

Supplementary Text

Annotations status at the beginning of modENCODE

At the beginning of modENCODE project, only 46% of the genes, 30% of transcripts, and 48% of exons were experimentally verified across their entire length. The annotation set that existed at the time was FlyBase version 2008_09 release 5.12 (FB5.12)¹ which we used as the starting point of our analysis. This annotation includes 15,139 genes, 22,258 transcripts, 68,739 exons, and 51,358 introns. Less than half (6,232) of the genes are well studied and named. The processed transcripts comprise 29,979,817 bp (21%) and span 78,919,140 bp (56%) of the assembled genome (Release 5; 139,731,881 bp) (Fig. 3c). Systematic experimental validation of these annotated features comes primarily from EST and cDNA sequences (reviewed in²), which are biased toward abundantly expressed and moderately sized transcripts.

FB5.12 contains 1,081 annotated ncRNA genes. Of these, 927 are named genes, primarily tRNAs, snoRNAs, snRNAs and miRNAs. The remaining 157 include both well-known and uncharacterized ncRNAs. From the total RNA-Seq data we detected expression of 171 of the 249 annotated snoRNAs, 14 of 32 annotated snRNAs, 115 of 292 annotated tRNAs, five of 11 annotated predicted non-coding RNAs³, one of 22 annotated mitochondrial tRNAs, 21 of 88 annotated miRNAs, 21 of 97 annotated RNAs (CR RNAs), seven of 40 annotated snmRNAs, as well as 7SLRNA and RNase P.

Summary of RNA-Seq Experiments

From the poly(A)+ RNA-Seq libraries sequenced on the Illumina GAIIx platform, a total of 3.5 billion reads (265 Gb of sequence) was generated comprising between 72 and 193 million reads from each of the 30 RNA samples (Supplementary Table 3). Between 51 and 112 million reads from each of the 30 samples were uniquely mapped to the genome (Release 5) or splice junctions (Methods) for a total of 2,286,173,597 mapped reads (171,180,620,475 bp) (Supplementary Table 3). From the total RNA-Seq libraries sequenced on the SOLiD platform, we generated between 5 and 9 million uniquely mapped reads for each time point for a total of 92,548,111 million mapped reads (4,627,405,550 bp) (Supplementary Table 4). In addition, we generated 3,225,745 single-end reads of ~250 bp (1,154,880,016 bp) using the Roche 454 platform from poly(A)+ RNA from mixed-stage embryos, adult males, and adult females (long poly(A)+ RNA-seq) (Supplementary Table 5).

For the short poly(A)+ RNA-Seq experiments, replicate lanes are highly reproducible regardless of the laboratory in which the library was sequenced or whether single read or paired-end data were compared (average Pearson Correlation Coefficient (R)=0.99, tau > 0.94). The results described throughout the paper are derived from merging all lanes sequenced from each biological sample. Furthermore, there was high concordance between the RNA-Seq data and 38 bp tiling arrays when comparing annotated sequences (Supplementary Figure 6).

Transposable Elements

Other annotated features that may be expressed include transposable elements (TE). Complete TEs and TE fragments comprise 7,450,875 bp or approximately 5% of the annotated genome^{4,5}, the density being significantly higher in heterochromatin than euchromatin^{6,7}. However, very few of these elements appear to be actively transposing⁸. Minimal evidence for expression has been obtained from EST sequencing. Expression is a function of *cis*-acting elements that reside within the TEs and some elements are expressed sex specifically. Embryonic spatial expression of TEs has been studied^{9,10} and shows restricted patterns. We observe measurable RPKM values at a given element is unknown as the number of expressed TEs is not known. As shown in Supplementary Figure 8, the expression of TEs shows stage-specificity. The highest overall expression element is *TART-B*, with especially high RPKM in adult female and embryonic stages. The *copia* element is distinguished by significantly higher expression in pupal stages, as is *Tirant* in adult female stages. The overall third highest RPKM element, *Tom1*, has only four annotated copies in the genome and the high RPKM values may be influenced by the low copy number.

Splice Junction Discovery and Validation

We used a variety of methods and the RNA-Seq data to predict novel splice junctions and then to evaluate the confidence level of those junctions. We aligned reads to databases of predicted splice junctions, and to models using TopHat¹⁴, STAR, and BLAT¹¹. To identify new splice junctions we derived a set of metrics that could be used to distinguish true splice junctions from false positive splice junctions. First, we required a minimum of 6 nt overhang across a splice junction. Moreover, rather than requiring a specific number of reads to map to a given splice junction, we measured the entropy of the reads that mapped to the splice junction. The entropy score is a function of both the total number of reads that map to a given junction and the number of different offsets to which those reads map and the number that map at each offset. Thus, junctions with multiple reads mapping at each of the possible windows across the junction will be assigned a higher entropy score, than junctions where many reads map to only one or two positions. For this analysis we required that a junction have an entropy score of two or greater in at least two biological samples for junctions with canonical splice sites, and an entropy score of three or greater in at least three biological samples for junctions with non-canonical splice sites. Entropy was calculated using the following equations:

$$p_i = \text{reads at offset } i / \text{total reads to junction window}$$

$$\text{Entropy} = - \sum_i (p_i * \log(p_i) / \log 2)$$

We produced a total of ~79,400 high confidence junctions, ~29,000 of which are novel junctions. 259,628,732 reads map uniquely to splice junctions. Adult males, 5 days post eclosion express the largest number of unique junctions

(66,637) while 0-2 hour embryos express the fewest (40,945). The set of ~79K splice junctions was further filtered to exclude junctions with an intron length less than 41 nt, or that inadvertently join paralogous genes. Junctions were also excluded if they had non-canonical splice sites with the exception of those that were previously annotated or had other experimental support. This yielded a final set of 67,317 high confidence splice junctions in the short poly(A)+ RNA-Seq data that correspond to 46,566 (93%) FB5.12 annotated splice junctions and 20,751 new splice junctions.

