Title

Revised annotations, sex-biased expression, and lineage-specific genes in the Drosophila melanogaster group.

Permalink

https://escholarship.org/uc/item/21h0t710

Journal

G3 (Bethesda, Md.), 4(12)

ISSN

2160-1836

Authors

Rogers, Rebekah L
Shao, Ling
Sanjak, Jaleal S
et al.

Publication Date

2014-10-01

DOI

10.1534/g3.114.013532

Peer reviewed

Revised Annotations, Sex-Biased Expression, and Lineage-Specific Genes in the *Drosophila melanogaster* Group

Rebekah L. Rogers,*[,1] Ling Shao,* Jaleal S. Sanjak,* Peter Andolfatto,[†] and Kevin R. Thornton*
*Ecology and Evolutionary Biology, University of California, Irvine, Irvine, California 92697, and [†]Ecology and Evolutionary Biology and the Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544

ABSTRACT Here, we provide revised gene models for *D. ananassae*, *D. yakuba*, and *D. simulans*, which include untranslated regions and empirically verified intron-exon boundaries, as well as ortholog groups identified using a fuzzy reciprocal-best-hit blast comparison. Using these revised annotations, we perform differential expression testing using the cufflinks suite to provide a broad overview of differential expression between reproductive tissues and the carcass. We identify thousands of genes that are differentially expressed across tissues in *D. yakuba* and *D. simulans*, with roughly 60% agreement in expression patterns of orthologs in *D. yakuba* and *D. simulans*. We identify several cases of putative polycistronic transcripts, pointing to a combination of transcriptional read-through in the genome as well as putative gene fusion and fission events across taxa. We furthermore identify hundreds of lineage specific genes in each species with no blast hits among transcripts of any other *Drosophila* species, which are candidates for neofunctionalized proteins and a potential source of genetic novelty.

KEYWORDS
gene annotation
sex-specific gene expression
polycistronic genes
gene fusion
D. simulans
D. yakuba
D. ananassae

Accurate models of gene structure including untranslated regions (UTRs), intron-exon boundaries, as well as coding sequences are essential for proper interpretation of molecular genetics (Fire *et al.* 1998, Jinek *et al.* 2012), demographic inference (Halligan and Keightley 2006, Parsch *et al.* 2010, Clemente and Vogl 2012), tests of selection (Mcdonald and Kreitman 1991), and comparative genomics (Chen *et al.* 2014). The *Drosophila* offer an excellent model for comparative genomics, with high-quality sequenced genomes for 12 species(Drosophila Twelve Genomes Consortium 2007) as well as draft genomes for an additional eight species (Chen *et al.* 2014) spanning a total of 63 million years (Tamura *et al.* 2004). Previous gene models provided for the 12 *Drosophila* genomes focused on gene prediction with the aid of homology to establish putative annotations of coding sequences across taxa with 15,000−16,000 genes for most species (Drosophila Twelve Genomes Consortium 2007). These gene models produce reliable annotations for conserved genes as well as genes that are present in multiple species but will lack lineage specific genes where sources of genetic novelty add to the genome, will misidentify rapidly evolving sequences, and will neglect isoforms that may offer alternative molecular functions.

Recent work has validated gene models in *D. melanogaster* through cross-species comparisons (Chen *et al.* 2014). Although aligned CDS sequences often display patterns of expression consistent with conservation, gene structure varies across taxa at UTRs, introns, and noncoding DNA (Chen *et al.* 2014). Furthermore, any gene families and functional classes subject to rapid evolution in gene structure that is unlikely to be reflected in homology-based annotations (Wasbrough *et al.* 2010) and *de novo* genes also will be absent, despite their role in developing functional diversity (Zhao *et al.* 2014).

Beyond the evolution of gene structure, expression patterns of genes across tissues is expected to influence their evolutionary constraints and rates of sequence evolution (Van Dyken and Wade 2010) as well as their functional roles within the organism. Sex-biased genes are often rapidly evolving (Ranz *et al.* 2003; Ellegren and Parsch 2007), especially among male reproductive proteins (Haerty *et al.* 2007; Zhang *et al.* 2007; Meisel 2011). Moreover, sex-specific expression patterns can influence the extent to which genes contribute to sexual antagonism, influencing their role in the evolution of sexual

doi: 10.1534/g3.114.013532
Manuscript received July 31, 2014; accepted for publication September 30, 2014; published Early Online October 1, 2014.

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.114.013532/-/DC1
[1]Corresponding author: Department of Integrative Biology, University of California, Berkeley, 4151 Valley Life Sciences Building, Berkeley, CA 94720. Email: bekah@berkeley.edu

dimorphism Mank *et al.* (2008). Correctly identifying sex-specific and tissue specific expression is therefore essential for studying the molecular and evolutionary impacts of genes within species. Changes in tissue specific or sex-specific expression may point to rapid evolution of gene functions and candidate loci to search for signals of evolutionary change and differential selective effects between the sexes.

