ABSTRACT

Title of dissertation: COMPARATIVE TRANSCRIPTOMICS OF LONG INTERGENIC NONCODING RNAS IN *DROSOPHILA*

Kevin G. Nyberg, Doctor of Philosophy 2015

Dissertation directed by: Carlos A. Machado, Associate Professor, Department of Biology

Without the constraints of the amino acid code, long intergenic noncoding RNAs (lincRNAs) can be expected to evolve along different trajectories than protein-coding genes. Most studies of lincRNA evolution analyze evolution only at the sequence level without ascertaining whether the lincRNA is expressed. Over 2,000 lincRNAs (and counting) have already been identified in the classic model system *Drosophila melanogaster*. Here, using RNA-Seq and computational identification of protein-coding ability, we identify 1,768 lincRNA transcripts at 1,586 unique loci in a second species of *Drosophila – D. pseudoobscura*. These lincRNAs are expressed in every surveyed developmental stage (1st instar larva, 3rd instar larva, pupa, and adult) in both sexes, with a large number increasing in expression as male development proceeds. This male bias can largely be explained by overrepresentation of lincRNAs in the testes. Unequal distributions of sex-biased lincRNAs on the X chromosome and autosomes are consistent with selection-based models of gene trafficking on or off the X chromosome, implying function for some of these lincRNAs. Finally, reciprocal blast searches between

annotated lincRNAs in the *D. pseudoobscura* and *D. melanogaster* transcriptomes identify 80 conserved lincRNAs. Interestingly, direct coordinate conversions between the two genomes reveal another 54 *D. pseudoobscura* lincRNAs that are expressed in the same position as a *D. melanogaster* lincRNA but have low enough sequence conservation to preclude alignment via blast. Whether these positionally equivalent lincRNAs are true homologs with similar functions in both genomes is unclear, but we look at other transcript features, such as transcript orientation, gene structure, and developmental expression profiles to explore this possibility. We find 22 high-confidence lincRNA homologs with conservation of multiple transcript-level features, and we designate these as high-confidence homologs that warrant further biological investigation. This work represents the first comparative transcriptomic analyses of lincRNAs in *Drosophila*.

COMPARATIVE TRANSCRIPTOMICS OF LONG INTERGENIC NONCODING RNAS IN *DROSOPHILA*

by

Kevin G. Nyberg

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Associate Professor Carlos A. Machado, Chair
Associate Professor Karen L. Carleton
Associate Professor Eric S. Haag
Professor Thomas D. Kocher
Professor Leslie Pick
Professor Gerald S. Wilkinson

For my mother, Nancy M. Nyberg, and

my father, Arthur T. Nyberg

ACKNOWLEDGEMENTS

The process of completing a dissertation is a monumental task that is not possible without the support of many, many people whom I will feebly try to acknowledge here.

First, I would like to thank my advisor Carlos Machado for providing me the opportunity to research in his lab. His knowledge and encouragement have pushed me to become a better scientist, and I am quite proud of the work that we have accomplished together. I would also like to thank each of the members of my committee – Karen, Tom, Leslie, Jerry, and Eric – all of who have provided valuable feedback that has strengthened the final product of this long journey.

I also need to thank the other members of the Machado Lab who have provided an extraordinarily supportive and pleasant working environment for the past four years: Kawther, Carlos (Flores), Laura, Zi-Feng, Tian, Li-Yuan, Erik, Henry, and Abby. I want to thank specifically three very talented undergraduates who worked with me on specific aspects of my lincRNA biology projects: Levi, Yiressy, and Kacie.

Four people helped tremendously in the final stages of dissertation preparation. I am immensely indebted to James, Kawther, Gavin, and Jackson for critical and supportive feedback.

I would like to thank Yuji Kageyama and Sachi Inagaki, both now at Kobe University, for providing an unforgettable research and cultural experience during my summer studying lincRNA biology in Japan. I would also like to thank the NSF and JSPS for funding that fellowship (NSF Award #1107742).

Omics projects can rarely be performed without technical assistance from others. Specifically, I would like to thank Suwei Zhao of UMD's Institute for Bioscience and Biotechnology Research sequencing center for providing me consistently high-quality sequence. I would like to thank Yan Wang and Zi-Feng Jiang for generating the proteomics data used in Chapter 1, and I would like to thank Ryan Hardy and Mohamed Noor for going out of their way to extract thousands of exon alignments from Pseudobase specifically for this project.

I would also like to thank the institutions that have kept me salaried and working. Specifically, I acknowledge the NSF, who provided funding for most of the work detailed here (#1330766 to Carlos A. Machado and #1401503 to Carlos A. Machado and Kevin G. Nyberg). I would also like to acknowledge UMD's Graduate School, and Dean Charles Caramello in particular, for a Flagship Fellowship, a Summer Research Fellowship, and a Dean's Dissertation Fellowship. I have been so fortunate to receive such support.

So many friends and colleagues have been supportive as I have tried to wind my way through graduate school. In particular, I need to thank Alana, Will, Matt, Bud, Judy, Jamie, Diane, Wayne, Paul, Elif, Travis, Carl i.e CJ, Adam, Matthew, Matt, Kerrie-Ann, Jeff, MK, Rick, Ilaria, Amanda, Jenny, Natalia, Sushmita, and Maureen for keeping me sane and occasionally distracted from work.

I need to thank my family for providing unconditional support: my sister Kim, my brother Art, my sister-in-law Alicia, my nieces Harper and Madison, and my nephews Joe and Jacob. Most of all, I thank my parents, Arthur and Nancy, for their unwavering belief in my success.

TABLE OF CONTENTS

INTRODUCTION: Biological frontiers of long noncoding RNAs

Biologists have long striven to understand the genetic mechanisms that underlie phenotypic diversity. Both changes in the functions of genes themselves, via alteration of sequence, and changes in gene expression result in phenotypic evolution (Carroll, 2005; Hoekstra and Coyne, 2007; King and Wilson, 1975). The genes that have been studied, however, are almost exclusively protein-coding genes. Revolutionary advances in sequencing technologies and coordinated efforts like the human ENCODE project and the *Drosophila* modENCODE project have deepened our understanding of the diversity of genomic elements found throughout the genome (Celniker et al., 2009; Consortium, 2012). From these efforts, we know that protein-coding exons actually make up a very small fraction of the eukaryotic genome (e.g. 1.5% in humans and 20% in *D. melanogaster*), but the vast majority of the genome is transcribed (Consortium, 2012; Graveley et al., 2011). Introns and untranslated regions (UTRs) account for some of this non-protein-coding transcribed sequence, but there are also large fractions of the genome that are noncoding and independently transcribed. Comparatively little is known about the function and evolution of these long noncoding RNAs (lncRNAs) (Ulitsky and Bartel, 2013).

The term "noncoding RNA" typically evokes a short molecule with a conserved secondary structure and a very specific biological role, but lncRNAs tend to have much more in common with mRNAs (Erdmann et al., 2000; Numata et al., 2003; Ota et al., 2004; Rymarquis et al., 2008). Indeed, sometimes they are referred to as "mRNA-like noncoding RNAs" in the literature (Jiang et al., 2011). They are typically longer than

classic short noncoding RNAs, with most annotation efforts using an arbitrary cutoff of 200 nucleotides. They can possess introns and have multiple isoforms being expressed from a single locus. They are also polyadenylated, which first led to their identification in cDNA libraries and facilitates easy identification in poly(A+) RNA-Seq libraries (Cabili et al., 2011; Calzone et al., 1988; Derrien et al., 2012; Jiang et al., 2011; Numata et al., 2003; Ota et al., 2004; Young et al., 2012).

At this point, thousands of lncRNAs have been identified in dozens of species, mostly in vertebrates, and several lncRNA expression properties have emerged (Kapusta and Feschotte, 2014; Necsulea et al., 2014; Ulitsky and Bartel, 2013). LncRNAs, in general, are expressed at lower levels than protein-coding genes (Brown et al., 2014; Cabili et al., 2011; Derrien et al., 2012; Young et al., 2012). They have higher rates of evolutionary turnover, resulting in higher proportions of lineage-specific lncRNAs than protein-coding genes (Kutter et al., 2012; Necsulea et al., 2014). They also tend to be expressed in a more tissue-specific manner than protein-coding genes, and conservation has been detected with respect to tissue-specificity and developmental expression profiles (Chodroff et al., 2010; He et al., 2014; Washietl et al., 2014).

Even with many thousands of lncRNAs identified, however, we still know very little about their overall biological relevance. We know that some lncRNAs have critical biological functions based on knockout and knockdown studies, and we know how even fewer work on a mechanistic level (Guttman et al., 2011; Wang and Chang, 2011). We do not know whether they are involved in only a few specific biological processes or whether they have been integrated more universally into gene regulatory networks. Some have suggested that lncRNAs play a fundamental role in the evolution of developmental

complexity in eukaryotes, but it is still not clear whether the majority of lncRNAs have any function at all (Kung et al., 2013; Mattick, 2009; Ulitsky and Bartel, 2013).

Critics of big-science, genome-scale projects will point out that the "functional" elements being studied often lack empirical evidence of function, greatly overestimating the proportion of the genome which is actually functional (Graur et al., 2013). It is a fair criticism. With respect to lncRNAs, we do know, empirically, that some lncRNAs play critical and even indispensable roles in eukaryotic biology. The appropriate question is not *if* lncRNAs have function, but rather *how many*? For example, the human ENCODE project resulted in the annotation of 9,277 human lncRNAs, but the proportion of those that are functional remains a mystery (Derrien et al., 2012).

The biological relevance of lncRNAs: evidence from functional studies

Mechanistic functional studies have been conducted on very few lncRNAs. In order to illustrate the diversity of lncRNAs, we discuss functional and evolutionary data for a select few of the most well studied lncRNAs.

Perhaps the best known of all lncRNAs are the dosage compensators. *Xist* in mammals and the two *roX* RNAs in *Drosophila* both are involved in dosage compensation of the X chromosome and autosomes between the homogametic and heterogametic sexes (Brockdorff et al., 1992; Franke and Baker, 1999). The mechanisms of dosage compensation in these species evolved independently and act in opposition; in mammals, a single copy of the X is silenced in females while the X is hypertranscribed in flies (Marin et al., 2000). *Xist* and *roX* evolved independently and act through different mechanisms. *Xist* is a 17kb transcript expressed on the silenced X chromosome. It has a chromatin binding domain that allows it to coat the entire silenced X chromosome and a

conserved secondary-structure protein-binding domain that recruits Polycomb repressor proteins to the chromosome (Brockdorff et al., 1991; Brockdorff et al., 1992; Clemson et al., 1996; Zhao et al., 2008). Studies of *Xist* transcripts between humans and rodents show conserved intron-exon structure but only five small domains of sequence conservation over the entire transcript (Nesterova et al., 2001). The *roX* genes, on the other hand, do not have chromatin binding domains but rather act as scaffold RNAs for the MSL ribo-protein complex, which hypertranscribes the single X chromosome in males (Hamada et al., 2005; Meller et al., 2000). The two *roX* genes differ in length by a few thousand nucleotides and are functionally redundant despite sharing only a small stem loop structure at their 3' ends (Park et al., 2007). Within mammals and flies, respectively, *Xist* and the *roX* genes have very limited sequence conservation and have conserved secondary structure, and the two dosage compensator lncRNAs act through different mechanisms.

LncRNAs, however, have been implicated in more processes than just dosage compensation. A knockdown screen of lncRNAs in zebrafish identified a lncRNA, *cyrano*, that results in deformations in nervous system development when silenced (Ulitsky et al., 2011). Only a few hundred bases of its 4.5kb sequence are highly conserved. As opposed to the dosage-compensator lncRNAs, there is no evidence of conserved secondary structure in *cyrano*, but there is conservation of intron-exon structure. Targeted inhibition of the first splice site results in developmental defects, and despite little conservation in sequence or length, the mouse homolog of *cyrano* can rescue these defects in the zebrafish embryo. It is not clear how *cyrano* functions, but the authors suggest an association with the microRNA *miR-7*.

Like *Xist*, the human lncRNA *HOTAIR* also binds to the Polycomb Repressor Complex 2 (Rinn et al., 2007; Tsai et al., 2010). Instead of functioning in *cis* as a silencer for an entire chromosome, the 2.2kb *HOTAIR* is expressed at the *HoxC* locus, but acts in *trans* and induces repression of the *HoxD* locus on an entirely different chromosome. Sequence and gene structure between mice and humans are poorly conserved, though synteny is strongly conserved (Schorderet and Duboule, 2011). Functional secondary-structure protein-binding sites in human are not present in mouse. That said, both the human and murine *HOTAIR* transcripts induce *HoxD* silencing in their respective genomic environments (Li et al., 2013).

Though the aforementioned lncRNAs all have at least some sequence conservation, the mammalian lncRNA *Airn* functions as a transcriptional repressor of the downstream gene *Igf2r* (Latos et al., 2012). It does not recruit the Polycomb proteins, but instead operates via transcriptional interference, blocking access of RNA polymerase II o the *Igf2r* promoter. Mutations that alter sequence content, length, and intron-exon structure of *Airn* have no effect on its function so long as the transcript is long enough to overlap the *Igf2r* promoter.

These select lncRNAs illustrate many of the challenges facing investigations of lncRNA biology. Some of these lncRNAs have conserved secondary structures that seem to be important; others do not. None of them have high sequence conservation, but some have at least short stretches of high conservation. Synteny seems to be important as illustrated by the *cis*-acting *Airn*, but *HOTAIR* acts in a *trans* fashion. Conserved intron-exon structure can be important on occasion, but other homologous lncRNAs have highly variable gene structures.

We tend to define short noncoding RNAs based on their length and structure and their specific biological roles, but we do not apply this same reasoning to lncRNAs (Eddy, 2001). The lncRNAs appear to be incredibly diverse in terms of length, gene structure, secondary structure, and mechanism of function, but those differences are rarely appreciated. One distinction we do make is between lncRNAs that are expressed in the intergenic spaces of the genome and those that are expressed antisense to a known locus (Kung et al., 2013). This distinction is important methodologically, as RNA-Seq analyses with unstranded sequence reads, like the RNA-Seq data that we have generated, are very poor at discriminating between an antisense lncRNA and the locus it overlaps. Therefore, we restrict our studies to long intergenic noncoding RNAs (lincRNAs).

The biological relevance of lncRNAs: evidence from genome-wide studies

Biological relevance on a genome-wide scale can be assayed either by knockdown/knockout screens or with tests that look for evidence of selection using nucleotide sequence or expression. To date, almost all of the evidence for lncRNA functionality comes from the latter.

Tests for selection have been performed on lncRNA sequences, and these have shown significant but often weak signals of purifying selection in mice and flies (Haerty and Ponting, 2013; Marques and Ponting, 2009; Ponjavic et al., 2007; Young et al., 2012). Evidence of selection in humans is mixed, with one study showing weak evidence of purifying selection and another, using the same frequency site spectrum methods that identified purifying selection in *Drosophila*, failing to find any evidence of selection (Haerty and Ponting, 2013; Ward and Kellis, 2012, 2013). Selection seems to be stronger on lncRNA promoters than on the exonic sequence itself (Guttman et al., 2009; Ponjavic

et al., 2007). When purifying selection is detected in the transcript sequence itself, it is often limited to small regions within the larger transcript (Bhartiya et al., 2014).

Thus far, only a single high-throughput analysis of lncRNA function has been performed. Guttman et al. attempted knockdown via short hairpin RNAs of the entire complement of lncRNAs then annotated in mouse embryonic stem cells (Guttman et al., 2011). They achieved successful knockdown in 147 of the 226 targeted lncRNAs and found that knockdown of 137 lncRNAs (93.2%) resulted in significant changes to the global expression state. Despite most functionally characterized lncRNAs acting in *cis*, the majority of these lncRNAs had *trans* effects, and lncRNAs were implicated in both maintaining the pluripotent stem cell state and driving the stem cells toward differentiation.

lncRNA biology in Drosophila

T. H. Morgan first pioneered the use of the fruit fly *D. melanogaster* to study genetic inheritance (Morgan, 1911). Since then, the organism has often been at the forefront of innovation in genetic research. *D. melanogaster* was the first eukaryotic organism to have its genome sequenced using the whole-shotgun method, and the genus-wide genomic resources make *Drosophila* one of the best systems for studying evolution, particularly on shorter time scales (Adams et al., 2000; Drosophila 12 Genomes et al., 2007; Richards et al., 2005). Curiously, the powerful comparative resources of *Drosophila* have not yet been fully utilized to study lncRNA biology.

Thus far, lncRNAs have been identified only in *D. melanogaster*. Slightly over 100 lncRNAs were initially identified from cDNA libraries, but RNA-Seq data has caused that number to increase rapidly of late (Brown et al., 2014; Inagaki et al., 2005;

Tupy et al., 2005; Young et al., 2012). Over 2,000 lncRNAs are now annotated in the *D. melanogaster* FlyBase annotations, and that number rises with each successive annotation release (St Pierre et al., 2014). LncRNAs in *D. melanogaster* show some of the same properties as seen in vertebrates: low expression levels, low but significant evidence of purifying selection, and high-levels of tissue-specificity (Brown et al., 2014; Haerty and Ponting, 2013; Young et al., 2012). Despite the large numbers of annotated lncRNAs and the general ease of genetic manipulation in flies, functional analyses have been performed on relatively few lncRNAs, most of which have been shown to have neural functions (Gummalla et al., 2012; Lakhotia et al., 2001; Li and Liu, 2014; Li et al., 2012; Mulvey et al., 2014; Petruk et al., 2006; Soshnev et al., 2011).

The genome resources within the *Drosophila* genus are unparalleled, but the transcriptomic resources lag behind (Drosophila 12 Genomes et al., 2007). At this point, extensive developmental transcriptome data is available only for *D. melanogaster* (Graveley et al., 2011). Because lncRNAs often have greater conservation in transcript-level features like intron-exon structure and tissue-specificity than sequence, comparative transcriptomic data is necessary to identify biologically relevant lncRNAs.

D. pseudoobscura *as a model for lincRNA biology*

D. pseudoobscura, which diverged from *D. melanogaster* 25-55 million years ago, has long been used as a model for comparative biology (Richards et al., 2005). Dobzhansky first studied hybrid incompatibilities and investigated causes of hybrid male sterility in *D. pseudoobscura* and its sympatric sister species *D. persimilis* (Dobzhansky, 1936; Dobzhansky, 1937). *D. pseudoobscura* was also the second species of *Drosophila* to have its genome sequenced, facilitating genome-scale comparisons of genomic features, like

cis-regulatory elements, that evolve more quickly than protein-coding sequence (Richards et al., 2005). Numerous evolutionary questions have been investigated in the *pseudoobscura* species subgroup including, but not limited to: causes of reproductive isolation between species (Noor et al., 2001a; Noor et al., 2001b), effects of chromosomal inversions on genetic introgression between species (Kulathinal et al., 2009; Machado et al., 2007; Noor et al., 2007), the divergence of sex-biased gene expression (Jiang and Machado, 2009), and the role of gene misexpression in hybrid dysfunction (Noor, 2005; Ortiz-Barrientos et al., 2007; Reiland and Noor, 2002).

More recently, three male-biased and testes-expressed lincRNAs were discovered that were differentially expressed between *D. pseudoobscura* and *D. persimilis* (Jiang et al., 2011). As male sterility is the primary hybrid dysfunction that keeps *D. pseudoobscura* and *D. persimilis* genetically isolated, this observation of lincRNA expression divergence raises questions about their relationship to transcriptome divergence in general and hybrid dysfunction in particular.

In order to facilitate genus-wide comparisons with *D. melanogaster* and analyses of lncRNA evolution in a classic evolutionary model, we chose to methodically annotate lncRNAs in *D. pseudoobscura*. We used unstranded RNA-Seq to generate developmental and tissue-specific transcriptome data and computationally identified intergenic lncRNAs from unannotated transcripts. We then characterized the expression dynamics of these lincRNAs throughout sex-specific development and in adult gonad and carcass tissues. Finally, we cross-referenced the *D. pseudoobscura* lincRNAs identified here and the existing set of *D. melanogaster* lincRNAs and used various sequence and transcript-level features to identify the first set of high confidence homologous lincRNAs in *Drosophila*.

CHAPTER 1: Identification of long intergenic noncoding RNAs in *Drosophila pseudoobscura* via RNA-Seq

---

ABSTRACT

Extensive annotations of long intergenic noncoding RNAs (lincRNAs) are now available in many species, but few comparative lincRNA datasets have been generated outside of vertebrates. Over 2,000 lincRNAs (and counting) have already been identified in the classic model system *Drosophila melanogaster*. Here, using RNA-Seq and computational identification of protein-coding ability, we identified 1,771 lincRNA transcripts at 1,589 unique loci in a second species of *Drosophila – D. pseudoobscura*. We show that *D. pseudoobscura* lincRNAs share many of the same transcript and sequence properties (i.e. length, alternative transcription, GC content) as lincRNAs from other systems. *D.* pseudoobscura lincRNA sequence, however, does not have the extensive transposable element content seen in vertebrate lincRNAs. In addition, we find that D. *pseudoobscura* lincRNAs are overrepresented on the autosomal 4th chromosome. Finally, we identified 35 multi-locus lincRNA families, with evidence of origination via both tandem duplication and transposition.

INTRODUCTION

Functional noncoding RNAs have long been recognized to have specific and limited roles in eukaryotic cells, both in support of protein translation (i.e. tRNAs and rRNAs) and in a few other isolated processes (e.g. *Xist* and *roX* in dosage compensation) (Brockdorff et al., 1992; Franke and Baker, 1999; Stuckenholz et al., 2003). The advent of genome-wide tiling microarrays and later RNA-Seq demonstrated that transcription can be quite

pervasive in eukaryotic genomes (Bertone et al., 2004; Consortium, 2012; Djebali et al., 2012). Many of these novel transcripts are transcribed at independent loci and share many properties with mRNAs, but do not appear to possess protein-coding ability. Since then, these long intergenic noncoding RNAs (lincRNAs) have been identified in dozens of eukaryotic species, from human to *Plasmodium* (Billerey et al., 2014; Boerner and McGinnis, 2012; Broadbent et al., 2011; Brown et al., 2014; Derrien et al., 2012; Inagaki et al., 2005; Jenkins et al., 2014; Kapusta and Feschotte, 2014; Kutter et al., 2012; Li et al., 2014a; Li et al., 2014b; Liu et al., 2012; Lu et al., 2011; Nam and Bartel, 2012; Necsulea et al., 2014; Pauli et al., 2012; Qu and Adelson, 2012; Weikard et al., 2013; Xie et al., 2014; Young et al., 2012).

High-throughput sequencing is only the first step in a lincRNA identification effort. Typically, once transcript sequences have been identified, protein-coding ability is assayed using various computational methods. Early protein-coding identification methods were often simple, relying strictly on open reading frame (ORF) length, dN/dS ratios, and homology to annotated protein-coding genes in other systems (Inagaki et al., 2005; Ravasi et al., 2006; Tupy et al., 2005). Current methods are more sophisticated. Instead of using just ORF length, some methods incorporate ORF coverage along with other sequence properties (Kong et al., 2007; Wang et al., 2013). Strict dN/dS methods have given way to methods that search for signals of ORF conservation using not only dN/dS ratios, but also INDEL and nonsense mutation information (Lin et al., 2011; Washietl et al., 2011). Transcriptomic and proteomic databases are constantly increasing, both in depth and phylogenetic breadth. New technologies, like Ribo-Seq, can even infer protein-coding ability by looking for physical associations between transcripts and the

ribosome (Ingolia et al., 2009). All of these methods specifically search for evidence of protein-coding ability. Noncoding sequence, therefore, is determined by the lack of protein-coding signal.

The vast majority of these lincRNA identification efforts have been performed either in vertebrates, which tend to have particularly large lincRNA complements, and classic genetic models like *Drosophila melanogaster* and *Caenorhabditis elegans*. Comparative lincRNA transcriptomics have only been performed in vertebrates, so our current knowledge of lincRNA evolution is limited and carries a strong vertebrate bias (Kutter et al., 2012; Necsulea et al., 2014; Qu and Adelson, 2012; Ulitsky et al., 2011). To learn more about lincRNA evolution, it is necessary to sample lincRNAs in a broader swath of the eukaryotic phylogeny. The genome-wide resources in the *Drosophila* genus are unparalleled, but lincRNAs have only been extensively identified in *D. melanogaster* (Brown et al., 2014; Young et al., 2012). Here, we have expanded the lincRNA identification efforts in *Drosophila* to *D. pseudoobscura*, a classic evolutionary model used extensively by Dobzhansky, that diverged from *D. melanogaster* roughly 40 million years ago (Dobzhansky, 1936; Dobzhansky, 1937).

We describe a set of 1,589 lincRNA loci in *D. pseudoobscura* identified using RNA-Seq datasets sampled from multiple developmental stages and adult tissues. We detail how the battery of protein-coding identification methods that we employ allows us to computationally identify proteins that are annotated or novel, long or short, and conserved or lineage-specific. We show that many of the transcript and sequence properties of *D. pseudoobscura* lincRNAs are typical of lincRNAs identified in other

species, but we do find interesting patterns in the genomic organization and transposable element content of *D. pseudoobscura* lincRNAs.

RESULTS

RNA-Seq sample generation and transcriptome assembly

In order to identify lincRNAs in the *D. pseudoobscura* transcriptome and characterize their expression, we performed poly(A+) RNA-Seq on 12 distinct samples from the inbred MV2-25 line, the same line previously used to construct the reference *D. pseudoobscura* genome (Richards et al., 2005). RNA was collected from whole-body flies in a sex-specific manner at four developmental stages: 1st instar larva, wandering 3rd instar larva, mid-pupa, and 7-day adult. While not a comprehensive developmental series, this subset of stages includes three of the four major developmental stages in the life cycle of *Drosophila* and, crucially, includes a developmental stage, the 1st instar larva, that precedes gonad development in both sexes (Bate

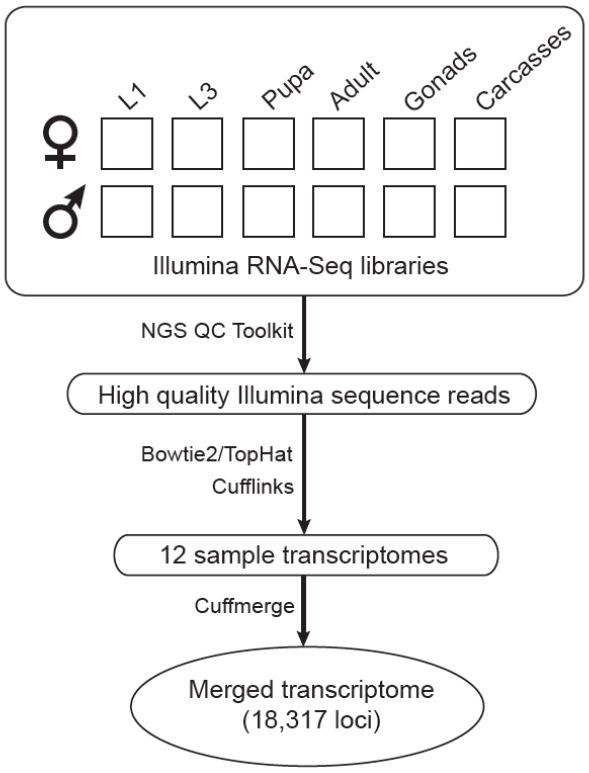


Figure 1 - Workflow of RNA sequencing and transcriptome assembly. RNA from 12 different samples of *D. pseudoobscura* were sequenced, filtered for quality, mapped to the *D. pseudoobscura* genome, and assembled into a single comprehensive transcriptome.

and Martinez Arias, 1993). Because the majority of sex-biased gene expression can be attributed to gene expression in the gonads, we also performed RNA-Seq on isolated testes and ovaries and the remaining carcasses from 7-day adult flies (Parisi et al., 2004).

A workflow for the sequencing, quality control, and transcriptome assembly is shown in Figure 1. The initial 100bp, paired-end RNA-Seq libraries for lincRNA identification were run on an Illumina HiSeq1000, with each library filling 1/3 to 1/2 of a lane on the flow cell. Sequence reads were generally of high quality, with 88.8% to 92.3% of raw reads having an average PHRED quality score above 20 (Table 1). A single sequencing run produced extremely poor base quality in the last 7bp of read 2, so we trimmed all base pairs with a PHRED quality score less than 20 on the 3' ends of all reads. High-quality mate pairs (filtered and trimmed) were then aligned to the *D. pseudoobscura* reference genome (FlyBase r2) using TopHat2/Bowtie2, with 87.0% to 94.9% of the high-quality mate pairs having at least one read mapping to the genome (Table 1) (Kim et al., 2013; Langmead and Salzberg, 2012; St Pierre et al., 2014).

Transcriptomes were assembled individually for each of the twelve samples using Cufflinks and then merged into a comprehensive transcriptome using Cuffmerge (Trapnell et al., 2010). This comprehensive transcriptome consists of 50,459 transcripts at 18,317 gene loci. 42,910 transcripts across 12,475 loci match exonic sequence at a previously annotated gene locus in the *D. pseudoobscura* annotation (r2.29, Cuffmerge class codes "=", "j", and "o") (St Pierre et al., 2014). We identified 6,499 novel intergenic transcripts at 5,478 loci (Cuffmerge class code "u"). The *D. pseudoobscura* genome is poorly annotated with respect to lincRNAs, with only three lincRNAs included

in the r2.29 annotation. Therefore, we used the set of novel intergenic transcripts as the primary source for identifying lincRNAs.

Sample	Raw mate pairs	HQ mate pairs	Mapped fragments
L1M_A	58,282,056	53,298,011 (91.4%)	49,079,587 (92.1%)
L1F_A	55,193,916	50,274,850 (91.1%)	46,722,242 (92.9%)
L3M_A	51,356,548	45,837,820 (89.3%)	41,282,341 (90.1%)
L3F_A	67,852,303	60,285,270 (88.8%)	52,565,652 (87.2%)
PupM_A	58,842,987	53,584,620 (91.1%)	50,634,906 (94.5%)
PupF_A	53,535,245	48,436,580 (90.5%)	45,472,451 (93.9%)
AdM_A	90,031,787	81,796,082 (90.9%)	75,729,598 (92.6%)
AdF_A	93,387,489	84,635,697 (90.6%)	79,914,990 (94.4%)
carcM_A	89,222,887	82,122,717 (92.0%)	71,423,775 (87.0%)
test_A	95,123,169	87,593,042 (92.1%)	83,160,873 (94.9%)
carcF_A	89,322,583	82,471,706 (92.3%)	74,204,686 (90.0%)
ov_A	94,094,047	86,855,374 (92.3%)	82,116,093 (94.5%)

Table 1: RNA-Seq sample statistics – Shown are sequencing, quality control, and mapping statistics for each of the twelve RNA-Seq libraries used to construct the *D. pseudoobscura* transcriptome. "Raw mate pairs" refers to the total number of fragments sequenced with Illumina paired-end sequencing. "HQ mate pairs" refers to the number of raw mate pairs with average PHRED score > 20. "Mapped fragments" refers to the number of high-quality mate pairs, either both mate pairs or only one mate pair, that map to the *D. pseudoobscura* genome.

Computational identification of lincRNAs

We set out to identify a conservative set of lincRNAs in *D. pseudoobscura* by computationally screening the set of 6,499 novel intergenic transcripts (Cuffmerge class code "u") for evidence of protein-coding ability. 3,075 of these transcripts at 2,644 loci do not map to the major chromosome scaffolds (XL, XR, 2, 3, and 4) in the *D. pseudoobscura* genome (FlyBase r2) but to "Unknown_groups" or "Unknown_singletons". We chose not to consider these transcripts further in our analyses. Further, Cufflinks classifications are determined at the level of the transcript and not the locus. We found a small number of novel intergenic transcripts (265 at 189 loci) that derive from a locus that also contains an annotated transcript (Cuffmerge class code "="or "j"). We also chose not to consider these transcripts any further in our analyses.

Thus, we screened 3,159 novel intergenic transcripts from 2,645 loci for protein-coding ability.

Computational screening was performed with four established and complementary methods for identifying protein-coding ability (Figure 2): (1) blastx alignment to the vast NCBI non-redundant protein database (Altschul et al., 1990); (2) local alignments to proteomics datasets, including the *D. melanogaster* PeptideAtlas dataset and a *D. pseudoobscura* testes proteomics dataset (Desiere et al., 2006; Jiang et

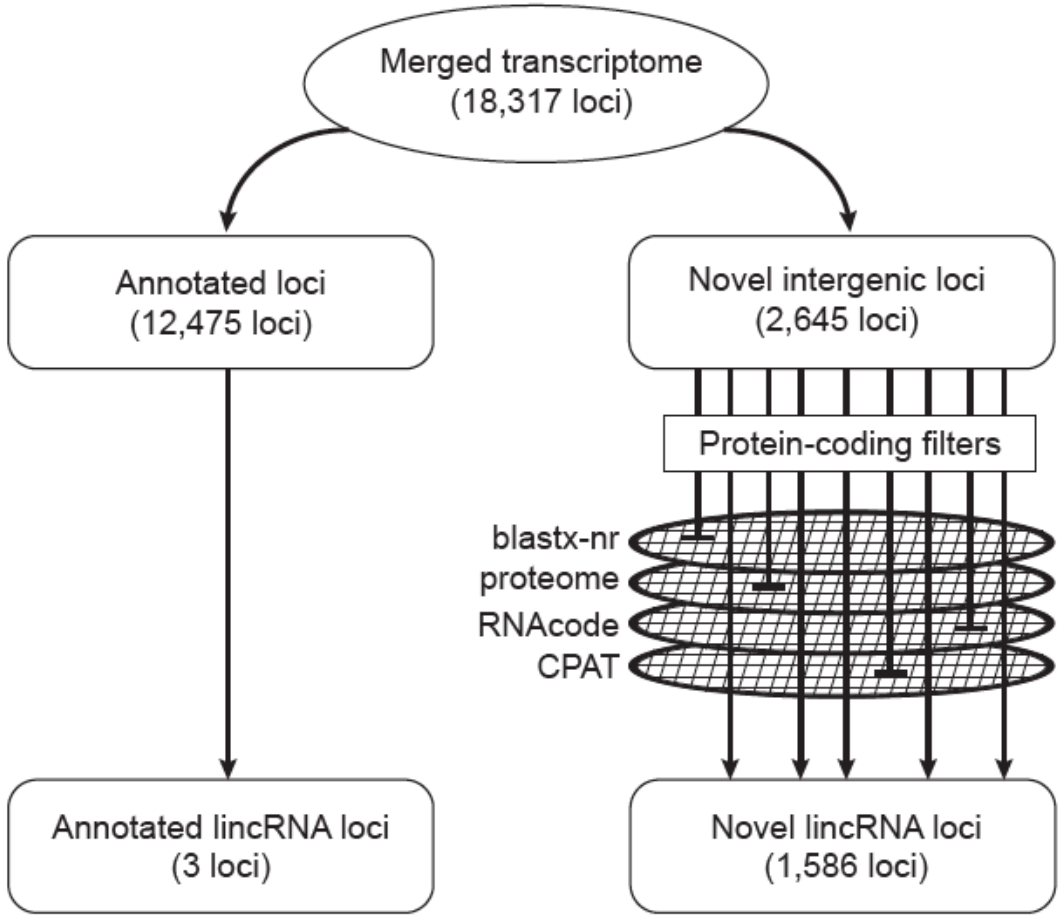


Figure 2 - Workflow of computational lincRNA identification. Novel loci were screened with four different methods of protein-coding identification. In total, 1,589 putative lincRNA loci were identified.

al., 2011); (3) identification of conservation of open reading frames in multiple sequence alignments using RNAcode (Washietl et al., 2011); and (4) identification of coding RNA

sequence features using the Coding Potential Assessment Tool (CPAT) (Wang et al., 2013). Novel intergenic loci with no evidence of protein-coding ability in any of their transcripts using any of these four methods were then classified as putative lincRNAs.

In total, evidence of protein-coding ability was found at 1,059 loci. Specific strategies for each method are further detailed below. Protein-coding ability was detected at 276 loci using the NCBI nr database, 42 loci using proteomics datasets, 777 loci using RNAcode's identification of conserved ORFs, and 233 loci using CPAT's identification of unique coding sequence features (Figure 3). In most cases, putative protein-coding ability was identified using only a single method, usually by ORF conservation via RNAcode. Only 5 loci showed evidence of protein-coding ability using all four methods, though the limited sensitivity of the proteomics databases is largely to blame. 47 loci showed evidence of protein-coding ability using the NCBI nr database, RNAcode, and CPAT.

After filtering out the 1,059 putative protein-coding loci, a set of 1,768 putative transcripts at 1,586 loci remained. In addition, the *D. pseudoobscura* annotation (FlyBase r2.29) includes three lincRNAs: *RNaseP:RNA* (GA29345), *SRP* (GA29352), and *HSR-omega* (GA30101) (St Pierre et al., 2014). Put together, we have identified a set of 1,771 putative lincRNA transcripts at 1,589 independent loci in the *D. pseudoobscura* genome.

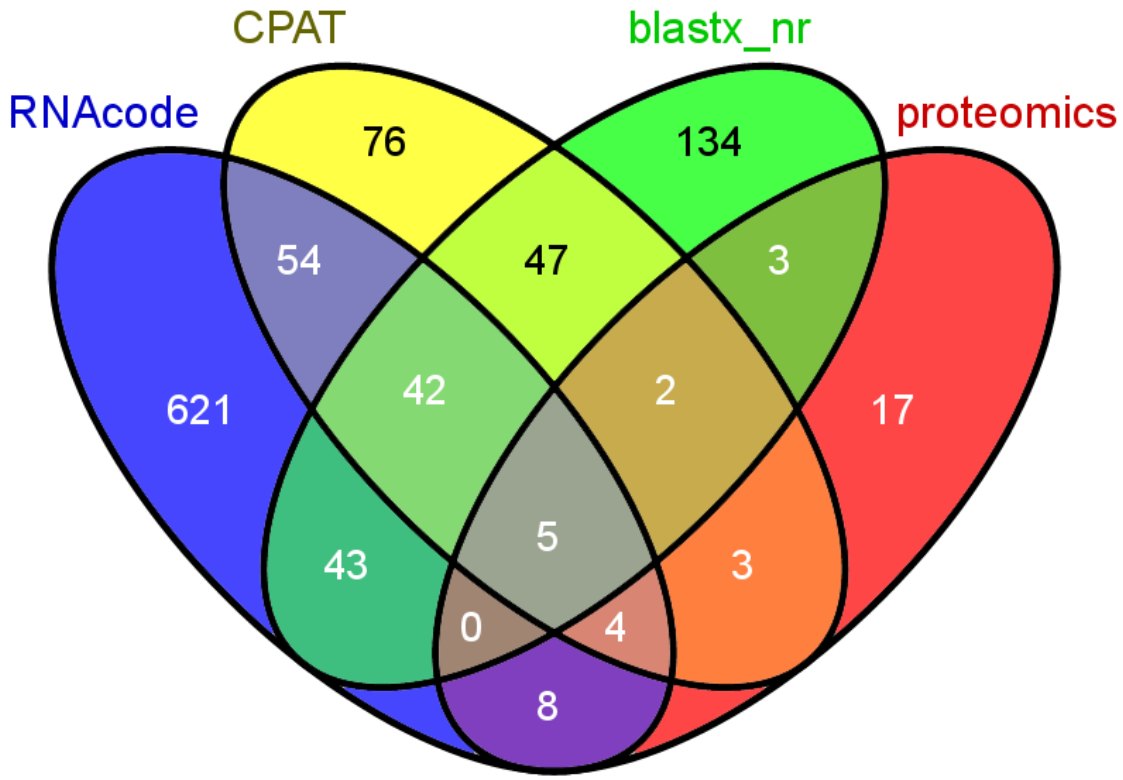


Figure 3 – Detection of protein-coding ability in novel intergenic transcripts. Venn diagram shows the total number of loci with evidence of protein-coding ability using four methods: blastx against NCBI nr database, alignment to proteomics datasets, detection of ORF conservation via RNAcode, and detection of unique noncoding sequence features using CPAT.

Individual results for each of the four different methods of protein-coding ability identification are detailed below. When appropriate, these methods were screened against a set of 10,415 annotated protein-coding loci in *D. pseudoobscura* (r2.29) (all Cuffmerge class code "=") and 699 noncoding loci, consisting mostly of short RNAs, to obtain estimates of the sensitivity and specificity of each of these methods. Maximum sensitivity was our goal even at a moderate cost to specificity, as the biological properties of a high-confidence set of lincRNAs are more informative than a larger repertoire of lower-confidence lincRNAs in *D. pseudoobscura*. We also recognize that locus length is a factor with some of these methods, notably CPAT, and that using a set of predominantly

short RNAs will likely overestimate the specificity of these protein-coding identification methods.

(1) Blastx against the NCBI non-redundant protein database

The NCBI non-redundant protein (nr) database is the most comprehensive set of protein sequences available, including CDS translations from GenBank and protein sequences from SWISS-PROT, the Protein Data Bank, the Protein Information Resource, and the Protein Research Foundation. Individual transcript sequences were aligned against non-*D. pseudoobscura* sequences in the nr database via blastx with an E-value cutoff of 1e-10 (Altschul et al., 1990). Calculated sensitivity (i.e. true positive rate) of this method is 0.956 (9,961/10,415) and specificity (i.e. true negative rate) is 0.993 (694/699). In total, 2,645 novel intergenic loci were screened using the blastx method, and evidence of protein-coding ability was found at 276 loci.

(2) Local alignments to Drosophila *proteomics datasets*

Proteomics datasets offer perhaps the most direct evidence of protein-coding ability, as they rely on observations at the peptide level. To that end, we cross-referenced our novel intergenic transcripts with two *Drosophila* proteomics datasets. First, *D. pseudoobscura* transcript sequences were aligned against a *D. melanogaster* proteomics dataset from the PeptideAtlas database (Aug. 2012) that contains 58,746 distinct peptides using blastx (Desiere et al., 2006). Calculated sensitivity for this method is 0.263 (2738/10,415) and specificity is 1.0 (699/699). In total, 2,645 novel intergenic loci were screened against the *D. melanogaster* PeptideAtlas proteomics dataset, and we found evidence of protein-coding ability at 7 loci.

Second, we queried several testes proteomics datasets from the MV2-25 line of *D. pseudoobscura,* the Susa6 line of *D. ps. bogotana*, and hybrid offspring of the two by matching the longest predicted peptide from each *D. pseudoobscura* transcript (minimum length 10 amino acids) to observed peptide sequences (Jiang et al., 2011). Calculated sensitivity for this method is 0.099 (1,033/10,415) and specificity is 0.977 (683/699). In total, 2,645 novel intergenic loci were screened against the *D. pseudoobscura* testes proteomics dataset, and we found evidence of protein-coding ability at 36 loci, only one of which was also identified via the *D. melanogaster* PeptideAtlas.

