Comparative Expression Dynamics of Intergenic Long Noncoding RNAs in the Genus *Drosophila*

Kevin G. Nyberg[1] and Carlos A. Machado*,[1]

[1]Department of Biology, University of Maryland, College Park

*Corresponding author: E-mail: machado@umd.edu.

Accepted: May 6, 2016

Data deposition: All sequence data for this project have been deposited at NCBI's SRA database under accession PRJNA291085.

Abstract

Thousands of long noncoding RNAs (lncRNAs) have been annotated in eukaryotic genomes, but comparative transcriptomic approaches are necessary to understand their biological impact and evolution. To facilitate such comparative studies in *Drosophila*, we identified and characterized lncRNAs in a second *Drosophilid*—the evolutionary model *Drosophila pseudoobscura*. Using RNA-Seq and computational filtering of protein-coding potential, we identified 1,589 intergenic lncRNA loci in *D. pseudoobscura*. We surveyed multiple sex-specific developmental stages and found, like in *Drosophila melanogaster*, increasingly prolific lncRNA expression through male development and an overrepresentation of lncRNAs in the testes. Other trends seen in *D. melanogaster*, like reduced pupal expression, were not observed. Nonrandom distributions of female-biased and non-testis-specific male-biased lncRNAs between the X chromosome and autosomes are consistent with selection-based models of gene trafficking to optimize genomic location of sex-biased genes. The numerous testis-specific lncRNAs, however, are randomly distributed between the X and autosomes, and we cannot reject the hypothesis that many of these are likely to be spurious transcripts. Finally, using annotated lncRNAs in both species, we identified 134 putative lncRNA homologs between *D. pseudoobscura* and *D. melanogaster* and find that many have conserved developmental expression dynamics, making them ideal candidates for future functional analyses.

Key words: *Drosophila pseudoobscura*, long noncoding RNA, lncRNAs, sex-biased expression, transcriptome evolution.

Introduction

With advances in sequencing technologies and efforts like the human ENCODE and the *Drosophila* modENCODE projects to identify functional elements in the genome, we understand that the majority of the eukaryotic genome is both nonprotein-coding and transcriptionally active (Celniker et al. 2009; Encode Project Consortium 2012). Noncoding regions of the genome, however, are far less likely to be under purifying selection than coding regions; thus, it is unclear to what extent noncoding RNAs impact eukaryotic biology (Sella et al. 2009; Rands et al. 2014). Noncoding RNAs themselves are diverse. Typically, the term evokes a short molecule with a conserved secondary structure and a very specific biological role (e.g. miRNAs), but thousands of long noncoding RNAs (lncRNAs) have also been documented in complex eukaryotes (Ulitsky and Bartel 2013; Kapusta and Feschotte 2014). lncRNAs share many properties with mRNAs: lengths of hundreds or thousands of nucleotides, introns, multiple isoforms from a single locus, and polyadenylation, which facilitates easy identification in poly(A+) RNA-Seq libraries (Cabili et al. 2011; Ulitsky et al. 2011; Derrien et al. 2012; Young et al. 2012; Brown et al. 2014). In contrast, lncRNAs tend to be expressed at lower levels than protein-coding genes and in a more tissue-specific manner (Cabili et al. 2011; Derrien et al. 2012; Young et al. 2012; Brown et al. 2014; Washietl et al. 2014). They also have higher rates of evolutionary turnover, both in sequence and expression, resulting in higher proportions of lineage-specific lncRNAs than protein-coding genes (Kutter et al. 2012; Necsulea et al. 2014).

Early efforts to detect purifying selection on lncRNA exonic sequence produced mixed results, although stronger signals were detected in *Drosophila* than in humans and other vertebrates (Ponjavic et al. 2007; Marques and Ponting 2009; Ward and Kellis 2012; Young et al. 2012; Haerty and Ponting 2013; Schuler et al. 2014). Many of these early efforts, however, only have expression data for a single taxon; thus, it is not clear whether the RNA is even expressed in other taxa. Recent studies in vertebrates more explicitly integrate expression characteristics into comparative analyses (Necsulea et al. 2014; Washietl et al. 2014; Hezroni et al. 2015; Chen et al. 2016). Although thousands of lncRNAs have been annotated, only a small fraction of them have been functionally characterized, with documented roles, among others, in dosage compensation (i.e., the classic lncRNAs *Xist* in mammals and *roX* in *Drosophila*), development, and cell biology (Kung et al. 2013; Bassett et al. 2014). In the absence of empirical molecular data or convincing evidence of selection on the RNA transcript, one cannot reject the possibility that large numbers of lncRNAs are spuriously transcribed with little impact on organismal biology.

The fruit fly *Drosophila* is an ideal system to study the evolution and function of lncRNAs. With extensive genomic resources and knowledge of evolutionary history and population genetics in *Drosophila*, evolutionary analyses of lncRNAs are straight-forward. Mechanistic investigations in *Drosophila* are facilitated by well-characterized development and ample tools for genetic manipulation. About 100 lncRNAs were initially identified in *Drosophila melanogaster* from cDNA libraries, but RNA-Seq data have caused that number to increase rapidly, with over 2,000 lncRNAs now annotated in the *D. melanogaster* FlyBase annotations (Inagaki et al. 2005; Tupy et al. 2005; Young et al. 2012; Brown et al. 2014). lncRNAs in *D. melanogaster* show some of the same properties seen in vertebrates: low expression levels, low but significant evidence of purifying selection, and high-levels of tissue-specificity (Young et al. 2012; Haerty and Ponting 2013; Brown et al. 2014). 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 (Li and Liu 2014).

Despite vast genomic resources throughout the genus, lncRNAs have yet to be characterized in other *Drosophila* species, and this is necessary to fully understand lncRNA evolution and identify candidates for future functional studies. *D. pseudoobscura*, which diverged from *D. melanogaster* 25–55 Ma, has long been used as a model for evolutionary biology. Dobzhansky (1936, 1937) first studied hybrid incompatibilities and investigated causes of hybrid male sterility in *D. pseudoobscura* and its sympatric sister species *Drosophila persimilis*. *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 faster than protein-coding sequences (Richards et al. 2005). A recent pilot study on a small number of lncRNAs in *D. pseudoobscura* and *D. persimilis* showed instances of differential expression between males of the two species and raises questions about lncRNA contributions to transcriptome divergence and hybrid incompatibilities (Jiang et al. 2011).

Extensive identification of lncRNAs in *D. pseudoobscura* and subsequent comparisons to previously annotated lncRNAs in *D. melanogaster* will provide key insights and further test existing hypotheses of lncRNA evolution and function. In vertebrates, most lncRNAs have been found to be lineage-specific, but development and morphology often vary wildly in the surveyed vertebrate taxa (Necsulea et al. 2014; Hezroni et al. 2015). Because development and morphology in *D. pseudoobscura* and *D. melanogaster* are quite similar, it is possible to conduct meaningful comparisons between the expression dynamics of lncRNAs throughout development over more moderate evolutionary timescales. Further, the strong similarity of gene content across homologous chromosome arms combined with conserved microsynteny among *Drosophila* species (Schaeffer et al. 2008) enables the identification of mutually transcribed regions that may be homologous lncRNAs even when sequence conservation is poor or not detectable, as is often the case with lncRNAs (Richards et al. 2005; Chen et al. 2016). While sequence features like chromatin binding sites or RNA secondary structures have been documented as critical for function in some lncRNAs, they are not universally observed in functional lncRNAs. We would expect, however, that truly homologous lncRNAs would have similar expression characteristics, and the extensive transcriptome profiling of *D. melanogaster* facilitates these comparisons (Graveley et al. 2011).

It is well documented that sex-biased genes in *Drosophila*, especially male-biased genes, are unequally distributed between the X chromosome and the autosomes (Parisi et al. 2003; Sturgill et al. 2007; Vibranovski et al. 2009; Bachtrog et al. 2010; Meisel et al. 2012). Several mechanisms have been offered to explain these observations; all suggest selective pressures to optimize the genomic location of sex-biased genes. A recent study reported that male-biased intergenic noncoding RNAs are also underrepresented on the X in *D. melanogaster*, which suggests that similar evolutionary forces act on both protein-coding genes and lncRNAs (Gao et al. 2014). Because the right arm of the *D. pseudoobscura* X chromosome is not homologous to the *D. melanogaster* X (Schaeffer et al. 2008), similar observations in *D. pseudoobscura* would suggest that sex-biased lncRNAs are indeed subject to selection. It is thus possible to use chromosomal distributions of lncRNAs to get insights into their potential functionality.

In order to facilitate genus-wide comparisons with *D. melanogaster* and test these evolutionary hypotheses, we annotated and characterized intergenic lncRNAs (lncRNAs) in *D. pseudoobscura*. We used unstranded RNA-Seq to generate developmental and tissue-specific transcriptome data and computationally identified lncRNAs from unannotated intergenic transcripts. We then characterized the expression dynamics of these lncRNAs throughout sex-specific development and in adult gonad and carcass tissues and considered whether selection could explain chromosomal distributions of sex-biased lncRNAs throughout the genome. Finally, we cross-referenced the *D. pseudoobscura* lncRNAs identified here with those in *D. melanogaster* and used conservation of developmental expression profiles to identify the first set of high confidence homologous lncRNAs in *Drosophila.*

Materials and Methods

More detailed descriptions of methodologies can be found in supplementary file S1, Supplementary Material online.

RNA Sequencing and Transcriptome Assembly

A comprehensive *D. pseudoobscura* transcriptome assembly was generated using RNA-Seq from multiple developmental timepoints and isolated adult tissues of the inbred genome reference line, MV2-25 (Richards et al. 2005). Four unique developmental timepoints were sampled in whole-body flies, with male and female samples collected separately: (1) 1st-instar larva, (2) wandering 3rd-instar larva, (3) mid-stage pupa, and (4) 6-day post-eclosion virgin adults. Sex was determined using morphological characters in adults and via PCR with Y-chromosome specific primers for earlier developmental stages (Carvalho and Clark 2005). Adult ovaries and testes and their resulting carcasses were also isolated and sequenced, resulting in four additional tissue samples. Twenty individuals were used for all samples except the much smaller 1st-instar larvae, where between 28 and 49 individuals were used to generate enough total RNA for RNA-Seq library prep. Total RNA was generated using a standard Trizol extraction protocol.

DNase-treated total RNA was used as input to create indexed poly(A+) RNA-Seq libraries using the Illumina TruSeq RNA Prep Kit. A single deep replicate (replicate A) was generated for transcriptome assembly, and two lower-depth biological replicates (replicates B and C) were generated for expression analyses. RNA sequencing was performed on an Illumina HiSeq1000 to generate 100 bp, unstranded paired-end libraries. Multidimensional scaling plots were generated to assess consistency of replicates (supplementary fig. S2, Supplementary Material online). Contamination for testes RNA was detected in the male carcass A sample, likely due to poor dissections (supplementary fig. S2*B*, Supplementary Material online), so an additional male carcass and testes replicate (replicate D) was generated with similar depth to the A sample and used for all expression analyses (supplementary fig. S2*C*, Supplementary Material online). Low-quality reads were filtered out, and low-quality nucleotides at the 3′ ends were trimmed using the NGS QC Toolkit (Patel and Jain 2012). Filtered, trimmed reads from each sample were then aligned to the *D. pseudoobscura* genome (FlyBase r2) using TopHat v2.0.5 (Langmead and Salzberg 2012; Kim et al. 2013; St Pierre et al. 2014). Transcriptomes were then assembled for each sample using Cufflinks v2.0.2, and a single merged transcriptome assembly was generated using Cuffmerge and the *D. pseudoobscura* FlyBase r2.29 annotation (Trapnell et al. 2010). Sequencing and mapping statistics and NCBI SRA accession numbers for all RNA-Seq samples can be found in supplementary table S3, Supplementary Material online.

Computational Identification of Intergenic lncRNAs

All loci in the transcriptome were classified in Cuffmerge as annotated or novel based on the *D. pseudoobscura* FlyBase r2.29 annotations. The 2,645 novel intergenic loci that map to the five major *D. pseudoobscura* chromosome arms (XL, XR, 2, 3, and 4) were screened for protein-coding ability using three approaches. Any locus that showed evidence of protein-coding ability using any of the three approaches was removed from the list of putative intergenic lncRNA loci.