Here, we describe RNA-seq based gene annotations for *Drosophila* outgroup species *D. simulans*, *D. yakuba*, and *D. ananassae* based on whole transcriptome sequencing of male and female adult tissues. These revised gene models capture hundreds of lineage-specific genes on major chromosomal arms in *D. yakuba* and *D. simulans*. We also describe 5′ and 3′ UTRs, and intron-exon structure for genes throughout the *D. melanogaster* group. Finally, we describe sex-biased expression across species, identifying thousands of genes that are differentially expressed across tissues. These revised gene models as well as results of sex-biased and tissue-biased differential expression testing should serve as a resource for the *Drosophila* evolutionary and molecular genetics community interested in evolution, conservation, and gene expression.

MATERIALS AND METHODS

Sample preparation

Fly stocks were incubated under controlled conditions at 25° and ~40% humidity. Virgin flies were collected within 2 hr of eclosion, then aged 2−5 d posteclosion before dissection. We dissected samples in isotonic Ringers solution using female ovaries and headless gonadectomized carcass from two adult flies as well as testes plus glands and male headless gonadectomized carcass for four adult flies for each sample RNA prep. We collected three biological replicates of the *D. yakuba* reference, three biological replicates of the *D. simulans* w^{501} reference, and one replicate of the *D. ananassae* reference and one replicate of the *D. melanogaster* reference (stock numbers in Table 1). Samples were flash frozen in liquid nitrogen immediately after dissection, and and stored in 0.2 mL of Trizol at −80°. All samples were homogenized in 0.5 mL of Trizol Reagent (Invitrogen) with plastic pestle in a 1.5-mL tube, mixed with 0.1 mL of chloroform, and centrifuged 12,000*g* for 15 min at 4°C, as Trizol RNA extraction protocol. The RNAs in the supernatant about 0.4 mL were then collected and purified with Direct-Zol RNA MiniPrep Kit (Zymo), following the standard protocol. The total RNAs were eluted in 65 μL of RNase-Free H_2O. Approximately 1 μg of purified RNAs were treated with 2 μL of Turbo DNase (Invitrogen) in a 65-μL reaction, incubated for 15 min at room temperature with gentle shaking. These RNAs were further purified with RNA Clean and Concentrator-5 (Zymo). One extra wash with fresh 80% ethanol after the final wash step was added into the original protocol. The treated RNAs were eluted with 15 μL of RNAse-Free H_2O, and stored at −80°.

The amplified cDNAs were prepared from 100 ng of DNase treated RNA with Ovation RNA-Seq System V2 (Nugen) and modified protocol. The preparations followed the protocol to the step of SPIA Amplification (Single Primer Isothermal Amplification). The amplified cDNAs were first purified with Purelink PCR Purification Kit (Invitrogen, HC Binding Buffer) and eluted in 100 μL of EB (Invitrogen). These cDNAs were purified again to 25 μL of Buffer EB with DNA Clean and Concentrator -5 Kit (Zymo) for Nextera library preparation. About 43 ng of cDNA was used to construct libraries with Nextera DNA Sample Preparation Kit (Illumina) and modified protocol. After tagmentation, a Purelink PCR Purification Kit with HC Binding Buffer was used for purification and eluted with 30μL of EB or H_2O. The products (libraries) of final polymerase chain reaction (PCR) amplification were purified with DNA Clean and Concentractor-5 and eluted in 20 μL of EB. The average library lengths of approximately 500 bp were estimated from profiles on a Bioanalyzer 2100 (Agilent) with DNA HS Assay. All libraries were normalized to 2−10 nM based on the real-time PCR method with Kapa Library Quant Kits (Kapa Biosystems). The qualities and quantities of these RNAs, cDNAs and final libraries were measured from Bioanalyzer with the Qubit RNA HS or DNA HS Assay Kit (Invitrogen) with RNA HS or DNA HS Reagents, respectively. Samples were barcoded and sequenced using 76-bp reads in 4-plex on an Illumina HiSequation 2500 using standard Illumina barcodes, resulting in high coverage (Figure S1, Figure S2, Figure S3, Figure S4, Figure S5, Figure S6). Coverage in RNA-seq data is dependent on expression level, but the majority of sites have coverage less than 50 reads. One replicate of female tissues was sequenced using single end reads. All other libraries were sequenced using paired end reads and dual indexing. Samples were sequenced as they became available, across the equivalent of one eight lane flow cell for the 32 samples in total.