(3) Identification of conserved ORFs using RNAcode

Signatures of ORF conservation can be powerful in identifying protein-coding nucleotide sequence, even when annotations are poor or peptides are short. Using multiple sequence alignments, the program RNAcode is able to discriminate between protein-coding and noncoding sequence by identifying nucleotide substitutions that significantly alter the biochemical properties of potentially translated amino acids, INDELs that disrupt potential ORFs, and substitutions that would result in premature stop codons (Washietl et al., 2011). We performed RNAcode analyses using multiple sequence alignments of *Drosophila* sequences from two sources: (1) the UCSC 15-species multiple genome alignment of *Drosophila* and several insect outgroups and (2) a *D. pseudoobscura* subgroup specific multiple genome alignment available from Pseudobase (Kuhn et al., 2007; McGaugh et al., 2012; McGaugh and Noor, 2012; Noor, 2012).

The UCSC *Drosophila* alignment contains genome sequences from 12 species of *Drosophila* along with three insect outgroups: *Anopheles gambiae*, *Apis mellifera*, and *Tribolium castaneum*, with all genome sequences aligned to the *D. melanogaster* BDGP

release 5 genome (Drosophila 12 Genomes et al., 2007; Holt et al., 2002; Honeybee Genome Sequencing, 2006; Hoskins et al., 2007; Kuhn et al., 2007; Richards et al., 2005; Tribolium Genome Sequencing et al., 2008). We converted *D. pseudoobscura* transcript coordinates to *D. melanogaster* transcript coordinates using the liftOver tool and extracted the resulting multiple sequence alignment from the larger multiple genome alignment (Hinrichs et al., 2006). Because *D. melanogaster* is the reference, only *D. pseudoobscura* loci with high-quality alignments to the *D. melanogaster* genome will be able to be converted and extracted. Conversions were attempted on all annotated protein-coding and noncoding loci (Cufflinks class code "=") and novel intergenic loci (Cufflinks class code "u"). Conversions were successful at 0.674 (7,022/10,415) of annotated protein-coding loci, 0.788 (551/699) of annotated noncoding loci, and 0.278 (736/2,645) of novel intergenic loci. In most cases, these conversions place the loci on the same Muller element in both *Drosophila* species: 0.964 (6,770/7,022) of annotated protein-coding loci, 0.938 (517/551) of annotated noncoding loci, and 0.984 (724/736) of novel intergenic loci (Muller, 1940).

We searched for evidence of conserved ORFs using RNAcode ($p<0.05$) at the loci that we were able to extract from the UCSC 15-species alignment (Washietl et al., 2011). Calculated sensitivity for this method is 0.979 (6,874/7,022), and calculated specificity is 0.953 (525/551). Of the 736 novel intergenic loci that we screened using RNAcode, we found evidence of protein-coding ability at 154 loci.

Though RNAcode performs well when alignments can be extracted from the UCSC 15-species alignment, the majority of novel intergenic loci do not survive the conversion process, likely due to poor alignments between *D. melanogaster* and *D.*

pseudoobscura. Therefore, we also chose to run RNAcode ($p<0.05$) using a multiple genome alignment from Pseudobase that include 11 lines of *D. pseudoobscura*, two lines of *D. pseudoobscura bogotana*, four lines of *D. persimilis*, three lines of *D. miranda*, and a single line of *D. lowei*, all aligned to the *D. pseudoobscura* MV2-25 reference genome (McGaugh et al., 2012; McGaugh and Noor, 2012; Noor, 2012). Calculated sensitivity for RNAcode using the Pseudobase alignment is 0.804 (8,370/10,415), and calculated specificity is 0.984 (688/699). Of the 2,645 novel intergenic loci that we screened using RNAcode with the Pseudobase alignment, we found evidence of protein-coding ability in 516 loci.

The calculated sensitivity of RNAcode using the full Pseudobase alignment (0.804) is not ideal. Because RNAcode relies on assessments of sequence variation and the Pseudobase alignment contains multiple lines of many species where variation is expected to be minimal, we re-ran RNAcode using a reduced version of the Pseudobase alignment with only a single line from *D. pseudoobscura*, *D. persimilis*, *D. miranda,* and *D. lowei*. Calculated sensitivity with this approach is marginally higher at 0.811 (8,444/10,415), and calculated specificity is 0.990 (692/699). Using this approach, we found evidence of protein-coding ability in 403 of the 2,645 loci that we screened.

Interestingly, only 8,017 annotated protein-coding loci were detected using both the full and reduced Pseudobase alignments, with 353 and 427 loci, respectively, being detected using only a single approach. Combining both, calculated sensitivity of RNAcode on Pseudobase alignments is 0.845 (8,797/10,415), and calculated specificity is 0.976 (682/699).

The computational intensity of running RNAcode prevented further optimization of the Pseudobase approach, but we did further examine the quality of the sequences in the reduced Pseudobase alignment to assess the reduced sensitivity of the Pseudobase alignment as compared to the UCSC alignment. The Pseudobase alignments lack INDELs of any type, which RNAcode would find informative. Furthermore, many sequences have large runs of Ns, which are not informative to RNAcode though did not cause the program to terminate. The UCSC alignment, using near-complete genomes from large-scale genome sequencing projects, has very few indeterminate bases and does include INDELs, both of which are informative to RNAcode. We calculated that 841 annotated protein-coding loci have Pseudobase alignments where at least one of the four sequences is comprised of >50% Ns. If these loci are ignored, then the sensitivity of

Figure 4 – Detecting protein-coding ability using RNAcode and three multiple genome alignments. Venn diagram shows the total number of loci with evidence of protein-coding ability using RNAcode and the UCSC 15-species *Drosophila* alignment (UCSC), the Pseudobase 5-species *D. pseudoobscura* subgroup specific alignment with all lines from each species (Pbase_all), and the Pseudobase 4-species *D. pseudoobscura* subgroup alignment with only a single line per species.

RNAcode using the reduced Pseudobase alignment rises from 0.811 to 0.867 (8,305/9,576).

Combining the results of RNAcode using both the UCSC alignment and the two iterations of the Pseudobase alignment, we identify evidence of protein-coding ability in 777 of the 2,645 novel intergenic loci (Figure 4). Only 274 of these loci were identified using more than one approach, with the greatest amount of overlap occurring between the two iterations of Pseudobase.

(4) Identification of noncoding sequence features using the Coding Potential Assessment Tool

Protein-coding sequence is constrained by the biochemical properties of the polypeptide chains that they encode and the frequency of the tRNAs that recruit the necessary amino acids. Noncoding sequence, not subject to these constraints, will be compositionally distinct from protein-coding sequence. The Coding Potential Assessment Tool (CPAT) builds a logistic regression model based on species-specific protein-coding sequence features and uses that to discriminate between coding and noncoding transcripts (Wang et al., 2013). CPAT uses four sequence features: (1) ORF length; (2) ORF coverage, or the percentage of the transcript that the ORF covers; (3) the Fickett TESTCODE statistic, which considers nucleotide composition and codon-usage bias (Fickett, 1982); and (4) in-frame hexamer usage bias, or the frequency with which amino-acid coding triplets are adjacent to each other within an ORF (Fickett and Tung, 1992).

CPAT requires a set of training protein-coding and noncoding sequences in order to build an adequate logistic regression model. Because ORF length is a factor, the set of *D. pseudoobscura* noncoding RNAs, which includes only three lincRNAs, is inadequate

for training purposes. CPAT does supply a pre-built model using *D. melanogaster* transcripts, and we consider whether that would be a suitable model to identify protein-coding sequences in *D. pseudoobscura*. Codon usage bias and GC content are largely similar between *D. pseudoobscura* and *D. melanogaster*, which supports the use of the *D. melanogaster* model (Moriyama and Powell, 1997; Powell and Moriyama, 1997; Rodriguez-Trelles et al., 2000). That said, the lengths of untranslated regions (UTRs) are significantly longer in *D. pseudoobscura*, which will impact measures of ORF coverage (Palmieri et al., 2012). With that in mind, we performed CPAT using the *D. melanogaster* model on our control locus sets and found a sensitivity of 0.978 (10,186/10,415) and specificity of 0.996 (696/699). Using this approach, we screened 2,645 loci and detected protein-coding ability in 126 loci (Figure 5).

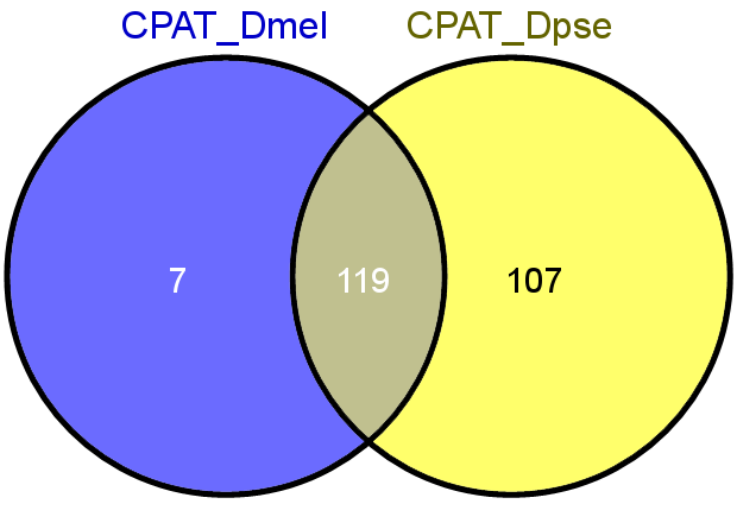


Figure 5 – Detecting protein-coding ability using CPAT with two logistic regression models. Venn diagram shows the total number of loci with evidence of protein-coding ability using CPAT and the default *D. melanogaster* model (CPAT_Dmel) and the custom *D. pseudoobscura* model (CPAT_Dpse).

Still, we wanted to run CPAT using a *D. pseudoobscura* specific model. At this point, we had already performed protein-coding searches using all aforementioned

methods, including RNAcode using the UCSC alignments. A set of 418 transcripts at 409 loci passed all previous filters, including blastx against the NCBI nr database, both proteomics database searches, RNAcode using all three approaches, and CPAT using the *D. melanogaster* model. We used this high-confidence set of lincRNAs along with the set of annotated protein coding transcripts to train a *D. pseudoobscura* model. Sensitivity and specificity were calculated using the CPAT training sets, with a sensitivity of 0.986 (10,270/10,415) and a specificity of 0.985 (403/409) using the high-confidence lincRNA dataset. After removal of the high-confidence lincRNAs, we screened the remaining loci using CPAT with a *D. pseudoobscura* model and found evidence of protein-coding ability at 226 loci. 94.4% (119/126) of the loci identified using CPAT with the *D. melanogaster* model were also identified using the *D. pseudoobscura* model, but CPAT with the *D. pseudoobscura* model identified protein-coding ability at an additional 107 loci (Figure 5). Together, we found evidence of protein-coding ability at 233 loci using CPAT.

Transcript and sequence properties of lincRNAs

Between novel transcripts uncovered by our RNA-Seq datasets and previously annotated transcripts, we have a total of 1,771 putative lincRNA transcripts at 1,589 independent loci throughout the *D. pseudoobscura* genome. Here, we describe transcript and sequence properties of these lincRNAs and detail how lincRNAs are differentiated from the 10,415 annotated protein-coding genes in *D. pseudoobscura*.

Total exonic length of lincRNAs

LincRNA transcripts tend to be shorter than transcripts from protein-coding loci (Figure 6). We combined all non-redundant exons from all isoforms at each locus and found a

median length of 772 nucleotides in lincRNAs and 3,165 nucleotides in protein-coding loci. Very few lincRNAs approach the longest total exonic lengths seen in protein-coding loci, with only 27 lincRNA loci possessing lengths above the protein-coding median of 3,165 nucleotides and only 2 lincRNA loci above 10,000 nucleotides (0.1%). In contrast, 1,333 (12.8%) of the protein-coding loci have total exonic length above 10,000 nt. The distributions of lincRNA and protein-coding exonic length are significantly different (Mann-Whitney, p<2.2e-16). Taken together, we identify 1,500,896 nucleotides of lincRNA exonic sequence and 50,502,941 nucleotides of protein-coding exonic sequence, including UTRs and CDS but not introns. This represents 1.0% (1,500,896/152,738,921) and 33.1% (50,502,941/152,738,921), respectively, of the sequenced *D. pseudoobscura* FlyBase r2 genome.

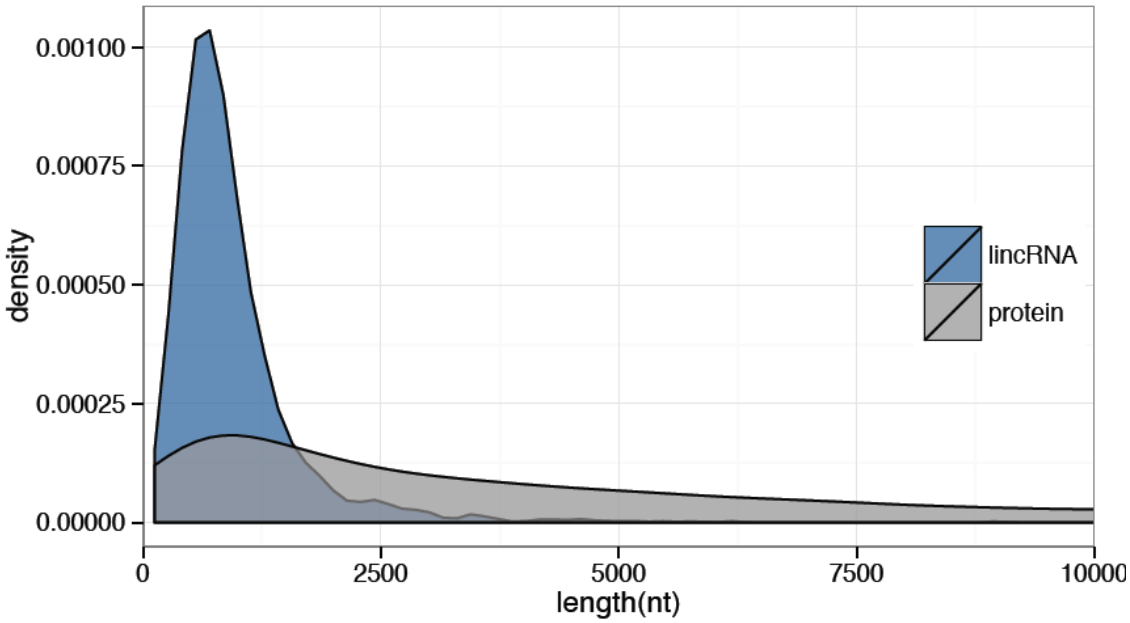


Figure 6 – Distributions of total exonic length in lincRNA and protein-coding loci. Shown are density distributions of total exonic length (nt) for 1,589 lincRNA loci and 10,415 protein-coding loci. Distributions are significantly different (Mann-Whitney, p<2.2e-16).

Splicing and alternative transcription in lincRNAs

LincRNA loci tend to have fewer exons than protein-coding loci, with the majority that we detected being single-exon transcripts (Figure 7). Mean exon number per locus is 1.50 for lincRNA loci and 6.04 for protein-coding loci. 1,088 (68.5%) lincRNA loci contain only a single exon, while 1,492 (14.3%) protein-coding loci contain a single exon. The distributions of exon number are significantly different via the Mann-Whitney test ($p<2.2e-16$).

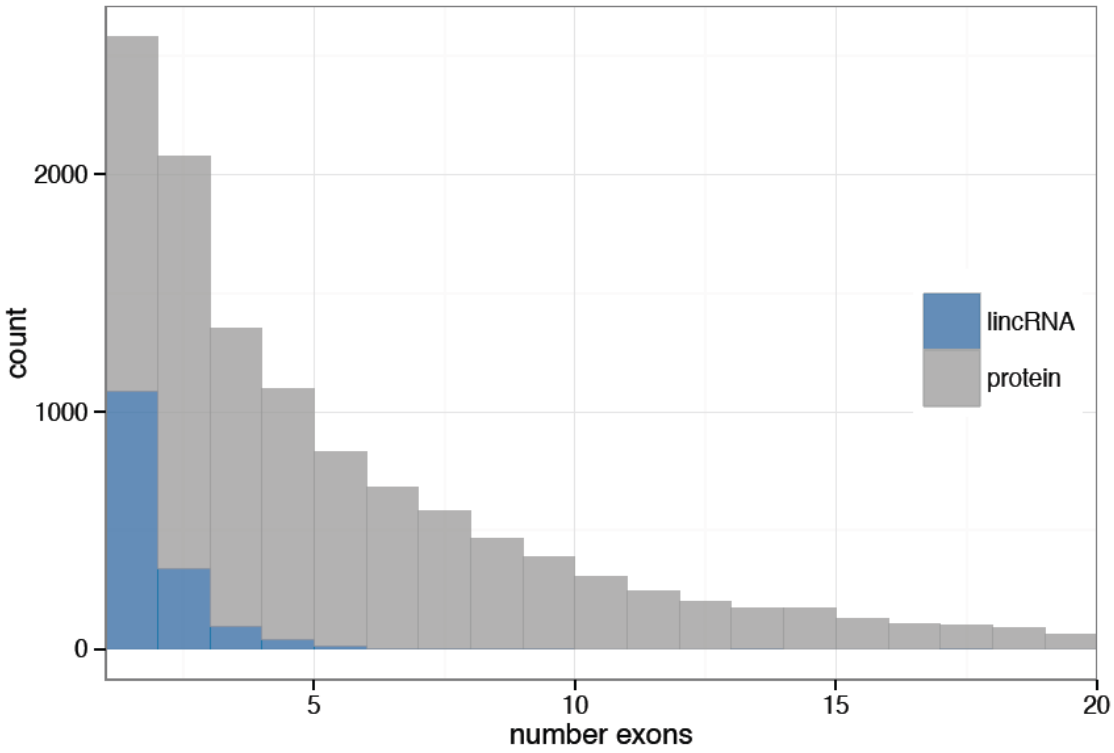


Figure 7 – Distributions of exon number in lincRNA and protein-coding loci. Shown are histograms of total exon number for 1,589 lincRNA loci and 10,415 protein-coding loci. Distributions are significantly different (Mann-Whitney, $p<2.2e-16$).

Consequently, alternative transcription, while present, is detected significantly less frequently in lincRNA loci (Figure 8, Mann-Whitney, $p<2.2e-16$). We found evidence of multiple isoforms in only 149 (9.4%) of lincRNA loci, while we found evidence of multiple isoforms in 7,264 (69.7%) of protein-coding loci. We detect no

more than 9 isoforms at any single lincRNA locus, while we detect more than 9 isoforms at 827 protein-coding loci.

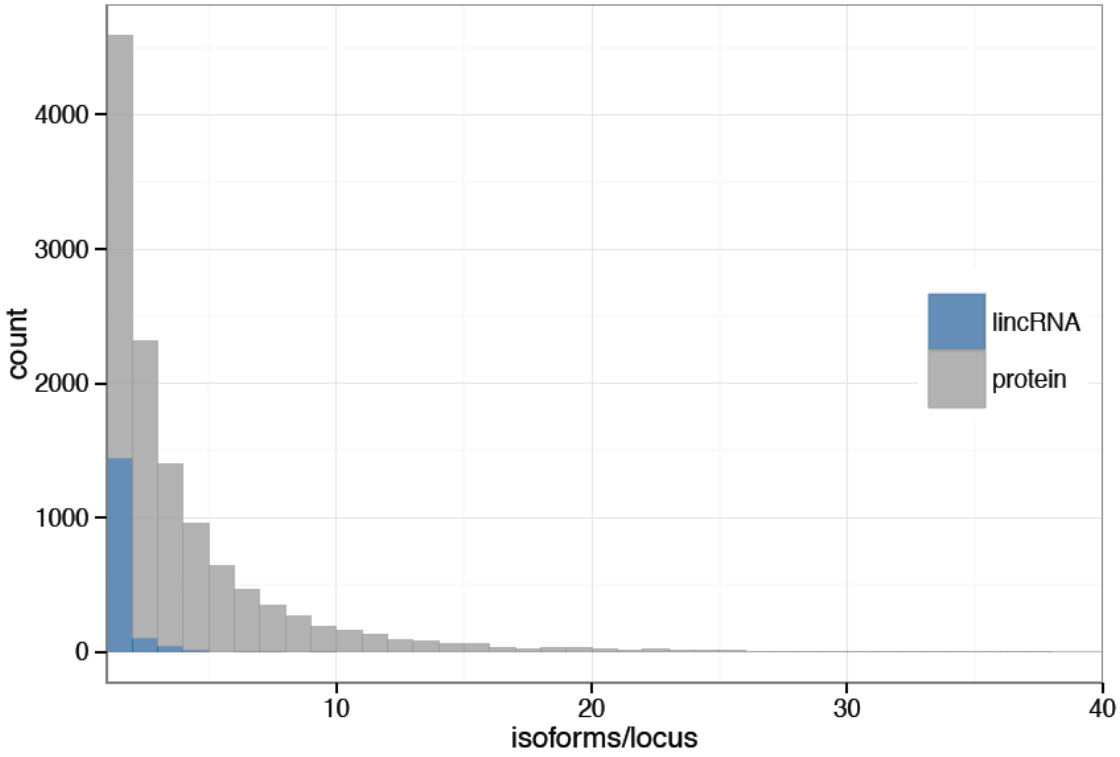


Figure 8 – Distributions of total detected isoform number in lincRNA and protein-coding loci. Shown are histograms of total detected isoform number for 1,589 lincRNA loci and 10,415 protein-coding loci. Distributions are significantly different (Mann-Whitney, p<2.2e-16).

Genomic location of lincRNA loci

We tested whether lincRNA loci are distributed differently among the chromosomes than protein-coding loci. Table 2 lists the number of lincRNA loci and protein-coding loci found on each of the five major chromosome scaffolds in the *D. pseudoobscura* genome.

Figure 9 shows the distributions of lincRNA and protein-coding loci on both arms of the X chromosome and the three major autosomes. Distributions of lincRNA and protein-coding loci are similar among both arms of the X chromosome and chromosomes

2 and 3. However, chromosome 4 shows a significant overrepresentation of lincRNA loci to protein-coding loci (X^2 test, modified Bonferroni corrected $p < 0.04$).

genomic element	XL	XR	2	3	4
lincRNA	238	324	386	267	374
protein	1624	2362	2510	1903	2016

Table 2 – Interchromosomal genomic location of lincRNA and protein-coding loci. Shown are the genomic locations of 1,589 lincRNA loci and 10,415 protein-coding loci on the five major *D. pseudoobscura* chromosome scaffolds.

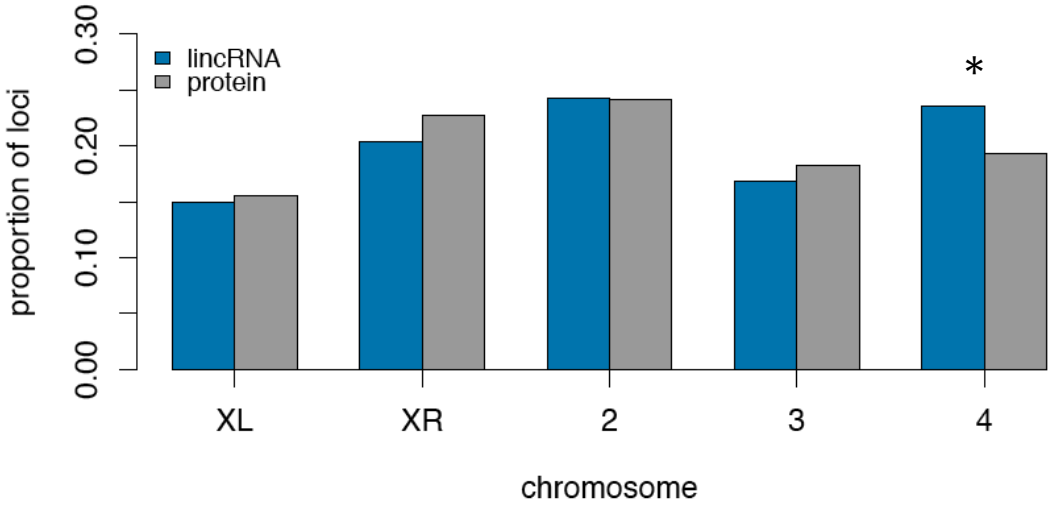


Figure 9 – Interchromosomal genomic distributions of lincRNA and protein-coding loci. Shown are the genomic distributions of 1,589 lincRNA loci and 10,415 protein-coding loci on the five major *D. pseudoobscura* chromosome scaffolds, with the proportion of the total loci in each class shown on the y-axis. *LincRNAs are significantly overrepresented on chromosome 4 as compared to protein-coding loci (X^2 test, modified Bonferroni corrected $p < 0.04$).

D. pseudoobscura has a submetacentric X chromosome, with XR slightly longer than XL, and three major telocentric autosomes (Tan, 1935). Because gene density often varies across a chromosome, we tested whether lincRNA loci are differentially represented in regions near the centromere or telomere as compared to protein-coding loci. We chose to focus on the 3Mb closest to each centromere and telomere, as that is the

distance where recombination rates have been observed to be lowest (McGaugh et al., 2012). Table 3 lists the number of lincRNA loci and protein-coding loci found within the 3Mb nearest the centromere, the 3Mb nearest the telomere, and the middle of the chromosome for each of the five major chromosome arms of *D. pseudoobscura*.

genomic element	3Mb centromere	middle	3Mb telomere
lincRNA	204	1214	171
protein	1197	8003	1215

Table 3 – Intrachromosomal genomic location of lincRNA and protein-coding loci. Shown are the genomic locations of 1,589 lincRNA loci and 10,415 protein-coding loci in regions 3Mb nearest the centromeres, 3Mb nearest the telomeres, and in the middle of the arms of each of the five major chromosome arms of *D. pseudoobscura*.

Figure 10 shows the intrachromosomal distributions of lincRNA and protein-coding loci on the five major chromosome arms. We find no significant differences between the distributions of lincRNA and protein-coding loci among regions near the centromeres and telomeres and the middle of the chromosome arms (X^2 test, p=0.1607).

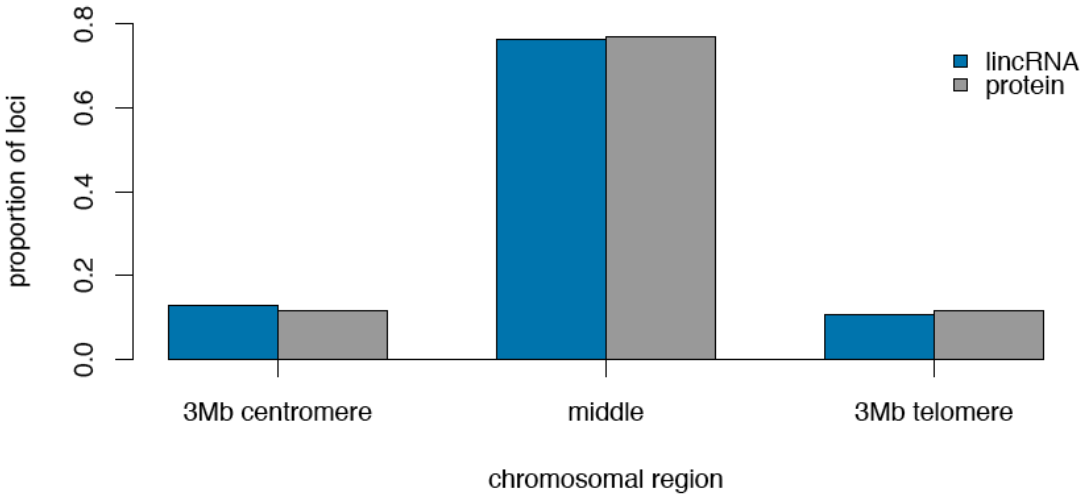


Figure 10 – Intrachromosomal genomic distributions of lincRNA and protein-coding loci. Shown are the genomic distributions of 1,589 lincRNA loci and 10,415 protein-coding loci in regions 3Mb nearest the centromeres, 3Mb nearest the telomeres, and in the middle of the arms of each of the five major chromosome arms of *D. pseudoobscura*, with the proportion of the total loci in each class shown on the y-axis. We find no significant differences between the distributions of lincRNA and protein-coding loci (X^2 test, p = 0.1607).

Sequence composition of lincRNAs

Sequence composition varies among different types of genomic elements. The total GC content of the *D. pseudoobscura* genome (r2.29) is 43.2% (66,041,991/152,738,921). We calculated GC content in lincRNA exons to be 43.7% (656,558/1,500,896), while GC content at protein-coding exons is 49.0% (24,754,282/50,502,941) (Table 4, Figure 11). Protein-coding exons contain both CDS and UTR, and we reason that GC content will be non-uniform across these different elements. Using *D. pseudoobscura* FlyBase r2.29 annotations, we calculated GC content in the following genomic elements: CDS at 55.8% (14,061,584/25,200,782), 5' UTR at 43.0% (62,091/144,411), 3' UTR at 38.6% (103,793/269,236), introns at 42.5% (15,838,001/37,243,879), and intergenic sequence at 40.7% (32,487,969/79,924,265). This suggests that the higher GC content of protein-coding loci can be attributed to GC content in the coding regions.

genomic element	GC	simple repeats	low complexity
lincRNA	43.7%	5.09%	0.73%
protein	49.0%	3.53%	0.54%
CDS	55.8%	2.22%	0.17%
5'UTR	43.0%	3.20%	0.81%
3'UTR	38.6%	5.21%	1.22%
intron	42.5%	5.88%	0.75%
intergenic	40.7%	5.22%	0.72%
genome	43.2%	4.49%	0.59%

Table 4 – Sequence composition of lincRNAs and other genomic elements. Shown are the GC content, simple repeat content, and low-complexity sequence content (all in % of total nucleotides) of non-redundant exons in 1,589 lincRNA loci and 10,415 protein-coding loci. Also shown are sequence composition statistics for multiple annotated genomic elements from the *D. pseudoobscura* FlyBase r2.29 annotations.

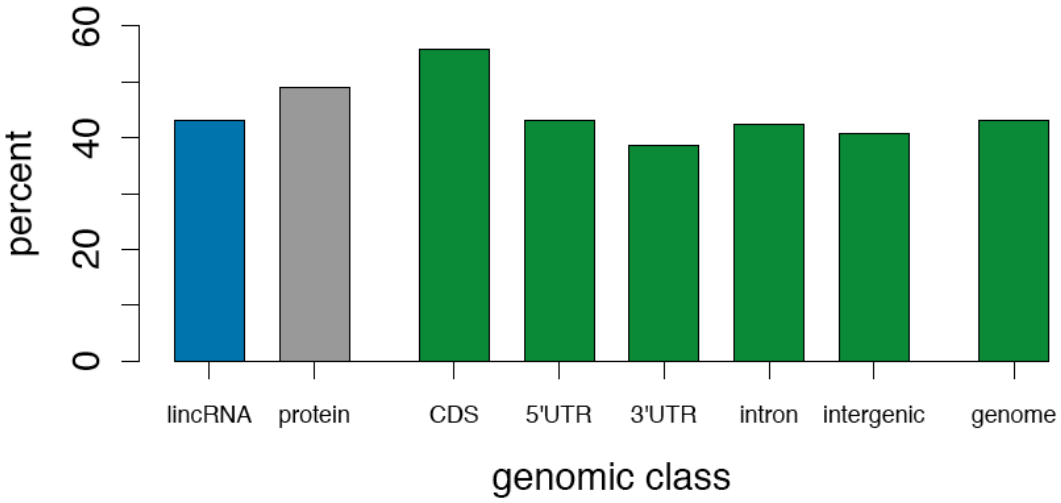


Figure 11 – GC content of lincRNAs and other genomic elements. Shown is GC content (%) for non-redundant lincRNA and protein-coding exons. Also shown is GC content (%) for five different genomic elements annotated in the *D. pseudoobscura* FlyBase r2.29 genome (CDS, 5' UTR, 3 UTR, introns, and intergenic sequence) and the complete genome.

We also looked at the contributions of simple repeats and low-complexity sequence (i.e. homopolymer and poly-purine/poly-pyrimidine stretches) to lincRNA sequence. Using the non-redundant exon sets, we find that 5.09% of lincRNA locus sequence and 3.53% of the protein-coding locus sequence is comprised of simple repeats, and that 0.73% of lincRNA locus sequence and 0.54% of protein-coding locus sequence is comprised of low-complexity sequence (Table 4, Figure 12). The reduced content of both simple repeats and low-complexity sequence is likely due to reductions in the CDS, although the 5'UTR does show lower levels of simple repeats.

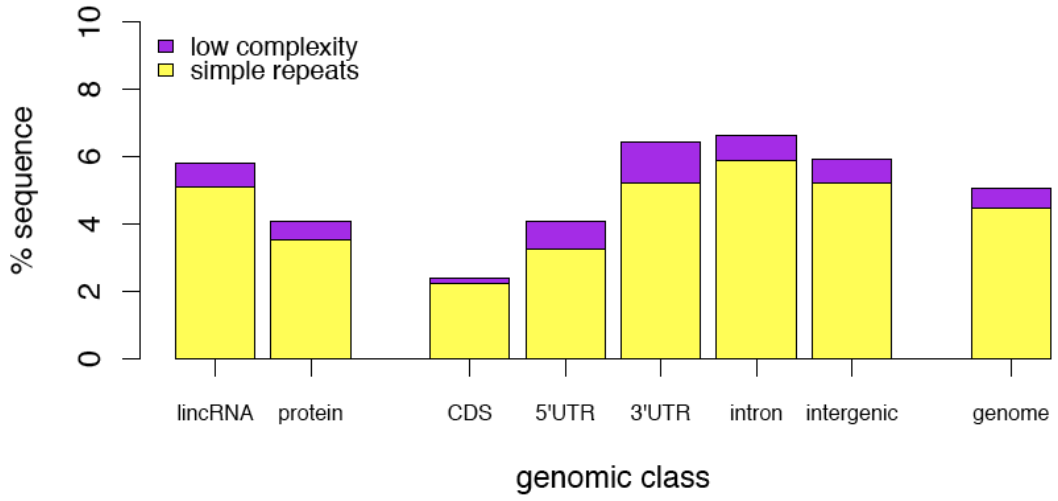


Figure 12 – Simple repeat and low-complexity sequence content of lincRNAs and other genomic elements. Shown are simple repeat content (%) and low-complexity sequence content (%) for non-redundant lincRNA and protein-coding exons. Also shown are percentages for five different genomic elements annotated in the *D. pseudoobscura* FlyBase r2.29 genome (CDS, 5' UTR, 3 UTR, introns, and intergenic sequence) and the complete genome.

Finally, we looked at the contributions of sequence fragments originally derived from transposable elements (TEs) to lincRNA sequence. Using the set of annotated TEs from *D. melanogaster*, we find that TE-fragment sequence comprises 11.99% of the *D. pseudoobscura* genome sequence (FlyBase r2). Total TE composition in the non-redundant exons at lincRNA and protein-coding loci is lower, at 3.49% and 1.63%, respectively (Table 5, Figure 13). Interestingly, when TE content is broken down into subclasses, we find a higher content of LINE retroelements in the protein-coding loci, and this can be attributed to their presence in the CDS. Altogether, TEs were identified in 14.4% (229/1,589) of lincRNA loci and 28.5% (2,970/10,415) of protein-coding loci.

genomic element	LINE	LTR	DNA	Unclassified	Total
lincRNA	0.65%	0.95%	0.69%	1.20%	3.49%
protein	0.77%	0.41%	0.26%	0.20%	1.63%
CDS	1.37%	1.52%	0.03%	0.01%	2.93%
5'UTR	0.72%	0.13%	0.04%	0.00%	0.88%
3'UTR	0.69%	0.14%	0.06%	0.02%	0.91%
intron	1.15%	1.09%	0.47%	0.85%	3.57%
intergenic	1.95%	3.40%	0.92%	1.28%	7.56%
genome	3.15%	6.56%	0.88%	1.41%	11.99%

Table 5 – TE composition of lincRNAs and other genomic elements. Shown are the contributions (in % total nucleotide sequence) of four classes of TEs – LINEs, LTRs, DNA transposons, and unclassified TEs – to the non-redundant exons of 1,589 lincRNA loci and 10,415 protein-coding loci. Also shown are TE content for multiple annotated genomic elements from the *D. pseudoobscura* FlyBase r2.29 annotations.

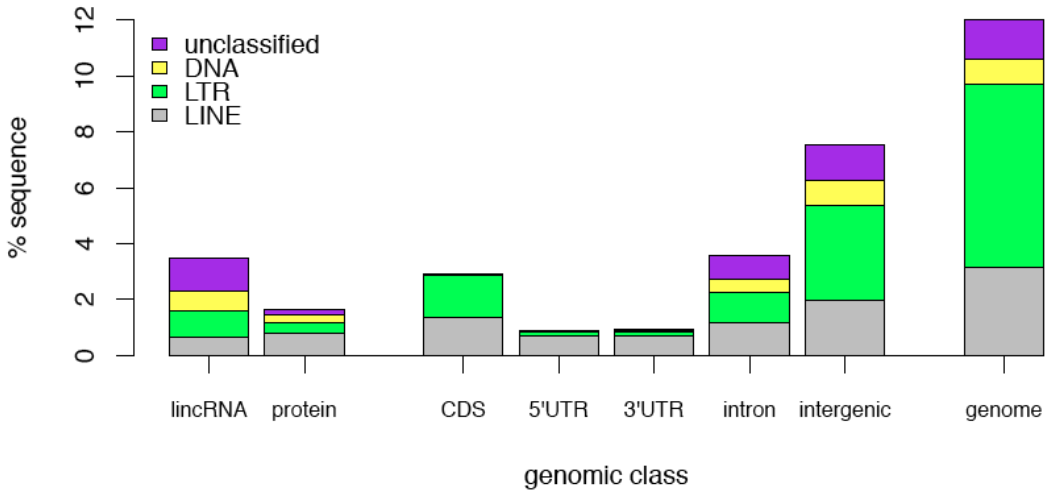


Figure 13 – TE content of lincRNAs and other genomic elements. Shown is TE content (% total sequence) for four classes of TEs (LINEs, LTRs, DNA transposons, and unclassified TEs) for non-redundant lincRNA and protein-coding exons. Also shown is TE content for five different genomic elements annotated in the *D. pseudoobscura* FlyBase r2.29 genome (CDS, 5' UTR, 3 UTR, introns, and intergenic sequence) and the complete genome.

Identification of multi-locus lincRNA families

We identified intraspecific multi-locus lincRNA families in *D. pseudoobscura* with sequence similarity via blastn (E-value < 1e-5). We identified 25 multi-locus lincRNA families in *D. pseudoobscura,* with most (60.0%, 15/25) consisting of two member loci. Six families had three loci, and a single family had five, six, 14, and 48 loci (Appendix, Table 1). Eight of these families had member loci located within 100kb of each other, suggesting that they could have arisen from tandem duplications. Seventeen families had member loci located far apart (>100kb) on the same chromosome or located on different chromosomes suggesting origination by transposition. Two families had member loci that could have arisen by both tandem duplication and transposition. We scanned the 17 transposition families for shared TEs that could explain their homology. Seven of the 17 families had shared TE sequence. Notably, of the 48 member loci in family 8, 35 (72.9%) include sequence from the HelitronN-1_DPe element.

We also looked for interspecific multi-locus families using the Rfam database (E-value <0.01), but found relatively few matches (Nawrocki et al., 2014; Nawrocki and Eddy, 2013). Five lincRNA loci showed matches to other annotated lincRNAs, including the three previously annotated lincRNAs in *D. pseudoobscura* FlyBase r2.29 (*RNaseP:RNA, SRP,* and *HSR-omega*). The two other loci matched a mammalian antisense long noncoding RNA in the *Hox* cluster (*HOTAIRM1_1*) and a mammalian lincRNA near the *Six*3 locus (*Six3os1_7*). We also found similarity between a number of our lincRNA loci and short noncoding RNAs. 14 lincRNA loci had matches to microRNAs, 13 lincRNA loci had matches to snoRNAs, seven lincRNA loci had matches to bacterial and yeast short RNAs, and a single lincRNA locus had a match to a tRNA.

DISCUSSION

High-throughput sequencing and increasingly sophisticated protein-coding identification methods have enabled the identification of thousands of lincRNAs. The vast majority of these efforts, however, have been in vertebrates or in classic genetic models, leaving the lincRNA biology of most of the eukaryotic phylogeny unexplored (Billerey et al., 2014; Boerner and McGinnis, 2012; Broadbent et al., 2011; Brown et al., 2014; Derrien et al., 2012; Inagaki et al., 2005; Jenkins et al., 2014; Kapusta and Feschotte, 2014; Kutter et al., 2012; Li et al., 2014a; Li et al., 2014b; Liu et al., 2012; Lu et al., 2011; Nam and Bartel, 2012; Necsulea et al., 2014; Pauli et al., 2012; Qu and Adelson, 2012; Weikard et al., 2013; Xie et al., 2014; Young et al., 2012). As an important and pervasive model system, extensive efforts have been made to identify lincRNAs in *D. melanogaster*, yet despite the vast genomic resources throughout the entire genus, almost nothing is known about lincRNAs in other *Drosophila* species (Brown et al., 2014; Young et al., 2012). Here, using extensive RNA-Seq datasets and the most current methods for the identification of protein-coding ability, we identified and characterized a large set of lincRNAs in a second species of *Drosophila*: the important evolutionary model *D. pseudoobscura*.

Identification of lincRNAs from RNA-Seq data

We assembled a transcriptome using RNA-Seq datasets taken from four developmental stages and adult gonad and carcass samples in both sexes. Since the *D. pseudoobscura* genome is poorly annotated with respect to lincRNAs, with only three annotated lincRNAs in FlyBase r2.29, we screened all novel intergenic transcripts that map to the major chromosome scaffolds for evidence of protein-coding ability using four different methods: (1) blastx alignment to the NCBI nr database, (2) alignments to *Drosophila*

proteomics databases, (3) conserved ORF identification using RNAcode, and (4) identification of noncoding sequence features using CPAT (Altschul et al., 1990; Desiere et al., 2006; Jiang et al., 2011; Wang et al., 2013; Washietl et al., 2011). We wanted to use a conservative approach to identify a high-confidence set of lincRNAs, so protein-coding signal from even a single method eliminated a transcript from contention as a putative lincRNA. After filtering and inclusion of previously annotated lincRNAs, we have identified a set of 1,771 lincRNA transcripts at 1,589 loci in *D. pseudoobscura*.

When tested against the annotated set of *D. pseudoobscura* protein-coding transcripts, three of these methods perform particularly well. Blastx against the NCBI nr database, RNAcode using the UCSC *Drosophila* alignments, and CPAT using both the *D. melanogaster* and *D. pseudoobscura* training sets all have sensitivities higher than 0.95. RNAcode using the Pseudobase alignments performs moderately well, with a sensitivity of 0.845. Only the proteomics dataset searches perform poorly, with sensitivities less than 0.3 using both the PeptideAtlas and a custom *D. pseudoobscura* dataset. This is particularly disappointing considering these proteomics searches are the only method that relies on direct observations at the peptide level.