Search against Existing Protein Databases

Novel intergenic loci were screened against three protein databases: the NCBI nr database, the PeptideAtlas *D. melanogaster* protein database, and a custom *D. pseudoobscura/D. ps. bogotana* testes proteomic dataset. Transcript sequences were aligned to the NCBI nr database (*E*-value $<1\mathrm{e}{-10}$) and PeptideAtlas database (*E*-value $<1\mathrm{e}{-5}$) using BLASTx (Desiere et al. 2006; Camacho et al. 2009). The longest ORF for each transcript (minimum of 10 amino acids) was calculated and translated using custom perl scripts. Putative peptide sequences were then matched against testes proteomic datasets from the MV2-25 line of *D. pseudoobscura*, the Susa6 line of *D. ps. bogotana*, and hybrid crosses between the two subspecies.

Identification of Conserved ORFs Using RNAcode

RNAcode v0.3 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 ($P < 0.05$) (Kuhn et al. 2007; Washietl et al. 2011; Noor 2012). Alignments for all loci were extracted from the UCSC alignment by converting *D. pseudoobscura* FlyBase r2 coordinates to *D. melanogaster* FlyBase r5 coordinates using liftOver and the maf_parse tool from Phast v1.1 (Hinrichs et al. 2006; Hubisz et al. 2011). Alignments for all loci from the Pseudobase alignment, which include multiple lines for several species, were kindly extracted for us by the developers. A second iteration of the Pseudobase alignment was then created using only a single line from four species: *D. pseudoobscura*, *D. persimilis*, *D. miranda*, and *D. lowei*.

Identification of Noncoding Sequence Features Using the Coding Potential Assessment Tool (CPAT)

CPAT was used to identify sequence features specific to protein-coding transcripts using the provided logistic regression model and hexamer frequency tables trained on *D. melanogaster* as well as a custom set created for *D. pseudoobscura* (Wang et al. 2013). The custom 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) (St Pierre et al. 2014). The logistic regression model was 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* lncRNA transcripts that had passed all previously mentioned filters.

Transcript and Sequence Properties of lncRNAs

A set of nonredundant exons was generated from the transcriptome assembly for each locus with a custom R script. For the 10,415 annotated protein-coding loci from *D. pseudoobscura*, all transcripts with the Cuffmerge class codes of "=", "j", and "o" were used. A fasta file was then generated using gffread, and total exonic length was determined using the perl script fastaNamesSizes.pl (http://www.molecularevolution.org/molevolfiles/exercises/QC_of_NGS/fastaNamesSizes.pl). Isoform number was calculated from the merged.gtf file. Genome coverage was calculated using the genomeCoverageBed utility from bedtools v2.17.0 (Quinlan and Hall 2010).

GC, simple repeat, low-complexity sequence, and TE content were calculated using RepeatMasker v4.0.5 with cross_match v0.990329 and *Drosophila* TEs (Smit et al. 2014). Genome and CDS GC content were calculated using annotated fasta files from the *D. pseudoobscura* genome (FlyBase r2.29) that included only sequences from the major chromosome arms (XL, XR, 2, 3, and 4) (St Pierre et al. 2014).

Expression Properties and Dynamics of lncRNAs

Locus fragment counts were generated for each replicate using samtools v0.1.18 and HTSeq-count v0.6.1p1 (Li et al. 2009; Anders et al. 2015). Locus fragment were then converted into the TPM metric (number of transcripts per million) (Li and Dewey 2011). $\log_2$(meanTPM) was then calculated across all three replicates for each sample and compared between samples. For subsequent expression analyses, loci with expression values less than an empirically determined threshold of 0.3 cpm (fragment counts per million fragments mapped) in all samples were removed (supplementary fig. S4, Supplementary Material online). Fragment counts were scale normalized across all samples separately for the whole-body developmental series and the adult tissue samples using calcNormFactors in the edgeR package v3.6.8 (Robinson et al. 2010). Fragment counts were then converted to $\log_2$(cpm) with precision weights using voom and fit to a linear model, all within the limma package v3.20.9 (Smyth 2005; Robinson et al. 2010; Law et al. 2014). Loci were clustered via Pearson's correlation using the hcluster function in the amap package v0.8-12, and heatmaps were generated using the heatmap.2 function in gplots v2.15.0 (Lucas 2014; Warnes et al. 2014). $\log_2$(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 0 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 (supplementary fig. S5*A*, Supplementary Material online). The optimal fuzzifier *m* of 1.436711 was calculated using the mestimate function. After clustering, all loci with membership values <0.5 were removed from clusters. Relationships between clusters are depicted in a Principal Components Analyses plot (supplementary fig. S5*B*, Supplementary Material online). lncRNA over- or underrepresentation was determined using a two-tailed Fisher's exact test with a Benjamini–Hochberg adjustment. Significant Biological Process GO terms ($\text{FDR} < 0.05$) for protein-coding genes in each cluster were identified using GeneCodis3 (Tabas-Madrid et al. 2012).

Differential Expression Analyses

Significant sex-bias was detected using limma-voom (Smyth 2005; Law et al. 2014). After being fit to a linear model as previously described, $\log_2$(cpm) expression values were used to calculate differential expression. Pairwise contrasts were made between male and female equivalents of all four developmental samples, gonads, and carcasses. After Benjamini–Hochberg correction for multiple tests, significant expression bias was determined using an adjusted *P*-value <0.01.

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

To identify a list of testis-specific and ovary-specific genes, we identified a set of genes (both lncRNA and protein-coding) with cpm >0.3 in only the testes or ovaries among all tissue samples.

Genomic Distributions of Sex-Biased Genes

To determine whether sex-biased lncRNAs are depleted or enriched on the X chromosome as compared to the autosomes, we calculated the odds ratio (OR) between the unbiased gene distributions (autosomes/X) and the sex-biased gene distributions (autosomes/X) (Gao et al. 2014). An OR below 1.0 indicates that the X-chromosome is depleted for that class of genes, and an OR above 1.0 indicates that the X-chromosome is enriched for that class of genes.

Genomic coordinates were determined by concatenating and modifying FlyBase r2.29 scaffolds where necessary (Schaeffer et al. 2008). Chromosomes 2 and 3 consist of a single scaffold, so no modifications were necessary. Chromosomes 4, XL, and XR scaffolds were concatenated in the order shown in published cytogenetic maps. 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. BLASTn (*E*-value $<1e-10$) was performed using testis-specific and male-biased non-testis-specific locus sequences against four *Drosophila* genome assemblies: *D. miranda* [SRA:SRX105954], *D. lowei* [SRA:SRX091467], *D. affinis* [SRA:ERX103525], and *D. melanogaster* (FlyBase r6.02) (Camacho et al. 2009; St Pierre et al. 2014). Genome assemblies for *D. miranda*, *D. lowei*, and *D. affinis* were created by aligning reads to the *D. pseudoobscura* genome (Flybase r3.2) using bwa v0.7.9a-r786 (Li and Durbin 2009; St Pierre et al. 2014). Random sets of intergenic sequences were generated using shuffleBed and testis-specific sets of lncRNAs and protein-coding genes (Quinlan and Hall 2010).

Pearson's correlations were calculated between developmental expression profiles (eight whole-body samples) of lncRNA loci and nearest protein-coding neighbor, protein-coding loci and nearest protein-coding neighbor, and 1,000 permutations of lncRNA loci and randomly-associated protein-coding loci for unbiased and sex-biased gene sets. The nearest single nonoverlapping protein neighbors to the genes from unbiased or sex-biased gene sets were identified using the closest utility in bedtools (Quinlan and Hall 2010).

Identification of Putative lncRNA Homologs between *D. pseudoobscura* and *D. melanogaster*

Reciprocal BLASTn searches were performed using parameters that are more tolerant of gaps and mismatches than default parameters (Mount et al. 2007; Camacho et al. 2009). The first BLASTn search queried the set of 1,771 *D. pseudoobscura* lncRNA transcripts against the full *D. melanogaster* transcriptome (FlyBase r6.02) (St Pierre et al. 2014). *D. melanogaster* best hits were 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* lncRNA transcripts. Reciprocal best hits that matched were retained as putative lncRNA homologs.

We also searched for putative homology between the 1,589 *D. pseudoobscura* lncRNA loci and the 2,359 annotated lncRNA loci in *D. melanogaster* (FlyBase r6.02) using coordinate overlap (St Pierre et al. 2014). *D. pseudoobscura* lncRNA coordinates (FlyBase r2 or UCSC dp4) were first converted to *D. melanogaster* FlyBase r5 coordinates (i.e., UCSC Dm3) with the UCSC liftOver tool, and *D. melanogaster* coordinates were then converted from FlyBase r5 to r6 in FlyBase (Hinrichs et al. 2006; Kuhn et al. 2007; St Pierre et al. 2014). Overlap was detected using intersectBed (Quinlan and Hall 2010). Unambiguous one-to-one lncRNA locus matches were retained as putative lncRNA homologs.

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). *D. melanogaster* developmental stages are roughly equivalent to those collected in *D. pseudoobscura*, and are mixed single-end and paired-end 75-bp Illumina sequence reads, though only adult samples are sex-specific. Reads were QC filtered and mapped to *D. melanogaster* genome (FlyBase r6), and fragment counts were obtained using the same pipeline described above for *D. pseudoobscura*.

Because the *D. melanogaster* data are not sex-specific in early developmental stages, male and female *D. pseudoobscura* datasets were pooled to create an approximation of the five stages available in *D. melanogaster*: 1st-instar larvae, 3rd-instar larvae, mid-pupae, adult males, and adult females. We generated $\log_2$(cpm) expression values for all annotated genes individually for both species (*D. pseudoobscura* r2.29 plus lncRNAs, *D. melanogaster* r6.02) using limma-voom as previously described (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; Waterhouse et al. 2013; St Pierre et al. 2014). Pearson correlation coefficients between 134 putative lncRNA homologs, 7,451 orthologous protein-coding genes, and a set of randomly associated lncRNAs between the two species were generated in R (R Core Team 2014). To generate the randomized control, the 65 lncRNAs with homologs in both species were randomly associated, and a *P*-value was generated by comparing the median correlation coefficients from each of 1,000 permutations with the empirically derived median correlation coefficient of 0.7435 for the putative lncRNA homologs. The nearest single and five nonoverlapping neighbors to the *D. melanogaster* lincRNA homologs were identified using the closest utility in bedtools (Quinlan and Hall 2010).

Results

Identification of lncRNAs in *D. pseudoobscura*

Using RNA-Seq data from 12 sex-specific developmental (whole-body 1st-instar larvae, 3rd-instar larvae, pupae, and adults) and adult tissue samples (gonads and carcasses) of *D. pseudoobscura*, we generated a single merged transcriptome that contains 50,459 transcripts expressed from 18,317 genomic loci (fig. 1*A*). Only three of these loci are annotated as lncRNAs in FlyBase r2.29: *RNaseP:RNA* (GA29345), *HSR-omega* (GA30101), and *SRP* (GA29352) (St Pierre et al. 2014). When compared with the *D. pseudoobscura* gene annotations (FlyBase r2.29), 5,261 loci are annotated as both novel and intergenic. The 2,616 novel, intergenic loci that do not map to one of the major *D. pseudoobscura* chromosome scaffolds but instead to an "Unknown_group" or "Unknown_singleton" are not considered further.

The remaining 2,645 novel, intergenic loci were screened for protein-coding potential using multiple filters (fig. 1*A*). In total, 1,059 loci showed evidence of protein-coding potential using at least one filter method and were removed from consideration as lncRNA loci. Matches to existing protein databases were found for 308 loci (fig. 1*B*). *De novo* searches for protein-coding potential using conserved ORFs via RNAcode and noncoding sequence features via CPAT found hits in 777 and 233 loci, respectively (Washietl et al. 2011; Wang et al. 2013). Protein-coding potential is evident in only 51 loci using all three approaches, and RNAcode alone accounts for the majority of loci that were eliminated from consideration as lncRNAs. Results of specific iterations for each approach can be found in supplementary fig. S6, Supplementary Material online. After filtering, we are left with 1,586 novel putative intergenic lncRNA loci in addition to the three previously annotated lncRNA loci (supplementary table S7, Supplementary Material online).