In addition, we identified 32,917 splice junctions from the 454 data of which 28,860 (87.6%) correspond to FB5.12 annotated splice junctions and 4,057 new splice junctions. We amplified and sequenced 2,589 RT-PCR products and cloned and sequenced 3,204 cDNAs to validate splice junctions and exons discovered in the short and long poly(A)+ RNA-Seq datasets and annotated junctions that had previously not been experimentally validated. We confirmed 45,153 annotated FB5.12 junctions and validated 4,008 new splice junctions. Importantly, we identified 1,643 new splice junctions in the cDNA, EST, and RT-PCR datasets, highlighting the value of focused validation of transcript models. Conserved splicing patterns are a powerful indication that a particular junction is functional. We therefore performed short poly(A)+ RNA-Seq on transcript rich adult heads for *D. mojavensis* or *D. pseudoobscura* (Supplementary Table 31). 1,647 of the splice junctions are supported by the species RNA-Seq data (Supplementary Table 32)

There are 3,168 junctions from 1,853 gene models that were not detected. Only 299 are named and well-characterized genes such as the odorant or gustatory receptor family genes. It remains to be seen how many of these predicted genes and transcripts will ultimately be retained in a reannotated genome.

Analysis of Novel Transcripts via Interspecific Comparison

We characterized the NTRs using a comparative genomics approach. We identified 1,229 NTRs (containing 1,262 transcripts) that did not overlap FB 5.12, 67 were validated with cDNA evidence, 126 were novel annotations included in MB8 or FlyBase 5.23, and the remaining 1,037 were novel transcribed regions identified by Cufflinks¹². 291 start with a methionine codon and end with a stop-codon, 190 have a start-codon but lack a stop-codon, and 41 lack a start-codon but have a stop-codon. We compared their patterns of sequence divergence to randomly extracted genomic sequence (see Methods). We compared Kimura 2 parameter corrected genetic distance¹³ between *D. melanogaster* and its sibling species, *D. simulans* (2.5-5.4 MYD)¹⁴, for identifiable NTR longest transcript orthologs (1070) and performed the same analysis on 5,000 randomly extracted orthologous coding exonic, intronic, and intergenic sequences, the size distribution of each of which was chosen to match the pooled size distribution of all NTRs (see Methods). We found that the distributions of the pairwise divergence estimates of randomly extracted sequence are significantly different from one another and follow a pattern of divergence expected from the proportion of functional sequence they contain: exons < introns < intergenic sequence

(pairwise permuted Kruskal-Wallis test, $P < 2.2 \times 10^{-16}$; note that all pairwise tests are Bonferroni corrected) (Figure 1A). NTRs are significantly more conserved than both random intergenic or intronic sequence ($P < 2.2 \times 10^{-16}$). While not as highly conserved as random coding exon sequence, our results suggest that NTRs are constrained.

In addition to testing for sequence conservation, we assessed whether the NTRs detected in *D. melanogaster* were also expressed in other, distantly related *Drosophila* species. Whereas expression of non-functional NTRs due to phenomena such as adventitious promoters and stably accessible chromatin confirmation may be detectable in closely-related species due to lack of sufficient sequence divergence¹⁵, it is likely that such NTRs would degenerate neutrally and disappear given sufficient evolutionary time. We used RNA-seq data generated from dissected adult heads of *D. pseudoobscura* (25-55 million years diverged [MYD])¹⁶ and *D. mojavensis* (~60 MYD)¹⁷, and assessed expression of sequences orthologous to the *D. melanogaster* NTRs. We identified orthologous genomic regions corresponding to the 1,262 NTR transcripts in *D. melanogaster* (see Methods)¹¹, resulting in 952 and 802 orthologs being identified in *D. pseudoobscura* and *D. mojavensis*, respectively.

We then mapped the RNA-seq reads generated from dissected heads of the three species to their respective genomes and determined the level of expression of each NTR transcript in terms of FPKM (see Methods) using a minimum of FPKM > 1 . Using this threshold, 295 and 185 transcripts were detected in the heads of both *D. melanogaster* and *D. pseudoobscura*, or *D. mojavensis*; 148 transcripts were detectably expressed in the heads of all three species). We performed pairwise comparisons of the mean log₂ transformed (FPKM + 1) expression level estimates between *D. melanogaster* and each of the two other species for the orthologous NTR transcripts detected in all three species in order to determine whether NTRs showed conserved levels of expression among different species (Supplementary Figure 4c,d). FPKMs were significantly correlated in both pairwise comparisons, with higher coefficients of correlation in the comparisons between *D. melanogaster* and *D. pseudoobscura*, as compared to the former and *D. mojavensis*, as is expected based on phylogenetic distance (Spearman correlation test, $\rho = 0.652$ and 0.534 , $P = 2.2 \times 10^{-16}$ and $= 2.2 \times 10^{-16}$ for the comparison between *D. melanogaster* and *D. pseudoobscura* or *D. mojavensis*, respectively).

Feature validation in other *Drosophila* species: Validation of Exons

We used RNA-Seq results from two non-*melanogaster* species to provide supporting evidence of biological function of all exons, both novel and annotated. Reads from RNA-Seq experiments on *D. melanogaster*, *D. pseudoobscura* and *D. mojavensis* male and female dissected heads (described in GEO entries GSE20348, GSE19989 and GSE20882 respectively) were mapped to their respective genomes using Tophat¹² accepting only uniquely-mapped reads (Supplementary Table 31). (Unique mapping was performed. Complete parameters supplied to Tophat: (-g 1 -F 0 -G [FB5.28 (D.mel), r1.3 (D.moj), r2.10

(D.pse)] --solexa1.3quals). Genomic coordinates of each putative exon in *D. pseudoobscura* (DP4/FB r.2 assembly) and *D. mojavensis* (droMoj3, FB r.1 assembly) were obtained using the UCSC genome browser liftover tool¹⁸. Liftover coordinates were filtered to exclude spurious mappings, removing mappings that spanned less than half, or more than double, the range of the *D. melanogaster* exon. After filtering, 88.7% of *D. melanogaster* exons were successfully lifted-over to *D. pseudoobscura*, and 81.4% lifted-over to *D. mojavensis*.