Each method has its own particular strengths and weaknesses. NCBI nr and proteomics methods screen against previously annotated proteins, which are likely to be long and conserved. RNAcode and CPAT can identify proteins with no previous annotation. RNAcode analyzes variation across taxa for signals consistent with ORF conservation but has no direct dependency on length; thus, RNAcode can uniquely identify short peptides of only a few amino acids but does require a high quality multiple sequence alignment. CPAT does not require sequence from other taxa but does require

protein-coding and noncoding training sets and assumes complete transcript models. Our transcript models were generated using RNA-Seq without any targeted capture of the 5' and 3' ends of the locus. Low coverage locus models are likely to be incomplete. Even so, CPAT offers perhaps the best means to identify lineage-specific protein-coding transcripts.

We considered alternative methods that have been used previously to discriminate between protein-coding and noncoding transcripts. ORF length alone is an extremely poor predictor of protein-coding ability, as the majority of annotated *D. melanogaster* lincRNAs have complete ORFs greater than 50 amino acids. PhyloCSF identifies ORF conservation in ways very similar to RNAcode but requires a phylogenetic model provided by the developers (Lin et al., 2011). A *D. pseudoobscura*-based phylogeny is not provided, so we opted to use RNAcode. The Coding Potential Calculator (CPC), which uses homology and ORF properties, has previously been used to identify protein-coding ability in *D. melanogaster* transcripts, but we find its sensitivity lower than optimal on the *D. pseudoobscura* annotated protein-coding loci (Kong et al., 2007). Further, the features that CPC uses are redundant with features of RNAcode and CPAT, both of which perform better on our data. Ribosome profiling is another possible method for future protein-coding transcript identification, but is costly and not without criticism (Ingolia et al., 2009; Michel et al., 2014).

Of the 2,645 novel intergenic loci that we screened, protein-coding ability was evident in 1,059. Most of these loci were identified using only a single of our protein-coding identification methods. With low sensitivities against annotated protein-coding genes, it is not surprising that the proteomics methods found the fewest number (42) of

protein-coding transcripts among the novel intergenic loci. CPAT and the NCBI nr database performed similarly, finding 233 and 276 protein-coding loci respectively. RNAcode overwhelmingly found the largest number of loci with protein-coding signal (777), 621 of which were uniquely found via RNAcode. The calculated specificities for RNAcode using the annotated short noncoding RNA loci are high using both UCSC alignment (0.953) and the Pseudobase alignment (0.976), so we are skeptical that the unique performance of RNAcode is due to a substantially elevated false positive rate. We speculate that these loci are unannotated, avoiding detection from both NCBI nr and the proteomics datasets. They also either code for short peptides or have incomplete transcript models, avoiding detection via CPAT.

Existing methods for detecting protein-coding ability from transcript models are adequate. To better improve performance, we would focus on improving the existing genome resources. RNA-PET, CAGE, or SuperSAGE would enable the capture of the 5' and 3' ends of a transcript and facilitate the completion of transcript models (Fullwood et al., 2009; Matsumura et al., 2010; Takahashi et al., 2012). This should greatly improve CPAT performance. Further, RNAcode using the UCSC-based alignment performed better on the set of annotated transcripts than RNAcode using the Pseudobase alignments but was limited by the number of loci whose coordinates could successfully be converted from *D. pseudoobscura* to *D. melanogaster*. We reason that the increased divergence times between taxa in the UCSC alignment (e.g. 25-55 million years for *D. pseudoobscura-D. melanogaster* divergence versus 5-11 million years for *D. pseudoobscura-D. lowei*) provided more useful variation for RNAcode to discriminate (Beckenbach et al., 1993; Richards et al., 2005). Thus, a multiple genome alignment

using taxa from the entire *Drosophila* genus but aligned to the *D. pseudoobscura* genome would be optimal for RNAcode performance.

D. pseudoobscura *lincRNAs display many typical lincRNA features*

Large sets of lincRNAs have now been described in a number of eukaryotic species (Kapusta and Feschotte, 2014). The *D. pseudoobscura* lincRNAs that we describe here display a number of features that are typical of lincRNAs in other systems. While longer than "classic" noncoding RNAs, the *D. pseudoobscura* lincRNAs, on the whole, are shorter than protein-coding transcripts (Derrien et al., 2012; Li et al., 2014b; Pauli et al., 2012; Young et al., 2012). They tend to have fewer exons, and while alternative splicing is observed, it is rare (Derrien et al., 2012; Li et al., 2014b; Pauli et al., 2012; Young et al., 2012). LincRNA exonic sequence is lower in GC content than protein-coding sequence and contains higher proportions of simple sequence repeats and low-complexity sequence (Niazi and Valadkhan, 2012). There is still little consensus on how, or even if, the majority of lincRNAs function. Interestingly, the common features of lincRNAs across diverse taxa suggest that there are distinct forces that drive lincRNA evolution, even if they are primarily derived from the lack of amino acid coding constraint.

LincRNA loci are overrepresented on the 4th chromosome compared to protein-coding loci

We compared the distribution of lincRNA loci to protein-coding loci among the five major chromosome arms and within each chromosome in *D. pseudoobscura*. We found no significant differences in distributions among four of the chromosomal arms (XL, XR, 2, and 3), indicating that there is no apparent bias for lincRNAs for or against the X chromosome, and no significant differences between regions near the centromeres and

telomeres and middle regions of the chromosomes. We did, however, find an overrepresentation of lincRNA loci on the 4th chromosome. To our knowledge, only one key feature differentiates the 4th chromosome from the other autosomes. It is the only major chromosome arm in *D. pseudoobscura* that lacks a fixed or nearly fixed inversion that severely reduces gene flow between *D. pseudoobscura* and its sympatric sister species *D. persimilis* (Machado et al., 2007; Noor et al., 2007; Noor et al., 2001b). The 4th chromosome, thus, would have a larger effective population size and an evolutionary history more strongly driven by selection than any of the other chromosomes. Considering the paucity of knowledge about the process of the origination or maintenance of lincRNAs in *Drosophila*, we speculate that their overrepresentation in a given chromosome is maintained by selection. To further explore this, one could compare signals of selection for lincRNAs found on the 4th chromosome versus lincRNAs found elsewhere in the genome.

Of course, the overrepresentation of lincRNAs on the 4th chromosome could also be an artifact of poor genome assembly. While there are no reported assembly issues specific to the 4th chromosome, that does not necessarily mean they do not exist. A larger number of gaps or repetitive elements on the 4th could easily lead to fragmented gene models that artificially inflate the numbers of newly annotated loci.

Finally, we mention that we do not analyze any lincRNAs that may appear on the largely heterochromatic Y or dot (i.e. 5th) chromosomes, as their scaffolds were not identified in the *D. pseudoobscura* (FlyBase r2) genome. In *D. melanogaster* (FlyBase r6.02), there are higher densities of lincRNAs on the nonhomologous Y and the homologous dot (i.e. 4th) chromosomes, with the Y actually having a higher number of

lincRNA loci than protein-coding loci (St Pierre et al., 2014). We would expect to see similar patterns in the heterochromatic Y and dot chromosomes in *D. pseudoobscura*.

***TEs are not major contributors to lincRNA sequence in* D. pseudoobscura**

A recent study in vertebrates showed that TEs were major contributors to lincRNA sequence, with TEs being found in more than 65% of annotated lincRNA transcripts in humans, mice, and zebrafish and comprising between 15.3% and 35.1% of total lincRNA exon sequence (Kapusta et al., 2013). The contributions of TEs to *D. pseudoobscura* lincRNAs, however, are more modest. Using a set of annotated *Drosophila* TEs, we find that only 14.4% (229/1,589) of all *D. pseudoobscura* lincRNA loci contain a TE sequence, and TEs cover only 3.49% of lincRNA exon sequence. To confirm the reduction in TE content, we ran the same analyses on the set of *D. melanogaster* lincRNAs (FlyBase r6.02) and found similarly depressed values for TE content. TEs are found in 9.6% (267/2776) of lincRNA transcripts and cover 1.79% of *D. melanogaster* lincRNA exon sequence. Interestingly, *D. pseudoobscura* protein-coding transcripts, particularly the CDS, are tolerant of LTR and LINE retrotransposons while vertebrate CDS are devoid of virtually any TEs. Considering that TE content is seen in both classes of genes, we speculate that unique TE contributions to lincRNA biology in *D. pseudoobscura* are minimal.

Genic sequence is necessarily a product of its genomic environment, so perhaps it is not surprising that TEs are less prevalent in *D. pseudoobscura* lincRNAs. TEs cover only 11.99% of the entire *D. pseudoobscura* genome while TE coverage in vertebrate genomes is far higher, with TEs covering 49.3% of the human genome (release hg19), 40.8% of the mouse genome (release mm10), and 46.2% of the zebrafish genome (release

danRer7) (Kapusta et al., 2013). The TE-rich vertebrates have both larger proportions of lincRNAs with TEs but also larger numbers of lincRNAs themselves. For example, 9,518 lincRNA transcripts from 5,094 loci are listed in the human GENCODE v7 annotation (Derrien et al., 2012). Kapusta et al. (2013) argue that TEs are key drivers in the origination and diversification of lincRNAs in vertebrates.

If this holds for all eukaryotes, then we expect that all taxa with relatively low numbers of lincRNAs will also have low TE content. This is largely observed in the few taxa where both types of data are available. *D. melanogaster*, the nematode *C. elegans*, the mosquito *A. gambiae*, the budding yeast *S. cerevisiae*, the mushroom *G. lucidum*, and plasmodium all have less than 3,000 annotated lincRNAs and genomic TE coverage less than 25% (Kapusta and Feschotte, 2014). Even so, it is clear that there are other mechanisms that drive lincRNA origination and evolution. In plants, over 20,000 lincRNAs are seen in *Z. mays*, which has a high genomic TE coverage of over 85%, but high numbers of lincRNAs (>6,000) are also seen in *A. thaliana*, which has a much more modest genomic TE coverage of about 10% (Arabidopsis Genome, 2000; Li et al., 2014b; Liu et al., 2012; Schnable et al., 2009).

Intra- and interspecies *D. pseudoobscura* multi-locus lincRNA families are rare

Of the 1,589 lincRNA loci in *D. pseudoobscura*, only 125 belong to a multi-locus lincRNA family, suggesting that the vast majority (92.1%, 1,464/1,589) evolved *de novo*. These 125 loci are grouped into 25 multi-locus lincRNA families, although 48 belong to a single family whose transposition is likely driven by the HelitronN-1_DPe TE. There is evidence of both tandem duplication and transposition between family loci, with transposition occurring more frequently. Because of this, we cross-referenced the

transposed multi-locus families with the TEs identified via Repeatmasker and found seven families with common TEs across multiple loci. While the overall contributions of TEs are modest in *D. pseudoobscura* lincRNAs, it does appear that the presence of TEs can drive some level of lincRNA diversification.

Using the Rfam database, we also wanted to see if we could identify interspecies lincRNA families but were largely unsuccessful. We identified five lincRNAs with significant hits (E-value < 0.01) to annotated lincRNAs in the Rfam database. The three strongest of these hits were, not coincidentally, the three previously annotated lincRNAs in the *D. pseudoobscura* genome (r2.29): *RNaseP:RNA, SRP,* and *HSR-omega*. The other two hits, *HOTAIRM1_1* and *Six3os1_7*, had E-values just under the cutoff (0.0038 and 0.0073, respectively). LincRNA annotations are rare outside of vertebrates, and we need greater phylogenetic sampling before we can ascertain whether these are true homologous lincRNAs. We revisit the idea of finding interspecific lincRNA families in Chapter 3 via a more targeted search with *D. melanogaster*.

Our Rfam searches also found a small number of alignments to multiple types of short noncoding RNAs. Most short noncoding RNAs, including microRNA, snoRNAs, and piRNAs, are processed from longer transcripts that fit the classic definition of a lincRNA though are often excluded from lincRNA datasets. The vast majority of our lincRNA loci, however, did not match a short noncoding RNA, which leads us to two conclusions: (1) since we only searched for putative lincRNAs in novel transcripts, the low number of Rfam hits suggests that the *D. pseudoobscura* annotation is of high quality with respect to short noncoding RNAs; and (2) the biological roles of the majority of

these lincRNAs, if any, will likely not involve the biogenesis of most types of short noncoding RNAs, the Rfam-deficient piRNAs excepted.

METHODS

Fly rearing and RNA extraction

All *D. pseudoobscura* flies used to generate RNA-Seq libraries were from the MV2-25 line that was originally collected in Mesa Verde, Colorado and inbred for 15 generations (Richards et al., 2005). Flies were kept in incubators maintained at 20°C on a 12h/12h light/dark cycle. They were maintained in polypropylene bottles and fed a diet of molasses/agar/cornmeal/yeast supplemented with Tegosept and propionic acid to minimize mold. To generate developmentally-staged flies for sequencing, embryos were collected over 4-hour windows in the same bottles, and embryos were then physically transferred to fresh bottles at a density of 100 embryos per bottle to maintain uniform density and prevent crowding. All flies were collected between the hours of 8pm and 12am, shortly after the start of the dark cycle, and ground in Trizol reagent (Life Technologies #15596-026) with polyacryl (Molecular Research Center PC 152) (1:1000) on ice. Flies were collected using the following criteria: (1) 1st-instar larva – flies were collected 34-40 hours post-laying, shortly after hatching. (2) wandering 3rd-instar larva – bottles were monitored for first evidence of wandering 3rd-instar larvae in the morning; flies were collected later that evening. (3) mid-pupa – bottles were monitored for new yellow pupae in the morning, and mid-pupae were collected two days later (post-head eversion with green malphigian tubules and no visible eye pigmentation). (4) 7-day adult – Virgin males and females were isolated shortly after eclosion, and adults were collected on the 7th day after eclosion. Ovary and testis dissections were also performed on 7-day

adult flies, with the gonads and resulting carcasses all ground up in Trizol:polyacryl (1:1000). Male accessory glands were removed from the testes and included with the male carcass samples.

Sex was determined visually in adults using genitalia and testes pigmentation. Sex-specific whole-body adult flies and dissected gonads and carcasses (n=20) were all pooled and ground separately in Trizol:polyacryl and frozen at -80°C overnight. RNA extractions were performed using the standard Trizol protocol.

For all other stages, flies were collected individually and genotyped using a pair of X-chromosome primers to indicate a good extraction (CG10274-F1 5'-CTGTGGCAAGCGGTTCGTG-3', CG10274-R2 5'-CACGTCGCGGATCCTTGGGTA-3') and a pair of Y-chromosome primers to distinguish males from females (CG12218Y-F 5'-GCAGTCGAACCAGTGCAAT-3', CG12218Y-R 5'- GTGCGGGCAATGGATAAT-3') (Carvalho and Clark, 2005). After collection, flies were frozen overnight at -80°C. Trizol phase separation was performed with chloroform/isoamyl alcohol (Sigma #25668-100ML) followed by centrifugation per standard protocol. The aqueous phase containing RNA was carefully removed and stored at -80°C until genotyping was complete. DNA from the leftover organic/interphase was carefully cleaned with the following protocol: (1) 0.5uL polyacryl and 0.3X volume 100% ethanol was added to the organic/interphase, mixed and incubated for 3 min. at room temperature, and centrifuged for 15 minutes at 2200 rcf. (2) Supernatant was removed, and pellet was washed three times with 0.1M sodium citrate/10% ethanol, pH 8.0 for 30 minutes. Samples were centrifuged for 15 minutes at 2200 rcf after each wash. (3) The pellet was washed one time with 75% ethanol for 20 minutes and then

centrifuged for 15 minutes at 2200 rcf. (4) DNA was resuspended in 20uL of 8mM sodium hydroxide and left on a rotater overnight. Samples were then heated for 10 minutes at 50°C, and the final pH was adjusted by adding 1.3uL of 0.1M HEPES. PCR was performed on this cleaned up DNA with the aforementioned primers: 2.5uL 10x PCR buffer, 0.6uM each of CG10274-F1, CG10274-R2, CG12218Y-F, and CG12218Y-R primers, 10uL DNA, 0.5uL 10mM dNTPs, 0.1uL HotMaster Taq polymerase (5 Prime #2200300) in a 25uL reaction. PCR reaction conditions were as follows: (1) initial denaturation at 94°C for 2 minutes; (2) 40 cycles at denaturation at 94°C for 15 sec., annealing at 60°C for 30 sec., and extension at 65°C for 1 minute; (3) final extension at 65°C for 5 min. Individual flies that showed one clear band at ~500bp were classified as female, and individual flies that showed two clear bands at ~500bp and ~700bp were classified as male. Individual flies not showing any bands were discarded. Male and female RNA was then pooled (n=20 for L3 and pupa, n=35 for L1 male, n=49 for L1 female) and cleaned using standard isopropanol washing protocols.

Poly(A+) library construction and RNA sequencing

Total RNA was DNase treated (Life Technologies #18068-015) and then cleaned using a Qiagen RNeasy Mini cleanup column (#74106). Total RNA concentration and quality were then determined using a Bio-Rad Experion Total RNA Stdsens Assay (RQI > 8.0) (#700-7103). Poly(A+) RNA-Seq libraries were constructed using the TruSeq RNA prep kit (Illumina RS-122-2001), with 1ug of input RNA for gonad and carcass samples (i.e. testes, ovaries, male carcass, female carcass) and 750ng of input RNA for all whole-body samples from the developmental series (i.e. male and female 1st-instar larvae, 3rd-instar larvae, mid-pupae, and 7-day adults). 100bp, paired-end sequencing was performed on an

Illumina HiSeq1000 machine at the University of Maryland's Institute for Bioscience and Biotechnology Research Sequencing Core.

Transcriptome assembly

Raw sequence reads were filtered for quality using the NGS QC Toolkit (Patel and Jain, 2012). Raw reads with an average PHRED quality score less than 20 were thrown out using IlluQC_PRLL.pl. These filtered reads were then trimmed of low-quality bases (PHRED < 20) from the 3' end using TrimmingReads.pl. For each of the 12 samples, filtered, trimmed reads were aligned to the *D. pseudoobscura* genome (FlyBase r2) using TopHat v2.0.5 (mate inner distance = -20, mate standard deviation = 50, minimum intron length = 20, minimum segment intron length = 20, max multihits = 1, all other options = default) (Kim et al., 2013; Langmead and Salzberg, 2012; St Pierre et al., 2014). Transcriptomes were then assembled from the TopHat accepted_hits.bam file for each sample using Cufflinks v2.0.2 (minimum intron length = 20, overlap radius = 20, all other options = default) (Trapnell et al., 2010). The 12 individual transcriptomes were then merged into a single, comprehensive transcriptome using the Cuffmerge command via Cufflinks and the *D. pseudoobscura* FlyBase r.2.29 annotation as a guide. All transcripts were labeled as either annotated or unannotated via Cuffmerge.

Computational identification of lincRNAs

All transcripts from the 2,645 novel intergenic loci (Cuffmerge class code "u") that map to the five major *D. pseudoobscura* chromosomal scaffolds (XL, XR, 2, 3, and 4) (FlyBase r2) were screened using the following four methods. Any locus with one or more transcripts that show evidence of protein-coding ability using any of the four methods was eliminated as a putative lincRNA locus. Except where otherwise specified,

sensitivity (i.e. true positive rate) was calculated using the 16,761 transcripts at 10,415 annotated protein-coding loci (Cuffmerge class code "="), and specificity (i.e. true negative rate) was calculated using the 718 transcripts at 699 annotated noncoding loci (*D. pseudoobscura* r2.29). Noncoding loci include rRNAs, tRNAs, miRNAs, snRNAs, snoRNAs, and three long noncoding RNAs. All command options assumed to be default unless otherwise specified. All Venn diagrams were created using VENNY (Oliveros, 2007).

(1) Blastx against the NCBI non-redundant protein database

Transcripts were aligned to the NCBI nr database using blastx (BLAST+ 2.2.28, E-value < 1e-10, output format = '6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore staxids') (Altschul et al., 1990; Camacho et al., 2009). In theory, novel intergenic loci should not match annotated loci in *D. pseudoobscura*. To avoid throwing out possible degenerated paralogs of protein-coding genes, we only considered blastx matches to non-*D. pseudoobscura* genes.

(2) Local alignments to Drosophila *proteomics datasets*

Transcripts were aligned to the PeptideAtlas *D. melanogaster* proteomics dataset (Aug. 2012 build) using blastx (BLAST+ 2.2.28, E-value < 1e-5, output format = 6, maximum target sequences = 1) (Camacho et al., 2009; Desiere et al., 2006).

The longest ORF for each transcript (minimum of 10 amino acids) was calculated and translated using custom perl scripts written by Josie Reinhardt. These putative peptide sequences were then matched against unpublished testes proteomics datasets from the MV2-25 line of *D. pseudoobscura*, the Susa6 line of *D. ps. bogotana*, and hybrid crosses between the two subspecies that were generated by Zi-Feng Jiang. Each

peptide fraction was injected into a Pepmap C18 trapping cartridge (0.3 × 5 mm, Dionex, Bannockburn, IL) with autosampler, and desalted with 100% solvent A (0.1% formic acid with 5% acetonitrile) at 10 µL/min for 15 min. Peptides were separated using a Zorbax 300 SB-C18 nano column (3.5 µm, 0.075 × 150 mm, Agilent Technologies, Palo Alto, CA) with a binary gradient consisting of solvent A (0.1% formic acid with 5% acetonitrile) and B (0.1% formic acid and 95% acetonitrile) at a flow rate of 300 nL/min. The gradient was run from 10% B to 45% B over 60 minutes, followed by a 5-minute wash step with 80% B and 5-minute equilibration at 0% B. Positive ion mass spectra of Nano LC eluents were acquired with a Thermo Finnigan LTQ Orbitrap XL mass spectrometer. A full Scan FT analysis of 400-2000 Daltons with resolution at 60,000 (m/z 400) was followed by up to 5 MSMS analysis in the linear ion trap (CID) at unit mass resolution.

Putative peptide sequences were searched against the in-house curated peptide database using Mascot and Sequest HT through Proteome Discoverer 1.4 software package (Thermo Fisher) in MUDPIT mode, where spectra from all data files are combined and searched as one file to maximize protein identification. Cysteine carbomidomethylation was set as fixed modification. Peptide mass tolerance was 50 ppm and fragment mass tolerance was 0.8 Da. Peptide identification was validated by Percolator and peptides with false discovery rate < 1% are reported.

(3) Identification of conserved ORFs using RNAcode

RNAcode (v0.3) ($p < 0.05$, number of simulations to calculate p-value = 1000, output format = GTF, stop-early = YES) was used to identify signatures of ORF conservation using both the UCSC 15-species *Drosophila* and two iterations of the Pseudobase *D.*

pseudoobscura subgroup multiple genome alignments (Kuhn et al., 2007; McGaugh et al., 2012; McGaugh and Noor, 2012; Noor, 2012; Washietl et al., 2011).

To extract alignments from the UCSC alignment, we needed to convert genome coordinates of transcripts from *D. pseudoobscura* (FlyBase r2 or UCSC Dp4) to *D. melanogaster* (BDGP r5 or UCSC Dm3). First, *D. pseudoobscura* coordinates in the Cuffmerge merged.gtf file were converted to the recommended genePred format using the UCSC gtfToGenePred tool. Coordinates were then converted using liftOver and the dp4ToDm3.over.chain file (input = genePred, allow multiple output regions = YES, minimum match = 0.1) (Hinrichs et al., 2006). Finally, individual transcript alignments were extracted using the maf_parse tool from the Phast package (v1.1) (Hubisz et al., 2011). Final input for RNAcode was in MAF format.

To extract alignments from Pseudobase, alignments for all exons in multi-fasta format were generously batch extracted by Ryan Hardy using *D. pseudoobscura* (FlyBase r2) coordinates taken from the Cuffmerge merged.gtf file. Exons were joined to form complete transcripts using a custom perl script. The multi-fasta format was then converted to ClustalW format using a custom perl script that utilized the Align:IO module from BioPerl (Stajich et al., 2002). A second iteration of the Pseudobase alignment was created using only a single line from four species using a custom perl script. The lines kept in the reduced alignment are: *D. pseudoobscura* MV2-25, *D. persimilis* MSH1993, *D. miranda* MAO, and the single *D. lowei* line. Final input for RNAcode was in ClustalW format.

(4) Identification of noncoding sequence features using the Coding Potential Assessment Tool (CPAT)

CPAT was used to identify sequence features specific to protein-coding transcripts (Wang et al., 2013). When transcript orientation was known, CPAT was only run on the three reading frames in the direction of transcription. Otherwise, CPAT was run on all six possible reading frames. CPAT was initially run using the provided *D. melanogaster* logistic regression model and hexamer frequency tables and a coding probability cutoff of 0.39. A *D. pseudoobscura* specific logistic regression model was also built. A hexamer frequency table was built using the set of *D. pseudoobscura* CDS and a set of all noncoding sequences including 5' and 3' UTRs, introns, and all annotated noncoding RNAs (FlyBase r2.30). The logistic regression model was then built using this hexamer frequency table and trained on the set of 16,761 annotated protein-coding transcripts (Cuffmerge class code "=") and a high-confidence set of 418 *D. pseudoobscura* lincRNA transcripts that had passed all previously mentioned filters (blastx to NCBI nr, both proteomics database searches, RNAcode using the UCSC alignment and both iterations of the Pseudobase alignments, and CPAT using the *D. melanogaster* logistic regression model). A coding probability cutoff of 0.93 was determined empirically from the training set using a TG-ROC R script provided by Liguo Wang utilizing the ROCR package (Sing et al., 2005).

Transcript and sequence properties of lincRNAs

Total exonic length, exon numbers, and alternative transcription of lincRNAs

A set of non-redundant exons was generated from the merged.gtf file for each locus with a modified custom R script obtained from Devon Ryan. A fasta file was then generated using the gffread utility in Cufflinks, and total exon length was determined using the perl script fastaNamesSizes.pl written by Lionel Guy (Trapnell et al., 2010). Isoform number

was calculated from the merged.gtf file. For the 10,415 annotated protein-coding loci from *D. pseudoobscura*, all transcripts with the Cuffmerge class codes of "=", "j", and "o" were used.

Genomic locations of lincRNAs

Genomic coordinates were determined by concatenating and modifying FlyBase r2.29 scaffolds where necessary. Chromosomes 2 and 3 consist of a single scaffold, so no modifications were necessary. Chromosomes 4, XL, and XR scaffolds were concatened in the order shown in published cytogenetic maps (Schaeffer et al., 2008; Schaeffer et al.). Note that some scaffolds needed to be broken before concatenation and that portions of XL_group3a and XL_group1a and all of XL_group3b actually map to XR. Chi-square tests were performed using all variables. If significance ($p<0.05$) was found, pair-wise comparisons were tested using a modified Bonferroni-corrected p-value (Keppel, 1991).

Sequence composition of lincRNAs

GC content was calculated using the count_fasta.pl script written by Joseph Fass. A fasta file containing the set of non-redundant exons was used for calculating GC content at lincRNA and protein-coding loci. GC content for all other classes was calculated using annotated fasta files from the *D. pseudoobscura* genome (FlyBase r2.29).

Simple repeat, low-complexity sequence, and TE content was calculated using RepeatMasker v4.0.5 with cross_match v0.990329 and the *D. melanogaster* set of transposable elements against the set of non-redundant exons for lincRNA and protein-coding loci. TE content for all other classes was calculated using annotated fasta files from the *D. pseudoobscura* genome (FlyBase r2.29).

Identification of lincRNA families

To identify lincRNA families within *D*. pseudoobscura, we aligned the set of non-redundant lincRNA exons against themselves using blastn (BLAST+ 2.2.28, E-value < 1e-5, output format = '6 std stitle') (Altschul et al., 1990; Camacho et al., 2009). Self-hits and bidirectional loci were removed, and any matches were grouped into families using custom perl scripts.

We also searched the Rfam database for any matches to our lincRNA loci (non-redundant exon set) using cmscan in Infernal v1.1 (E-value < 0.01) (Nawrocki et al., 2014; Nawrocki and Eddy, 2013).

CHAPTER 2: Expression dynamics of *D. pseudoobscura* lincRNAs

ABSTRACT

Long intergenic noncoding RNAs (lincRNAs) have been shown to have distinct expression properties in a number of species. Here, we analyze lincRNA expression in *D. pseudoobscura* using RNA-Seq datasets collected in different sexes, at multiple developmental stages, and in isolated adult gonad tissues. In agreement with studies from other species, we find that lincRNAs are expressed at much lower levels than protein-coding genes. After clustering all expressed genes by their developmental expression profiles, we find an overrepresentation of lincRNAs among genes that progressively increase in expression in male development and among genes that are most-highly expressed in the pupal stage in both sexes. Gene Ontology analysis of these clusters finds lincRNAs co-expressed with protein-coding genes that have roles in male-specific processes as well as more ubiquitous and fundamental biological processes like cell adhesion and transcriptional regulation. While overall levels of sex-bias are comparable between lincRNAs and protein-coding genes throughout development, sex-bias is skewed overwhelmingly towards males and can be explained by high levels of lincRNA expression in the testes. Finally, we detect both underrepresentation of male-biased lincRNAs and overrepresentation of female-biased lincRNAs on the X chromosome, consistent with models of selection that favor demasculinization and feminization of the X. Testis-specific lincRNAs, however, are distributed evenly between the X and autosomes, suggesting there are functional subdivisions within the lincRNA repertoire.

INTRODUCTION

Analyses of lincRNA expression in numerous, though mostly vertebrate, species indicate that lincRNAs have distinct expression properties from protein-coding genes (Akbari et al., 2013; Brown et al., 2014; Cabili et al., 2011; Derrien et al., 2012; Necsulea et al., 2014; Ulitsky and Bartel, 2013; Young et al., 2012). In general, lincRNAs are expressed at lower levels than mRNAs; these expression differences cannot be explained by increased instability of the long noncoding molecules (Clark et al., 2012). LincRNA expression has been observed in every tissue thus surveyed but tend to display more tissue-specific expression than protein-coding transcripts (Derrien et al., 2012; Necsulea et al., 2014; Pauli et al., 2012; Young et al., 2012). From these studies, two tissue types repeatedly show high levels of lincRNA expression: nervous system tissue and testis.

Our knowledge of lincRNA expression dynamics in *Drosophila* comes primarily from the extensive *D. melanogaster* RNA-Seq data generated by the modENCODE project, which includes various isolated tissues and a 30-time point developmental series covering all four major stages of the fly life cycle (Graveley et al., 2011). Two independent efforts were made to identify lincRNAs from these data; together, they show that lincRNAs in *Drosophila* display many of the same properties as seen in vertebrates (Brown et al., 2014; Young et al., 2012). They tend to be expressed at lower levels than protein-coding genes. Likewise, lincRNAs are more likely to be expressed in a developmental or tissue-specific context, and an overabundance show highest expression in testes.

More recently, a study of sex-biased lincRNAs in *D. melanogaster* found that, like male-biased protein-coding genes, male-biased lincRNAs are also underrepresented

on the X chromosome (Gao et al., 2014). The three major models that could explain this demasculinization of the X chromosome all invoke selection. Inactivation of the X chromosome during meiosis could favor the accretion of functionally-important male-biased genes on the autosomes (Gao et al., 2014). X chromosome dosage compensation via hypertranscription could constrain the upper limits of gene expression, preventing further upregulation and favoring accumulation of male-biased genes on the autosomes (Bachtrog et al., 2010; Vicoso and Charlesworth, 2009). Lastly, the sexual antagonism hypothesis predicts that sex-biased genes with advantageous fitness effects in one sex and detrimental fitness effects in the other would be favored to have different chromosomal distributions, with dominant male-biased genes accumulating on autosomes and female-biased genes accumulating on the X, as the X spends two-thirds of its time in females (Charlesworth et al., 1987; Rice, 1984). By these models, an underrepresentation of *D. melanogaster* male-biased lincRNAs on the X would imply that they possess some sort of advantageous male function.

Analyses of lincRNA expression in non-*melanogaster* species of *Drosophila* are almost nonexistent. Thorough analyses of the expression dynamics of lincRNAs in *D. pseudoobscura* would serve as another data point that expands the scope of lincRNA research beyond vertebrates and facilitate comparisons of lincRNA expression between two moderately-diverged species within the *Drosophila* genus, *D. pseudoobscura* and *D. melanogaster*. The *pseudoobscura* subgroup contains both an allopatric subspecies pair (*D. ps. pseudoobscura* and *D. ps. bogotana*) with incipient reproductive isolation and a sympatric species pair (*D. ps. pseudoobscura* and *D. persimilis*) that shows hybrid male sterility (Ayala and Dobzhansky, 1974; Dobzhansky, 1936; Dobzhansky, 1937;

Dobzhansky et al., 1963; Noor et al., 2001b; Orr, 1989a, b; Orr and Irving, 2001). The *pseudoobscura* species complex is an ideal system for studying the contributions of lincRNAs to species divergence, and while we do not yet expand our analyses to *D. persimilis* and *D. ps. bogotana*, we conduct our analyses of lincRNA expression dynamics in *D. pseudoobscura* with that in mind.

Previous work in our lab using cDNA and microarray technology identified 10 novel lincRNAs in *D. pseudoobscura* and its sympatric sister species *D. persimilis* (Jiang et al., 2011). Three of these lincRNAs were expressed at high levels in the testes. All three of these testes-biased genes were also differentially expressed between the two species, raising intriguing questions about possible contributions of lincRNAs to the hybrid male sterility that isolates these two species.

We set out to analyze lincRNA expression dynamics using replicated RNA-Seq both over the course of development and in the fully-developed gonads, both because of their unique lincRNA expression properties in other species and their importance for future evolutionary work. We chose four key stages: 1st-instar larvae, wandering 3rd-instar larvae, mid-pupae, and 7-day adults. The two larval stages are important with respect to gonad development. The former precedes gonad development, while the latter roughly coincides with gonad development in both *D. melanogaster* and *D. pseudoobscura* (Bate and Martinez Arias, 1993; Noor et al., 2001b; Orr, 1989a, b; Orr and Irving, 2001). In fact, spermatogenesis is already underway by the 3rd-instar. Because sex-transcriptomes can significantly diverge, we thought it prudent to collect our developmental samples in a sex-specific manner (Abdilleh, 2014; Jiang and Machado, 2009). As the modENCODE group only collected sex-specific RNA in adults, these data

will offer key insights not only into how lincRNAs are expressed through development but also how global sex-specific transcriptomes change throughout development (Graveley et al., 2011).

With the set of *D. pseudoobscura* lincRNAs previously identified in Chapter 1 and three biological RNA-Seq replicates for each sample, we explore the differences in lincRNA and protein-coding gene expression throughout development. We examine how sex-biased lincRNA and protein-coding gene expression changes as development proceeds, and we examine lincRNA and protein-coding contributions to both the testes and the ovaries transcriptomes. Finally, we analyze how sex-biased expression influences the distributions of genes on the autosomes and X chromosome and discuss the models that might explain those observations.

RESULTS

Generating expression datasets via RNA-Seq

Sequencing of additional RNA-Seq replicates

In order to adequately analyze the expression dynamics of lincRNAs through development and in different tissues, we needed to generate replicate datasets for each of the samples collected in Chapter 1. To that end, we performed poly(A+) RNA-Seq on two additional replicates of all 12 distinct samples from the inbred MV2-25 line at a slightly lower sequencing depth (1/8 of an Illumina lane, Table 1) (Richards et al., 2005). These include whole body samples of 1st instar larvae, wandering 3rd instar larvae, mid-pupae, and 7-day adults and dissected gonads and carcasses of 7-day adults, all of which are separated by sex. In sum, we have three biological RNA-Seq replicates of the MV2-25 line for each of 12 sample types.

Read quality and mapping efficiencies were similar for the lower-depth RNA-Seq replicates as for the initial RNA-Seq datasets described in Chapter 1 (Table 1). Between 86.7% and 93.1% of all raw mate pairs had an average PHRED score greater than 20, and between 88.3% and 96.4% of all high quality mate pairs had at least one read that mapped to the *D. pseudoobscura* genome (FlyBase r2) (St Pierre et al., 2014).

Sample	Raw mate pairs	HQ mate pairs	Mapped fragments
L1M_B	26,051,697	23,345,792 (89.6%)	21,498,101 (92.1%)
L1F_B	20,757,012	18,916,448 (91.1%)	17,554,289 (92.8%)
L3M_B	15,905,301	14,371,101 (90.4%)	13,134,148 (91.4%)
L3F_B	25,224,420	22,255,479 (88.2%)	20,221,486 (90.9%)
PupM_B	23,931,300	21,731,345 (90.8%)	20,539,397 (94.5%)
PupF_B	18,403,541	15,952,928 (86.7%)	14,937,457 (93.6%)
AdM_B	20,241,339	18,280,356 (90.3%)	16,711,617 (91.4%)
AdF_B	23,932,296	21,644,600 (90.4%)	20,228,980 (93.5%)
carcM_B	23,002,966	21,412,524 (93.1%)	18,905,500 (88.3%)
test_B	27,591,265	25,578,655 (92.7%)	24,616,176 (96.2%)
carcF_B	28,273,602	26,170,638 (92.6%)	23,918,463 (91.4%)
ov_B	24,643,356	22,764,629 (92.4%)	21,770,693 (95.6%)
L1M_C	20,012,101	18,270,111 (91.3%)	16,865,522 (92.3%)
L1F_C	23,698,742	21,664,507 (91.4%)	19,977,142 (92.2%)
L3M_C	22,517,230	20,117,471 (89.3%)	18,350,232 (91.2%)
L3F_C	16,425,999	14,620,450 (89.0%)	13,091,880 (89.5%)
PupM_C	17,963,951	16,432,851 (91.5%)	15,465,700 (94.1%)
PupF_C	20,696,117	18,853,309 (91.1%)	17,710,530 (93.9%)
AdM_C	23,633,057	21,329,701 (90.3%)	19,459,420 (91.2%)
AdF_C	20,965,737	18,681,114 (89.1%)	17,382,249 (93.0%)
carcM_C	22,331,736	20,717,591 (92.8%)	18,505,160 (89.3%)
test_C	24,013,590	22,238,254 (92.6%)	21,431,986 (96.4%)
carcF_C	22,687,950	21,058,104 (92.8%)	19,114,225 (90.8%)
ov_C	19,432,333	17,984,626 (92.6%)	16,952,136 (94.3%)

Table 1 – RNA-Seq replicate sample statistics – Shown are sequencing, quality control, and mapping statistics for each of the additional 24 RNA-Seq libraries generated for expression analyses. "Raw mate pairs" refers to the total number of fragments sequenced with Illumina paired-end sequencing. "HQ mate pairs" refers to the number of raw mate pairs with average PHRED score > 20. "Mapped fragments" refers to the number of high-quality mate pairs, either both mate pairs or only one mate pair, that map to the *D. pseudoobscura*

Quality checks of RNA-Seq datasets

To determine the quality and consistency of the biological replicates, we created multidimensional scaling (MDS) plots using locus expression values for each replicate. Fragment counts for every gene locus were calculated using HTSeq-count v0.6.1p1 (Anders et al., 2014). Because we did not use strand-specific RNA-Seq, we only counted fragments that map unequivocally to a single locus. Fragment counts were scale normalized across samples, and normalized fragment counts were used as input for MDS. High-quality replicates should cluster distinctly on the MDS plot. MDS plots using all three replicates from the eight development datasets (Figure 1) and the four adult tissue datasets (Figure 2) are shown below.

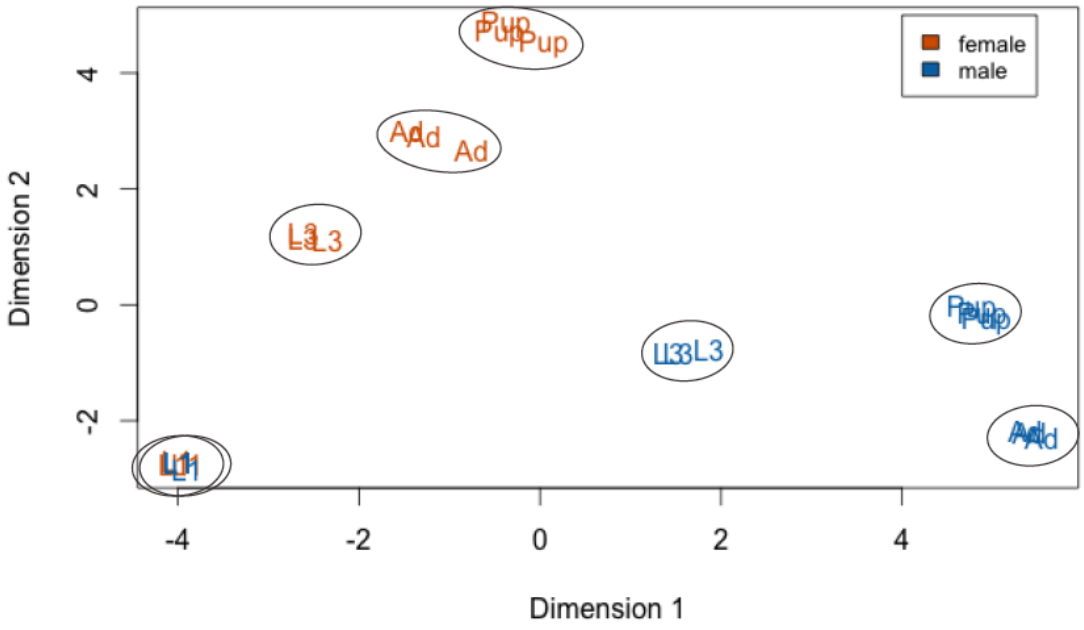


Figure 1 – MDS plot for development RNA-Seq datasets – The MDS plot shows distinct clustering of the male (blue) and female (orange) samples of the four development RNA-Seq datasets. The four stages are represented as: "L1" – 1st instar larvae, "L3" – wandering 3rd instar larvae, "Pup" – mid-pupae, and "Ad" – seven-day adults.

Replicates cluster distinctly for each sample with the exception of the male and female 1st instar larvae samples, which are mixed together. Because the 1st instar precedes gonad development and, presumably, the majority of sex-biased gene expression, we are not concerned with the overlap between these two samples (Bate and Martinez Arias, 1993; Parisi et al., 2004). Very few genes are expressed in a sex-specific manner this early in development, but one for which we have clear orthology and good expression data, *roX2*, is expressed solely in the male 1st instar samples (Franke and Baker, 1999). We do not see strong evidence of a batch effect between the higher-depth A replicates and the lower-depth B and C replicates, which were generated a full year apart.

