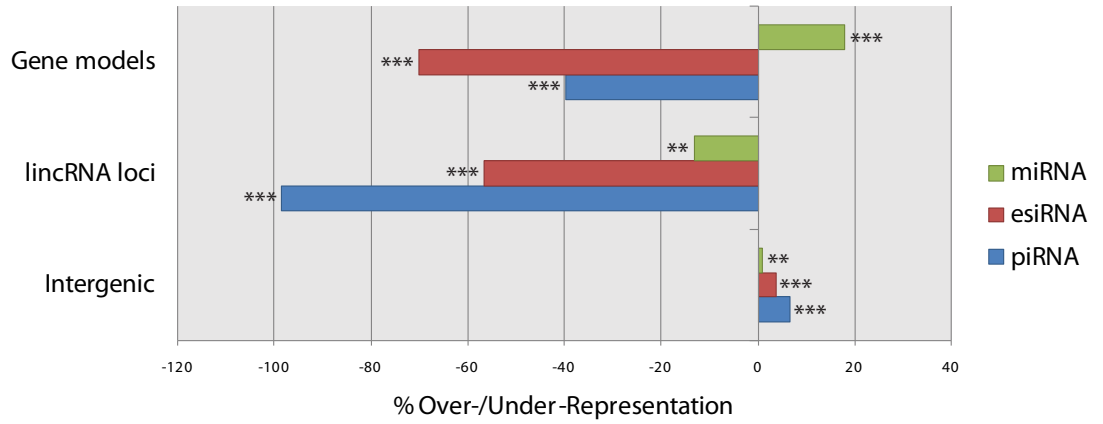

Figure 4.7 Enrichments or deficits of different short RNA classes within gene models, lincRNA loci, and untranscribed intergenic sequence relative to genome-wide random expectations (** indicates $p < 0.01$ and *** indicates $p < 0.001$).

4.5.5 lincRNAs show evolutionary signatures of functionality

As in **Chapter 3**, I investigated the functionality of these lincRNA loci by testing their transcribed sequences for evolutionary constraint. LincRNAs, like gene models, tolerate fewer substitution mutations (**Figure 4.8**) between *D.*

melanogaster and its sister species *D. simulans* and *D. yakuba*, relative to the non-expressed, intergenic regions. There is no significant difference between lincRNA and gene model substitution rates when aligned to *D. simulans* (Mann-Whitney, $p = 0.067$), but all other comparisons were highly significant (Mann-Whitney $p < 1.0\text{x}10^{-12}$).

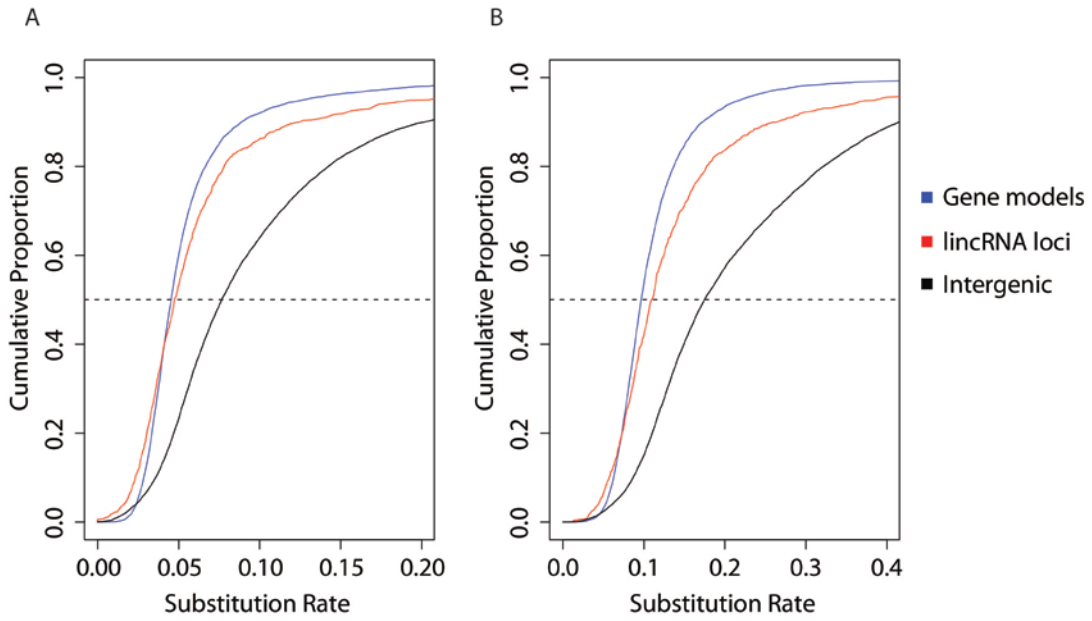

Figure 4.8 Cumulative frequency distributions of nucleotide substitution rates of protein-coding genes (blue), lincRNA loci (red) and intergenic regions (black). The dashed line indicates the 50th percentile. A. Alignments between *D. melanogaster* and *D. simulans*. B. Alignments between *D. melanogaster* and *D. yakuba*.

I repeated this analysis to compare novel lincRNAs which were identified here, and those which overlapped transcripts which were also predicted by the modENCODE consortium. As transcript models from loci expressed at lower levels are thought to be less robust (Mortazavi et al. 2008), I split the novel lincRNAs into those which are highly expressed (observed in one developmental time point at FPKM ≥ 1) and those which are lowly expressed

(observed only at FPKM < 1). Their substitution rate distributions are similar (**Figure 4.9**), but novel lowly expressed lincRNAs have a significantly lower substitution rate than either of the other two groups of lincRNAs (Mann-Whitney, $p = 2.3\text{x}10^{-8}$ and $4.8\text{x}10^{-10}$ when comparing novel lowly expressed lincRNA and modENCODE-overlapping lincRNA alignments between *D. melanogaster* and *D. simualans* and *D. yakuba*, respectively). There is, therefore, no evidence that these novel loci which were missed by the modENCODE consortium are less likely to represent functional lincRNAs; indeed they appear to be slightly more functionally constrained.

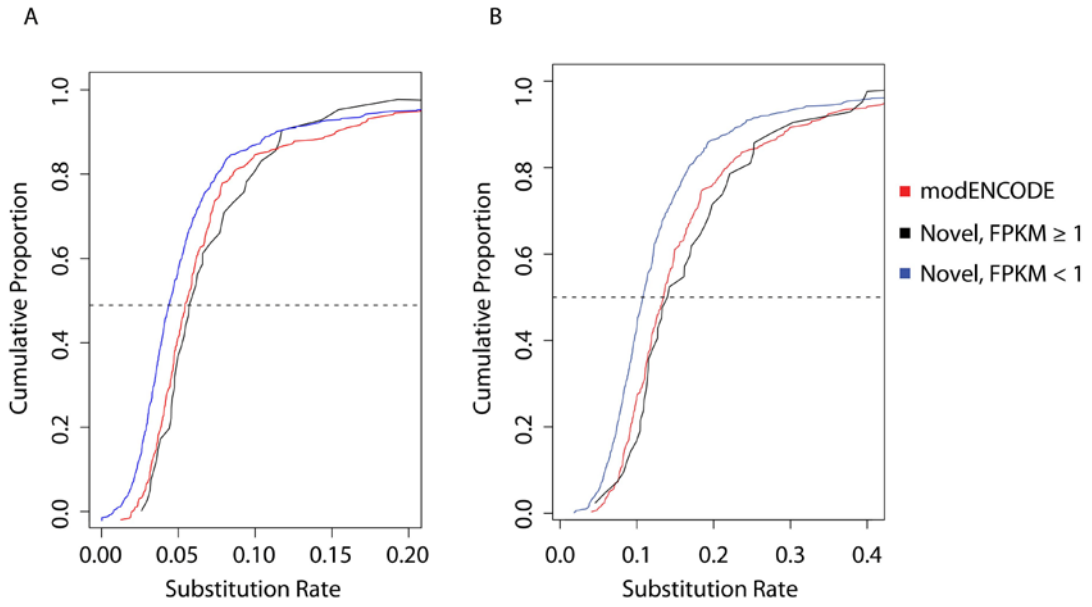

Figure 4.9 Cumulative frequency distributions of nucleotide substitution rates of lincRNAs identified by modENCODE (red), novel highly-expressed lincRNAs (with an FPKM ≥1, black) and novel lowly-expressed lincRNAs (FPKM < 1, blue). The dashed line indicates the 50th percentile. A. Alignments between *D. melanogaster* and *D. simulans*. B. Alignments between *D. melanogaster* and *D. yakuba*.

LincRNA loci tolerate fewer indel mutations, as shown in **Figure 4.10** by their significant 4.5% ($p = 1.0\text{x}10^{-4}$) enrichment in Indel-Purified Segments

(IPSs) between *D. melanogaster* and *D. simulans*. At a greater evolutionary distance, lincRNAs are also significantly enriched (14.0%, $p = 1.x10^{-4}$) in deeply conserved Multi-Species Conserved Segment (MCS) regions (**Figure 4.10**). Taken together, it is clear that lincRNA loci share several indicators of functionality with transcripts covering well-established gene models.

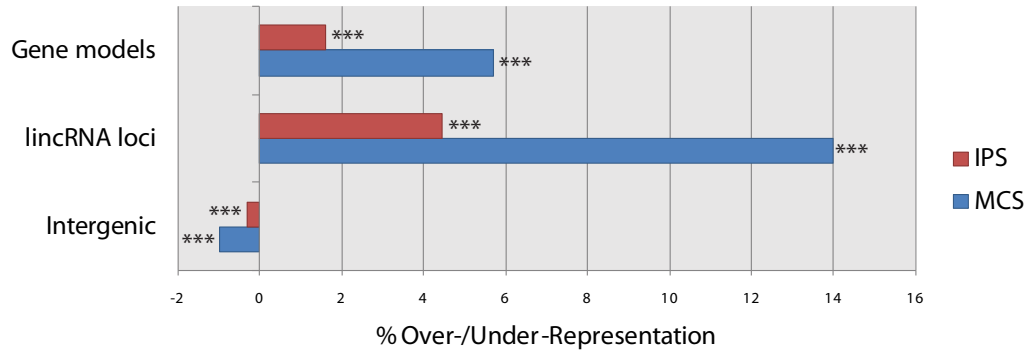

Figure 4.10 Enrichments or deficits of conserved sequence (IPS and MCS) within exonic sequences from gene models and lincRNA loci, and intergenic space, relative to genome-wide random expectations (*** indicates $p < 0.001$).

I partitioned lincRNAs into constrained, neutral and fast-evolving sets using the method described in **Chapter 3.4.1**. 1,074 (96%) lincRNAs were annotated as constrained and therefore I decided not to use this categorisation. Instead, I consider all 1,119 loci as a single group in further analyses.

4.5.6 LincRNAs may function in developmental regulation

I next examined the importance of lincRNA transcription to the overall transcriptome at each of the 30 developmental time points. LincRNA expression tends to be substantially lower than that of gene models, as is clear from the very different scales on which their total $\log_2$(FPKM) values are

plotted on **Figure 4.11**, and there are three orders of magnitude between these values' medians (-55 for lincRNAs compared to 25,834 for gene models). I also tested if there were any significant differences in expression across the life cycle for the two sequence classes. The total $\log_2$(FPKM) values for each class are approximately normally distributed (data not shown) and so a linear regression was fitted, using the embryonic stage as a baseline because it has the largest number of time points. For gene models, there are significant increases in the mean $\log_2$(FPKM) in the pupal stage (1.1-fold, $p = 6.3\text{x}10^{-5}$), and for adult males (1.2-fold, $p = 9.6\text{x}10^{-6}$). The lincRNA loci show a significant decrease during the pupal stage (-1.7-fold, $p = 1.5\text{x}10^{-2}$) but a significant increase in adult males (1.5-fold, $p = 4.9\text{x}10^{-5}$). LincRNAs therefore may be more important in the regulation of early developmental and male-specific processes.

This increased expression early in the life cycle, and the observation that lincRNAs are generally more specifically expressed (**Table 4.1**), led me to examine the relationship between the breadth of expression and the evolutionary rates of lincRNAs. The substitution rates presented in **Figure 4.8** for alignments with *D. yakuba* were $\log_2$-transformed so that they approximated a normal distribution. A linear regression on lincRNA $\log_2$(substitution rates) then showed that this rate increased with the number of stages in which expression is observed, as shown in **Figure 4.12**. For gene models, only those 'house-keeping' genes which were found in all four stages showed a significantly reduced substitution rate, relative to the median; an

observation which has previously been made in mammals (Zhang and Li 2004). It appears that the more restricted a lincRNA is in its expression, the slower its evolutionary rate and therefore the more likely that this constrained sequence is important for its function. The relatively high substitution rate observed for lincRNAs expressed in all major stages, when contrasted with the median substitution rate in red, suggests that these lincRNAs may be less constrained in both their expression and function. LincRNA function therefore tends to be specific to individual developmental stages and, I hypothesise, may involve the regulation of processes specific to each of these stages. The increased expression of lincRNAs in the earlier stages also suggests that this regulation is likely to be particularly important to these developmental processes.

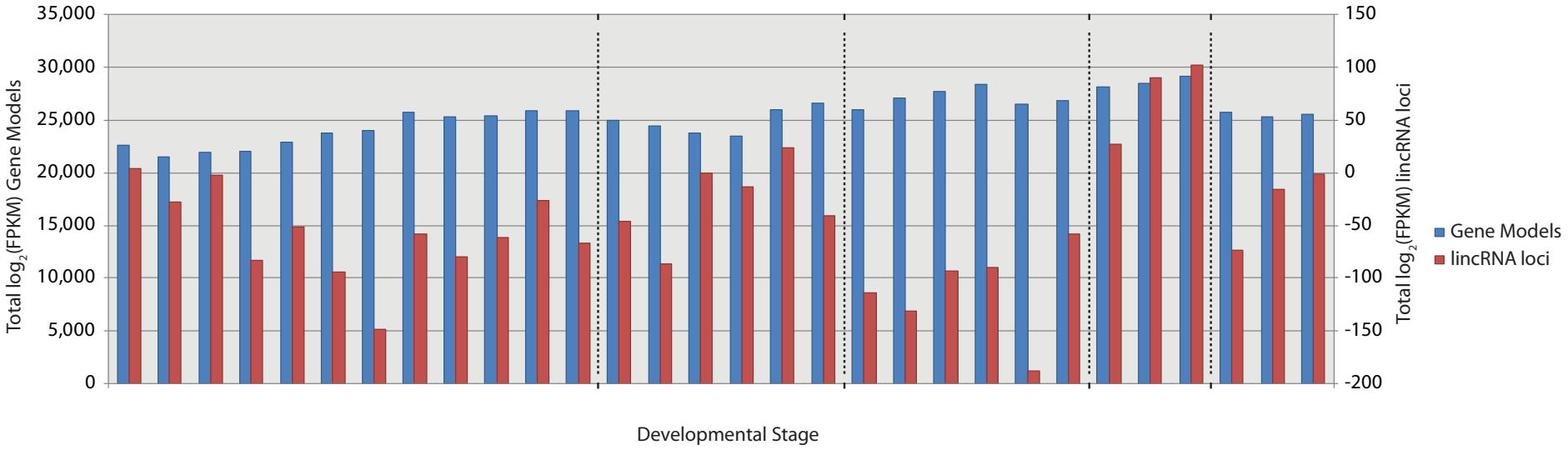

Figure 4.11 Expression levels of gene models and lincRNA loci across 30 developmental time points. Summed $\log_2$(FPKM) values for each time point are plotted for gene models (blue; left vertical axis) and lincRNA loci (red; right axis).

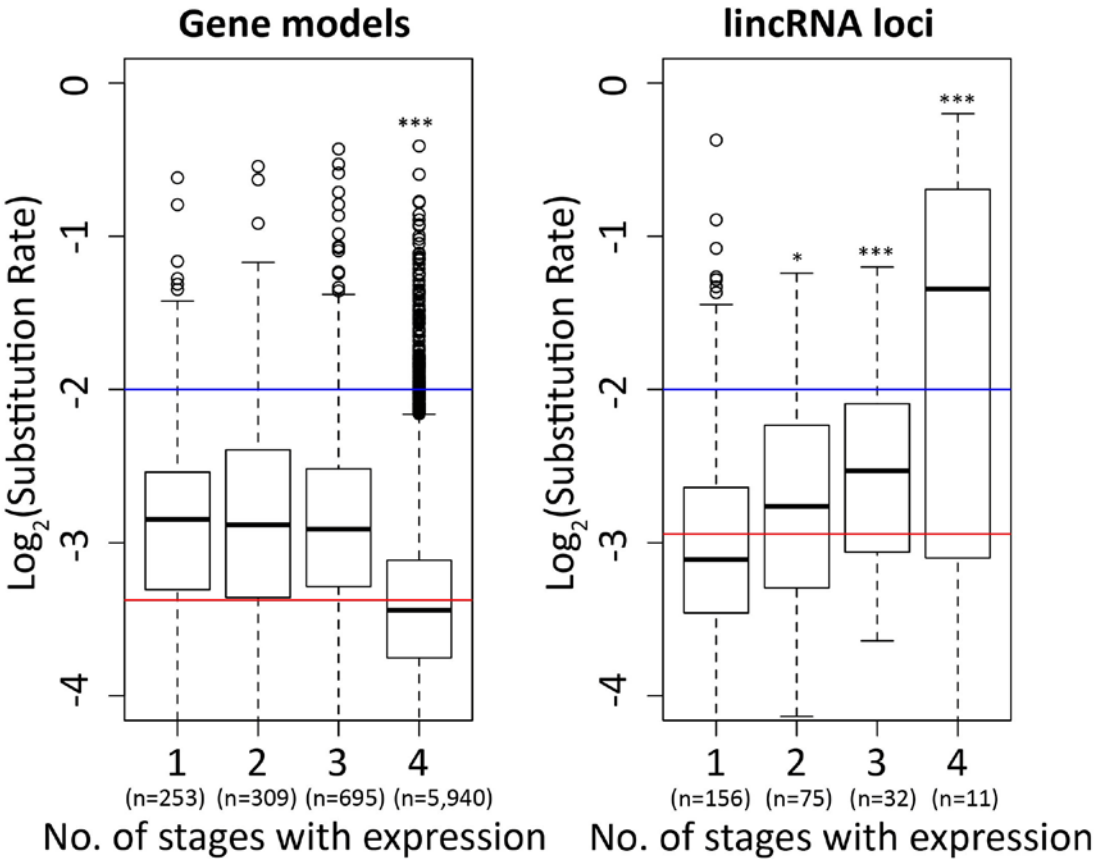

Figure 4.12 Box-and-whiskers plot of $\log_2$(substitution rates) for gene models (left) and lincRNA loci (right) for increasing breadth of expression across 1 or more of 4 developmental stages (linear regression, * indicates $p < 0.05$, and *** indicates $p < 0.001$). Red lines indicate $\log_2$(mean substitution rate) for the sequences examined here. Blue lines indicate the $\log_2$(mean substitution rate) for presumed neutrally evolving short introns. Note that only genes and lincRNAs which are expressed at greater than 1 FPKM in at least one developmental stage are plotted here.

To further test my hypothesis that lincRNAs are important in developmental regulation, I looked at their distribution across a set of genome-wide chromatin domains (**Chapter 2.1.9**). I found that lincRNA loci are significantly enriched within euchromatic regions which display a regulated expression (71.8%, $p = 4.0\text{x}10^{-4}$) and those heterochromatic regions which are marked by the Polycomb group of proteins (45.3%, $p = 1.0\text{x}10^{-4}$). 444 (39.7%) of my lincRNA

loci are found within one of these two types of chromatin domains. Conversely, gene models are enriched in euchromatic regions with both regulated and specific expression (16.6%, $p = 1.0\text{x}10^{-4}$ and 22.2%, $p = 1.0\text{x}10^{-4}$, respectively), which would be consistent with them having both regulatory and structural roles. The novel heterochromatic regions are slightly enriched only in intergenic regions (0.56%, $p = 3.3\text{x}10^{-2}$, data not shown).

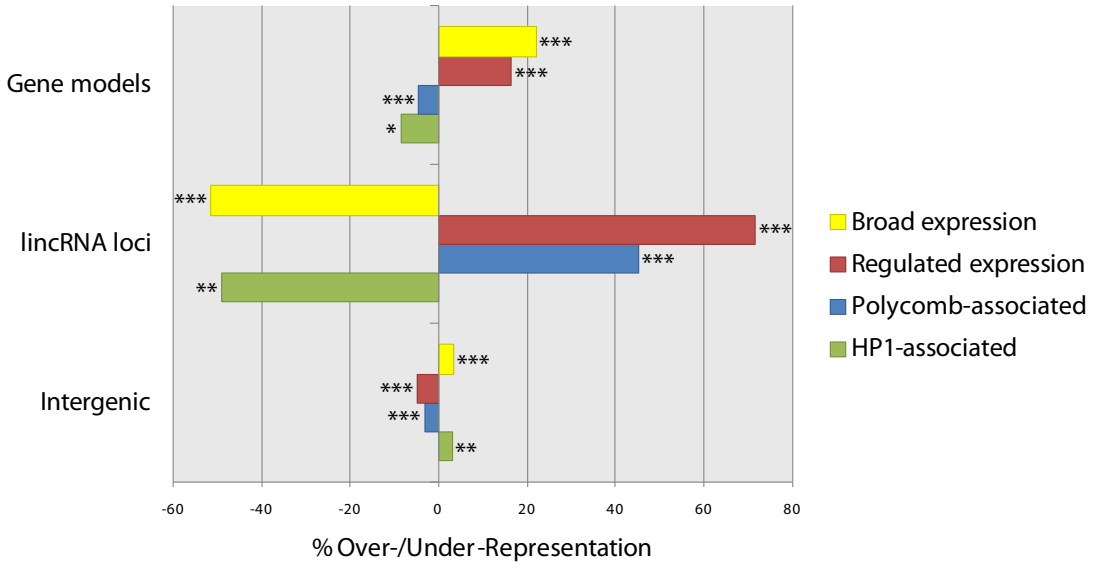

Figure 4.13 Enrichments or deficits of different chromatin types within gene models, lincRNA loci, and untranscribed intergenic sequence relative to genome-wide random expectations (* indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$).

4.5.7 Sex-specific behaviour of lincRNAs

Due to the presence of adult male- and female-specific time points in the RNA-seq data, I was interested in investigating the role, if any, of this lincRNA set in sexual differentiation. I defined gene models and lincRNA loci as being male- or female-specific if they were found to be expressed at FPKM ≥ 1 in one sex in at least one of the three time points for which there are sex-specific

sequence data, but not found to be expressed in any of the samples from the other sex (e.g. **Figure 4.14**). There are significantly more sex-specific lincRNAs than sex-specific gene models – 151, of which 139 are male-specific, against 121 sex-specific gene models (two-tailed chi-squared $p < 2.2\text{x}10^{-16}$). This is surprising given that there are almost seven times as many gene models as lincRNA loci in the *Drosophila* genome. Interestingly, male-specific gene models show an increased substitution rate when aligned to *D. yakuba* (median increase 1.5-fold, Mann-Whitney $p < 2.2\text{x}10^{-16}$) as shown in **Figure 4.15**, relative to those which show no specificity, which would be consistent with their roles in sexual selection (Haerty et al. 2007). Male-specific lincRNAs, on the other hand, show no such bias (Mann-Whitney $p = 0.21$) and I suggest that this is because their roles in the fly are to regulate the development of the male body plan, rather than being involved in male-specific selective processes.

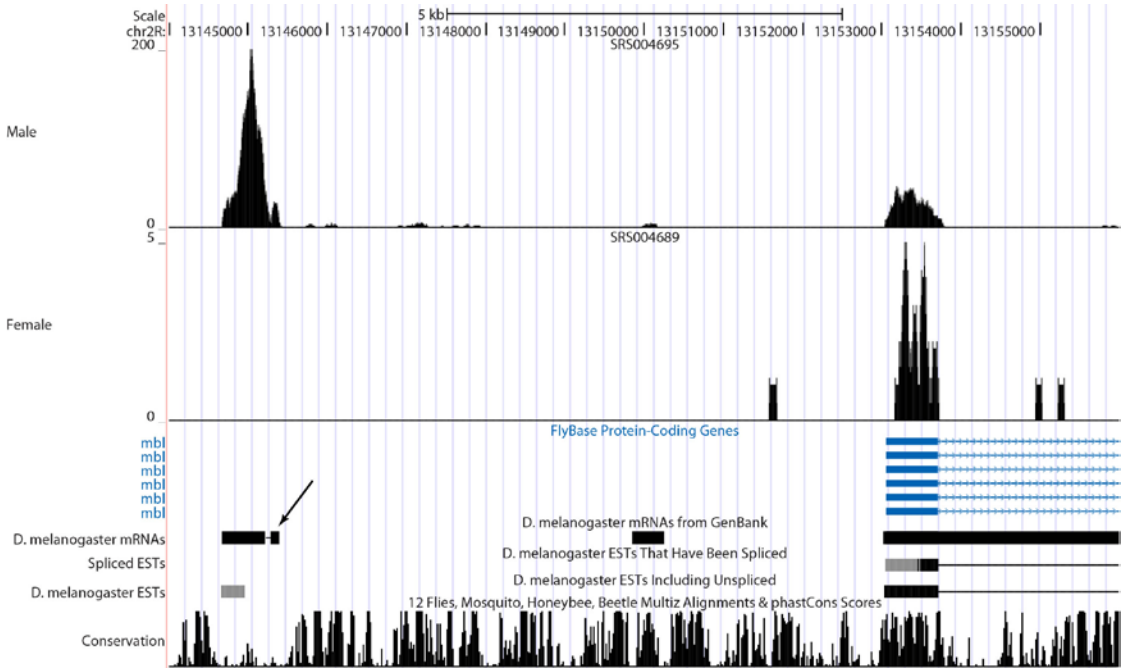

Figure 4.14 Example UCSC genome browser view of a spliced lincRNA locus in the vicinity of the *mbl* (FBgn0261642) protein-coding gene which has read support for expression in one sex, but not the other. The small exon at the right of the lincRNA (indicated by an arrow) is supported by mRNA but not EST evidence.

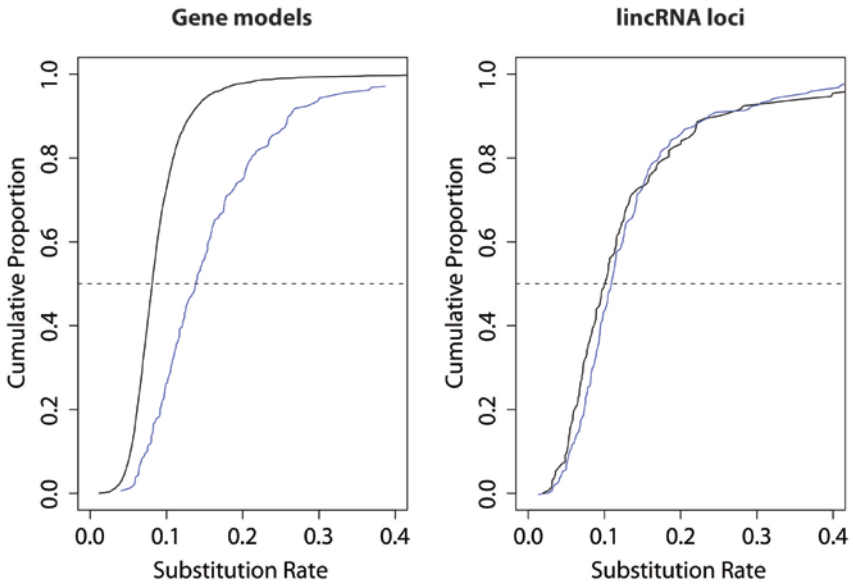

Figure 4.15 Cumulative distributions of the nucleotide substitution rate for gene models (left) and lincRNA loci (right) with different sex-specific expression profiles. Blue – male-specific, solid black – no sex specificity.