Transcript, Sequence, and Expression Properties of *D. pseudoobscura* lncRNAs

Exons from the 1,589 annotated lncRNA loci cover only 1.2% of the major chromosome scaffolds in the *D. pseudoobscura* genome (1,483,658/127,291,806 bp, FlyBase r2), as opposed to the 26.5% (40,419,012 bp) coverage of exons, including CDS and UTR, from a reference set of 10,415 protein-coding loci present in our transcriptome data. Including introns, lncRNA and protein-coding loci cover 1.9% (2,357,269 bp) and 74.9% (95,338,393 bp) of the major scaffolds, respectively. As observed in other eukaryotes, transcript and sequence properties of *D. pseudoobscura* lncRNA loci distinguish them from protein-coding loci (Cabili et al. 2011; Derrien et al. 2012; Niazi and Valadkhan 2012; Pauli et al. 2012; Young et al. 2012; Brown et al. 2014; Li et al. 2014). Total exonic length at a locus is shorter for lncRNA loci than protein-coding loci (median 772 vs. 3,165 bp, Mann–Whitney $P < 2.2e{-}16$, fig. 2*A*). lncRNA loci typically contain fewer exons than protein-coding loci (mean 1.5 vs. 6.0 exons per locus, Mann–Whitney $p < 2.2e{-}16$, fig. 2*B*). A majority of lncRNA loci (68.5%, 1,088 loci) contain only a single exon. Consequently, alternative transcription at an lncRNA locus is rare, with only 9.4% (149 loci) showing evidence of multiple isoforms as compared with 69.7% (7,264) of protein-coding loci.

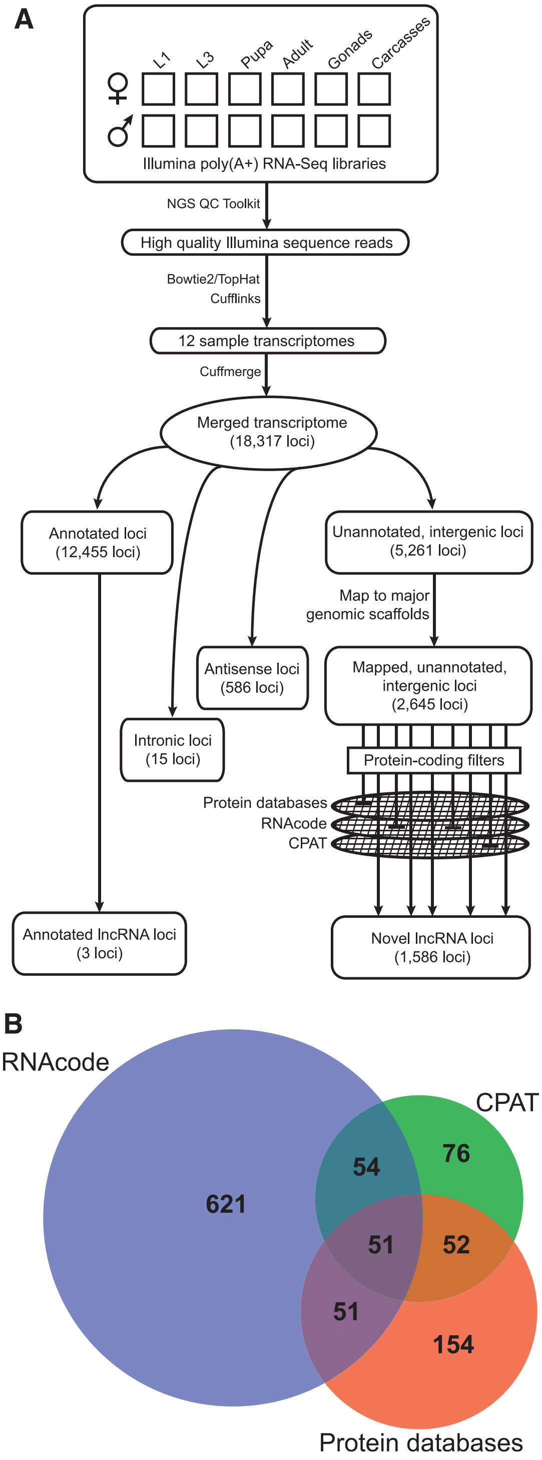
A
L1
L3
Pupa
Adult
Gonads
Carcasses
♀
♂
Illumina poly(A+) RNA-Seq libraries
NGS QC Toolkit
High quality Illumina sequence reads
Bowtie2/TopHat
Cufflinks
12 sample transcriptomes
Cuffmerge
Merged transcriptome (18,317 loci)
Annotated loci (12,455 loci)
Unannotated, intergenic loci (5,261 loci)
Map to major genomic scaffolds
Antisense loci (586 loci)
Intronic loci (15 loci)
Mapped, unannotated, intergenic loci (2,645 loci)
Protein-coding filters
Protein databases
RNAcode
CPAT
Annotated lncRNA loci (3 loci)
Novel lncRNA loci (1,586 loci)
B
RNAcode
CPAT
621
54
76
51
52
51
154
Protein databases

Fig. 1.—Computational identification of *D. pseudoobscura* lncRNAs from RNA-Seq data. (*A*) 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. Unannotated, intergenic loci were screened for protein-coding ability using three complementary approaches, with the numbers of protein-coding loci identified via each approach detailed in (*B*). In total, 1,589 putative lncRNA loci were identified.

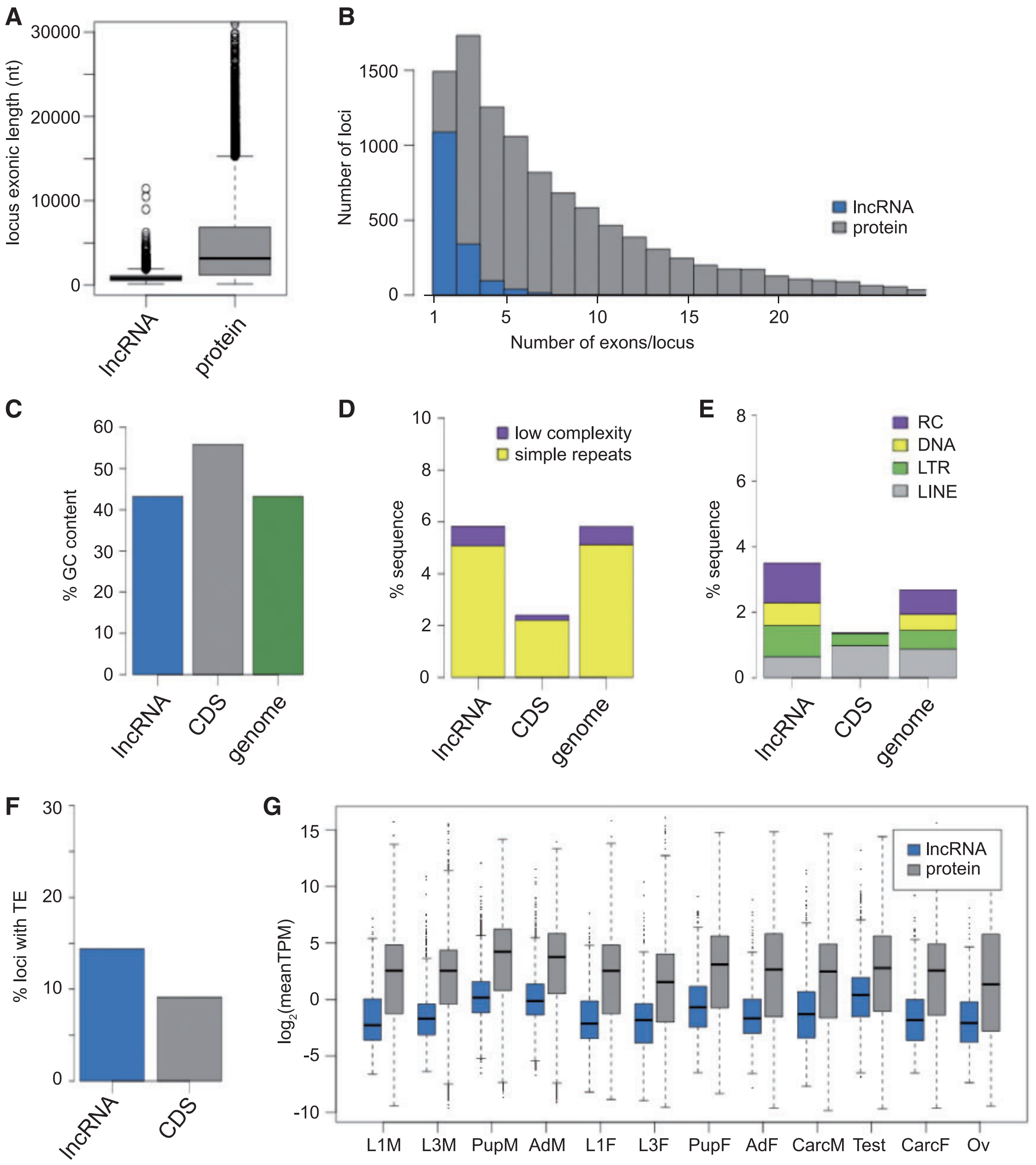

Fig. 2.—Transcript, sequence, and expression properties of *D. pseudoobscura* lncRNAs. (*A*) Total exonic length of 1,589 lncRNA loci and 10,415 protein-coding loci. (*B*) Distribution of exon number per locus for lncRNA and protein-coding loci. (*C*) GC content of lncRNA exons, CDS, and whole genome (XL, XR, 2, 3, and 4 scaffolds only). (*D*) Low complexity and simple repeat sequence content of lncRNA exons, CDS, and whole genome (XL, XR, 2, 3, and 4 scaffolds only). (*E*) TE content (LINE and LTR retrotransposons, DNA transposons, and rolling-circle transposons) of lncRNA exons, CDS, and whole genome (XL, XR, 2, 3, and 4 scaffolds only). (*F*) Percent of lncRNA loci and CDS with detectable TE content. (*G*) Expression values for lncRNA and protein-coding loci in each of 12 samples in $\log_2$(meanTPM). All loci with TPM >0 are included, and all comparisons between classes are significant (Mann–Whitney, $P < 2.2e-16$).

Sequence properties of lncRNA and CDS differ, with lncRNA sequence showing less deviation from genome-wide levels. lncRNA loci have lower GC content (43.8% vs. 56.0%, fig. 2C) and higher levels of low-complexity (0.73% vs. 0.18%, fig. 2D) and simple repeat sequence (5.09% vs. 2.28%, fig. 2D) than protein-coding loci . Three classes of transposable elements (TEs) are found in a greater percentage of lncRNA sequence than CDS (fig. 2E): LTR retrotransposons (0.95% vs. 0.36% of total sequence), DNA transposons (0.69% vs. 0.02%), and rolling-circle transposons (1.20% vs. 0.01%). LINE retrotransposons, however, comprise 0.65% of lncRNA sequence and 0.98% of CDS sequence (fig. 2E). Altogether, TEs are found in 14.4% of lncRNA loci and 9.1% of all CDS (fig. 2F).

Expression levels observed at lncRNA loci are significantly lower than at protein-coding loci in each of the 12 samples surveyed via RNA-Seq (Mann–Whitney, $P < 2.2e{-}16$, fig. 2G).