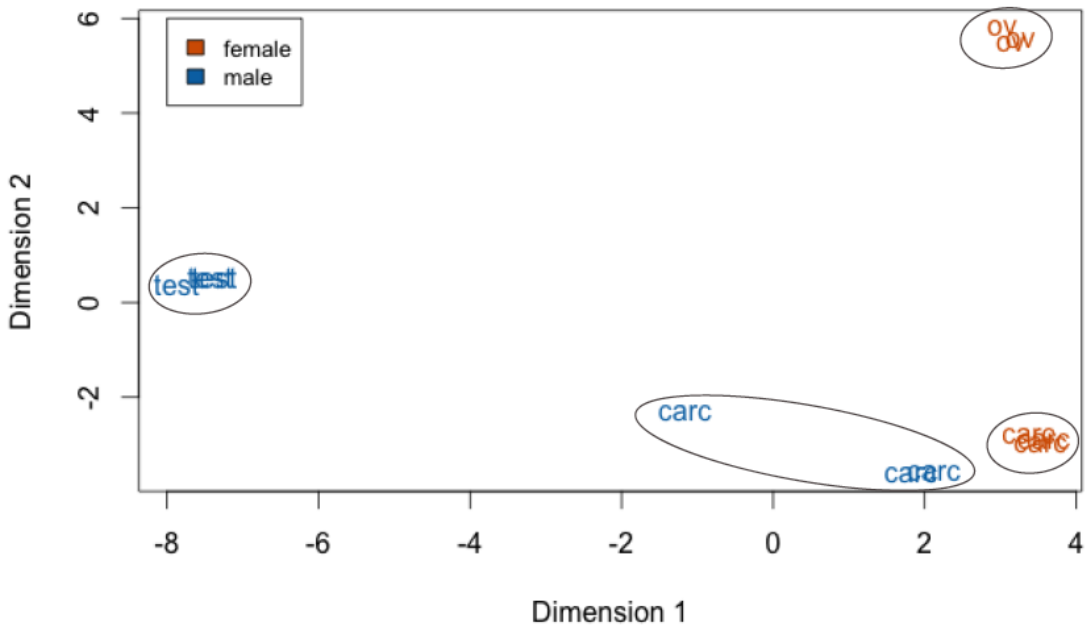


Figure 2 – MDS plot for adult tissue RNA-Seq datasets – The MDS plot shows distinct clusters of male (blue) and female (orange) samples of the adult gonad and carcass RNA-Seq datasets. The tissues are represented as: "test" – testes, "ov" – ovaries, and "carc" – carcasses.

The male carcass sample does have a single replicate, the high-depth A replicate (carcM_A), that does not cluster tightly with the other two replicates from that sample. Because of the position this replicate takes on the MDS plot, we investigated whether this

separation was due to contamination from testes. Sure enough, we found dozens of expressed loci in carcM_A that were silent in carcM_B and carcM_C and expressed highly in all three testes replicates. To correct for this, we re-collected RNA from male testes and carcass samples and re-sequenced them at high-depth. Sequencing and mapping statistics for this replicate D are shown in Table 2, and the MDS plot with testes and male carcass replicate D replacing replicate A is shown in Figure 3.

Sample	Raw mate pairs	HQ mate pairs	Mapped fragments
carcM_D	107,162,620	97,304,188 (90.8%)	86,651,475 (89.1%)
test_D	93,630,281	84,881,901 (90.7%)	81,683,540 (96.2%)

Table 2 – RNA-Seq sample statistics for testes and male carcass replicate D – Shown are sequencing, quality control, and mapping statistics for testes and male carcass replicate D. "Raw mate pairs" refers to the total number of fragments sequenced with Illumina paired-end sequencing. "HQ mate pairs" refers to the number of raw mate pairs with average PHRED score > 20. "Mapped fragments" refers to the number of high-quality mate pairs, either both mate pairs or only one mate pair, that map to the *D. pseudoobscura* genome.

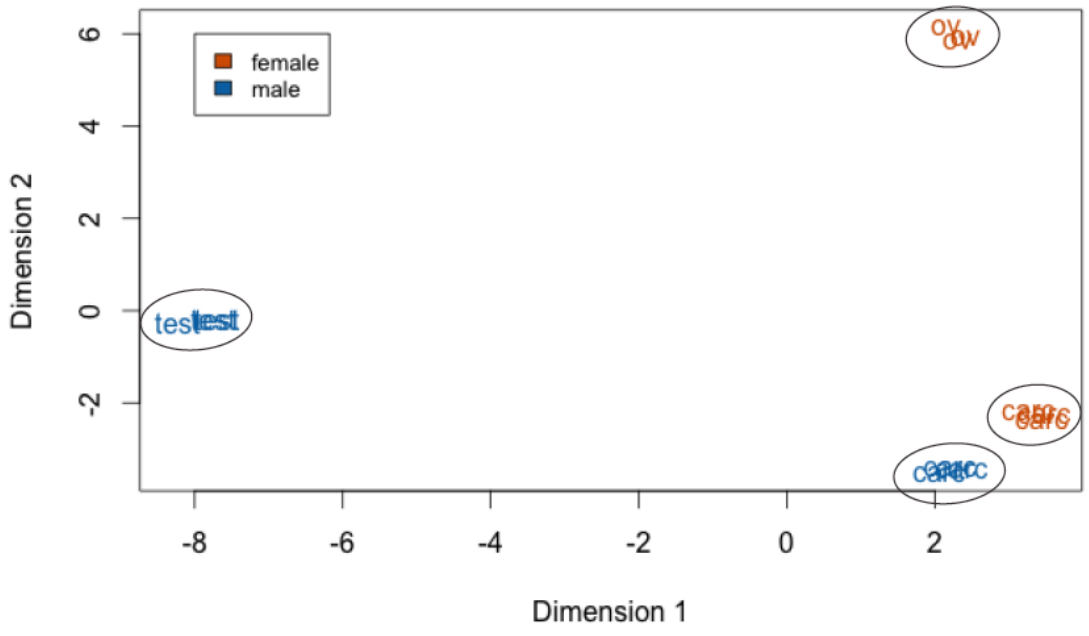


Figure 3 – MDS plot for adult tissue RNA-Seq datasets with male replicates D – This MDS plot shows tight and distinct clusters of the male (blue) and female (orange) samples of the adult gonad and carcass RNA-Seq datasets with male replicates D replacing replicates A. The tissues are represented as: "test" – testes, "ov" – ovaries, and "carc" – carcasses.

The mapping efficiency of the carcM_A reads to the *D. pseudoobscura* genome is high (87.0%), even with evidence of contamination from the testes. We believe that testes contamination from the same MV2-25 line of *D. pseudoobscura* was the only significant source of contamination. This probably occurred during dissections, as testes easily tear and rupture. Because of this, we opted not to redo transcriptome assembly from Chapter 1 using the D replicates, but the testes and male carcass D replicates are used in place of the A replicates for all expression analyses from this point forward.

Choosing an expression threshold

The digital nature of RNA-Seq makes it easy, in theory, to determine which genes are expressed or not expressed in any given sample. Expression values are represented as number of fragment counts per locus normalized to total library size (fragment counts per million fragments mapped, or cpm) or the number of fragment counts per locus normalized to both library size and locus length (fragment counts per kilobase of exon per million fragments mapped, or FPKM). In practice, low levels of fragment mismapping or stochastic transcription can blur this distinction, so a threshold value is often employed. Loci with expression values above the threshold are considered to be expressed, while loci with expression values below the threshold are not. These expression thresholds, however, are often completely arbitrary.

Choosing an appropriate expression threshold when analyzing lincRNA expression is particularly important because lincRNA loci, as a whole, are expressed at much lower levels than protein-coding loci (Derrien et al., 2012; Necsulea et al., 2014; Pauli et al., 2012; Young et al., 2012). This has been observed in a number of species, and we observe a similar trend in *D. pseudoobscura* (Figure 4, Mann-Whitney test, $p <$

2.2e-16 for all samples). Here, expression values are represented as $\log_2$(cpm), and all loci with at least a single mapped fragment in any replicate are included.

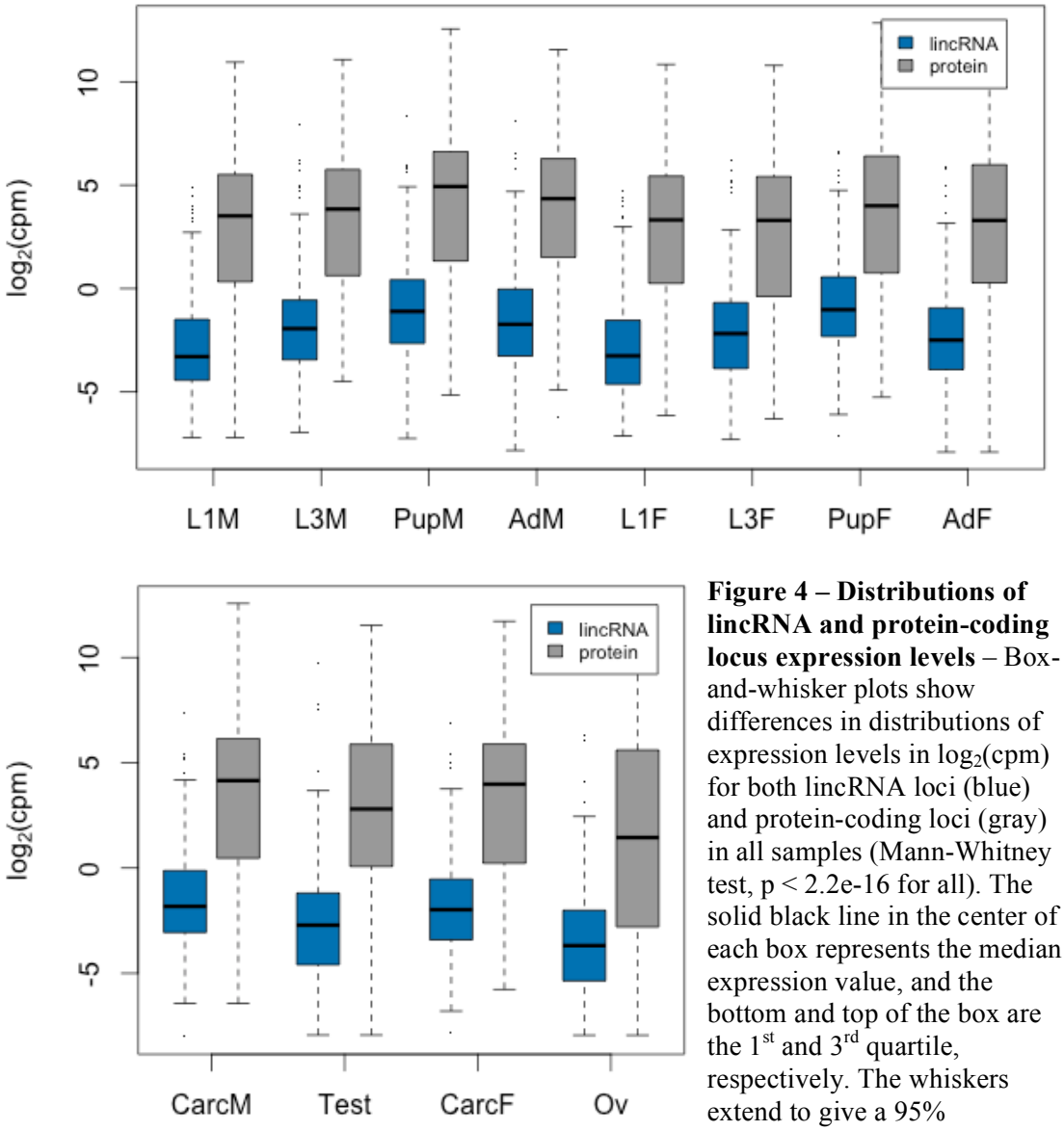


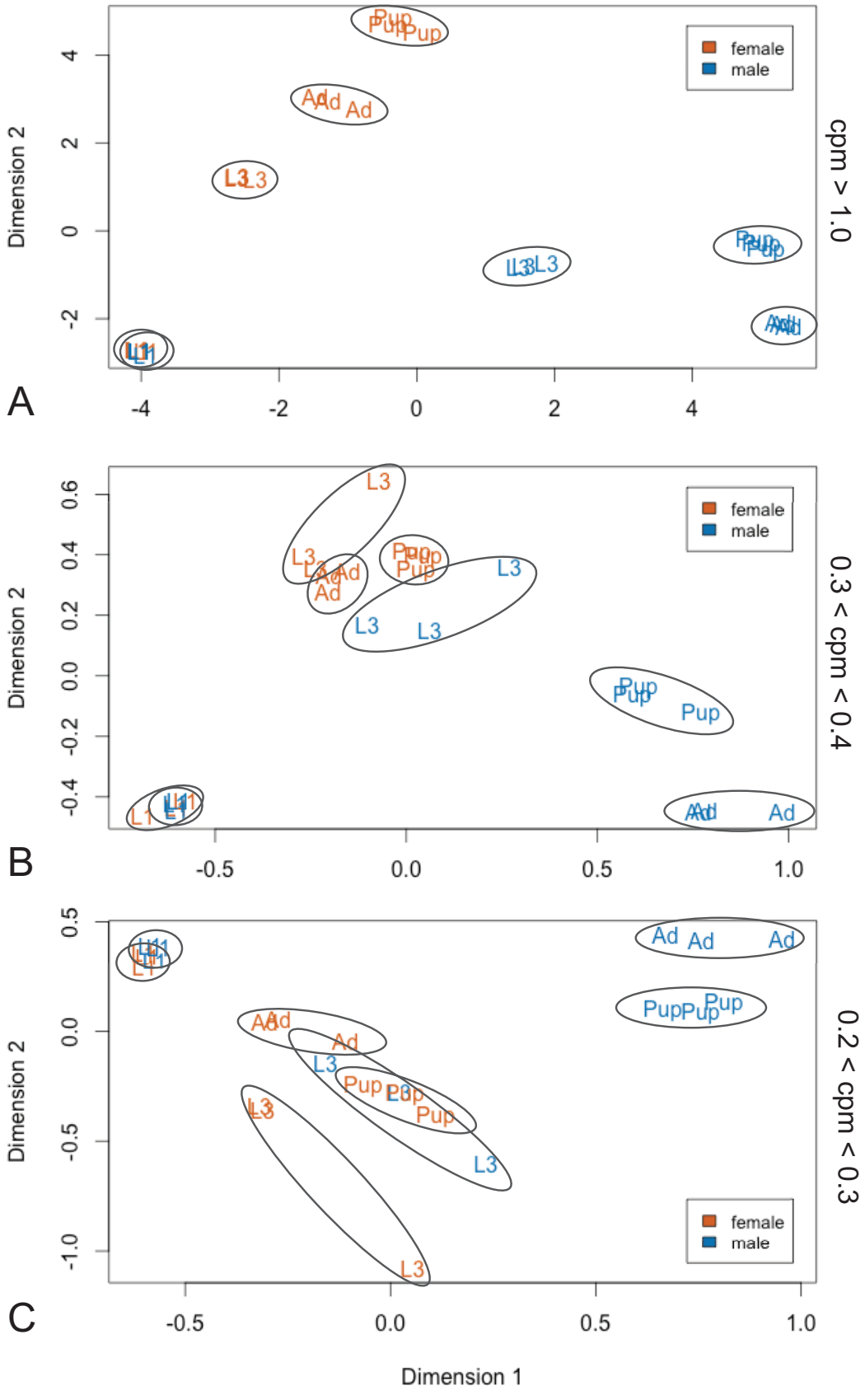
Figure 4 – Distributions of lincRNA and protein-coding locus expression levels – Box-and-whisker plots show differences in distributions of expression levels in $\log_2$(cpm) for both lincRNA loci (blue) and protein-coding loci (gray) in all samples (Mann-Whitney test, $p < 2.2e\text{-}16$ for all). The solid black line in the center of each box represents the median expression value, and the bottom and top of the box are the 1st and 3rd quartile, respectively. The whiskers extend to give a 95% confidence interval.

A threshold cpm of 1.0 (i.e. $\log_2$(cpm) = 0) is recommended for limma-voom, the package that we later use to look for differentially expressed genes, which is greater than the median lincRNA expression value in all 12 of our samples (Law et al., 2014). A

threshold value of 1.0 would only retain 449 of the 1,589 lincRNA loci (28.3%), while 6,828 of the 10,415 protein-coding loci (65.6%) would be included. We reason that this threshold is unnecessarily conservative and eliminates many genes that show legitimate expression signal that is distinguishable from transcriptional noise.

To find a more appropriate threshold value, we grouped all loci from the merged transcriptome into batches according to their maximum mean expression level in any single sample in the developmental series, with increments of 0.1 cpm from cpm = 0 to cpm = 1.0 and a final batch of all loci with cpm > 1.0. We generated MDS plots for all batches and looked for the batch with the lowest cpm for which we could still distinguish distinct developmental clusters. The minimum cpm value in that batch was then used as our expression threshold.

The cpm > 1.0 MDS plots largely mirrors the MDS plot that we used to quality check our RNA-Seq replicates and incorporates all loci (Figure 5A, Figure 1). All replicates cluster tightly and all sample clusters, with the exception of male and female L1, are distinct. MDS plots for cpm values between 1.0 and 0.4 also show seven distinct clusters, though replicates cluster less tightly and the distinct clusters are often closer as the cpm value drops (data not shown). With cpm between 0.3 and 0.4, distinct clusters are still apparent, though the fringes of these clusters begin to overlap and replicates are much less compact (Figure 5B). The female pupae and male L3 clusters completely overlap in the MDS plot with cpm between 0.2 and 0.3, and more clusters collapse as cpm drops to zero (Figures 5C, 5D, 5E).


A
B
C
Dimension 1
Dimension 2
female
male
cpm > 1.0
0.3 < cpm < 0.4
0.2 < cpm < 0.3
L1
L3
Pup
Ad

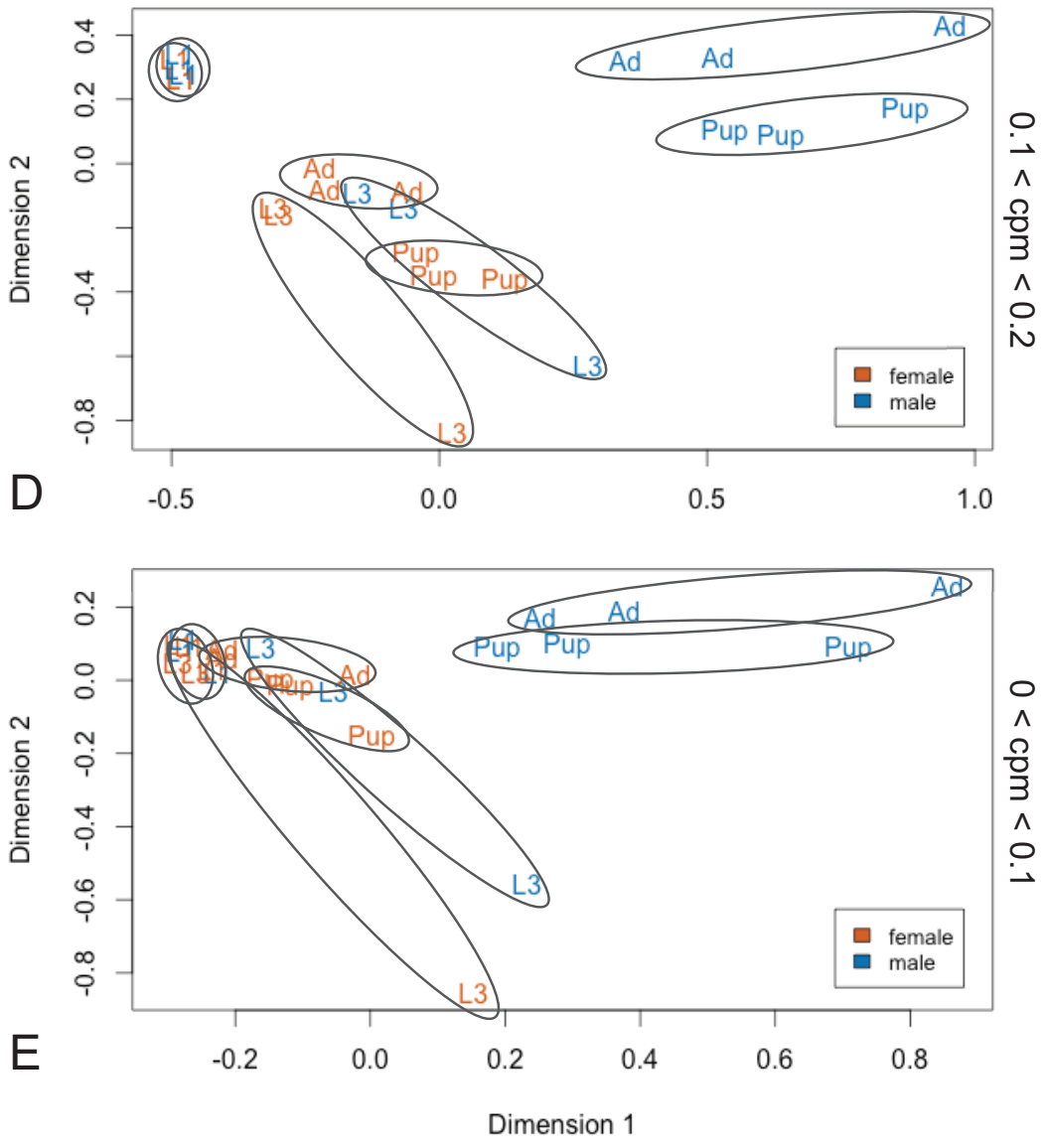


Figure 5 – Determining expression threshold using MDS plots of loci with decreasing expression values – Shown are MDS plots of all male (blue) and female (orange) RNA-Seq samples in the developmental series using only loci with a maximum mean expression value in at least one sample of (A) cpm > 1.0, (B) cpm between 0.3 and 0.4, (C) cpm between 0.2 and 0.3, (D) cpm between 0.1 and 0.2, and (E) cpm between 0 and 0.1. The four stages are represented as: "L1" – 1st instar larvae, "L3" – wandering 3rd instar larvae, "Pup" – mid-pupae, and "Ad" – seven-day adults.

Because the MDS plot with cpm between 0.3 and 0.4 contains the lowest expression values while still showing intact and mostly distinct clusters, we chose cpm = 0.3 as our expression threshold. In the lowest-depth replicate libraries, 0.3 cpm is roughly

equivalent to four mapped fragments at a locus, and a minimum mean of 12 mapped fragments across the three replicates. With this threshold, 925 of the 1,589 lincRNA loci (58.2%) will be retained, as will 7,649 of the 10,415 protein-coding loci (73.4%).

An overview of lincRNA expression throughout development and in adult gonadal tissues

To gain a broad understanding of the expression dynamics of lincRNAs throughout development and in adult gonadal tissues, we generated heatmaps using $\log_2$-transformed cpm values for the 8,574 lincRNA and protein-coding loci that fell above our 0.3 mean cpm expression threshold (Figures 6 and 7). We also determined how many lincRNA and protein-coding loci were expressed above the 0.3 mean cpm threshold in each sample. We then used these heatmaps as the starting point for more rigorous statistical analyses of lincRNA expression.

The number of expressed lincRNA increases in both sexes as development proceeds from the first-instar larval (115 in male, 110 in female) through third-instar larval stage (242 in male, 177 in female) and into the mid-pupal stage (452 in male, 279 in female), though increasingly higher numbers are seen as male development proceeds (Figure 6). The highest numbers of lincRNAs are expressed in the adult males (481) and at seemingly higher levels, but the number of expressed lincRNAs drops drastically in adult females (140). These overall trends are mirrored in the protein-coding loci, though the magnitude of these changes throughout development appears to be lower. Very few lincRNA loci appear to be expressed at all developmental stages, particularly when compared to the numbers of broadly expressed protein-coding loci. On the whole, lincRNAs are expressed in far fewer developmental samples than protein-coding loci

(lincRNA mean = 2.70, protein mean = 6.34, Mann-Whitney test, $p < 2.2e-16$). Two other trends stand out in the heatmap: (1) there is a group of lincRNAs that become more highly expressed as male development proceeds, and (2) there is a group of lincRNAs that are highly expressed in the 3rd instar larvae and pupae in both sexes.

LincRNA and protein-coding locus expression follows a similar pattern in the gonadal and carcass tissues, with parallels in the relative differences between tissues but seeming differences in magnitude (Figure 7). The highest number of expressed lincRNAs is seen in the testes (525), and the lowest seen in the ovaries (77). The carcass samples show intermediate levels of lincRNA expression, with 272 expressed in the male carcass and 172 expressed in the female carcass. Carcass samples for both lincRNAs and protein-coding loci show very similar expression profiles, suggesting that major differences in sex-specific gene expression in the adults are due to expression in the gonads. LincRNA representation is significantly overrepresented in the testes and underrepresented in the ovaries as compared to protein-coding gene representation (Fisher's exact test, $p < 0.05$). In general, lincRNAs are expressed in far fewer tissues than protein-coding loci (lincRNA mean = 1.39, protein mean = 3.04, Mann-Whitney test, $p < 2.2e-16$).

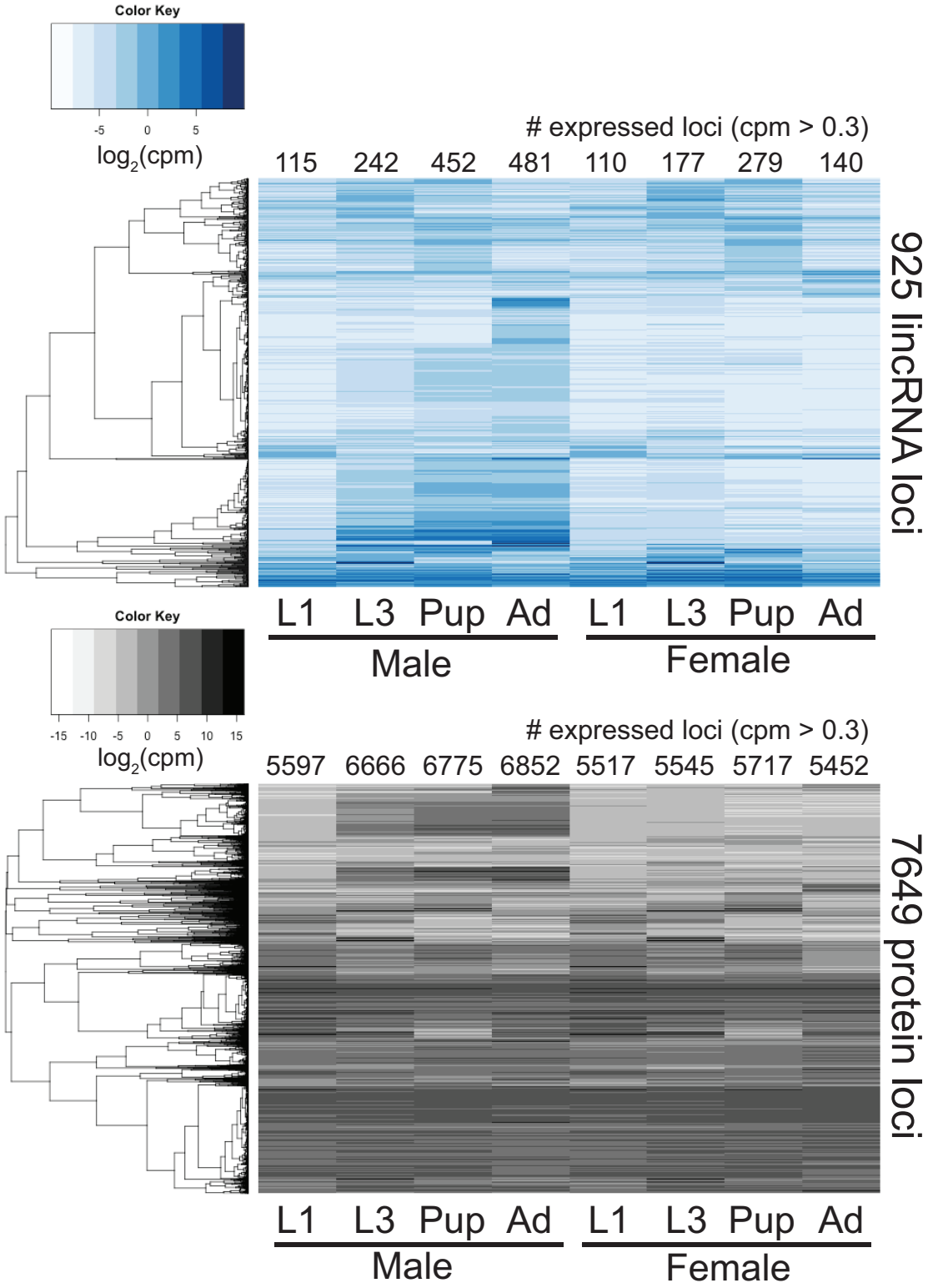


Figure 6 – Heatmaps of lincRNA and protein-coding locus expression through development – $\log_2$(cpm) values were used to generate heatmaps for the 958 lincRNA loci and 7,752 protein-coding loci included in our expression analyses. Each row represents an individual locus, and row clustering was done using Pearson's correlations. The numbers of expressed loci (mean cpm > 0.3) for each sample are located above the respective heatmap column.

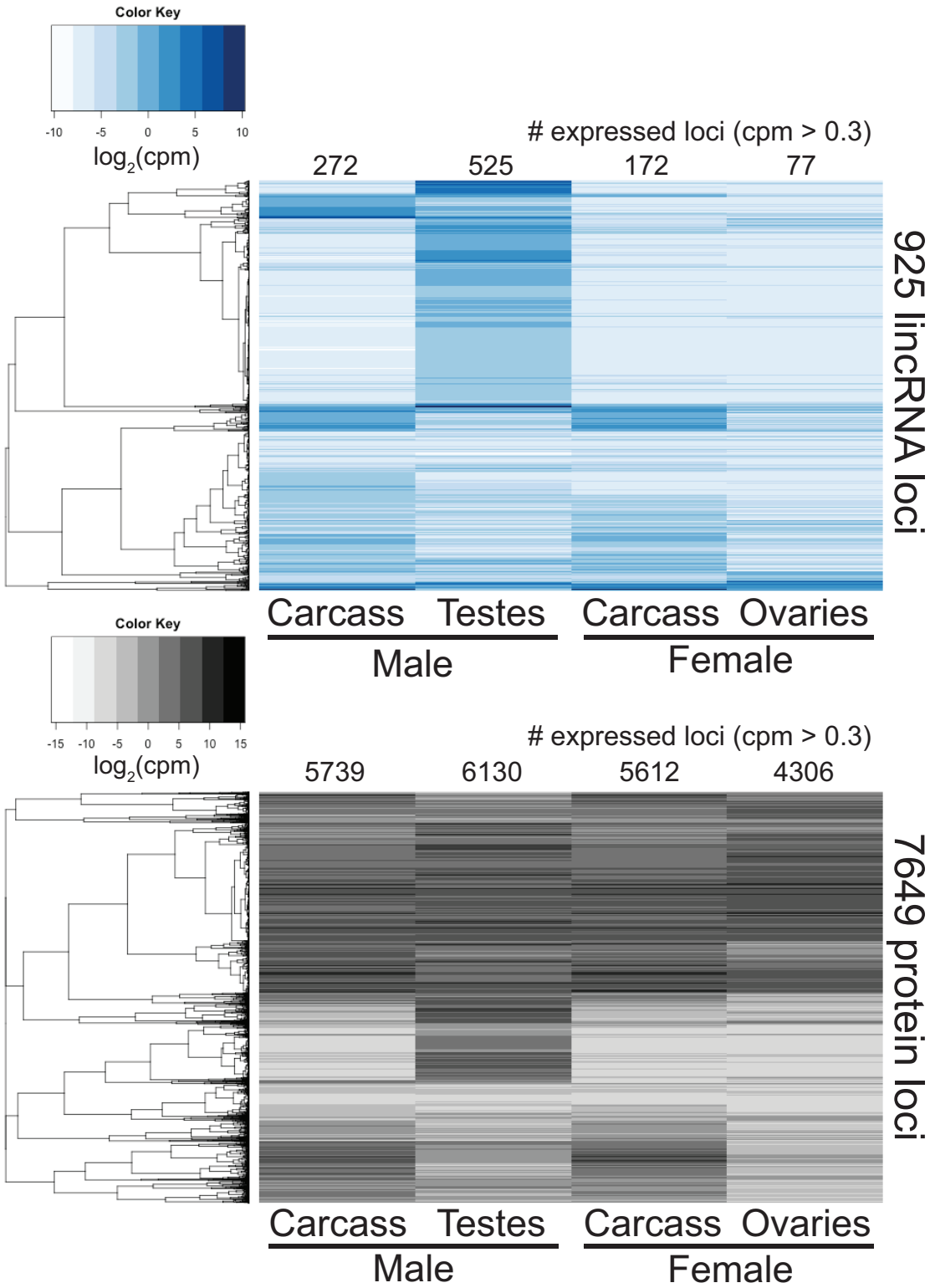


Figure 7 – Heatmaps of lincRNA and protein-coding locus expression in adult gonads and carcasses – $\log_2$(cpm) values were used to generate heatmaps for the 958 lincRNA loci and 7,752 protein-coding loci included in our expression analyses. Each row represents an individual locus, and row clustering was done using Pearson's correlations. The numbers of expressed loci (mean cpm > 0.3) for each sample are located above the respective heatmap column.

Developmental clustering of lincRNA expression

To facilitate a more rigorous statistical analysis of lincRNA expression through *D. pseudoobscura* development, we performed soft cluster analysis (i.e. fuzzy c-means) using the $\log_2$(cpm) expression values for all lincRNA and protein-coding loci at each sex-specific developmental stage (Kumar and Futschik, 2007). A soft clustering approach using the R package Mfuzz groups loci together by relative expression levels across all samples into an empirically determined number of clusters. A membership value between 0 and 1.0 for each locus gives an indication as to how closely that locus matches the cluster core. As opposed to hard (i.e. k-means) clustering, loci that do not have a great fit in any cluster will not be clustered, and loci can potentially be placed into more than one cluster.

To determine the most appropriate number of clusters, we performed repeated soft clustering with a varying number of clusters, from 4 to 40, and calculated the minimum centroid distance, or the minimum distance between cluster cores, for each iteration (Figure 8). As the number of clusters increases, the minimum centroid distance decreases and begins to plateau when the cluster number reaches 16. At this point, differences in cluster content will be minimal; thus, we chose 16 clusters for our developmental expression analyses.

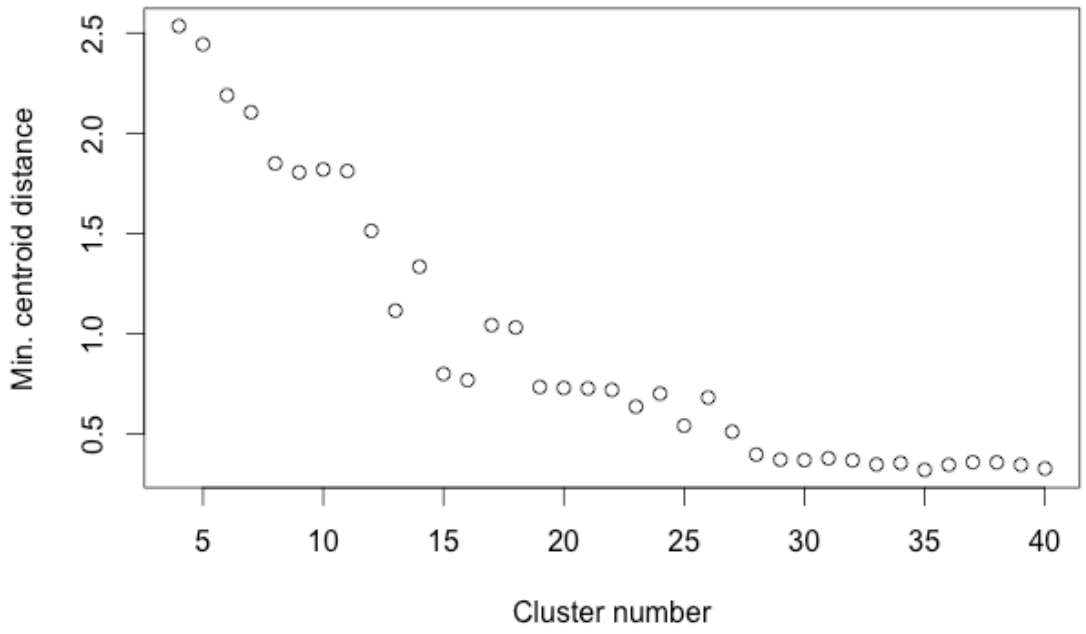


Figure 8 – Determining an appropriate cluster number using minimum centroid distance. Repeated soft clustering using varying numbers of clusters shows little separation between clusters when greater than 16 clusters are used.

We performed soft clustering of our 8,574 expressed loci across development using 16 clusters and requiring a moderately stringent cluster membership value of 0.5. With these parameters, 6,445 loci (75.2%) were placed into a cluster (Figure 9). This includes 729 of the 925 expressed lincRNA loci (78.8%) and 5,716 of the 7,649 expressed protein-coding loci (74.7%). No loci were placed in multiple clusters at the 0.5 membership level. The largest cluster, cluster 6, contained 1,170 loci (Table 3). All other clusters had between 168 and 577 loci.

Clusters are clearly defined by developmental stage and sex. Clusters 2, 6, and 14 show increasing gene expression in males as development proceeds, with the differentiation in the clusters based on when in development that increase begins, and constant, presumably absent, expression in females. Cluster 9 shows an increase in gene

expression in adult females and presumably absent expression in males. There are clusters that group loci with elevated expression at a single developmental stage in both sexes: 1st instar larvae (cluster 16), 3rd instar larvae (cluster 5), and pupae (cluster 4). There are clusters that show similar expression profiles through multiple developmental stages in both sexes. Cluster 3, for example, shows elevated expression in the male and female 1st-instar larvae and pupae with decreased levels in the 3rd-instar larvae and adults, and cluster 13 shows highest expression in the 1st-instar larvae with subsequent decreasing expression. Two clusters, cluster 1 and cluster 10, have sex-specific profiles that cannot be explained by complete inactivation in one of the sexes. In cluster 1, for example, gene expression is initially highest in both male and female 1st-instar larvae and decreases by the 3rd-instar stage; in males, expression stays low, while expression increases by the adult stage in females.

Each cluster contains at least 100 protein-coding loci, but the numbers of lincRNA loci within each cluster vary considerably (Table 3). Clusters 2, 6, and 14, which are the three clusters with male-specific expression increases, contain the highest numbers of lincRNAs, with 175, 104, and 175 lincRNA loci, respectively. Eight different clusters contain less than 20 lincRNA loci, with only a single lincRNA locus in cluster 1.

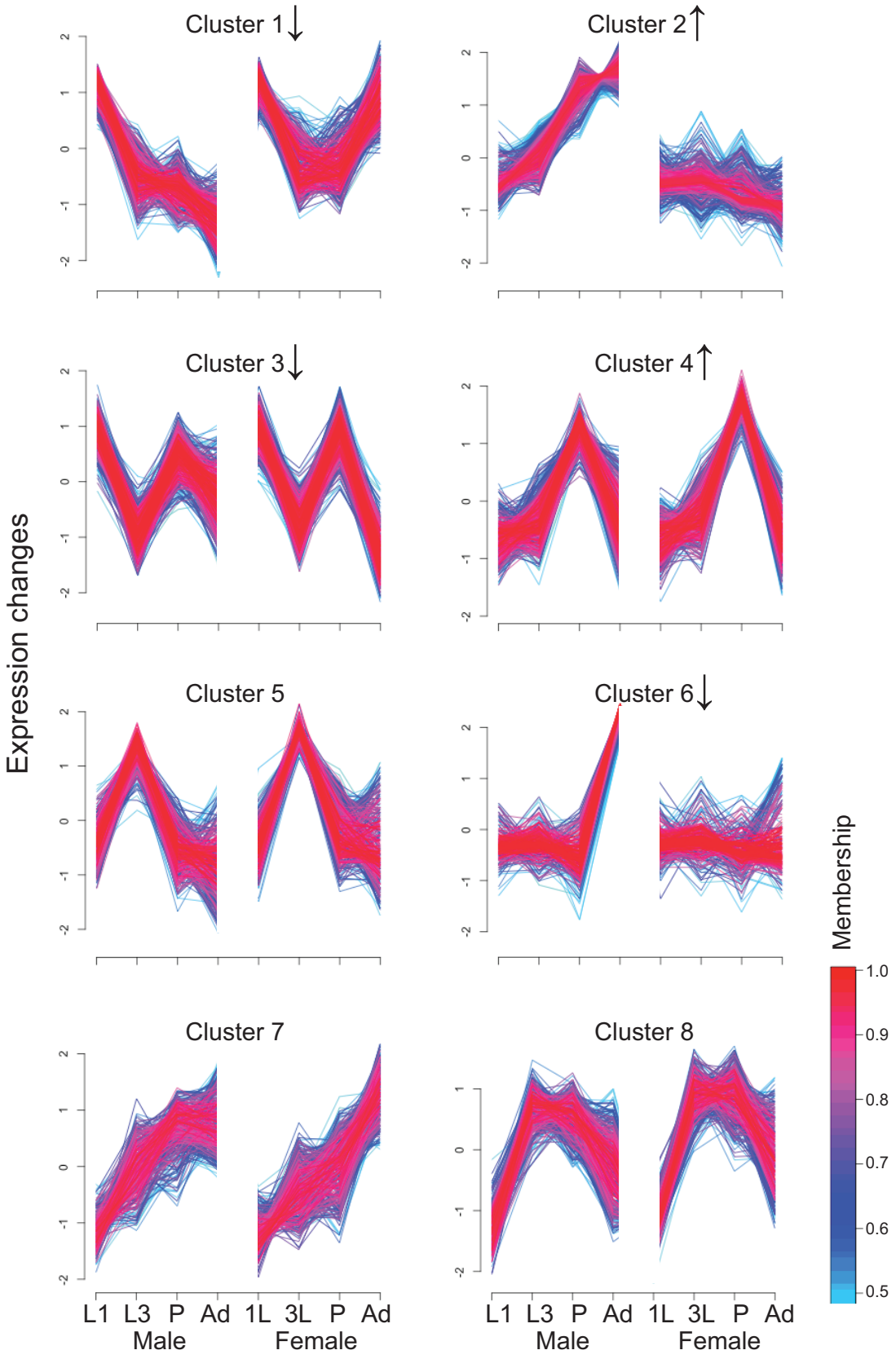
We next looked at whether lincRNAs are overrepresented in any clusters with respect to protein-coding loci. Table 3 details the numbers of lincRNA and protein-coding loci found in each cluster as well as whether lincRNAs are significantly over- or underrepresented within the cluster (Fisher's exact test with modified Bonferroni correction, $p < 0.035$). We find significant overrepresentation of lincRNA loci in two clusters that show male-specific expression increases (clusters 2 and 14) and a cluster that

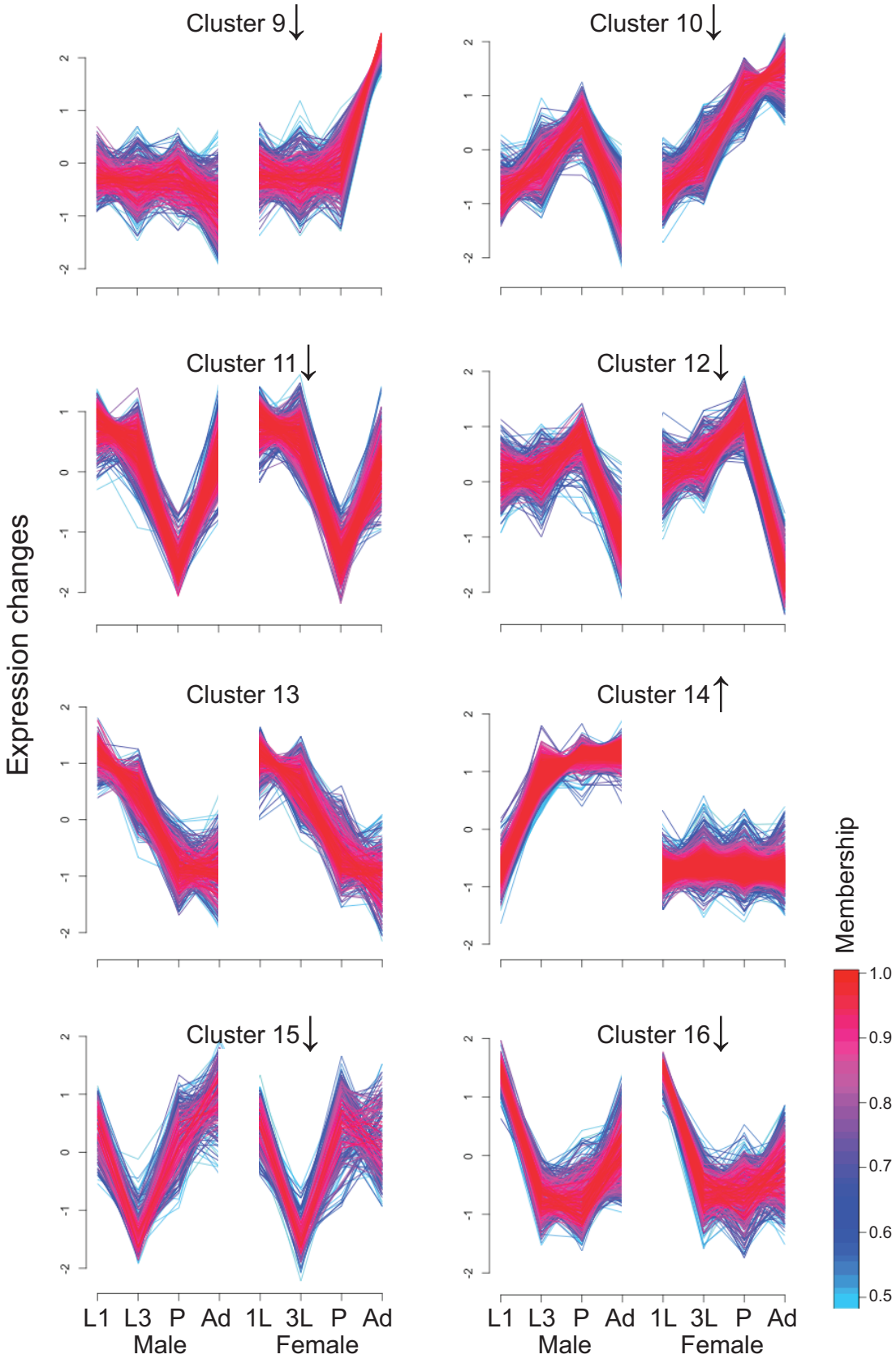
shows increased expression in the pupal stage in both sexes (cluster 4). LincRNAs have equal representation with protein coding loci in four clusters, all of which show similar expression profiles between males and females (clusters 5, 7, 8, and 13). LincRNAs are significantly underrepresented in all other clusters, particularly those that show female specific expression and 1st-instar larval expression in both sexes (Fisher's exact test with modified Bonferroni correction, $p < 0.036$). Despite containing 104 lincRNA loci, cluster 6 has a statistically significant underrepresentation of lincRNAs.