4.5.8 Analogous lincRNAs found in mouse

It was observed that over 30% of lincRNAs when aligned to *D. simulans* showed a greater suppression in their substitution rates than gene models, and that they show a greater enrichment in deeply conserved MCS regions. It thus appears that some of them may be as conserved as protein-coding genes. To examine this further, I looked for regions of orthologous lincRNA expression between *Drosophila* and mouse, a very divergent species for which a catalogue of lincRNA expression also exists. Noncoding sequence similarity is not expected to be conserved across such a great evolutionary distance (Woolfe et al. 2004), and so orthologous protein-coding genes were used to define regions of orthologous expression. Protein-coding gene territories were labelled as orthologous in both species if they contained an orthologous protein-coding gene, as defined by the InParanoid database (**Chapter 4.3.1**). The distribution of *Drosophila* lincRNAs within these *Drosophila* territories relative to those territories which also had a mouse lincRNA within them was then determined. I observed 42 *Drosophila* lincRNAs within territories whose orthologous territory also contain a mouse lincRNA – a 56.7% enrichment ($p = 2.4\text{x}10^{-2}$) relative to genome-wide random expectations. A chi-squared test also showed that there was a significant 45.5% increase (two-tailed, $p = 4.5\text{x}10^{-2}$) in the number of orthologous territories containing a lincRNA in both species, relative to those containing a lincRNA in only one. An example of such a shared lincRNA between orthologous protein-coding gene territories is shown in **Figure 4.16**.

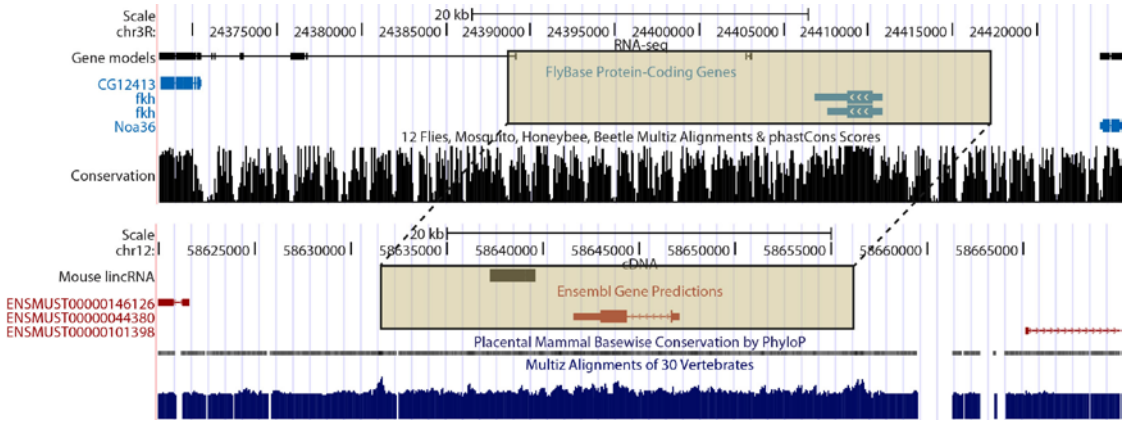

Figure 4.16 An example of analogous lincRNA loci in *D. melanogaster* and *M. musculus*. The boxed genomic regions indicate the orthologous protein-coding gene neighbourhoods for *D. melanogaster* (*fkh*) and *M. musculus* (*Foxa1*). Note that only multi-exonic transcripts are shown for the *D. melanogaster* gene models.

4.6 Discussion

Using these RNA-seq data, and my definition of a gene locus as a collection of overlapping transcripts, I have been able to increase the number of annotated loci in *Drosophila* by a further 15% (from 7,540 to 8,659). LincRNAs are, however, generally shorter than protein-coding genes and so this increase in the number of loci is not accompanied by a corresponding increase in the number of bases which have been annotated; instead I have increased this number by 2% (from 91 Mb to 93 Mb).

These lincRNAs share little overlap with the novel transcripts identified by the modENCODE consortium which produced this RNA-seq dataset, which is likely due to the increased power of my mapping protocol. I identify many novel lincRNA loci, which often appear to be expressed at very low levels and in a restricted developmental range. These do, however, show similar evolutionary constraint to those also identified by modENCODE, suggesting

that these also contain a number of previously unrecognised, functional lincRNA loci.

The lincRNAs identified here also appear to represent a different subset of the *D. melanogaster* lincRNA complement compared to those described in **Chapter 3**. Only 14% of lincRNAs defined by RNA-seq showed any overlap with the set defined using EST evidence. Many ESTs have been isolated from single tissues and it may be that the lincRNAs they correspond to are expressed in particularly narrow ranges, which are missed by the RNA-seq sampling of entire individuals. Alternatively, I noticed in **Table 4.1** that the vast majority of the RNA-seq lincRNAs are spliced, as few single-exonic loci are expressed at a sufficiently high level to distinguish them from genomic DNA contaminants. Due to the incompleteness of EST models, I did not attempt to quantify how many of these lincRNAs were also spliced but it may be that they preferentially detect single-exon lincRNAs, as opposed to the multi-exonic bias observed here for RNA-seq data. Nevertheless, I believe that it is unlikely that the *D. melanogaster* lincRNA catalogue has yet been saturated.

The evolutionary constraint seen in these lincRNAs, like for those defined by EST evidence, is consistent with them being functionally important and having a conserved role across several species. LincRNA loci tolerate fewer indel mutations than gene models, and a similar proportion of substitution mutations, which is far fewer than the rate observed in other intergenic regions. The lincRNA loci studied here are enriched in MCS regions deeply

conserved across the entire *Drosophila* phylogeny, and actually contain a greater frequency of these elements than the gene set. When taken together, these observations support lincRNA conservation, and thus functionality, across all *Drosophilid* species.

My data suggest a major biological role for lincRNAs in transcriptional regulation. This study is the first in which lincRNA expression has been comprehensively followed throughout the life cycle at a genome-wide scale. LincRNAs are more prominently expressed at earlier life stages and are enriched in Polycomb-protein marked heterochromatic domains, which are known to contain developmentally relevant genes. Interestingly, and in contrast to known gene models, the mean substitution rate decreases with increasing specificity in the number of stages in which a lincRNA is expressed. I suggest that the lincRNAs which are expressed in a restricted number of stages are more likely to require conservation of their primary sequence for function and to function in those stages in which they are most prominently expressed. The distribution of lincRNAs across the euchromatic regions of the genome supports this hypothesis, as they are enriched in the domains associated with specific, rather than broad, expression. These evolutionarily constrained, specifically expressed lincRNAs found within regulatory chromatin regions make good candidates for an initial experimental characterisation of this novel gene set.

LincRNAs are also frequently sex-specific, although these sex-specific loci do not mimic the increased substitution rate seen in their protein-coding

counterparts. These lincRNAs therefore do not appear to be involved in sexual selection, but may be important in the development of the male-specific body plan. This is in contrast to previous reports of small numbers of ncRNAs, which have suggested that the expression of such genes was highly divergent even between closely-related *Drosophilid* species (Yang et al. 2007; Jiang et al. 2011). Two previously identified lincRNAs, *roX1* and *roX2*, play a related role in the *Drosophila* male, where they form a complex with several proteins to mediate dosage compensation by causing hypertranscription of the X chromosome (**Chapter 1.2.1**). It is not clear from this work whether the male-specific lincRNAs identified here act in concert with these two genes, or whether they are involved in other sex-specific processes, such as spermatogenesis or epigenetic reprogramming (Amaral and Mattick 2008).

At a much greater evolutionary timescale, I have detected significant conservation of lincRNA expression between *Drosophila* and mouse. I found an increased frequency of *Drosophila* lincRNAs in gene neighbourhoods where the orthologous gene neighbourhood in mouse also contained a lincRNA (no such bias was seen for the EST lincRNAs, data not shown). Previously, there have been only two reports of analogous lincRNA action between these two species (Deng and Meller 2006; Jolly and Lakhotia 2006) but they were not particularly similar in the two species. In both instances, the lincRNAs appear to be involved in chromatin remodelling, whether in dosage compensation or the heat shock response, but they have little else in common. For example, the dosage compensation lincRNAs *Xist* and *roX1*/*roX2* do not even carry out this

compensation in the same direction – *Xist* silences one X chromosome in the female, while *roX1* and *roX2* act to upregulate the single X chromosome in males (see **Chapter 1.2.1**). In contrast, my approach has made it possible to individually pair *Drosophila* and mouse lincRNAs as having a similar location of transcription, although there is no primary sequence conservation. It is therefore not yet possible to resolve whether these pairs do indeed share a common ancestor, and have accumulated many mutations since their last common ancestor, or whether they have unique origins. If the second scenario is correct, it may be that the orthologous protein-coding genes involved may require *cis*-regulation by a neighbouring lincRNA locus, which has led to their independent evolution along different lineages. Due to this ambiguity, I termed these lincRNAs 'analogous' (between these two species). It does seem possible that lincRNAs could be conserved across much greater timescales than previously considered, and I think that future experiments could focus on investigating this possibility simultaneously in both species.

This chapter introduces a novel set of lincRNAs in *D. melanogaster* which is similar in a number of ways to the set defined in **Chapter 3**, yet still possesses several distinct features. Substitution and indel mutations are also suppressed within these lincRNAs, and they contain an excess frequency of MCS regions conserved across the *Drosophila* clade. Unlike the EST set, few of these lincRNAs appear to be the precursors of short RNA species. Many of these lincRNAs may be involved in the regulation of developmental processes, as they make a greater contribution to the transcriptome in early

developmental stages and they are specifically enriched in euchromatic regions showing a regulated expression profile. Their evolutionary constraint with respect to substitutions also appears to increase with increasing specificity of expression. A subset of these lincRNAs may be important in the development of male-specific tissues, such as these testes. Finally, using this dataset I have identified putative analogous lincRNAs shared between *D. melanogaster* and mouse. With these more sophisticated data, it has been possible to propose a number of different functions for subsets of lincRNAs contained within this set which can now be tested experimentally.

Chapter 5: *CIS*-REGULATION OF NEARBY TRANSCRIPTION FACTORS

5.1 Abstract

LincRNAs have been previously implicated in transcriptional regulation in several species (Ponting et al. 2009), but the importance of this has not yet been comprehensively studied in *D. melanogaster*. Using the set of 2,788 lincRNAs defined in **Chapter 3**, I observed that evolutionarily constrained, non-embryonic lincRNAs are enriched in the genomic neighbourhoods of transcription factor genes and therefore may interact with these genes by a *cis*-regulatory mechanism.

An RT-PCR screen of 43 genomically adjacent lincRNA/transcription factor gene pairs showed that only eight are co-expressed and, of these, four lincRNAs actually represent unannotated exons or UTRs of the neighbouring transcription factor. I studied one of the remaining four lincRNAs in more detail and demonstrated that it is an independent lincRNA capable of positively regulating expression of its adjacent transcription factor gene, *Dll*, in S2 cells. It was not possible to detect the lincRNA *in vivo*, perhaps due to its low expression level or cell specificity. The lincRNA is named *dEvf-2* due to its experimental similarities to the mouse lincRNA *Evf-2*, which appears to regulate the *Dll* orthologues *Dlx5* and *Dlx6*, possibly in a similar manner. The two lincRNAs may indeed even share a common evolutionary origin.

Although *cis*-regulation of genes involved in transcriptional regulation does not appear to explain the functionality of a large proportion of lincRNAs in *D. melanogaster,* it may be one general function of lincRNAs across distantly related species.

5.2 Introduction

Once lincRNAs have been shown to contain a number of predictors of functionality (as described in **Chapters 3** and **4**), it next becomes interesting to determine the mechanisms by which they might function. One of the most frequently cited functions of lincRNAs is regulation of protein-coding gene expression, which can take place at a number of different levels in the cell (Prasanth and Spector 2007). Here I investigate the possibility of lincRNA regulation of genomically encoded adjacent protein-coding genes, which I will call *cis*-regulation. It has been previously reported that many *cis*-regulatory elements, such as promoters and enhancers, are transcribed (Amaral and Mattick 2008). It is not yet clear how many of these require transcription to function, although several examples of well-studied lincRNAs which regulate the transcription of nearby genes on the chromosome have been identified (**Chapter 1.2.2**).

Here, I investigate the potential for members of the *D. melanogaster* lincRNA set defined in **Chapter 3** to regulate the expression of their genomically adjacent protein-coding gene. Evolutionarily constrained, non-embryonic lincRNAs are found to be preferentially transcribed from within the

neighbourhoods of genes encoding transcription factors, but only four of these are transcribed independently from their neighbouring transcription factor and share their expression profile, that of being ubiquitously found in all tissues examined here. One of these, which I call *dEvf-2*, is shown to positively regulate the mRNA level of the adjacent transcription factor gene, *Dll*, in S2 cells. This lincRNA appears to act in an analogous manner to *Evf-2* in mice, which positively regulates the *Dll* orthologous genes *Dlx5* and *Dlx6*. While *cis*-regulation may not explain the majority of the functionality observed in these loci, it is one interesting hypothesis under which their behaviour can be dissected experimentally in more detail.

5.3 Materials

5.3.1 Transfrags

Microarray technology can be used to assay DNA samples on a genome-wide scale. A microarray slide consists of many thousands of short DNA oligonucleotide probes which are complementary to the DNA sequences (or cDNA sequences when looking at gene expression) of interest. The sample is fluorescently labelled and allowed to hybridise to the slide, before sequences which have bound non-specifically are washed off. The fluorescence at each probe location is then related to the quantity of the complementary sequence in the original sample. The ability to carry out many tests in parallel has led to this technology being used in a variety of applications, such as high-

throughput genotyping of individuals and investigating gene expression differences in disease (reviewed in Heller 2002).

Transcription from the *D. melanogaster* genome has been assayed using tiling microarray slides, which cover all non-repeat regions of the genome (Manak et al. 2006). Regions of transcription known as transfrags are identified as intervals of continuous transcription. Genome-wide transcription during embryogenesis was assayed using this approach (Manak et al. 2006). Total RNA $>$ 200 bp was extracted at 2 hour intervals during the first 24 hours of development, and the numbers of transfrags called by Manak et al. at each interval are shown in **Table 5.1**. This information is used to annotate lincRNAs as being 'embryonic' or 'non-embryonic' in **Chapter 5.5.1**.

5.3.2 Gene Ontology (GO) terms

Genes and the products they encode can be described, or annotated, using a controlled vocabulary known as an ontology. Several of these have been produced by the GO consortium (Ashburner et al. 2000) which are both human- and computer-readable, and provide a consistent, structured and species-independent description of the many thousands of genes deposited in various public databases. The GO consortium has created three distinct ontologies which describe the molecular function (usually a biochemical activity, such as 'enzyme'); the biological process; and the cellular component (i.e. location) of genes.

The evidence for gene annotations can come from several sources. These sources can be classed as experimental, computational, indirect (inferred from one of the first two) or unknown. In 2007, over 95% of the gene annotations had been computationally derived (Rhee et al. 2008), and were expected to show a higher rate of false positive annotations than those which had manually curated. 9,563 (67.7%) of *D. melanogaster* genes had been annotated as of 2008, and 2,790 (29.2%) of those annotated possessed at least one experimental annotation. 246 genes contained a 'NOT' annotation, to indicate the lack of a specific property (Rhee et al. 2008).

GO databases are updated regularly and frequently as the ontologies and the annotations are improved. In this chapter, I make use of the 8th March 2008 release for *D. melanogaster* to identify genes with particular GO terms whose territories (**Chapter 2.2.4**) are enriched with constrained, non-embryonic lincRNAs.

5.4 Methods

5.4.1 LiftOver

As genome assemblies are improved and new versions are released, the coordinate numbering system between them can become inconsistent, such as when gaps are filled in. Coordinates and annotations can be converted between different assemblies of the same genome using the LiftOver tool from UCSC, which uses the CHAIN data created as part of the BLASTZ alignment

between the two genome assemblies (**Chapter 2.1.3**). It can also be used to convert between genomes of different species and therefore identify orthologous coordinates but this is not recommended, particularly for distantly related species. Here, I only use LiftOver to adapt transfrag coordinates generated using the BDGP4 assembly of *D. melanogaster* for use with the current BDGP5 assembly.

5.4.2 Rapid Amplification of cDNA Ends (RACE)

Full-length cDNA sequences can be cloned using the RACE protocol described below. This is required to define transcript boundaries and can be useful when analysing EST sequences as these frequently do not extend to the full length of the transcript. The RLM-RACE (RNA ligase-mediated RACE) kit (Ambion) is used here to identify full-length RNA molecules from total RNA extractions from mixed-sex adult flies created as described in **Chapter 2.2.6.1**. Separate libraries are made for identifying the 5' and 3' ends of transcripts and these libraries are used to amplify the ends of a specific transcript using nested PCR. Relevant PCR products are purified and cloned, and the resulting plasmids sequenced.

5.4.2.1 **5' RACE Library Generation**

A 5' RACE library is created as described in the manufacturer's instructions and using 10 µg total RNA. Free 5' phosphates are removed from ribosomal RNA (rRNA), genomic DNA, tRNA and fragmented mRNA molecules by treatment with calf intestine alkaline phosphatase (CAP). The 5' cap is then

removed from full-length mRNAs by the addition of tobacco acid pyrophosphatase (TAP) to leave a 5'-monophosphate which is ligated to a 45 base RNA oligonucleotide by T4 RNA ligase. Random primers are then used to direct reverse transcription of these tagged RNA molecules.

5.4.2.2 **3' RACE Library Generation**

The manufacturer's instructions were also followed to create a 3' RACE library using RNA extracted from the same tissue. cDNA is synthesised using 1 µg of total RNA and the 3' RACE adaptor included with the kit. This adaptor contains a stretch of thymine bases which match the polyA tail at the end of a mature RNA molecule and a unique sequence specific to the 3' RACE primers (see **Figure 5.1**) for later PCR amplification steps.

5.4.2.3 **Nested PCR and PCR purification**

Those cDNA molecules in either the 5' or 3' libraries which correspond to the sequences of interest are purified using the nested PCR protocol summarised in **Figure 5.1**. Outer primers are used for the first PCR reaction and the product of this (which can often not be visualised on an agarose gel) is used as the template for the second PCR reaction using the inner primers. PCR reactions are set up as in the manufacturer's instructions, using the expand high fidelity polymerase blend (Roche) for 5' *dEvf-2* RACE and *Taq* DNA polymerase (Bioline) for all other experiments.

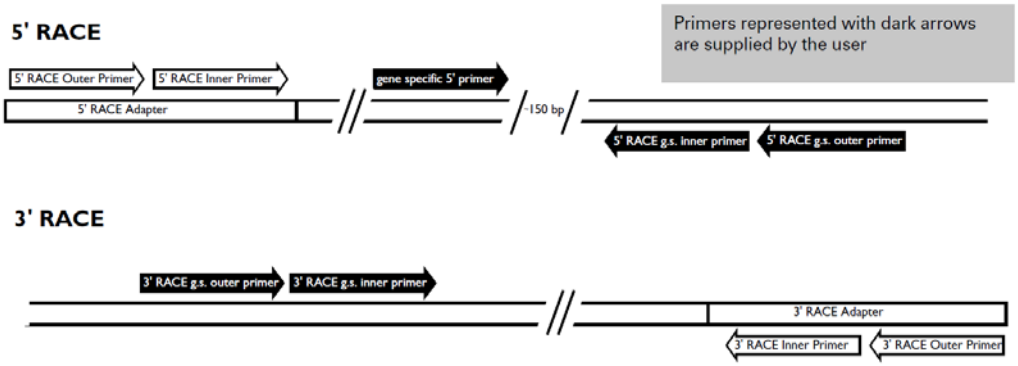

Figure 5.1 Primer positions for 5' and 3' RACE. Taken from RLM-RACE manual.

The products of the second round of PCR are purified prior to cloning (Qiaquick PCR purification kit, Qiagen). The manufacturer's instructions were followed where the DNA concentration was increased by eluting the DNA in 30 µl elution buffer and allowing the column to stand for 1 min before the final centrifugation step.

5.4.2.4 **TOPO TA Cloning**

The purified PCR products produced above are cloned into the pCR 2.1-TOPO vector (Invitrogen), by incubating the maximum volume of 4 µl with the vector and salt solution (1.2 M NaCl and 0.06 M $MgCl_2$) for 5 mins at room temperature. The 3' adenosine added to the end of Taq-produced PCR products is complementary to the overhanging 3' thymidine residue in the linearised vector, which allows these products to be efficiently inserted into the vector. The cloned plasmids are transformed into chemically competent DH5α-T1R cells (Invitrogen 18263-012) by mixing the cells with the plasmid and placing them on ice for ~30 mins. The cells are then heat-shocked at 42°C for

30 s (without shaking), before adding 250 µl of S.O.C. medium (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose) and shaking the cells at 200 rpm and 37°C for 1 hour. 40 µl of 40 mg/ml bromo-chloro-indolyl-galactopyranoside (X-gal) in dimethylformamide is spread onto plates containing 50 µg/ml ampicillin in Luria-Bertani medium (made as described by Invitrogen). 50 µl and 200 µl aliquots of each transformation are spread onto each plate and cultured overnight at 37°C where surviving colonies contain the ampicillin resistant vector. Those that are also white are unable to metabolise the X-gal because they have disrupted lacZ expression, due to insertion of the original PCR product in the vector. White colonies are picked and cultured overnight in 10 ml Luria-Bertani medium and 100 µg/ml ampicillin at 200 rpm and 37°C.

5.4.2.5 **Extraction of plasmid DNA**

Once selected and cultured, plasmids are purified using the Qiaprep Spin Miniprep kit (Qiagen). Briefly, cultured cells are pelleted by centrifugation (8,000 rpm for 3 mins) and then resuspended in buffer P1 containing RNase A. The P2 buffer then lyses the cells and denatures DNA. Fast renaturation with the N3 buffer leads to genomic, but not plasmid, DNA aggregation. The genomic DNA can then be separated from the plasmid DNA by centrifugation, where the genomic DNA and some other cell debris form a pellet. The plasmid-containing supernatant is added to a spin column to capture the plasmid. After washing, the plasmid DNA is then eluted.

Insertions within the purified plasmid DNA are tested by restriction digestion. Approximately 1 µg plasmid DNA is incubated with 20 units *Eco* RI (New England Biolabs) and 1 x NEBuffer *Eco* RI (New England Biolabs) in a 10 µl reaction volume for at least 1 hour at 37°C. The reaction mix is analysed using gel electrophoresis (**Chapter 2.2.6.3**). If the PCR product has inserted correctly into the plasmid, a band is expected at approximately the same size as the original product. Those plasmids which contained an insert of the expected size were sequenced from the M13 primer positions that flank the insert.

5.4.3 *Drosophila* Schneider 2 (S2) cells

Drosophila S2-DRSC (*Drosophila* RNAi Screening Centre) cells are cultured in M3+BPYE medium. This is made by adding 0.5 g yeast extract and 1.25 g bacteriological peptone to 500 ml Shields and Sang M3 media by sterile filtration before also adding 50 ml heat-inactivated fetal bovine serum (Invitrogen 10082147).

5.4.4 RNAi

The expression of a particular gene can be almost completely abolished by a technique known as RNA interference (RNAi) to try to reveal the normal function of that gene's expression (Alberts et al. 2002). This effect is mediated by short RNA molecules which bind to the complementary sequence in the target mRNA and cause it to be degraded or, when bound nearer the 3' end of the mRNA, prevent translation of the mRNA (Hannon 2002). These short

molecules can be generated by processing of exogenous long double-stranded RNA (dsRNA) molecules within the cell. In this section, I describe how I introduce such molecules to *Drosophila* S2 cells to investigate the effect of abolishing lincRNA expression.

dsRNA is produced through *in vitro* transcription of a DNA template. These templates are generated by PCR amplification of S2 cDNA. Primers are designed using the SnapDragon tool (http://www.flyrnai.org/cgi-bin/RNAi_find_primers.pl) to reduce off-target effects of the short RNA molecules produced by processing of the long dsRNA. All parameters are left at their defaults, and primer pairs which minimised the pair penalty while still producing a product > 300 bp and containing no off-target regions ≥ 19 bp are selected. The T3 promoter sequence is added to the 5' of each primer to allow the resulting PCR products to be transcribed by T3 RNA polymerase. PCR is carried out as in **Chapter 2.2.6**, using 50 µl reactions, rather than 25 µl. Approximately 70 ng of cDNA is added to each reaction.

In vitro transcription is then carried out using 30 µl of the resulting PCR reaction product. The reaction mix contains 1 mM C/A/GTP nucleotides, 0.8 mM UTP, 1 x transcription buffer (Agilent), 5 µM T3 primer, 2.5 units T3 polymerase (Agilent), and is made up to a final volume of 120 µl with RNase-free H_20. The reaction is incubated for 2 hours at 37°C. After this, 5 volumes (~600 µl) of RNase-free H_20 are added and transcription is terminated by incubation at 95°C for 5 mins. The reaction mix is allowed to cool slowly to room temperature to allow the dsRNA to anneal. The dsRNA is precipitated

by adding 2 volumes (1440 µl) of 100% ethanol and centrifuging it for 5 mins at 7,500 rpm and 4°C. The resulting pellet is washed twice in 70% ethanol and allowed to air-dry. Finally, the pellet is resuspended in RNase-free H_20.

For each transfection, 25 µg of dsRNA is incubated with 12.5 µl Fugene HD (Roche) in 1.125 ml of RNase-free H_20 for 15 mins. This mix is then added to ~$1x10^7$ S2 cells in 10 ml of medium, where the cells soak up the dsRNA. After 3-4 days, RNA is extracted from all cells in the 10 ml sample. This protocol for *in vitro* transcription and transfection was adapted from Björklund et al. 2006.

5.4.5 Imaging

The expression of *Dll* and its neighbouring lincRNA *dEvf-2* in the eye-antennal disc of third instar larvae is investigated using a combination of immunohistochemistry and fluorescent *in situ* hybridisation (FISH). Discs are dissected at room temperature in Grace's Insect Medium (Invitrogen) and fixed for 10 mins in 4% paraformaldehyde in phosphate buffered saline solution (PBS, 135 mM NaCl, 2.5 mM KCl, 4.3 mM Na_2HPO_4, 1.5 mM KH_2PO_4, pH 7.2). They are then washed twice in PBS and left at room temperature for 30 mins. The protocol then diverges, depending on whether the Dll protein or the mRNA transcripts of *Dll* and *dEvf-2* are being investigated. Discs are imaged on a Zeiss LSM 510 META confocal microscope.