To estimate expression of exons in the other species, we used BedTools¹⁹ to intersect genomic coordinates with base-level read coverage produced by Tophat. These results were parsed to quantify the read coverage for each lifted-over region. The percent of the non-melanogaster lifted-over region with coverage ≥ 1 (in either sex) was used as the metric to define 'validation,' or evidence of conserved expression, with a conservative cutoff of 95%.

The complete list of annotated exons was first examined in *D. melanogaster* heads, and exons that that detected by the threshold of 95% of exon length with coverage ≥ 1 in either sex were then compared to the other species. Overall, 60.7% and 73.4% of all exons expressed in *D. melanogaster* heads showed conservation of expression in *D.pseudoobscura* and *D.mojavensis*, respectively. Putative novel exons less frequently validated than annotated ones, with 51.9% novel exons showing conservation of expression in either species.

Validation of Splice Junctions

Biological validation for splicing was performed by comparing Tophat predicted splice junctions. Tophat results were obtained as described above for *D. pseudoobscura*, *D. mojavensis*, and *D. melanogaster*. Splice junctions detected by Tophat in *D. melanogaster* heads were used to find the corresponding junction region in the other species using the liftover tool¹⁸ with the junction spanning region from the *D. melanogaster* "junctions.bed" file. Thus, each junction examined was detected in *D. melanogaster* heads. These results were then filtered to include only those splice junctions included in the junction database used to analyze the developmental timecourse. A junction was considered validated if at least one junction-spanning read was present in male or female heads of another species.

A summary of these results is presented in Supplementary Table 32. Most splice junctions annotated in FB5.12 are detected in at least one other species (78.3%). These data provide support for 1,647 putative novel splice junctions, although these validate with a smaller percentage (31.8%).

Analysis of Expression Dynamics

To understand the dynamics of gene expression over the developmental time course we clustered genes using a fuzzy c-means clustering algorithm²⁰. We identified 20 distinct clusters representing a variety of developmental gene expression patterns (Supplementary Fig. 11). By analyzing genes that were predominantly associated with each cluster, we identified between 113 and 2,573

(mean = 677) genes in each cluster (Supplementary Table 33). We found that clusters 1-7 are strongly associated with early-to-mid embryogenesis and adult females, the latter likely the result of our profiling of whole-body adult females containing a substantial amount of oocytes. Clusters 11-17 appear to be primarily associated with late embryogenesis and the larval stage, though several of these show significant peaks of expression during later stages as well. Clusters 18-20 show low expression during early development followed by a monotonic increase in expression level terminating in adult males – and no significant expression in adult females.

The female-biased or -specific genes are known to be enriched for expression in the ovary and larval CNS^{28,21}, suggesting that most have ovary specific functions and/or have pleiotropic functions throughout development. Genes with male-biased or -specific expression tended to be single transcript genes and from array studies are expressed in the testis, brain and larval CNS^{28,21}.

Expression Dynamics of Non-Coding RNAs

Expression of the new ncRNAs ranges between FPKM 0.16 and 24,000 and displays a median fold-increase throughout embryogenesis of 6.5 and a maximum of 300.

We also examined the expression dynamics of 1026 putative non-coding RNAs among the new transcribed regions (NTRs). These transcripts have similar expression profiles as putative protein-coding NTRs, and also tend to be observed in male-specific stages (Supplementary Fig. 12).

As most non-coding RNAs are not polyadenylated, they can easily be missed by examining the poly(A)+ fraction alone. We detected expression for 838 non-coding RNAs in the SOLiD total RNA-seq data, of which only half (405) are also detected in the poly(A)+ RNA-seq despite the fact that the poly(A)+ samples had more than twice as many time points and stages. In particular, the snRNAs, snoRNAs, and tRNAs are highly expressed and only detectable by total RNA-seq (Supplementary Figs. 12 and 13). We expect to find hundreds more non-coding RNAs by interrogating Total RNA in more stages of development.

SUPPLEMENTARY METHODS

Animal Staging and Collection

The *Drosophila melanogaster* isogenic sequenced⁸ strain *y*¹; *cn bw*¹ *sp*¹ was used for the majority of our studies. Population cages were maintained at 24°C on a 24-hour light cycle (14 hours light/10 hours dark). Embryos were collected from the cages on plates (3.3% agar, 13% unsulfured molasses, and 0.15% Tegasept) at 2-hour intervals during a light cycle and aged at 24° for the required period. Aged embryos were passed through an 850 µm screen, and collected on a 75 µm screen, then dechorionated by treatment with 50% bleach (3% sodium hypochlorite), 0.2% NaCl, and 0.02% Triton-X-100 for five minutes and washed twice with 0.2% NaCl, 0.02% Triton buffer. For quality control, approximately 5% of the sample was removed, fixed²² and staged^{23,24}. The remainder was rinsed with water under mild vacuum, frozen on dry ice and stored at -80°C.

For the larval stages, a 2-hr embryo collection was aged 20 hr. The larvae were allowed to feed until harvesting 42-44 hr after egg-lay (L1), 66-70 hr after egg-lay (L2), or 83-85 hr after egg-lay (L3+12 hr), using the same procedure used for embryos, without dechorionation.

Later L3 stages were collected by culturing larvae in fly food supplemented with 0.05% bromophenol blue to mark the guts of feeding animals²⁵. Wandering larvae were collected at gut stage 1 (dark blue gut), 2 (light blue gut), or 3 (clear gut), corresponding to puff stages 1, 2-6, and 7-9, respectively²⁶.