Table 1 Fly stocks used for RNA-seq

Species	Strain
D. melanogaster	14021-0231.36
D. simulans	w^{501}
D. yakuba	14021-0261.01
D. ananassae	14024-0371.13

Trinity and Augustus gene annotation

RNA sequencing data for ovary, female carcass, testes, and male carcass of *D. yakuba*, *D. simulans*, and *D. ananassae* were concatenated into a single fastq file and digitally normalized to remove excess redundant reads for highly expressed transcripts. This step results in tractable runtimes with no loss of information in transcript annotations. We ran Trinity (http://trinityrnaseq.sourceforge.net/; Grabherr *et al.* 2011; Haas *et al.* 2013). For a single sample:

1. FASTQ files for left and right reads were concatenated.
2. The concatanated files were subject to "digital normalization" using the following command in the Trinity package:

```
normalize_by_kmer_coverage.pl --seqType fa --JM 100G --max_cov 30 --left left.fa
--right right.fa --pairs_together --PARALLEL_STATS --JELLY_CPU $CORES
```

where $CORES is how many CPU we had available

1. The resulting normalized left and right fastq files where then used in the following command:

```
Trinity.pl --seqType fa --bflyHeapSpaceInit 1G --bflyHeapSpaceMax 8G --JM 7G --left $LEFTFILE --right $RIGHTFILE --output trinity_output
```

where the variables are the normalized fastq files and the number of cores available.

1. Further detail is at https://github.com/ThorntonLab/annotation_methods

The resulting annotations were used as input in the Augustus v2.5.5 gene prediction software http://bioinf.uni-greifswald.de/augustus/ with command line options

```
--species=fly --hintsfile=$INFILE_BASE.hints.E.gff --extrinsicCfgFile=augustus.2.5.5/config/extrinsic/extrinsic.ME.cfg $INFILE --gff3=on --uniqueGeneId=true > $INFILE_BASE.gff3
```

Alignment and annotation

We matched Augustus gene models to previous annotations from FlyBase r1.3 for *D. ananassae*, and *D. yakuba*, or to the *D. simulans* w^{501} reference annotations. CDS sequences were required to physically overlap with the location of a current FlyBase or w^{501} gene model and were required to have matches to 85% or more CDS sequence with 90% or greater amino acid similarity in an all-by-all BLASTp of translated sequences at a cutoff of $E \leq 10^{-10}$ with low-complexity filters turned off (-F F). We mapped RNA-seq data to known gene models annotated in FlyBase, *D. yakuba* r1.3, *D. ananassae* r1.3, *D. melanogaster* r.5.45, and the *D. simulans* w^{501} gene models annotated by Hu *et al.* (2012). GFF Files were reformatted using gffread from Cufflinks suite. Sequences were mapped to the genome using Tophat v.2.0.6 (Trapnell *et al.* 2009, Kim *et al.* 2013) and Bowtie2 v.2.0.2 (Langmead *et al.* 2009), using reference annotations as a guide, with no attempt to identify novel transcripts using reads which fell outside reference annotations (-G) and all other parameters set to default. We used Cufflinks 2.0.2 (Trapnell *et al.* 2012) to calculate expression levels across genes and transcripts, normalizing expression by the upper quantile (-N) and ignoring reads which fall outside known gene models (-G) with all other options set to default. Orphaned gene models from FlyBase or from Hu *et al.* (2012) which had fragments per kilo bases of exons for per million mapped read (FPKM) ≥2 but which had no assembled gene model match from Augustus were included in the final annotations used for differential expression testing. Some annotations contain polycistronic transcripts encompassing multiple independent open reading frames. A portion of these polycistronic transcripts may reflect only low-level polycistronic transcription, rather than polycistronic transcripts serving as the dominant isoforms but the rate of polycistronic transcription cannot be readily determined with available data. For genes with polycistronic transcripts but no 1:1 transcript match with FlyBase gene models, we included annotations for both the polycistronic Augustus gene models and the gene models from FlyBase supporting independent transcripts.

Sex-bias and tissue bias

The union of gene models from Augustus and orphaned gene models from FlyBase were combined into a single GFF containing transcript and CDS annotations for each species. We then re-mapped RNA-seq reads to gene models in the reannotated GFF for *D. yakuba*, *D. ananassae*, and *D. simulans*, as well as unmodified gene models for *D. melanogaster* r.5.45 with Tophat and performed differential expression testing at a false discovery rate ≤0.1 normalizing expression by the upper quantile (-N) and ignoring reads that fall outside known gene models (-G) with all other options set to default using the Cufflinks suite according to the same criteria described previously. We compared female carcass with female ovaries, male carcass with male testes, female carcass with male carcass, and female ovaries with male testes for each species, grouping replicates for reference genomes.