5.4.5.1 **Immunohistochemistry**

Discs are washed for 30 mins in PBT (1 x PBS, 0.5% horse serum, 0.3% Triton X-100). Washed discs are incubated with the primary Dll antibody (Duncan et al. 1998) overnight at room temperature at a concentration of 1:300 in PBT. Discs are washed again for 30 mins in PBT before being incubated overnight at room temperature with the secondary antibody (Alexa fluor-488 conjugated donkey anti-mouse) diluted 1:1000 in PBT and 1 µg/ml Hoechst 33342 (Molecular Probes). The discs are then washed one more time for 30 mins in PBS before being placed in mounting solution (80% glycerol, 20% PBS) for microscopy.

5.4.5.2 **FISH**

FISH probes are generated by *in vitro* transcription in a similar manner to that described in **Chapter 5.4.4**. Synthetic oligonucleotides containing the DNA sequence encoding the transcribed sequence with the T3 promoter sequence added to the 3' end are used as the template instead of a PCR product and a 20 µl reaction mix is used for *in vitro* transcription. 0.7 mM UTP is added along with Alexa fluor-488-UTP or Alexa fluor-Cy5-UTP (both at 0.1 mM). After the 2 hour incubation at 37°C, the probe is purified through a hydrated Centri Spin-10 column (Princeton Separations) by centrifugation at 3,000 rpm for 2 mins. A 1:10 stock is made by adding 180 µl of *in situ* mix (ISM, 50% formamide, 5 x SSC [1 x SSC is 0.15 M NaCl, 0.15 M Na citrate, pH 7], 10 mM citric acid, 50 µg/ml heparin, 500 µg/ml yeast tRNA, 0.1%

tween-20 and 1 µg/ml 4',6-diamidino-2-phenylindole (DAPI)). Discs are incubated for 1 hour at 42°C at a final probe concentration of 1:250 and 1:500 for *Dll* or *dEvf-2*, respectively. They are mounted onto microscopy slides directly in ISM.

The probes used in the example images shown here are described in **Appendix C**.

5.5 Results

5.5.1 LincRNA expression can be defined as embryonic or non-embryonic

Sets of transcribed regions known as transfrags were defined at 2 hour intervals for the first 24 hours of *D. melanogaster* embryogenesis by the microarray experiment described in **Chapter 5.3.1**. The numbers of regions at each stage are summarised in **Table 5.1**. I merged these into a group of consensus transfrags which consists of transfrags at different intervals joined by at least 1 bp overlap, in a similar manner to the clustering of EST sequences in **Chapter 3**. 81,300 transfrag clusters which cover 29 Mb (24.4%) of the genome were created. The coordinates of these were converted from the BDGP4 assembly, for which they were defined, to the current BDGP5 assembly using the LiftOver tool (**Chapter 5.4.1**). I classified the gene models defined in **Chapter 3** as being 'embryonic' if they had at least 1 bp overlap with a transfrag cluster. All other models were classed as 'non-embryonic'.

9,088 (77.2%) of protein-coding genes and 836 (30.0%) of lincRNA loci are embryonic, and this difference in coverage is highly significant (two-tailed chi-squared test, $p < 2.2\text{x}10^{-16}$). This may be because a smaller proportion of lincRNAs are genuinely expressed in the embryo, although this was not seen in the RNA-seq data in **Chapter 4**, or because microarrays are unable to detect the generally more lowly expressed lincRNAs. Those which are most lowly expressed in the embryo could therefore be misclassified as being non-embryonic.

Time period (hours)	**Numbers of transfrags**
0-2	31,082
2-4	34,451
4-6	30,516
6-8	43,442
8-10	46,541
10-12	52,161
12-14	41,509
14-16	45,164
16-18	48,607
18-20	37,662
20-22	39,496
22-24	42,739

Table 5.1 Numbers of transfrags mapped during embryogenesis. Data from Manak et al. 2006.

5.5.2 Constrained, non-embryonic lincRNAs tend to be encoded near to transcription factors

The distribution of constrained, non-embryonic lincRNAs across different protein-coding gene territories was investigated using the Annotator (**Chapter 2.2.4**). 808 constrained, non-embryonic lincRNAs were selected for this analysis. This was because previous work (**Chapter 3**) suggested that

constrained lincRNAs were among the most likely to be functional. Non-embryonic lincRNAs were selected because correlations between lincRNA loci and neighbouring protein-coding genes can only be deduced using data from non-embryonic tissues.

The distribution of these lincRNAs within protein-coding gene territories annotated with particular GO terms was tested as described in **Chapter 2.2.4**. A correction was made for multiple testing by selecting a *p*-value cut-off of 0.01. This maintained the average and median numbers of expected false positive annotations below one for all three ontologies tested. No terms within the 'cellular component' ontology were deemed significant, while 'regulation of transcription, DNA-dependent' was significant in the 'biological process' ontology, and 'transcription factor activity' and 'RNA polymerase II transcription factor activity' were significant within the 'molecular function' ontology. The magnitudes of these enrichments are shown in **Table 5.2**. These do not represent three independent enrichments as 'transcription factor activity' and 'RNA polymerase II transcription factor activity' are both children of the parent term 'transcription regulator activity', which is not significant, and many genes are annotated with more than one of these terms.

Gene Ontology Term	**Enrichment (%)**	***p*-value**
Regulation of transcription, DNA-dependent	37.1	8.9×10^{-3}
Transcription factor activity	47.6	2.0×10^{-4}
RNA polymerase II transcription factor activity	75.5	4.8×10^{-3}

Table 5.2 Significant enrichments of constrained non-embryonic lincRNAs within the territories of genes with particular GO terms. Expected number of false positive terms = 0.5.

As these lincRNA loci are found near to this protein-coding gene class more frequently than expected, I considered adjacent lincRNA-transcription factor pairs which could be potentially functionally related. LincRNA loci were discarded if they were found partially in the territories of genes with these GO terms but if there was less intergenic distance between the locus endpoints and the other protein-coding gene whose territory they overlap. There are 43 constrained, non-embryonic lincRNAs in my set whose nearest protein-coding gene either upstream or downstream is annotated with at least one of the terms shown in **Table 5.2**. All 43 of these pairs are described in **Appendix B** and were investigated experimentally for a possible functional interaction.

5.5.3 LincRNA/transcription factor expression across the life cycle

I first tested each of the 43 lincRNA-transcription factor pairs for co-expression by RT-PCR (**Chapter 2.2.6**) at several developmental stages throughout the *Drosophila* life cycle. The stages selected were larva, early pupa, late pupa, adult body and head (both adult stages were mixed sex). The embryonic stages were not included as these lincRNAs were selected as being specifically non-embryonic in **Chapter 5.5.1**. Adult heads were separated from the bodies to identify head-specific lincRNAs, which could be important in the central nervous system (CNS). The other samples cover a diverse range of developmental time points. It was hoped that, if the expression of any of these lincRNAs was differentially regulated, this could be detected by the

presence of the lincRNA transcript in some tissues and its absence in others. A related expression profile between a lincRNA and its adjacent transcription factor (whether co-expressed or mutually exclusive) would be one line of evidence suggesting a regulatory interaction between them.

A typical RT-PCR result is shown in **Figure 5.2**. This transcription factor *croc* (FBgn0014143) is ubiquitously expressed, while the expression of the lincRNA BI621778 is head-specific. In this example, there is no correlated expression between these transcripts.

Each lincRNA/transcription factor pair was tested once in this way, and the results of all 43 experiments are summarised in **Table 5.3**. Expression of 31 (72%) of these lincRNAs were verified and for 20 of these a regulated expression pattern was shown: expression was apparent in some tissues but not in others. Four of the 20, such as BI621778 (**Figure 5.2**), are expressed only in the head. Two others are expressed in the adult head and other stages, but not the adult body, which suggests they could be specifically expressed in the central nervous system throughout several developmental stages.

Relationship	**Number of lincRNA/TF pairs**
Ubiquitous expression	8
Head-specific lincRNA	4
Other regulated, but not co-expressed, lincRNA	16
Ubiquitously-expressed lincRNA	3
lincRNA not expressed	12

Table 5.3 Results of an RT-PCR screen showing the relationship between the expression of 43 genomically adjacent transcription factor and lincRNA locus pairs.

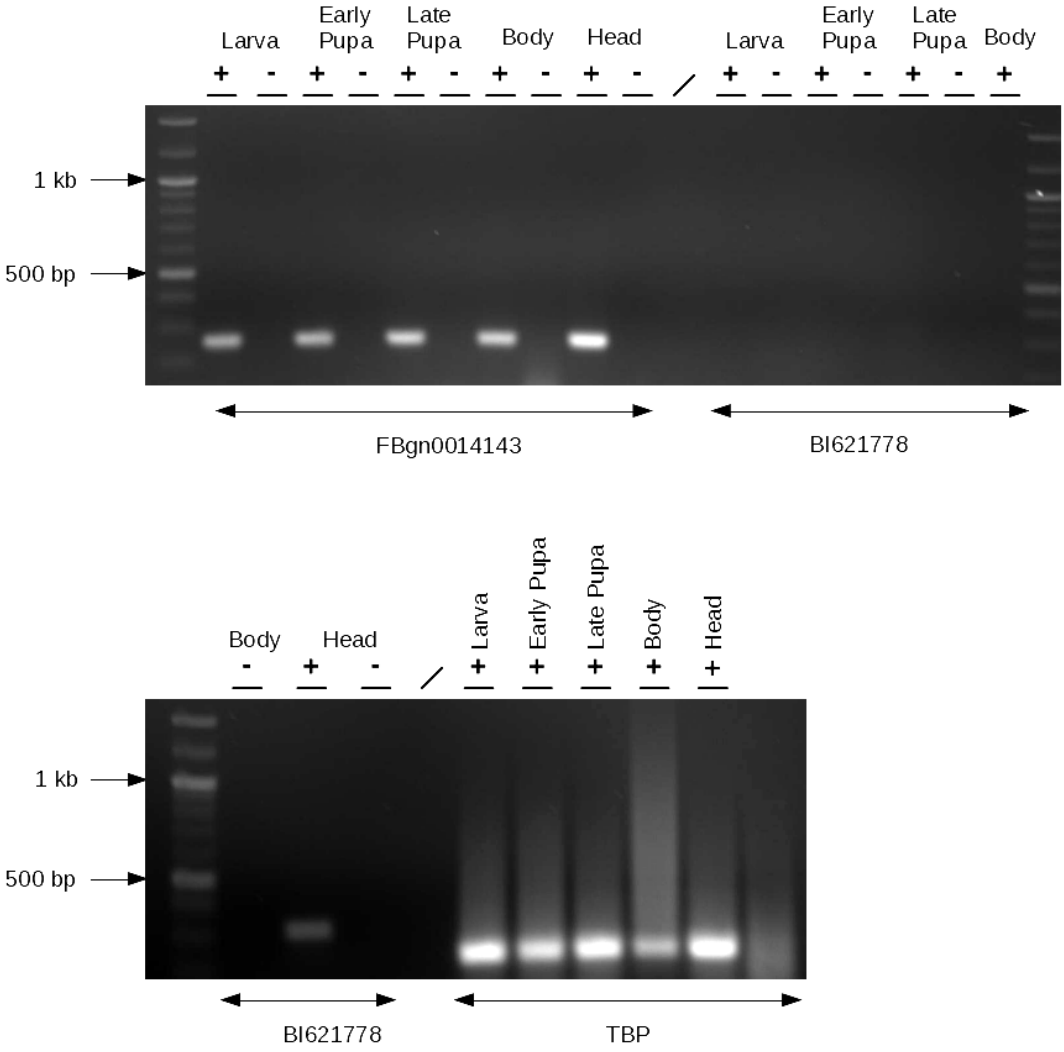

Figure 5.2 Example RT-PCR gel showing expression of a transcription factor (*croc*, FBgn0014143) and lincRNA (BI621778) during the life cycle. Expected product sizes: FBgn0014143 – 263 bp, BI621778 – 337 bp, TBP – 239 bp. Positive (+) lanes and negative (-) reactions for each PCR reaction are shown.

No regulated lincRNAs shared an identical or mutually exclusive differential expression profile with their adjacent transcription factor; hence it is unlikely that any of these share a functional relationship. Eight pairs exhibited ubiquitous expression in all five samples, and for two replicates, and these were investigated in subsequent experiments.

5.5.4 Four of eight candidate lincRNAs are UTRs of the neighbouring transcription factor.

The lincRNA models defined by ESTs are unlikely to represent full-length transcripts. LincRNA loci which are constrained, adjacent to transcription factors, and display the same expression pattern as the transcription factor, may represent unannotated exons or UTRs of that transcription factor. The ends of a transcript can be defined precisely using RACE (**Chapter 5.4.2**) but this is a demanding and relatively time-consuming protocol. To reduce the number of such experiments required, I first used RT-PCR to test whether I could amplify across one transcript joining the lincRNA to the transcription factor. I used the same primers as for the RT-PCR screen (described in **Chapter 5.5.3**) and combined one primer for the lincRNA locus and one for the neighbouring transcription factor in the same reaction. The cDNA template used was created from a mixed-sex, whole-adult RNA extraction.

Four of my eight candidates – BI61123, AI541814, EC088598, and CO309004 – produced a positive result (**Figure 5.3**), implying that these lincRNA models are unannotated 5' (BI611233) and 3' (AI541814, EC088598, and CO309004) regions of their neighbouring transcription factor genes; they were thus discarded.

The literature describing these remaining four transcription factors was briefly reviewed and one of these, *Distal-less* (*Dll,* FBgn0000157), was noted to be orthologous to two mouse transcription factors (*Dlx5* and *Dlx*6) which are

regulated in *cis* by a lincRNA, termed *Evf-2* (Feng et al. 2006). This is described in more detail in **Chapter 5.6**. Due to its potential similarity with the mouse lincRNA, it was decided to focus solely on *Dll* and its neighbouring lincRNA, AI945277, in future experiments. I renamed AI945277 as *dEvf-2* and will refer to it by this name throughout the rest of this thesis.

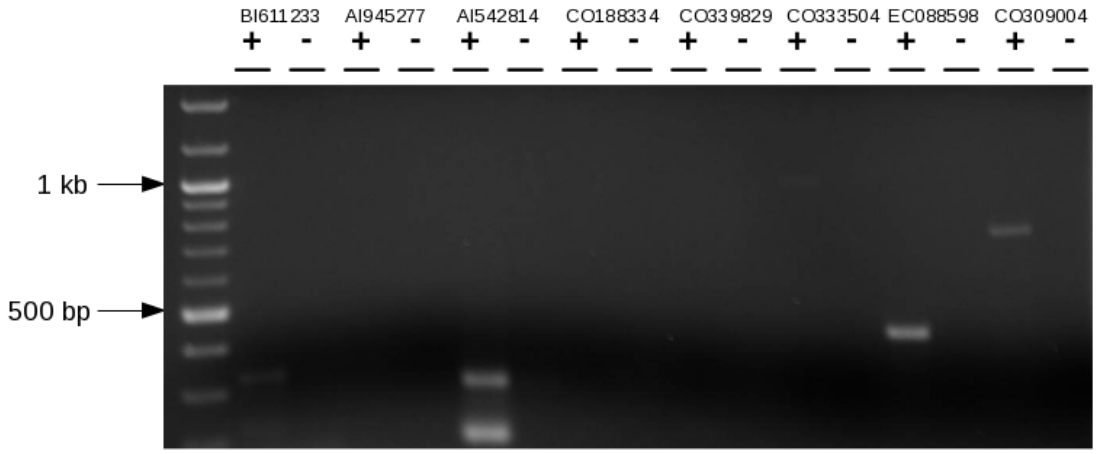

Figure 5.3 RT-PCR gel which confirms that four remaining lincRNAs (BI611233, AI541814, EC088598 and CO309004) represent previously unknown exons of a neighbouring transcription factor transcript. Positive (+) lanes and negative (-) reactions for each PCR reaction are shown.

5.5.5 *dEvf-2* encodes two independent, single-exonic transcripts.

The gene structure of *dEvf-2* was characterised using 5' and 3' RACE, which also conclusively tested if it is produced by a transcript independent of *Dll*.

A single 5' PCR product of ~400 bp was observed after the second round of PCR using *Dll*-specific primers (**Figure 5.4**). After cloning and sequencing, two sequences were mapped by BLAT to the 5' region of *Dll*. Both extended from the genomic coordinates chr2R:20702344-20702709, resulting in a mapped

sequence of 365 bp. This confirms that the *Dll* gene model is broadly correct (**Figure 5.5**), although the TSS appears to begin 9 bp upstream of the currently annotated position (chr2R: 20702353).

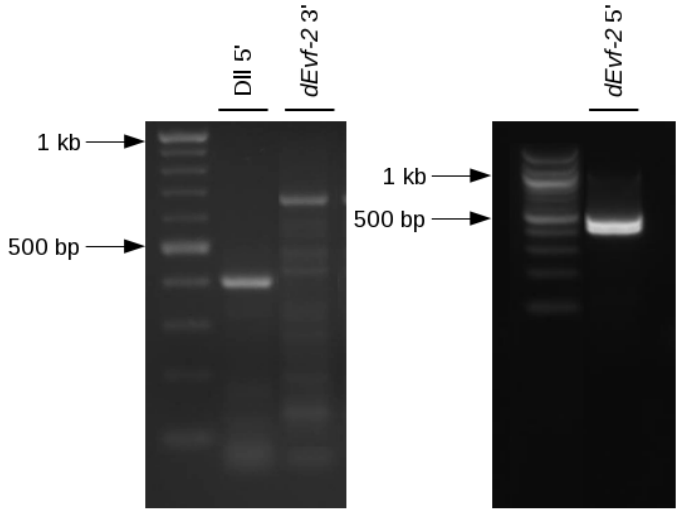

Figure 5.4 PCR products produced after amplification using inner primers for *Dll* (5'), *dEvf-2* (3' and 5'). RT-negative controls are not shown.

The *dEvf-2* TSS was similarly studied by Charlotte Tibbit where a single ~400 bp band was observed after the second round of PCR (**Figure 5.4**). However, this wide band obscured the presence of two different size products which were detected upon sequence analysis of the resulting clones. Five sequences were found to map to the *dEvf-2* locus – one (CT_43) extended from chr2R:20696847-20697249 (432 bp) while the other four mapped to chr2R:20696925-20697270 (345 bp). *dEvf-2* therefore contains two TSSs, 78 bp apart, at bases 20696847 and 20696925, as shown in **Figure 5.5**. The 5' most TSS is 24 bp upstream of the 5' end of the EST AI945277, while the other TSS is 54 bp downstream. Subsequent RT-PCR experiments using primers

spanning the 5' most TSS of *dEvf*-2 confirmed that transcription does not extend beyond this site (data not shown).

The 3' end of the *dEvf-2* transcript was defined using 3' RACE. After two rounds of PCR, a clear product was detected at ~700 bp, while several smaller, but weaker, products could also be seen. Only three cloned sequences mapped to the *dEvf-2* locus (**Figure 5.5**), and all three corresponded to the largest 700 bp band. It is therefore considered likely that the smaller, weaker products were generated by non-specific amplification. The three cloned sequences extended from chr2R:20697000-20697632 (632 bp). There was slight variation of ~10 bp in the end point of the sequences, but all overlapped the genomically encoded polyadenylation signal of 5'-AAUAAA-3' which begins at chr2R: 20697611. This polyadenylation signal appears genuine as the three sequences also contained a non-genomic polyA tail of ~13 adenosine bases downstream of this signal.

Despite a recent FlyBase annotation (CG42851, **Figure 5.5**) suggesting that *dEvf-2* may contain a small intron, I found no evidence to support this. All 5' RACE sequences for *dEvf-2* span the putative intron location. No gaps were observed in their alignments to the *D. melanogaster* genome, which would suggest that they had been spliced. There also does not appear to be an intron elsewhere in the transcript. An RT-PCR reaction using primers spanning the 5' and 3' RACE sequences produced a single product of the expected size (Charlotte Tibbit, personal communication), with no variants implying splice events.

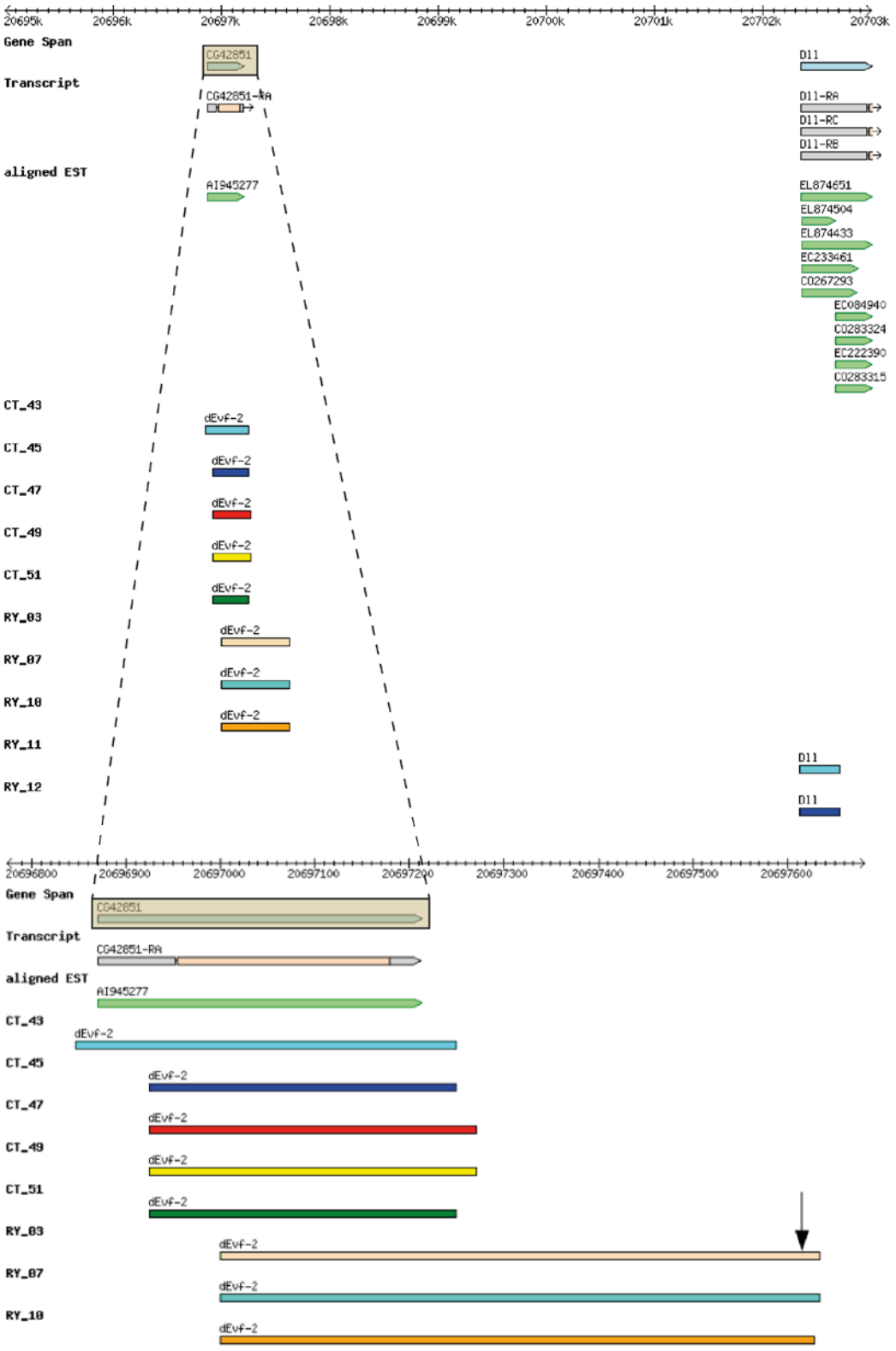

Figure 5.5 RACE sequences describing the 5' of *Dll* and the complete *dEvf-2* structure, annotated as CG42851-RA by FlyBase. The arrow in the zoomed image indicates the site of the *dEvf-2* polyadenylation signal.

5.5.6 *dEvf-2* positively regulates *Dll* expression in S2 cells.

dEvf-2 is a constrained, independently transcribed lincRNA which appears to be co-expressed with its adjacent transcription factor gene *Dll*. I was next interested in the functional consequence, if any, of this relationship. I tested whether *dEvf-2* was able to regulate *Dll* transcription, by reducing *dEvf-2* expression in S2 cell culture.

I first verified by RT-PCR that both these genes were expressed in these cells. Four biological replicates of S2 cells were then soaked in dsRNA targeting both *dEvf-2* and an off-target control for three days, and the expression of *dEvf-2*, *Dll*, CG3611 and *Gapdh* were assayed by real time RT-PCR. CG3611 is a gene that is upstream of *dEvf-2* and was used as a control to determine whether the RNAi treatment had an effect on transcription across the entire chromosomal region. The results are summarised in **Figure 5.6**.

The 82% and 98% knockdowns observed for *dEvf-2* and *Dll*, respectively are significant (Mann-Whitney, $p = 3.4\text{x}10^{-2}$ and $1.7\text{x}10^{-5}$, respectively). CG3611 is unaffected by the RNAi treatment (Mann-Whitney, $p = 0.54$). The *Dll* knockdown is associated with a lower p-value than *dEvf-2*, partly because it has a much reduced standard error due to it being expressed at a higher level in S2 cells.

This experiment shows not only that these two genes share a positive functional relationship but also that it is the mature *dEvf-2* transcript which mediates this function. The RNAi treatment does not affect lincRNA

transcription. Consequently, if it was the act of transcription which influenced *Dll* expression, this effect would not be observed in this experiment. Instead, these data show that the mature *dEvf-2* lincRNA is able to positively regulate the transcription of the *Dll* mRNA in S2 cells.

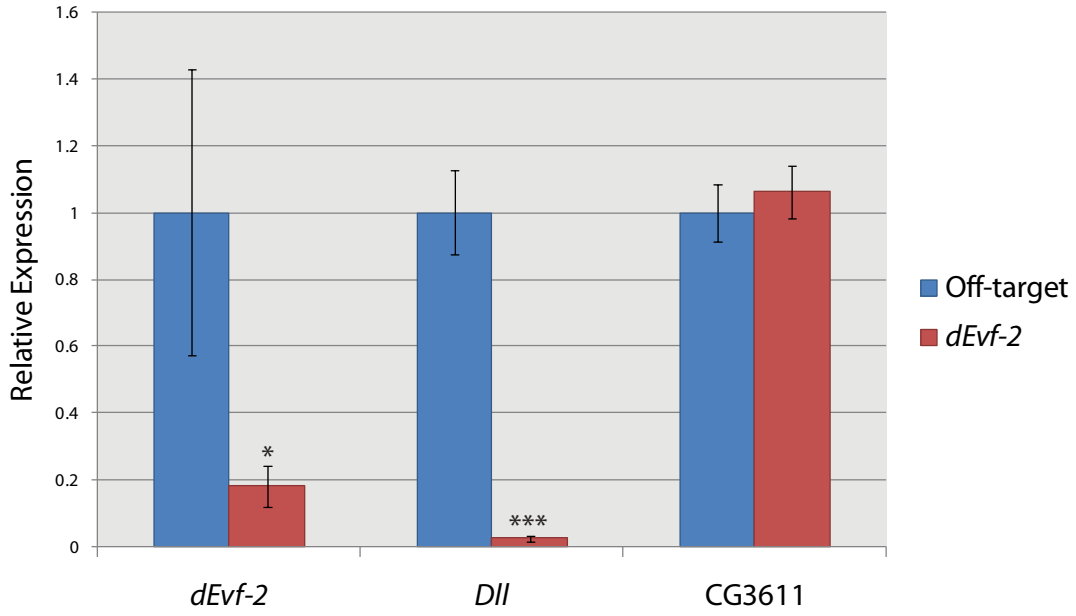

Figure 5.6 Significant and specific knockdown of *dEvf-2* and *Dll* upon four biological replicates of RNAi treatment against *dEvf-2*. Error bars represent the standard error of the mean (* indicates $p < 0.05$ and *** indicates $p < 0.001$).