Prepupae, pupae, and adults were collected from bottles containing standard medium. White prepupae were collected from the walls of the bottles. For later prepupal and pupal stages the positions of white prepupae in bottles were marked, then these animals were aged for 12 hr, 1, 2, 3, or 4 days.

Adults were either collected immediately after eclosion (adults eclosion + 1 day), or transferred to fresh bottles and aged for 4 (eclosion + 5 days) or 29 days (eclosion + 30 days).

RNA extraction from whole animals

Frozen samples were homogenized and extracted using the TRIzol reagent protocol (Invitrogen). RNA was purified on an RNeasy spin column (Qiagen), and DNase treated. poly(A)+ RNA was prepared from 250-500 µg aliquots, using the Oligotex mRNA Midi kit (Qiagen). RNA sample quality was assessed by Northern blots probed with RpL11-RA. RNA samples were stored at -80.

Affymetrix Tiling Arrays

RNA from three biological replicates of each developmental sample was independently hybridized on 38-bp Affymetrix arrays (Affymetrix GeneChip® *Drosophila* Tiling 2.0R Array), using standard procedures²⁷. Raw signals from the replicates were combined and smoothed using a 3-probe sliding window (bandwidth=50), with the intensity of each probe calculated as its background-corrected pseudomedian. Transcribed fragments (transfrags) were classified and “gene expression scores” for both novel transfrags and annotated exons calculated as described previously²⁸.

Illumina Library Preparations and Sequencing

Libraries were generated using the mRNA-Seq preparation kit as recommended by the manufacturer (Illumina) and sequenced on the Illumina Genome Analyzer IIx using either single or paired-end chemistry and 75-76 bp cycles per read.

SOLiD Library Preparations and Sequencing

To construct libraries for SOLiD sequencing, total RNA was treated with the RiboMinus™ Eukaryote Kit for RNA-Seq (Invitrogen); although the treatment failed to give significant depletion of ribosomal RNA. The samples were fragmented and ligated with SOLiD-specific adaptors using SOLiD Whole Transcriptome Analysis Kit (Ambion). Twelve samples were uniquely barcoded during PCR amplification using SOLiD Multiplexing Kit (Ambion). Sequencing was carried out to 50 bases using SOLiD V3 chemistry.

454 Library construction and Sequencing

For 454 sequencing, poly(A)⁺ RNA from male and female Oregon R adults and mixed-staged y1; cn bw1 sp1 embryos was used. mRNA was fragmented and first strand cDNA made using of Superscript II enzyme (Invitrogen Corporation, Carlsbad, CA). After ligation of adaptors the ligated material was recovered using Sera-Mag 30 beads (Seradyn Inc, Indianapolis, IN). The sscDNA libraries were purified with RNAClean and quantified with the Quant iT Ribogreen RNA Assay Kit (Invitrogen Corporation, Carlsbad, CA) on a Synergy HT (Bio Tek Instruments Inc, Winooski, VT) instrument. The libraries were taken through emulsion PCR with both kit II and kit III reagents (Roche/454 Life Sciences, Branford, CT) allowing for bidirectional sequencing of the library fragments. A total of three sequencing runs were performed on a Genome Sequencer FLX instrument (Roche/454 Life Sciences).

Targeted RT-PCR and cDNA Isolation and Sequencing

Standard procedures were used for RT-PCR³. Targeted cDNA isolation and sequencing was performed as described previously²⁹.

ANALYSIS

Identification of Unannotated Transcribed Regions from RNA-Seq Data

We used Cufflinks¹² to identify regions of active transcription that do not overlap, nor are directly adjacent to existing gene annotations. In total, we identified 5,636 NTRs, of which 2,428 show evidence of being spliced to existing, annotated transcripts, and thus likely represent previously unannotated exons or untranslated regions (UTRs) within existing transcripts. Those NTRs that showed no evidence of splicing association with existing transcripts were collapsed into transcripts themselves, based on whether they possessed splice junction-spanning reads connecting them to one another (658 the remaining 3208 NTRs possessed such junction spanning reads). We therefore collapsed NTRs with evidence of splice junctions into 2,879 transcripts, of which 1,269 fall within the

introns of annotated genes, while the remaining 1,610 lie within intergenic segments.

Identification of NTR Orthologs

We identified orthologs to individual NTRs within the genomes of *D. simulans* (University of California at Santa Cruz [UCSC] Genome Browser release DroSim1), *D. pseudoobscura* (UCSC Genome Browser release dp4), and *D. mojavensis* (UCSC Genome Browser release droMoj3) using the UCSC Genome Browser LiftOver tool with default parameters¹⁸. We rejected any ortholog whose length in base pairs was greater or equal than 2-fold, or less than or equal to 1/2 that of *D. melanogaster*. The remaining individual exons identified in other species were then reassembled into transcripts based on splicing evidence from *D. melanogaster*. We further retained only those orthologous NTRs where all exons identified in *D. melanogaster* could be identified under the conditions listed above in the other species. Orthologs for 1,070, 981, and 836 NTR transcripts were detected in *D. simulans*, *D. pseudoobscura*, and *D. mojavensis*, respectively.

Sequence Divergence Analysis

We obtained the complete list of predicted coding exons, introns, and intergenic sequence according to the FB5.12 annotation and extracted random sequences matching the same size distribution the NTRs using a custom PERL script. We then used the UCSC liftover tool to obtain orthologous coordinates for 5,000 random sequences in each category. We aligned the orthologous *D. melanogaster* and *D. simulans* sequences for each of the random sequences as well as the three categories of novels using Dialign-TX³⁰ and then calculate pairwise Kimura 2-Parameter genetic distances for each pair of sequences using the APE package in R³¹.