Orthologs and lineage specific genes

Orthologs were identified using fuzzy reciprocal best hit BLASTp comparisons of all translations across reference genomes for gene model predictions of *D. ananassae*, *D. yakuba*, *D. simulans*, as well as *D. melanogaster*. Orthologs are similar to those previously used to annotate gene families in *Drosophila* (Drosophila Twelve Genomes Consortium 2007, Hahn *et al.* 2007). Putative orthologs must be putative reciprocal hits of the same rank order, where genes with an E-value within a single log-unit of one another are assigned the same rank, using the best E-Value for a gene with a cutoff of $E \leq 10^{-10}$. Lineage-specific genes were defined as genes with no hit in an all-by-all BLASTp of translations against translations of the other outgroups (*e.g.*, *D. yakuba*, *D. ananassae*, and *D. melanogaster* for *D. simulans*) at a cutoff of $E \leq 10^{-10}$ with low-complexity filters turned off (-F F).

Gene ontology

We used DAVID gene ontology analysis software http://david.abcc.ncifcrf.gov/ to determine whether any functional categories were overrepresented among genes with sex specific or tissue specific expression. Functional data for *D. ananassae*, *D. yakuba*, and *D. simulans* are not readily available in many cases, and thus we identified functional classes in the *D. melanogaster* orthologs as classified in Flybase. Gene ontology clustering threshold was set to Low. Genes with tissue specific expression were based on genes with differential expression from cufflinks for comparisons of female carcass *vs.* female ovaries and male carcass *vs.* male testes at a genomewide false discovery rate ≤0.10, according to cufflinks default settings.

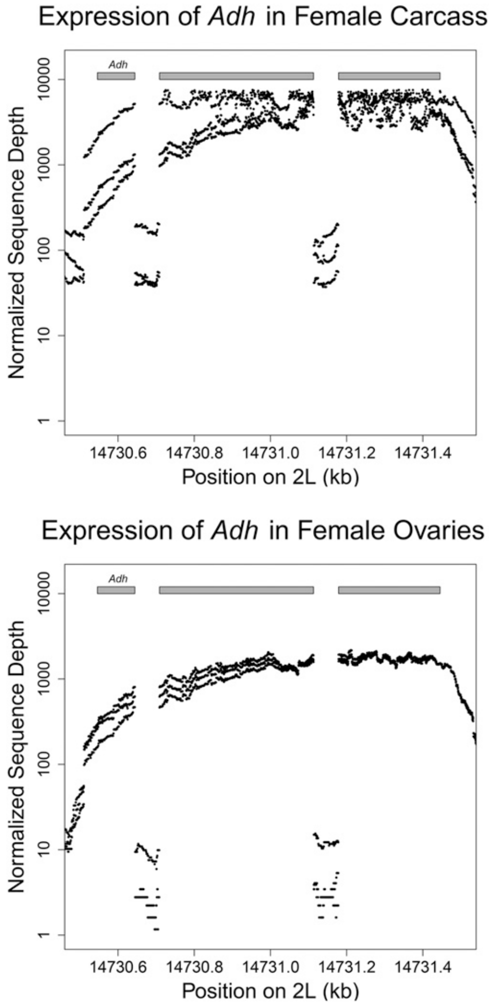

Figure 1 Quantile-normalized coverage for reference strains at the *Adh* locus in *D. yakuba*. Coverage shows clear distinctions between introns and exons, and coverage that spans both 5′ and 3′ UTRs in ovaries and carcass. Low coverage of intron sequence points to partial sequencing of low levels of unprocessed transcripts. Top, carcass; bottom, ovaries.

Data access

Gene annotations, ortholog calls, gene ontology calls, and Cuffdiff differential expression testing output files for all samples are available at http://github.com/ThorntonLab/GFF. RNA-seq−based annotations as well as first=order orthologs in comparison with *D. melanogaster* can be viewed in the UCSC browser on the Thornton Lab public track hub at http://genome.ucsc.edu. Sequencing fastq files were deposited in SRA with accession numbers PRJNA196536, PRJNA193071, PRJNA257286, and PRJNA257287.