I also attempted to test the reciprocal relationship and determine whether RNAi treatment against the *Dll* transcript affected *dEvf-2* transcription. Despite using two different dsRNA molecules, I was unable to detect any significant *Dll* knockdown and therefore I could not determine the effect, if any, that this might have on *dEvf-2*.

5.5.7 *Dll* mRNA and *dEvf*-2 lincRNA were not detectable in the larval antennal disc.

That RNAi against *dEvf-2* results in a decrease in *Dll* expression in S2 cells does not prove that the same effect occurs *in vivo*. For the lincRNA to be able to stimulate mRNA transcription, the two genes would need to be expressed in the same tissues. My RT-PCR screen in **Chapter 5.5.3** showed that both *dEvf-2* and *Dll* are transcribed in the five major developmental stages (larva, early pupa, late pupa, adult body and adult head) but a more detailed description of their regions of expression was still lacking.

I stained third instar larval antenna discs with an antibody against the Dll protein, as shown in **Figure 5.7a-c**. The specific expression of Dll in the antenna disc, and not in the eye disc (data not shown), confirms previously reported Dll expression (Duncan et al. 1998). The protein is also restricted to the cell nuclei which is consistent with its role as a transcription factor.

I next used FISH to investigate the distribution of the *Dll* mRNA in the same tissue. A typical result for the antennal disc is shown in **Figure 5.7d-f**. U2, a spliceosomal RNA (Liu and Gall 2007), which I used as a positive control, is clearly visible in the cell nuclei in panel D and demonstrates that the *in situ* protocol has been successful. However, there is no clear coincident pattern of *Dll* mRNA and Dll protein expression, or a pattern that would be suggestive of post-transcriptional regulation. Despite using six different probes against *Dll*, and varying the incubation time (from 30 mins to overnight), temperature

(from room temperature to 72°C) and probe concentration (from 1:100 to 1:2,000), a wide range of inconsistent results were observed which likely reflects non-specific probe binding. I also attempted to use this protocol to detect the mRNA of the protein-coding gene *Prospero*, which is localised in the ganglion mother cells of the CNS (Lee et al. 2006), but was unable to replicate the published expression pattern.

I simultaneously attempted to visualise *dEvf-2* expression in the same tissues using FISH, but was also unable to detect a consistent pattern between 10 different probes. A typical picture for the antennal disc is shown in **Figure 5.7g-i**. *dEvf-2* is likely to be expressed at a much lower level than *Dll*, which would result in its detection being more difficult. Alternatively, *dEvf-2* may not be expressed in the larval CNS. Given the failure of the *Dll* FISH and the time constraints of this project it was decided not to further pursue this technique, either through optimisation or testing of other tissues.

Consequently, the potential *in vivo* interaction between *dEvf-2* and *Dll* remains a subject for further investigation and I suggest possible future experiments in **Chapter 5.6**.

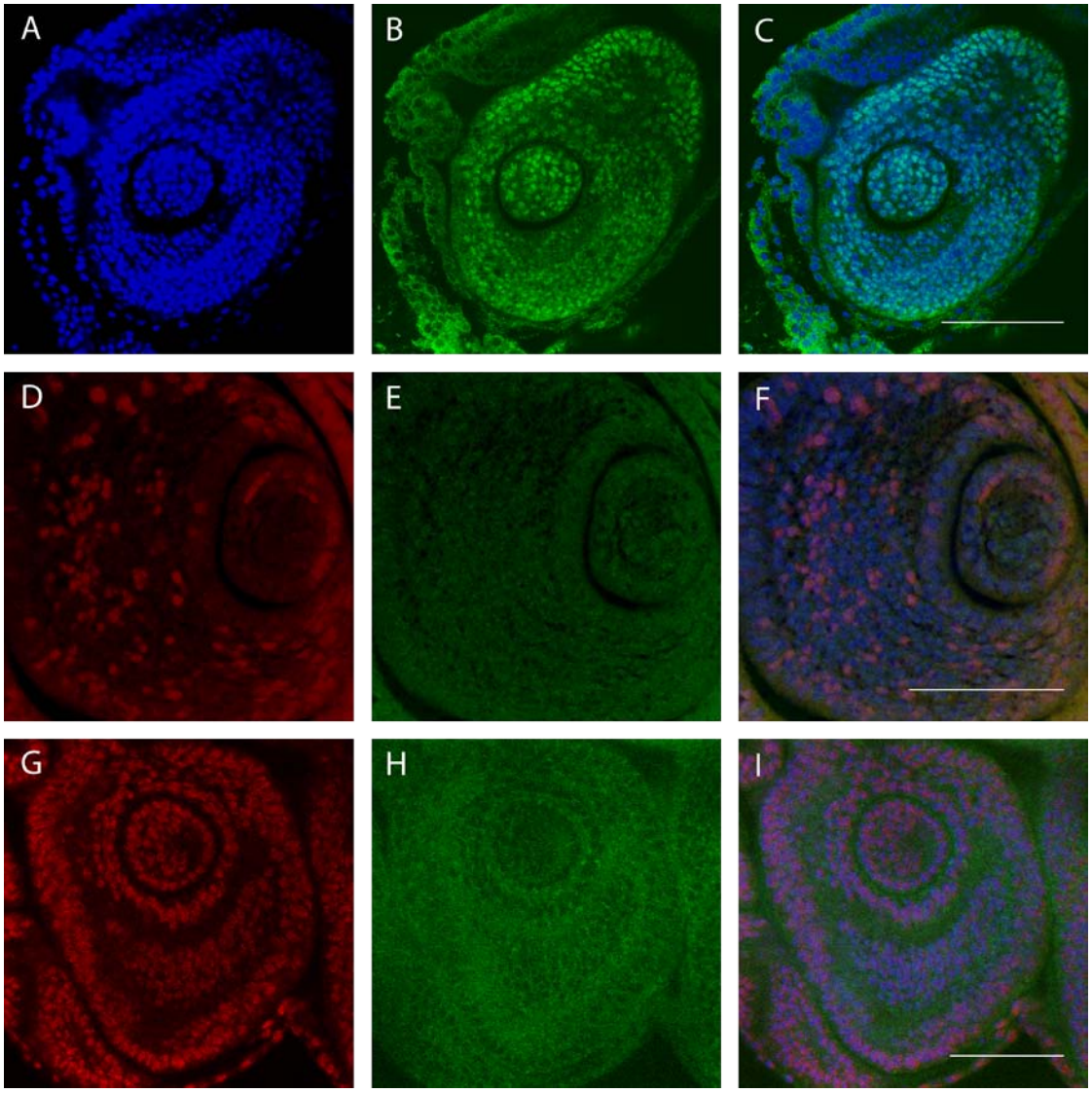

Figure 5.7 Expression of Dll protein and *Dll* mRNA and *dEvf-2* lincRNA expression in the third instar larval antenna disc. A. Hoechst staining (DNA). B. Dll antibody staining. C. Merged image. D. U2 FISH (positive control and nuclear marker). E. *Dll* mRNA FISH. F. Merged image. G. U2 FISH (positive control and nuclear marker). H. *dEvf-2* lincRNA FISH. I. Merged image. All scale bars are 50 µm.

5.6 Discussion

My observation that lincRNA locus locations are not randomly distributed throughout the *D. melanogaster* genome but, instead, are biased to be in the vicinity of genes encoding transcription factors replicates the results of similar studies in mice (Guttman et al. 2009; Ponjavic et al. 2009). These authors observed different enrichments of GO terms, depending on which subset of lincRNAs was selected (e.g. brain expressed), but all tests yielded at least one term that described transcription regulation. This bias of *Drosophila* lincRNA loci to lie near transcriptional regulator genes thus appears to be a general phenomenon across lincRNAs and organisms. Similar studies in other animals should reveal how widespread this relationship is. Further testing of the lincRNAs which give rise to this enrichment in *D. melanogaster* can test their biological relevance.

I tested the expression of all constrained, non-embryonic lincRNA loci whose nearest protein-coding locus encodes a transcription factor using RT-PCR at various developmental stages. Despite no primer optimisation, I was able to verify 31 (72%) of these loci. This is a greater proportion than was detected when attempting to validate smaller lincRNA catalogues previously defined in *D. melanogaster.* RT-PCR at similar developmental stages has been used to verify 7 (58%) lincRNAs defined by possession of a conserved intron (Hiller et al. 2009), while northern blotting detected 28 (39%) lincRNAs defined using a smaller cDNA dataset (Tupy et al. 2005). My RT-PCR analysis may be more

sensitive than northern blotting (two-tailed chi-squared test, $p = 1.1\text{x}10^{-3}$) but the stability of transcripts detected in this way can be questioned. Transcripts known as cryptic unstable transcripts (CUTs) have been identified in yeast (Wyers et al. 2005). These are produced by fortuitous promoter sequences within intergenic regions, are rapidly degraded by the exosome, and are not thought to possess a biological function. The Trf4p protein involved in this degradation is also found in *Drosophila*, suggesting that CUTs may also be transcribed in *D. melanogaster.* RT-PCR can be used to identify these CUTs (Wyers et al. 2005) and it may be that my increased validation rate results from identification of several CUTs, rather than true lincRNAs. The 31 transcribed loci could be investigated by northern blot using the developmental stages in which they are now known to be transcribed, as this would indicate a more stable transcript.

Although none of the 31 transcribed lincRNA loci shared a common expression profile with their adjacent transcription factor, 20 exhibited a regulated expression pattern: they exhibited expression at some developmental stages, but not in others. It has been suggested previously that this variable expression implies that a lincRNA is functional (Ravasi et al. 2006), but this need not necessarily be true. A particular sequence might be ubiquitously expressed at a very low level near the limit of what can be detected by PCR and, depending on the quality of individual RNA extractions, might be observed in certain samples but not others. If these 20 lincRNAs are, however, functional, then these regulated expression patterns describe the developmental

stages at which these functions could be experimentally characterised. *In situ* hybridisation could reveal in more detail where these lincRNAs are expressed, while RNAi could be used to examine any phenotypic consequences of knocking down lincRNA expression during the stages in which their expression has been detected. The lack of correlated expression suggests that these lincRNAs are unlikely to function as *cis*-regulators of the neighbouring transcription factor, although this function may be tissue-specific. These lincRNAs may regulate the expression of the transcription factor in one or two tissues, but the transcription factor may be regulated by other factors in other contexts. Here, one would not expect to see related expression profiles between the lincRNA and the transcription factor, despite the lincRNA playing a regulatory role. I suggest that disrupted expression of the transcription factor could be one of the first possible phenotypes which should be investigated in any future RNAi experiments.

Of the eight lincRNAs whose ubiquitous expression was shared with their adjacent transcription factor, RT-PCR confirmed that four (50%) represented unannotated exons of the transcription factor gene (**Figure 5.3**). Only *dEvf-2* has been conclusively shown by RACE to be transcribed from an independent locus. Previous work suggested that only five (18% of those sampled) lincRNAs represented unannotated exons or UTRs of a neighbouring protein-coding gene (Tupy et al. 2005). The reduced frequency of independent transcripts in my group may reflect the selection of only constrained, non-embryonic lincRNAs that are co-expressed with their neighbouring

transcription factor. It is clear that EST sequences and RNA-seq data do not always represent the full-length transcript, and that many gene models annotated in *D. melanogaster* remain incomplete. Indeed, microarray experiments have shown that unannotated exons can be found on average up to 20 kb from the 5' end of current gene models (Manak et al. 2006). A technique such as RACE which identifies the location of the 5' cap and the 3' polyA tail in a mature (Pol II-transcribed) lincRNA transcript is therefore an essential step in identifying a true lincRNA before investing in further detailed experimental investigation.

The one lincRNA of the remaining four, *dEvf-2*, which I have studied further, does appear to be able to enhance transcription of the adjacent transcription factor gene. RNAi treatment targeting *dEvf-2* in S2 cells led to a significant depletion of *Dll* mRNA levels, but it remains to be seen if and where *dEvf-2* has the same effect *in vivo*.

The *Dll* gene is surrounded by a number of enhancer elements (**Figure 5.8**) which can drive expression of a reporter in different, overlapping, regions of the embryo (Vachon et al. 1992). One of these, *Dll*-304, also contains sequence elements responsible for repressing *Dll* expression in the embryonic abdomen (Gebelein et al. 2002). *Dll*-179, which entirely overlaps the *dEvf-2* locus, drives expression in a subset of the *Dll*-expressing cells in the labium (which gives rise to head structures), the maxilla (sense organs) and the labrum (also in the head). This enhancer activity may be mediated by the mature lincRNA which it encodes, particularly as the S2 cell line in which I tested its function is also

embryonically-derived. Nevertheless, it should be remembered that this lincRNA was originally defined as non-embryonic. The RNA-seq data discussed in **Chapter 4** also indicate that there is little, if any, *dEvf-2* expression in embryonic tissues.

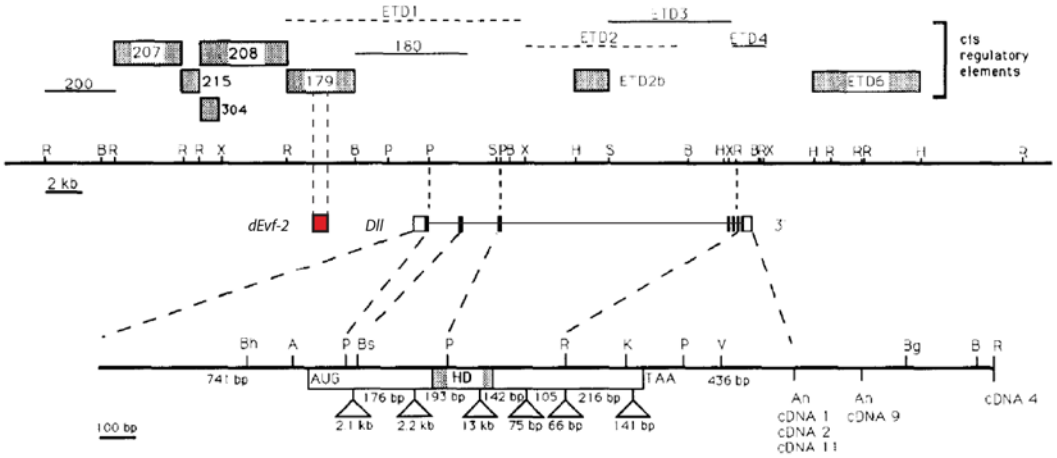

Figure 5.8 Physical organisation of the *Dll* locus. *dEvf-2* overlaps the *Dll*-179 enhancer. Adapted from Vachon et al. 1992.

The RNA-seq data suggest that *dEvf-2* is highly and specifically expressed in male tissues, while *Dll* is expressed relatively consistently throughout development after about the first 2 hours (data not shown). *Dll* has been extensively studied as it is required for distal limb development and is also expressed in the nervous system (reviewed in Panganiban and Rubenstein 2002), but no evidence has yet been published to suggest a role in any male-specific processes. It may be that this high expression of *dEvf-2* in males is unrelated to *Dll* and instead conveys a function distinct from its putative embryonic *cis*-regulation of *Dll*.

These observations suggest that further characterisation of *dEvf-2* and its relationship with *Dll* would be worthwhile. Subsequent experiments have already revealed that over-expression of *dEvf-2* in S2 cells leads to a similar over-expression of *Dll* mRNA (Andrew Bassett, personal communication), suggesting that these genes may share a dose-response relationship.

Due to the failure of my FISH experiments (**Figure 5.7**) it is still unclear where *dEvf-2* is expressed; knowing this is likely to help delineate its possible places of action and, if these are coincident with those of *Dll*, where it could regulate *Dll* in *cis*. *dEvf-2* expression could be up- or down-regulated in specific tissues by placing it, or an antisense hairpin, respectively, under the control of the Gal4/UAS system (Brand and Perrimon 1993). This involves cloning the sequence of interest so that its expression is regulated by an Upstream Activation Sequence (UAS) element. These elements are activated by binding of the Gal4 protein and many lines already exist where Gal4 is expressed in a known, regulated manner, for example in the developing CNS. By crossing the Gal4 flies to flies containing the UAS construct it is possible to express any sequence in a wide range of patterns. *dEvf-2* expression should be first disrupted in both embryonic and male tissues to assay the effect on the Dll protein.

It would also be interesting to determine if *dEvf-2* and Dll physically interact with each other. The Dll protein could be purified by immunoprecipitation, and RT-PCR using *dEvf-2* primers attempted to determine whether the lincRNA was part of the purified complex. Alternatively, and complementarily,

a *dEvf-2* construct which places a biotin tag at the 5' end of the lincRNA could be constructed. Once *dEvf-2* was isolated using streptavidin beads, a western blot could be used to test if the Dll protein could be isolated through its interaction with the lincRNA.

dEvf-2 is particularly intriguing due to its similarity with *Evf-2* (**Figure 5.9**), a lincRNA identified in mouse which similarly regulates two members of the *Dlx* family in *cis*. This type of analogous behaviour was previously predicted for the RNA-seq lincRNA dataset in **Chapter 4**, although *dEvf-2* does not overlap any of the lincRNAs defined there. There are six *Dlx* genes identified in mouse (Panganiban and Rubenstein 2002), which are found in three pairs of adjacent loci. Both members of one of these pairs, *Dlx5* and *Dlx6*, are up-regulated in cell culture by a lincRNA known as *Evf-2* which overlaps an intergenic enhancer found between *Dlx5* and *Dlx6* (Feng et al. 2006). *Evf*-2, unlike *dEvf*-2, is a multi-exonic transcript with a large intron spanning the entire *Dlx6* locus. The lincRNA and the two protein-coding transcripts share a similar expression pattern in the ventral forebrain and *Evf-2* can drive reporter expression in two neural cell lines in a dose-dependent manner. Increased expression of these two protein-coding genes appears to require a physical interaction between *Evf-2* and a third protein in the *Dlx* family, Dlx2 (Feng et al. 2006). *Evf-2* is likely to function as a single-stranded RNA (ssRNA) molecule, as only injection of ssRNA stimulated *Dlx5* and *Dlx6* transcription in the cell lines. Mice expressing a truncated *Evf-2* transcript show a reduced number of GABAergic neurons in the early postnatal hippocampus and

reduced *Dlx5* expression; however, it appears that *Dlx6* is actually up-regulated in these mice in contrast to what was observed in cell culture (Bond et al. 2009). *Dlx5* and *Dlx6* are also expressed in male fetal testis where they are required for masculinisation and testicular steroidogenesis (Nishida et al. 2008). It is possible that their expression here could also be regulated by *Evf-2* and, if indeed it does share a common regulatory mechanism with *dEvf-2*, this could explain the high expression of *dEvf-2* observed in male *Drosophila* tissues.

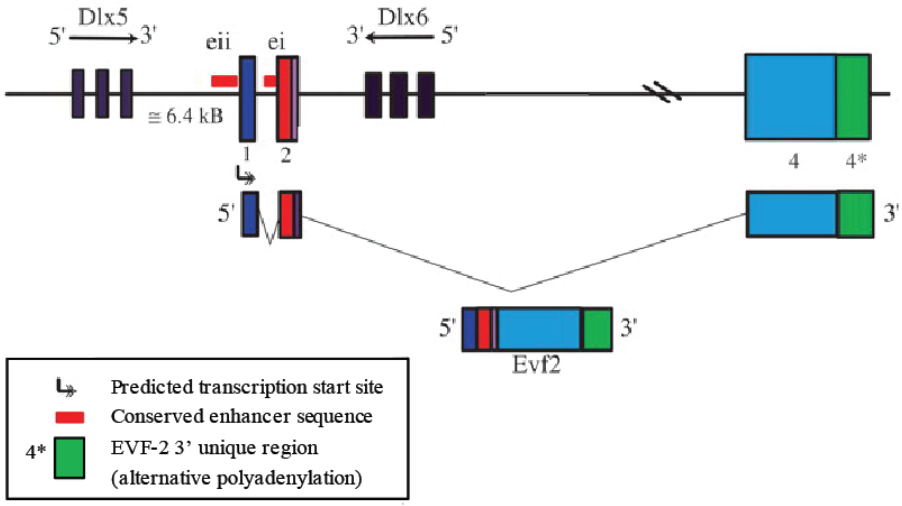

Figure 5.9 Genomic organisation of the *Dlx5/Dlx6* locus containing the mouse lincRNA *Evf-2*. Adapted from Feng et al. 2006.

It is tempting to speculate that *dEvf-2* is orthologous to *Evf-2* and that they share a common evolutionary origin, making this locus the first example of such a deeply conserved functional lincRNA. There is, however, no sequence similarity apparent between the *Drosophila* and the mouse sequences despite them both being highly conserved in their own phylogenetic clades. Sequences

could be collected across a range of species which are intermediately related to both *Drosophila* and mouse and their mutations tracked to try to identify an evolutionary pathway linking the two sequences. If *dEvf-2* physically interacts with Dll, then the conservation of this physical interaction could be examined by testing whether *dEvf-2* can interact with the mouse Dlx2 and, reciprocally, whether *Evf-2* can interact with Dll. Without this evidence, these loci should be considered to be analogous, because they both positively regulate the expression of adjacent Dll/Dlx transcription factor(s). The only other noncoding RNAs which are known to act similarly in *Drosophila* and mouse are the dosage compensation (*roX1*/*roX2* and *Xist*) and heat shock response (hsr-ω and sat III transcripts) noncoding RNAs discussed in **Chapter 1.2**. To my knowledge, this is the first example of such striking similarity between two lincRNAs in these distantly related species.

In this chapter I have shown that lincRNA loci in *D. melanogaster* are found preferentially near to loci encoding transcription factors, and that this can reflect a functional relationship between the lincRNA and transcription factor. I validated 31 (72% of those tested) genuine lincRNAs, many of which appear to be expressed in a regulated manner. One of these lincRNAs, *dEvf-2*, can positively regulate expression of its genomically neighbouring transcription factor *Dll* in cell culture. Three remaining candidates have not yet been tested for this function. This type of *cis*-regulation has now been identified in both mouse and *Drosophila* and, as such, may be a general animal phenomenon rather than an indicator of increasing biological complexity. Further inspection

of this mechanism may even reveal other analogous lincRNAs, beyond *dEvf-2/Evf-2*, which are shared between diverse species. The regulation of genomically adjacent protein-coding genes by lincRNAs therefore appears to be a useful hypothesis to follow when investigating lincRNA function.

Chapter 6: SEARCH FOR lincRNA KNOCKDOWNS

6.1 Abstract

The biological roles of nearly all *Drosophila* lincRNA loci remain experimentally untested. One way to address this would be to mutate promoters at these loci and examine the resulting mutants for any phenotypic consequences of disrupting lincRNA transcription. *Drosophila* is ideal for this type of experiment due to the vast collection of *Drosophila* insertion mutation lines already available from various stock centres.

A set of 249 lincRNA promoters was predicted from a large collection of TSSs that I called using a set of genome-wide 5'-SAGE tags. My findings suggest that lincRNAs are particularly important in male-specific developmental processes and that lincRNAs make less frequent use of canonical promoter motifs than protein-coding genes.

Five mutations available from the Bloomington Stock Centre were identified within putative lincRNA promoters, but only one of these causes a significant increase in transcription of the relevant lincRNA. Current stock collections thus may not be the best place to begin a search for lincRNA knockdowns for phenotypic analysis, but they could be useful if a mutation is identified within a promoter near to a particular lincRNA locus of interest.

6.2 Introduction

Reverse genetics represents an approach to discovering gene function that starts from the DNA and then creates mutants to analyse the gene's function (Alberts et al. 2002). This can be done in a targeted manner, whereby the sequence of a particular gene of interest is mutated, or in a more naïve way, where a large set of mutants are generated indiscriminately, for example through chemical mutagenesis, and then screened for mutations in the gene(s) of interest.

A range of different mutagens exists for the *Drosophila* genome, with transposable elements being perhaps the most useful for reverse genetics experiments. One particular class of DNA transposons, known as P-elements, are the most frequently used type of transposable elements. These are not naturally present in lab strains of *Drosophila* and hence their transposition can be controlled by crossing them to strains containing a transposase enzyme (Cooley et al. 1988). P-elements have been shown to confer a significant mean 55% fitness cost when inserted and made homozygous in flies previously lacking such elements (Mackay 1986). These types of transposable elements are also rarely found within transcriptional units (TUs) in *Drosophila* species in nature, despite it being known that they are able to insert into these regions (Charlesworth et al. 1994). These observations are taken as evidence that such transposable element insertions in functional sequences are detrimental to the

organism carrying it and, as such, have been selectively purged from the population.

Transposable element mutagenesis is a powerful technique for reverse genetics. Many thousands of mutant lines are available at relatively little cost from stock centres around the world, such as the Bloomington Stock Centre or the Exelixis Collection at Harvard Medical School. These centres can be queried for mutations within a genomic region of interest, making it possible for individual researchers to carry out naïve reverse genetics. The transposable elements stored within these centres are particularly stable due to their modified genetics described in **Chapter 6.3.2**.

Transposable element mutagenesis should also be useful for the investigation of lincRNA function. Our lack of knowledge about the structure and the functional domains of these sequences suggests that it is not clear which mutations, if any, within them will disrupt their sequence sufficiently to impair their function. By inserting a transposable element into the lincRNA promoter, it is hoped that the transcription termination signal encoded in the element will disrupt lincRNA transcription and thereby its function. P-elements have a recognised preference to insert into gene promoters (Spradling et al. 1995) and are therefore ideal when trying to mutate lincRNA promoters.

This chapter describes an attempt to identify fly lines carrying mutated lincRNA promoters which interfere with the expression of the associated lincRNA. A set of putative lincRNA TSSs and promoters is defined, and their

characteristics compared to those of protein-coding genes. Four mutant lines that carry a P-element insertion in a putative lincRNA promoter are described, but only one of these results in significant disruption of lincRNA transcription.

6.3 Materials

6.3.1 SAGE Tags

I defined a set of TSSs using a 5'-end mRNA dataset available from the MachiBase service (Ahsan et al. 2009). A modified version of the 5'-SAGE (Serial Analysis of Gene Expression) method (Hashimoto et al. 2004) which is summarised in **Figure 6.1** was used to generate a large set of 26 bp sequences. These sequences, known as SAGE tags, correspond to the 5' end of mRNA molecules from various tissues. The non-redundant set of sequences produced by this method for various *D. melanogaster* tissues were downloaded from MachiBase (http://download.utgenome.org/pub/machibase/) and are summarised in **Table 6.1**.