Interspecific Expression Profile Analysis

We collected RNA-Seq data from dissected head tissue of eight day old males and females of *D. melanogaster* (OregonR), *D. pseudoobscura* (14011-0121.94), and *D. mojavensis* (15081-1352.22) using the Illumina GAll platform. A single biological sample of male and female reads were pooled before analysis, resulting in 77.1 million high quality 75 bp reads from *D. melanogaster*, 86.6 million high quality reads for *D. pseudoobscura*, and 114 million high quality reads for *D. mojavensis* (full information regarding procedures can be found in GEO Series numbers GSE20348, GSE19989, and GSE20882).

The reads were mapped to the genomes of the species from which they were obtained using Tophat³², version 1.0.13, using the parameters "-F 0 -i 40 -solexa1.3quals -g 1", thus retaining *de novo* splice junction predictions derived from uniquely mapping reads. The output from Tophat was then used as input for Cufflinks¹², version 0.8.3, with default parameters, and constraining the isoform abundance estimates, Fragments per Kilobase per Million Mapped reads (FPKM), to a Gene Transfer Format (GTF) file containing the NTRs - or their orthologs in the case of non-*D. melanogaster* species. Restricting our analysis

only to those NTRs detectably expressed in the head data (FPKM > 1), we detected 364, 295, and 185 NTRs in *D. melanogaster*, *D. pseudoobscura*, and *D. mojavensis* heads, respectively.

Gene Model Annotation

Our new models include FB5.23 annotations, N-SCAN-EST³³ and PASA³⁴ gene predictions, as well as alternate transcripts based on cDNA clusters and Cufflinks transcript models. N-SCAN-EST gene predictions³³ were done on the whole genome, but only the predictions that did not overlap any FlyBase annotation were kept. To generate PASA clusters, ESTs and full insert cDNAs were obtained from GenBank and any vector sequence removed. They were clustered using the PASA pipeline, which aligns sequences to the genome, validates the splice sites, and keeps alignments that have at least 90% identity to the genome over 90% of their length. These alignments are clustered into non-redundant pseudo transcripts, which can then be used to extend and alter genome annotations³⁴. PASA also determines open reading frames for clusters that contain a full insert cDNA. If these annotations were not part of the FB5.23 set, or predicted by N-SCAN, they were added. The PASA clusters were also used to extend UTRs and add alternative splices to the N-SCAN predictions, and to the FB5.23 annotations for protein coding genes. Noncoding blocks from FB 5.23 that overlapped a gene prediction were designated as noncoding genes and employed the gene structure of the underlying gene prediction. Transcripts were assigned a confidence score based on available cDNA evidence. Annotations that were confirmed by a full insert cDNA have a lower confidence score than ones that are only partially confirmed by a PASA cluster. Known and predicted coding and noncoding gene and transcript models were assembled into a single gtf file (Supplementary Table 34).

TE expression

Illumina reads mapped to annotated TE regions were tabulated and each TE instance was assigned a score equaling the reciprocal of the number of times that read was mapped. For each developmental sample, the scores were summed among all instances of each TE to determine an expression score for each class. Each score was divided by the total length (in kb) of annotated sequence for each class and the number of mapped reads.

Identification of new non-coding RNAs

Potential miRNA precursors were identified by manually inspecting the genomic loci containing each miRNA annotated in miRBase v15 on a genome browser. The coordinates of continuous regions of coverage of the short total RNA-Seq data that spanned the annotated miRNAs were used in Cufflinks to calculate the expression level of each potential miRNA precursor.

The snoRNA-like RNAs were discovered by identifying regions of the genome with continuous coverage from the short total RNA-Seq data that did not overlap annotated exons. FPKM levels were calculated from these intervals using Cufflinks for each of the 12 samples. Intervals with an average FPKM >20

were manually inspected on a genome browser and the boundaries refined based on the coverage in the RNA-Seq data. MFold³⁵ was used to determine the presence of known snoRNA motifs.

RNA Editing

Short poly(A)⁺ RNA-Seq reads which mapped to annotated transcripts were compared to the reference sequence to find instances in which a 'G' aligned to an 'A' (for plus strand transcripts) or a 'C' aligned to a 'T' (for minus strand). We examined all but the first or last six base pairs of each read as an empirical inspection of the bases in these positions showed that they were easily mismapped to introns rather than across splice junctions.

We required at least five reads from each developmental sample to show a substitution to be considered further. Each site was examined to count the number of reads that aligned to the site with an exact nucleotide match. We created a dataset of potential editing sites with the number of edited reads and the number of non-edited read for each sample. Potential editing sites were defined as having at least 5% of the reads in two adult samples showing evidence of editing, and, to prevent a polymorphism in the sequenced strain from being reported as editing, at least 100 reads having an exact match to the genome. Expression levels and ratios of edited reads were mapped in rank order. The complete list of editing sites is included in the Supplementary materials.

Identifying alternatively spliced exons

The Junction Based Analysis of Splicing Events (JuncBASE) package³⁶ was used to identify the following alternative splicing events: cassette exon, alternative 5' splice sites, alternative 3' splice sites, alternative first exon, alternative last exon, mutually exclusive exons, intron retention and coordinate cassette exons³⁶. These events were counted as follows:

Cassette Exon: The number of cassette exons in the reference annotation or "new" is shown in Table 2. Each cassette exon is considered a separate cassette exon event.

Alternative 5' Splice Sites, Alternative 3' splice sites, Alternative First Exons, and Alternative Last Exon Events: These alternative events were grouped by an "anchor" splice site. For example, alternative donor events were assumed to have the same acceptor site ("anchor" splice site) and grouped together in this way. The number of total splicing events in a group is equal to the number of introns involved in the group - 1. This is due to the method of calculating PSI values, where every proximal alternative splice site (inclusion isoform) is compared to more distal splice sites (exclusion isoform(s)). When comparing events between two different annotation sets, the same group of introns, with the same anchor splice site, were compared. If both annotation sets had the same anchor point, then identical introns were not considered and the remaining intron(s) equaled the number of "novel" event(s). If the reference (FB5.12)

annotation did not have the same anchor point, then the number of new events was equal to the number of events in the "new" annotation (number of introns in the event - 1).