Expression Dynamics of lncRNAs

The expression dynamics of lncRNAs were analyzed in a sex-specific manner throughout development in whole-body flies and in dissected adult gonads. Developmental time points were chosen around major developmental and sexual milestones: a single time point before extensive proliferation of gonadal precursors with presumably little differences between the sexes (1st-instar larvae); a time point where gonadal proliferation has occurred and spermatogenesis is underway (3rd-instar larvae); a time point in the midst of metamorphosis (pupae); and a time point at sexual maturity (adults) (King 1970; Bate and Martinez Arias 1993; Hartenstein 1993). In these expression analyses, we included 925 lncRNA loci (58.2%) and 7,649 protein-coding loci (73.4%) with a minimum mean expression value of 0.3 fragment counts per million mapped fragments (cpm) across three biological replicates in any of the 12 development or tissue samples.

lncRNA Expression throughout Development

The numbers of expressed lncRNAs increase in both sexes as development proceeds from the 1st-instar larval (115 in male, 110 in female) through 3rd-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 (fig. 3). The highest number of expressed lncRNAs is observed in the adult males (481), but the number of expressed lncRNAs 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. As seen in *D. melanogaster*, few (42/925, 4.5%) lncRNA loci are expressed at all developmental stages, particularly when compared with the numbers of broadly expressed protein-coding loci (4,587/7,649, 60.0%) (Brown et al. 2014). On the whole, lncRNAs are expressed in far fewer developmental samples than protein-coding loci (lncRNA mean = 2.70, protein mean = 6.34, Mann–Whitney test, $P < 2.2e{-}16$).

Total expression of lncRNAs is more variable than protein-coding expression throughout *D. melanogaster* development, with peak expression in adult males and minimum expression in the pupae (Graveley et al. 2011; Young et al. 2012). To facilitate a more rigorous statistical analysis of lncRNA expression through *D. pseudoobscura* development, we performed fuzzy *c*-means cluster analysis of developmental expression profiles for all lncRNA and protein-coding loci (Kumar and Futschik 2007). 729 of the 925 expressed lncRNA loci (78.8%) and 5,716 of the 7,649 expressed protein-coding loci (74.7%) clustered into 16 clusters that are nonoverlapping when only genes with membership values >0.5 are considered (fig. 4). Clusters are clearly defined by developmental stage and sex. Three clusters with increasing male-specific expression (clusters 2, 6, and 14) contain the highest numbers of lncRNA loci. The numbers of lncRNA loci are significantly overrepresented as compared with protein-coding loci in two of these (clusters 2 and 14, Fisher's exact test, Benjamini–Hochberg adjusted $P < 0.05$). lncRNAs are also significantly overrepresented in a cluster (cluster 4) with nonsex-biased increases in expression in pupal stages (Fisher's exact test, Benjamini–Hochberg adjusted $P = 1.49e{-}6$). lncRNAs are significantly underrepresented in a number of clusters (Fisher's exact test, Benjamini–Hochberg adjusted $P < 0.05$), particularly those that show female-specific expression (clusters 1, 9, and 10) and expression in both 1st-instar larval stages (clusters 1, 3, 11, and 16). Numbers of lncRNA and protein-coding genes for all clusters are listed in supplementary table S8, Supplementary Material online.

Top Biological Process Gene Ontology (GO) matches were determined for each cluster (Tabas-Madrid et al. 2012). Top GO hits for the three clusters with an overrepresentation of lncRNAs are: spermatogenesis and sensory-perception of smell (cluster 2), homophilic cell adhesion (cluster 4), and microtubule-based movement (cluster 14) (supplementary table S9, Supplementary Material online). Top GO hits for the nine clusters with an underrepresentation of lncRNAs are listed in supplementary table S10, Supplementary Material online.

lncRNA Expression in Gonads

lncRNA and protein-coding locus expression follows a similar pattern in the gonadal and carcass tissues, with parallels in direction of change but differences in magnitude (fig. 3). As also seen in *D. melanogaster*, the highest number of expressed lncRNAs is seen in the testes (525) and the lowest seen in the ovaries (77) (Brown et al. 2014). The carcass samples show intermediate levels of lncRNA expression, with 272 expressed in the male carcass and 172 expressed in the female carcass. Carcass samples for both lncRNAs 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. lncRNAs are 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, lncRNAs are expressed in significantly fewer of the four tissue types than protein-coding loci (lncRNA mean = 1.39, protein mean = 3.04, Mann–Whitney test, $P < 2.2e{-}16$).

Sex-Biased Expression of *D. pseudoobscura* lncRNAs

Sex-specific expression data are limited in *D. melanogaster*, with sex-specific RNA-Seq data available only for adult stages in the modENCODE datasets (Graveley et al. 2011).

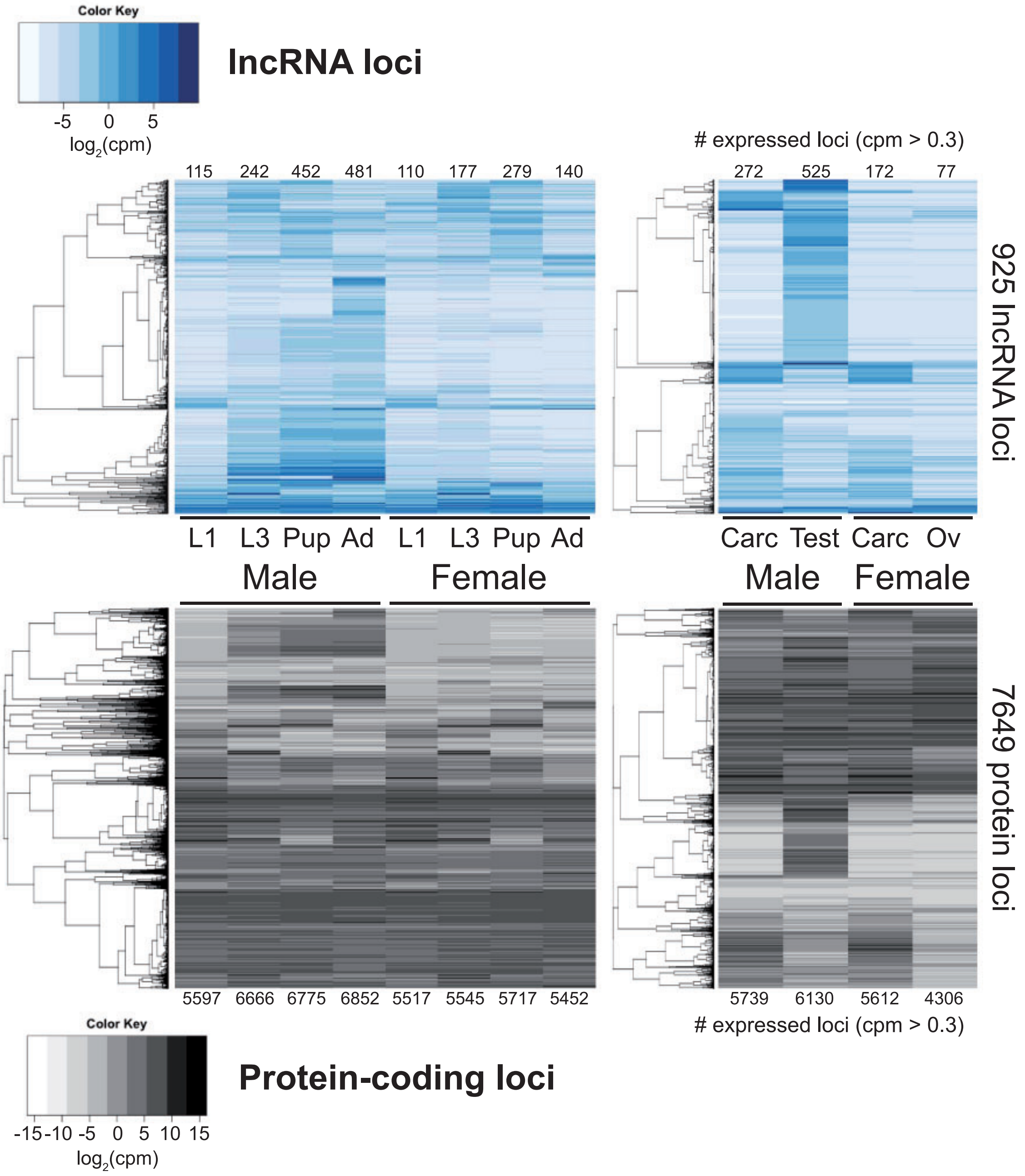

FIG. 3.—Heatmaps of lncRNA and protein-coding locus expression through development and in adult tissues. $\log_2$(cpm) values were used to generate heatmaps for the 925 lncRNA loci and 7,649 protein-coding loci included in expression analyses. Each row represents an individual locus, and row clustering was performed using Pearson's correlation. The numbers of expressed loci (mean cpm > 0.3) for each sample are located above (lncRNA) or below (protein-coding) each sample column.

In *D. melanogaster* adults, lncRNA expression is higher in males than females (Young et al. 2012). Developmental clustering in *D. pseudoobscura* strongly suggests that a large number of loci, both lncRNA and protein-coding, show sex-biased expression in earlier developmental stages as well. To tease apart whether lncRNAs and protein-coding loci show the same patterns of sex-bias, we analyzed differential expression using the limma-voom package (adj. $P < 0.01$) for each developmental stage and the carcass and gonad tissues (Law et al. 2014).

We found few sex-biased genes of either class in the 1st-instar larvae, and progressive increases in the numbers of sex-

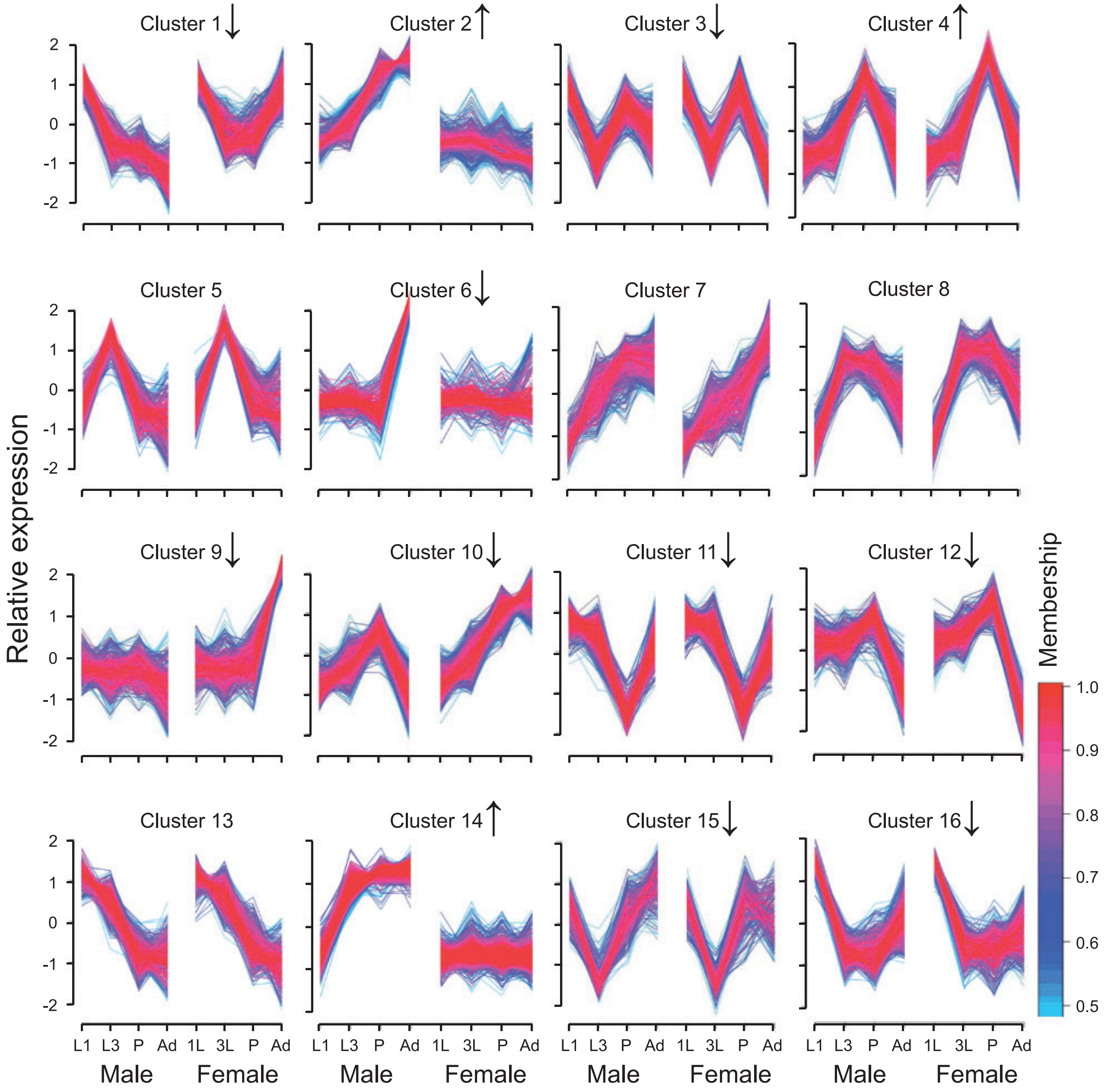

FIG. 4.—Soft clustering of expression profiles through development. Soft clustering of developmental expression profiles for the combined set of lncRNA and protein-coding loci. The *y*-axis of each chart represents relative expression changes, with the mean expression value for each locus centered on 0. The color of an individual locus' expression profile indicates its membership value in the cluster. Up and down arrows next to the cluster name indicate whether lncRNAs are significantly overrepresented or underrepresented, respectively, in the cluster (Fisher's exact test with Benjamini-Hochberg correction for multiple comparisons, $P < 0.05$).

biased genes thereafter (fig. 5*A*, supplementary table S11, Supplementary Material online). While the total proportions of sex-biased lncRNA and protein-coding loci remain comparable in every developmental stage, with roughly 80% of genes showing sex-biased expression in adults, the expression bias in lncRNAs is significantly skewed toward males in the 3rd-instar larval, pupal, and adult stages (Fisher's exact test, $P < 0.05$). Male expression bias in lncRNAs is more frequent as development proceeds; female expression bias in lncRNAs remains relatively low throughout development, with a maximum 10.1% of all lncRNAs showing female expression bias in the pupal stage while over 30% of protein-coding genes are female-biased in the pupal and adult stages.