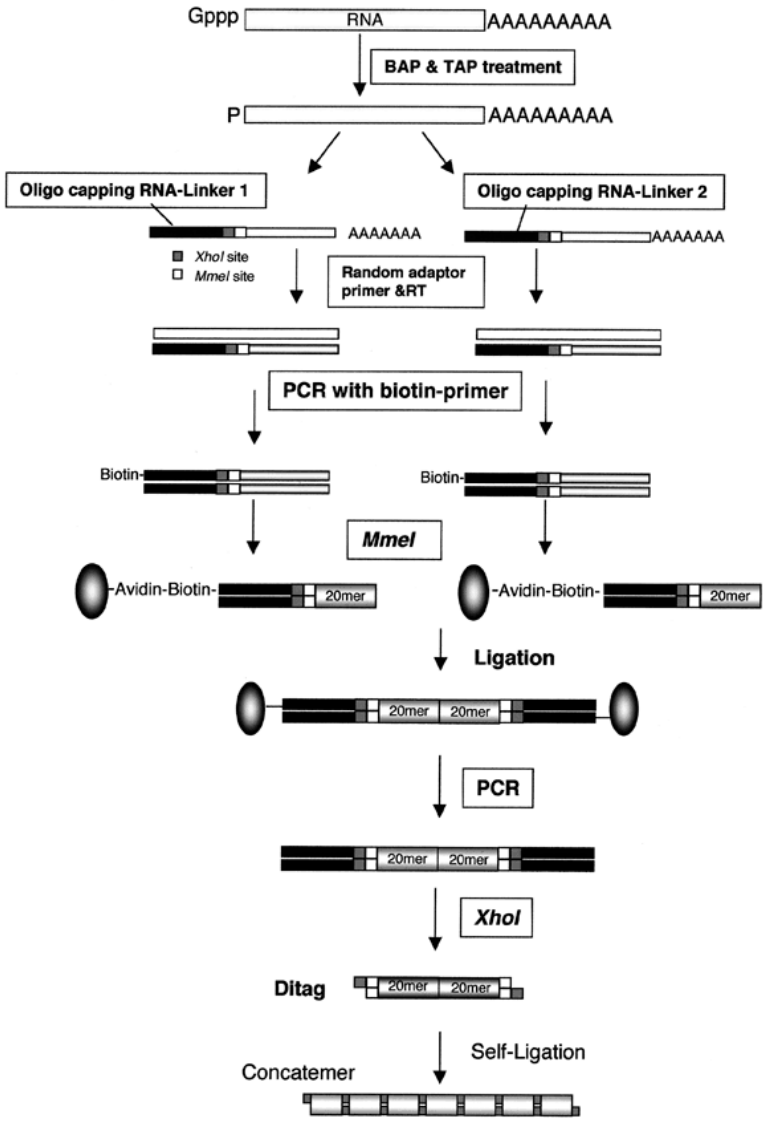

Figure 6.1. Scheme for construction of a 5'-end SAGE library. RNA is extracted as described in **Chapter 2.2.6.1**, and the sample split in two. The 5' cap is removed from each RNA molecule by treatment with Bacterial Alkaline Phosphatase (BAP) and Tobacco Acid Pyrophosphatase (TAP) to allow a synthetic oligonucleotide containing two restriction endonuclease sites (*Mme*I and *Xho*I) to be added enzymatically. The sequences of the oligonucleotide are different between the two samples, although both contain the same two restriction sites. Random primers are used to synthesise the first cDNA strand by reverse transcription, before the second strand is synthesised using the same primers and a biotin-bound 5' primer which recognises the synthetic oligonucleotide. This cDNA is cleaved with *Mme*I to remove the cDNA sequence around 20 bp from the recognition site and the biotin-bound fragments are captured on streptavidin beads. Both samples are then combined where fragments from opposing samples are ligated. These ligated fragments are amplified using PCR to select only ligations between opposing samples with the primers positioned in both synthetic oligonucleotides. The second restriction enzyme *Xho*I is then used to remove these oligonucleotide sequences and the remaining 'sticky ends' allow the fragments to be concatenated into longer DNA molecules. Finally, the DNA molecules are sequenced using the Illumina/Solexa system. Taken from Hashimoto et al. 2004.

6.3.2 Transposable element insertions

Information about a large number of insertion mutations in the *D. melanogaster* genome is available from FlyBase. Many of these have been generated by the Berkeley *Drosophila* Genome Project (BDGP) Gene Disruption Project (Bellen et al. 2004). This project aims to create a mutation in every single gene in the *Drosophila* genome using a genetically engineered P-element as the mutagenic agent which results in stable lines, each containing a single mutation. As of 2004, the BDGP reported they had deposited 7,140 of these lines (Bellen et al. 2004) in the Bloomington Stock Centre at Indiana University.

By August 2008, FlyBase reported that the Bloomington Stock Centre contained 19,553 mutant insertion lines of which 13,141 transposable element insertions (of all types) had been mapped to a specific location (i.e. one or more genomic coordinates). It is these 13,141 insertions which are used in this work.

6.4 Methods

6.4.1 McPromoter

Promoter motifs were identified around TSSs using the McPromoter program (Ohler 2006). This uses a generalised HMM, which is similar to a regular HMM of the type described in **Chapter 2.1.6**, but it tests for the presence of a series of neighbouring features. In this case, the series is a run of consecutive

bases that are part of the same promoter motif. McPromoter identifies several different parts of the promoter, including two regions of distinct GC content upstream of the promoter, the core motif, spacer, initiator and the downstream region. Unlike other motif prediction programs (e.g. Reese 2001) core motifs of more than one state can be found using this program. McPromoter is able to identify the following five motifs: Initiator Motif (Inr); Downstream Replication Element (DRE); Downstream Promoter Element (DPE)/Inr; TATA/Inr; and Motif6/Motif1. Motif6 and Motif1 have been predicted computationally (Ohler et al. 2002), although their functional relevance remains unclear. The best model is chosen by comparing the log likelihood of the best promoter model with the best non-promoter model and, in this way, it is possible to determine the position of the core promoter as well as its motif.

McPromoter was trained on a previously well-defined set of curated core promoters (Ohler et al. 2002). Its equal recognition rate (the level at which the rate of true positives equals the rate of true negatives) was identified as 94.1%. This is an improvement on the 89.9% obtained from the previous iteration of McPromoter, which could identify only TATA elements. With a sensitivity of 64%, ~35,000 core promoters were predicted in the *D. melanogaster* genome and, at the well-studied *Adh* locus, 52% of known promoters were found with a false positive only once every 12 kb. This relatively low sensitivity may be caused by a still unrecognised diversity of core promoter motifs; 15% of the well-studied promoter set did not match any of the motifs contained within the current McPromoter models.

McPromoter looks at a 300 bp sliding window that moves in steps of 10 bp. It is therefore unable to identify promoter motifs within the first 250 bp of a sequence. I use it to analyse 1 kb of sequence, which is ±500 bp of each TSS identified in **Chapter 6.5.1**. I use a minimum cut-off score of 0.03 to call a promoter as genuine, giving a sensitivity (65%) comparable to that discussed above (http://tools.igsp.duke.edu/generegulation/McPromoter/).

6.4.2 Genomic DNA Extraction

Genomic DNA is extracted for genotyping both *wt* and Bloomington stock centre lines. Two flies are crushed with a pipette tip in the presence of 50 µl of genomic lysis buffer. This buffer contains 10 mM Tris-HCl (pH 8.2), 1 mM EDTA, 25 mM NaCl, and 200 µg proteinase-K. The proteinase-K is added fresh to the buffer every day. The sample is then incubated at 37°C for 30 mins before the proteinase-K is inactivated by incubation at 95°C for 5 mins. Samples are diluted to 200 µl with RNase-free H_2O.

6.5 Results

6.5.1 Promoter Prediction

A critical first step in the identification of putatively mutated lincRNA promoters is the definition of a set of lincRNA promoters themselves. As current gene models and particularly EST-based models of the type described in **Chapter 3** are likely to be incomplete (Manak et al. 2006), I used a set of 5'-SAGE tags to define TSSs from which promoter motifs could be predicted.

The non-redundant set of 5' sequences available for a variety of developmental stages and S2 cells were downloaded from MachiBase (Ahsan et al. 2009). These were converted into FASTA format (they were not supplied with quality scores) and mapped using Bowtie (**Chapter 2.2.1.2**). Only the uniquely mapping hits, which accounted for 18-34% of all reads, were retained, as shown in **Table 6.1**.

Tissue	**Reads**	**Reads Mapped Uniquely**
S2 Cells	2,803,878	751,905
Embryo	2,123,688	691,983
Larva	2,231,078	716,969
Young Female	3,398,754	788,400
Young Male	7,859,645	1,398,589
Old Female	3,258,523	1,099,938
Old Male	3,442,886	1,030,416

Table 6.1. Numbers of 5'-SAGE reads before and after mapping with Bowtie.

The mapped reads were clustered on the genome separately for each developmental stage and each DNA strand if they shared at least one overlapping base. A TSS was then called within these clusters if the cluster was supported by at least three reads and the 5' most locations of at least two of these reads were found at the same position (Rach et al. 2009). This filtering process ensures that only high-quality, reproducible TSSs are included in subsequent steps. Promoters were predicted within the genomic sequence stretching from 500 bp upstream to 500 bp downstream of each TSS using McPromoter.

These predicted promoters were associated with their nearest gene, whether it was a protein-coding gene or a lincRNA locus defined in **Chapter 3**. This was done by defining a gene neighbourhood known as a territory, as discussed in **Chapter 2.2.4**, for each locus. Promoters were associated with each territory, provided that they did not overlap the gene model within the territory (to remove any complications of studying secondary promoters found within exonic sequence). The distribution of TSSs and predicted promoters in both protein-coding gene and lincRNA loci territories for each developmental stage is summarised in **Figure 6.2**.

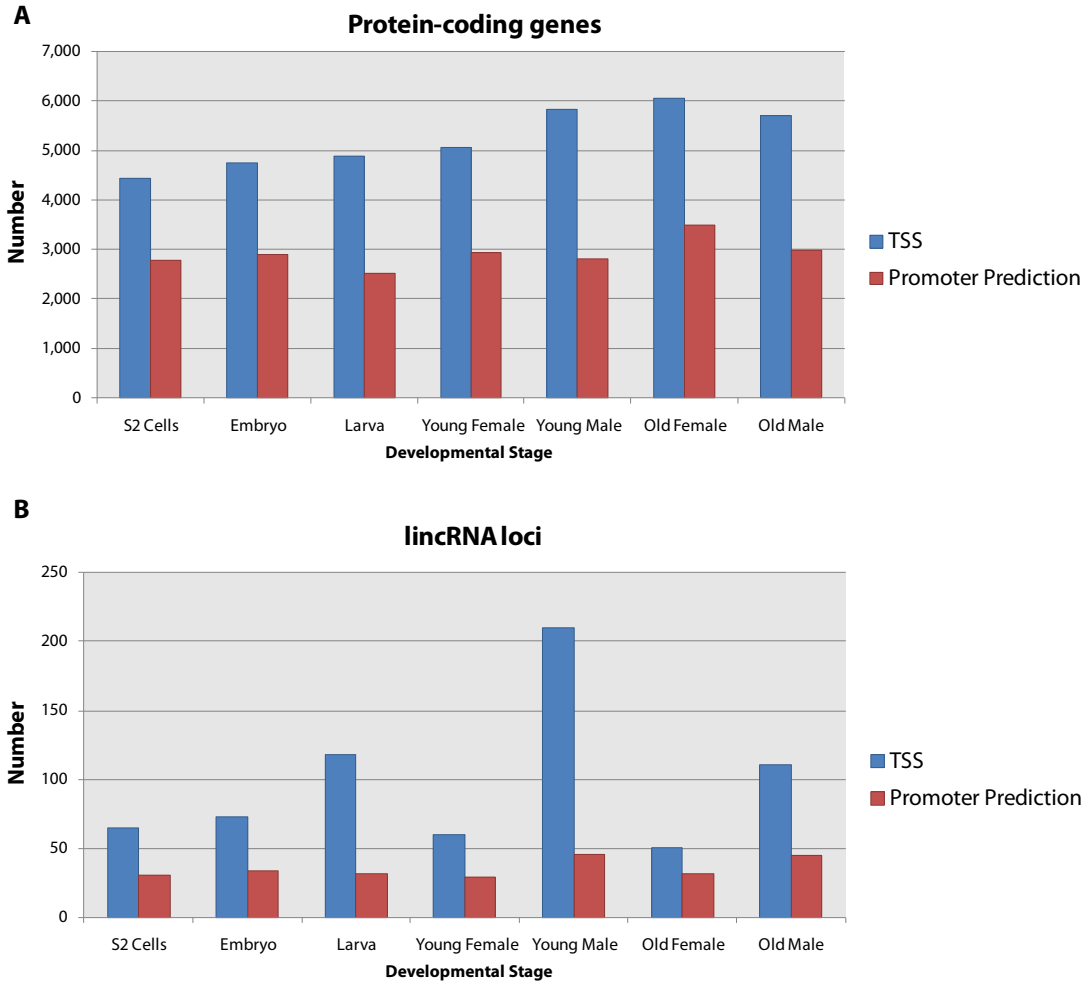

Figure 6.2. Distributions of TSS and promoter prediction numbers across territories for each of the developmental stages for which 5'-SAGE tags were available. A. Protein-coding gene territories. B. lincRNA locus territories.

As expected from the previously reported observation of lower expression of these lincRNAs relative to protein-coding genes (**Chapter 3**), far fewer putative TSSs and promoters were found within lincRNA territories. There is a 53-fold increase in the number of putative protein-coding TSSs relative to those found for lincRNA loci, which is matched by a comparable 82-fold increase in the number of promoters. A greater proportion of protein-coding TSSs (56%) gives rise to a predicted promoter than that of lincRNA TSSs

(36%), and this difference is statistically significant (Mann-Whitney, $p = 0.03$). This observation is not overly surprising as McPromoter was trained on protein-coding gene promoters but it does suggest that lincRNA loci are, indeed, less frequent users of these canonical promoter motifs whose presence is examined using McPromoter.

It is also apparent that the frequency of TSSs and promoters between these two sequence classes varies across the developmental stages sampled here. During the early developmental stages there is a gradual increase in the number of TSSs associated with both protein-coding gene and lincRNA loci, although the number of promoters peaks for both classes at the embryonic stage. However, a clear difference becomes apparent when the samples are split by sex. While protein-coding TSSs and promoters are found in similar numbers in males and females, those found near lincRNA loci are predominantly found in males, both young and old (**Figure 6.2**).

The core promoter in *D. melanogaster* has previously been reported to stretch from 50 bp upstream to 50 bp downstream of the TSS (Ohler 2006), and so core promoters were defined here as ±50 bp of the 249 promoter sites within 80 lincRNA loci that were predicted by McPromoter. These putative core promoters were cross-referenced with the set of 13,141 transposable element insertions stored at the Bloomington Stock Centre. Seven of these insertions were found to lie within promoters within lincRNA locus territories, and thus may represent insertions into the genuine lincRNA promoter, while 334 overlap putative protein-coding gene promoters. Upon checking the FlyBase genome

browser it was shown that the protein-coding gene model neighbouring one of these lincRNAs (CK658523) had since been extended to include the lincRNA and so it was discarded at this stage. The relevant genomic coordinates of the remaining six lincRNAs are summarised in **Table 6.2**.

6.5.2 LincRNA expression verification

The expression of the remaining six lincRNAs and their neighbouring protein-coding genes was examined by RT-PCR on a mixed-sex, whole-adult *wt* cDNA library. After 40 cycles, five out of six lincRNAs were found to be expressed (BI243900 was not detected) and all 12 protein-coding genes were detected, as shown in **Figure 6.3**. The positive TBP controls are not shown here. All PCR product sizes were in the expected range except FBgn0033703, which produced three bands. The smallest of the bands in the FBgn0033703 lane corresponds to the size of product predicted by the primer locations, while the other two represent probable non-specific amplification. This is because the primers are placed in the same exon of the mRNA, and so it is unlikely that a complex alternative splicing event has taken place to move the primer locations further apart on the transcript.

lincRNA	**Genomic coordinates**	**Territory coordinates**	**Promoter coordinates**	**Insertion position**	**Insertion ID**	**Stock genotype**	**Stock ID**
CO327356	2R:8116082-8116815	2R:8115217-8121679	2R:8116765-8116865	2R:8116829	FBti0007611	w^{1118}; P{EP}EP502	FBst0017172
BI243900	2R:9055008-9055534	2R:9054910-9055769	2R:9055464-9055564	2R:9055547	FBti0021571	y^{1} w^{67c23}; P{SUPor-P} KG04129	FBst0013520
EC069024	3L:16574793-16575141	3L:16574567-16575301	3L:16574674-16574774	3L:16574733	FBti0070777	y^{1} w^{67c23}; P{EPgy2}EY13479	FBst0021094
CK658697	X:2186781-2187444	X:2186781-2201665	X:2187403-2187503	X:2187477	FBti0015945	P{w[+mC]=lacW}G0226a w^{67c23} P{lacW}G0226b, $l(1)Go226^{GO226}$/FM7c	FBst0012232
BQ103275	X:20362238-20381679	X:20362186-20382126	X:20381691-20381791	X:20381713	FBti0017214	w^{1118} P{GT1}BG01461	FBst0012471
BI640269	X:1179817-1180446	X:1178050-1181262	X:1181146-1181246	X:1181220	FBti0021960	y^{1} P{SUPor-P} KG05635	FBst0013908

Table 6.2. Possible lincRNA knockdown coordinates, P-element insertion locations and stock information.

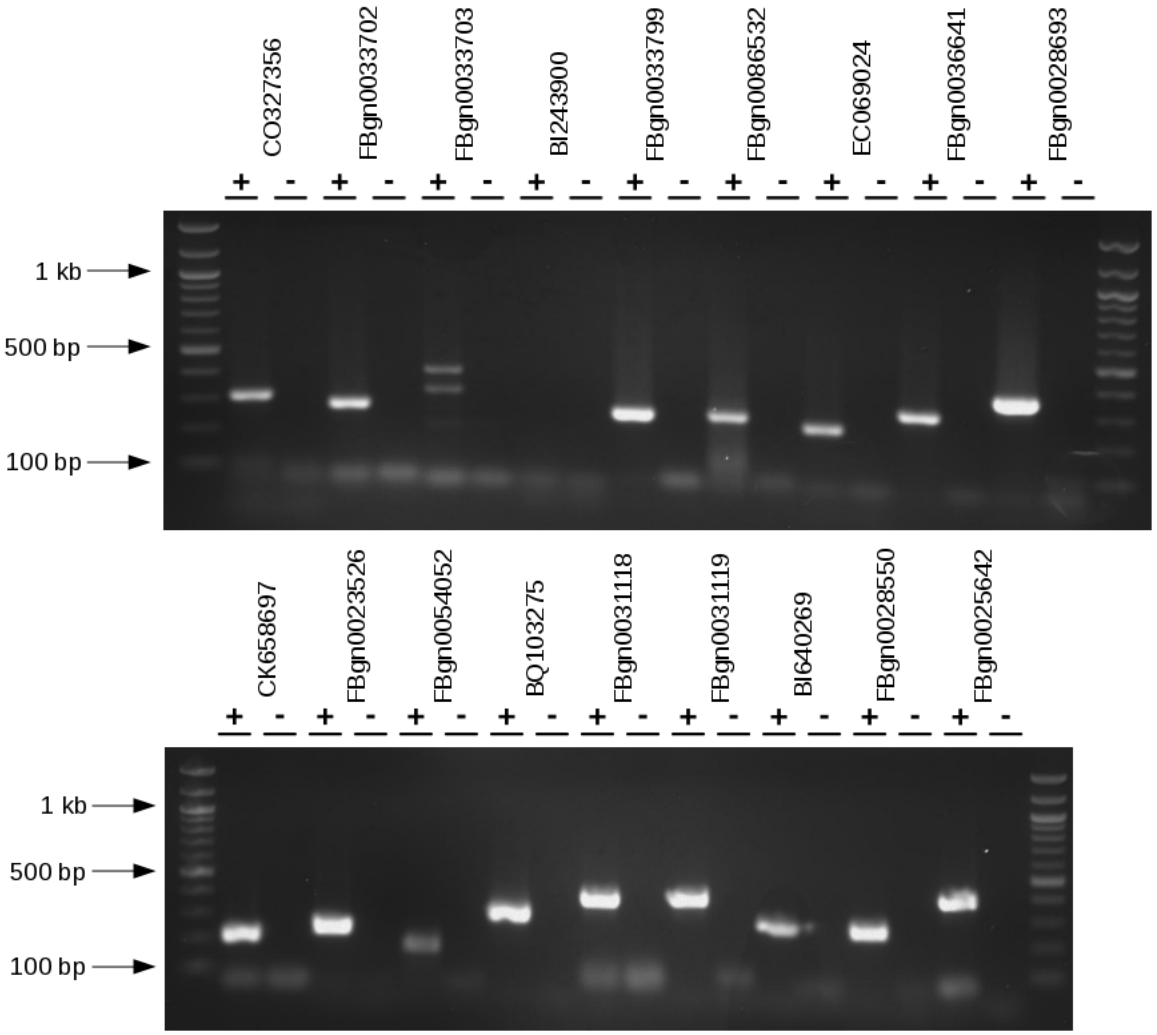

Figure 6.3. Expression of experimental lincRNAs and neighbouring protein-coding genes in *wt* flies. Expected product sizes: CO327356 – 314 bp, FBgn0033702 – 282 bp, FBgn0033703 – 215 bp, BI243900 – 397 bp, FBgn0033799 – 326 bp, FBgn0086532 – 325 bp, EC069024 – 240 bp, FBgn0036641 – 294 bp, FBgn0028693 – 367 bp, CK658697 – 221 bp, FBgn0023526 – 253 bp, FBgn0054052 – 180 bp, BQ103275 – 346 bp, FBgn0031188 – 378 bp, FBgn0031119 – 385 bp, BI640269 – 264 bp, FBgn0028550 – 250 bp, FBgn0025642 – 391 bp. Positive (+) lanes and negative (-) reactions for each PCR reaction are shown.

6.5.3 LincRNA expression in mutants

The RT-PCR experiment described in **Chapter 6.5.2** was repeated using cDNA libraries from the mutant stock relevant for each lincRNA. Expression of all five lincRNAs and 12 protein-coding genes was still observed to produce PCR products similar to those seen in **Figure 6.3** suggesting that, even if these P-elements had inserted into the genuine lincRNA promoter, they had not completely abolished its transcription.

The insertion labelled FBti0015945 in the CK658697 territory was homozygous lethal but, as the stock contains two P-element insertions, it remains unclear which is responsible for this lethality. This line was therefore discarded at this point.

Instead, these insertion lines may only show reduced expression of their associated lincRNAs and this was investigated using real-time RT-PCR. The expression of each of these four lincRNAs, relative to *Gapdh*, was tested for each of four biological replicates (individually isolated cDNA libraries). The relative quantity of each lincRNA in both the *wt* line and the mutant line was calculated as described in **Chapter 2.2.7**, with the *wt* sample being considered the control. Surprisingly, only FBti0070777 resulted in a significant (one-tailed T-test, $p = 0.04$), approximately 1.4-fold, increase in lincRNA (EC069024) expression. The genomic neighbourhood of EC069024 is shown in **Figure 6.5**.

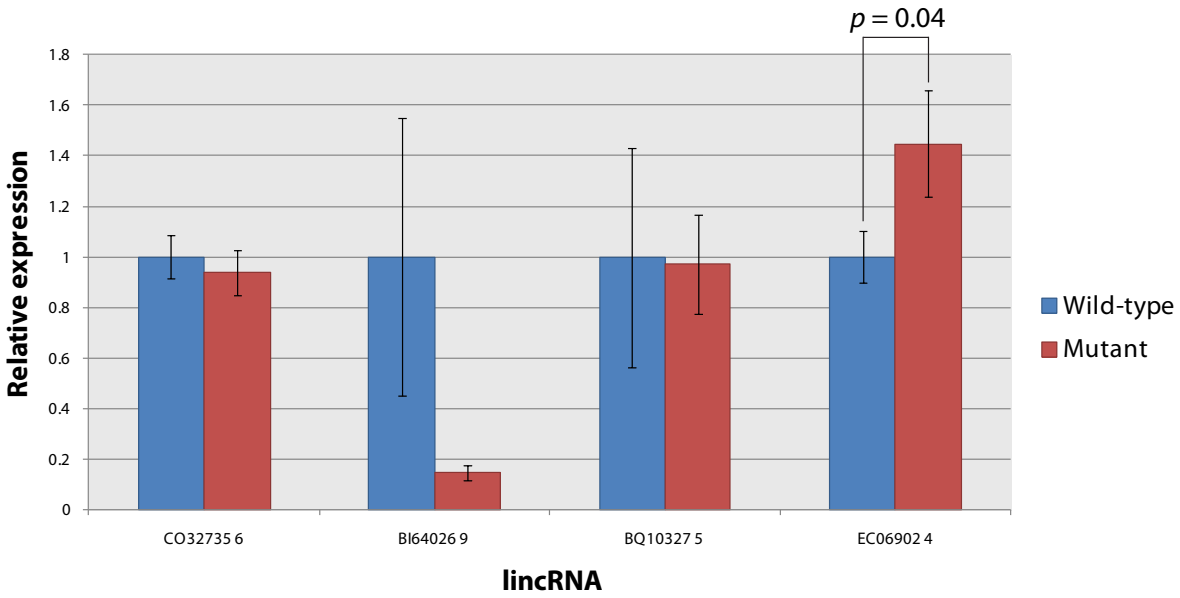

Figure 6.4. Relative expression of four lincRNAs in *wt* flies and P-element mutant lines with mutations in the putative core promoter. Error bars represent the standard error of the mean.

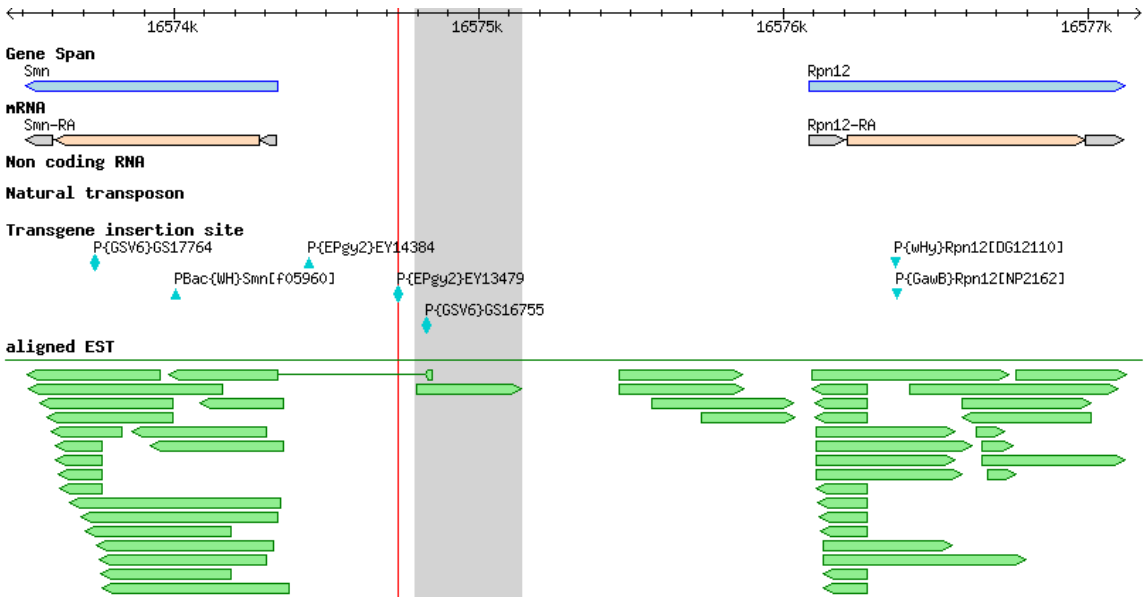

Figure 6.5. Genomic region surrounding EC069024. The neighbouring genes (Survival motor neuron (FBgn0036641) and *Rpn12* (FBgn0028693)) are shown, while EC069024 is highlighted in grey. The red line represents the location of P-element insertion FBti0070777.