Mutually exclusive exons: The number of exons involved in a mutually exclusive exon event is shown in Table 2. "New" mutually exclusive exons were exons that were not part of a mutually exclusive event in the reference annotation (FlyBase r5.12). *n.b.* If there was evidence of an isoform that skipped all exons in a mutually exclusive event, it was still considered mutually exclusive. "Percent spliced in" (PSI) values were calculated by treating each mutually exclusive exon as an inclusion isoform of an event and all other exons in the event as possible exclusion isoforms. If there are only two exons involved in a mutually exclusive event, then a PSI was calculated for only one of the exons.

Coordinate cassette exons: The number of exons involved in a coordinate cassette exon event is shown in Table 2. "New" coordinate cassette exons are exons that were not part of a coordinate cassette exon event in the reference annotation (FB5.12). A PSI was calculated for all observed coordinate cassette exon events given a set of exons. For example, there may be four consecutive exons that have evidence that all are skipped but also evidence of additional events with a subset (2 or 3 of the 4) of exons that are included while the rest are skipped.

Intron Retention: The number of introns that show evidence of intron retention is given in Table 2. Intron retention events observed in the RNA-Seq data, and subsequently incorporated into the modENCODE annotation, cannot be distinguished between unprocessed RNA; therefore we report two numbers for the modENCODE intron retention events and the short read poly(A)+ RNA-Seq data: (1) intron retention events that are present in either FlyBase or from modENCODE cDNAs (2) All intron retention events.

JuncBASE was also used to identify exons that were differentially spliced. Read counts to every exclusion and inclusion isoform from each alternative splicing event for each developmental sample were obtained from JuncBASE and subsequently used to perform pairwise comparisons of differential splicing between samples. Fisher's exact test was used to identify differential splicing between each pair of samples and a Benjamini-Hochberg correction was performed for all pairwise comparisons within each type of splicing event.

Exon Conservation Analysis

To analyze and compare the conservation of constitutive and cassette exons, we extracted the PhastCons scores of each individual exon >50 bp and calculated the average PhastCons score at each nt position across the exons within each class.

Clustering alternatively spliced exons

For each alternatively spliced exon and for each sample, a percent spliced in (PSI) was calculated: $PSI = (\text{number of reads to inclusion isoform}) / (\text{number of reads to inclusion isoform} + \text{number of reads to exclusion isoform})$. The PSI was calculated only for events that had a total read count > 25. The delta PSI for an event was the difference between the highest and lowest PSI across all time points. For clustering analysis, a Z-score for each event was calculated. A negative Z-score is interpreted as an exon that is more skipped when compared to the average PSI and a positive Z-score corresponds to an exon that is more included when compared to the average PSI. We performed hierarchical clustering with optimal leaf ordering [ref: K-ary clustering with optimal leaf ordering for gene expression data. Bar-Joseph et al. Bioinformatics 2003]. Clusters were viewed with Java TreeView (<http://jtreeview.sourceforge.net/>). GO analysis was performed using Funcassociate 2.0³⁷.

SUPPLEMENTARY REFERENCES

- 1 Drysdale, R. FlyBase : a database for the Drosophila research community. *Methods Mol Biol* 420, 45-59, (2008).
- 2 Celniker, S. E. & Rubin, G. M. The Drosophila melanogaster genome. *Annu Rev Genomics Hum Genet* 4, 89-117 (2003).
- 3 Yandell, M. *et al.* A computational and experimental approach to validating annotations and gene predictions in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A* 102, 1566-1571 (2005).
- 4 Kaminker, J. *et al.* The transposable elements of the Drosophila melanogaster euchromatin - a genomics perspective. *Genome Biology* 3, research0084 (2002).
- 5 Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol* 1, 166-175, doi:10.1371/journal.pcbi.0010022 (2005).
- 6 Hoskins, R. A. *et al.* Heterochromatic sequences in a Drosophila whole genome shotgun assembly. *Genome Biology* 3, research0085 (2002).
- 7 Smith, C. D., Shu, S., Mungall, C. J. & Karpen, G. H. The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science* 316, 1586-1591, (2007).
- 8 Celniker, S. E. *et al.* Finishing a whole genome shotgun: Release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biology* 3, research0079 (2002).
- 9 Parkhurst, S. M. & Corces, V. G. Developmental expression of Drosophila melanogaster retrovirus-like transposable elements. *EMBO J* 6, 419-424 (1987).
- 10 Ding, D. & Lipshitz, H. D. Spatially regulated expression of retrovirus-like transposons during Drosophila melanogaster embryogenesis. *Genet Res* 64, 167-181 (1994).
- 11 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-664, (2002).
- 12 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28, 511-515, (2010).
- 13 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16, 111-120 (1980).
- 14 Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol* 21, 36-44, (2004).
- 15 Khaitovich, P. *et al.* Functionality of intergenic transcription: an evolutionary comparison. *PLoS Genet* 2, e171, (2006).
- 16 Richards, S. *et al.* Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* 15, 1-18, (2005).
- 17 Hahn, M. W., Han, M. V. & Han, S. G. Gene family evolution across 12 Drosophila genomes. *PLoS Genet* 3, e197, (2007).

18 Rhead, B. *et al.* The UCSC Genome Browser database: update 2010.
19 *Nucleic Acids Res* 38, D613-619, (2010).

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

22 Bezdek, J. C. *Pattern recognition with fuzzy objective function algorithms.*
23 (Plenum Press, 1981).

24 Chintapalli, V. R., Wang, J. & Dow, J. A. Using FlyAtlas to identify better
25 *Drosophila melanogaster* models of human disease. *Nat Genet* 39, 715-
26 720, (2007).