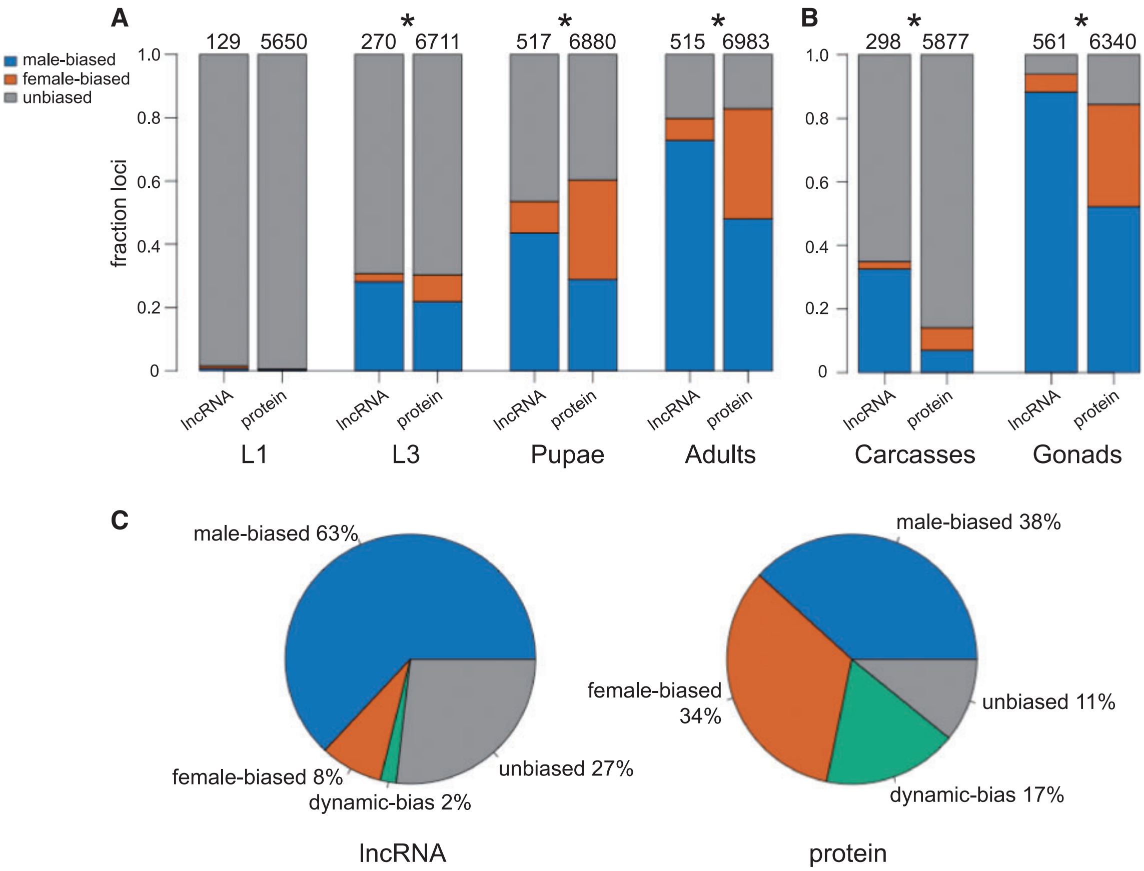

Fig. 5.—Sex-biased expression of lncRNAs. Fractions of loci that show significant sex-biased expression (adj. $P < 0.01$) in (*A*) whole-body developmental samples and (*B*) adult gonadal tissue samples. Numbers above the columns indicate the total number of expressed loci at that stage (mean cpm >0.3). * indicates a significant difference between proportions of sex-biased genes via Fisher's exact test ($P < 0.05$). (*C*) Overall proportions of sex-biased genes integrated across all samples.

Patterns of sex-bias in the dissected gonads and carcasses suggest a basis for the developmental sex-bias patterns (fig. 5*B*). Female-biased lncRNAs are rare in the carcasses, as are both male and female-biased protein-coding genes (all <7.2%). On the other hand, 32.6% of expressed lncRNAs show male-biased expression in the carcasses. Patterns of sex-bias in the gonads mirror the patterns seen in the whole-body adults. Gonad sex-bias is high for both lncRNAs and proteins, with a significant skew towards male-biased expression for the lncRNAs. lncRNA and protein-coding proportions of sex-biased genes are significantly different (Fisher's exact test, $P < 0.05$).

When we integrate data from all six different sample types, both whole-body developmental stages and dissected adult tissues, we find 583 lncRNA loci that show male-biased expression in at least one sample but no female-biased expression in any sample (i.e., "male-biased", fig. 5C, supplementary table S11, Supplementary Material online). 75 lncRNA loci show female-biased expression in at least one sample but no male-biased expression in any sample (i.e., "female-biased", fig. 5C). 248 lncRNAs have no sex-biased expression in any sample (i.e., "unbiased lncRNAs", fig. 5C). 19 lncRNA loci show evidence of both male and female expression bias in different samples (i.e., "dynamic-bias lncRNAs", fig. 5C). 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 (fig. 5C). The cumulative proportions of sex-biased loci in lncRNAs and protein-coding genes are significantly different (Fisher's exact test, $P < 0.05$). The majority of the dynamic-bias genes, in both lncRNAs (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.

Genomic Localization of Sex-Biased lncRNAs

Nonrandom distributions of sex-biased genes among chromosomes, particularly the demasculinization of the X chromosome, have been observed in *Drosophila* for both protein-coding genes and intergenic noncoding RNAs (Parisi et al. 2003; Sturgill et al. 2007; Vibranovski et al. 2009; Bachtrog et al. 2010; Meisel et al. 2012; Gao et al. 2014). These patterns have been explained using evolutionary models that invoke selection to, for example, localize male-beneficial genes off of a precociously silenced X during meiosis in testes or localize female-beneficial genes on the relatively more-abundant X in females (Rice 1984; Charlesworth et al. 1987; Vibranovski et al. 2009; Bachtrog et al. 2010). Here, we explore the effects of sex-biased expression on *D. pseudoobscura* lncRNA chromosomal localization using an Odds Ratio (OR) of the genomic distributions of unbiased and sex-biased genes and significance determined using the Fisher's exact test ($P < 0.05$).

Male-biased lncRNAs and male-biased protein-coding genes, as determined from the integrated sex-bias analyses (fig. 5C), are both significantly underrepresented on the X chromosome (OR = 0.71 for both, fig. 6*A*, supplementary table S12, Supplementary Material online). Female-biased lncRNAs display the opposite trend: significant enrichment on the X chromosome (OR = 2.15, $P < 0.001$). In contrast, there is no evidence of significant female-biased protein-coding gene enrichment on the X (OR = 1.10, $P = 0.2415$).

We further divided our male-biased genes into two sets: testis-specific genes and non-testis-specific genes. Male-biased genes that are not testis-specific are significantly underrepresented on the X chromosome, both for lncRNA (OR = 0.63, $P = 0.04042$) and protein-coding loci (OR = 0.60, $P = 7.6\text{e}{-9}$) (fig. 6*B*). On the other hand, testis-specific lncRNAs (OR = 0.74, $P = 0.0887$) and protein-coding genes (OR = 0.91, $P = 0.3502$) show statistically random distributions between the X and autosomes (fig. 6*B*). To see whether these random distributions can be explained by gene age, we performed BLASTn of male-biased genes as categorized in fig. 6*B* against genomes of four increasingly divergent species of *Drosophila*: *D. miranda*, *D. lowei*, *D. affinis*, and *D. melanogaster*. Testis-specific protein-coding genes have noticeably lower rates of BLASTn hits in the *D. affinis* (84.0%) and *D. melanogaster* (27.7%) genomes than male-biased but nontestis-specific protein-coding genes (98.4% and 68.4%, respectively), but this trend is not observed with lncRNAs (fig. 6C). Rates of BLASTn hits for testis-specific and non-testis-specific male biased lncRNAs are roughly equal for both the *D. affinis* (84.7% vs. 86.7%) and *D. melanogaster* (50.5% vs. 48.7%) genomes. One thousand BLASTn permutations of randomly-selected intergenic sequences, controlled for lncRNA length, against the *D. melanogaster* genome suggest that conservation levels of *D. pseudoosbcura* testis-specific lncRNA sequences and random intergenic sequences are equivalent (median = 51.7%, $P = 0.327$). The corresponding permutation analysis using random intergenic sequences adjusted for protein-coding lengths suggests that testis-specific protein-coding genes are significantly less conserved than random intergenic sequences (median = 61.4%, $P < 0.001$).

To address the possibility that the observed genomic distributions of lncRNAs could be the indirect result of spurious transcriptional events arising from the regulation of neighboring protein-coding genes, we estimated correlations between developmental expression profiles of lncRNA loci and their nearest protein-coding neighbor (fig. 6D). These correlations were then compared with correlations of lncRNAs and random protein-coding loci as well as correlations of protein-coding loci within the same sex-bias class and their nearest neighbors. Correlations between lncRNAs and their nearest neighbors were significantly higher than those for randomly associated genes in all classes (Mann–Whitney, $P < 6.453\text{e}{-4}$), though the testis-specific lncRNAs had the highest median correlation and lowest *P*-value by at least nine orders of magnitude ($P < 2.2\text{e}{-16}$). However, correlations between protein-coding genes and their nearest neighbors were also significantly higher for all classes. While median correlations for female-biased and both classes of male-biased lncRNAs were higher than their protein-coding counterparts, none were significantly different (fig. 6D). These results are consistent with general coexpression patterns observed among neighboring protein-coding genes (Ghanbarian and Hurst 2015) and in lncRNAs from humans and *D. melanogaster* (Derrien et al. 2012; Young et al. 2012), although here we have also looked at sex-biased transcription.

Homology of *Drosophila* lncRNAs

With annotated sets of lncRNAs in both *D. pseudoobscura* described here and *D. melanogaster* available through FlyBase, we set out to identify potentially homologous lncRNAs (St Pierre et al. 2014). Using a best-hit reciprocal BLASTn approach, we identified 80 putative lncRNA homologs (fig. 7*A*). We further identified evidence of homology in 114 lncRNA loci by looking for coordinate overlap in a whole genome alignment between the two species (fig. 7*A*). Sixty putative homologs were identified using both methods; taken together, we found 134 putative lncRNA homologs between *D. pseudoobscura* and *D. melanogaster* (supplementary table S13, Supplementary Material online).