6.5.4 Genotyping

The presence of the P-element insertions in these stock lines was confirmed by sequencing of PCR products across the insertion site. In three cases – BI640269, CO327356 and EC069024 – the insertion site was verified to be within 10 bp of the reported site and within the core promoters as defined in **Table 6.2**. The sequencing reaction for FBti0017214 (lincRNA BQ103275) failed. It may be that this line has truly lost its P-element and that the core promoter of the lincRNA did not contain this insertion. However, as these flies do not display any reduced lincRNA expression, I would suggest that investigating their genotype in greater detail is unlikely to lead to any further insight into potential lincRNA functionality.

The *wt* line was also genotyped to verify the absence of P-element insertions within these four lincRNA territories. This was done by PCR of genomic DNA across the insertion sites and the absence of all four insertions was supported by PCR products of the predicted size (data not shown).

6.6 Discussion

In this chapter, I describe how I identified 249 putative lincRNA promoters with 80 lincRNA loci territories. These indicate a high level of lincRNA expression in the male and, specifically, the young male, which is consistent with the observations in **Chapter 4** that RNA-seq based lincRNA loci are also involved in male-specific development.

LincRNA TSSs less frequently contain predicted promoter motifs than those associated with protein-coding genes, which may be limited by the selection of the motifs modelled by McPromoter. However, it is not yet clear how many of these lincRNA TSSs represent the true start site of the lincRNA whose territory they are found within, and this is a general problem with this work. In fact, many of these TSSs could themselves be false positives. A larger study of gene promoters carried out since this work as part of the modENCODE consortium has suggested that Cap Analysis of Gene Expression (CAGE, of which 5'-SAGE is a modified protocol) tags may contain many false positives and, as such, should not be used alone to identify TSS locations (Hoskins et al. 2011). Only 5' RACE (as described in **Chapter 5.4.2**) would be able to join lincRNA sequences with these putative TSSs. However, the time investment required for this would be significant and I decided to progress directly to associating all possible promoters with nearby transposable element insertions.

This work suggests that P-element insertions are unlikely to be a good source of mutated lincRNA promoters. Despite a large number of insertions (13,140) and predicted promoters (249) being cross-referenced, only five putatively mutated promoters could be identified. Furthermore, of the four which were tested, only one showed an appreciable disruption in lincRNA expression, giving a 25% success rate. The Bloomington Stock Centre is the largest source of insertion lines in *D. melanogaster* and this low success rate indicates that querying the other stock centres in a similar way is unlikely to yield many, if any, more lincRNA promoter mutants.

The mutant (FBst0021094) carrying a mutation in the lincRNA EC069024 promoter could be investigated in more detail in the future. It would be important to define the structure of EC069024 using 5' and 3' RACE. Real-time RT-PCR could then be used to test whether this mutation is specific in disrupting the expression of this lincRNA, and not the neighbouring protein-coding genes *Smn* and *Rpn12*. If the mutation was specific, then this true lincRNA mutant could be examined for possible phenotypic consequences of this lesion, although nothing conspicuous was observed when handling these flies during the above experiments. This insertion also contains a UAS element, making it possible to further over-express this lincRNA by crossing these flies to a GAL4-carrying line. However, I do not think this should be a priority for future research. This lincRNA was annotated as fast-evolving in **Chapter 3**, but it does not overlap any short RNA sequence suggesting that it may not represent the primary transcript of these types of short RNA species. This lincRNA is also not supported by any gene models called using the RNA-seq evidence discussed in **Chapter 4**.

These findings indicate that reverse genetics should be used to study specific lincRNA loci only when more information is available regarding their function. For example, the RNA-seq information presented in **Chapter 4** proposes a number of functional hypotheses which could be tested explicitly by generating targeted mutants in the reduced number of lincRNAs that are predicted to display these functions. Alternatively, the genomic region surrounding these lincRNAs could be examined for previous insertion mutations, although these

are likely to have only a small chance of reducing lincRNA transcription. Naïve screening of large numbers of previously produced mutants for disruption in lincRNA expression and related phenotypes does not appear to be a prudent course of action at this time.

Chapter 7: CONCLUSIONS AND FUTURE PERSPECTIVES

The work presented in this thesis aimed to define, for the first time, the extent of lincRNA transcription from the *D. melanogaster* genome and, importantly, to what degree such lincRNAs are biologically relevant. This was done using both *in silico* genome-wide approaches and *in vitro* functional characterisation of a subset of these loci.

7.1 The *Drosophila* genome contains a large number of previously unrecognised lincRNA loci.

Using both EST (**Chapter 3**) and RNA-seq (**Chapter 4**) evidence, I annotated several thousand novel lincRNA loci in *D. melanogaster*. This confirms that, even in such a well-studied model organism as this, the genome annotation endeavour remains unfinished. My results have shown that there are many transcribed loci outside of the currently annotated gene set in *D. melanogaster* and which are worthy of further characterisation and study.

I found that, consistent with previous observations in mammals (Halasz et al. 2006), lincRNA loci are generally simpler and expressed in a more restricted manner than protein-coding genes. They are, on average, shorter than their protein-coding counterparts, while unbiased transcriptome analysis revealed that they often have fewer alternative isoforms and fewer introns. As

previously observed in mammals (Mortazavi et al. 2008), lincRNA loci are supported by fewer sequencing reads than protein-coding genes, which suggests that they are expressed at a much lower level, and that they contribute less to the overall transcriptional output of the cell (**Chapter 4.5.6**). This low expression coupled with no clear DNA sequence motifs may explain why this class of genes has previously escaped computational and experimental scrutiny (Griffiths-Jones 2007).

As noted in **Chapter 4**, however, there is little overlap between the lincRNAs defined using EST and RNA-seq evidence. A comparison of mouse lincRNAs obtained by cDNA sequencing and chromatin state mapping has revealed a similarly limited degree of loci detected by both approaches (Marques and Ponting 2009). This suggests that the current catalogues of lincRNAs remain incomplete. Most ESTs are isolated from specific tissues, such as the testes, while the RNA-seq data came from temporally-defined isolates of complete animals. It may be that the majority of those transcript models built from EST evidence are expressed in such a spatially restricted pattern that, when RNA is isolated from the complete animal at individual time points as was done when building the RNA-seq libraries, these transcripts are expressed at too low a level to be detectable. Alternatively, as the bias of predicted RNA secondary structures observed in the EST lincRNAs was not reproduced in those defined by RNA-seq, these two sets may represent two functionally distinct lincRNA subsets.

The extent of this incompleteness is likely to become apparent from the results of future sequencing experiments of the type described in **Chapter 4.** This should also reveal further lincRNA loci in *D. melanogaster* and how the two sets defined here are related. RNA-seq using libraries isolated from more specific developmental stages, individual tissues, or even from individual cell types (when combined with fluorescence-activated cell sorting (FACS), e.g. Brunskill et al. 2011) should discover further lincRNAs. In particular, this should identify those lincRNAs which are expressed in a more restricted profile than those identified here and hopefully validate many more of those defined by EST evidence. Similar sequencing experiments proposed by modENCODE in related species, such as *D. simulans* and *D. pseudoobscura*, will reveal how conserved this lincRNA transcription is across these species. Other modENCODE experiments, such as the genome-wide profiling of histone modifications H3K4me3 and H3K36me3 (Nègre et al. 2011), can be used to test if my lincRNA loci, like those defined in mammals, carry the same chromatin marks as their protein-coding gene counterparts.

7.2 The majority of *Drosophila* lincRNA loci contain several, but distinct, indicators of functionality.

I have used evolutionary constraint and several other predictors of functionality to investigate how many lincRNAs are likely to be biologically meaningful, and how many are likely to represent uncontrolled 'transcriptional noise'.

The vast majority of lincRNA loci defined here show clear evolutionary constraint between *D. melanogaster*, *D. simulans* and *D. yakuba*. These loci are also enriched for IPSs and MCSs – two other signatures of evolutionary constraint. It is likely that the mature RNA is important, rather than just the act of transcription from these loci, as those defined by ESTs are also enriched for RNA secondary structures. Constrained lincRNA loci are frequently found within euchromatin associated with regulated expression and in Polycomb-group protein-associated heterochromatin, suggesting that these loci may be involved in the regulation of developmental processes. This evolutionary constraint and frequent predicted secondary structures argue that both the primary sequence and secondary transcript structure are important for the function of these lincRNAs.

The RNA-seq data set also allowed me to define in which developmental time points lincRNAs were expressed. Using this information, I observed that the number of stages in which a lincRNA is expressed is positively correlated with its substitution rate: lincRNAs whose expression is more specifically restricted appear to also be more functionally constrained. Intriguingly, I identified an excess of male-specific lincRNAs. These loci did not share the increased substitution rate observed for male-specific protein-coding genes. Rather than being involved in sexually-selected processes, these lincRNAs may be involved in the development of male-specific tissues, such as the testis, or they could even be targets of transcription factors involved in sexual differentiation, such as *doublesex* (Rideout et al. 2010).

A second set of lincRNAs have evolved at a significantly faster rate than expected for neutral sequence, and this may be because they have experienced positive natural selection. Their apparent rapid evolution may be a consequence of poor alignments in these regions (Markova-Raina and Petrov 2011), or this may actually be driven by an evolutionary arms race with the transposable elements whose activity esiRNAs and piRNAs are known to suppress (**Chapter 1.1.2**). Arguing for their functionality is the observation that these lincRNA loci are relatively common in HP1-associated heterochromatin and are enriched in miRNA, esiRNA, and piRNA candidate sequences, all types of short RNA species. Both HP1-associated heterochromatin (Filion et al. 2010) and piRNA sequences (Yin and Lin 2007) are commonly found in the pericentric regions of the chromosome, so their joint enrichment is perhaps not surprising. This result suggests that these lincRNAs may represent the primary transcripts of these short RNAs. Those fast-evolving lincRNAs which do not overlap these sequences may in fact represent a previously unrecognised source of short RNAs.

As my results provide evidence that the majority of lincRNA loci in *Drosophila* are likely to impart some biological function, the central problem is to identify exactly what that function could be. The lincRNA catalogues defined here possess a number of other features which hint at the functions of some of these, and these can now be tested experimentally. Many lincRNAs are predicted to be involved in the regulation of specific developmental stages and a subset of these may be required to regulate correct male differentiation

or development. Detailed expression profiling of these lincRNAs should reveal in which tissues or cell populations they are transcribed, and these data could be used to inform experiments designed to disrupt expression of individual lincRNAs, as it would already be known where they are normally transcribed. The development of novel techniques specific to the study of lincRNAs, such as tools for their efficient purification, should also help to dissect lincRNA functionality *in vivo*. Constrained lincRNA loci were found to be enriched within heterochromatic regions bound by Polycomb-group proteins. Whether these Polycomb-group proteins are regulating expression of these lincRNAs could be tested by examining the expression of these lincRNAs in flies carrying a mutation within the Polycomb locus (e.g. Classen et al. 2009) and determining whether this is different to that observed in *wt* flies. Knockdown experiments which target individual fast-evolving lincRNAs and monitor the effect on gene (for potentially miRNA-harbouring lincRNAs) or transposable element (candidate esiRNA- and piRNA-harbouring lincRNAs) activity could be used to test how many of these sequences truly function in this way *in vivo*.

The results presented in **Chapter 6** imply that utilising the power of *Drosophila* to carry out naïve reverse genetics experiments by targeting putative lincRNA promoters is unlikely to have much power to reveal lincRNA function. Of a total of 249 predicted lincRNA core promoters, only five contained a mutation available from the Bloomington Stock Centre. Only one of these had an appreciable effect on the associated lincRNA which was, unexpectedly, to increase its expression. More success may be achieved by

studying the phenotypic effects of transposable elements carrying premature transcription terminal signals which, when inserted into lincRNA exons, should lead to a truncated RNA being transcribed. This approach has already been used to study the mouse lincRNA *Evf-2* (Bond et al. 2009).

7.3 What function(s) do individual *Drosophila* lincRNAs possess?

In this thesis, I have also presented work which experimentally tested the possible functionality of individual lincRNA loci. In **Chapter 5**, I investigated the potential for lincRNAs to be involved in the regulation of the expression of genomically adjacent protein-coding genes (a mechanism I term *cis*-regulation). Evolutionarily constrained, non-embryonic lincRNAs are enriched in the neighbourhoods of protein-coding genes involved in transcriptional regulation. Expression profiling revealed a number of lincRNA loci whose expression appears to be regulated. Regulated expression is often cited as evidence of functionality (Ravasi et al. 2006) and, although the function(s) of these lincRNAs remains unclear, their expression profiles can be considered as further evidence of their functionality which could be studied in future experiments.

I investigated the function of one of four lincRNAs which shared a ubiquitous expression profile with their adjacent transcription factor gene. This lincRNA locus, which I named *dEvf-2*, is genomically adjacent to and positively regulates the expression of the transcription factor *Dll*. Two *Dll* orthologues in

mouse, *Dlx5* and *Dlx6*, are also positively regulated in cell culture by a neighbouring lincRNA locus known as *Evf-2* (Feng et al. 2006). However, in a mouse model, *Evf-2* appears to only positively regulate *Dlx5* while simultaneously suppressing *Dlx6* expression (Bond et al. 2009). In *Drosophila*, *dEvf-2* positively regulates expression of the adjacent protein-coding gene *Dll* in S2 cell culture. It remains to be seen what effect, if any, *dEvf-2* has in the fly. For this purpose, a UAS-controlled RNAi stock is currently being generated. Other experiments proposed in **Chapter 5** to investigate *dEvf-2* function and, in particular, how it is related to the mammalian *Evf-2* are a current priority. Due to the long evolutionary distance separating *Evf-2* and *dEvf-2* and the general lack of conservation observed in lincRNA sequences (Pang et al. 2006), it cannot be demonstrated that these lincRNAs in mouse and *Drosophila* share a common evolutionary ancestor and they are therefore considered as analogous. This is the first and, to my knowledge, only example of such an analogous lincRNA shared between such distantly related species.

Specifically using the RNA-seq data presented in **Chapter 4**, I tested whether this type of analogy could be a more general feature of lincRNAs. I observed an excess of lincRNAs in gene neighbourhoods whose orthologous neighbourhood in mouse also contain a lincRNA, which suggests that there may be many more examples of lincRNA analogy, like *dEvf-2*/*Evf-2*, waiting to be discovered. All lincRNA/protein-coding gene pairs in mouse and *Drosophila* make particularly interesting experimental candidates and several of these are being actively researched. In particular, the mouse lincRNA

AK032637 has already been demonstrated to negatively regulate expression of the genomically adjacent transcription factor Pax6, in N2A cells (Keith Vance, personal communication). The expression profile of the analogous *Drosophila* lincRNA, *lincRNA.927*, has been analysed by real-time RT-PCR, where it appears to be expressed constitutively, but at variable levels throughout development. A fly expressing an RNAi construct targeting *lincRNA.927* is currently in preparation to test whether it similarly regulates *toy*, a *Drosophila Pax6* orthologue (Charlotte Tibbit, personal communication).

7.4 Concluding Remarks

This thesis has demonstrated that large sets of functional lincRNAs are not a largely mammalian-specific phenomenon, but that they can also be found in a less complex model organism, namely the fruit fly *D. melanogaster*. The mouse genome is already known to contain more lincRNA loci than *D. melanogaster*, and estimates of this are likely to increase upon further RNA-seq interrogation of the mouse transcriptome. Further studies like this one in a range of model and non-model organisms will reveal how important lincRNAs are in different genomes and therefore how much they can explain differences between these species.

The presence of many lincRNAs in *D. melanogaster* opens an exciting avenue of research using this organism (**Chapter 1.5.2**) as a model for future functional characterisation of lincRNAs on a scale previously impossible in mammalian systems. The analyses presented here suggest that a large number

of these lincRNA loci are likely to encode functional RNAs and I have proposed a number of functional hypotheses which can be tested. As I have now shown that *Drosophila* lincRNA loci share a number of features with mammalian loci, the full power of *Drosophila* can now be harnessed as a model for studying lincRNA behaviour and, in several instances, such as that of *dEvf-2*, as a model of specific individual lincRNAs. The work that has formed this thesis provides the necessary foundations to justify such future experiments, and I await the results of these over the next few years with anticipation.

REFERENCES

Adams MD, Soares MB, Kerlavage AR, Fields C, and Venter JC. 1993. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet* **4**: 373-380.

Ahsan B, Saito TL, Hashimoto S.-I., Muramatsu K, Tsuda M, Sasaki A, Matsushima K, Aigaki T, and Morishita S. 2009. MachiBase: A Drosophila melanogaster 5'-end mRNA transcription database. *Nucleic Acids Res.* **37**.

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**: 927-935.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P. 2002. *Molecular Biology of the Cell*. 4th ed. Garland Science.

Altenhoff AM, and Dessimoz C. 2009. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Comput Biol* **5**: e1000262.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman D J. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol* **215**: 403-410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman D J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Amaral PP, and Mattick J S. 2008. Noncoding RNA in development. *Mamm. Genome* **19**: 454-492.

Amrein H, and Axel R. 1997. Genes expressed in neurons of adult male Drosophila. *Cell* **88**: 459-469.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437**: 1149-1152.

Aravin AA, Hannon GJ, and Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**: 761-764.

Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M., Marks D., Snyder B, Gaasterland T, Meyer J, and Tuschl T. 2003. The small RNA profile during Drosophila melanogaster development. *Dev. Cell* **5**: 337-350.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet* **25**: 25-29.

Babak T, Blencowe BJ, and Hughes TR. 2005. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics* **6**: 104.

Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig A, et al. 2007. RNAs

everywhere: Genome-wide annotation of structured RNAs. *J. Exp. Zool. Part B Mol. Dev. Evol.* **308**: 1-25.

van Bakel H, Nislow C, Blencowe BJ, and Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**: e1000371.

van Bakel H, and Hughes TR. 2009. Establishing legitimacy and function in the new transcriptome. *Brief Funct Genomic Proteomic* **8**: 424-436.

Bao S, Jiang R, Kwan W, Wang B, Ma X, and Song Y-Q. 2011. Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet* **56**: 406-414.

Barciszewski J, and Erdmann VA. 2003. *Noncoding RNAs: molecular biology and molecular medicine*. Springer.

Begun DJ, Holloway AK, Stevens K, Hillier L W, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CD, et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PloS Biol.* **5**: 2534-2559.

Bellen HJ, Levis RW, Batty G, He Y, Carlson J W, Tsang G, Evans-Holm M, Hiesinger PR, Schulze KL, Rubin GM, et al. 2004. The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. *Genetics* **167**: 761-781.

Beltran M, Puig I, Peña C, García JM, Alvarez AB, Peña R, Bonilla F, and de Herreros AG. 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**: 756-769.

Benjamini Y, and Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289-300.

Benson DA, Karsch-Mizrachi I, Lipman David J., Ostell J, and Wheeler DL. 2005. GenBank. *Nucleic Acids Res* **33**: D34-D38.

Berglund A-C, Sjolund E, Ostlund G., and Sonnhammer E. L. L. 2007. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research* **36**: D263-D266.

Bergman CM, Quesneville H, Anxolabéhère D, and Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol* **7**: R112.

Bergman CM, and Kreitman M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11**: 1335-1345.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn J L, Tongprasit W, Samanta M, Weissman S, et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**: 2242-2246.

Birney E, Stamatoyannopoulos J A, Dutta A, Guigó R, Gingeras T R, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman R E, et

al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.

Björklund M, Taipale M, Varjosalo M, Saharinen J, Lahdenperä J, and Taipale J. 2006. Identification of pathways regulating cell size and cell-cycle progression by RNAi. *Nature* **439**: 1009-1013.

Blencowe BJ. 2002. Transcription: surprising role for an elusive small nuclear RNA. *Curr. Biol* **12**: R147-149.

De Bona F, Ossowski S, Schneeberger K, and Rätsch G. 2008. Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**: i174-180.

Bond AM, Vangompel MJW, Sametsky EA, Clark MF, Savage JC, Disterhoft JF, and Kohtz JD. 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat. Neurosci* **12**: 1020-1027.

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

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, and Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**: 1089-1103.

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, and Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**: 1387-1392.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, and Rastan S. 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**: 515-526.

Brunskill E, Lai H, Jamison DC, Potter SS, and Patterson L. 2011. Microarrays and RNA-Seq identify molecular mechanisms driving the end of nephron production. *BMC Developmental Biology* **11**: 15.

Bulger M, and Groudine M. 2010. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol* **339**: 250-257.

Burrows M, and Wheeler DJ. 1994. A block-sorting lossless data compression algorithm. *Digital Equipment Corporation* **Technical Report 124**.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B., Wells C, et al. 2005. The Transcriptional Landscape of the Mammalian Genome. *Science* **309**: 1559-1563.

Carroll L. 1865. *Alice in Wonderland*. Macmillan.

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

Celniker S E, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen G H, Kellis M, Lai EC, Lieb JD, MacAlpine D M, et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927-930.

Celniker S E, and Rubin GM. 2003. The Drosophila melanogaster genome. *Annu Rev Genomics Hum Genet* **4**: 89-117.

Charlesworth B, Sniegowski P, and Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**: 215-220.

Chen F-C, Chen C-J, Li Wen-Hsiung, and Chuang T-J. 2007. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res* **17**: 16-22.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**: 1149-1154.

Chiaromonte F, Y, and Miller W. 2002. Scoring pairwise genomic sequence alignments. *Pacific Symposium on Biocomputing* 115-126.

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnár Z, and Ponting C P. 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol* **11**: R72.

Church DM, Goodstadt Leo, Hillier Ladeana W, Zody Michael C, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**: e1000112.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman T C, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203-218.

Classen A-K, Bunker BD, Harvey KF, Vaccari T, and Bilder D. 2009. A tumor suppressor activity of Drosophila Polycomb genes mediated by JAK-STAT signaling. *Nat Genet* **41**: 1150-1155.

Claude A. 1937. Preparation of an Active Agent from Inactive Tumour Extracts. *Science* **85**: 294-295.

Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**: 613-619.

Cooley L, Kelley R, and Spradling A. 1988. Insertional mutagenesis of the Drosophila genome with single P elements. *Science* **239**: 1121-1128.

Costa V, Angelini C, De Feis I, and Ciccodicola A. 2010. Uncovering the Complexity of Transcriptomes with RNA-Seq. *J Biomed Biotechnol* **2010**.

Crouse HV. 1960. The Controlling Element in Sex Chromosome Behavior in Sciara. *Genetics* **45**: 1429-1443.

Cunnington MS, Santibanez Koref M, Mayosi BM, Burn J, and Keavney B. 2010. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. *PLoS Genet* **6**: e1000899.

Czech B, Malone D, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. 2008. An endogenous small interfering RNA pathway in Drosophila. *Nature* **453**: 798-802.

Daines B, Wang H, Wang L, Li Y, Han Y, Emmert D, Gelbart W, Wang X, Li W, Gibbs R, et al. 2011. The Drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome Res* **21**: 315-324.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, and Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences* **103**: 5320 -5325.

Deng X, and Meller VH. 2006. Non-coding RNA in fly dosage compensation. *Trends Biochem. Sci.* **31**: 526-532.

Dinger M E, Amara PP, Mercer TR, Pang K C, Bruce S J, Gardiner BB, Askarian-Amiri MME, Ru K, Soldà G, Simons C, et al. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18**: 1433-1445.

Dubchak I, and Frazer K. 2003. Multi-species sequence comparison: the next frontier in genome annotation. *Genome Biol* **4**: 122-122.

Duncan DM, Burgess EA, and Duncan I. 1998. Control of distal antennal identity and tarsal development in Drosophila by spineless-aristapedia, a homolog of the mammalian dioxin receptor. *Genes Dev* **12**: 1290-1303.

Eddy S.R. 2002. Computational genomics of noncoding RNA genes. *Cell* **109**: 137-140.

Eissenberg JC, and Elgin SC. 2000. The HP1 protein family: getting a grip on chromatin. *Curr. Opin. Genet. Dev* **10**: 204-210.

Eyre-Walker A, and Hurst LD. 2001. The evolution of isochores. *Nat. Rev. Genet* **2**: 549-555.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol* **17**: 368-376.

Feng J, Bi C, Clark BS, Mady R, Shah P, and Kohtz JD. 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.

Ferragina P, and Manzini G. 2000. Opportunistic Data Structures with Applications. *Proc FOCS* 390--398.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**: 212-224.

Gaffney DJ, and Keightley PD. 2006. Genomic Selective Constraints in Murid Noncoding DNA. *PLoS Genet* **2**.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy Sean R, et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**: D136-140.

Gebelein B, Culi J, Ryoo HD, Zhang W, and Mann RS. 2002. Specificity of Distalless Repression and Limb Primordia Development by Abdominal Hox Proteins. *Developmental Cell* **3**: 487-498.

Gelbart ME, Larschan E, Peng S, Park P J, and Kuroda M I. 2009. Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. *Nat Struct Mol Biol* **16**: 825-832.

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science* **320**: 1077-1081.

Ghildiyal M, and Zamore PD. 2009. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet* **10**: 94-108.

González J, Lenkov K, Lipatov M, Macpherson JM, and Petrov DA. 2008. High rate of recent transposable element-induced adaptation in Drosophila melanogaster. *PloS Biol.* **6**: 2109-2129.

Graveley B R, Brooks AN, Carlson J W, Duff M O, Landolin JM, Yang L, Artieri C G, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**: 473-479.

Griffiths-Jones S. 2007. Annotating Noncoding RNA Genes. *Annu. Rev. Genom. Human Genet.* **8**: 279-298.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**: 223-227.

Guttman Mitchell, Garber Manuel, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum Chad, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol* **28**: 503-510.

Haddrill PR, Charlesworth B, Halligan DL, and Andolfatto P. 2005. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. *Genome Biol.* **6**: R67.

Haerty W, Jagadeeshan S, Kulathinal Rob J., Wong A, Ravi Ram K, Sirot LK, Levesque L, Artieri Carlo G., Wolfner MF, Civetta A, et al. 2007. Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila. *Genetics* **177**: 1321-1335.