Cluster	# total loci	# lincRNA	# protein	p-value	lincRNA representation
1	316	1	315	*6.98E-15	down
2	577	175	402	*2.20E-16	up
3	393	26	367	*0.004612	down
4	382	73	309	*7.46E-07	up
5	366	38	328	0.8634	equal
6	1170	104	1066	*0.02554	down
7	215	14	201	0.04414	equal
8	475	46	429	0.4931	equal
9	217	12	205	*0.01027	down
10	392	6	386	*6.39E-13	down
11	322	6	316	*6.41E-10	down
12	295	18	277	*0.007175	down
13	168	20	148	0.6153	equal
14	499	175	324	*2.20E-16	up
15	301	5	296	*9.91E-10	down
16	357	10	347	*1.47E-08	down
All	6445	729	5716		

Table 3 – lincRNA and protein-coding loci content of developmental clusters. Columns show total gene number per cluster as well as the number of lincRNA loci and protein-coding loci per cluster. LincRNA representation was determined using a two-tailed Fisher's exact test with a modified Bonferroni correction of $p < 0.035$, and significant differences are marked with a *.

Figure 9 – Soft clustering of expression profiles throughout development. Soft clustering of developmental expression values for the combined set of lincRNA and protein-coding loci produces 16 major clusters that account for 74.1% of expressed loci. The y-axis of each chart represents relative expression changes, with the mean expression value for each locus centered on zero. The color of an individual locus' expression profile indicates its membership value in that cluster. Up and down arrows next to the cluster name indicate whether lincRNAs are significantly overrepresented or underrepresented, respectively, in the cluster (Fisher's exact test with modified Bonferroni correction $p < 0.035$).


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Expression changes
L1 L3 P Ad 1L 3L P Ad
Male
Female
Membership
1.0
0.9
0.8
0.7
0.6
0.5


Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Cluster 15
Cluster 16
Expression changes
L1 L3 P Ad 1L 3L P Ad
Male
Female
Membership
1.0
0.9
0.8
0.7
0.6
0.5

Gene Ontology (GO) analysis of developmental expression clusters with an overrepresentation of lincRNAs

To follow up our soft clustering analyses, we identified Biological Process Gene Ontology (GO) terms that are significantly overrepresented in each of the 16 clusters using GeneCodis3 (Ashburner et al., 2000; Tabas-Madrid et al., 2012). We focus on the three clusters with an overrepresentation of lincRNA loci, and a list of significant GO terms for all clusters are listed in Appendix Tables 2-17.

Two of the clusters with lincRNA overrepresentation had predominantly male-specific expression increases. The top GO terms for Cluster 2, which contains genes that increase in expression between every developmental stage in males, are spermatogenesis and sensory perception of smell (Appendix Table 3). Cluster 14 has genes with elevated and fairly constant expression from the 3rd-instar larval stage through the adult stage, and its top GO hits are microtubule-based movement, translational initiation, tricarboxylic acid cycle, sperm motility, 'de novo' protein folding, and proteolysis involved in cellular protein catabolic process (Appendix Table 15). The third cluster with an overrepresentation of lincRNAs, cluster 4, has elevated expression in pupae of both sexes. The top GO terms for cluster 4 are various types of cell adhesion, regulation of transcription, steroid hormone mediated signaling pathway, compound eye morphogenesis, and several others (Appendix Table 5).

Four clusters did not show any significant differences in lincRNA or protein-coding representation. Clusters 5, 7, 8, and 13 contain expression profiles that are roughly equivalent in males and females. Cluster 5, which has peak expression in the 3rd-instar larvae, has top GO terms of chitin-based cuticle development and body morphogenesis

(Appendix Table 6). Cluster 7 contains genes that progressively increase in expression in both sexes and has top GO terms of mitotic spindle organization, mitosis, and microtubule-based movement (Appendix Table 7). Cluster 8, which contains genes with high expression in the 3rd-instar larval and pupal stages, has a top GO term of defense response (Appendix Table 8). The top GO term for cluster 13, which contains genes that progressively decrease in expression throughout development in both sexes, is chitin metabolic process (Appendix Table 9).

LincRNAs are significantly underrepresented in the other 9 clusters. The underrepresentation for cluster 6, which contains 104 lincRNA loci, is only slightly significant. Cluster 6 genes have elevated expression in adult males and have top GO terms of multicellular organism reproduction and sperm competition (Appendix Table 7). The single top hits for the other clusters are: cluster 1 – neurogenesis (Appendix Table 2); cluster 3 – ion transport (Appendix Table 4); cluster 9 – DNA-dependent DNA replication (Appendix Table 10); cluster 10 – dendrite morphogenesis (Appendix Table 11); cluster 11 – transmembrane transport (Appendix Table 12); cluster 12 – regulation of transcription DNA-dependent (Appendix Table 13); cluster 15 – neurotransmitter transport (Appendix Table 16); and cluster 16 – chitin-based cuticle development and body morphogenesis (Appendix Table 17).

Sex-bias of lincRNAs

Developmental clustering strongly suggests that a large number of loci, both lincRNA and protein-coding, show sex-biased expression. We performed differential expression analyses using limma-voom to detect significant sex-bias during development and in adult tissues (Law et al., 2014; Smyth, 2005). We detect significant differences (adj. p-

value < 0.01) between the male and female developmental expression profiles in 426 lincRNAs (57.6% of the 739 total lincRNA with cpm > 0.3 in the developmental series) and 5,728 protein-coding loci (75.7% of the total 7,570 protein-coding loci).

To tease apart whether lincRNAs and protein-coding loci show the same patterns of sex-bias, we performed differential expression analysis individually on each developmental stage and also in the combined carcass and gonad tissues. Results of these analyses are presented in Table 4 and Figures 10 and 11.

Sample	Total loci	MB	FB	UB	p-value
L1-lincRNA	129	1	1	127	0.1753
L1-protein	5650	21	11	5618	
L3-lincRNA	270	76	7	187	*0.0001153
L3-protein	6711	1469	568	4674	
Pup-lincRNA	517	225	52	240	*< 2.2e-16
Pup-protein	6880	1990	2160	2730	
Ad-lincRNA	515	375	36	104	*< 2.2e-16
Ad-protein	6983	3353	2432	1198	
Carc-lincRNA	298	97	7	194	*< 2.2e-16
Carc-protein	5877	413	419	5045	
Gonads-lincRNA	561	495	32	34	*< 2.2e-16
Gonads-protein	6340	3312	2038	990	

Table 4 – Sex expression bias in lincRNAs and protein-coding genes. Sex bias was determined using limma-voom with adjusted p-value < 0.01. "Total loci" refers to the number of loci with expression > 0.3 cpm in the union of the male and female samples. "MB", "FB", and "UB" refer to significant male-biased genes, female-biased genes, and unbiased genes, respectively. "p-value" compares bias distributions between lincRNAs and protein-coding genes using Fisher's exact test ($p < 0.05$). Significant deviations between two classes are indicated with *.

There are very few male and female-biased genes in the1st-instar larvae, for both lincRNAs and protein-coding genes (Figure 10). There are progressive increases in the levels of protein-coding male and female-biased expression thereafter. Numbers of male-biased protein-coding genes exceed numbers of female-biased genes in the 3rd-instar larvae, but by the pupal and adult stages, the levels are roughly equal. Male-biased lincRNAs similarly increase in frequency as development proceeds, but the numbers of

female-biased lincRNAs stay low, with the highest numbers of female-biased lincRNAs in the pupal stage at 10.1%. The total proportions of sex-biased genes in lincRNAs and protein-coding genes remain roughly equal in every developmental stage, with over 80% of all loci showing some level of sex-bias in adult flies. That said, the male versus female proportions at each stage after the 1st-instar larvae are significantly different (Fisher's exact test, $p < 0.05$); sex-biased lincRNAs are overwhelmingly male-biased.

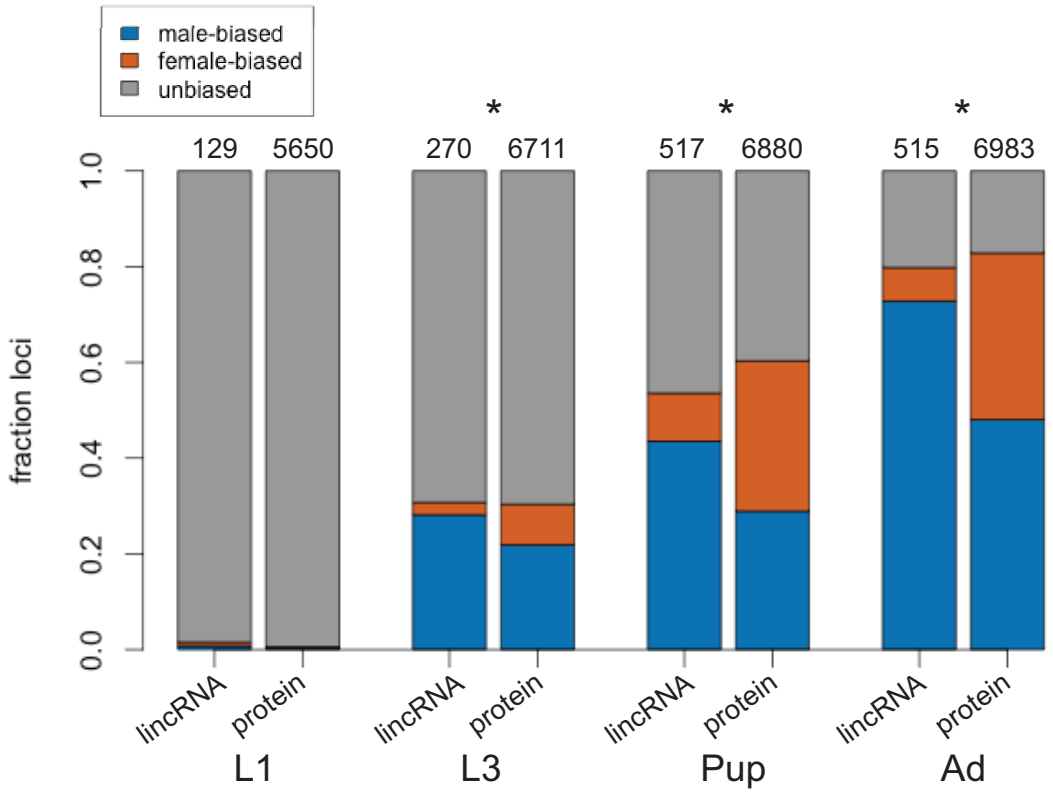


Figure 10 – Sex-bias in development. Shown are total fractions of significantly male-biased, female-biased, and unbiased genes at four developmental stages of *D. pseudoobcura*. Numbers above columns indicate the total number of expressed loci at that stage (cpm > 0.3). * indicates a significant difference between proportions of sex-biased genes via Fisher's exact test ($p < 0.05$).

An analysis of sex-bias in the dissected gonads and carcasses suggests a basis for the developmental sex-bias patterns (Table 4, Figure 11). The numbers of male and female-biased protein-coding genes are quite low in the carcass samples. Female-biased lincRNA expression is likewise low in the carcass, but 32.6% of lincRNAs show male-

biased expression in the carcasses. Patterns of sex-bias in the gonads mirror the patterns seen in the adults. Gonad sex-bias is high for both lincRNAs and proteins, but whereas the proportions of male-biased and female-biased protein-coding genes are not drastically different (52.2% male-biased and 32.1% female-biased), the proportions seen in lincRNA loci are (88.2% male-biased and 5.7% female-biased). LincRNA and protein-coding proportions of sex-biased genes are significantly different (Fisher's exact test, $p < 0.05$).

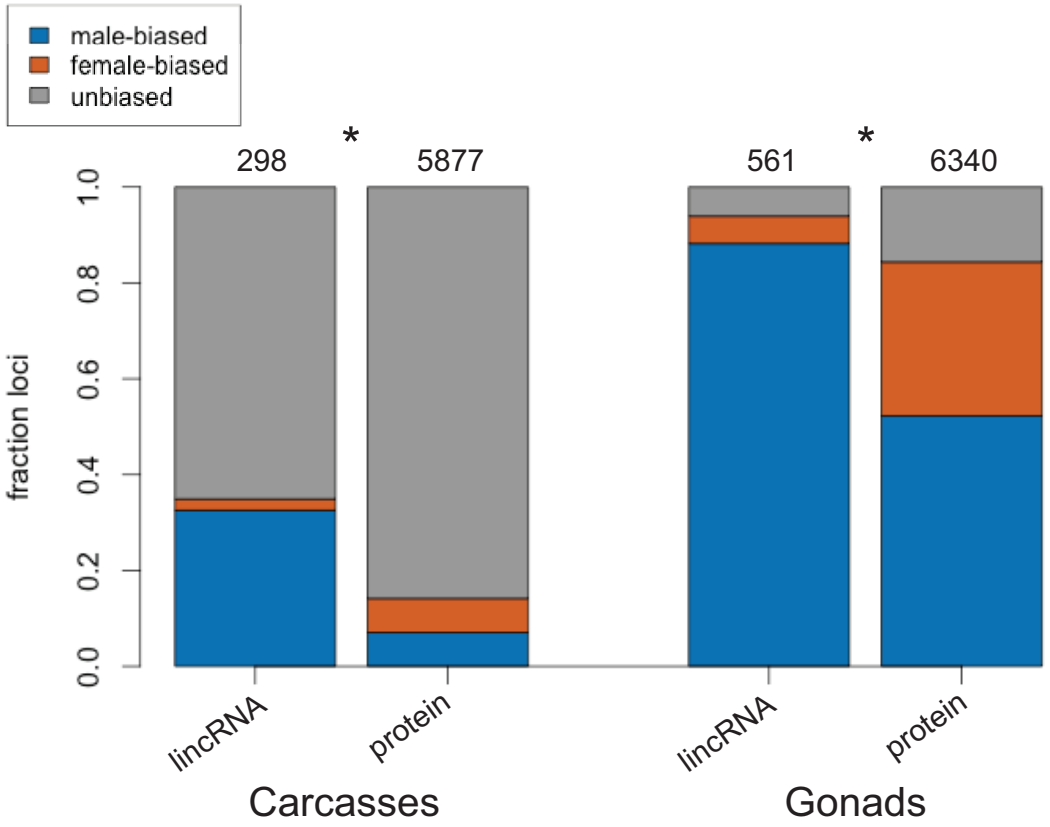


Figure 11 – Sex-bias in adult gonads and carcasses. Shown are total fractions of significantly male-biased, female-biased, and unbiased genes in adult gonads and carcasses of *D. pseudoobcura*. Numbers above columns indicate the total number of expressed loci at that stage (cpm > 0.3). * indicates a significant difference between proportions of sex-biased genes via Fisher's exact test ($p < 0.05$).

When we combine data from all six different sample types, both developmental stages and adult tissues, we find 583 lincRNA loci that show male-biased but not female-biased expression in at least one sample (i.e. a gene that shows male-bias in adults and is

unbiased in all other samples is classified as "male-biased"); we refer to these henceforth as "male-biased lincRNAs" (Table 5). 75 lincRNA loci show female-biased expression in at least one sample but no male-biased expression in any sample; we refer to these henceforth as "female-biased lincRNAs". 248 lincRNAs have no sex-biased expression in any sample; these will be referred to as "unbiased lincRNAs". 19 lincRNA loci show evidence of both male and female expression bias in different samples; we refer to these henceforth as "dynamic-bias lincRNAs". Using the same criteria for protein-coding genes across all samples, we identify 2,927 male-biased genes, 2,563 female-biased genes, 830 unbiased genes, and 1,329 dynamic-bias genes (Figure 12). The cumulative proportions of sex-biased loci in lincRNAs and protein-coding genes are significantly different (Fisher's exact test, $p < 0.05$).

class	total loci	MB	FB	DB	UB	p-value
lincRNA	925	583	75	19	248	*<2.2e-16
protein	7649	2927	2563	1329	830	

Table 5 – Overall levels of sex-bias across all samples. Overall sex bias was determined by parsing results from individual samples. "Total loci" refers to the number of loci with expression > 0.3 cpm in the union of the male and female samples. "MB", "FB", "DB", and "UB" refer to male-biased genes, female-biased genes, dynamic-bias genes, and unbiased genes, respectively. "p-value" compares bias distributions between lincRNAs and protein-coding genes using Fisher's exact test ($p < 0.05$). Significant deviations between two classes are indicated with *.

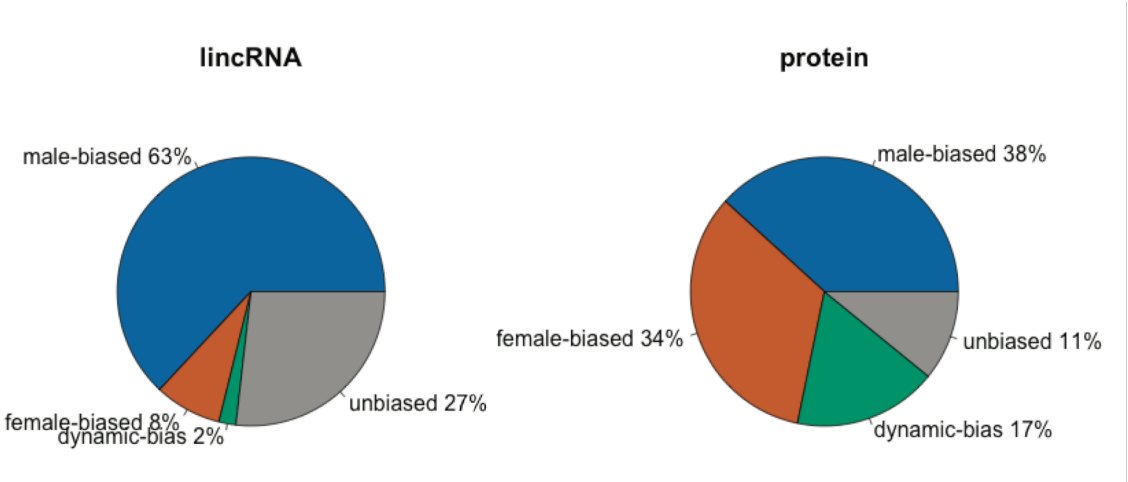


Figure 12 – Overall proportions of lincRNA and protein-coding sex-bias across all samples. Pie charts show the overall proportions of sex-biased genes across all four developmental series samples and two adult tissue samples.

The majority of the dynamic-bias genes, in both lincRNAs (78.9%) and protein-coding genes (77.1%), exhibit switches in sex-bias as development proceeds, with a switch from female-to-male bias more common than a male-to-female switch for both classes of genes. Only two protein-coding genes demonstrated a switch from female-to-male bias and then a reversion back to female-bias. The remainder of the dynamic-bias genes result from either: (1) inversions of sex-bias between the gonads and carcasses or (2) conflict between signal from the developmental series and signal from the adult tissues. Only four lincRNAs show either of these expression patterns.

LincRNA representation in the gonads

The patterns of sex-bias suggest that the gonad transcriptomes might show differences between lincRNA and protein-coding gene content. We performed differential expression analysis with limma-voom to detect significant gonad or carcasses expression biases in the dissected adult tissues (Law et al., 2014; Smyth, 2005). Results are presented in Table 6.

Sample	total loci	GB	CB	UB	p-value
Male tissues - lincRNA	711	449	203	59	*< 2.2e-16
Male tissues - protein	7041	3204	2756	1081	
Female tissues - lincRNA	202	36	126	40	*1.589e-11
Female tissues - protein	5780	2358	2746	676	

Table 6 – Tissue expression bias in lincRNAs and protein-coding genes. Tissue bias was determined using limma-voom with adjusted p-value < 0.01. "Total loci" refers to the number of loci with expression > 0.3 cpm in the union of the gonad and carcass samples. "GB", "CB", and "UB" refer to significant gonad-biased genes, carcass-biased genes, and unbiased genes, respectively. "p-value" compares bias distributions between lincRNAs and protein-coding genes using Fisher's exact test ($p < 0.05$). Significant deviations between two classes are indicated with *.

The vast majority of both lincRNAs and protein-coding genes have biased tissue expression in males and females. That said, lincRNAs show wildly different levels of gonad-bias and carcass-bias in males and females (Figure 13). 63.1% of lincRNAs in males exhibit testes expression bias, while only 17.8% of lincRNAs in females exhibit ovaries expression bias. Likewise, carcass-biased lincRNAs are less prevalent in males than females. Both lincRNA tissue-bias patterns are significantly different than the more equivalent patterns of tissue-bias seen with protein-coding genes (Fisher's exact test, $p < 0.05$).

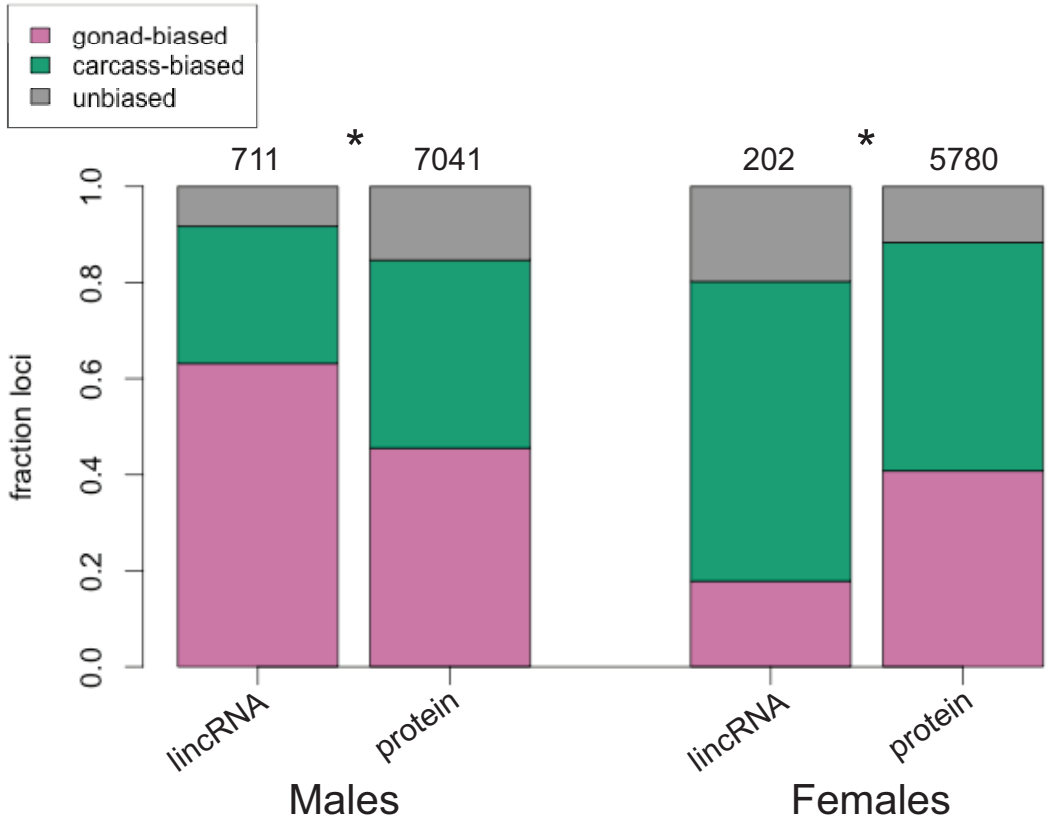


Figure 13 – Tissue-bias in adult gonads and carcasses. Shown are total fractions of significantly gonad-biased, carcass-biased, and unbiased genes in adult males and female *D. pseudoobcura*. Numbers above columns indicate the total number of expressed loci at that stage (cpm > 0.3). * indicates a significant difference between proportions of sex-biased genes via Fisher's exact test ($p < 0.05$).

Because genes that are expressed exclusively in the gonads often have unique properties, we identified testis-specific and ovary-specific lincRNAs and protein-coding genes. Of the 583 lincRNA loci with an overall male-biased expression, 424 (45.8% of all expressed lincRNA loci) are expressed exclusively in the testes (cpm > 0.3). Only 16 of the 75 female-biased lincRNAs (1.7% of all expressed lincRNAs) are expressed exclusively in the ovaries (cpm > 0.3). In the set of protein-coding genes, we found 1,145 testes-specific genes (15.0%) and 59 ovary-specific genes (0.8%).

Demasculinization and feminization of X-linked lincRNAs

Numerous studies in *Drosophila* have shown a dearth of male-biased protein-coding genes on the X chromosome, an observation that has been explained by several selection-based evolutionary models (Bachtrog et al., 2010; Gao et al., 2014; Meiklejohn and Presgraves, 2012; Meisel et al., 2012; Sturgill et al., 2007; Vibranovski et al., 2009). A recent study has shown similar X chromosome demasculinization for lincRNAs in *D. melanogaster* (Gao et al., 2014). Here, we explore the effects of sex-biased expression on *D. pseudoobscura* lincRNA chromosomal location.

To determine whether sex-biased lincRNAs are depleted or enriched on the X chromosome as compared to the autosomes, we calculated the odds ratio (OR) between the sex-biased gene distributions (autosomes/X) and the unbiased gene distributions (autosomes/X). An odds ratio above 1.0 indicates that the X-chromosome is depleted for that class of genes, and an odds ratio below 1.0 indicates that the X-chromosome is enriched for that class of genes. We used Fisher's exact test ($p < 0.05$) to determine whether the differences in chromosomal gene distributions are statistically significant.

First, we calculated the ORs for the sets of male-biased and female-biased lincRNAs and protein-coding genes (Table 7, Figure 14). We find that male-biased lincRNAs and male-biased protein-coding genes both have an OR of 1.41, and both are significantly underrepresented on the X chromosome (Fisher's exact test, $p < 0.05$). Female-biased lincRNAs display the opposite trend with a significant OR of 0.46, indicating that they are enriched on the X. In contrast, there is no evidence of female-biased protein-coding gene enrichment on the X (Fisher's exact test, $p = 0.2415$).

Class	Bias A	Bias X	Unbias A	Unbias X	Odds Ratio	p-value
MB-lincRNA	411	172	156	92	1.41	*0.03436
MB-protein	1966	961	492	338	1.41	*2.934e-05
FB-lincRNA	33	42	156	92	0.46	*1.056e-05
FB-protein	1459	1104	492	338	0.91	0.2415
MB-noTS-lincRNA	116	43	156	92	1.59	*0.04042
MB-noTS-protein	1268	523	492	338	1.66	*7.569e-09
TS-lincRNA	295	129	156	92	1.34	0.0887
TS-protein	698	438	492	338	1.09	0.3502

Table 7– Sex-bias effect on chromosomal locations. X chromosome depletion or enrichment was determined using odds ratios and Fisher's exact test ($p < 0.05$). "MB", "FB", and "TS" refer to significant male-biased genes, female-biased genes, and testis-specific genes, respectively. "A" refers to the autosomes: chromosomes 2, 3, and 4. "X" refers to both the XL and XR arms. Significant deviations between two classes are indicated with *.

Subsequent work suggests that testis-specific genes, despite being male-biased, do not show underrepresentation on the X chromosome but rather show random chromosomal distributions (Meiklejohn and Presgraves, 2012; Meisel et al., 2012). We divided our male-biased genes into two sets: testis-specific genes (424 lincRNAs, 1,136 protein-coding loci) and male-biased but not testis-specific genes (159 lincRNAs, 1,791 protein-coding loci). The non-testis-specific male-biased genes still show evidence of X chromosome demasculinization, with the lincRNAs having an OR of 1.59 and the protein-coding genes having an OR of 1.66. Both are significant (Figure 14, Fisher's exact test, $p < 0.05$). The testis-specific genes both have ORs above 1.0, but neither is statistically significant (Fisher's exact test, lincRNA $p = 0.0887$, protein $p = 0.3502$), indicating that testis-specific genes, including the large set of testis-specific lincRNAs are not depleted on the X chromosome.

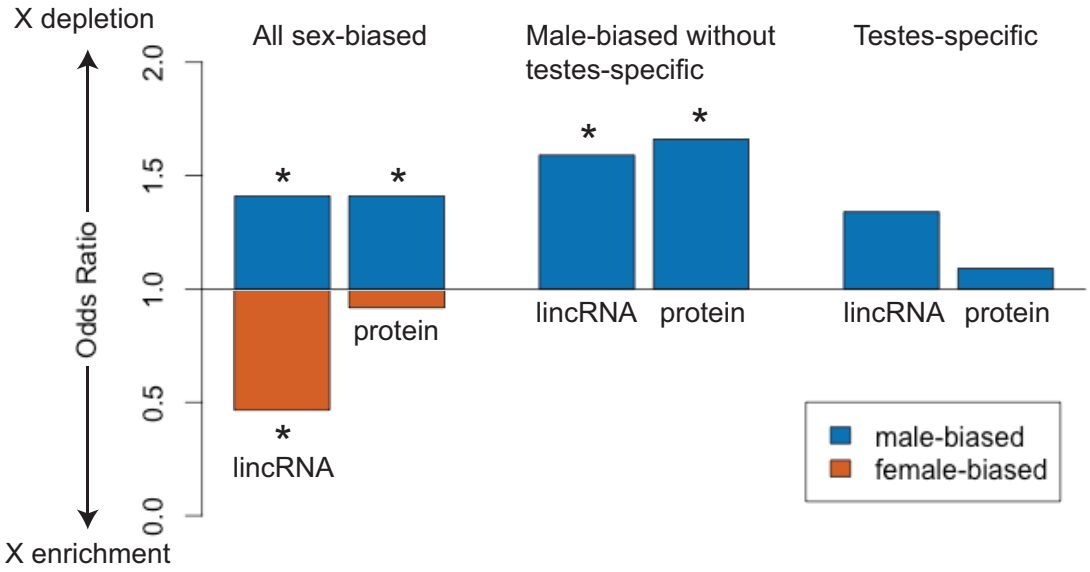


Figure 14 – Relationships between sex-bias and X-to-autosomes distributions. Odds ratios (ORs) between sex-biased and unbiased genes are shown for both lincRNAs and protein-coding genes. OR above 1.0 indicates relative depletion on the X chromosome. OR below 1.0 indicates relative enrichment on the X chromosome. * indicates differences are significant by Fisher's exact test, $p < 0.05$.

DISCUSSION

Our analyses of *D. pseudoobscura* lincRNA expression dynamics reveal key similarities and differences both with *D. pseudoobscura* protein-coding genes as well as the *D. melanogaster* lincRNAs previously described in the literature. Our analyses also give some limited insight into the biological relevance of lincRNAs in *D. pseudoobscura*.

Justification of RNA-Seq Expression Methodology

With the recent proliferation of RNA-Seq in modern biological research, there are many choices to be made with respect to high-throughput expression analyses, from how to count sequenced fragments to how to normalize data and determine significance in

differential expression. Here, we provide justification for the choices we made for our analyses of lincRNA expression in *D. pseudoobscura*.

First, we chose how to count sequence fragments for each locus using HTSeq-count v0.6.1p1 (Anders et al., 2014). Because of the low level of alternative transcription in lincRNAs and the comparatively fewer options for analyzing expression at the isoform level, we opted to count fragments at the locus level. Fragments that map to any and all transcripts at a locus will be counted for that locus. Fragments that map to overlapping loci on the same chromosome strand were labeled as ambiguous and not counted. Because our RNA-Seq data is unstranded, fragments that map to overlapping loci on opposite strands are likewise labeled as ambiguous and not counted. While this will eliminate some truly expressed genes from our analyses, we have a higher confidence in the expression counts for the loci that remain. We also note that this stringency can explain in part why the same RNA-Seq data can assemble 12,004 lincRNA and protein-coding loci but show detectable expression in only 8,574 of them (71.4%).

Fragment counts in libraries of varying sizes can be normalized to the total library size in different ways. The cpm metric normalizes fragment counts to just the library size (Law et al., 2014). The FPKM metric normalizes fragment counts both to library size and the length of the locus (Mortazavi et al., 2008). FPKM values typically are useful for comparing expression values between different loci with varying lengths. Because most of our analyses are comparing the same locus across samples (i.e. locus length is constant), we find the length normalization of FPKM unnecessary. Further, the differential-expression package we use, limma-voom, requires preservation of the order of counts across loci (Law et al., 2014). FPKM would distort this; cpm does not and is

thus appropriate for limma-voom. We use cpm in all analyses for the sake of consistency, even for the one analysis where the length normalization of FPKM might be more appropriate (i.e. comparison of $\log_2$(cpm) expression levels between lincRNAs and protein-coding genes, Figure 4).

Once we chose a suitable expression metric, we needed to choose an appropriate expression threshold to determine whether the observed expression signal is distinguishable from stochastic transcriptional noise or fragment mismapping. Limma-voom suggests a threshold of 1.0 cpm, and the threshold used in the 2012 study of lincRNA expression in *D. melanogaster* used a threshold of 1.0 FPKM (Law et al., 2014; Young et al., 2012). For a typical lincRNA of length 700nt, 1.0 FPKM translates to a cpm of 0.7. Both of these typical thresholds would eliminate most of the *D. pseudoobscura* lincRNAs from our expression analyses. Using MDS plots on batches of genes with decreasing expression, we find that sample-specific expression signal is distinguishable from noise down to 0.3 cpm. Only at expression levels lower than this does the signal collapse and samples become indistinguishable. Thus, we use a mean cpm of 0.3 across all three replicates as a threshold for expression.

The most appropriate choice of approach for detecting differential expression (DE) will vary based on experimental conditions like sample number and read distributions (Soneson and Delorenzi, 2013). The limma-voom package utilizes robust statistical methods originally designed for microarray analyses that assume normal distribution by log-transforming fragment count data and calculating a precision weight for each gene based on the relationship between the mean and variance (Law et al., 2014). In comparisons with 10 other differential-expression methods, limma-voom was among

the top performers in controlling for false discovery rate and Type I error (Soneson and Delorenzi, 2013). Three replicates were typically necessary for sufficient power to detect DE, but as we have three replicates, this was not a concern. Limma-voom continued to perform well with varying sample sizes and varying levels of DE. Only with high dispersions between sample distributions did limma-voom perform poorly. In most scenarios, we find limma-voom to be a robust choice for DE detection. A Benjamini-Hochberg adjusted p-value (i.e. FDR) of 0.01, calculated through limma-voom, is used as a cutoff for DE. With this, 1% of all detected DE genes are likely false positives, and we are comfortable with that level.

Finally, we discuss our choice to use fuzzy c-means clustering over hard k-means clustering for a co-expression analysis of gene expression in the developmental series. Hard clustering will place every gene into a cluster with a binary fit choice, in or out, and each gene can be assigned to only one cluster (Tavazoie et al., 1999). This is appropriate when clusters are quite distinct with little overlap, but this pattern is not common among developmental series data. Soft clustering with a fuzzy c-means algorithm assigns genes to a predetermined number of clusters, and genes can both belong to multiple clusters as well as be left out entirely (Futschik and Carlisle, 2005). Each gene is given a membership value between 0 and 1, with higher membership values indicating that the expression profile for that gene is strongly similar to the overall expression profile of the cluster core. With the chosen membership value of 0.5, we did not detect any multi-cluster genes, but 24.8% of all genes were left out of clusters. These orphans have expression profiles that are not representative of major developmental expression trends. Hard clustering would have forced these orphans into clusters with no way to analyze

their fit. Thus, we find soft clustering more appropriate for our developmental expression analyses.

LincRNA expression throughout development

We find that *D. pseudoobscura* lincRNAs and protein-coding genes display distinct expression properties through development. As previously observed with *D. melanogaster* lincRNAs, *D. pseudoobscura* lincRNAs tend to be more narrowly expressed than protein-coding genes (mean 2.70 stages versus 6.34), with very few expressed in all eight developmental stages (Brown et al., 2014; Young et al., 2012). The *D. melanogaster* modENCODE data only had sex-specific samples in adults, but the trend of overall increased lincRNA expression in adult males is consistent in both species. The overrepresentation of lincRNAs in two clusters that show increases in expression as male development proceeds, while not directly observed in *D. melanogaster* because of the lack of pre-adult sex-specific RNA-Seq libraries, is nonetheless consistent with this pattern.

The observation that *D. pseudoobscura* lincRNAs are overrepresented in a non-sex-biased pupae-enriched cluster is not seen in the *D. melanogaster* data. In fact, the pupal stages show the lowest lincRNA expression levels of the four major life cycle stages in *D. melanogaster* (Young et al., 2012). Key differences in methods of lincRNA identification and expression analyses might explain this difference, but studies in vertebrates have also shown evidence of high volatility in lincRNA expression evolution (Necsulea et al., 2014).

We observed equal representations of lincRNAs and protein-coding loci in several clusters that have similar expression profiles in males and females, suggesting that these

lincRNAs are expressed in somatic tissues that are similar between the two sexes. Because we largely assume that expression of protein-coding genes is biologically relevant, the observation of lincRNAs with the same expression profiles strengthens the case for biological relevance of lincRNAs.

We found significant underrepresentation of lincRNAs in two types of clusters. First, five of the six clusters that show elevated expression levels in the 1st-instar larval stage are depleted of lincRNAs. This also varies from the pattern seen in *D. melanogaster*, where lincRNA expression in the 1st-instar is higher than that seen in the pupal stage and is roughly equivalent to levels in the 3rd-instar larvae (Young et al., 2012). In general, we observe increases in the numbers of expressed lincRNAs as development proceeds, but most of that can be explained by increasing expression in the testes. We chose to sample the 1st-instar larvae so that we would have at least one pre-gonad sample in our developmental series, but we excluded entirely the embryonic stages of development. Embryos were included in analyses of *D. melanogaster* lincRNAs, and there are no clear trends of embryonic lincRNA expression. We would expect to see similar variation of lincRNA levels in *D. pseudoobscura* embryogenesis. In other words, with increased developmental sampling, we do not expect to see such a clear trend of increased lincRNA expression as development proceeds.

Second, lincRNAs are significantly underrepresented in all three clusters that show female-specific expression elevation in adults. Cluster 1, which has both elevated expression in the 1st-instar of both sexes and elevated expression in the adult female, only contains a single lincRNA.

The drop in the number of expressed genes in the adult females, which is more pronounced for lincRNAs than protein-coding genes, is curious, as it is the only reduction in expressed gene numbers seen over the course of development. Having collected the RNA for all samples ourselves, we observed that dissected ovaries yield far more total RNA than the resultant carcasses, and thus the whole-body adult female transcriptome is likely dominated by the ovaries transcriptome. Despite their high overall expression levels, the isolated ovaries contain the fewest number of both expressed lincRNAs and protein-coding genes. The tissue heterogeneity in adult females, with their mature ovaries with abundant RNA arising from small numbers of loci, likely underlies the seeming reduction in gene expression in adult females. Thus, the severe drop in lincRNA expression in adult females is most likely a drop in expression in the ovaries. We discuss gonad expression further in the next section.

The opposite trend is seen in males, where the testes contribute much less to the total RNA output of a whole body male than the carcass but have the highest number of expressed genes. 525 lincRNAs are expressed in the testes (cpm > 0.3), and 272 lincRNAs are expressed in the isolated male carcass. Combined, 711 lincRNA loci are expressed in either of these samples, yet only 481 lincRNAs are detected as expressed in the whole male body (cpm > 0.3). We reason that lincRNAs that are expressed near the 0.3 cpm threshold in isolated testes are likely to be considered unexpressed when observed as a smaller fraction of the male whole-body transcriptome. Tissue heterogeneity is a serious confounding factor in DE analyses. Careful experimental design is necessary to avoid misinterpreting trends due to tissue heterogeneity rather than

true expression dynamics. In our case, we use isolated gonads and their resulting carcasses to gain a more detailed understanding of the complex transcriptomes in adults.

Sex-biased expression of lincRNAs diverges drastically from protein-coding sex-bias

Developmental clustering suggests significant differences in levels of male-biased and female-biased lincRNA and protein-coding gene expression. DE analyses between male and female equivalents at each stage of development confirm this. Very little sex-biased expression in either direction is detected in 1st-instar larvae, but the proportion of all genes that are sex-biased increases in each progressive stage of development. Interestingly, the overall levels of sex-bias remain fairly constant between lincRNAs and protein-coding genes at each stage, but there is a heavy skew towards male-bias in lincRNAs at later stages of development.