RESULTS

Annotations

For moderately to highly expressed genes, we recover gene structure with intron-exon boundaries and UTR sequences for full-length transcripts including novel exons that were previously unannotated based on comparative genomics (Figure 1 and Figure 2). Many genes are part of polycistronic transcripts including one from *D. melanogaster*. We identify 2529 putative polycistronic transcripts in *D. yakuba*, 2379 in *D. ananassae*, and 561 in *D. simulans*. For such genes, we offer gene models from FlyBase as well as fused models from Augustus. The extent to which such transcription of multiple genes is functional as opposed to a stochastic byproduct of transcriptional errors is unclear. We also include FlyBase gene models expressed in the reference genomes with no 1:1 match in gene models from Augustus. The addition of FlyBase gene models results in an additional 4265 annotations in *D. yakuba*, 5367 in *D. ananassae*, and 7419 in *D. simulans*. We identify a total of 22,989 transcripts for 16,278 genes in *D. simulans*, 20,315 transcripts for 17,579 genes in *D. yakuba*, and 22,420 transcripts for 20,580 genes in *D. ananassae* (Table 2). Compared with *D. yakuba*, for *D. ananassae* an additional 1173 FlyBase gene models failed to match sufficiently well with RNA-seq supported gene models and these were added to the annotations, explaining some of the excess in the number of genes and also highlighting the difficulties of annotation through comparative genomics across large phylogenetic distances. In *D. ananassae*, 72% of gene models have RNA-seq data supporting 60% or more gene features (exons, UTRs) compared with 79.4% in *D. yakuba* and 80.1% in *D. simulans* (Table 3).

As defined by a fuzzy reciprocal best-hit blast (Drosophila Twelve Genomes Consortium 2007, Hahn *et al.* 2007) we identify 12,127 genes in *D. melanogaster* with first-order orthologs in *D. simulans*, 11,425 with first-order orthologs in *D. yakuba*, and 11,348 with first-order orthologs in *D. ananassae* (Table 4). The increase in the number of genes with orthologs in *D. simulans* is the product of improved annotations as well as the improved assembly of the w^{501} *D. simulans* reference (Hu *et al.* 2012). We observe a 1:1 concordance for 48% of FlyBase gene models in *D. yakuba* and 46% of FlyBase gene models for *D. ananassae*. These annotations typically include UTR sequences in addition to empirically supported intron-exon and coding sequence boundaries, an improvement over previous gene models from release r1.3 which lack UTRs (Drosophila Twelve Genomes Consortium 2007). We further identify thousands of lineage specific genes in each species of *Drosophila* with no matching gene model in other outgroup species. Although lineage-specific genes identified on minor chromosomes could result from assembly issues, we identify hundreds on major chromosomal arms (Table 5), suggesting that many of these are in fact cases of lineage-specific gene formation.

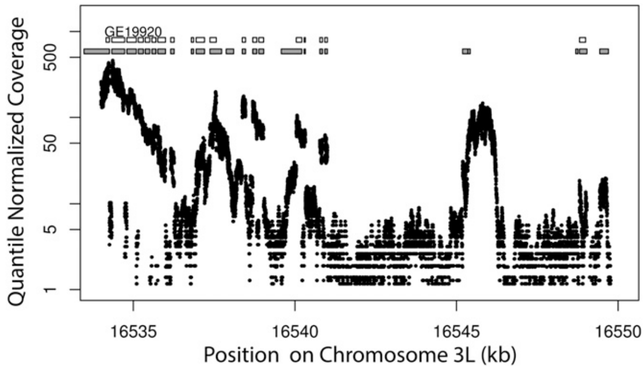

Figure 2 Quantile-normalized RNA-seq data for three replicates of the *D. yakuba* reference with an example of FlyBase gene model (white) and revised gene model (gray).

Table 2 Number of transcripts and genes identified

Species	Previous Release	Transcripts		Genes	
		Revised	FlyBase	Revised	FlyBase
D. simulans	r1.3	18,781	15,415	16,278	15,413
D. yakuba	r1.3	20,239	16,082	17,579	16,077
D. ananassae	r1.3	22,418	15,070	20,580	15,069

Sex-biased and tissue-biased expression

We observe thousands of genes with sex-biased or tissue-biased expression in *D. yakuba* and *D. simulans* but hundreds of genes with sex-biased or tissue-biased expression in *D. melanogaster* and *D. ananassae*, a direct product of increasing power to detect differences in RNA levels with biological replicates. Gene ontology categories overrepresented between ovary and female carcass reflect differences in genes involved in reproduction, chromosome segregation, and DNA synthesis or repair, whereas genes differentially expressed between testes and carcass reflect sperm development, cell division, and energy production (Supporting Information, File S1). We used reciprocal best hit orthologs to identify genes with similar regulation across species, focusing on *D. yakuba* and *D. simulans* where biological replicates increase power for differential expression testing (Table 6). A total of 10,369 genes in *D. yakuba* have reciprocal best-hit orthologs in *D. simulans* and were retained to compare differential expression across the two species. We have collected replicates for both *D. yakuba* and *D. simulans*, contributing to the greater power in differential expression testing. Roughly 60% of genes with tissue biased expression in *D. yakuba* that have a reciprocal best hit ortholog in *D. simulans* exhibit the same tissue-specific bias in *D. simulans*, with marginally greater agreement in genes biased toward the carcass than the reproductive tissues in both males and females (Figure 3). We additionally observe evidence of differential expression in at least one of the four comparisons of male and female germline or somatic tissues for 118 lineage specific genes in *D. ananassae*, 334 in *D. yakuba*, and 222 in *D. simulans* (Table 5), suggesting that they are not solely artifacts of gene annotation software.