27 Matthews, K. A., Miller, D. F. & Kaufman, T. C. Functional implications of
28 the unusual spatial distribution of a minor alpha-tubulin isotype in
29 *Drosophila*: a common thread among chordotonal ligaments, developing
30 muscle, and testis cyst cells. *Dev Biol* 137, 171-183, (1990).

31 Hartenstein, V. & Campos-Ortega, J. A. *The embryonic development of*
32 *Drosophila melanogaster*. Second edition edn, (Springer-Verlag Berlin,
33 1997).

34 Wilson, R. J., Goodman, J. L. & Strelets, V. B. FlyBase: integration and
35 improvements to query tools. *Nucleic Acids Res* 36, D588-593, (2008).

36 Maroni, G. & Stamey, S. Use of blue food to select synchronous, late
third-instar larvae. *Drosophila Inf. Serv.* 59, 142-143 (1983).

Andres, A. J. & Thummel, C. S. Methods for quantitative analysis of
transcription in larvae and prepupae. *Methods Cell Biol* 44, 565-573
(1994).

Manak, J. R. *et al.* Biological function of unannotated transcription during
the early development of *Drosophila melanogaster*. *Nat Genet* 38, 1151-
1158 (2006).

Cherbas, L. The Transcriptional Diversity of 25 *Drosophila* Cell Lines.
Genome Res (2010).

Wan, K. H. *et al.* High-throughput plasmid cDNA library screening. *Nat*
Protoc 1, 624-632 (2006).

Subramanian, A. R., Kaufmann, M. & Morgenstern, B. DIALIGN-TX:
greedy and progressive approaches for segment-based multiple sequence
alignment. *Algorithms Mol Biol* 3, 6, (2008).

Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics
and Evolution in R language. *Bioinformatics* 20, 289-290 (2004).

Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice
junctions with RNA-Seq. *Bioinformatics* (2009).

Wei, C. & Brent, M. R. Using ESTs to improve the accuracy of de novo
gene prediction. *BMC Bioinformatics* 7, 327, (2006).

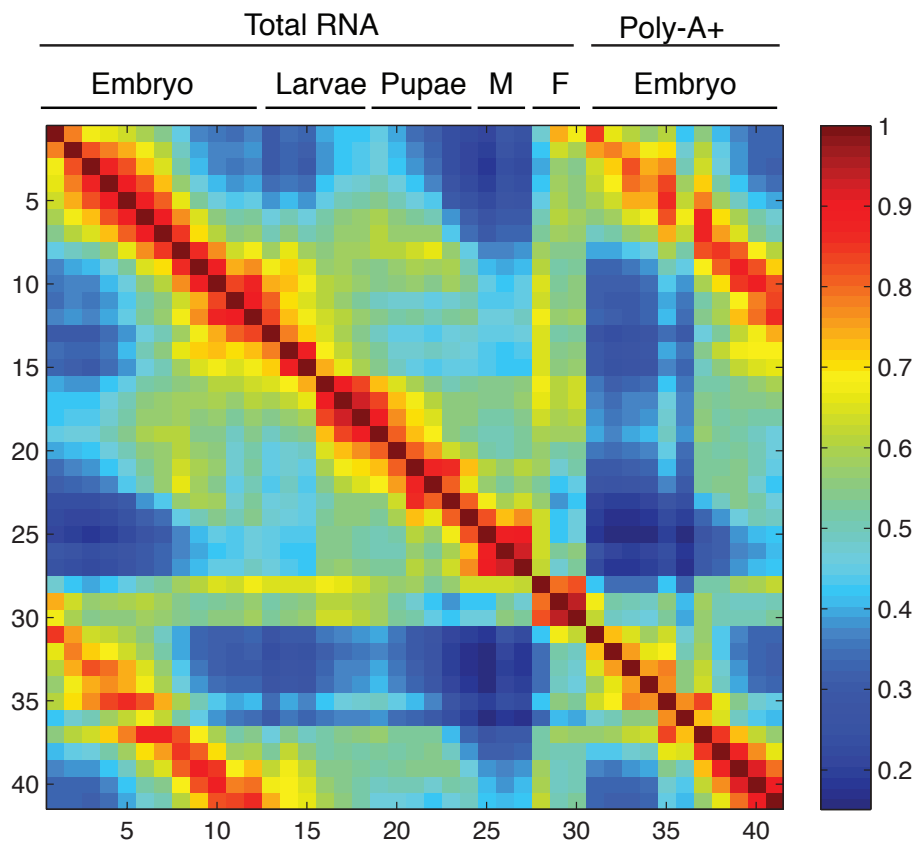
Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using
maximal transcript alignment assemblies. *Nucleic Acids Res* 31, 5654-
5666 (2003).

Zuker, M. Mfold web server for nucleic acid folding and hybridization
prediction. *Nucleic Acids Res* 31, 3406-3415 (2003).