In *D. melanogaster*, we know that the majority of sex-biased expression can be attributed to gene expression differences in the gonads (Parisi et al., 2004). The patterns of increasing numbers of male-biased and female-biased protein-coding genes from the 3rd-instar forward are consistent with this observation, as gonad development in both sexes begins around the 3rd-instar stage (Bate and Martinez Arias, 1993). Like protein-coding genes, lincRNA levels of male-bias also increase from the 3rd-instar forward, though to a much higher degree. LincRNA female-bias, however, stays low throughout development while female-biased protein-coding gene expression progressively increases.

By explicitly looking at sex-bias between the gonads and the remaining somatic carcass tissues, we confirm that the majority of sex-biased expression seen in the whole-body flies for all genes can be explained by sex-biases in the gonads. We detect sex-bias in 94.0% of lincRNAs and 84.4% of protein-coding genes in the gonads, with the sex-

bias in the testes being skewed overwhelmingly toward males, just like in adults. Very little protein-coding sex-biased expression is seen in the carcass samples, which is a little surprising considering the male accessory glands were removed from the testes and included with the carcass samples. We detect a moderate degree of male bias in lincRNAs (36.8% of expressed lincRNA loci). It would be interesting to see whether, like testes, lincRNA expression is elevated in male accessory glands, or whether the somatic male-biased expression is due to expression in common tissues between males and females.

By integrating sex-bias data from all developmental stages and adult tissues, we were able to assign each gene an overall sex-bias classification. Not surprisingly, a higher proportion of lincRNA loci (63.0%) were designated as male-biased as compared to protein-coding genes (38.3%). We see far more complexity in the expression dynamics of protein-coding genes. 17.4% of all protein-coding loci show conflicting levels of sex-bias in different samples, whether they be switches from female-biased to male-biased expression as development progresses, inverted bias between the gonads and carcasses, or sex-bias signals from the developmental series that disagree with sex-bias in the adult tissues. In contrast, only 19 lincRNAs (2.1%) show dynamic expression biases through development or between tissues. The majority of these show switches from female-biased expression to male-biased expression later in development.

Finally, while the focus of this work is on lincRNA expression dynamics, we point out that our observation of slightly higher numbers of male-biased genes than female-biased genes in adults is inverted from observations previously made in the lab using microarray technology (Abdilleh, 2014; Jiang and Machado, 2009). We attribute

this difference as mostly methodological. RNA-Seq expression analyses are, in general, more sensitive than microarray technology. Coupled with our low expression thresholds, we have considerably more power to detect DE in lowly-expressed genes, many of which we expect to be expressed in testes.

Divergent lincRNA content in the gonads

When comparing lincRNA expression bias between gonad types and the corresponding carcasses, we observe levels of DE consistent with the previously observed levels of gonadal sex-bias. High lincRNA male-bias in testes and lower male-bias in the carcasses suggest that more lincRNAs will be differentially upregulated in the testes than the male carcass, and this is confirmed via DE analysis. We see a similar trend in the female tissue samples, where the majority of expression bias is in favor of the female carcass, suggesting low numbers of ovary-expressed lincRNAs. The more equitable levels of protein-coding sex-bias in both the carcasses and gonads suggest more even distributions of DE between the two tissue types, and this is largely observed.

Using the sets of genes that show overall male or female-bias and the expression threshold of cpm > 0.3 in all tissue types, we determined that the majority of male-biased lincRNAs (424/583) were expressed exclusively in the testes. LincRNAs expressed in the ovaries tend to be expressed more broadly across tissues; only 16 of the 75 female-biased lincRNAs were ovary-specific. All told, we found that 45.8% and 1.7% of lincRNAs had testis-specific or ovary-specific expression, as compared to 15.0% and 0.8% for protein-coding genes.

High lincRNA content in testes has been observed in both *D. melanogaster* and numerous vertebrate species (Akbari et al., 2013; Brown et al., 2014; Derrien et al., 2012;

Necsulea et al., 2014). An open chromatin state that is permissive to widespread transcription has been observed in both *Drosophila* and vertebrates and has been implicated in the high lincRNA content of the testes (Hennig and Weyrich, 2013; Kimmins and Sassone-Corsi, 2005; Kleene, 2005).

The low lincRNA content in *D. pseudoobscura* ovaries has not been described in *D. melanogaster* or vertebrates, and lincRNA content is actually observed to be enriched in ovaries of the non-blood fed mosquito *Aedes aegypti* (Akbari et al., 2013; Brown et al., 2014; Derrien et al., 2012; Necsulea et al., 2014). In isolated *D. pseudoobscura* ovaries, we observed significant reductions in expressed gene number (cpm > 0.3) for both lincRNAs (77) and proteins (4306) as compared to the female carcasses, but the reduction is more severe in the lincRNAs. As opposed to the open chromatin of the testes, chromatin modifications in *Drosophila* oocytes restrict and tightly regulate gene expression (Iovino, 2014; Ivanovska and Orr-Weaver, 2006). Because of this, we suspect that lincRNAs expressed in the ovaries, though small in number, are far more likely to be actively regulated and maintained by selection than lincRNAs expressed in the testes. Further analyses are necessary to explore this possibility.

Functional implications of lincRNA expression

Can we draw any inferences about the biological roles, or lack thereof, of *D. pseudoobscura* lincRNAs based on their expression dynamics? The majority of lincRNAs show evidence of developmental, sex-biased, or tissue-biased regulation, but that in of itself is not evidence of function.

Gene Ontology (GO) analyses of the developmental expression clusters show that many lincRNAs are co-expressed with protein-coding genes that play important roles in

D. pseudoobscura biology (Ashburner et al., 2000). GO analyses, however, are insufficient to assign particular functions to any lincRNA, as co-expression can be a poor indicator of functional relationships. For example, co-expression between unrelated genes in the same open chromatin region would present similar expression signals as a gene that acts as a *cis*-activator for an adjacent gene. That said, we see two benefits of the developmental clustering and subsequent GO analyses. Together with their more narrowly-defined developmental expression profiles, the relative dearth of lincRNA-associated GO terms suggests that the scope of lincRNA biological function, if any, is smaller than that seen in protein-coding genes. Secondly, a lincRNA-associated GO term could expand the range of possible phenotypes in future lincRNA functional screens. LincRNAs were overrepresented in the pupae-enriched cluster, which had a significant GO hit to cell adhesion. Thus, it would make sense to screen knockdowns or knockouts of the lincRNAs from this cluster for adhesion defects, which might not always result in obvious lethality.

The strongest evidence of *D. pseudoobscura* lincRNA function via expression dynamics is implied from the depletion of male-biased lincRNAs and accumulation of female-biased lincRNAs on the *D. pseudoobscura* X chromosome. Similar observations have been made in *D. melanogaster* for both protein-coding genes and lincRNAs, and the most cogent models that explain either observation all invoke selection. Under the meiotic sex chromosome inactivation (MSCI) model, a silenced X chromosome during meiosis favors the buildup of advantageous testes-expressed genes on the autosomes (Betran et al., 2002). The dosage compensation hypothesis posits a different mechanism to explain the same observation (Bachtrog et al., 2010; Vicoso and Charlesworth, 2009).

The male X in *Drosophila* is hypertranscribed in order to maintain equal dosage levels between the X and autosomes in both males and females. Because of its hypertranscribed state, there is little room for modulation or further upregulation of X-linked male-bias genes, and selection favors the movement of beneficial loci to the autosomes. Neither of these models, however, adequately addresses the overrepresentation of female lincRNAs on the X. With two copies of the X in females and just one copy in males, the X spends relatively more time in female flies, thus encouraging the accumulation of advantageous dominant female-biased genes on the X (Charlesworth et al., 1987; Rice, 1984). Likewise, the reduced time of the X spends in males encourages accumulation of advantageous male-biased genes on the autosomes. Our data is consistent with all these models, and they are not necessarily mutually exclusive. Only one that invokes a sexual antagonism, however, can explain both the demasculinization and feminization of the X chromosome with respect to lincRNAs, provided that the lincRNA alleles are acting in a dominant manner.

Several studies that have analyzed the demasculinization of the X chromosome in *D. melanogaster* have shown that the subset of male-biased genes that are testis-specific are in fact evenly distributed across the X and the autosomes (Meiklejohn and Presgraves, 2012; Meisel et al., 2012). After dividing our set of male-biased genes into those that are testis-specific and those that are not, we find similar chromosomal distributions for both lincRNAs and protein-coding genes. Testis-specific genes show statistically equal distributions between the X and the autosomes, while non-testis-specific male-biased genes, both lincRNAs and protein-coding genes, are significantly underrepresented on the X. Under the selection-based models, gene movement off the X occurs only when the

gene has a fitness effect. The contrapositive, therefore, is that genes that do not show preferential movement off the X likely have no fitness effects. By this logic, the 295 testis-specific lincRNAs, which are 31.9% of the total number of expressed lincRNAs, are putatively nonfunctional.

The testis is a unique tissue in two ways. (1) The aforementioned open chromatin state permits broader transcription of the genome than in most other tissues (Hennig and Weyrich, 2013; Kimmins and Sassone-Corsi, 2005; Kleene, 2005). (2) Selective pressures like sexual conflict, sperm competition, and germline pathogens can result in the rapid evolution of genes expressed in the testis (Haerty et al., 2007; Jagadeeshan and Singh, 2005). The testis, therefore, provides an ideal environment for the origination of new genes, with both a plentiful source to draw from and pressures to keep them around. Under the "out of the testis" hypothesis for new gene origination, new transcripts with beneficial fitness effects may be selectively maintained in the testis and are more likely to evolve more efficient regulatory elements (Kaessmann, 2010; Kaessmann et al., 2009). Now in possession of gene-specific regulatory elements, the new gene has a greater probability of acquiring expression and function in other tissues. Sure enough, many newly-evolved genes show expression in the testes (Reinhardt et al., 2013; Zhao et al., 2014).

We consider the possibility that testes-specific lincRNAs are simply new genes. They may even be functional, and the lack of trafficking off the X chromosome may be a consequence of lack of time for selection to act rather than lack of function for selection to act on. That said, it is still possible that many of these lincRNAs are truly nonfunctional and that elimination will have no detrimental fitness effects to the fly, but

as we mentioned before, biological innovation requires raw material to act on. To borrow Sydney Brenner's terminology, these lincRNAs may not be transcriptional "garbage" that is eliminated from the genome, but rather transcriptional "junk" that is useless and harmless and available for innovation (Brenner, 1998).

Unlike short noncoding RNAs that are classified by structure and function, we tend to group all lincRNAs together despite little knowledge of the biological roles, if any, they play. Here, we provide support for the recognition of two subclasses of lincRNAs: (1) the testis-specific lincRNAs, which are most likely young transcripts that are expressed in open chromatin or actively regulated but either without function or with a recently-acquired role in spermatogenesis, and (2) the non-testis-specific lincRNAs, which are far less likely to be transcribed in open chromatin and, according to the "out-of-the-testes" model, far more likely to play important biological roles than the testis-specific lincRNAs.

Jiang and Machado identified three lincRNAs that were highly expressed in the testes of *D. pseudoobscura* and differentially expressed between *D. pseudoobscura* and *D. persimilis* (Jiang et al., 2011). Quantitative PCR of two of these lincRNAs also shows low, though detectable, expression in the male carcass. They speculated that these transcripts could be important for male-specific processes in *D. pseudoobscura*. Because these transcripts are broadly expressed, we find that their interpretation continues to be valid. We will consider tissue-specificity as we continue to look at lincRNA expression differences between closely related species, subspecies, and hybrids within the *pseudoobscura* subgroup.

METHODS

Generating RNA-Seq expression datasets

Additional poly(A+) library construction and RNA sequencing

Biological replicate RNA-Seq datasets were prepared using the same methods as described in Chapter 1, though multiplexed with eight samples to a lane for replicates B and C (versus two or three samples to a lane for replicate A and two to a lane for replicate D). The only other deviations from the Chapter 1 methodology are as follows: 28 flies were used for L1M_B, 35 flies were used for L1M_C, 33 flies were used for L1F_B, and 41 flies were used for L1F_C; gonad and carcass Illumina libraries for replicates B and C were generated using 500ng of starting total RNA; developmental series Illumina libraries were generated using 200ng of starting total RNA.

Quality checks of RNA-Seq replicates using multidimensional scaling (MDS) plots

Multidimensional scaling (MDS) plots were generated to assess the consistency of replicates. TopHat2 output bam files were first sorted by read names and then converted into sam files using samtools v0.1.18 (Kim et al., 2013; Li et al., 2009). Fragment counts for each locus were then obtained using HTSeq-count v0.6.1p1 on the sorted sam file (stranded = no, minimum quality score = 20, all other options default) (Anders et al., 2014). Fragment counts were scale normalized across all samples separately for the developmental series and the adult tissue samples using the calcNormFactors function from the edgeR package v3.6.8, and MDS plots were created using the plotMDS function from the edgeR package v3.6.8 (Robinson et al., 2010). Ellipses around clusters were manually added.

Choosing an expression threshold

Locus fragment counts for each replicate were converted into the cpm metric (fragment counts per million mapped fragments). Mean cpm was then calculated across all three replicates for each sample. Boxplots were made in R using all loci with a positive mean cpm in a given sample.

To find an appropriate threshold value for expression analyses, MDS plots were generated using batches of loci with varying expression levels. Loci were batched by their highest mean cpm value in any given sample from the developmental series, from cpm = 0 to cpm =1.0 in increments of 0.1 Once loci in a given batch were identified, fragment counts were scale normalized across all samples using the calcNormFactors function from the edgeR package v3.6.8, and MDS plots were created using the plotMDS function from the edgeR package v3.6.8 (Robinson et al., 2010). Ellipses around clusters were manually added. A threshold cpm of 0.3 was chosen because the batch with cpm between 0.3 and 0.4 was the lowest value batch for which distinct clusters appeared without extensive overlap. 925 lincRNA and 7,649 protein-coding loci were expressed above the 0.3 cpm threshold in at least one sample and were retained for further analysis.

Visualizing expression patterns

Heatmaps were generated using R (Team, 2014). Fragment counts of all loci above the 0.3 cpm threshold were scale normalized using calcNormFactors from the edgeR package v3.6.8 and then converted to $\log_2$(cpm) with precision weights using voom and fit to a linear model using lmfit and eBayes, all from the limma package v3.20.9 (Law et al., 2014; Robinson et al., 2010; Smyth, 2005). Log_2(cpm) values were generated independently for the developmental series and the adult tissues. LincRNA and protein-

coding loci were parsed out and clustered via Pearson's correlation using the hcluster function in the amap package v0.8-12 (Lucas, 2014). Heatmaps were then generated using the heatmap.2 function in gplots v.2.15.0 and color palettes from the RColorBrewer package v.1.1-2 (Neuwirth, 2014; Warnes et al., 2014).

Fuzzy c-means clustering of developmental expression profiles

Log_2(cpm) values for the developmental series were soft clustered using a fuzzy c-means algorithm via the R package Mfuzz (Kumar and Futschik, 2007). Expression values were standardized using the standardise function so that mean expression for each gene is zero with a standard deviation of one. Optimal cluster number *c* of 16 was determined by looking for a plateau in the minimum centroid distance using the Dmin function. The optimal fuzzifier *m* of 1.436711 was calculated using the mestimate function. After clustering, all loci with membership values less than 0.5 were removed from clusters. LincRNA over- or underrepresentation was determined using a two-tailed Fisher's exact test with a modified Bonferroni correction for multiple tests (Keppel, 1991).

Gene Ontology (GO) analyses of developmental expression clusters

Significant Biological Process GO terms for each Mfuzz-produced developmental expression cluster were identified using the web-based GeneCodis3 (Ashburner et al., 2000; Tabas-Madrid et al., 2012). *D. melanogaster* orthologs of all clustered *D. pseudoobscura* protein-coding loci were identified via OrthoDB, and the *D. melanogaster* Ensembl IDs were used as input for GeneCodis3 (Cunningham et al., 2014; Waterhouse et al., 2013). Modular enrichment analysis was performed in GeneCodis3, and lowest level GO annotations were obtained. To determine whether a GO term (minimum three genes) was significantly overrepresented in a cluster, a hypergeometric test was

performed with a false-discovery rate (FDR) correction for multiple hypothesis testing. GO terms with FDR-corrected p-values less than 0.05 were determined to be overrepresented in the cluster.

Differential expression analyses

Significant sex-bias and tissue-bias were detected using the limma-voom differential expression package (Law et al., 2014; Smyth, 2005). Fragment counts of all loci above the 0.3 cpm threshold were scale normalized using calcNormFactors from the edgeR package v3.6.8 and then converted to $\log_2$(cpm) with precision weights using voom and fit to a linear model that incorporates both sample type as well as sequencing batch using lmfit and eBayes, all from the limma package v3.20.9 (Law et al., 2014; Robinson et al., 2010; Smyth, 2005). Log_2(cpm) values were generated independently for the developmental series and the adult tissues. Pairwise contrasts were made between male and female equivalents of all four developmental samples and both adult tissue samples along with gonads and carcasses in both sexes. After Benjamini-Hochberg correction for multiple tests, significant expression bias was determined using an adjusted p-value < 0.01. Comparisons of patterns of expression bias between lincRNAs and protein-coding genes were performed using Fisher's exact test in R.

To assign an overall sex-bias designation to genes, we parsed out sex-bias observations from all six samples. Genes with male-biased expression (adj. p-value < 0.01) in at least one sample without any female-biased expression are designated "male-biased". Genes with female-biased expression (adj. p-value < 0.01) in at least one sample without any male-biased expression are designated "female-biased". Genes with both male and female bias in different samples are "dynamic-bias" genes. Genes that are

unbiased in all samples are “unbiased”. Note that overall male-biased and female-biased genes may exhibit unbiased expression in some samples.

To identify a list of testis-specific and ovary-specific genes, we first identified a set of genes (both lincRNA and protein-coding) that exclusively had expression cpm > 0.3 in only the testes or ovaries among all tissue samples. The overlap between these sets and the overall male and female-biased genes produced the set of 424 testis-specific and 16 ovary-specific lincRNAs and the set of 1,145 testis-specific and 59 ovary-specific protein-coding genes.

Demasculinization and feminization of the X chromosome

To determine whether sex-biased lincRNAs are depleted or enriched on the X chromosome as compared to the autosomes, we calculated the odds ratio (OR) between the sex-biased gene distributions (autosomes/X) and the unbiased gene distributions (autosomes/X). An odds ratio above 1.0 indicates that the X-chromosome is depleted for that class of genes, and an odds ratio below 1.0 indicates that the X-chromosome is enriched for that class of genes. We used Fisher’s exact test ($p < 0.05$) to determine whether the differences in chromosomal gene distributions are statistically significant.

CHAPTER 3: Homology of long intergenic noncoding RNAs (lincRNAs) between *D. pseudoobscura* and *D. melanogaster*

ABSTRACT

Annotated lincRNA sets from two species within the *Drosophila* genus (*pseudoobscura* and *melanogaster*) provide the opportunity to identify homologous lincRNAs using both sequence features and transcript features. We identified a set of 134 putative lincRNA homologs using reciprocal best hit local alignments via blast and positional equivalence in the genome using coordinate conversion. Then, we examined these putative homologs for evidence of conservation of several transcript-level features, including transcriptional orientation, gene structure, and developmental expression profile. Several of these putative lincRNAs homologs had TE insertions within lincRNA exonic regions that resulted in homozygous lethality, suggesting a possible function that requires further investigation. We found 65 putative lincRNAs with evidence of conservation in at least one transcript feature, and 22, including all three previously annotated *D. pseudoobscura* lincRNAs, that had evidence of conservation of two or more transcript-level features. With evidence of homology at the sequence and transcript levels, these 22 high-confidence lincRNA homologs are the best candidates for specific functional studies to start exploring directly the biological roles of lincRNAs in *Drosophila*.

INTRODUCTION

As opposed to protein-coding genes, little is known about the biological relevance of lincRNAs and the factors that constrain their evolution (Guttman et al., 2009; Haerty and Ponting, 2013; Kapusta and Feschotte, 2014; Marques and Ponting, 2009; Ponjavic et al.,

2007; Ulitsky and Bartel, 2013). We know that protein-coding sequence is primarily constrained by the amino acid code to the extent that large portions of evolutionary theory are built on differences between synonymous and nonsynonymous substitution rates (Li et al., 1985; McDonald and Kreitman, 1991). Moderate signals of purifying selection have been detected in lincRNA exonic sequence but are typically weaker than in protein-coding genes, UTRs, and smaller noncoding RNAs like microRNAs (Bhartiya et al., 2014; Guttman et al., 2009; Haerty and Ponting, 2013; Marques and Ponting, 2009; Ponjavic et al., 2007; Ward and Kellis, 2012, 2013). Even so, scant few of these comparative studies integrate transcriptomic data from multiple species into their analyses. In other words, it is unclear whether any specific lincRNA in question is even present in all the species being compared. As there would be less constraint on sequences that are not expressed in multiple species within a phylogeny, expression information is first required before analyzing the evolution of lincRNAs in more than one species.

The few evolutionary studies of lincRNAs that do consider expression information, all performed in vertebrates, reveal that the rates of lincRNA gains and losses tend to be higher than protein-coding gene turnover (Guttman et al., 2010; He et al., 2014; Homolka et al., 2011; Kutter et al., 2012; Necsulea et al., 2014; Paralkar et al., 2014; Ulitsky et al., 2011). The evolutionary origins for many are recent with, for example, 8% of human lincRNAs surveyed from six tissues originating after the split with chimpanzees and 20% not being detectable outside of chimpanzees (Washietl et al., 2014). Surprisingly, not all transcript features are well-conserved (Diederichs, 2014). Tissue-specificity and developmental-expression profiles are highly conserved (He et al., 2014; Necsulea et al., 2014; Ulitsky et al., 2011; Washietl et al., 2014). Syntenic

relationships with protein-coding genes have mixed levels of conservation, with high syntenic conservation in human-chimpanzee brains (Qu and Adelson, 2012). Similarly, intron-exon structure conservation is mixed, with some lincRNAs showing high levels of constraint in splice site sequence and others displaying rapid turnover (Guttman et al., 2010; Ulitsky et al., 2011; Washietl et al., 2014). Sequence is often poorly conserved, though small stretches of high-sequence conservation can be detected in some transcripts (Ulitsky et al., 2011; Washietl et al., 2014). Conservation of RNA structure is still poorly understood, although it is observed in a handful of lincRNAs like *HOTAIR* and the *roX* genes (He et al., 2011; Ilik et al., 2013; Johnsson et al., 2014; Schorderet and Duboule, 2011).

Few lincRNAs have been functionally investigated, particularly at a mechanistic level, but several show functional conservation despite lack of conservation of some of these transcript features. The mammalian dosage-compensator *Xist* is functionally-conserved between humans and rodents, with conservation seen in intron-exon structure and only five small sequence domains over its 17kb length (Brockdorff et al., 1991; Nesterova et al., 2001). That said, the first exon of *Xist* contains the majority of the known functional elements for the gene but has only weak sequence conservation, and the highly-conserved fourth exon appears entirely dispensable (Caparros et al., 2002). The *HOTAIR* lincRNA transcriptionally represses the *HOXD* cluster in mammals, but has low sequence conservation and poorly-conserved intron-exon structure between humans and mice (He et al., 2011; Schorderet and Duboule, 2011). This repression is achieved through the *trans*-binding of *HOTAIR* to the Polycomb Repressor Complex 2 at known sites through secondary structure, but these sites are missing in the mouse homolog.

Similarly, the 2.4kb lincRNA *megamind* is necessary for proper brain development in vertebrates (Ulitsky et al., 2011). It has poor sequence conservation in all but 200 nucleotides of its length and exhibits poor intron-exon structure conservation but strong syntenic conservation between human, mice, and zebrafish.

To this point, all comparative transcriptomic work in lincRNAs has been performed in vertebrates, so we know little about the evolution of lincRNAs in *Drosophila* or any other eukaryotes. LincRNA exonic sequences show moderate levels of conservation within the genus, less than high levels shown by protein-coding exons but higher than those shown by random intergenic sequence (Young et al., 2012). Likewise, evidence of purifying selection was found for lincRNA exons using polymorphism data taken from genomic sequence (Haerty and Ponting, 2013). Expression polymorphism data has not been considered. Few lincRNAs, like the dosage compensation *roX* genes, have been transcriptionally and functionally characterized in flies outside *D. melanogaster* (Ilik et al., 2013; Park et al., 2007).

This work is the first instance of high throughput lincRNA annotation in a non-*melanogaster* Drosophilid. Here, we cross-reference the set of *D. pseudoobscura* lincRNAs identified in Chapter 1 and the annotated set of *D. melanogaster* lincRNAs in FlyBase (r6.02) (Brown et al., 2014; St Pierre et al., 2014; Young et al., 2012). We use sequence and genome coordinate information to try to identify potentially homologous lincRNAs and examine transcript features and expression data to qualify the strength of that classification. We then discuss the strongest cases for lincRNA homology between *D. pseudoobscura* and *D. melanogaster*.

RESULTS

Identification of putative D. pseudoobscura *lincRNA homologs in* D. melanogaster

We set out to identify conserved lincRNA homologs between *D. pseudoobscura* and *D. melanogaster* using a two-pronged approach. (1) We performed reciprocal blastn searches using the full *D. pseudoobscura* and *D. melanogaster* transcriptomes and considered reciprocal best hits as putative homologs. (2) Using a *Drosophila* genus multiple genome alignment, we looked for coordinate overlap between annotated *D. pseudoobscura* and *D. melanogaster* lincRNAs (Kuhn et al., 2007).

(1) Identification through reciprocal best hits blastn searches

Using blastn and a set of parameters optimized for detecting homology in snRNAs, we queried the set of *D. pseudoobscura* lincRNAs identified in Chapter 1 against the full *D. melanogaster* transcriptome (FlyBase r6.02) (Altschul et al., 1990; Camacho et al., 2009; Mount and Nguyen, 2005; St Pierre et al., 2014). We then performed a second blastn search with the same parameters querying the best hits from *D. melanogaster* against the full *D. pseudoobscura* transcriptome (FlyBase r2.30), including newly identified lincRNAs. LincRNA loci with identical best hits in both directions were considered to be putative homologs. Using this approach, we identified 80 putative lincRNA homologs between *D. pseudoobscura* and *D. melanogaster*. Results for individual loci can be seen in Table 1.

(2) Identification through coordinate conversion from a multiple genome alignment

The UCSC Genome Browser provides multiple genome alignments for several species of *Drosophila*, including *D. melanogaster* and *D. pseudoobscura*, and the liftOver tool to convert coordinates between genome assemblies (Hinrichs et al., 2006; Kuhn et al., 2007).

We previously used this tool in Chapter 1 and verified its efficacy. We converted the genome coordinates of our *D. pseudoobscura* lincRNAs (FlyBase r2) to *D. melanogaster* coordinates (FlyBase r6) and then searched for overlap between the converted coordinates and *D. melanogaster* coordinates. We found *D. melanogaster* lincRNA matches at 174 *D. pseudoobscura* lincRNA loci. However, not all of these were one-to-one matches. Twelve *D. pseudoobscura* loci matched multiple *D. melanogaster* loci, and 10 *D. melanogaster* loci matched multiple *D. pseudoobscura* loci. Because we could not precisely determine which of these assignments were correct, we have not considered these multi-matchers further.

After filtering, we were left with one-to-one matches at 115 loci. The majority of these coordinate conversion matches agreed with the results of the reciprocal blastn search; only a single *D. melanogaster* locus, FBgn0031778, matched a different *D. pseudoobscura* locus, albeit an adjacent one. In this case, we reason that the local alignment is more reliable than a whole genome alignment, and we retained the reciprocal blast match between *D. melanogaster* FBgn0031778 and *D. pseudoobscura* XLOC_006569. Therefore, we found 114 unambiguous putative lincRNA homologs via coordinate conversion; results for individual loci can be found in Table 1.

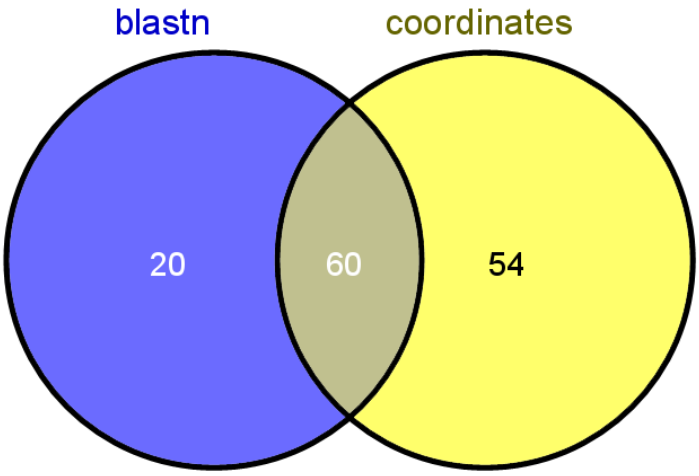


Figure 1 – Putative lincRNA homologs identified using two methods. Putative lincRNA homologs between *D. pseudoobscura* and *D. melanogaster* were identified via reciprocal blastn searches (blastn) and coordinate conversions using a whole-genome alignment (coordinates).

We found good agreement between the reciprocal blast matches and the coordinate conversion matches, with 60 putative lincRNA homologs found using both methods (Figure 1). Fifty-four putative lincRNA homologs were identified solely using coordinate conversions. Taken together, we have identified 134 putative lincRNA homologs between *D. pseudoobscura* and *D. melanogaster*, which is 8.4% of the total number of *D. pseudoobscura* lincRNA loci.

Assessing evidence of homology

For each putative lincRNA, we assessed other potential features that suggest homology including: (1) conservation of transcriptional orientation, (2) conservation of gene structure, (3) conservation of developmental expression profile, and (4) lethal transposable element insertions in the exonic sequence.

(1) Conservation of transcriptional orientation

Functional mechanisms of described lincRNAs vary, but in all cases the orientation of transcription appears important. Because we used unstranded RNA-Seq, we only know

the transcriptional orientation for multi-exon transcripts in which orientation can be inferred from canonical splice donor/acceptor sites. Of the 134 putative lincRNA homologs, 55 are multi-exon and 79 are single exon. We used the nearest neighboring orthologous protein-coding genes as references to ascertain the transcriptional orientation of the putative lincRNA homolog. We found conservation of transcriptional orientation in 35 putative homologs (63.6%), and these are listed in Table 1.

We scrutinized the local genomic environment in more detail for the 20 putative lincRNAs for which we could not find evidence of transcriptional orientation conservation. For eight of these putative homologs, the nearest neighboring protein-coding orthologs in *D. pseudoobscura* were not detectable in the *D. melanogaster* genome, suggesting genomic rearrangements that altered gene synteny. Likewise, the lincRNA XLOC_002512 is located within a cluster of paralogous genes, the *Ccp84A* genes, and is located between different pairs of paralogs in either genome; whether that is due to transposition or incorrect assignment of orthology, either at the protein-coding or lincRNA loci, is unclear. Three putative homologs were located adjacent to another lincRNA transcribed in the opposite direction, possibly confounding our results. Finally, eight putative homologs had unambiguously inverted transcriptional orientations.

(2) Conservation of gene structure

We searched for evidence of conservation of gene structure in the 35 putative lincRNA homologs with conserved transcriptional orientation between *D. pseudoobscura* and *D. melanogaster*. We considered both overlap at the 5' and 3' ends of transcript models as well as conservation of intron-exon structure, and we found 10 putative homologs that share at least one end within 40 nt. Seven of these were at the 5' end, five were at the 3'

end, and two shared both ends (Figure 2). These two are the previously annotated lincRNA *RNaseP:RNA* and XLOC_000186. We did not detect any evidence of conserved intron-exon structure. Results are listed in Table 1.

Figure 2 – Local alignments show conserved transcript boundaries between putative *D. pseudoobscura*/*D. melanogaster* lincRNA homologs. (A) Both the 5' and 3' ends of the XLOC_000186 transcript are weakly conserved with its putative homolog, FBgn0262993. FBtr0306833 and Fbtr0306834 are two isoforms expressed at the FBgn0262993 locus. (B) The 3' ends of XLOC_002366 and FBgn0261814 are strongly conserved. FBtr0303300 is a transcript of the FBgn0261814 locus. * represent 100% consensus sites.

(3) Conservation of developmental expression profile

With developmental poly(A+) RNA-Seq now available in both *D. pseudoobscura* and *D. melanogaster*, we looked for correlations between developmental expression profiles (Brown et al., 2014; Graveley et al., 2011; Young et al., 2012). Because the *D. melanogaster* RNA-Seq data is not sex-specific before the adult stage, we pooled our male and female data together for the 1st-instar larval, 3rd-instar larval, and pupal stages. We included these three stages and whole-body adult males and females in our analyses.

We generated $\log_2$(cpm) expression values for both *D. pseudoobscura* and *D. melanogaster* loci as described in Chapter 2 and determined Pearson's correlation coefficients for all 134 putative lincRNA homologs. Sixty-nine of these lincRNA homologs fell below the 0.3 cpm threshold in at least one species, so no correlation coefficient was generated. Of the 65 with expression data in both species, 21 (32.3%) have correlation coefficients above 0.9, indicating strong correlation. 36 (55.4%) and 43 (66.2%) have correlation coefficients greater than 0.7 and 0.5, respectively (Figure 3). To compare, we also determined correlation coefficients for 7,845 orthologous protein-coding genes. Of these, 2,048 were not expressed above our threshold. Expression correlations for protein-coding orthologs were only moderately higher than those for lincRNAs, though not significantly so (Mann-Whitney test, $p = 0.4385$), with 1,940 (33.5%) with $r > 0.9$, 3,422 (59.0%) with $r > 0.7$, and 4,157 (71.7%) with $r > 0.5$. Pearson's correlation coefficients for all putative lincRNA homologs are listed in Table 1.

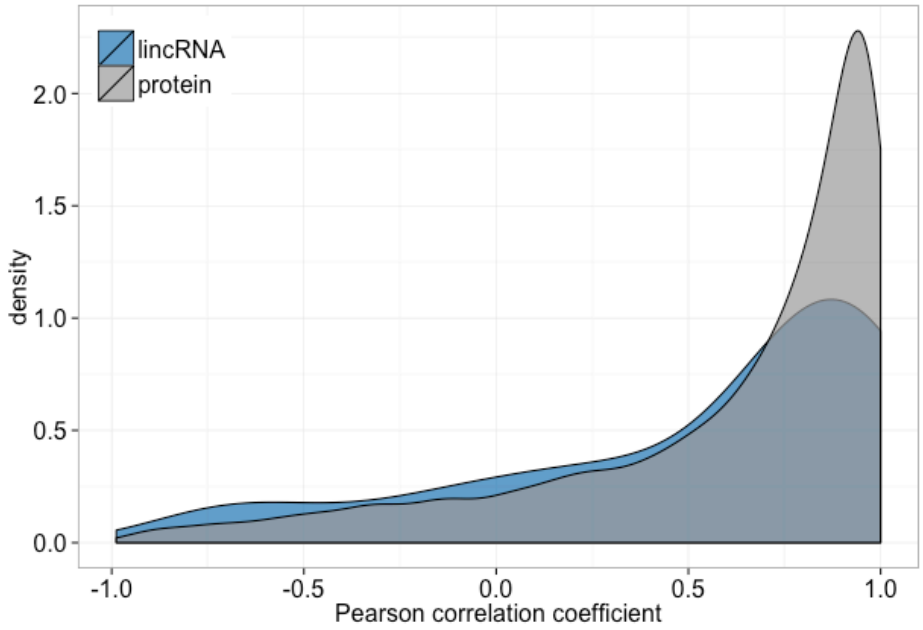


Figure 3 – Density distribution of Pearson correlation coefficients for developmental expression profiles between *D. pseudoobscura* and *D. melanogaster*. Shown are density distributions for Pearson correlation coefficients from 65 putative lincRNA homologs and 5,797 protein-coding orthologs between *D. pseudoobscura* and *D. melanogaster*. Distributions are not significantly different (Mann-Whitney, p=0.4385).

Finally, we cross-referenced the set of putative lincRNA homologs with high correlation coefficients ($r > 0.5$) with the developmental expression clusters generated via soft clustering in Chapter 2 to ascertain whether these potential homologs have diverse expression dynamics or are limited to narrow profiles (i.e. testes expression). While we did find the highest numbers of homologs in clusters that had an overrepresentation of lincRNAs (clusters 2, 4, and 14; male development and pupal expression), we also found putative homologs in 10 of the total 16 clusters, including clusters that were underrepresented for lincRNAs (clusters 11 and 16; pupal inactivation and L1 expression). Putative lincRNAs, therefore, can have quite varied developmental profiles, as seen by a few specific lincRNAs with high correlation coefficients (Figure 4).

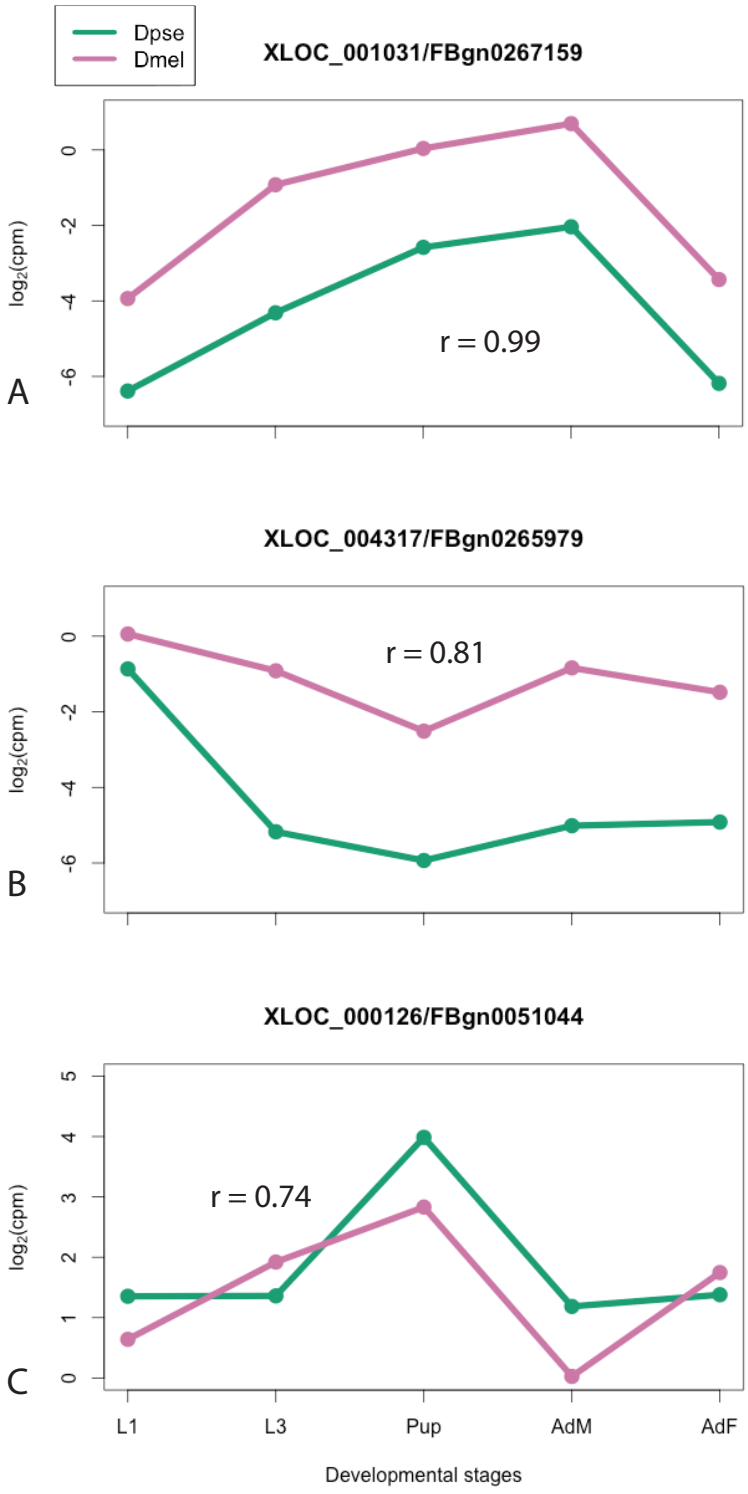


Figure 4 – High correlations between putative lincRNA homolog expression profiles. Expression profiles in $\log_2$(cpm) of putatively homologous lincRNAs in *D. pseudoobscura* and *D. melanogaster* can be highly correlated (Pearson correlation coefficients are shown). (A) *D. pseudoobscura* locus XLOC_001031 and *D. melanogaster* locus FBgn0267159. (B) XLOC_004317 and FBgn0265979. (C) XLOC_000126 and FBgn0051044.

(4) Identification of lethal transposable element (TE) insertions in exonic sequence of putative lincRNA homologs

Resources for genetic manipulation in *D. pseudoobscura* are minimal, but there are many tools to manipulate gene function in *D. melanogaster*. Thousands of genome-wide transposable element (TE) insertion stocks are available through major stock centers around the world (Bellen et al., 2004). We cross-referenced the set of putative lincRNA homologs with existing TE insertion stocks in the Bloomington *Drosophila* Stock Center. We found 21 putative lincRNA homologs with TE insertions in their exonic sequence.

Bloomington also provides crude phenotypic information for each of their TE insertion stocks, typically whether the TE insertion results in lethality when homozygosed. Of the 21 loci with TE insertions, 16 carry only TE insertions that are homozygous viable. The other five loci have at least one TE insertion that is fatal when homozygous. One of these loci is the heat shock response *HSR-omega* lincRNA, which has already been functionally characterized in *D. melanogaster* (McColl and McKechnie, 1999). We obtained the annotated homozygous lethal TE stocks for the other four putative homologs and verified their lethality by screening for genotype ratios. Three of these lincRNAs are located on the *D. melanogaster* third chromosome, and screening was as simple as checking to make sure all flies in the population carried dominant balancer markers like Tubby, Serrate, and Stubble (Greenspan, 2004). For lincRNA XLOC_018262/FBgn0263039, we identified zero homozygotes in total sample sizes of 49 (stock #10240), 79 (stock #10154), 133 (stock #11677), and 100 (stock #10155). For lincRNA XLOC_000126/FBgn0051044, we identified zero homozygotes in total sample sizes of 96 (stock #33239) and 95 (stock #11637). For lincRNA

XLOC_017214/Fbgn0267665, we identified zero homozygotes in a total sample size of 93 (stock #36157).

LincRNA XLOC_005786/FBgn0263504 is X-linked and carries a P-element insertion on the X chromosome marked with a mini-white eye color marker (stock #11872). The stock is balanced with an FM7c chromosome that lacks a dominant marker. We screened 45 females, all with the mini-white marker and thus had at least a single copy of the TE. All fourteen males, however, had white eyes and the recessive X-linked singed bristle marker, suggesting that the TE insertion is truly lethal in a homozygous or hemizygous environment. The full set of lincRNAs with lethal and viable TE insertion stocks is listed in Table 1.