Table 3 Percent of revised gene models with ≥60% of features supported by RNA-seq data

Species	Percent Suported
D. ananassae	72.3
D. yakuba	79.4
D. simulans	80.05

Table 4 Genes with a first-order ortholog identified

Genes in	With an Ortholog in	Revised	FlyBase
D. melanogaster	*D. simulans*	12,199	10,705
D. melanogaster	*D. yakuba*	11,472	11,556
D. melanogaster	*D. ananassae*	11,451	10,938
D. simulans	*D. melanogaster*	13,295	–
D. simulans	*D. yakuba*	12,299	–
D. simulans	*D. ananassae*	11,994	–
D. yakuba	*D. melanogaster*	12,868	–
D. yakuba	*D. simulans*	12,831	–
D. yakuba	*D. ananassae*	12,337	–
D. ananassae	*D. melanogaster*	12,897	–
D. ananassae	*D. simulans*	12,612	–
D. ananassae	*D. yakuba*	12,723	–

Table 5 Putative Lineage Specific Genes on Major Chromosomes

Species	Major chromosomes	Total	Diff Exp
D. ananassae	–	2977	118
D. yakuba	230	1340	334
D. simulans	369	1314	222

DISCUSSION

Gene models and ortholog calls

Correct interpretations of gene expression changes and gene family evolution depend on accurate gene models. Among the *Drosophila*, *D. melanogaster* has received the most attention with empirically verified gene annotations through high-throughput expressed sequence tags and RNA-seq data as well as detailed manual or molecular curation of single genes. Until present, gene models for outgroup species offer only CDS sequences, with no information concerning 5′ or 3′ UTRs or alternative isoforms even for well-studied genes such as *Adh*. Establishing more complete gene models based on RNA-seq data will allow us to correctly identify coding and noncoding sequences during functional assays, correctly identify putatively neutral *vs.* non-neutral mutations, and correctly define new mutations including the origins of new genes, expansion of gene families, and gross modification of coding sequences through rearrangement, duplication, and deletion. Moreover, identifying putative isoforms provides a more complete portrait of gene structure and function across species. Here, we provide updated gene annotations based on high coverage RNA-seq data for the reference genomes of three species of *Drosophila*: *D. ananassae*, *D. yakuba*, and *D. simulans*, which should serve as an excellent springboard and initial resource to the *Drosophila* molecular and evolutionary genetic community.

Additionally, lineage specific genes are unlikely to be identified in previous annotation efforts that focused largely on conserved amino acid sequences. With RNA-seq annotations we can identify lineage specific genes, which are expected to be important in the evolution of genomes and emergence of genetic novelty. The increased power to identify genes independently from conservation is a major step toward studying the evolution of genome content and rapidly evolving gene sequences. The ability to identify transcribed genes independently from conservation will facilitate evolutionary and functional analyses of the most rapidly changing segments of the genome. There may be additional gene models and isoforms in other tissue types or time points, which deserves to be explored. Particularly, these annotations are unlikely to reflect the full diversity of sequences and isoforms expressed during embryonic development postfertilization, pupal development, or in larvae. These alternative time points deserve future exploration and the gene models offered here will serve as a lower bound on the full diversity of sequences that are transcribed within each species.

Polycistronic genes

We observe thousands of putatively polycistronic annotations in *D. ananassae* and *D. yakuba*, as well as hundreds in *D. simulans* w^{501} where fewer gene model annotations were previously aligned and power to identify polycistronic genes is limited. In *D. melanogaster*, hundreds of genes are known to show signs of polycistronic transcription (Lin *et al.* 2007), offering a means of co-regulation across genes with similar functions (Blumenthal 1998; Slone *et al.* 2007). With very high sequencing coverage, we are able to recover a greater number of polycistronic transcripts, though some of these may be false positives resulting from annotation algorithms. Some genes may differ in the frequency with which they are transcribed as polycistronic *vs.* independent transcripts, but these results imply that at least low levels of polycistronic transcription may be common for many genes in the genome and future validation may explore the extent of their functionality. *Adh* and *Adhr* show evidence of differing polycistronic status