With developmental poly(A+) RNA-Seq now available in both *D. pseudoobscura* and *D. melanogaster*, we looked for correlations between developmental expression profiles. Because the *D. melanogaster* RNA-Seq data are 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. Sixty-five of the 134 putative lncRNA homologs have expression levels above the 0.3 cpm threshold in both species. Of these, 21 (32.3%) have Pearson's correlation coefficients above 0.9, indicating strong correlation. Thirty-six (55.4%) and 43 (66.2%)

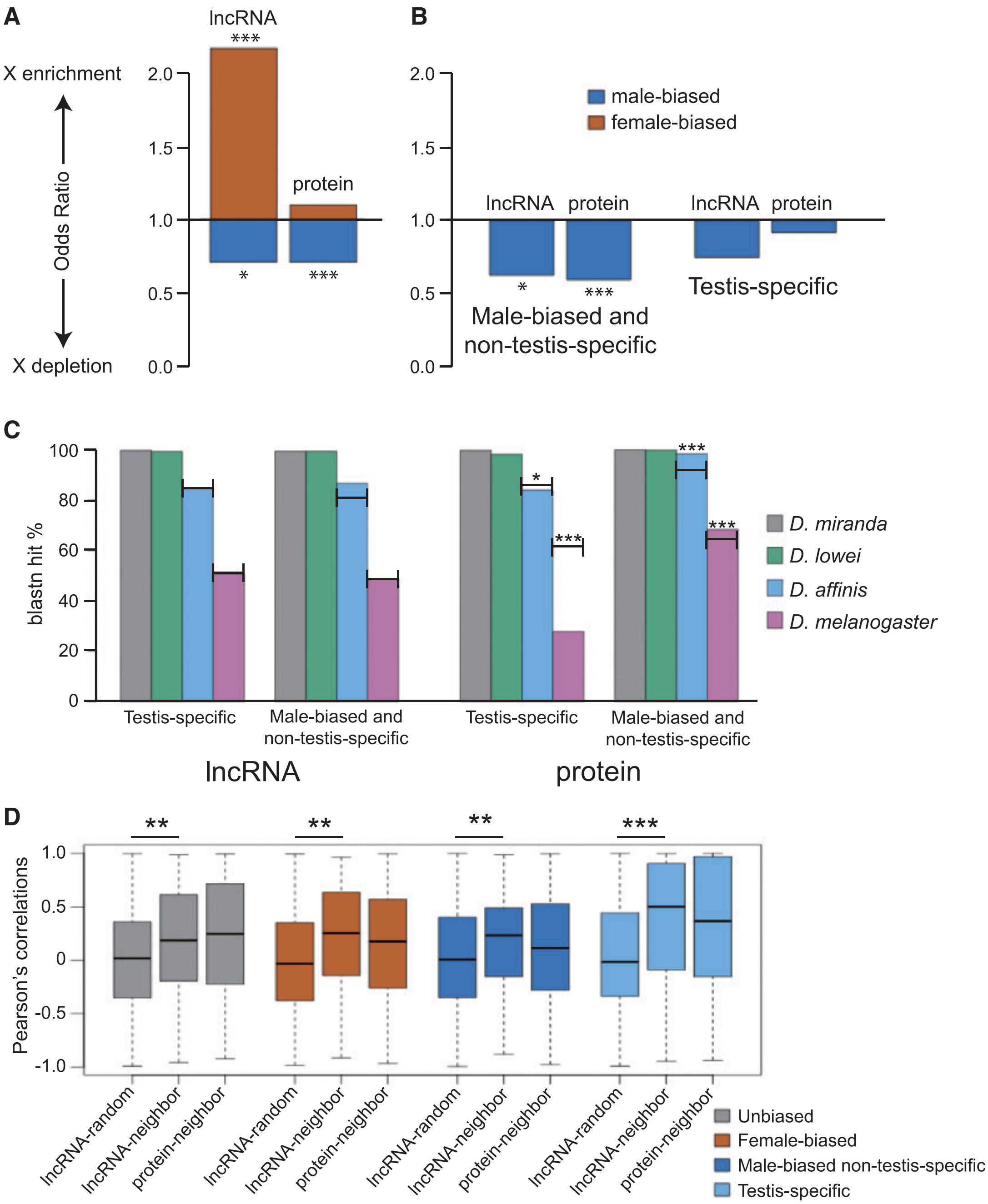

Fig. 6.—Sex-biased expression and genomic localization. Odds Ratios (ORs) between genomic distributions of unbiased and sex-biased genes shown for (*A*) all male-biased and female-biased lncRNA and protein-coding loci (Fisher's exact test, ***$P < 0.001$, *$P < 0.05$) and (*B*) male-biased genes subdivided by tissue specificity in adults. OR above 1.0 indicates relative enrichment on the X chromosome, and OR below 1.0 indicates relative depletion on the X chromosome. *P*-values from Fisher's exact test are indicated above or below each column in (B). (*C*) BLASTn hit rates of testis-specific and all male-biased non-testis-specific genes (lncRNAs and protein-coding genes) against four increasingly divergent *Drosophila* genomes: *D. miranda*, *D. lowei*, *D. affinis*, and *D. melanogaster*. Black bars in each column for *D. affinis* and *D. melanogaster* indicate the median BLASTn hit rate from 1,000 permutations of random intergenic sequence with the same length distributions as the queried sequence. Significance of the observed BLASTn hit rate as compared with these 1,000 permutations is noted (***$P \leq 0.001$; *$P < 0.01$). (*D*) Distributions of Pearson's correlations between developmental expression profiles (eight samples) for: (1) lncRNA loci and random protein-coding loci (1,000 permutations), (2) lncRNA loci and nearest protein-coding neighbor, and (3) protein-coding loci and their nearest protein-coding neighbor. Correlations were determined for unbiased, female-biased, male-biased but nontestis-specific, and testis-specific gene sets. (***$P < 2.2e{-}16$; **$P < 6.453e{-}4$).

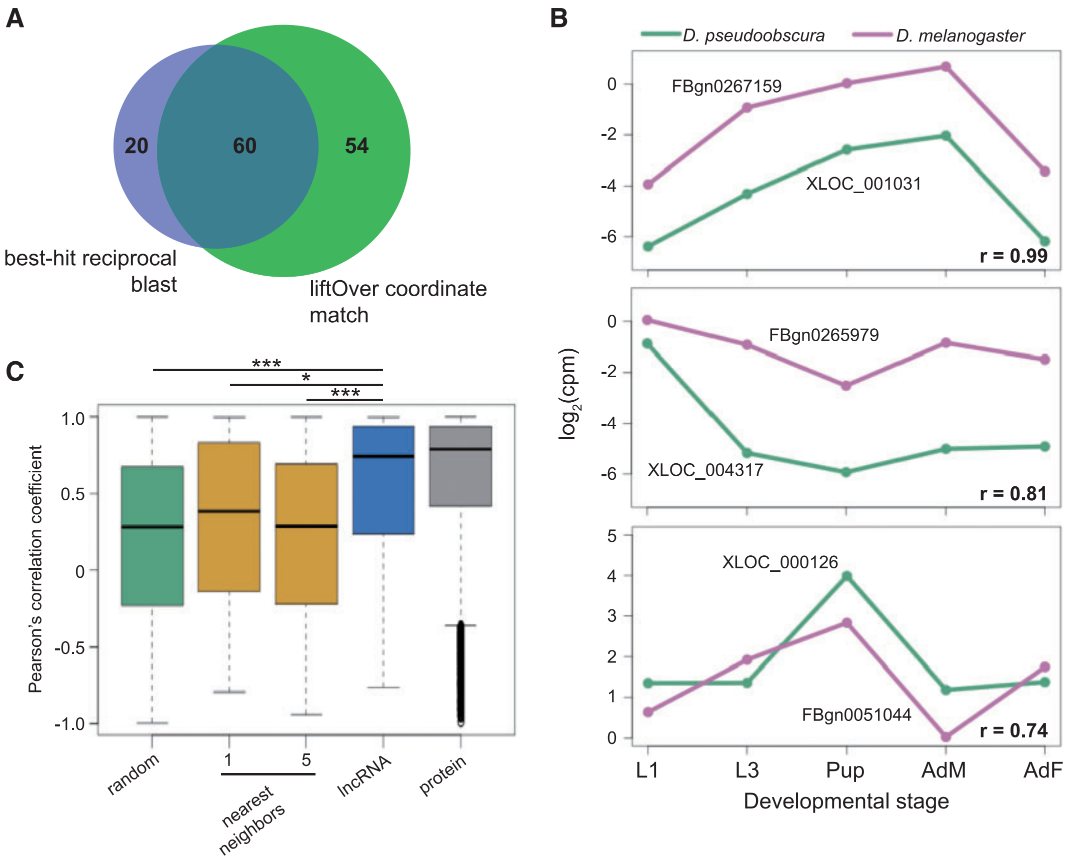

Fig. 7.—Identification of putative lncRNA homologs between *D. pseudoobscura* and *D. melanogaster*. (*A*) Numbers of putative lncRNA homologs identified between *D. pseudoobscura* and *D. melanogaster* using a best-hit reciprocal BLAST approach and a genome coordinate overlap approach. (*B*) Developmental expression profiles of three pairs of putative lncRNA homologs with high Pearson's correlation coefficients. (*C*) Distributions of Pearson's correlation coefficients among a set of randomly paired lncRNA loci between *D. pseudobscura* and *D. melanogaster*, 65 *D. pseudoobscura* lncRNA homologs and the nearest single and five neighboring protein-coding genes to their corresponding *D. melanogaster* homologs, 65 putative lncRNA homologs, and 5,797 protein-coding orthologs. Distributions between putative lncRNA loci and randomly-paired lncRNA loci over 1,000 permutations are significantly different (*$P < 0.01$, ***$P < 0.001$).

have correlation coefficients >0.7 and 0.5, respectively (fig. 7*B*). Seventy of the 134 putative lncRNA homologs have expression levels above the 0.3 cpm threshold in the adult tissue samples of *D. pseudoobscura*. Of these, 34 (48.6%) have testis-specific expression in *D. pseudoobscura*, which is less than the 56.9% (427/750) of all lncRNAs that have testis-specific expression in *D. pseudoobscura*, albeit not significantly so (Fisher's exact test, $P = 0.2078$). Sixty-three putative lncRNA homologs are included in the developmental expression profile clusters (fig. 4). Of these, 54.0% (34/63) are present in clusters with male-bias (clusters 2, 6 and 14). This is less than the 62.3% (454/729) of all lncRNAs that are included in the cluster analysis, though again, not significantly so (Fisher's exact test, $P = 0.5147$).

To compare developmental expression profiles of putative lncRNA homologues with those of protein coding orthologs, we also determined correlation coefficients for 7,845 orthologous protein-coding genes and a control set of lncRNA loci coupled at random. Of the protein-coding orthologs, 2,048 are not expressed above our threshold. Expression correlations for protein-coding orthologs are only moderately higher than those for lncRNAs, though not significantly so (Mann–Whitney, $P = 0.4385$, fig. 7C), with 1,940 (33.5%) with $r > 0.9$, 3,422 (59.0%) with $r > 0.7$, and 4,157 (71.7%) with $r > 0.5$. The distribution of expression correlations for the putative lncRNA homologs is significantly higher than the control of randomly paired lncRNA loci (1,000 permutations, $P < 0.001$, fig. 7C). These high correlations among putative lncRNA homologs are locus-specific, as expression correlations between the *D. pseudoobscura* lncRNA homologs and the nearest neighbors of their corresponding *D. melanogaster* lncRNA homologs are significantly lower (Mann–Whitney, $P = 0.008063$ for the nearest single neighbor, $P = 5.034e{-}06$ for the nearest five neighbors, fig. 7C).

Discussion

Identification of lncRNAs from RNA-Seq Data

Using RNA-Seq from multiple developmental stages and adult tissue samples, we computationally identified and characterized intergenic lncRNAs in a second species of *Drosophila*: the important evolutionary model *D. pseudoobscura*. Since annotation of the *D. pseudoobscura* genome has lagged behind *D. melanogaster*, with only three annotated lncRNAs in FlyBase r2.29, we screened all novel intergenic transcripts that map to the major chromosome scaffolds for evidence of protein-coding ability using three different approaches: (1) local alignments to existing protein databases, (2) conserved ORF identification using RNAcode, and (3) identification of noncoding sequence features using CPAT (Desiere et al. 2006; Camacho et al. 2009; Jiang et al. 2011; Washietl et al. 2011; Wang et al. 2013). We wanted to use a stringent approach to identify a high-confidence set of lncRNAs, so protein-coding signal from even a single method eliminated a transcript from contention as a putative lncRNA. After filtering and inclusion of previously annotated lncRNAs, we identified a set of 1,771 intergenic lncRNA transcripts at 1,589 loci in *D. pseudoobscura* (fig. 1*A*).