Halasz G, van Batenburg MF, Perusse J, Hua S, Lu X-J, White K P, and Bussemaker HJ. 2006. Detecting transcriptionally active regions using genomic tiling arrays. *Genome Biol* **7**: R59.

Hamada FN, Park P J, Gordadze PR, and Kuroda M I. 2005. Global regulation of X chromosomal genes by the MSL complex in Drosophila melanogaster. *Genes Dev* **19**: 2289-2294.

Hannon GJ. 2002. RNA interference. *Nature* **418**: 244-251.

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13**: 13-26.

Hasegawa M, Kishino H, and Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol* **22**: 160-174.

Hashimoto S-I, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, and Matsushima K. 2004. 5'-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22**: 1146-1149.

He L, and Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet* **5**: 522-531.

Heger A, and Ponting C P. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* **17**: 1837-1849.

Heller MJ. 2002. DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications. *Annu. Rev. Biomed. Eng.* **4**: 129-153.

Hiller M, Findeiss S, Lein S, Marz M, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G, et al. 2009. Conserved introns reveal novel transcripts in Drosophila melanogaster. *Genome Res* **19**: 1289-1300.

Hogga I, and 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 PM, Abramson RD, Watson R, and Gelfand DH. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'-3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A* **88**: 7276-7280.

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, and Zamir A. 1965. Structure of a Ribonucleic Acid. *Science* **147**: 1462-1465.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**: 129-149.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**: 931-949.

Hosack DA, Dennis G, Sherman BT, Lane HC, and Lempicki RA. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol* **4**: R70.

Hoskins R, Landolin J, Brown J, Sandler J, Takahashi H, Lassmann T, Yu C, Booth B, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**: 182-192.

Hulsen T, Huynen MA, de Vlieg J, and Groenen PM. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7**: R31-R31.

Hüttenhofer A, Schattner P, and Polacek N. 2005. Non-coding RNAs: Hope or hype? *Trends Genet.* **21**: 289-297.

Inagaki S, Numata K, Kondo T, Tomita M, Yasuda K, Kanai A, and Kageyama Y. 2005. Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila. *Genes Cells* **10**: 1163-1173.

Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, et al. 2003. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**: 8031-8041.

Jiang Z-F, Croshaw DA, Wang Y, Hey J, and Machado CA. 2011. Enrichment of mRNA-like Noncoding RNAs in the Divergence of Drosophila Males. *Mol. Biol. Evol* **28**: 1339-1348.

Johnson JM, Edwards S, Shoemaker D, and Schadt EE. 2005. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21**: 93-102.

Jolly C, and Lakhotia SC. 2006. Human sat III and Drosophila hsr omega transcripts: a common paradigm for regulation of nuclear RNA processing in stressed cells. *Nucleic Acids Res* **34**: 5508-5514.

Jukes T, and Cantor C. 1969. *Evolution of Protein Molecules*. Academy Press.

Jung C-H, Hansen MA, Makunin IV, Korbie DJ, and Mattick J S. 2010. Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. *BMC Genomics* **11**: 77.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Aerts S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**: 331-342.

Kawamura Y, Saito K, Kin T, Ono Y, Asai K, Sunohara T, Okada TN, Siomi MC, and Siomi H. 2008. Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* **453**: 793-797.

Kelley RL, Meller VH, Gordadze PR, Roman G, Davis RL, and Kuroda M I. 1999. Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. *Cell* **98**: 513-522.

Kent WJ. 2002. BLAT-the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Research* **12**: 996 -1006.

Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, and Pääbo S. 2006. Functionality of intergenic transcription: an evolutionary comparison. *PLoS Genet* **2**: e171.

Khalil AM, Guttman Mitchell, Huarte Maite, Garber Manuel, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein Bradley E, van Oudenaarden A, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A* **106**: 11667-11672.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan Erica, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**: 480-485.

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread

transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182-187.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol* **16**: 111-120.
Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, and Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**.
Kozomara A, and Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**: D152-157.
Lai EC. 2003. microRNAs: Runts of the Genome Assert Themselves. *Current Biology* **13**: R925-R936.
Lakhotia SC, and Sharma A. 1996. The 93D (hsr-omega) locus of Drosophila: non-coding gene with house-keeping functions. *Genetica* **97**: 339-348.
Lander E S, Linton LM, Birren B, Nusbaum C, Zody M C, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
Langmead B, Trapnell C, Pop M, and Salzberg S L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**.
Lee C-Y, Wilkinson BD, Siegrist SE, Wharton RP, and Doe CQ. 2006. Brat is a Miranda cargo protein that promotes neuronal differentiation and inhibits neuroblast self-renewal. *Dev. Cell* **10**: 441-449.
Lee JT, Davidow LS, and Warshawsky D. 1999. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet* **21**: 400-404.
Lee RC, Feinbaum RL, and Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843-854.
Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**: 168-176.
Lempradl A, and Ringrose L. 2008. How does noncoding transcription regulate Hox genes? *Bioessays* **30**: 110-121.
Lewin B. 2003. *Genes VIII*. 1st ed. Pearson Education.
Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM, et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**: 509-521.
Li H, Ruan J, and Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851-1858.
Li R, Li Y, Kristiansen K, and Wang J. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**: 713-714.
Li Wen-Hsiung. 1997. *Molecular Evolution*. Sinauer Associates.
Li Z, Liu M, Zhang L, Zhang W, Gao G, Zhu Z, Wei L, Fan Q, and Long M. 2009. Detection of intergenic non-coding RNAs expressed in the main

developmental stages in Drosophila melanogaster. *Nucleic Acids Res* **37**: 4308-4314.

Lin R, Maeda S, Liu C, Karin M, and Edgington TS. 2007. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* **26**: 851-858.

Liu J, Gough J, and Rost B. 2006. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet.* **2**.

Liu J-L, and Gall JG. 2007. U bodies are cytoplasmic structures that contain uridine-rich small nuclear ribonucleoproteins and associate with P bodies. *Proc. Natl. Acad. Sci. U. S. A.* **104**: 11655-11659.

Lu J, Fu Y, Kumar S, Shen Y, Zeng K, Xu A, Carthew R, and Wu C-I. 2008. Adaptive evolution of newly emerged micro-RNA genes in Drosophila. *Mol. Biol. Evol.* **25**: 929-938.

Lunter G, Ponting C P, and Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* **2**.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, and Hein J. 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Res.* **18**: 298-309.

MacIntosh GC, Wilkerson C, and Green PJ. 2001. Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol* **127**: 765-776.

Mackay TFC. 1986. Transposable Element-Induced Fitness Mutations in Drosophila Melanogaster. *Genetics Research* **48**: 77-87.

Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, Cheng J, Bell I, Ghosh S, Piccolboni A, et al. 2006. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat. Genet.* **38**: 1151-1158.

Marco A, Hui JHL, Ronshaugen M, and Griffiths-Jones S. 2010. Functional shifts in insect microRNA evolution. *Genome Biol Evol* **2**: 686-696.

Mardis ER. 2007. ChIP-seq: welcome to the new frontier. *Nat Meth* **4**: 613-614.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387-402.

Marguerat S, and Bähler J. 2010. RNA-seq: from technology to biology. *Cell. Mol. Life Sci* **67**: 569-579.

Markova-Raina P, and Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res* **21**: 863-874.

Marques A C, and Ponting C P. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10**: R124.

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

Mattick J S. 2003. Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* **25**: 930-939.

McManus CJ, Coolon JD, Duff Michael O, Eipper-Mains J, Graveley Brenton R, and Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**: 816-825.

Meader S, Ponting C P, and Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20**: 1335-1343.

Meller VH, Wu KH, Roman G, Kuroda M I, and Davis RL. 1997. roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. *Cell* **88**: 445-457.

Meller VH, and Rattner BP. 2002. The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *EMBO J* **21**: 1084-1091.

Mercer TR, Dinger M E, Sunkin SM, Mehler MF, and Mattick J S. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U. S. A.* **105**: 716-721.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet* **11**: 31-46.

Mikkelsen Tarjei S, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T-K, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**: 553-560.

Miyoshi K, Miyoshi T, Hartig JV, Siomi H, and Siomi MC. 2010. Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in Drosophila. *RNA* **16**: 506-515.

Mohler J, and Pardue ML. 1982. Deficiency mapping of the 93D heat-shock locus in Drosophila melanogaster. *Chromosoma* **86**: 457-467.

Monk M, and Harper MI. 1979. Sequential X chromosome inactivation coupled with cellular differentiation in early mouse embryos. *Nature* **281**: 311-313.

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

Mukherjee AS, and Beermann W. 1965. Synthesis of ribonucleic acid by the X-chromosomes of Drosophila melanogaster and the problem of dosage compensation. *Nature* **207**: 785-786.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of Drosophila. *Science* **287**: 2196-2204.

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, and Fraser P. 2008. The Air Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin. *Science* **322**: 1717 - 1720.

Navarro P, Pichard S, Ciaudo C, Avner P, and Rougeulle C. 2005. Tsix transcription across the Xist gene alters chromatin conformation

without affecting Xist transcription: implications for X-chromosome inactivation. *Genes Dev* **19**: 1474-1484.

Nesterova TB, Slobodyanyuk SY, Elisaphenko EA, Shevchenko AI, Johnston C, Pavlova ME, Rogozin IB, Kolesnikov NN, Brockdorff N, and Zakian SM. 2001. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res* **11**: 833-849.

Nguyen VT, Kiss T, Michels AA, and Bensaude O. 2001. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414**: 322-325.

Nishida H, Miyagawa S, Vieux-Rochas M, Morini M, Ogino Y, Suzuki K, Nakagata N, Choi H-S, Levi G, and Yamada G. 2008. Positive regulation of steroidogenic acute regulatory protein gene expression through the interaction between Dlx and GATA-4 for testicular steroidogenesis. *Endocrinology* **149**: 2090-2097.

Numata K, Okada Y, Saito R, Kiyosawa H, Kanai A, and Tomita M. 2007. Comparative analysis of cis-encoded antisense RNAs in eukaryotes. *Gene* **392**: 134-141.

Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. 2011. A cis-regulatory map of the Drosophila genome. *Nature* **471**: 527-531.

Ogawa Y, and Lee JT. 2002. Antisense regulation in X inactivation and autosomal imprinting. *Cytogenet. Genome Res* **99**: 59-65.

Ohhata T, Hoki Y, Sasaki H, and Sado T. 2008. Crucial role of antisense transcription across the Xist promoter in Tsix-mediated Xist chromatin modification. *Development* **135**: 227-235.

Ohler U, Liao G-C, Niemann H, and Rubin GM. 2002. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**: RESEARCH0087.

Ohler U. 2006. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res.* **34**: 5943-5950.

Okamura K, Ishizuka A, Siomi H, and Siomi MC. 2004. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes & Development* **18**: 1655 -1666.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563-573.

Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**: 46-58.

Ørom UA, and Shiekhattar R. 2011. Long non-coding RNAs and enhancers. *Curr Opin Genet Dev* **21**: 194-198.

Ostlund Gabriel, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, and Sonnhammer Erik L L. 2010. InParanoid 7: new

algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**: D196-203.

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet* **36**: 40-45.

Oudejans CB, Westerman B, Wouters D, Gooyer S, Leegwater PA, van Wijk IJ, and Sleutels F. 2001. Allelic IGF2R repression does not correlate with expression of antisense RNA in human extraembryonic tissues. *Genomics* **73**: 331-337.

Pang KC, Frith MC, and Mattick J S. 2006. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet.* **22**: 1-5.

Pang KC., Stephen S, Dinger Marcel E., Engström PG, Lenhard Boris, and Mattick JS. 2007. RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res* **35**: D178-D182.

Panganiban G, and Rubenstein JLR. 2002. Developmental functions of the Distal-less/Dlx homeobox genes. *Development* **129**: 4371-4386.

Pasmant E, Laurendeau I, Héron D, Vidaud M, Vidaud D, and Bièche I. 2007. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res* **67**: 3963-3969.

Pasmant E, Sabbagh A, Vidaud M, and Bièche I. 2011. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* **25**: 444-448.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander E S, Kent J, Miller W, and Haussler D. 2006. *PLoS Comput. Biol.* **2**: 251-262.

Penny GD, Kay GF, Sheardown SA, Rastan S, and Brockdorff N. 1996. Requirement for Xist in X chromosome inactivation. *Nature* **379**: 131-137.

Perkins DO, Jeffries C, and Sullivan P. 2005. Expanding the "central dogma": The regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia. *Mol. Psychiatry* **10**: 69-78.

Peters L, and Meister G. 2007. Argonaute Proteins: Mediators of RNA Silencing. *Molecular Cell* **26**: 611-623.

Petruk S, Sedkov Y, Brock HW, and Mazo A. 2007. A model for initiation of mosaic HOX gene expression patterns by non-coding RNAs in early embryos. *RNA Biology* **4**: 1-6.

Pfeffer S, Zavolan Mihaela, Grässer FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks Debora, Sander C, et al. 2004. Identification of virus-encoded microRNAs. *Science* **304**: 734-736.

Phillippy AM, Schatz MC, and Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol* **9**: R55.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, and Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**: 1033-1038.

Pollard DA, Bergman CM, Stoye J, Celniker S.E., and Eisen MB. 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinform.* **5**.

Ponjavic J, Oliver PL, Lunter G, and Ponting C P. 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* **5**: e1000617.

Ponjavic J, Ponting C P, and Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**: 556-565.

Ponjavic J, and Ponting C P. 2007. The long and the short of RNA maps. *BioEssays* **29**: 1077-1080.

Ponting C P, Oliver PL, and Reik W. 2009. Evolution and Functions of Long Noncoding RNAs. *Cell* **136**: 629-641.

Ponting C P, and Belgard TG. 2010. Transcribed dark matter: meaning or myth? *Hum. Mol. Genet* **19**: R162-168.

Ponting C. 2008. The functional repertoires of metazoan genomes. *Nat. Rev. Genet* **9**: 689-698.

Ponting CP, and Hardison R. 2011. What fraction of the human genome is functional? *Genome Res*.

Powell JR. 1997. *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, USA.

Prasanth K V, Rajendra TK, Lal AK, and Lakhotia SC. 2000. Omega speckles - a novel class of nuclear speckles containing hnRNPs associated with noncoding hsr-omega RNA in Drosophila. *J. Cell. Sci* **113 Pt 19**: 3485-3497.

Prasanth K V, and Spector DL. 2007. Eukaryotic regulatory RNAs: An answer to the "genome complexity" conundrum. *Genes Dev.* **21**: 11-42.

Putney SD, Herlihy WC, and Schimmel P. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302**: 718-721.

Qureshi IA, Mattick J S, and Mehler MF. 2010. Long non-coding RNAs in nervous system function and disease. *Brain Res* **1338**: 20-35.

Rach EA, Yuan H-Y, Majoros WH, Tomancak P, and Ohler U. 2009. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol* **10**: R73.

Ravasi T, Suzuki H, Pang K.C., Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM, et al. 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**: 11-19.

Reese MG. 2001. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Computers & Chemistry* **26**: 51-56.

Reiss D, Josse T, Anxolabéhère D, and Ronsseray S. 2004. aubergine mutations in Drosophila melanogaster impair P cytotype determination by telomeric P elements inserted in heterochromatin. *Mol. Genet. Genomics* **272**: 336-343.

Rhee SY, Wood V, Dolinski K, and Draghici S. 2008. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet* **9**: 509-515.

Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, Bucher G, et al. 2008. The genome of the model beetle and pest Tribolium castaneum. *Nature* **452**: 949-955.

Rideout EJ, Dornan AJ, Neville MC, Eadie S, and Goodwin SF. 2010. Control of sexual differentiation and behavior by the doublesex gene in Drosophila melanogaster. *Nat. Neurosci* **13**: 458-466.

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, Cobianchi F, Riva S, and Biamonti G. 2004. Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. *Mol. Biol. Cell* **15**: 543-551.

Rodriguez A, Griffiths-Jones S, Ashurst JL, and Bradley A. 2004. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research* **14**: 1902 -1910.

Rodriguez C, Borgel J, Court F, Cathala G, Forné T, and Piette J. 2010. CTCF is a DNA methylation-sensitive positive regulator of the INK/ARF locus. *Biochem. Biophys. Res. Commun* **392**: 129-134.

Ronsseray S, Lehmann M, Nouaud D, and Anxolabéhère D. 1996. The Regulatory Properties of Autonomous Subtelomeric P Elements Are Sensitive to a Suppressor of Variegation in Drosophila Melanogaster. *Genetics* **143**: 1663-1674.

Rose D, Hackermüller J, Washietl S, Reiche K, Hertel J, Findeiss S, Stadler PF, and Prohaska SJ. 2007. Computational RNomics of drosophilids. *BMC Genomics* **8**.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, and Lai EC. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Research* **17**: 1850 -1864.

Sackton TB, Kulathinal R J, Bergman CM, Quinlan AR, Dopman EB, Carneiro M, Marth GT, Hartl DL, and Clark AG. 2009. Population Genomic Inferences from Sparse High-Throughput Sequencing of Two Populations of Drosophila melanogaster. *Genome Biology and Evolution* **1**: 449 -465.

Salemi M, and Vandamme A-M. 2003. *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press.

Sanchez-Elsner T, Gou D, Kremmer E, and Sauer F. 2006. Noncoding RNAs of trithorax response elements recruit Drosophila ash1 to Ultrabithorax. *Science* **311**: 1118-1123.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei C-L, and Natoli G. 2010. A large fraction of

extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**: e1000384.

Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, and Cavalli G. 2007. Genome Regulation by Polycomb and Trithorax Proteins. *Cell* **128**: 735-745.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13**: 103-107.

Seemann S, Gilchrist M, Hofacker I, Stadler P, and Gorodkin J. 2007. Detection of RNA structures in porcine EST data and related mammals. *BMC Genomics* **8**: 316.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Abe K, Clawson H, Spieth J, Hillier L W, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.

Slater GSC. 2000. Algorithms for the Analysis of Expressed Sequence Tags. University of Cambridge, Cambridge.

Sleutels F, Zwart R, and Barlow DP. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**: 810-813.

Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D, Matthews L, Rogers J, Pask AJ, Shaw G, VandeBerg JL, et al. 2008. Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. *Nat Genet* **40**: 971-976.

Sparmann A, and van Lohuizen M. 2006. Polycomb silencers control cell fate, development and cancer. *Nat. Rev. Cancer* **6**: 846-856.

Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, and Rubin GM. 1995. Gene disruptions using P transposable elements: An integral component of the Drosophila genome project. *PROC. NATL. ACAD. SCI. U. S. A.* **92**: 10824-10830.

Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, Misra S, and Rubin GM. 1999. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics* **153**: 135-177.

Stanley SM, Bailey TL, and Mattick J S. 2006. GONOME: measuring correlations between GO terms and genomic positions. *BMC Bioinformatics* **7**: 94-94.

Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, et al. 2002. A Drosophila full-length cDNA resource. *Genome Biol* **3**: RESEARCH0080.

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, and Kellis M. 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.* **17**: 1865-1879.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson J W, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**: 219-232.

van Steensel B, and Henikoff S. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol* **18**: 424-428.

Stolc V, Gauhar Z, Mason C, Halasz G, Van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, et al. 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* **306**: 655-660.

Storz G. 2002. An expanding universe of noncoding RNAs. *Science* **296**: 1260-1263.

Straub T, Gilfillan GD, Maier VK, and Becker PB. 2005. The Drosophila MSL complex activates the transcription of target genes. *Genes Dev* **19**: 2284-2288.

The UniProt Consortium. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**: D142-148.

Trapnell C, Pachter L, and Salzberg S L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105 -1111.

Trapnell C, Williams B A, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg S L, Wold BJ, and Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol* **28**: 511-515.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma Alok, Bubulya PA, et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39**: 925-938.

Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, Misra S, Celniker S E, and Rubin GM. 2005. Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **102**: 5495-5500.

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**: D555-559.

Vachon G, Cohen B, Pfeifle C, McGuffin ME, Botas J, and Cohen SM. 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.

van Valen L. 1973. A New Evolutionary Law. *Evolutionary Theory* **1**: 1-30.

Venken KJT, and Bellen HJ. 2005. Emerging technologies for gene manipulation in Drosophila melanogaster. *Nat. Rev. Gen.* **6**: 167-178.

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, and Hunkapiller M. 1998. Shotgun sequencing of the human genome. *Science* **280**: 1540-1542.

Visel A, Zhu Y, May D, Afzal V, Gong E, Attanasio C, Blow MJ, Cohen JC, Rubin EM, and Pennacchio LA. 2010. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**: 409-412.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang Lu, Mayr C, Kingsmore SF, Schroth GP, and Burge CB. 2008. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **456**: 470-476.

Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, and Wong GJ. 2004. Mouse transcriptome: neutral evolution of "non-coding" complementary DNAs. *Nature* **431**.

Washietl S, Hofacker IL, and Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **102**: 2454-2459.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* **17**: 852-864.

Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, and Peterson KJ. 2009. The deep evolution of metazoan microRNAs. *Evolution & Development* **11**: 50-68.

Wilhelm BT, Marguerat S, Goodhead I, and Bähler J. 2010. Defining transcribed regions using RNA-seq. *Nat Protoc* **5**: 255-266.

Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, and Schultz PG. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**: 1570-1573.

Wolpert L, Smith J, Jessell T, Lawrence P, Robertson E, and Meyerowitz E. 2006. *Principles of Development*. 3rd ed. OUP Oxford.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2004. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biol* **3**: e7.

Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, and Barlow DP. 1997. Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* **389**: 745-749.

Wyers F, Rougemaille M, Badis G, Rousselle J-C, Dufour M-E, Boulay J, Régnault B, Devaux F, Namane A, Séraphin B, et al. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**: 725-737.

Xue C, and Li F. 2008. Finding noncoding RNA transcripts from low abundance expressed sequence tags. *Cell Res.* **18**: 695-700.

Yang Y, Li Z, Fan Q, Long M, and Zhang W. 2007. Significant divergence of sex-related non-coding RNA expression patterns among closely related species in Drosophila. *Chin. Sci. Bull.* **52**: 748-754.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol* **39**: 306-314.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**: 1586 -1591.

Yang Z, Zhu Q, Luo K, and Zhou Q. 2001. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414**: 317-322.

Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ, and Zhou M-M. 2010. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* **38**: 662-674.

Yin H, and Lin H. 2007. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature* **450**: 304-308.

Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, and Cui H. 2008. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**: 202-206.

Zhang L, and Li W-H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol* **21**: 236-239.

Zhang Z, Harrison PM, Liu Y, and Gerstein M. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13**: 2541-2558.

Zieve G, and Penman S. 1976. Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell* **8**: 19-31.

Zipper H, Brunner H, Bernhagen J, and Vitzthum F. 2004. Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic Acids Res* **32**: e103.

APPENDIX A: SUMMARY OF SHORT READ ASSEMBLY PIPELINE

Stage	Number of reads	Number of reads mapped	Number of transcripts	Mean read depth (fold coverage)
Embyros, 0-2 hour after egg laying	108,437,006	70,549,513	17,225	44.19
Embyros, 2-4 hours after egg laying	83,209,765	67,194,068	22,988	43.11
Embyros, 4-6 hours after egg laying	227,851,364	42,514,758	26,503	28.36
Embyros, 6-8 hours after egg laying	141,469,542	79,907,107	26,140	54.22
Embyros, 8-10 hours after egg laying	127,491,218	70,065,605	26,969	47.75
Embyros, 10-12 hours after egg laying	181,475,614	91,260,284	27,163	64.36
Embyros, 12-14 hours after egg laying	252,990,263	112,726,687	28,348	75.12
Embyros, 14-16 hours after egg laying	166,996,796	80,351,409	28,016	54.93
Embyros, 16-18 hours after egg laying	142,729,615	92,378,132	28,712	62.99
Embyros, 18-20 hours after egg laying	142,396,405	99,754,488	29,672	66.63
Embyros, 20-22 hours after egg laying	75,758,667	62,468,301	24,179	42.37
Embyros, 22-24 hours after egg laying	194,363,331	84,918,915	27,811	60.4
L1 stage larvae	135,149,889	97,265,284	25,967	69.0
L2 stage larvae	220,279,711	120,679,616	26,217	81.56
L3 stage larvae, 12 hour post-molt	74,322,783	63,895,420	25,238	47.11
L3 stage larvae, dark blue gut, puff stage 1-2	93,220,713	63,249,944	27,097	51.62
L3 stage larvae, light blue gut, puff stage 3-6	73,777,616	52,572,231	27,584	35.66
L3 stage larvae, clear gut puff stage 7-9	120,420,263	85,765,885	26,087	58.91
White prepupae	119,724,308	95,926,322	23,738	61.85
Pupae, 12 hours after	148,029,481	90,801,408	23,554	61.23

white prepupae				
Pupae, 24 hours after white prepupae	127,167,221	81,034,553	33,340	53.91
Pupae, 2 days after white prepupae	110,182,007	82,781,553	33,232	53.06
Pupae, 3 days after white prepupae	178,863,564	102,952,025	32,173	69.96
Pupae, 4 days after white prepupae	112,051,595	85,625,100	31,208	57.77
Adult male, one day after eclosion	102,297,246	74,318,284	32,647	53.54
Adult male, 5 days after eclosion	129,507,559	101,588,116	30,832	71.86
Adult male, 30 days after eclosion	136,579,522	65,378,791	31,954	45.32
Adult female, one day after eclosion	127,226,000	98,210,535	30,746	66.84
Adult female, 5 days after eclosion	97,798,582	66,812,323	32,969	40.81
Adult female, 30 days after eclosion	102,949,757	73,720,063	33,445	46.28

APPENDIX B: GENOMICALLY ADJACENT TRANSCRIPTION FACTORS AND lincRNA LOCI

Protein-coding gene	Gene coordinates	lincRNA	lincRNA coordinates
FBgn0000015	3R:12752932-12797958	CO269838	3R:12741034-12741422
FBgn0000095	3R:2721975-2824950	BI628297	3R:2709800-2710411
FBgn0000157	2R:20702353-20722686	AI945277	2R:20696870-20697212
FBgn0000210	X:1504545-1555575	CO284672	X:1486943-1487470
FBgn0000233	X:9588221-9591606	BI620480	X:9585183-9583328
FBgn0000591	3R:21866046-21866585	EL879184	3R:21864358-21865213
FBgn0000606	2R:5866746-5868284	EL878947	2R:5865095-5865991
FBgn0000964	2L:19464762-19466291	BP550759	2L:19469125-19469788
FBgn0001320	3L:20685430-20688463	CO333504	3L:20683002-20683538
FBgn0002734	3R:21823343-21824358	CO269994	3R:21824572-21824849
FBgn0003339	3R:2648842-2675886	BI633062	3R:2679247-2679760
FBgn0003396	2R:7061325-7106718	EC267184	2R:7060057-7060530
FBgn0003513	3R:12200150-12229405	BI374757	3R:12233903-12234428
FBgn0003963	2L:523467-540541	CO152933	2L:521764-522337
FBgn0004053	3R:2578586-2579921	EL877621	3R:2577128-2578028
FBgn0004567	2L:3836842-3839201	BI626657	2L:3835338-3836024
FBgn0004595	3R:7198510-7219326	AI260690	3R:7224482-7225133
FBgn0004606	3R:26591648-26614205	AA817471	3R:26587440-26588248
FBgn0004854	X:17208614-17218195	BI605392	X:17234970-17235617
FBgn0004896	2R:18754044-18758229	AA697410	2R:18752652-18753746
FBgn0005659	3R:23562262-23572797	BF505180	3R:23556753-23557350
FBgn0008651	3R:17235699-17259845	EC210957	3R:17262072-17262414
FBgn0010313	3R:909343-912408	AI107660	3R:880951-882956
FBgn0011701	3R:14061916-14067906	BI616929	3R:14058796-14059516
FBgn0014143	3L:21467135-21469656	BI621778	3L:21471647-21472306
FBgn0016076	2L:5299647-5310997	CO309004	2L:5298789-5299090
FBgn0016694	3L:7807435-7860472	EC253607	3L:7803209-7804221
FBgn0023417	3L:21591958-21606453	BG640055	3L:21588891-21589379
FBgn0026411	X:8651268-8699808	EC216573	X:8650757-8651108
FBgn0027339	3L:22789076-22826488	CO339829	3L:22788421-22788890
FBgn0028789	3L:9034477-9038144	BI639135	3L:9032154-9032843
FBgn0029504	X:7577491-7598851	CO325142	X:7576670-7577141
FBgn0029775	X:5488306-5506266	BI626973	X:5473097-5473745