Table 6 Differentially Expressed Genes by Tissue and Species at FDR ≤0.1

Species	Tissue	Tissue	Significant	Tested
D. ananassae	Female ovary	Female carcass	203	7537
	Female ovary	Male testes	200	8282
	Male carcass	Male testes	1013	8349
	Male carcass	Female carcass	175	8417
D. yakuba	Female ovary	Female carcass	5420	8689
	Female ovary	Male testes	5868	10,202
	Male carcass	Male testes	3065	10,412
	Male carcass	Female carcass	724	9430
D. simulans	Female ovary	Female carcass	5053	8967
	Female ovary	Male testes	5741	10,222
	Male carcass	Male testes	4628	10,679
	Male carcass	Female carcass	611	9566
D. melanogaster	Female ovary	Female carcass	112	11,326
	Female ovary	Male testes	370	12,890
	Male carcass	Male testes	220	13,268
	Male carcass	Female carcass	286	12,502

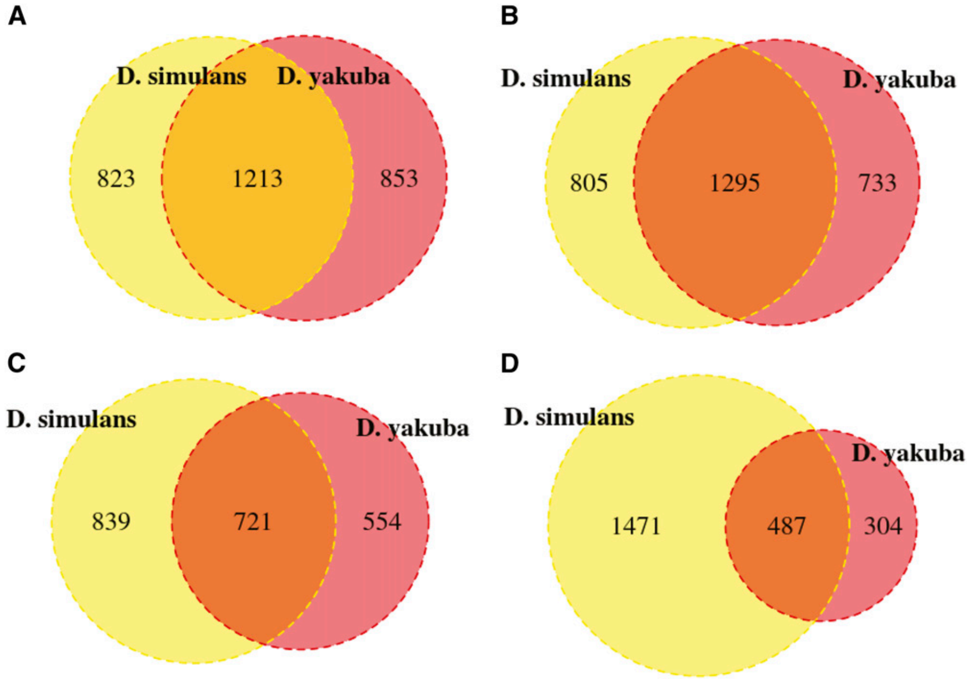

Figure 3 Genes with tissue biased expression in both *D. yakuba* and *D. simulans* in (A) female ovary, (B) female carcass, (C) male testes, and (D) male carcass. Numbers shown include only genes with a reciprocal best-hit ortholog in the sister species.

across *Drosophila* species (Betran and Ashburner 2000) and plant genomes are thought to split and fuse genes at rates of roughly $10^{-11}-10^{-10}$ per gene per year (Nakamura *et al.* 2007). Switching the rates at which genes are cotranscribed has the potential to alter regulatory patterns across species (Blumenthal 1998; Slone *et al.* 2007) and thereby produce novel phenotypes. These differences in polycistronic transcription therefore represent potential sources of genetic change that may be important in evolutionary change.

Tissue-specific expression

We identify thousands of genes that are differentially expressed across tissues in *D. yakuba* and *D. simulans*, where biological replicate samples for each tissue are available. A large fraction of the genome appears to display tissue-biased expression between germiline and carcass, with many fewer genes showing differential expression between male and female gonadectomized carcass, consistent with early microarry-based assays in *D. melanogaster* (Parisi *et al.* 2003; 2004). We have sequenced samples to extremely high coverage, generating transcript annotations with 5′ and 3′ UTRs with empirically supported intron-exon structures, improving accuracy in differential expression testing. Biological replicates for *D. yakuba* and *D. simulans* result in much greater power to identify differentially expressed genes in comparison with *D. ananassae* and *D. melanogaster*. However, even in *D. ananassae* and *D. melanogaster*, where only one replicate was available per tissue, we are still able to identify hundreds of genes that are differentially expressed across tissues with high coverage.