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 approach (fig. 1*B*). This is not surprising, as each approach has its own particular strengths and weaknesses. We identified 308 protein-coding loci (154 uniquely) through protein databases that contain predominantly long and conserved peptides (Camacho et al. 2009). CPAT and RNAcode, however, can identify proteins with no previous annotation (Washietl et al. 2011; Wang et al. 2013). 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; thus, low coverage locus models are likely to be incomplete. Even so, CPAT offers perhaps the best means to identify lineage-specific protein-coding transcripts, and we identified 233 protein-coding loci (76 uniquely) using CPAT.

We identified 777 protein-coding loci (621 uniquely) using RNAcode, by far the largest number of any approach. 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 like *tarsal-less* but does require a high quality multiple sequence alignment (Washietl et al. 2011). The calculated specificities for RNAcode using the annotated short noncoding RNA loci are high using both the 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 existing protein databases, and may also code for short peptides or have incomplete transcript models, avoiding detection via CPAT. Like lncRNAs, the number of short peptides in the proteome is proving to be larger than previously thought, and short peptide-producing loci have been misidentified as noncoding multiple times (Kondo et al. 2007; Hanyu-Nakamura et al. 2008; Andrews and Rothnagel 2014). As ORF-conservation identification programs like RNAcode and PhyloCSF can reliably identify short ORFs, we recommend including these tools whenever possible in lncRNA identification efforts (Lin et al. 2011; Washietl et al. 2011).

Recent updates to the *D. pseudoobscura* genome annotations (FlyBase r3.03) include more extensive lncRNA annotations generated using the NCBI Gnomon pipeline (Souvorov et al. 2010; St Pierre et al. 2014). While there is some overlap between Gnomon-identified lncRNAs and lncRNAs identified here, we also find evidence of protein-coding ability in almost 200 Gnomon-annotated lncRNA loci. The Gnomon pipeline is likely not as good at identifying short, conserved coding regions as RNAcode, and whether these loci are truly noncoding remains unclear.

Universal and Unique Features of *D. pseudoobscura* lncRNAs

Large sets of lncRNAs have now been described in a number of eukaryotic species (Ulitsky and Bartel 2013; Kapusta and Feschotte 2014). The *D. pseudoobscura* lncRNAs that we describe here display a number of features that are typical of lncRNAs in other systems (fig. 2). While longer than “classic” noncoding RNAs, the *D. pseudoobscura* lncRNAs, on the whole, are shorter than protein-coding transcripts (Derrien et al. 2012; Pauli et al. 2012; Young et al. 2012; Li et al. 2014). They tend to have fewer exons, and alternative splicing is rare (Derrien et al. 2012; Pauli et al. 2012; Young et al. 2012; Li et al. 2014). lncRNA exonic sequence has lower GC content and contains higher proportions of simple sequence repeats and low-complexity sequence (Niazi and Valadkhan 2012). lncRNAs tend to be expressed at lower levels than protein-coding genes (Derrien et al. 2012; Pauli et al. 2012; Young et al. 2012; Necsulea et al. 2014). There is still little consensus on how, or even if, the majority of lncRNAs function. Interestingly, the common features of lncRNAs across diverse taxa suggest that there are distinct forces that drive lncRNA evolution, even if they are primarily derived from the lack of amino acid coding constraint.

That said, most lncRNA studies have been performed in vertebrates, and generalized lncRNA properties observed in vertebrates may not hold for all eukaryotes. For example, TE sequences are found in most lncRNAs in several vertebrate species with high genomic TE content, and TEs are hypothesized to be major drivers of lncRNA evolution and function (Kapusta et al. 2013; Johnson and Guigo 2014). Even though

TEs are detectable at slightly higher levels in lncRNA sequence than coding sequence, we still find relatively minimal TE contributions to lncRNA sequence in the low-TE content *D. pseudoobscura* genome (fig. 2E and *F*). We do point out that 2,616 (49.7%) of novel intergenic transcripts map to the TE-rich "Unknown" genomic contigs and are not considered in this study, leaving open the possibility that the true number of TE-associated lncRNAs in *D. pseudoobscura* is higher than presented here. Further exploration of TE contributions to lncRNAs in other low-TE genomes will reveal whether TEs are fundamental to eukaryotic lncRNA biology.

Expression Dynamics of lncRNAs

The overall expression properties of *D. pseudoobscura* lncRNAs are not identical to those observed for protein-coding genes. *D. pseudoobscura* lncRNAs tend to be more narrowly expressed than protein-coding genes (mean 2.70 developmental stages vs. 6.34), with few expressed in all eight developmental stages in both sexes. Both gene classes show similar developmental trends, with the lowest numbers expressed in the 1st-instar larvae and adult females and the highest numbers expressed in adult males (fig. 3). That said, a larger proportion of lncRNAs display developmental variability, with the number of expressed lncRNAs in males quadrupling over developmental time. Only a minority of lncRNAs are expressed in any female stage, while a majority of protein-coding genes are expressed. This, along with the significant over- or underrepresentation of lncRNAs in specific developmental clusters, suggests that lncRNAs may be deployed to varying degrees in different biological processes (fig. 4). Associated GO annotations may provide useful hypotheses for further functional investigations. Interestingly, while most functionally characterized lncRNAs in *D. melanogaster* have neural functions, nervous system-related GO terms are associated with multiple clusters with lncRNA underrepresentation (Li and Liu 2014)

Previous studies in *Drosophila* show extensive sex-biased gene expression for protein-coding genes, and changes in sex-bias during development largely follow the differentiation and proliferation of the gonads (Parisi et al. 2004; Jiang and Machado 2009; Abdilleh 2014). Similarly, the sex-specific developmental trends of lncRNAs can largely be explained by sex-biased expression in the gonads, both of which undergo proliferation by the 3rd-instar larval stage (fig. 5A and *B*) (King 1970; Bate and Martinez Arias 1993; Hartenstein 1993; Parisi et al. 2004). Tissue heterogeneity in adult females, with large mature ovaries containing abundant RNA transcribed from comparatively small numbers of loci, likely underlies the seeming reduction in the number of expressed genes in adult females. Thus, the severe decline in the number of expressed lncRNAs in adult females is a reflection of the general reduction in the number of genes expressed in the ovaries. The opposite trend is seen in males, where the small but transcriptionally active testes contribute much less to the total RNA output of the whole body. Spermatogenesis is understood to be underway by the late 3rd-instar stage, with the transcript-rich primary spermatocytes already present, and the consistent increase of lncRNA expression through development can largely be attributed to the development of the testes (Bate and Martinez Arias 1993; Hartenstein 1993).

The observed increases in tissue- or developmental stage-specificity of *D. melanogaster* lncRNAs versus protein-coding genes are mirrored in *D. pseudoobscura* (Young et al. 2012; Brown et al. 2014). The *D. melanogaster* modENCODE data only had sex-specific samples in adults, but the trend of overall increased lncRNA expression in adult males is consistent in both species (Graveley et al. 2011; Young et al. 2012). The observation that *D. pseudoobscura* lncRNAs are overrepresented in a non-sex-biased pupae-enriched cluster, however, is not seen in the previously published *D. melanogaster* data (Young et al. 2012). In fact, the pupal stages show the lowest lncRNA expression levels of the four major life cycle stages in *D. melanogaster* (Young et al. 2012). Key differences in methods of lncRNA identification and expression analyses might explain this difference, but studies in vertebrates have also shown evidence of high volatility in lncRNA expression evolution (Necsulea et al. 2014).

Sex-Biased lncRNAs Are Unevenly Distributed in the Genome

By integrating sex-bias data from all developmental stages and adult tissues, we were able to assign each gene an overall sex-bias classification (fig. 5C). Not surprisingly, a higher proportion of lncRNA loci (63.0%) were designated as male-biased as compared with protein-coding genes (38.3%), and a lower proportion of lncRNA loci (8.1% vs. 33.5%) were designated as female-biased. Previous studies in *Drosophila* and other species have shown that sex-biased genes are often unequally distributed between the sex chromosome and the autosomes (Parisi et al. 2003; Khil et al. 2004; Sturgill et al. 2007; Vibranovski et al. 2009; Bachtrog et al. 2010; Meisel et al. 2012; Gao et al. 2014). In *Drosophila*, both male-biased protein-coding genes and intergenic noncoding RNAs (though not specifically long intergenic ncRNAs) have been shown to be underrepresented on the X chromosome.

Different selection-based models have been posited to explain these observations. Under the meiotic sex chromosome inactivation model, a precociously silenced X chromosome during meiosis favors the buildup of advantageous testes-expressed alleles on the autosomes (Vibranovski et al. 2009). Under the dosage compensation model, the male X 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-biased genes, and selection favors the movement of beneficial alleles to the autosomes (Bachtrog et al. 2010). Neither of these models, however, adequately addresses unequal distributions of female-biased genes. With two copies of the X in *Drosophila* females and just one copy in males, the X spends relatively more time in females, encouraging the accumulation of advantageous dominant female-biased alleles on the X (Rice 1984; Charlesworth et al. 1987). Likewise, the reduced time that the X spends in males favors the accumulation of advantageous male-biased alleles on the autosomes.

In *D. pseudoobscura*, we observe a significant depletion of male-biased lncRNA loci on the X, consistent with patterns seen for protein-coding loci, and a significant enrichment of female-biased lncRNAs in the X chromosome that is not seen for protein-coding loci (fig. 6*A*). These patterns are consistent with all three previously proposed selection-based models and may be interpreted as indirect evidence of functional significance of lncRNAs as a whole, though we lack sufficient power in this dataset to offer further support in favor of a particular model. Because this evidence is based purely on expression data, this evidence cannot be confounded by unannotated sequence features and neatly complements previous studies that show evidence of selection on lncRNA exonic sequence in *D. melanogaster* (Young et al. 2012; Haerty and Ponting 2013). While correlations between expression profiles of lncRNAs and their nearest protein-coding loci are higher than would be expected by chance, similarly high correlations between neighboring protein-coding loci impede our ability to distinguish between spurious transcription and active regulation of lncRNAs (fig. 6D).

Several studies of *D. melanogaster* have shown that, in contrast to patterns observed for male-biased genes in general, testis-specific genes are not significantly depleted from the X (Meiklejohn and Presgraves 2012; Meisel et al. 2012; Gao et al. 2014). This observation has been rationalized as a consequence of gene age, with an excess of young genes in the testes that have not existed long enough for selection to act upon (Gao et al. 2014). Consistent with those observations, we find that neither testis-specific lncRNAs nor protein-coding genes are significantly underrepresented on the X in *D. pseudoobscura*, while male-biased and non-testis-specific genes of both classes are (fig. 6*B*).

Gao et al. (2014) showed that young testis-specific protein-coding genes are more likely to be X-linked in *D. melanogaster* than older testis-specific coding genes, and this could reasonably explain the lack of X-chromosome demasculinization in *D. pseudoobscura* testis-specific lncRNAs and protein-coding genes. The testes provide an ideal environment for the origination of new genes, with an open chromatin environment that permits broad transcription of the genome and selective pressures like sexual conflict, sperm competition, and germline pathogens that can result in the rapid evolution of new genes (Kaessmann et al. 2009; Kaessmann 2010). Indeed, many newly-evolved genes are expressed in the testes (Reinhardt et al. 2013; Zhao et al. 2014). Using simple BLASTn searches, we show that *D. pseudoobscura* testis-specific protein-coding gene sequences are far less likely to be present in more divergent *Drosophila* genomes than male-biased but non-testis-specific coding sequences and even random intergenic sequences, suggesting that testis-specific protein-coding genes indeed tend to be younger (Gao et al. 2014). However, testis-specific lncRNA sequences have similar BLASTn hit rates as male-biased non-testis-specific lncRNA sequences in *D. affinis* and *D. melanogaster*, both of which are similar to BLASTn hit rates for random intergenic sequence (fig. 6C). We posit two potential explanations: (1) as opposed to protein-coding genes, many of which contain *de novo* sequence, new lncRNA transcription tends to originate from existing genomic sequence; or (2) testis-specific lncRNAs are not more likely to be new genes at all, and the absence of their underrepresentation on the X chromosome cannot be explained by invoking a lack of time for selection to act.