Compiling evidence of lincRNA homology

Having performed two analyses to first identify potential lincRNA homologs and four analyses to assess that classification, we compiled all these data into Table 1. Of the 134 putative lincRNA homologs, 65 (48.5%) have additional support by at least one transcript-level feature. Sixteen putative homologs have support from two transcript-level features, and six had support from three different features. There were not any putative homologs that had, all together, conserved transcriptional orientation, conserved gene structure, conserved developmental expression profile, and a lethal TE insertion within its exonic region.

Dpse_homolog	Dmel_homolog	blastn	coordinates	orientation	gene structure	Pearson	TE stocks
XLOC_000009	FBgn0046696	+	+	+	5'&3'	NA	NA
XLOC_000086	FBgn0267782	-	+	+	-	0.67	viable
XLOC_000104	FBgn0265859	+	-	-	NA	NA	NA
XLOC_000126	FBgn0051044	+	-	+	-	0.74	lethal
XLOC_000186	FBgn0262993	-	+	+	5'&3'	NA	viable
XLOC_000292	FBgn0001234	+	+	+	-	NA	lethal
XLOC_000566	FBgn0265162	-	+	-	NA	NA	NA
XLOC_000718	FBgn0267222	-	+	-	NA	NA	NA
XLOC_000848	FBgn0267149	+	+	+	-	0.98	NA
XLOC_000967	FBgn0261504	+	+	+	-	0.60	NA
XLOC_001031	FBgn0267159	+	+	-	NA	0.99	NA
XLOC_001208	FBgn0264905	-	+	-	NA	0.64	NA
XLOC_001368	FBgn0262025	+	+	+	5'	1.00	NA
XLOC_001695	FBgn0266257	+	+	+	-	0.98	NA
XLOC_002193	FBgn0265204	+	+	-	NA	NA	viable
XLOC_002345	FBgn0266248	+	+	+	5'	0.46	NA
XLOC_002366	FBgn0261814	+	+	+	3'	0.94	NA
XLOC_002512	FBgn0063127	+	+	-	NA	0.93	NA
XLOC_002664	FBgn0267678	-	+	-	NA	NA	NA
XLOC_003033	FBgn0267112	+	-	NA	NA	0.93	NA
XLOC_003075	FBgn0265157	+	+	NA	NA	0.88	NA
XLOC_003099	FBgn0266260	+	+	NA	NA	NA	NA
XLOC_003104	FBgn0266243	-	+	NA	NA	0.96	viable
XLOC_003125	FBgn0265198	+	+	NA	NA	NA	NA
XLOC_003180	FBgn0267485	-	+	NA	NA	NA	NA
XLOC_003205	FBgn0263388	-	+	NA	NA	0.97	NA

Dpse_homolog	Dmel_homolog	blastn	coordinates	orientation	gene structure	Pearson	TE stocks
XLOC_003212	FBgn0266232	+	+	NA	NA	NA	NA
XLOC_003227	FBgn0054046	-	+	NA	NA	NA	NA
XLOC_003232	FBgn0267124	+	+	NA	NA	-0.76	NA
XLOC_003250	FBgn0020556	-	+	NA	NA	-0.02	NA
XLOC_003260	FBgn0051084	-	+	NA	NA	NA	NA
XLOC_003265	FBgn0265379	-	+	NA	NA	NA	NA
XLOC_003274	FBgn0261503	-	+	NA	NA	NA	NA
XLOC_003319	FBgn0263623	+	-	NA	NA	NA	NA
XLOC_003404	FBgn0267196	+	+	NA	NA	-0.73	NA
XLOC_003411	FBgn0051386	-	+	NA	NA	0.74	NA
XLOC_003495	FBgn0266818	+	+	+	-	NA	viable
XLOC_003972	FBgn0265644	-	+	+	-	NA	NA
XLOC_004013	FBgn0263290	-	+	+	-	-0.11	NA
XLOC_004037	FBgn0267647	+	-	-	NA	NA	NA
XLOC_004123	FBgn0264601	+	-	-	NA	0.98	NA
XLOC_004189	FBgn0265763	+	+	-	NA	NA	NA
XLOC_004201	FBgn0265840	-	+	+	-	NA	NA
XLOC_004312	FBgn0265654	+	+	+	5'	NA	NA
XLOC_004317	FBgn0265979	-	+	-	NA	0.81	NA
XLOC_004482	FBgn0052835	-	+	+	-	0.93	NA
XLOC_004508	FBgn0265044	-	+	-	NA	-0.67	NA
XLOC_004540	FBgn0266763	-	+	-	NA	0.94	NA
XLOC_004806	FBgn0265661	+	+	+	-	0.27	viable
XLOC_005423	FBgn0265338	-	+	+	-	0.69	NA
XLOC_005758	FBgn0265104	-	+	NA	NA	NA	NA
XLOC_005759	FBgn0266816	+	+	NA	NA	NA	NA

Dpse_homolog	Dmel_homolog	blastn	coordinates	orientation	gene structure	Pearson	TE stocks
XLOC_005761	FBgn0265635	+	+	NA	NA	NA	NA
XLOC_005780	FBgn0265639	+	+	NA	NA	NA	NA
XLOC_005786	FBgn0263504	-	+	NA	NA	0.74	lethal
XLOC_005818	FBgn0265985	+	-	NA	NA	0.28	NA
XLOC_005854	FBgn0266867	-	+	NA	NA	NA	NA
XLOC_005931	FBgn0266766	+	-	NA	NA	NA	viable
XLOC_005937	FBgn0265765	+	+	NA	NA	0.78	NA
XLOC_005956	FBgn0265106	+	+	NA	NA	0.99	NA
XLOC_005974	FBgn0265938	-	+	NA	NA	-0.60	NA
XLOC_006405	FBgn0266140	-	+	+	-	0.78	NA
XLOC_006482	FBgn0265945	-	+	+	-	NA	NA
XLOC_006569	FBgn0031778	+	-	NA	NA	0.98	viable
XLOC_006572	FBgn0266827	+	+	NA	NA	NA	NA
XLOC_006585	FBgn0265255	+	+	NA	NA	NA	NA
XLOC_006587	FBgn0266223	+	-	NA	NA	NA	NA
XLOC_006590	FBgn0266225	-	+	NA	NA	NA	NA
XLOC_006593	FBgn0266902	-	+	NA	NA	NA	NA
XLOC_006646	FBgn0263866	+	+	NA	NA	0.99	NA
XLOC_006999	FBgn0264549	+	+	-	NA	0.74	NA
XLOC_007513	FBgn0264370	+	+	+	-	NA	NA
XLOC_007643	FBgn0266844	+	-	+	3'	NA	NA
XLOC_007662	FBgn0267271	-	+	-	NA	0.79	NA
XLOC_007832	FBgn0264994	+	+	NA	NA	-0.42	NA
XLOC_007837	FBgn0262353	-	+	NA	NA	0.19	NA
XLOC_007893	FBgn0267294	+	+	NA	NA	NA	NA
XLOC_007896	FBgn0263019	+	+	NA	NA	0.34	viable

Dpse_homolog	Dmel_homolog	blastn	coordinates	orientation	gene structure	Pearson	TE stocks
XLOC_007900	FBgn0266886	+	+	NA	NA	0.16	viable
XLOC_007907	FBgn0266894	-	+	NA	NA	NA	NA
XLOC_007911	FBgn0266323	-	+	NA	NA	NA	NA
XLOC_007916	FBgn0266032	+	+	NA	NA	-0.37	NA
XLOC_007927	FBgn0265585	+	+	NA	NA	NA	NA
XLOC_007930	FBgn0267579	+	+	NA	NA	NA	NA
XLOC_007944	FBgn0266823	+	+	NA	NA	0.85	NA
XLOC_008032	FBgn0264943	-	+	-	NA	NA	NA
XLOC_008647	FBgn0265947	+	-	NA	NA	NA	NA
XLOC_008670	FBgn0264944	+	+	NA	NA	NA	NA
XLOC_008679	FBgn0266158	+	+	NA	NA	0.96	NA
XLOC_008706	FBgn0263331	+	+	NA	NA	NA	NA
XLOC_008862	FBgn0265587	+	+	+	-	NA	NA
XLOC_008897	FBgn0266313	-	+	+	-	0.96	NA
XLOC_013009	FBgn0267087	+	+	-	NA	0.17	viable
XLOC_013865	FBgn0264507	-	+	NA	NA	-0.21	NA
XLOC_013875	FBgn0265454	-	+	NA	NA	0.54	NA
XLOC_013876	FBgn0265922	-	+	NA	NA	0.97	NA
XLOC_013895	FBgn0266095	+	+	NA	NA	0.57	NA
XLOC_014954	FBgn0265700	+	+	+	-	NA	NA
XLOC_014967	FBgn0264384	-	+	+	-	0.24	NA
XLOC_015225	FBgn0265865	-	+	NA	NA	-0.69	NA
XLOC_015260	FBgn0267099	-	+	NA	NA	NA	viable
XLOC_015311	FBgn0267090	+	+	NA	NA	NA	NA
XLOC_015320	FBgn0265902	-	+	NA	NA	-0.02	NA
XLOC_015359	FBgn0261522	+	+	+	5'	0.94	NA

Dpse_homolog	Dmel_homolog	blastn	coordinates	orientation	gene structure	Pearson	TE stocks
XLOC_015570	FBgn0267173	-	+	NA	NA	NA	viable
XLOC_015573	FBgn0030911	+	+	NA	NA	NA	NA
XLOC_015578	FBgn0265918	+	+	NA	NA	-0.08	NA
XLOC_015585	FBgn0266199	-	+	NA	NA	NA	NA
XLOC_015592	FBgn0267166	+	+	NA	NA	NA	viable
XLOC_015635	FBgn0265967	+	-	+	-	NA	NA
XLOC_016186	FBgn0266537	+	+	-	NA	0.85	NA
XLOC_016287	FBgn0266952	+	-	+	-	NA	NA
XLOC_016322	FBgn0266985	-	+	+	-	NA	NA
XLOC_016488	FBgn0267615	-	+	+	5'	0.84	NA
XLOC_016553	FBgn0265913	+	+	+	3'	0.78	NA
XLOC_016923	FBgn0264462	-	+	+	-	NA	viable
XLOC_016985	FBgn0265932	+	-	-	NA	NA	NA
XLOC_017119	FBgn0266947	+	+	NA	NA	NA	NA
XLOC_017139	FBgn0267219	+	+	NA	NA	0.97	NA
XLOC_017144	FBgn0265745	+	+	NA	NA	0.98	NA
XLOC_017189	FBgn0266255	+	-	NA	NA	NA	NA
XLOC_017214	FBgn0267665	+	-	NA	NA	0.56	lethal
XLOC_017220	FBgn0267666	+	-	NA	NA	0.87	NA
XLOC_017254	FBgn0266786	-	+	NA	NA	0.30	NA
XLOC_017263	FBgn0267794	-	+	NA	NA	NA	viable
XLOC_018032	FBgn0266966	+	-	+	-	NA	NA
XLOC_018228	FBgn0266979	+	+	NA	NA	NA	NA
XLOC_018262	FBgn0263039	-	+	NA	NA	NA	lethal
XLOC_018273	FBgn0264705	-	+	NA	NA	-0.13	NA
XLOC_018274	FBgn0266771	+	+	NA	NA	NA	NA

Dpse_homolog	Dmel_homolog	blastn	coordinates	orientation	gene structure	Pearson	TE stocks
XLOC_018293	FBgn0265893	+	+	NA	NA	NA	NA
XLOC_018294	FBgn0262690	+	-	NA	NA	NA	NA
XLOC_018307	FBgn0267473	+	+	NA	NA	NA	NA
XLOC_018311	FBgn0265719	-	+	NA	NA	0.89	NA

Table 1 – Evidence of potential homology for 134 putative lincRNAs between *D. pseudoobscura* and *D. melanogaster*. The *D. pseudoobscura* lincRNA locus ID and its *D. melanogaster* counterpart are listed in the first two columns. "blastn" refers to results from reciprocal blast searches for homology. "coordinates" refers to matches from the direct genome coordinate conversion. "orientation" refers to the conservation of direction of transcription relative to syntenic genes. "gene structure" refers to conservation of transcript model ends or intron-exon structure. "Pearson" is the Pearson correlation coefficient between developmental profiles of *D. pseudoobscura* and *D. melanogaster*. "TE stocks" refers to available TE insertion stocks in the lincRNA exonic region in *D. melanogaster*, with the broad lethal/viable phenotype given. If analysis was not possible for a particular locus, it was assigned "NA".

DISCUSSION

Our comparative transcriptomic analyses have revealed the first large set of putative lincRNA homologs within the *Drosophila* genus. Here, we discuss the efficacy of the different methods we used to support a classification of homology between lincRNAs, and we discuss in more detail the particular loci that have strong evidence in support of homology.

Building a case for homology between lincRNAs

As mentioned before, lincRNA sequence is often poorly conserved over large areas of a transcript with only short domains of conservation (Brockdorff et al., 1991; Brockdorff et al., 1992; Diederichs, 2014; He et al., 2011; Ulitsky et al., 2011). Even so, both of the methods we used to identify a pool of putative lincRNA homologs rely on sequence conservation. Local alignments via blastn are problematic when conservation is low. Therefore, we used blastn parameters that were more tolerant of gaps and mismatches to compensate for this (Mount and Nguyen, 2005). The coordinate conversion approach relies on a multiple genome alignment, though the sequence alignment in the precise region of the lincRNA, even if poor, can be anchored by strong conservation in adjacent regions. Therefore, we are not surprised that we identified more potential homologs via the coordinate conversion approach. Recall that we initially identified 174 *D. pseudoobscura* lincRNA loci with coordinate matches to *D. melanogaster* lincRNAs. A fair number of those, however, matched to multiple lincRNA loci, so that a precise assignment of homology could not be made. Whether these are paralogous lincRNAs within a genome, true lincRNA-dense regions, or a consequence of poor transcript models is unclear. The annotated *D. melanogaster* models, however, have been finished

with RNA-PET while our *D. pseudoobscura* models have not (Brown et al., 2014). Therefore, the possibility that neighboring *D. melanogaster* lincRNAs are incompletely assembled transcripts is low. The 114 identified lincRNA are likely an underestimate for the total number of positionally-equivalent lincRNAs between the two genomes.

With a set of 134 putative lincRNA homologs, we used transcript-level features to strengthen the case for homology for select loci. Somewhat surprisingly, more lincRNA loci showed evidence of developmental expression conservation than any other type of conservation. While several studies have shown high tissue and developmental stage-specificity, lincRNA expression has also been characterized as dynamic over evolutionary time (He et al., 2014; Kutter et al., 2012; Necsulea et al., 2014; Qu and Adelson, 2012; Washietl et al., 2014). Also consider the methodological differences between the two sets of RNA-Seq data. They were sequenced with different read lengths, with different numbers of replicates, with vastly different library sizes, in different labs by different people, with pooled male and female reads in only one species, and almost certainly with imprecise staging of developmental stages between species. Despite all this, 66.2% of all lincRNA loci with detectable expression in both species have a correlation coefficient above 0.5, which is only slightly less than the 71.7% seen with protein-coding orthologs and is not different statistically.

It is true that a larger percentage of lincRNA loci than protein-coding loci (51.5% versus 26.1%) were not included in the correlation analysis because of expression below threshold, likely the result of low sampling rather than lincRNA turnover. The evolutionary patterns we present for lincRNAs with detectable expression in both *D. pseudoobscura* and *D. melanogaster* are similar to those we see for protein-coding genes.

We were also surprised with the low conservation seen in transcript orientation between putative homologs. Only 35 of the 55 *D. pseudoobscura* lincRNAs with orientation information showed conservation with its *D. melanogaster* homolog, assuming the orientation identification made by Cufflinks for the *D. pseudoobscura* lincRNAs is correct (Trapnell et al., 2010). Some of these 20 could be explained by genome rearrangements or other genomic messiness, but eight of them appeared unambiguously reversed. It is unlikely that a noncoding transcript would retain its function after such a switch and that these "homologs" may actually be independently evolved transcripts. Strand-specific RNA-Seq could directly resolve these ambiguities.

We found little evidence for conservation of gene structure, including no instances of conserved intron-exon structure using 35 multi-exon lincRNAs from *D. pseudoobscura*. This is consistent with other work that suggests that splice site position may not be highly constrained in lincRNAs (Washietl et al., 2014). We hesitate to even call the ends of our transcript models transcription start and stop sites, as we currently lack any type of transcript end sequencing like RNA-PET that could provide clearer transcript boundaries (Fullwood et al., 2009). Since we have ample total RNA leftover from generating our RNA-Seq libraries, obtaining this transcript end and, likewise, strand-specificity information should be prioritized.

While evidence of functionality in both species would be ideal for ascribing homology, evidence of functionality in even one provides a selective rationale for maintaining a gene over time. To that end, we examined a handful of fly lines with TE insertions in lincRNA exonic regions and confirmed a lethal phenotype. This is encouraging data but far from conclusive about specific lincRNA functionality, as the TE

could also be disrupting other crucial genomic elements in the region. Further experiments with RT-PCR of mutant and wild-type individuals to correlate transcript level with TE disruption, rescue experiments, and complementary knockout experiments are all necessary to definitely link the lincRNA to the lethality. CRISPR technology in particular holds much promise for lincRNA research, as CRISPRs would facilitate removal or alteration of small target regions that may be crucial for lincRNA function (i.e. transcription start sites or splice sites) (Mali et al., 2013).

Finally, we do not suggest that those lincRNAs with viable TE insertions are necessarily nonfunctional. Non-lethal mutations are certainly possible, though much more difficult to screen for. Knowledge of developmental expression and associated GO terms could provide possible phenotypes to assay.

Integrating sequence and transcript-level features to uncover lincRNA homology

Several studies that have attempted to identify homologous lincRNAs between species of vertebrates rely exclusively on sequence homology, which we point out again is often poorly conserved in lincRNAs, and positional equivalence (Guttman et al., 2010; Necsulea et al., 2014). We chose to incorporate other transcript features into our assessments of homology (Diederichs, 2014; Washietl et al., 2014).

Before we delve into the full list of putative lincRNA homologs, we note that three lincRNAs were previously annotated in the *D. pseudoobscura* genome prior to our analyses (St Pierre et al., 2014). They include: *RNaseP:RNA* (XLOC_000009)*, HSR-omega* (XLOC_000292), and *SRP* (XLOC_000967). All three of these lincRNAs were retained in our initial list of putative lincRNA homologs. Further, homology for each is supported by two additional transcript-level features. All three have conserved

transcriptional orientation with their *D. melanogaster* homolog. In addition to that, *RNaseP:RNA* has shared 5' and 3' ends with its homolog's transcript model. We previously mentioned that a TE insertion into the *HSR-omega* exonic sequence causes a lethal disruption, and *SRP* has a moderately high expression correlation coefficient of 0.6 with its *D. melanogaster* homolog. Because we knew *a priori* that these genes should show strong evidence of homology, we argue that the 19 other lincRNA loci with two or more conserved transcript features, using an expression correlation coefficient greater than 0.5, should be considered high-confidence homologs.

Like the previously-annotated lincRNAs, the vast majority of the other 19 high-confidence homologs have conserved transcriptional orientation with their *D. melanogaster* homologs. All 10 of the loci with evidence of gene structure conservation are necessarily in this list, as conserved transcriptional orientation was a requirement for that assay. Four of the five loci with lethal TE insertions are similarly on this list, and 16 of the 19 high-confidence homologs have expression correlation coefficients above 0.5. In terms of developmental expression clustering, all high-confidence homologs cluster in groups with increasing male bias through development or unbiased elevated expression during mid-development. All of these clusters contain an equal or overrepresentation of lincRNAs as compared to protein-coding genes; there are no high-confidence homologs in clusters with an underrepresentation of lincRNAs.

Six of the high-confidence homologs have three levels of support. For five of the six, that includes conservation of transcriptional orientation, gene structure, and developmental expression profiles. Three of these cluster into groups with progressively increasing levels of expression in males after the 3rd-instar larval stage. Interestingly, all

three of these loci are testes-specific, with no expression detected in the male carcass. While the high expression correlations could be consistent with convergent expression due to open chromatin, we think the evidence of conserved gene structure suggests active regulation inside the testes.

The last high-confidence homolog, XLOC_000126, has conserved transcriptional orientation and correlated expression with its *D. melanogaster* homolog. It clusters into cluster 4, which has genes with elevated unbiased expression in the pupal stage. It also has several lethal TE insertions within its exonic region. One of these stocks, #33239, carries a dominant Tubby balancer that is visible by the pupal stage. All pupal casings in the vials from this stock exhibit the Tubby phenotype, indicating that they are heterozygous for the balancer chromosome and the chromosome with the TE insertion. The homozygotes never make it to the pupal stage, suggesting that the lethality occurs in the larval stages or prior. Interestingly, this is one of the few *D. pseudoobscura* loci for which we obtained an rfam match, suggesting that this lincRNA could serve as a primary miRNA transcript for *mir-996* (Nawrocki et al., 2014).

As these lincRNAs display the strongest evidence for lincRNA homology within the genus, they will be the focus as we pivot to evolutionary expression and population genetics studies of lincRNAs in the *pseudoobscura* subgroup. As we generate more expression data from more species within *Drosophila*, as we currently have for *D. persimilis*, we can continue the search for homologous lincRNAs. Perhaps some of the transcript features, like gene structure, will be more amenable to analysis at shorter evolutionary distances.

While we have not directly answered the ever-present questions of biological relevance of *Drosophila* lincRNAs, we have generated data that will better inform those future attempts. The broad genome sequence and transcriptome data that we have generated for multiple populations of both *D. pseudoobscura* and *D. persimilis* will facilitate powerful evolutionary analyses that will enable us to detect evidence of selection for these lincRNA loci. Further, functional analyses of the 19 newly-described high-confidence homologs will begin to shed light on the biological roles of lincRNAs in invertebrates.

METHODS

Reciprocal blastn searches

Reciprocal blastn searches were performed using parameters optimized for identification of snRNA homologs between human and *D. melanogaster* (Altschul et al., 1990; Camacho et al., 2009; Mount and Nguyen, 2005). The first blastn search queried the set of 1,771 *D. pseudoobscura* lincRNA transcripts against the full *D. melanogaster* transcriptome (FlyBase r6.02) with parameters: word_size = 7, gapopen = 10 , gapextend = 6, reward = 5, penalty = -4 (St Pierre et al., 2014). Fasta sequence was obtained for the *D. melanogaster* best hits and used as query for a blastn search with identical parameters against a database with all transcripts from the *D. pseudoobscura* annotation (r2.30) and the set of 1,771 *D. pseudoobscura* lincRNAs. Best hits that matched were retained as putative lincRNA homologs.

D. pseudoobscura ***to*** **D. melanogaster** ***coordinate conversion***

We searched for putative lincRNA homology between the 1,589 *D. pseudoobscura* lincRNA loci and the 2,359 annotated lincRNA loci in the *D. melanogaster* genome

(FlyBase r6.02) (St Pierre et al., 2014). As the *D. melanogaster* FlyBase r6 assembly was new at the time of analysis, we converted *D. pseudoobscura* lincRNA coordinates (FlyBase r2 or UCSC dp4) first to *D. melanogaster* FlyBase r5 coordinates (i.e. UCSC Dm3) with the UCSC liftOver tool and the dp4ToDm3.over.chain chain file (Hinrichs et al., 2006; Kuhn et al., 2007). Conversion parameters for liftOver were: input = gff, allow multiple output regions = YES, minimum match = 0.1. We then converted *D. melanogaster* coordinates from FlyBase r5 to r6 in FlyBase, and looked for overlap using the intersectBed utility from BEDtools (Quinlan and Hall, 2010; St Pierre et al., 2014). Loci with multiple matches in either direction were not considered further, and unambiguous one-to-one lincRNA locus matches were retained as putative lincRNA homologs. A Venn diagram was generated with VENNY (Oliveros, 2007).

Detecting conservation of transcriptional orientation

We compared the transcriptional orientation of putative lincRNA homologs in relation to the nearest neighboring protein-coding genes with orthologs in both species. Protein orthologs were identified using OrthoDB designations available through FlyBase annotations, and orientation calls were made in a high-throughput manner using custom perl scripts (Kriventseva et al., 2008; St Pierre et al., 2014; Waterhouse et al., 2013). For putative homologs that lacked conservation of transcriptional orientation, we manually inspected local synteny using the GBrowse function in FlyBase (St Pierre et al., 2014).

Analysis of gene structure conservation

We compared transcript models for the 35 putative lincRNAs with conserved transcriptional orientation to look for evidence of conservation of gene structure. We aligned both the *D. pseudoobscura* lincRNA transcript sequence and the homologous *D.*

melanogaster transcript to the *D. pseudoobscura* genome sequence (FlyBase r2) using Clustal-Omega v1.2.0 with RNA specified as the biomolecule (Sievers et al., 2011; St Pierre et al., 2014). This was sufficient to make alignments for most loci. For loci that spanned large genomic regions, we broke down the transcript models into their constituent exons to help facilitate better alignment. Transcript ends were considered to be conserved if within 40 nucleotides of each other. Introns were considered to be conserved if within 10 nucleotides of each other with a splice donor/acceptor site on either side.

***Correlation of developmental expression profiles between* D. pseudoobscura *and* D. melanogaster**

To generate developmental expression data from *D. melanogaster*, we used RNA-Seq datasets originally generated for the modENCODE project and available through the Sequence Read Archive (Graveley et al., 2011; Kodama et al., 2012; Leinonen et al., 2011). We chose datasets at developmental stages roughly equivalent to those we collected in *D. pseudoobscura*. These are mixed single-end and paired-end 75 bp Illumina sequence reads and include: 1st-instar larvae (7 datasets – SRR023597, SRR023646, SRR023661, SRR023666, SRR023706, SRR023835, and SRR035410); 3rd-instar larvae, light blue gut PS(3-6) (7 datasets – SRR023505, SRR023676, SRR023683, SRR023690, SRR023692, SRR023742, and SRR027108); pupae, 2 days after white prepupae (6 datasets – SRR023667, SRR023721, SRR023743, SRR023785, SRR023829, and SRR026431); 5-day adult males (10 datasets – SRR023605, SRR023606, SRR023642, SRR023658, SRR023672, SRR023679, SRR023713, SRR029176, SRR029231, SRR029233, an SRR029235); and 5-day adult females (8 datasets –

SRR023547, SRR023607, SRR023645, SRR023651, SRR023717, SRR023730, SRR029230, and SRR029234). Total library size ranged from 1,248,148 fragments to 50,925,810 fragments.

D. melanogaster sequence fragments were first quality filtered using the NGS QC Toolkit (Patel and Jain, 2012). Between 13.7% and 87.0% of fragments had PHRED > 20 and were retained for mapping. Sequence fragments were mapped to the *D. melanogaster* genome (FlyBase r6) with Bowtie2/TopHat2, and fragments were counted at *D. melanogaster* loci from the FlyBase r6.02 annotation using HTSeq-count v0.6.1p1 using the same parameters used for our *D. pseudoobscura* data as described in Chapters 1 and 2 (Anders et al., 2014; Kim et al., 2013; Langmead and Salzberg, 2012). Between 82.7% and 98.8% of high-quality fragments successfully mapped to the *D. melanogaster* genome.

Because the *D. melanogaster* data is not sex-specific in early developmental stages, we pooled our male and female datasets to create an approximation of the five stages available in *D. melanogaster*: 1st-instar larvae, 3rd-instar larvae, mid-pupae, adult male, and adult females. We generated $\log_2$(cpm) expression values for all annotated genes individually for both species (*D. pseudoobscura* r2.29 plus lincRNAs, *D. melanogaster* r6.02) using limma-voom as described in Chapter 2 (Law et al., 2014). A minimum cpm of 0.3 was required in at least three replicates for the locus to be retained for correlation analysis. Protein-coding orthologs between *D. pseudoobscura* and *D. melanogaster* were identified using OrthoDB data from FlyBase (Kriventseva et al., 2008; St Pierre et al., 2014; Waterhouse et al., 2013). Pearson correlation coefficients

between 134 putative lincRNA homologs and 7,451 orthologous protein-coding genes were generated in R (Team, 2014).

Identifying transposable element (TE) insertion stocks in exonic regions of putative lincRNA homologs

We screened the following transposable element (TE) insertion stocks for lethality (all Bloomington stock numbers): 10240, 10154, 11677, 10155, 33239, 11637, 36157, and 11872. All stocks are available from the Bloomington *Drosophila* Stock Center at Indiana University. Lethality was confirmed by screening for lack of homozygotes in the stable TE insertion stock.

CONCLUSIONS: Towards a better understanding of lincRNA biology in *Drosophila*

In sum, we have identified and documented 1,586 novel lincRNA loci in *D. pseudoobscura*. This is only the second large-scale lincRNA annotation effort in *Drosophila* and one of still only very few to be undertaken in a non-vertebrate eukaryote (Kapusta and Feschotte, 2014; Ulitsky and Bartel, 2013). We have shown that a large number of these lincRNAs have male-biased expression throughout development and that this male-bias can be largely, but not exclusively, attributed to expression in the testes. We have also shown that very few are expressed in the ovaries, even when factoring in the overall lower gene expression content of the ovaries compared to other tissues. Many lincRNAs also have unbiased developmental regulation, with the largest numbers being expressed most highly in the pupal stage. We examined the genomic distributions of lincRNAs on the X chromosome and autosomes and found, like protein-coding genes, an underrepresentation of non-testis-specific male-biased genes and an overrepresentation of female-biased genes on the X chromosome. The major models that have been put forward to explain this observation all invoke selection, and as the trends in the lincRNAs mirror those of the protein-coding genes, we interpret these distributions as evidence of biological relevance. Lastly, we integrated sequence and transcript features from annotated sets of lincRNAs in *D. pseudoobscura* and *D. melanogaster* in order to identify potentially homologous lincRNAs that are suitable candidates for future functional assays.

Uncovering the full extent of biological relevance is still the greatest challenge in lincRNA biology, and we have just begun to scratch the surface in *Drosophila*. We have

presented two observations that suggest biological relevance in at least some lincRNAs: (1) the observation of unequal distributions of sex-biased lincRNAs between the X and the autosomes, and (2) the documentation of dozens of putative lincRNAs with conserved sequence and transcript features. On the other hand, our data also suggests that the large set of testis-specific lincRNAs shows no evidence of proximate biological relevance.

This dataset also has value in opening doors to analyses that are more direct indicators of biological function. We detail three approaches to exploring lincRNA biology that we will be pursuing.

(1) Natural variation within and between species can be used to look for evidence of natural selection. Evidence of purifying selection has already been found in *D. melanogaster* lincRNA sequence using population-level data (Haerty and Ponting, 2013). However, there has not been a single study, in *Drosophila* or otherwise, that tests for evidence of selection by using expression polymorphism and divergence data. We have already generated RNA-Seq data from adult gonads and carcasses for six additional inbred populations of *D. pseudoobscura* and two inbred populations of its sympatric sister species *D. persimilis*. Together with genome sequence from each of these lines and the set of lincRNAs annotated in this doctoral dissertation, we will perform the most thorough analyses of lincRNA evolution yet.

(2) The Machado Lab's interest in lincRNA biology blossomed from observations of expression divergence between a handful of male-biased lincRNAs in *D. pseudoobscura* and *D. persimilis* (Jiang et al., 2011). Because reproductive isolation between *D. pseudoobscura* and *D. persimilis* stems largely from hybrid

male sterility, finding expression divergence in the testes is always enticing. We plan to examine mode of inheritance and regulatory divergence in the testes of *D. pseudoobscura*/*D. persimilis* hybrids, paying particular attention to the biological implications of testis-specificity as suggested by this current work. We have already collected hybrid samples and have received funding for this project from an NSF Doctoral Dissertation Improvement Grant. It stands to be the first high-throughput analysis of lincRNA expression in hybrids of any species.

(3) Finally, we are interested not only in what biological processes lincRNAs are involved in, but also how they function and how that function constrains their evolution. To do so, we need to identify conserved lincRNAs in multiple genetically-amenable species. Using natural and induced variation, we will be able to tease apart what features are critical for the core function of the lincRNA. The 22 high-confidence lincRNA homologs that we have identified between *D. pseudoobscura* and *D. melanogaster* are a natural starting point for this type of project.

APPENDIX

Family	Loci
1	XLOC_000549, XLOC_001850, XLOC_002112, XLOC_003058, XLOC_003270, XLOC_004781, XLOC_005749, XLOC_005849, XLOC_005891, XLOC_013574, XLOC_015254, XLOC_015599, XLOC_017291, XLOC_018231
2	XLOC_001175/XLOC_002691, XLOC_002862, XLOC_006015, XLOC_014914, XLOC_015237
3	XLOC_001646, XLOC_008372
4	XLOC_001759, XLOC_003064
5	XLOC_002433, XLOC_013539
6	XLOC_002617, XLOC_017222
7	XLOC_002940, XLOC_003377
8	XLOC_001321, XLOC_003082, XLOC_003253, XLOC_003285, XLOC_003345, XLOC_003354, XLOC_003398, XLOC_003406, XLOC_004684, XLOC_005752, XLOC_005759, XLOC_005786, XLOC_005801, XLOC_005805, XLOC_005809, XLOC_005829, XLOC_005838, XLOC_005879, XLOC_005884, XLOC_005993, XLOC_007861, XLOC_007873, XLOC_007926, XLOC_008694, XLOC_008715, XLOC_013178, XLOC_013264, XLOC_013862, XLOC_013866, XLOC_013869, XLOC_013879, XLOC_013880, XLOC_013885, XLOC_013916, XLOC_013936, XLOC_015058, XLOC_015197, XLOC_015206, XLOC_015293, XLOC_015511, XLOC_015598, XLOC_017109, XLOC_017476, XLOC_017515, XLOC_017986, XLOC_017987, XLOC_018240, XLOC_018244
9	XLOC_003194, XLOC_013899
10	XLOC_003300, XLOC_013861
11	XLOC_003465, XLOC_005778, XLOC_013142
12	XLOC_005068, XLOC_013069, XLOC_013793
13	XLOC_006482, XLOC_006622
14	XLOC_006786, XLOC_008709, XLOC_015186
15	XLOC_006984, XLOC_007846, XLOC_007853, XLOC_007876, XLOC_008398, XLOC_017272
16	XLOC_007065/XLOC_007599, XLOC_007600, XLOC_007602
17	XLOC_007989, XLOC_014630
18	XLOC_008650, XLOC_013918
19	XLOC_005795, XLOC_013080
20	XLOC_013094, XLOC_013502, XLOC_013509
21	XLOC_014539, XLOC_014910
22	XLOC_015281, XLOC_015282
23	XLOC_016383/XLOC_016979, XLOC_017250
24	XLOC_017587/XLOC_018043, XLOC_018042
25	XLOC_007912, XLOC_017214, XLOC_018215

Table 1 – Multi-locus lincRNA families in *D. pseudoobscura*. Shown are the 25 multi-locus lincRNA families in *D. pseudoobscura*. Several lincRNA loci are unable to be resolved because they overlap on opposite strands.

Cluster 1 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
neurogenesis (BP)	41	282	556	14869	6.77E-14	2.50E-12
rRNA processing (BP)	6	282	25	14869	5.77E-06	1.07E-04
mitotic spindle organization (BP), translation (BP), mitotic spindle elongation (BP)	7	282	47	14869	2.69E-05	3.32E-04
mRNA processing (BP)	4	282	19	14869	3.92E-04	3.63E-03
mitotic spindle organization (BP), protein folding (BP)	3	282	9	14869	5.21E-04	3.85E-03
mitotic spindle organization (BP)	11	282	190	14869	1.02E-03	5.39E-03
negative regulation of JNK cascade (BP)	3	282	12	14869	1.31E-03	6.05E-03
regulation of circadian sleep/wake cycle, sleep (BP)	3	282	11	14869	9.95E-04	6.13E-03
cytoskeletal anchoring at plasma membrane (BP)	3	282	17	14869	3.77E-03	0.014
translation (BP)	19	282	505	14869	3.63E-03	0.015
protein folding (BP)	7	282	115	14869	6.30E-03	0.021
ribosome biogenesis (BP)	3	282	23	14869	9.03E-03	0.028

Table 2 – Cluster 1 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 1 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 1. "Total cluster" refers to the total number of loci in cluster 2 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 2 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
spermatogenesis (BP)	4	147	88	14869	0.011	0.034
sensory perception of smell (BP)	4	147	82	14869	8.84E-03	0.040

Table 3 – Cluster 2 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 2 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 2. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 3 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
ion transport (BP)	11	317	51	14869	7.75E-09	4.65E-07
sensory perception of chemical stimulus (BP)	10	317	71	14869	2.46E-06	7.37E-05
maintenance of presynaptic active zone structure (BP)	3	317	5	14869	9.30E-05	1.12E-03
axon guidance (BP), axon midline choice point recognition (BP)	4	317	12	14869	8.77E-05	1.31E-03
G-protein coupled receptor signaling pathway (BP)	11	317	126	14869	7.99E-05	1.60E-03
anesthesia-resistant memory (BP)	3	317	8	14869	4.96E-04	4.96E-03
small GTPase mediated signal transduction (BP), GTP catabolic process (BP)	4	317	20	14869	7.50E-04	5.00E-03
regulation of transcription, DNA-dependent (BP)	20	317	434	14869	1.06E-03	5.28E-03
lateral inhibition (BP), peripheral nervous system development (BP)	3	317	9	14869	7.33E-04	5.50E-03
mesoderm migration involved in gastrulation (BP)	3	317	9	14869	7.33E-04	5.50E-03
lateral inhibition (BP), axon guidance (BP)	3	317	11	14869	1.39E-03	5.58E-03
fibroblast growth factor receptor signaling pathway (BP)	3	317	11	14869	1.39E-03	5.58E-03
GTP catabolic process (BP)	5	317	36	14869	9.35E-04	5.61E-03
axon guidance (BP)	10	317	144	14869	1.04E-03	5.68E-03
motor axon guidance (BP)	5	317	39	14869	1.36E-03	6.26E-03
response to mechanical stimulus (BP)	3	317	12	14869	1.83E-03	6.86E-03
open tracheal system development (BP), tracheal outgrowth, open tracheal system (BP)	3	317	13	14869	2.34E-03	7.03E-03
ion transport (BP), transmembrane transport (BP)	3	317	13	14869	2.34E-03	7.03E-03
R7 cell fate commitment (BP)	3	317	13	14869	2.34E-03	7.03E-03
genital disc development (BP)	3	317	13	14869	2.34E-03	7.03E-03
mesoderm development (BP), heart development (BP)	3	317	15	14869	3.61E-03	0.010
regulation of transcription, DNA-dependent (BP), brain development (BP)	3	317	16	14869	4.37E-03	0.011

Cluster 3 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
imaginal disc-derived wing morphogenesis (BP), imaginal disc-derived wing vein specification (BP)	3	317	16	14869	4.37E-03	0.011
nervous system development (BP)	7	317	98	14869	4.95E-03	0.012
cytoskeletal anchoring at plasma membrane (BP)	3	317	17	14869	5.23E-03	0.013
peripheral nervous system development (BP)	6	317	78	14869	6.36E-03	0.014
epithelial cell migration, open tracheal system (BP), open tracheal system development (BP)	3	317	18	14869	6.18E-03	0.014
neuropeptide signaling pathway (BP)	4	317	38	14869	8.46E-03	0.018
small GTPase mediated signal transduction (BP)	6	317	84	14869	9.06E-03	0.019
synaptic vesicle exocytosis (BP)	3	317	21	14869	9.60E-03	0.019

Table 4 – Cluster 3 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 6 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 3. "Total cluster" refers to the total number of loci in cluster 3 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value. Table is truncated after the top 30 hits.

Cluster 4 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
homophilic cell adhesion (BP)	7	246	29	14869	3.56E-07	1.18E-05
cell adhesion (BP), homophilic cell adhesion (BP)	4	246	6	14869	1.07E-06	1.76E-05
homophilic cell adhesion (BP), cell-cell adhesion (BP)	3	246	4	14869	1.77E-05	1.46E-04
regulation of transcription, DNA-dependent (BP), steroid hormone mediated signaling pathway (BP)	5	246	2	14869	1.51E-05	1.66E-04
homophilic cell adhesion (BP), calcium-dependent cell-cell adhesion (BP), calcium-dependent cell-cell adhesion (BP), ommatidial rotation (BP)	3	246	6	14869	8.62E-05	5.69E-04
cell adhesion (BP)	9	246	118	14869	1.50E-04	7.06E-04
homophilic cell adhesion (BP), calcium-dependent cell-cell adhesion (BP)	4	246	17	14869	1.47E-04	8.08E-04
compound eye morphogenesis (BP)	6	246	78	14869	1.83E-03	6.73E-03
one-carbon metabolic process (BP)	3	246	15	14869	1.76E-03	7.25E-03
cell redox homeostasis (BP)	4	246	50	14869	9.28E-03	0.026
signal transduction (BP), defense response (BP)	3	246	26	14869	8.78E-03	0.026
wing disc dorsal/ventral pattern formation (BP)	4	246	48	14869	8.05E-03	0.027
myoblast fusion (BP)	3	246	28	14869	0.011	0.027
gonad development (BP)	3	246	31	14869	0.014	0.034

Table 5 – Cluster 4 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 4 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 4. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 5 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
chitin-based cuticle development (BP), body morphogenesis (BP)	9	363	22	14869	1.05E-09	3.15E-08
oxidation-reduction process (BP)	29	363	421	14869	5.16E-07	7.74E-06
transmembrane transport (BP)	23	363	293	14869	8.82E-07	8.82E-06
cellular amino acid metabolic process (BP)	4	363	19	14869	1.01E-03	7.59E-03
dephosphorylation (BP)	4	363	27	14869	3.93E-03	0.024

Table 6 – Cluster 5 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 5 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 5. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 6 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
multicellular organism reproduction (BP), sperm competition (BP)	3	94	7	14869	8.41E-06	5.89E-05
oxidation-reduction process (BP)	11	94	421	14869	6.88E-05	1.60E-04
detection of pheromone (BP)	3	94	12		5.17E-05	1.81E-04
multicellular organism reproduction (BP)	5	94	75	14869	1.10E-04	1.93E-04
sensory perception of smell (BP)	5	94	82	14869	1.69E-04	2.36E-04
sensory perception of chemical stimulus (BP)	4	94	71	14869	1.05E-03	1.23E-03
proteolysis (BP)	8	94	548	14869	0.023	0.023

Table 7 – Cluster 6 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 6 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 6. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 7 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
mitotic spindle organization (BP)	20	277	190	14869	4.11E-10	3.66E-08
mitosis (BP)	14	277	99	14869	4.03E-09	1.79E-07
oogenesis (BP)	18	277	200	14869	3.75E-08	1.11E-06
mitotic spindle organization (BP), protein localization (BP)	4	277	4	14869	1.18E-07	2.62E-06
mitotic spindle organization (BP), mitosis (BP)	8	277	36	14869	2.52E-07	4.49E-06
mitotic spindle organization (BP), microtubule cytoskeleton organization (BP)	4	277	5	14869	5.81E-07	8.62E-06
microtubule cytoskeleton organization (BP)	7	277	29	14869	7.93E-07	1.01E-05
mitotic spindle organization (BP), cytokinesis (BP)	5	277	11	14869	9.12E-07	1.01E-05
pronuclear fusion (BP)	4	277	6	14869	1.72E-06	1.70E-05
mitotic spindle organization (BP), neurogenesis (BP)	8	277	47	14869	2.20E-06	1.95E-05
female meiosis (BP)	6	277	24	14869	4.02E-06	3.25E-05
mitotic spindle organization (BP), mitosis (BP), protein localization (BP)	3	277	3	14869	6.40E-06	3.35E-05
mitotic spindle organization (BP), cytokinesis (BP), protein localization (BP)	3	277	3	14869	6.40E-06	3.35E-05
mitotic spindle organization (BP), pronuclear fusion (BP), pronuclear migration (BP)	3	277	3	14869	6.40E-06	3.35E-05
oogenesis (BP), negative regulation of oskar mRNA translation (BP)	3	277	3	14869	6.40E-06	3.35E-05
female meiosis (BP), mitosis (BP)pronuclear fusion (BP)	3	277	3	14869	6.40E-06	3.35E-05
negative regulation of oskar mRNA translation (BP)	4	277	8	14869	7.78E-06	3.85E-05
neurogenesis (BP), cytokinesis (BP)	5	277	15	14869	5.58E-06	4.14E-05
mitotic spindle organization (BP), cell cycle (BP), mitosis (BP)	3	277	4	14869	2.52E-05	1.02E-04

Cluster 7 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
circadian rhythm (BP), rhythmic behavior (BP), locomotor rhythm (BP), negative regulation of transcription from RNA polymerase II promoter (BP), regulation of circadian sleep/wake cycle sleep (BP)eclosion rhythm (BP)	3	277	4	14869	2.52E-05	1.02E-04
mitosis (BP), microtubule cytoskeleton organization (BP)	3	277	4	14869	2.52E-05	1.02E-04
spindle assembly involved in female meiosis (BP)	3	277	4	14869	2.52E-05	1.02E-04
cytokinesis (BP)	8	277	65	14869	2.64E-05	1.02E-04
protein localization (BP)	7	277	50	14869	3.63E-05	1.29E-04
mitotic cell cycle spindle assembly checkpoint (BP)	4	277	11	14869	3.51E-05	1.30E-04
microtubule-based movement (BP)	9	277	94	14869	6.33E-05	2.09E-04
oogenesis (BP), microtubule cytoskeleton organization (BP)	3	277	5	14869	6.22E-05	2.13E-04
mitotic spindle organization (BP), neurogenesis (BP), cytokinesis (BP)	3	277	6	14869	1.23E-04	3.64E-04
mitotic spindle organization (BP), mitosis (BP), microtubule-based movement (BP)	3	277	6	14869	1.23E-04	3.64E-04
neurogenesis (BP)	24	277	556	14869	1.19E-04	3.78E-04

Table 8 – Cluster 7 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 7 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 7. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value. Table is truncated after the top 30 hits.