A comparison of orthologs between *D. yakuba* and *D. simulans* where biological replicates should result in sufficient power to detect differential expression reliably across both species reveals that roughly 60% of genes with reciprocal best hit orthologs exhibit similar tissue biased expression between the two species. The remaining 30% represent either genes that are differentially regulated between tissues but with effect sizes beyond the limits of detection, or genes that have evolved independent expression patterns between the two species. We also observe tissue biased expression in hundreds of lineage specific genes, which may represent candidates for neofunctionalization and new gene origination. These genes that display changes in sex-biased or tissue-biased expression across taxa, as well as lineage specific genes that exhibit tissue-biased and sex-biased expression are important candidate loci for evolutionary change in genome content and function that will be useful in exploring the functional and selective impacts of genomic changes.

ACKNOWLEDGMENTS

We thank Rahul Warrior for offering incubator space for fly cultures and Rachel Martin for an emergency supply of liquid nitrogen. R.L.R. is supported by the National Institutes of Health Ruth Kirschstein National Research Service Award F32-GM099377. Research funds were provided by National Institutes of Health grant R01-GM085183 to K.R.T. and R01-GM083228 to P.A. All sequencing was performed at the UC Irvine High Throughput Genomics facility, which is supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA062203. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

LITERATURE CITED

Betran, E., and M. Ashburner, 2000 Duplication, dicistronic transcription, and subsequent evolution of the alcohol dehydrogenase and alcohol dehydrogenase−related genes in *Drosophila*. Mol. Biol. Evol. 17: 1344–1352.

Blumenthal, T., 1998 Gene clusters and polycistronic transcription in eukaryotes. BioEssays 20: 480–487.

Chen, Z. X., D. Sturgill, J. Qu, H. Jiang, S. Park *et al.*, 2014 Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res. 24: 1209–1223.

Clemente, F., and C. Vogl, 2012 Unconstrained evolution in short introns? An analysis of genome-wide polymorphism and divergence data from *Drosophila*. J. Evol. Biol. 25: 1975–1990.

Drosophila Twelve Genomes Consortium, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450: 203–218.

Ellegren, H., and J. Parsch, 2007 The evolution of sex-biased genes and sex-biased gene expression. Nat. Rev. Genet. 8: 689–698.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver *et al.*, 1998 Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391: 806–811.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8: 1494–1512.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. Ravi Ram *et al.*, 2007 Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. Genetics 177: 1321–1335.

Hahn, M. W., M. V. Han, and S. G. Han, 2007 Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3: e197.

Halligan, D. L., and P. D. Keightley, 2006 Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res. 16: 875–884.

Hu, T. T., M. B. Eisen, K. R. Thornton, and P. Andolfatto, 2013 A second generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23: 89–98.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna *et al.*, 2012 A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816–821.

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

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

Lin, M. F., J. W. Carlson, M. A. Crosby, B. B. Matthews, C. Yu *et al.*, 2007 Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. Genome Res. 17: 1823–1836.

Mank, J. E., L. Hultin-Rosenberg, M. Zwahlen, and H. Ellegren, 2008 Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. Am. Nat. 171: 35–43.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351: 652–654.

Meisel, R. P., 2011 Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution. Mol. Biol. Evol. 28: 1893–1900.

Nakamura, Y., T. Itoh, and W. Martin, 2007 Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. Mol. Biol. Evol. 24: 110–121.

Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley *et al.*, 2003 Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science 299: 697–700.

Parisi, M., R. Nuttall, P. Edwards, J. Minor, D. Naiman *et al.*, 2004 A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. Genome Biol. 5: R40.

Parsch, J., S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, and P. Andolfatto, 2010 On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. Mol. Biol. Evol. 27: 1226–1234.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl, 2003 Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 300: 1742–1745.

Slone, J., J. Daniels, and H. Amrein, 2007 Sugar receptors in *Drosophila*. Curr. Biol. 17: 1809–1816.

Tamura, K., S. Subramanian, and S. Kumar, 2004 Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol. Biol. Evol. 21: 36–44.

Trapnell, C., L. Pachter, and S. L. Salzberg, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7: 562–578.

Van Dyken, J. D., and M. J. Wade, 2010 The genetic signature of conditional expression. Genetics 184: 557–570.

Wasbrough, E. R., S. Dorus, S. Hester, J. Howard-Murkin, K. Lilley *et al.*, 2010 The *Drosophila melanogaster* sperm proteome-II (DmSP-II). J. Proteomics 73: 2171–2185.

Zhang, Y., D. Sturgill, M. Parisi, S. Kumar, and B. Oliver, 2007 Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature 450: 233–237.

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

Communicating editor: R. Kulathinal