In the absence of compelling evidence of selection, we cannot reject the hypothesis that many of these testis-specific lncRNAs, which comprise a near-majority (45.8%) of all detected lncRNAs in *D. pseudoobscura*, are not under selection and are likely to be spurious or "junk" transcription. Interestingly, of all gene classes, the highest correlations in expression are seen between testis-specific lncRNAs and their nearest protein-coding neighbors (fig. 6D). Whether this is a consequence of spurious transcription due to shared chromatin environment or evidence of active and elevated *cis* regulation of or by lncRNAs remains unclear.

Putative lncRNA Homolog Show Conservation of Developmental Expression Profiles

Our comparative transcriptomic analyses have revealed the first large set of putative lncRNA homologs within the *Drosophila* genus, with 134 putative lncRNA homologs identified between *D. melanogaster* and *D. pseudoobscura*. Eighty of these putative homologs were identified directly through local alignment of transcript sequences. However, lncRNA sequence is often poorly conserved over large areas of a transcript with only short domains of conservation (Brockdorff et al. 1992; He et al. 2011; Ulitsky et al. 2011; Diederichs 2014; Chen et al. 2016). By searching for coordinate overlap via a whole genome alignment, a method less reliant on primary sequence conservation, we were able to identify an additional 54 putative homologs.

Sufficient developmental RNA-Seq data in both species enable comparison of expression dynamics of putative homologs. Levels of expression correlation for putative lncRNA homologs mirror those found for protein-coding orthologs and are significantly higher than randomly-associated lncRNAs from both species (fig. 7C). Further, expression correlations are locus-specific; expression profiles of the *D. pseudoobscura* lncRNAs show little correlation with genes neighboring their putative *D. melanogaster* homologs. Thus, the high correlations between putative lncRNA homologs cannot be explained as a consequence of a shared open chromatin environment (Struhl 2007; Ulitsky and Bartel 2013). While we cannot exclude that some of these putative homologs are actually independently evolved transcriptional events, we consider this strong evidence in support of homology of these lncRNAs. Potentially incomplete lncRNA gene models in *D. pseudoobscura*, largely due to a lack of transcript end sequencing (e.g., 5′ CAGE, RNA-PET), complicate further comparisons of transcript start and stop sites and promoter characteristics between the two species that would also further support homolog classifications.

The statistically random distributions of testis-specific lncRNAs between the X and the autosomes led us to argue that they are more likely to be the result of spurious transcription than female-biased or more broadly-expressed male-biased lncRNAs. If this is the case, we expect to see a lower fraction of testis-specific lncRNAs in the set of putative lncRNA homologs than in the larger set of all lncRNAs in *D. pseudoobscura*. We observe this trend; 56.9% of all lncRNAs in *D. pseudoobscura* are testis-specific, while only 48.6% of putative homologs are the same. Similarly, 62.3% of all lncRNAs that are included in the fuzzy c-means cluster analysis are found in clusters with male-bias in development, but only 54.0% of putative lncRNA homologs are found in these clusters. These differences are not statistically significant, although the low numbers of putative lncRNA homologs may provide low power for detecting significance.

Conclusions

The extent of the biological relevance of lncRNAs remains unresolved. Using a comparative transcriptomic approach, we showed commonalities in lncRNA expression dynamics, like overrepresentation in males, in two species of *Drosophila*. By observing the genomic location of sex-biased lncRNAs, we also were able to construct testable hypotheses of the biological impact of lncRNAs in particular tissues. lncRNAs expressed in somatic tissues and the ovaries show indirect evidence of selection to optimize genomic location. This evidence was not observed, however, for the numerous testis-specific lncRNAs, and we cannot reject the possibility that these are more likely to be spurious transcriptional events. To satisfyingly explore not only the biological function of lncRNAs but also how that function influences their evolution, we need to identify conserved lncRNAs in multiple genetically amenable species. With catalogs of natural variations and the ability to induce disruptive mutations in *Drosophila*, we have the opportunity to tease apart what features are critical for the core function of lncRNAs. The lncRNAs identified here display the strongest evidence for lncRNA homology within the genus and are obvious candidates for further investigations into lncRNA biology.

Supplementary Material

Supplementary file, tables S3 and S7–S13, and figures S2 and S4–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Suwei Zhao and the University of Maryland's Institute for Bioscience and Biotechnology Research for sequencing, Yan Wang and Zi-Feng Jiang for generating proteomics data from *D. pseudoobscura* and hybrid testes, Josie Reinhardt for providing custom perl scripts to identify ORFs, Liguo Wang for providing a custom TG-ROC R script, Kawther Abdilleh and Carlo Artieri for assistance with bioinformatics analyses, Ryan Hardy and Mohamed Noor for extracting subalignments from Pseudobase, and Marta Wayne and two anonymous reviewers for helpful comments on the manuscript. This work was supported by National Science Foundation grant MCB-1330766 to C.A.M.

Literature Cited

Abdilleh KA. 2014. Patterns of sex-biased gene expression and gene pathway evolution in *Drosophila*. PhD Dissertation, Department of Biology, University of Maryland. http://drum.lib.umd.edu/handle/1903/15359.

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166–169.

Andrews SJ, Rothnagel JA. 2014. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. 15:193–204.

Bachtrog D, Toda NR, Lockton S. 2010. Dosage compensation and demasculinization of X chromosomes in *Drosophila*. Curr Biol. 20:1476–1481.

Bassett AR, et al. 2014. Considerations when investigating lncRNA function in vivo. Elife 3:e03058.

Bate M, Martinez Arias A. 1993. The development of Drosophila melanogaster. Plainview, NY: Cold Spring Harbor Laboratory Press.

Brockdorff N, et al. 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 JB, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. Nature 512:393–399.

Cabili MN, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25:1915–1927.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Carvalho AB, Clark AG. 2005. Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. Science 307:108–110.

Celniker SE, et al. 2009. Unlocking the secrets of the genome. Nature 459:927–930.

Charlesworth B, Coyne JA, Barton N. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am Nat. 130:113–146.

Chen J, et al. 2016. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. 17:19.

Derrien T, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22:1775–1789.

Desiere F, et al. 2006. The PeptideAtlas project. Nucleic Acids Res. 34:D655–D658.

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

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 University Press.

Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Gao G, et al. 2014. A long-term demasculinization of X-linked intergenic noncoding RNAs in *Drosophila melanogaster*. Genome Res. 24:629–638.

Ghanbarian AT, Hurst LD. 2015. Neighboring genes show correlated evolution in gene expression. Mol Biol Evol. 32:1748–1766.

Graveley BR, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473–479.

Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome Biol. 14:R49.

Hanyu-Nakamura K, Sonobe-Nojima H, Tanigawa A, Lasko P, Nakamura A. 2008. *Drosophila* Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. Nature 451:730–733.

Hartenstein V. 1993. Atlas of Drosophila development. Plainview, NY: Cold Spring Harbor Laboratory Press.

He S, Liu S, Zhu H. 2011. The sequence, structure and evolutionary features of HOTAIR in mammals. BMC Evol Biol. 11:102.

Hezroni H, et al. 2015. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 11:1110–1122.

Hinrichs AS, et al. 2006. The UCSC genome browser database: update 2006. Nucleic Acids Res. 34:D590–D598.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 12:41–51.

Inagaki S, et al. 2005. Identification and expression analysis of putative mRNA-like non-coding RNA in *Drosophila*. Genes Cells 10:1163–1173.

Jiang ZF, Croshaw DA, Wang Y, Hey J, Machado CA. 2011. Enrichment of mRNA-like noncoding RNAs in the divergence of *Drosophila* males. Mol Biol Evol. 28:1339–1348.

Jiang ZF, Machado CA. 2009. Evolution of sex-dependent gene expression in three recently diverged species of *Drosophila*. Genetics 183:1175–1185.

Johnson R, Guigo R. 2014. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA 20:959–976.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 10:19–31.

Kapusta A, et al. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 9:e1003470.

Kapusta A, Feschotte C. 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. 30:439–452.

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet. 36:642–646.

Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

King RC. 1970. Ovarian development in Drosophila melanogaster. New York: Academic Press.

Kondo T, et al. 2007. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol. 9:660–665.

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. 36:D271–D275.

Kuhn RM, et al. 2007. The UCSC genome browser database: update 2007. Nucleic Acids Res. 35:D668–D673.

Kumar L, Futschik ME. 2007. Mfuzz: a software package for soft clustering of microarray data. Bioinformation 2:5–7.

Kung JT, Colognori D, Lee JT. 2013. Long noncoding RNAs: past, present, and future. Genetics 193:651–669.

Kutter C, et al. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 8:e1002841.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15:R29.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Li H, et al. 2009. The sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li L, et al. 2014. Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. 15:R40.

Li M, Liu L. 2014. Neural functions of long noncoding RNAs in *Drosophila*. J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 201(9):921–926.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27:i275–i282.

Lucas A. 2014. amap: another Multidimensional Analysis Package. Version R package version 0.8-12. http://mulcyber.toulouse.inra.fr/projects/amap/.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. 10:R124.

Meiklejohn CD, Presgraves DC. 2012. Little evidence for demasculinization of the *Drosophila* X chromosome among genes expressed in the male germline. Genome Biol Evol. 4:1007–1016.

Meisel RP, Malone JH, Clark AG. 2012. Disentangling the relationship between sex-biased gene expression and X-linkage. Genome Res. 22:1255–1265.

Mount SM, Gotea V, Lin CF, Hernandez K, Makalowski W. 2007. Spliceosomal small nuclear RNA genes in 11 insect genomes. RNA. 13:5–14.

Necsulea A, et al. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505:635–640.

Niazi F, 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. http://pseudobase.biology.duke.edu/.

Parisi M, et al. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science 299:697–700.

Parisi M, et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila* melanogaster adults. Genome Biol. 5:R40.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619.

Pauli A, et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. 22:577–591.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. 17:556–565.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genet. 10:e1004525.

Reinhardt JA, et al. 2013. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 9:e1003860.

Rice W. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution. 38:735–742.

Richards S, et al. 2005. Comparative genome sequencing of *Drosophila* pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. 15:1–18.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Schaeffer SW, 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.

Schuler A, Ghanbarian AT, Hurst LD. 2014. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. Mol Biol Evol. 31:3164–3183.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5:e1000495.

Smit AFA, Hubley R, Green PJ. 2014. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Smyth GK. 2005. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irazarry R, Huber W, editors. Bioinformatics and computational biology solutions using R and bioconductor. New York: Springer. P. 397–420.

Souvorov A, et al. 2010. Gnomon—NCBI eukaryotic gene prediction tool. Bethesda, MD: National Center for Biotechnology Information.

St Pierre SE, Ponting L, Stefancsik R McQuilton P., FlyBas. 2014. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res. 42:D780–D788.

Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol. 14:103–105.

Sturgill D, Zhang Y, Parisi M, Oliver B. 2007. Demasculinization of X chromosomes in the *Drosophila* genus. Nature 450:238–241.

Tabas-Madrid D, Nogales-Cadenas R, Pascual-Montano A. 2012. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. 40:W478–W483.

Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.

Tupy JL, et al. 2005. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 102:5495–5500.

Ulitsky I, et al. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147:1537–1550.

Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. Cell 154:26–46.

Vibranovski MD, Lopes HF, Karr TL, Long M. 2009. Stage-specific expression profiling of *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet. 5:e1000731.

Wang L, et al. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 41:e74.

Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337:1675–1678.

Warnes GR, et al. 2014. gplots: Various R programming tools for plotting data. Version R package version 2.15.0.

Washietl S, et al. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17:578–594.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res. 24:616–628.

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41:D358–D365.

Young RS, et al. 2012. Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. Genome Biol Evol. 4:427–442.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. Science 343:769–772.

Associate editor: Marta Wayne