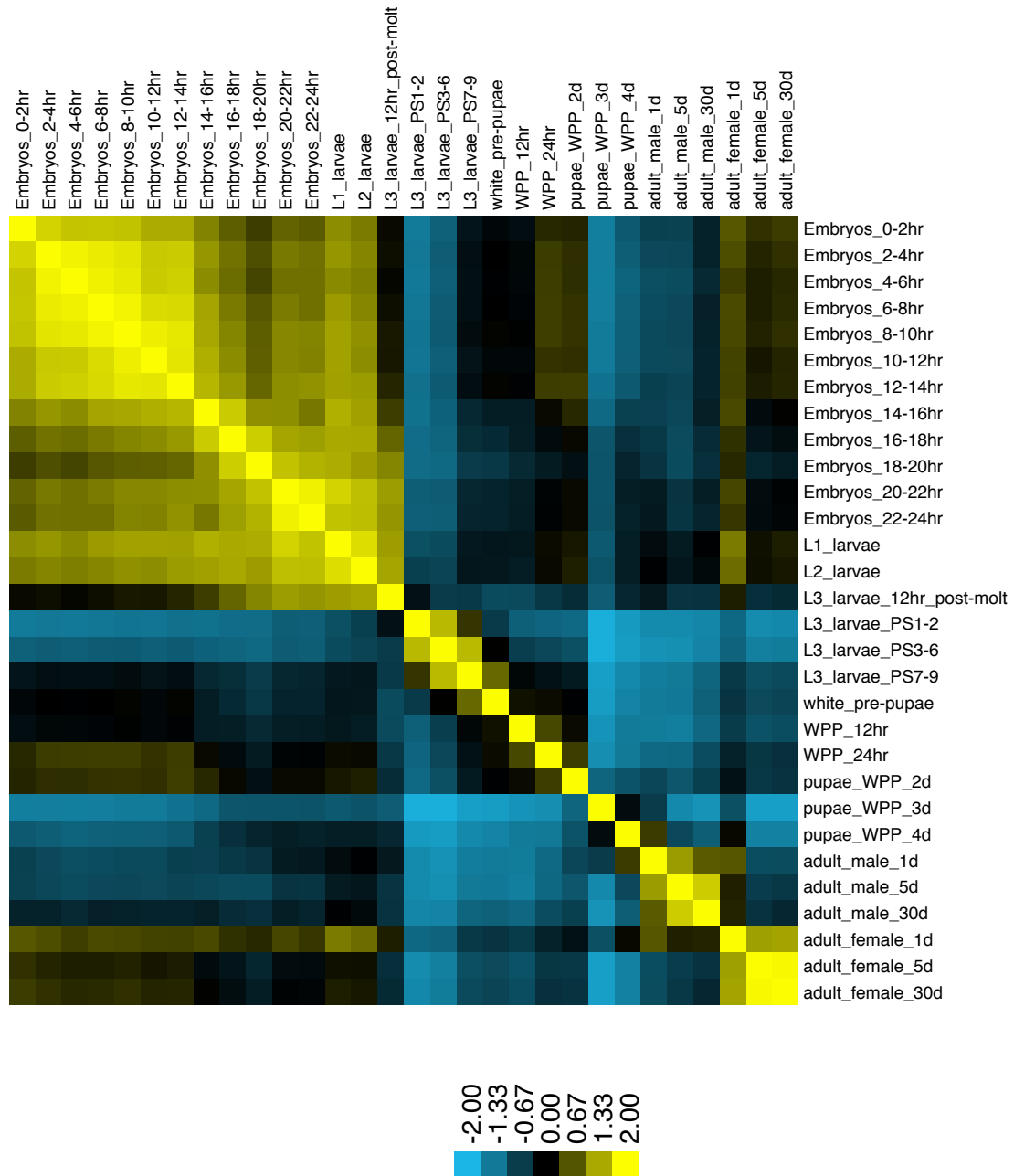
Brooks, A. N. *et al.* Conservation of an RNA Regulatory Map between
Drosophila and Mammals. . *Genome Research* (2010).

37

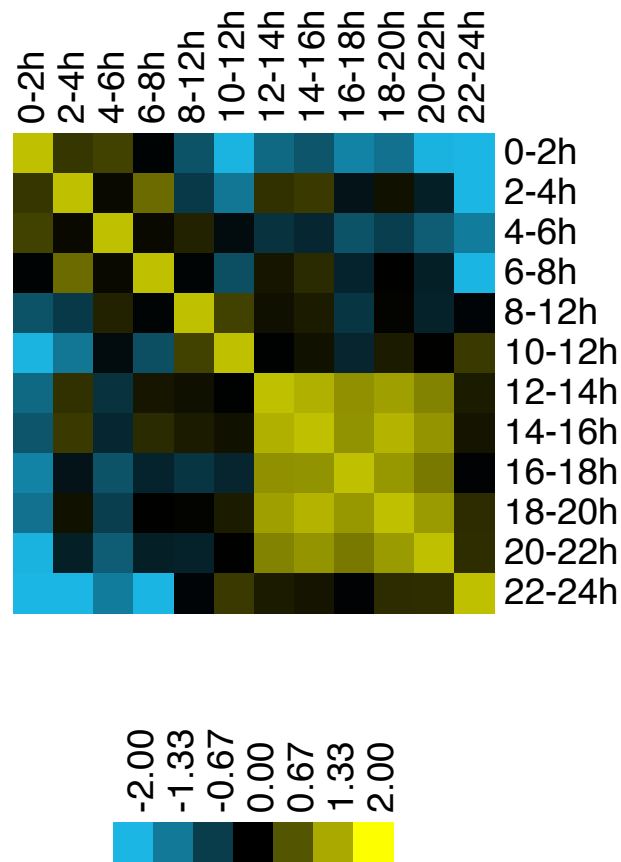
Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. *Bioinformatics* 25, 3043-3044, (2009).



Supplementary Figure 1. Reproducibility of gene expression between samples measured by microarray hybridization. Correlation of biological replicates of total RNA from 30 developmental stages throughout development (1-30) and from polyA(+) RNA from 12 embryonic samples (31-42) independently hybridized on 38-bp Affymetrix arrays. The scale bar indicates the correlation between samples.



Supplementary Figure 2. Correlation of gene expression levels between stages using poly(A)+ RNA-seq data. Each developmental stage is most highly correlated with its adjacent stage, particularly in the embryonic stages. A decrease in correlation is observable in the third instar transition. Adult males and adult female form their own highly correlated groups. The scale bar indicates the Z-scores of the correlation between samples.



Supplementary Figure 3. Correlation of gene expression levels between samples in the embryonic total RNA-seq samples. Each developmental stage is most highly correlated with its adjacent stage. The scale bar indicates the Z-scores of the correlation between samples.