Cluster 8 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
defense response (BP)	8	228	103	14869	1.84E-04	5.89E-03
response to bacterium (BP)	3	228	20	14869	3.35E-03	0.036
defense response (BP), Toll signaling pathway (BP)	3	228	19	14869	2.88E-03	0.046

Table 9 – Cluster 8 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 8 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 8. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 9 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
DNA-dependent DNA replication initiation (BP)	6	290	13	14869	8.00E-08	4.40E-06
pre-replicative complex assembly (BP), DNA-dependent DNA replication initiation (BP)	4	290	5	14869	6.98E-07	1.92E-05
oogenesis (BP), mRNA transport (BP)	3	290	3	14869	7.34E-06	1.01E-04
mitotic cell cycle G2/M transition DNA damage checkpoint (BP), DNA-dependent DNA replication initiation (BP)	3	290	3	14869	7.34E-06	1.01E-04
chromosome condensation (BP), DNA-dependent DNA replication initiation (BP)	3	290	4	14869	2.90E-05	3.18E-04
neurogenesis (BP)	25	290	556	14869	9.37E-05	6.44E-04
neurogenesis (BP), DNA-dependent DNA replication initiation (BP)	3	290	5	14869	7.13E-05	6.54E-04
vitelline membrane formation involved in chorion-containing eggshell formation (BP)	4	290	13	14869	8.82E-05	6.93E-04
oogenesis (BP)	13	290	200	14869	1.54E-04	8.49E-04
neurogenesis (BP), translational initiation (BP)	3	290	6	14869	1.41E-04	8.59E-04
DNA replication (BP), DNA-dependent DNA replication initiation (BP)	3	290	8	14869	3.82E-04	1.50E-03
oogenesis (BP), oocyte microtubule cytoskeleton organization (BP)	3	290	8	14869	3.82E-04	1.50E-03
vitellogenesis (BP)	3	290	8	14869	3.82E-04	1.50E-03
mitotic cell cycle G2/M transition DNA damage checkpoint (BP)	7	290	67	14869	3.19E-04	1.60E-03
telomere capping (BP)	3	290	10	14869	7.96E-04	2.92E-03
germ cell development (BP)	4	290	29	14869	2.29E-03	7.00E-03
DNA replication (BP)	5	290	47	14869	2.14E-03	7.36E-03
oogenesis (BP), germ cell development (BP)	3	290	14	14869	2.28E-03	7.37E-03

Cluster 9 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
neurogenesis (BP), mitotic cell cycle G2/M transition DNA damage checkpoint (BP)	3	290	17	14869	4.08E-03	0.012
pole cell formation (BP)	3	290	20	14869	6.55E-03	0.018
flight behavior (BP)	3	290	21	14869	7.53E-03	0.020
protein phosphorylation (BP)	11	290	244	14869	8.55E-03	0.021
border follicle cell migration (BP)	5	290	70	14869	0.012	0.028
mitotic spindle organization (BP), neurogenesis (BP)	4	290	47	14869	0.013	0.030
protein dephosphorylation (BP)	5	290	81	14869	0.021	0.046
translational initiation (BP)	4	290	55	14869	0.022	0.047
chromosome segregation (BP)	3	290	32	14869	0.024	0.047
female meiosis chromosome segregation (BP)	3	290	32	14869	0.024	0.047
regulation of transcription DNA-dependent (BP), neurogenesis (BP)	3	290	34	14869	0.028	0.050
negative regulation of apoptotic process (BP)	3	290	34	14869	0.028	0.050

Table 10 – Cluster 9 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 9 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 9. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 10 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
dendrite morphogenesis (BP)	17	334	135	14869	9.20E-09	1.04E-06
muscle organ development (BP), dendrite morphogenesis (BP)	10	334	41	14869	1.73E-08	1.30E-06
compound eye development (BP)	15	334	102	14869	8.15E-09	1.83E-06
smoothened signaling pathway (BP)	8	334	33	14869	5.07E-07	1.14E-05
regulation of transcription DNA-dependent (BP), muscle organ development (BP), dendrite morphogenesis (BP)	7	334	23	14869	4.88E-07	1.22E-05
oogenesis (BP)	18	334	200	14869	6.15E-07	1.26E-05
dendrite morphogenesis (BP), neuron development (BP)	8	334	32	14869	3.92E-07	1.26E-05
neuron development (BP)	9	334	44	14869	4.64E-07	1.30E-05
establishment or maintenance of cell polarity (BP)	8	334	31	14869	3.00E-07	1.35E-05
imaginal disc-derived wing morphogenesis (BP)	14	334	117	14869	3.62E-07	1.36E-05
muscle organ development (BP)	11	334	67	14869	2.61E-07	1.47E-05
asymmetric cell division (BP)	7	334	25	14869	9.20E-07	1.72E-05
peripheral nervous system development (BP)	11	334	78	14869	1.27E-06	2.03E-05
asymmetric cell division (BP), sensory organ precursor cell fate determination (BP)	4	334	5	14869	1.23E-06	2.13E-05
nervous system development (BP), ovarian follicle cell development (BP)	4	334	6	14869	3.62E-06	5.43E-05
imaginal disc-derived wing morphogenesis (BP), compound eye development (BP)	6	334	21	14869	5.01E-06	7.05E-05
mesoderm development (BP), peripheral nervous system development (BP)	4	334	7	14869	8.30E-06	1.10E-04
imaginal disc-derived wing morphogenesis (BP), compound eye development (BP), nervous system development (BP), ovarian follicle cell development (BP)	3	334	3	14869	1.12E-05	1.20E-04
protein localization (BP)asymmetric cell division (BP)sensory organ precursor cell fate determination (BP)	3	334	3	14869	1.12E-05	1.20E-04

Cluster 10 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
regulation of proteolysis (BP), smoothened signaling pathway (BP)	3	334	3	14869	1.12E-05	1.20E-04
negative regulation of smoothened signaling pathway (BP), smoothened signaling pathway (BP)	3	334	3	14869	1.12E-05	1.20E-04
regulation of transcription DNA-dependent (BP), dendrite morphogenesis (BP)	8	334	50	14869	1.41E-05	1.38E-04
regulation of transcription DNA-dependent (BP), negative regulation of transcription DNA-dependent (BP), dendrite morphogenesis (BP)	4	334	8	14869	1.63E-05	1.41E-04
negative regulation of transcription DNA-dependent (BP), dendrite morphogenesis (BP), neuron development (BP)	4	334	8	14869	1.63E-05	1.41E-04
negative regulation of transcription DNA-dependent (BP), dendrite morphogenesis (BP)	5	334	15	14869	1.38E-05	1.42E-04
signal transduction (BP)	14	334	160	14869	1.52E-05	1.43E-04
imaginal disc-derived wing morphogenesis (BP), peripheral nervous system development (BP)	5	334	16	14869	1.98E-05	1.53E-04
transcription initiation from RNA polymerase II promoter (BP), transcription from RNA polymerase II promoter (BP)	5	334	16	14869	1.98E-05	1.53E-04
regulation of mitotic cell cycle (BP)	6	334	26	14869	1.93E-05	1.61E-04
ovarian follicle cell development (BP)	8	334	53	14869	2.19E-05	1.64E-04

Table 11 – Cluster 10 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 10 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 10. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value. Table is truncated after the top 30 hits.

Cluster 11 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
transmembrane transport (BP)	21	316	293	14869	1.25E-06	1.50E-05
proteolysis (BP)	31	316	548	14869	7.09E-07	1.70E-05
proteolysis (BP), digestion (BP)	3	316	3	14869	9.51E-06	7.61E-05
lipid metabolic process (BP)	10	316	87	14869	1.53E-05	9.20E-05
oxidation-reduction process (BP)	20	316	421	14869	7.00E-04	3.36E-03
neurotransmitter transport (BP)	4	316	24	14869	1.52E-03	6.08E-03
intracellular signal transduction (BP), cyclic nucleotide biosynthetic process (BP)	3	316	12	14869	1.81E-03	6.22E-03
amino acid transmembrane transport (BP)	4	316	31	14869	4.00E-03	0.012
antimicrobial humoral response (BP)	4	316	41	14869	0.011	0.029
metabolic process (BP)	11	316	249	14869	0.018	0.043

Table 12 – Cluster 11 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 11 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 11. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 12 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
regulation of transcription DNA-dependent (BP)	31	238	434	14869	2.66E-12	4.51E-10
compound eye morphogenesis (BP)	12	238	78	14869	3.66E-09	3.11E-07
ommatidial rotation (BP)	7	238	27	14869	1.67E-07	9.44E-06
regulation of transcription DNA-dependent (BP), imaginal disc-derived wing morphogenesis (BP), compound eye morphogenesis (BP)	4	238	5	14869	3.16E-07	1.07E-05
regulation of transcription from RNA polymerase II promoter (BP)	12	238	115	14869	3.09E-07	1.31E-05
regulation of transcription DNA-dependent (BP), compound eye morphogenesis (BP)	5	238	12	14869	7.28E-07	1.77E-05
neuroblast development (BP)	5	238	12	14869	7.28E-07	1.77E-05
equator specification (BP)	4	238	6	14869	9.36E-07	1.99E-05
open tracheal system development (BP)	11	238	107	14869	1.13E-06	2.13E-05
anterior head segmentation (BP)	4	238	7	14869	2.16E-06	3.67E-05
compound eye morphogenesis (BP), ovarian follicle cell development (BP), anterior/posterior axis specification embryo (BP)	3	238	3	14869	4.05E-06	5.30E-05
specification of segmental identity antennal segment (BP)	3	238	3	14869	4.05E-06	5.30E-05
positive regulation of transcription from RNA polymerase II promoter (BP)	8	238	58	14869	3.68E-06	5.69E-05
homophilic cell adhesion (BP)	6	238	29	14869	5.51E-06	6.24E-05
compound eye development (BP)	10	238	102	14869	5.37E-06	6.52E-05
regulation of transcription DNA-dependent (BP), neuroblast development (BP)	4	238	9	14869	7.57E-06	7.15E-05
spiracle morphogenesis open tracheal system (BP)	5	238	18	14869	7.28E-06	7.28E-05
sensory organ development (BP)	8	238	63	14869	6.95E-06	7.38E-05
regulation of transcription DNA-dependent (BP), neuroblast development (BP), heart development (BP)	3	238	4	14869	1.60E-05	1.09E-04

Cluster 12 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
spiracle morphogenesis open tracheal system (BP), ovarian follicle cell development (BP)	3	238	4	14869	1.60E-05	1.09E-04
spiracle morphogenesis open tracheal system (BP), anterior head segmentation (BP)	3	238	4	14869	1.60E-05	1.09E-04
compound eye morphogenesis (BP), equator specification (BP)	3	238	4	14869	1.60E-05	1.09E-04
compound eye development (BP), lymph gland crystal cell differentiation (BP)	3	238	4	14869	1.60E-05	1.09E-04
heart development (BP), epidermis development (BP)	3	238	4	14869	1.60E-05	1.09E-04
spiracle morphogenesis open tracheal system (BP), open tracheal system development (BP)	4	238	10	14869	1.25E-05	1.12E-04
open tracheal system development (BP,)regulation of transcription from RNA polymerase II promoter (BP)	4	238	11	14869	1.93E-05	1.13E-04
compound eye development (BP), negative regulation of transcription from RNA polymerase II promoter (BP)	4	238	11	14869	1.93E-05	1.13E-04
mesodermal cell fate specification (BP)	4	238	11	14869	1.93E-05	1.13E-04
imaginal disc-derived wing morphogenesis (BP)	10	238	117	14869	1.82E-05	1.19E-04
regulation of transcription DNA-dependent (BP,)imaginal disc-derived wing morphogenesis (BP)	5	238	22	14869	2.12E-05	1.20E-04

Table 13 – Cluster 12 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 12 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 12. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value. Table is truncated after the top 30 hits.

Cluster 13 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
chitin metabolic process (BP)	19	272	74	14869	4.46E-17	1.21E-15
neuropeptide signaling pathway (BP)	6	272	38	14869	5.99E-05	8.08E-04
G-protein coupled receptor signaling pathway (BP)	8	272	126	14869	2.21E-03	0.015
asymmetric neuroblast division (BP)	4	272	29	14869	1.81E-03	0.016
transmembrane transport (BP)	12	272	293	14869	7.85E-03	0.042
gastrulation (BP)	3	272	25	14869	0.010	0.047
cation transport (BP)	3	272	27	14869	0.013	0.049

Table 14 – Cluster 13 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 13 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 13. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 14 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
microtubule-based movement (BP)	13	546	94	14869	3.78E-05	1.02E-03
translational initiation (BP)	8	546	55	14869	8.35E-04	0.011
tricarboxylic acid cycle (BP)	6	546	41	14869	3.59E-03	0.032
sperm motility (BP)	3	546	10	14869	4.87E-03	0.033
'de novo' protein folding (BP)	3	546	11	14869	6.52E-03	0.035
proteolysis involved in cellular protein catabolic process (BP)	3	546	12	14869	8.45E-03	0.038

Table 15 – Cluster 14 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 14 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 14. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 15 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
neurotransmitter transport (BP)	4	154	24	14869	1.00E-04	2.20E-03
protein phosphorylation (BP)	10	154	244	14869	2.29E-04	2.52E-03
protein phosphorylation (BP), intracellular signal transduction (BP)	3	154	17	14869	6.66E-04	4.88E-03
G-protein coupled receptor signaling pathway (BP)	6	154	126	14869	1.99E-03	7.29E-03
protein phosphorylation (BP), regulation of cell shape (BP)	3	154	23	14869	1.66E-03	7.29E-03
intracellular signal transduction (BP)	5	154	81	14869	1.52E-03	8.36E-03
cilium assembly (BP)	3	154	29	14869	3.27E-03	8.99E-03
neurotransmitter secretion (BP)	5	154	95	14869	3.06E-03	9.62E-03
olfactory behavior (BP)	3	154	40	14869	8.13E-03	0.016
cell adhesion (BP)	5	154	118	14869	7.65E-03	0.017
imaginal disc-derived leg morphogenesis (BP)	3	154	39	14869	7.58E-03	0.019
microtubule-based movement (BP)	4	154	94	14869	0.016	0.028
ion transport (BP)	3	154	51	14869	0.016	0.029
lateral inhibition (BP)	6	154	212	14869	0.023	0.036
salivary gland cell autophagic cell death (BP), autophagic cell death (BP)	3	154	61	14869	0.025	0.037

Table 16 – Cluster 15 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 15 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 15. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

Cluster 16 GO terms	GO cluster	Total cluster	GO reference	Total Reference	Hyp p	Hyp p*
chitin-based cuticle development (BP), body morphogenesis (BP)	9	363	22	14869	1.05E-09	3.15E-08
oxidation-reduction process (BP)	29	363	421	14869	5.16E-07	7.74E-06
transmembrane transport (BP)	23	363	293	14869	8.82E-07	8.82E-06
cellular amino acid metabolic process (BP)	4	363	19	14869	1.01E-03	7.59E-03
dephosphorylation (BP)	4	363	27	14869	3.93E-03	0.024

Table 17 – Cluster 16 Biological Process GO terms. Biological Process GO terms shown here are significantly overrepresented in the developmental expression cluster 16 (hypergeometric test with FDR correction, $p < 0.05$). "GO cluster" refers to the number of GO-annotated loci in cluster 16. "Total cluster" refers to the total number of loci in cluster 4 for which any GO term could be found. "GO reference" refers to the number of GO-annotated loci in the Ensembl reference, and "Total reference" is the total size of the Ensembl reference. "Hyp p" is the hypergeometric test p-value, and "Hyp p*" is the FDR-corrected hypergeometric test p-value.

REFERENCES

Abdilleh, K.A. (2014). Patterns of sex-biased gene expression and gene pathway evolution in Drosophila (drum.lib.umd.edu).

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Akbari, O.S., Antoshechkin, I., Amrhein, H., Williams, B., Diloreto, R., Sandler, J., and Hay, B.A. (2013). The developmental transcriptome of the mosquito Aedes aegypti, an invasive species and major arbovirus vector. G3 *3*, 1493-1509.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - a Python framework to work with high-throughput sequencing data (BioRxiv).

Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Ayala, F.J., and Dobzhansky, T. (1974). New Subspecies of Drosophila-Pseudoobscura (Diptera-Drosophilidae). Pan-Pac Entomol *50*, 211-219.

Bachtrog, D., Toda, N.R., and Lockton, S. (2010). Dosage compensation and demasculinization of X chromosomes in Drosophila. Current biology : CB *20*, 1476-1481.

Bate, M., and Martinez Arias, A. (1993). The Development of Drosophila melanogaster (Plainview, N.Y.: Cold Spring Harbor Laboratory Press).

Beckenbach, A.T., Wei, Y.W., and Liu, H. (1993). Relationships in the Drosophila obscura species group, inferred from mitochondrial cytochrome oxidase II sequences. Molecular biology and evolution *10*, 619-634.

Bellen, H.J., Levis, R.W., Liao, G., He, Y., Carlson, J.W., Tsang, G., Evans-Holm, M., Hiesinger, P.R., Schulze, K.L., Rubin, G.M.*, et al.* (2004). The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. Genetics *167*, 761-781.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., 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.

Betran, E., Thornton, K., and Long, M. (2002). Retroposed new genes out of the X in Drosophila. Genome research *12*, 1854-1859.

Bhartiya, D., Jalali, S., Ghosh, S., and Scaria, V. (2014). Distinct patterns of genetic variations in potential functional elements in long noncoding RNAs. Human mutation *35*, 192-201.

Billerey, C., Boussaha, M., Esquerre, D., Rebours, E., Djari, A., Meersseman, C., Klopp, C., Gautheret, D., and Rocha, D. (2014). Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing. BMC genomics *15*, 499.

Boerner, S., and McGinnis, K.M. (2012). Computational identification and functional predictions of long noncoding RNA in Zea mays. PloS one *7*, e43047.
Brenner, S. (1998). Refuge of spandrels. Current biology : CB *8*, R669.
Broadbent, K.M., Park, D., Wolf, A.R., Van Tyne, D., Sims, J.S., Ribacke, U., Volkman, S., Duraisingh, M., Wirth, D., Sabeti, P.C.*, et al.* (2011). A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. Genome biology *12*, R56.
Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P., Smith, S., McCabe, V.M., Norris, D.P., Penny, G.D., Patel, D., and Rastan, S. (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. Nature *351*, 329-331.
Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515-526.
Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M.*, et al.* (2014). Diversity and dynamics of the Drosophila transcriptome. Nature *512*, 393-399.
Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development *25*, 1915-1927.
Calzone, F.J., Lee, J.J., Le, N., Britten, R.J., and Davidson, E.H. (1988). A long, nontranslatable poly(A) RNA stored in the egg of the sea urchin Strongylocentrotus purpuratus. Genes & development *2*, 305-318.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC bioinformatics *10*, 421.
Caparros, M.L., Alexiou, M., Webster, Z., and Brockdorff, N. (2002). Functional analysis of the highly conserved exon IV of XIST RNA. Cytogenetic and genome research *99*, 99-105.
Carroll, S.B. (2005). Evolution at two levels: on genes and form. PLoS biology *3*, e245.
Carvalho, A.B., and Clark, A.G. (2005). Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. Science *307*, 108-110.
Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M.*, et al.* (2009). Unlocking the secrets of the genome. Nature *459*, 927-930.
Charlesworth, B., Coyne, J.A., and Barton, N. (1987). The relative rates of evolution of sex chromosomes and autosomes. American Naturalist *130*, 113-146.
Chodroff, R.A., Goodstadt, L., Sirey, T.M., Oliver, P.L., Davies, K.E., Green, E.D., Molnar, Z., and Ponting, C.P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. Genome biology *11*, R72.
Clark, M.B., Johnston, R.L., Inostroza-Ponta, M., Fox, A.H., Fortini, E., Moscato, P., Dinger, M.E., and Mattick, J.S. (2012). Genome-wide analysis of long noncoding RNA stability. Genome research *22*, 885-898.

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

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

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2014). Ensembl 2015. Nucleic acids research.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research *22*, 1775-1789.

Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N., and Aebersold, R. (2006). The PeptideAtlas project. Nucleic acids research *34*, D655-658.

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. Trends in genetics : TIG *30*, 121-123.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F.*, et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Dobzhansky, T. (1936). Studies on Hybrid Sterility. II. Localization of Sterility Factors in Drosophila Pseudoobscura Hybrids. Genetics *21*, 113-135.

Dobzhansky, T. (1937). Genetics and the origin of species (New York,: Columbia Univ. Press).

Dobzhansky, T., Hunter, A.S., Pavlovsky, O., Spassky, B., and Wallace, B. (1963). Genetics of natural populations. XXXI. Genetics of an isolated marginal population of Drosophila pseudoobscura. Genetics *48*, 91-103.

Drosophila 12 Genomes, C., Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W.*, et al.* (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203-218.

Eddy, S.R. (2001). Non-coding RNA genes and the modern RNA world. Nature reviews Genetics *2*, 919-929.

Erdmann, V.A., Szymanski, M., Hochberg, A., Groot, N., and Barciszewski, J. (2000). Non-coding, mRNA-like RNAs database Y2K. Nucleic acids research *28*, 197-200.

Fickett, J.W. (1982). Recognition of protein coding regions in DNA sequences. Nucleic acids research *10*, 5303-5318.

Fickett, J.W., and Tung, C.S. (1992). Assessment of protein coding measures. Nucleic acids research *20*, 6441-6450.

Franke, A., and Baker, B.S. (1999). The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. Molecular cell *4*, 117-122.

Fullwood, M.J., Wei, C.L., Liu, E.T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome research *19*, 521-532.

Futschik, M.E., and Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. Journal of bioinformatics and computational biology *3*, 965-988.

Gao, G., Vibranovski, M.D., Zhang, L., Li, Z., Liu, M., Zhang, Y.E., Li, X., Zhang, W., Fan, Q., VanKuren, N.W., *et al.* (2014). A long-term demasculinization of X-linked intergenic noncoding RNAs in Drosophila melanogaster. Genome research *24*, 629-638.

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

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

Greenspan, R.J. (2004). Fly pushing : the theory and practice of Drosophila genetics, 2nd edn (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press).

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

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., *et al.* (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295-300.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., *et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology *28*, 503-510.

Haerty, W., Jagadeeshan, S., Kulathinal, R.J., Wong, A., Ravi Ram, K., Sirot, L.K., Levesque, L., Artieri, C.G., Wolfner, M.F., Civetta, A., *et al.* (2007). Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics *177*, 1321-1335.

Haerty, W., and Ponting, C.P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome biology *14*, R49.

Hamada, F.N., Park, P.J., Gordadze, P.R., and Kuroda, M.I. (2005). Global regulation of X chromosomal genes by the MSL complex in Drosophila melanogaster. Genes & development *19*, 2289-2294.

He, S., Liu, S., and Zhu, H. (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. BMC evolutionary biology *11*, 102.

He, Z., Bammann, H., Han, D., Xie, G., and Khaitovich, P. (2014). Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. Rna *20*, 1103-1111.

Hennig, W., and Weyrich, A. (2013). Histone modifications in the male germ line of Drosophila. BMC developmental biology *13*, 7.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., *et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic acids research *34*, D590-598.

Hoekstra, H.E., and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. Evolution *61*, 995-1016.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R.*, et al.* (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science *298*, 129-149.

Homolka, D., Ivanek, R., Forejt, J., and Jansa, P. (2011). Differential expression of non-coding RNAs and continuous evolution of the X chromosome in testicular transcriptome of two mouse species. PloS one *6*, e17198.

Honeybee Genome Sequencing, C. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. Nature *443*, 931-949.

Hoskins, R.A., Carlson, J.W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K.H., Park, S., Mendez-Lago, M., Rossi, F.*, et al.* (2007). Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science *316*, 1625-1628.

Hubisz, M.J., Pollard, K.S., and Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in bioinformatics *12*, 41-51.

Ilik, I.A., Quinn, J.J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R.C., Luscombe, N., Backofen, R.*, et al.* (2013). Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. Molecular cell *51*, 156-173.

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 to cells : devoted to molecular & cellular mechanisms *10*, 1163-1173.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218-223.

Iovino, N. (2014). *Drosophila* epigenome reorganization during oocyte differentiation and early embryogenesis. Briefings in Functional Genomics *13*, 246-253.

Ivanovska, I., and Orr-Weaver, T.L. (2006). Histone Modifications and the Chromatin Scaffold for Meiotic Chromosome Architecture. Cell Cycle *5*, 2064-2071.

Jagadeeshan, S., and Singh, R.S. (2005). Rapidly evolving genes of Drosophila: differing levels of selective pressure in testis, ovary, and head tissues between sibling species. Molecular biology and evolution *22*, 1793-1801.

Jenkins, A.M., Waterhouse, R.M., Kopin, A.S., and Muskavitch, M.A.T. (2014). Long non-coding RNA discovery in *Anopheles gambiae* using deep RNA sequencing (BioRxiv).

Jiang, Z.F., Croshaw, D.A., Wang, Y., Hey, J., and Machado, C.A. (2011). Enrichment of mRNA-like noncoding RNAs in the divergence of Drosophila males. Molecular biology and evolution *28*, 1339-1348.

Jiang, Z.F., and Machado, C.A. (2009). Evolution of sex-dependent gene expression in three recently diverged species of Drosophila. Genetics *183*, 1175-1185.

Johnsson, P., Lipovich, L., Grander, D., and Morris, K.V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochimica et biophysica acta *1840*, 1063-1071.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome research *20*, 1313-1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nature reviews Genetics *10*, 19-31.

Kapusta, A., and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends in genetics : TIG *30*, 439-452.
Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS genetics *9*, e1003470.
Keppel, G. (1991). Design and analysis : a researcher's handbook, 3rd edn (Englewood Cliffs, N.J.: Prentice Hall).
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.
Kimmins, S., and Sassone-Corsi, P. (2005). Chromatin remodelling and epigenetic features of germ cells. Nature *434*, 583-589.
King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.
Kleene, K.C. (2005). Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Developmental biology *277*, 16-26.
Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database, C. (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic acids research *40*, D54-56.
Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic acids research *35*, W345-349.
Kriventseva, E.V., Rahman, N., Espinosa, O., and Zdobnov, E.M. (2008). OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic acids research *36*, D271-275.
Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., Thakkapallayil, A., Sugnet, C.W., Stanke, M., Smith, K.E., Siepel, A.*, et al.* (2007). The UCSC genome browser database: update 2007. Nucleic acids research *35*, D668-673.
Kulathinal, R.J., Stevison, L.S., and Noor, M.A. (2009). The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS genetics *5*, e1000550.
Kumar, L., and Futschik, M.E. (2007). Mfuzz: a software package for soft clustering of microarray data. Bioinformation *2*, 5-7.
Kung, J.T., Colognori, D., and Lee, J.T. (2013). Long noncoding RNAs: past, present, and future. Genetics *193*, 651-669.
Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS genetics *8*, e1002841.
Lakhotia, S.C., Rajendra, T.K., and Prasanth, K.V. (2001). Developmental regulation and complex organization of the promoter of the non-coding hsr(omega) gene of Drosophila melanogaster. Journal of biosciences *26*, 25-38.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Latos, P.A., Pauler, F.M., Koerner, M.V., Senergin, H.B., Hudson, Q.J., Stocsits, R.R., Allhoff, W., Stricker, S.H., Klement, R.M., Warczok, K.E*., et al.* (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science *338*, 1469-1472.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology *15*, R29.

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011). The sequence read archive. Nucleic acids research *39*, D19-21.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, J., Wu, B., Xu, J., and Liu, C. (2014a). Genome-wide identification and characterization of long intergenic non-coding RNAs in Ganoderma lucidum. PloS one *9*, e99442.

Li, L., Eichten, S.R., Shimizu, R., Petsch, K., Yeh, C.T., Wu, W., Chettoor, A.M., Givan, S.A., Cole, R.A., Fowler, J.E*., et al.* (2014b). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome biology *15*, R40.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A*., et al.* (2013). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. Cell reports *5*, 3-12.

Li, M., and Liu, L. (2014). Neural functions of long noncoding RNAs in Drosophila. Journal of comparative physiology A, Neuroethology, sensory, neural, and behavioral physiology.

Li, M., Wen, S., Guo, X., Bai, B., Gong, Z., Liu, X., Wang, Y., Zhou, Y., Chen, X., Liu, L*., et al.* (2012). The novel long non-coding RNA CRG regulates Drosophila locomotor behavior. Nucleic acids research *40*, 11714-11727.

Li, W.H., Wu, C.I., and Luo, C.C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular biology and evolution *2*, 150-174.

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

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. The Plant cell *24*, 4333-4345.

Lu, Z.J., Yip, K.Y., Wang, G., Shou, C., Hillier, L.W., Khurana, E., Agarwal, A., Auerbach, R., Rozowsky, J., Cheng, C*., et al.* (2011). Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome research *21*, 276-285.

Lucas, A. (2014). amap: Another Multidimensional Analysis Package.

Machado, C.A., Haselkorn, T.S., and Noor, M.A. (2007). Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in Drosophila pseudoobscura and D. persimilis. Genetics *175*, 1289-1306.

Mali, P., Esvelt, K.M., and Church, G.M. (2013). Cas9 as a versatile tool for engineering biology. Nature methods *10*, 957-963.

Marin, I., Siegal, M.L., and Baker, B.S. (2000). The evolution of dosage-compensation mechanisms. BioEssays : news and reviews in molecular, cellular and developmental biology *22*, 1106-1114.
Marques, A.C., and Ponting, C.P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome biology *10*, R124.
Matsumura, H., Yoshida, K., Luo, S., Kimura, E., Fujibe, T., Albertyn, Z., Barrero, R.A., Kruger, D.H., Kahl, G., Schroth, G.P.*, et al.* (2010). High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. PloS one *5*, e12010.
Mattick, J.S. (2009). Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. Annals of the New York Academy of Sciences *1178*, 29-46.
McColl, G., and McKechnie, S.W. (1999). The Drosophila heat shock hsr-omega gene: an allele frequency cline detected by quantitative PCR. Molecular biology and evolution *16*, 1568-1574.
McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652-654.
McGaugh, S.E., Heil, C.S., Manzano-Winkler, B., Loewe, L., Goldstein, S., Himmel, T.L., and Noor, M.A. (2012). Recombination modulates how selection affects linked sites in Drosophila. PLoS biology *10*, e1001422.
McGaugh, S.E., and Noor, M.A. (2012). Genomic impacts of chromosomal inversions in parapatric Drosophila species. Philosophical transactions of the Royal Society of London Series B, Biological sciences *367*, 422-429.
Meiklejohn, C.D., and Presgraves, D.C. (2012). Little evidence for demasculinization of the Drosophila X chromosome among genes expressed in the male germline. Genome biology and evolution *4*, 1007-1016.
Meisel, R.P., Malone, J.H., and Clark, A.G. (2012). Disentangling the relationship between sex-biased gene expression and X-linkage. Genome research *22*, 1255-1265.
Meller, V.H., Gordadze, P.R., Park, Y., Chu, X., Stuckenholz, C., Kelley, R.L., and Kuroda, M.I. (2000). Ordered assembly of roX RNAs into MSL complexes on the dosage-compensated X chromosome in Drosophila. Current biology : CB *10*, 136-143.
Michel, A.M., Andreev, D.E., and Baranov, P.V. (2014). Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. BMC bioinformatics *15*, 380.
Morgan, T.H. (1911). The Origin of Five Mutations in Eye Color in Drosophila and Their Modes of Inheritance. Science *33*, 534-537.
Moriyama, E.N., and Powell, J.R. (1997). Codon usage bias and tRNA abundance in Drosophila. Journal of molecular evolution *45*, 514-523.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.
Mount, S.M., and Nguyen, M.-C.L. (2005). blastn Parameters for noncoding queries.
Muller, H.J. (1940). Bearings of the *Drosophila* work on systematics. In The New Systematics, J. Huxley, ed. (Oxford (United Kingdom): Clarendon Press), pp. 185-268.

Mulvey, B.B., Olcese, U., Cabrera, J.R., and Horabin, J.I. (2014). An interactive network of long non-coding RNAs facilitates the Drosophila sex determination decision. Biochimica et biophysica acta *1839*, 773-784.
Nam, J.W., and Bartel, D.P. (2012). Long noncoding RNAs in C. elegans. Genome research *22*, 2529-2540.
Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J.*, et al.* (2014). Rfam 12.0: updates to the RNA families database. Nucleic acids research.
Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933-2935.
Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635-640.
Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N., and Zakian, S.M. (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 research *11*, 833-849.
Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes.
Niazi, F., and Valadkhan, S. (2012). Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. Rna *18*, 825-843.
Noor, M. (2012). Pseudobase: Genome Sequences of *Drosophila pseudoobscura* subgroup species.
Noor, M.A. (2005). Patterns of evolution of genes disrupted in expression in Drosophila species hybrids. Genetical research *85*, 119-125.
Noor, M.A., Garfield, D.A., Schaeffer, S.W., and Machado, C.A. (2007). Divergence between the Drosophila pseudoobscura and D. persimilis genome sequences in relation to chromosomal inversions. Genetics *177*, 1417-1428.
Noor, M.A., Grams, K.L., Bertucci, L.A., Almendarez, Y., Reiland, J., and Smith, K.R. (2001a). The genetics of reproductive isolation and the potential for gene exchange between Drosophila pseudoobscura and D. persimilis via backcross hybrid males. Evolution *55*, 512-521.
Noor, M.A., Grams, K.L., Bertucci, L.A., and Reiland, J. (2001b). Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences of the United States of America *98*, 12084-12088.
Numata, K., Kanai, A., Saito, R., Kondo, S., Adachi, J., Wilming, L.G., Hume, D.A., Hayashizaki, Y., Tomita, M., Group, R.G.*, et al.* (2003). Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome research *13*, 1301-1306.
Oliveros, J.C. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams.
Orr, H.A. (1989a). Genetics of Sterility in Hybrids between 2 Subspecies of Drosophila. Evolution *43*, 180-189.
Orr, H.A. (1989b). Localization of Genes Causing Postzygotic Isolation in 2 Hybridizations Involving Drosophila-Pseudoobscura. Heredity *63*, 231-237.

Orr, H.A., and Irving, S. (2001). Complex epistasis and the genetic basis of hybrid sterility in the Drosophila pseudoobscura Bogota-USA hybridization. Genetics *158*, 1089-1100.

Ortiz-Barrientos, D., Counterman, B.A., and Noor, M.A. (2007). Gene expression divergence and the origin of hybrid dysfunctions. Genetica *129*, 71-81.

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. Nature genetics *36*, 40-45.

Palmieri, N., Nolte, V., Suvorov, A., Kosiol, C., and Schlotterer, C. (2012). Evaluation of different reference based annotation strategies using RNA-Seq - a case study in Drososphila pseudoobscura. PloS one *7*, e46415.

Paralkar, V.R., Mishra, T., Luan, J., Yao, Y., Kossenkov, A.V., Anderson, S.M., Dunagin, M., Pimkin, M., Gore, M., Sun, D.*, et al.* (2014). Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. Blood *123*, 1927-1937.

Parisi, M., Nuttall, R., Edwards, P., Minor, J., Naiman, D., Lu, J., Doctolero, M., Vainer, M., Chan, C., Malley, J.*, et al.* (2004). A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. Genome biology *5*, R40.

Park, S.W., Kang, Y., Sypula, J.G., Choi, J., Oh, H., and Park, Y. (2007). An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the Drosophila X chromosome. Genetics *177*, 1429-1437.

Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS one *7*, e30619.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A.*, et al.* (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome research *22*, 577-591.

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

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

Powell, J.R., and Moriyama, E.N. (1997). Evolution of codon usage bias in Drosophila. Proceedings of the National Academy of Sciences of the United States of America *94*, 7784-7790.

Qu, Z., and Adelson, D.L. (2012). Identification and comparative analysis of ncRNAs in human, mouse and zebrafish indicate a conserved role in regulation of genes expressed in brain. PloS one *7*, e52275.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

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

Reiland, J., and Noor, M.A. (2002). Little qualitative RNA misexpression in sterile male F1 hybrids of Drosophila pseudoobscura and D. persimilis. BMC evolutionary biology *2*, 16.

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J., and Jones, C.D. (2013). De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS genetics *9*, e1003860.

Rice, W. (1984). Sex chromosomes and the evolution of sexual dimorphism. Evolution *38*, 735-742.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P.*, et al.* (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome research *15*, 1-18.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E.*, et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Rodriguez-Trelles, F., Tarrio, R., and Ayala, F.J. (2000). Evidence for a high ancestral GC content in Drosophila. Molecular biology and evolution *17*, 1710-1717.

Rymarquis, L.A., Kastenmayer, J.P., Huttenhofer, A.G., and Green, P.J. (2008). Diamonds in the rough: mRNA-like non-coding RNAs. Trends in plant science *13*, 329-334.

Schaeffer, S.W., Bhutkar, A., McAllister, B.F., Matsuda, M., Matzkin, L.M., O'Grady, P.M., Rohde, C., Valente, V.L., Aguade, M., Anderson, W.W.*, et al.* (2008). Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics *179*, 1601-1655.

Schaeffer, S.W., Machado, C.A., Anderson, W., Papaceit, M., Aguade, M., Segarra, C., and Noor, M.A. Cytogenetic Maps of the Six Muller's Elements in *Drosophila pseudoobscura*.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A.*, et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112-1115.

Schorderet, P., and Duboule, D. (2011). Structural and functional differences in the long non-coding RNA hotair in mouse and human. PLoS genetics *7*, e1002071.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940-3941.

Smyth, G.K. (2005). Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irazarry, and W. Huber, eds. (New York: Springer), pp. 397-420.

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC bioinformatics *14*, 91.
Soshnev, A.A., Ishimoto, H., McAllister, B.F., Li, X., Wehling, M.D., Kitamoto, T., and Geyer, P.K. (2011). A conserved long noncoding RNA affects sleep behavior in Drosophila. Genetics *189*, 455-468.
St Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P., and FlyBase, C. (2014). FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic acids research *42*, D780-788.
Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H.*, et al.* (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome research *12*, 1611-1618.
Stuckenholz, C., Meller, V.H., and Kuroda, M.I. (2003). Functional redundancy within roX1, a noncoding RNA involved in dosage compensation in Drosophila melanogaster. Genetics *164*, 1003-1014.
Sturgill, D., Zhang, Y., Parisi, M., and Oliver, B. (2007). Demasculinization of X chromosomes in the Drosophila genus. Nature *450*, 238-241.
Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012). GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic acids research *40*, W478-483.
Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nature protocols *7*, 542-561.
Tan, C.C. (1935). Salivary Gland Chromosomes in the Two Races of Drosophila Pseudoobscura. Genetics *20*, 392-402.
Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. Nature genetics *22*, 281-285.
Team, R.C. (2014). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.
Tribolium Genome Sequencing, C., Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W., Gibbs, R., Beeman, R.W., Brown, S.J.*, et al.* (2008). The genome of the model beetle and pest Tribolium castaneum. Nature *452*, 949-955.
Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689-693.
Tupy, J.L., Bailey, A.M., Dailey, G., Evans-Holm, M., Siebel, C.W., Misra, S., Celniker, S.E., and Rubin, G.M. (2005). Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America *102*, 5495-5500.
Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell *154*, 26-46.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537-1550.
Vibranovski, M.D., Lopes, H.F., Karr, T.L., and Long, M. (2009). Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS genetics *5*, e1000731.
Vicoso, B., and Charlesworth, B. (2009). Effective population size and faster-X effect: An extended model. Evolution *63*, 2413-2426.
Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Molecular cell *43*, 904-914.
Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research *41*, e74.
Ward, L.D., and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science *337*, 1675-1678.
Ward, L.D., and Kellis, M. (2013). Response to Comment on "Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions". Science *340*, 682-b.
Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Moeller, S., Schwartz, M., and Venables, B. (2014). gplots: Various R programming tools for plotting data.
Washietl, S., Findeiss, S., Muller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F., and Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. Rna *17*, 578-594.
Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome research *24*, 616-628.
Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M., and Kriventseva, E.V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic acids research *41*, D358-365.
Weikard, R., Hadlich, F., and Kuehn, C. (2013). Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing. BMC genomics *14*, 789.
Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., and Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic acids research *42*, D98-103.
Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., and Ponting, C.P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. Genome biology and evolution *4*, 427-442.
Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750-756.
Zhao, L., Saelao, P., Jones, C.D., and Begun, D.J. (2014). Origin and spread of de novo genes in Drosophila melanogaster populations. Science *343*, 769-772.