FBgn0029920	X:6844543-6845046	AI135141	X:6817067-6817700
FBgn0031121	X:20411908-20424910	BI611233	X:20429289-20429742
FBgn0032651	2L:17589507-17591267	EC195259	2L:17589046-17589171
FBgn0034821	2R:18977626-18981785	AI945498	2R:18982179-18982896
FBgn0038063	3R:8378648-8386676	CO188334	3R:8388465-8389711
FBgn0039039	3R:18846944-18852435	BG633226	3R:18853773-18854418
FBgn0039411	3R:21902311-21924433	CK136128	3R:21901606-21902253
FBgn0045852	2L:18763494-18782304	AA140999	2L:18785946-18786610
FBgn0085424	2L:12587871-12628143	EC088598	2L:12572407-12572853
FBgn0085448	3R:13385144-13393648	AI542814	3R:13380369-13380916

APPENDIX C: DNA OLIGONUCLEOTIDES

Co-expression screen

Primer Name	Genomic Location	Sequence 5' → 3'
AA140999_L	2L:18785950-18785969	ACTCCGGTTTGTCTCTGTGG
AA140999_R	2L:18786148-18786167	ACTCGAGATCCGAACAATGG
AA697410_L	2R:18753266-18753285	GACAGGAACATGGATACGG
AA697410_R	2R:18753632-18753651	CTGGAAAATGCTGGAAATCG
AA817471_L	3R:26587792-26587811	TAATTTCCCTCGAACGCAAC
AA817471_R	3R:26588066-26588085	TGGCACTGATTTTGTGGTTC
AI107660_L	3R:882730-882749	TCGGAGCTCAAGAAAGAAGC
AI107660_R	3R:882930-882949	CTGGCGGAAAAAGAAATCTG
AI135141_L	X:6817212-6817231	TTACAGCAACATCAGCAGCC
AI135141_R	X:6817447-6817466	TTTCTTATCGGCGATTAGG
AI260690_L	3R:7224506-7224525	ATTCTCTTACGAGCAGGCCC
AI260690_R	3R:7224798-7224817	ATTGCTGGGCTATTGTCTGG
AI542814_L	3R:13380542-13380561	CTAACTGCAAGTTCCCCCAC
AI542814_R	3R:13380726-13380745	TAATTAAACCGGCAAGTGCG
AI945277_L	2R:20696977-20696996	GTCTGCCGATGGAAACAAAC
AI945277_R	2R:20697167-20697187	TGGCTCATCTACAACTGGAGC
AI945498_L	2R:18982215-18982234	GCAGGATGTGTCCGTTGTAG
AI945498_R	2R:18982416-18982435	TAAACACAGCTGCAGAACGC
BF505180_L	3R:23557009-23557028	CACATCCAGTCAAGCATTCG
BF505180_R	3R:23557193-23557212	CTTTTGCATCCTACGCTTCC
BG633226_L	3R:18853791-18853810	CTGCAATTGCGATAGTGACG
BG633226_R	3R:18854072-18854091	GCAGCCGATCTAAACCATTC
BG640055_L	3L:21589028-21589047	GTAGTCATGCAAGTCCTGCG
BG640055_R	3L:21589263-21589282	CAGCCACAATTACATCCACG
BI374757_L	3R:12233913-12233932	TGTTTAGCTGGAGGAGCAGG
BI374757_R	3R:12234159-12234178	GATCAATCTTATTTGCGCCC
BI605392_L	X:17235177-17235196	GTGAAGCGGAATTTCCAGTC
BI605392_R	X:17235423-17235442	CAGTTCGAACCTCTTGCCTC
BI611233_L	X:20429337-20429356	GGCCAAAGCCAAATCTTATC
BI611233_R	X:20429569-20429588	CGGAATCCAGCCTACTTCAC
BI61929_L	3R:14059276-14059295	CTACCAGCCCATTTTATCCC
BI61929_R	3R:14059480-14059499	TACTCGTGATCCTCGAACCC
BI620480_L	X:9585591-9585610	AGATTTGCAAACATCCAGGC
BI620480_R	X:9585774-9585793	AATGGAAGAAAAGTCGGCAG
BI621778_L	3L:21471676-21471695	AATTTCGTCTTCGACATCGG
BI621778_R	3L:21471993-21472012	TTCAGCTTTCACTGCATTCG

BI626657_L	2L:3835679-3835698	GTTCGGTTCGTTTCGGTTAG
BI626657_R	2L:3835874-3835893	AATGCTGTTCTACGGCCAAG
BI626973_L	X:5473390-5473409	GACCCACTGGTGGTTGTTTC
BI626973_R	X:5473604-5473623	AAGACAGAGAGCCCACGATG
BI628297_L	3R:2710173-2710192	TCGAAATCGATCATCGTCAC
BI628297_R	3R:2710378-2710397	AATATGCGTGTGTGTGGTCC
BI633062_L	3R:2679294-2679313	AAAGTCCCGACTGACTGCTC
BI633062_R	3R:2679594-2679613	CACATGCAGCTTAGTCCTGC
BI639135_L	3L:9032571-9032590	GCATGACGCAATTACACCTG
BI639135_R	3L:9032774-9032793	GACTGGAGATTTCTGGCAGC
BP550759_L	2L:19469281-19469300	CATTTCGGAACTCTGGCTTC
BP550759_R	2L:19469637-19469656	AATCATCCCAGTTGTCCGAG
CK136128_L	3R:21901713-21901732	TTCGTCGTCACCATGAAGTG
CK136128_R	3R:21901897-21901916	TGCACTAATGCCTTGCACTC
CO152933_L	2L:521990-522009	GCTGGCAAGAAAGGCTAAAG
CO152933_R	2L:522243-522262	AGATTCATTAAACGTGCCGC
CO188334_L	3R:8388513-8388532	CTGTGGTTGTTGCCAAAATG
CO188334_R	3R:8388788-8388807	GAAAATCAACAACCCTTCGC
CO269838_L	3R:12741106-12741125	CGAAACGCACACATTCATTC
CO269838_R	3R:12741326-12741345	GCAGGATATGGCCACGTAAC
CO269994_L	3R:21824637-21824656	CAAGTCAATTGAACTCTCCC
CO269994_R	3R:21824743-21824762	CTCGCTGTGGACACACTTTC
CO284672_L	X:1487083-1487102	CATCTCGACCCACATTTTTG
CO284672_R	X:1487269-1487288	TCTATGCGGTTTATGGCCTC
CO309004_L	2L:5298846-5298866	CAGCCAGAAATTGTTTATCAG
CO309004_R	2L:5299070-5299090	GTCACATGTAGAAATGTCTGC
CO325142_L	X:7576739-7576758	ACGTACATCTACGTATGTAC
CO325142_R	X:7577066-7577085	CCTACACACAATGTACAGAC
CO333504_L	3L:20683286-20683305	GCTTGTCTGCGGACATTTTC
CO333504_R	3L:20683510-20683529	GTTCGTTGCCCAAGACTTTC
CO339829_L	3L:22788557-22788578	CAGATCATTCTGATCAAGTCCC
CO339829_R	3L:22788759-22788778	AATTACGCCTACGGTTACGG
EC088598_L	2L:12572593-12572612	CCAGCTTAACTGGATCGGAG
EC088598_R	2L:12572785-12572804	CCCAAGTCTAATCCACGAGC
EC195259_L	2L:17589049-17589068	GCTGAAACTGTGACTCATTC
EC195259_R	2L:17589135-17589154	TGGTTGACGTTTAGTGGTTG
EC210957_L	3R:17262141-17262160	AAAGTCATTGCCGTTGTGTC
EC210957_R	3R:17262372-17262391	GTTGGCTTGAGGATAGCTGC
EC216573_L	X:8650819-8650836	TTGCTGCCGTTTCGTTTG
EC216573_R	X:8651004-8651023	AAACGAGACAAAAATCCCCC
EC253607_L	3L:7803841-7803860	GTTGTCTGTTAGCGGTTGCC
EC253607_R	3L:7804064-7804083	GACCTGAGGCAAAGCAATTC
EC267184_L	2R:7060247-7060266	ATCCACTTGACATCGCCATC
EC267184_R	2R:7060465-7060484	ACGTTTAAATGAAGGCGTCC
EL877621_L	3R:2577557-2577576	TGTCATCAAATGTGTTCCGC

EL877621_R	3R:2577874-2577893	TGGTGTGGTATGGTGTGGAG
EL878947_L	2R:5865499-5865518	TTTAGACGGAATCGAGGGAC
EL878947_R	2R:5865683-5865702	TGAAAAGGACAAGGGATTGG
EL879184_L	3R:21864813-21864832	GATTCCGAAACCAAAAGCAC
EL879184_R	3R:21865081-21865100	CTTGTGCCGGAATGTCTACC
FBgn0000015_L	3R:12755137-12755156	GACCCATGTTCAGGCTAAGG
FBgn0000015_R	3R:12755458-12755477	CAAGCCGTACTCGAAGTTCC
FBgn0000095_L	3R:2734248-2734267	AAACTGACTTCGCATCCAGG
FBgn0000095_R	3R:2735110-2735129	TGGTGGACTGGGTATGGTTC
FBgn0000157_L	2R:20707341-20707360	CTCCATCTCCGATAAGTGCG
FBgn0000157_R	2R:20721046-20721065	ATTTGATTGGGGCTATGCTG
FBgn0000210_L	X:1518556-1518575	TGGACGACACACAGCACTTC
FBgn0000210_R	X:1529925-1529944	TAGATGAACTCCACCAGGGC
FBgn0000233_L	X:9589131-9589150	CCAACTCCAGCATACAATCG
FBgn0000233_R	X:9589911-9589930	ACCGATTGCCGTAATAATC
FBgn0000591_L	3R:21866078-21866097	GAAGGTGAAGAAGCCAATGC
FBgn0000591_R	3R:21866303-21866322	CGTTGACGGCATTCATGTAG
FBgn0000606_L	2R:5867692-5867712	TGGTCCTCATATGCATCATCC
FBgn0000606_R	2R:5868055-5868073	AATCACAGTTGTCGTCGGC
FBgn0000964_L	2L:19465936-19465955	AGGATTGCCTCAATGACGAC
FBgn0000964_R	2L:19466264-19466283	AGAATTCCGGAGAGCTTTGG
FBgn0001320_L	3L:20686952-20686971	GCAGCAGACAATGGATCTTG
FBgn0001320_R	3L:20687420-20687439	TGAACCAGACATGCAAAGTG
FBgn0002734_L	3R:21823717-21823736	TAACCTGGACAAATTTCGGG
FBgn0002734_R	3R:21823917-21823936	CTCTTCTCGCGGAGACTTTG
FBgn0003339_L	3R:2667564-2667583	CGTGTAGTCCACCATATCCG
FBgn0003339_R	3R:2667869-2667888	TTTTGCGATGTCCTCGTACC
FBgn0003396_L	2R:7097035-7097054	GCAGCCAACAACAGTGAGTC
FBgn0003396_R	2R:7097442-7097461	TTGATTGGCTCCTCGTAAGG
FBgn0003513_L	3R:12209771-12209790	GTCATGTGCGTGTATGTGGG
FBgn0003513_R	3R:12216591-12216610	AAGGATGGCGTCACAAAGAG
FBgn0003963_L	2L:536671-536690	ACGAGGCTGGCAGTGATAAG
FBgn0003963_R	2L:537056-537075	GAACCGGATATCACAGTGGC
FBgn0004053_L	3R:2578758-2578777	TTGTAGGAGGTTACGCACGG
FBgn0004053_R	3R:2579091-2579110	GAGATCTTGGAGCATCTGGC
FBgn0004567_L	2L:3837250-3837269	TCCACTTTCAGCACCATCAC
FBgn0004567_R	2L:3837560-3837579	GTGACATGGAGGTCACATCG
FBgn0004595_L	3R:7202109-7202128	TACAACTCGATGACCGGACC
FBgn0004595_R	3R:7202406-7202424	GTTGTTGCTGCTGCTGTTG
FBgn0004606_R	3R:26610881-26610900	CAGACGAGCTGCTATCATTC
FBgn0004854_L	X:17217004-17217023	ATGCTCAGCGGCTACTATGC
FBgn0004854_R	X:17217295-17217314	TCCTCCTCATCATCCTCGTC
FBgn0004896_L	2R:18756086-18756105	TCATCAAAAGCTCCAGAGGC
FBgn0004896_R	2R:18756747-18756766	TCGTTCAGGCTGAGATTGTG
FBgn0005659_L	3R:23570031-23570050	CCACAGAATCTGCAGGACAC

FBgn0005659_R	3R:23570428-23570447	CTCGCTGAGTAATGCCAGTG
FBgn0008651_L	3R:17259461-17259480	TTCTCTCCGGCTTCTTCTCC
FBgn0008651_R	3R:17259643-17259662	AACTCTGATGTTTCGGTGGG
FBgn0010313_L	3R:910910-910929	ACTGTTGCTGGGGATAGTGG
FBgn0010313_R	3R:911193-911212	GCTACTTCCGCAAGATGGAG
FBgn0011701_L	3R:14062805-14062823	TGGACCATTTGGATCATGG
FBgn0011701_R	3R:14063598-14063617	ATACGGAGCACGTTCAAAGG
FBgn0014143_L	3L:21469015-21469034	TCAGGGTCACCTTCTTGTCC
FBgn0014143_R	3L:21469258-21469277	CGACCAGAACTCCTTTACGC
FBgn0016076_L	2L:5307806-5307825	CACCAACAGCCACAACAAAG
FBgn0016076_R	2L:5308289-5308308	TCTTCTCTCGACTCCGCTTG
FBgn0016694_L	3L:7809436-7809455	TGGAATTTCGATAGACGTGC
FBgn0016694_R	3L:7809728-7809747	CGATGATCAAAAAGTCACGC
FBgn0023417_L	3L:21602526-21602547	TCAGTGGACATGGGAGTCTAGG
FBgn0023417_R	3L:21602888-21602907	TGGTTGTAGTGGTGCAGTGG
FBgn0026411_L	X:8662152-8662170	CTTCGGCGTCTGATTGAAG
FBgn0026411_R	X:8662812-8662831	AAAGCATGAGCACCGATAGC
FBgn0027339_L	3L:22791274-22791293	AGGATGTTGGCCTTGTTGTC
FBgn0027339_R	3L:22791964-22791983	CGAGAAGCCGTACAAGTGTG
FBgn0028789_L	3L:9034645-9034664	TTCCATGCAAGAGCTAATGG
FBgn0028789_R	3L:9035258-9035277	ATCTCGGTGCCGATTTTATG
FBgn0029504_L	X:7580219-7580238	AATAAAGCCATCTTGCTGCC
FBgn0029504_R	X:7580623-7580642	TGCTGCCTACGATAGCAGTG
FBgn0029775_L	X:5488811-5488830	CAACAGCAGGCGATTGATAG
FBgn0029775_R	X:5500185-5500204	AACTGCTATTCTTGCTGGGC
FBgn0029920_L	X:6844566-6844585	GCGCGATACGGAAGAATATG
FBgn0029920_R	X:6845027-6845046	TTAGGCGGTGGTAACTTTGG
FBgn0031121_L	X:20412062-20412081	CTGCCAGCAGCAATAACAAC
FBgn0031121_R	X:20412895-20412914	CTTGAAGGCGATCGGTAGAG
FBgn0032651_L	2L:17590483-17590502	GAGAAGCCATTGAGTCCCAC
FBgn0032651_R	2L:17590910-17590929	AAATGCACCCAAATCCAGTG
FBgn0034821_L	2R:18977972-18977990	ACCACTCGATCCAAATCCC
FBgn0034821_R	2R:18978470-18978489	AGGTCTTCGAACGCACTCAC
FBgn0038063_L	3R:8386338-8386357	ATTCAGTTCGTGTGTGGCAG
FBgn0038063_R	3R:8386618-8386637	CAAAGGCATCGAAATCGAAC
FBgn0039039_L	3R:18851081-18851100	TCTGTGTTTGCTGTTCCTGC
FBgn0039039_R	3R:18851291-18851310	TCTCTACCAACAGGGATCGG
FBgn0039411_L	3R:21913893-21913912	TGGAAGTAGTTGGCTTTCCG
FBgn0039411_R	3R:21922069-21922088	AGCAACATCACTTGCAGCAC
FBgn0045852_L	2L:18769101-18769120	CAAAGGACGATGGCAAAATC
FBgn0045852_R	2L:18771365-18771384	TGGTGACGAATTCTCCAAAG
FBgn0046906_L	3R:26610625-26610644	CAGCACGAGAAGGTGCTTTG
FBgn0085424_L	2L:12626020-12626039	GTTCCATCAGCTATTGCAGC
FBgn0085424_R	2L:12626255-12626274	TGCGTCATCTTGTGGTTCAG
FBgn0085448_L	3R:13391177-13391196	TTGGGTTTCATATCGCAGTG

FBgn0085448_R	3R:13392968-13392986	GTTGTTGCTGCTGCTGTTG
TBP_L	3R:19578930-19578950	CACCGAAAAGATCAAGGTCAA
TBP_R	3R:19579213-19579232	CTTTGTTGACTCCGACCAGA

RACE

RACE location	**Primer Name**	**Genomic Location**	**Sequence 5' → 3'**
Dll 5'	Dll_5_outer	2R:20702692-20702712	CTTGATTTGGCGGAACTGTTG
	Dll_5_inner	2R:20702689-20702709	GATTTGGCGGAACTGTTGAGC
dEvf-2 5'	dEvf-2_5_outer	2R:20697267-20697286	CGATGTGATTTACGAGCAGC
	dEvf-2_5_inner	2R:20697250-20697270	CGACAAATGTCACACATGCTG
dEvf-2 3'	dEvf-2_3_outer	2R:20696992-20697011	CAAACGAAGGAGAGGAGTGC
	dEvf-2_3_inner	2R:20697000-20697019	GGAGAGGAGTGCAGAAATGG

dsRNA Generation

Primer Name	**Genomic Location**	**Sequence 5' → 3'**
FBgn0000157_L_dsRNA	2R:20702532-20702551	AATTAACCCTCACTAAAGGGCG CAAAACGGAGTTGAAAAT
FBgn0000157_R_dsRNA	2R:20702901-20702920	AATTAACCCTCACTAAAGGGCC CTAGGTCCGTGAAATTGA
AI945277_L_dsRNA	2R:20696873-20696892	AATTAACCCTCACTAAAGGGGC GGAACCAAATTCAACAAC
AI945277_R_dsRNA	2R:20697087-20697106	AATTAACCCTCACTAAAGGGAG CCTCTTGGCTGTGACACT

FISH probes

Transcript	Probe sequence 5' → 3'
U2	ACAACAAAUGUUAAACUGAUUUUUGGAAU
Dll	AGCAGCACGCCGCCGCCGGGUACGGCGGCAUCCGGAG CACCUAUCAGCAU
dEvf-2	UCAGAAAACCUACGAUUCAUAGGUUUCAAGGGGCCUU GCAGAUCCAU

Knockdown expression

Primer Name	Genomic Location	Sequence 5' → 3'
BI243900_L	2R:9055136-9055155	CTAGATCGCGACACATAGCG
BI243900_R	2R:9055513-9055532	TTTCCATTGAACTGCTGCTG
BI640269_L	X:1179854-1179873	GTCAGGAGATTGAGCCTTCG
BI640269_R	X:1180098-1180117	TGAATTCGACTTGCTGTTGC
BQ103275_L	X:20369922-20369941	CATGACTGGACGTTTTGTGG
BQ103275_R	X:20370248-20370267	CCATCACGTTGTCAAACCAG
CK658697_L	X:2186861-2186880	AAGGAAGTCACACCACTCGG
CK658697_R	X:2187062-2187081	AAGACAACAACAAAAGCGGC
CO32735_R	2R:8116657-8116676	AGTGACCGACGGCATAAAAC
CO327356_L	2R:8116363-8116382	AGTTCTCTCGCATACCTGGG
EC069024_L	3L:16574873-16574892	TGAAGTTAAGCACGGCACTG
EC069024_R	3L:16575093-16575112	TCGATACTCGATTCCCCATC
FBgn0023526_L	X:2176384-2176403	ACGATGATCCGTTTGAGGAG
FBgn0023526_R	X:2186424-2186443	AGCGTCTGCATCAACAACAC
FBgn0025642_L	X:1182351-1182370	CATTTTGGTGCCACAGTTTG
FBgn0025642_R	X:1182722-1182741	AACTGGCAGTACCATCTCGG
FBgn0028550_L	X:1141873-1141892	TGCTGCCGCTATAGTTGTTG
FBgn0028550_R	X:1145266-1145285	CAGGACAGCAGTCATTCCAG
FBgn0028693_L	3L:16576559-16576578	ATCCGACTTTCACACGGAAC
FBgn0028693_R	3L:16576906-16576925	CTCCTTGGGTTTGACGCTAC
FBgn0031118_L	X:20346639-20346658	TGCAGAAGTCCAAACAGACG

FBgn0031118_R	X:20347067-20347086	GATCAAGGGAGATCTGCTGC
FBgn0031119_L	X:20384497-20384516	TGAGGAACATGAACTGCTCG
FBgn0031119_R	X:20384941-20384960	CAGTAGCAGAAGCCAAAGGGV
FBgn0033702_L	2R:8112610-8112628	ATAGCTGCCGACCATGTTG
FBgn0033702_R	2R:8113331-8113350	AGCGATATTTACGGTGTGCC
FBgn0033703_L	2R:8128026-8128045	ACTGCAACTGGAACTCCACC
FBgn0033703_R	2R:8128221-8128240	AACATTCGATGACGAGGAGG
FBgn0033799_L	2R:9051777-9051796	AGAGAGAGCCGCTTCAACAC
FBgn0033799_R	2R:9052083-9052102	CATTGGCTATGCAACACCAG
FBgn0036641_L	3L:16573610-16573629	TGGGTGTCTTTTTCTTTCCG
FBgn0036641_R	3L:16573884-16573903	GTCGCGAGCAGTTCCTTATC
FBgn0054052_L	X:2216040-2216059	TGGGCTGAAGTACAGTGAGG
FBgn0054052_R	X:2216200-2216219	GCAACAAGAAGGCCAAGAAC
FBgn0086532_L	2R:9057196-9057215	ATGAATACGGGCGAGATTTG
FBgn0086532_R	2R:9057501-9057520	CGAGAAGATGTAGCCCAAGC

Knockdown P-element absence verification

lincRNA	**Primer Name**	**Genomic Location**	**Sequence 5' → 3'**
BI640269	BI640269_L_G	X: 1181187-1181206	GCATGTCTCCATTGGTTGTG
	BI640269_R_G	X: 1181380-1181399	TTGGAGGGCAAGGTAAAGTG
BQ103275	BQ103275_L_G	X: 20381669-20381688	GCTTTCCACTCGATTTCGTC
	BQ103275_R_G	X: 20381974-20381993	AATACGCGTTCCATTCTTGG
CO327356	CO327356_L_G	2R: 8116668-8116687	TCGGTCACTTGTCACAGAGC
	CO327356_R_G	2R: 8116882-8116901	GCTCAGACCGAGGATCAAAC
EC069024	EC069024_L_G	3L: 16574487-16574506	TCTGAAAGCGAACGAAATGG
	EC069024_R_G	3L: 16574844-16574863	AACTCGTTTTCACATTGCCC

Knockdown P-element presence verification

lincRNA	Primer Name	Genomic Location	Sequence 5' → 3'
BI640269	FBst0013908_L	X: 1181187-1181206	GCATGTCTCCATTGGTTGTG
	FBst0013908_R	N/A	TTTGGGAGTTTTCACCAAGG
BQ103275	FBst0012471_L	X: 20381669-20381688	GCTTTCCACTCGATTTCGTC
	FBst0012471_R	N/A	TTTGAGGGGGCAATAAACAG
CO327356	FBst0017172_L	2R: 8116668-8116687	TCGGTCACTTGTCACAGAGC
	FBst0017172_R	N/A	CCTCTCAACAAGCAAACGTG
EC069024	FBst0021094_L	3L: 16574487-16574506	TCTGAAAGCGAACGAAATGG
	FBst0021094_R	N/A	CCTCTCAACAAGCAAACGTG

Real-time PCR

Target	**Primer Name**	**Genomic Location**	**Sequence 5' → 3'**	**Efficiency (%)**	***c***	***m***
CG3611	CG3611	QuantiTect Primer Assay	N/A	95.25	39.41	-3.44
Dll	*Dll*	QuantiTect Primer Assay	N/A	84.87	32.74	-3.66
Gapdh	GAPDH	QuantiTect Primer Assay	N/A	97.39	26.09	-3.39
BI640269	BI640269_L_Q	X: 1179954- 1179975	GCATCCAGATCCTTGGAAACTC	72.14	38.25	-4.24
	BI640269_R_Q	X: 1180016- 1180034	TTGCAGTGCATGCCTGAAG			
BQ103275	BQ103275_L_Q	X: 20366645- 20366664	TGATCTGGCCCCTCGTTATC	95.73	33.97	-3.43
	BQ103275_R_Q	X: 20366739- 20366760	AGTGTACGGCTATTCCGATTCC			
CO327356	CO327356_L_Q	2R: 8116661- 8116680	TATGCCGTCGGTCACTTGTC	84.13	30.81	-3.77
	CO327356_R_Q	2R: 8116784- 8116804	GGTCGAAGAAGCGAACACAAC			
dEvf-2	dEvf-2_L_Q	2R: 20696945-20696966	CAGATCCATGTGGCTTCTCTTG	73.50	30.58	-4.18
	dEvf-2_R_Q	2R: 20697087-20967106	AGCCTCTTGGCTGTGACACT			
EC069024	EC069024_L_Q	3L: 16574844- 16574864	GGGCAATGTGAAAACGAGTTC	99.04	35.74	-3.35
	EC069024_R_Q	3L: 16574960- 16574980	CAGGAGGTGCAAAATGTTTCG			