Supplementary Text for "Deep annotation of *Drosophila melanogaster* microRNAs yields insights into their processing, modification, and emergence"

Eugene Berezikov[1], Nicolas Robine[2], Anastasia Samsonova[3], Jakub O. Westholm[2], Ammar Naqvi[2], Jui-Hung Hung[4], Katsutomo Okamura[2], Qi Dai[2], Diane Bortolamiol-Becet[2], Raquel Martin[2], Yongjun Zhao[5], Phillip D. Zamore[6], Gregory J. Hannon[7], Marco A. Marra[5], Zhiping Weng[8], Norbert Perrimon[3], and Eric C. Lai[2,9]

Supplementary Materials and Methods

Small RNA datasets

A number of our datasets, produced as part of the modENCODE project, were described in previous publications (Chung et al. 2008; Flynt et al. 2009; Okamura et al. 2009; Okamura et al. submitted). 48 additional datasets produced for this project, and described for the first time herein, have already been deposited in the modENCODE Data Coordination Center and are also available from NCBI GEO. We cloned 18-28nt RNAs according to the method of Hannon and colleagues (Czech et al. 2008) and generated single Illumina GA lanes of sequence for each RNA sample. These datasets comprise a developmental series (2-6 hr or 12-24 embryos, 1st or 3rd instar larvae, mass-isolated imaginal disc/brain/salivary gland/larval gonad fraction, 0-1 or 2-4 day pupae, whole males or females, male or female bodies, male or female heads) and a variety of cultured cells (S2-GMR, Kc167, CMEL1, MLDmD20c5, Sg4, GM2, ML-DmD21, 1182-4H, CMEW1 cl.8+, ML-DmBG1-C1 and ML-DmBG3-C2 cells). We also prepared libraries from S2-GMR cells or Kc cells exposed to 5x10[-6] 20-hydroxyecdysone for 48 hours prior to RNA harvest, and a paired set of 2-18 hr wildtype embryos subjected to 4K rads irradiation or mock treatment. Many of these libraries were prepared from independent biological samples. Detailed analysis of miRNA content of these individual libraries will be described elsewhere. A summary of the read characteristics of these datasets and their accession numbers is provided in Supplementary Table 1.

We combined these modENCODE datasets with 111 other published *Drosophila* small RNA datasets (Brennecke et al. 2007; Ruby et al. 2007; Brennecke et al. 2008; Czech et al. 2008; Ghildiyal et al. 2008; Kawamura et al. 2008; Lu et al. 2008; Czech et al. 2009; Hartig et al. 2009; Li et al. 2009; Malone and Hannon 2009; Shi et al. 2009;

Ameres et al. 2010; Ghildiyal et al. 2010; Marques et al. 2010). Where possible (virtually all cases) we began from the raw sequence files so that all data were processed uniformly. 3' linker sequences were stripped using the fastx-toolkit developed by the Hannon lab (http://hannonlab.cshl.edu/fastx_toolkit/). For most purposes, we used Bowtie (Langmead et al. 2009) to map to the dm3 genome assembly, using parameters to restrict to perfectly matching reads ≥18 nt and all genomic hits reported. Analyses of untemplated additions and editing involved specialized procedures described below. A summary of all these datasets and their read characteristics is provided in Supplementary Table 1.

Analysis of miRNA 5' variation

We mapped reads to the canonical *D. melanogaster* miRNAs in miRBase 15 using Bowtie (Langmead et al. 2009). We selected those loci with >1000 reads that were ≥18 nt and had exclusively unique genomic mappings. In particular, we excluded members of the following miRNA families (*mir-2a-1/mir-2a-2/mir-2b-1/mir-2a-2/mir-2c*; *mir-6-1/mir-6-2/mir-6-3*; *mir-13b-1/mir-13b-2*; *mir-276a/mir-276b*; *mir-983-1/mir-983-2*) to avoid ambiguity of genomic assignment created by reads matching multiple loci. This left 135 canonical miRNAs for analysis. We tabulated the frequency of alternative 5' ends and their position (nt 5' or 3' to the base end) in Table S4. The corresponding panels in Figure 4 were drawn using the R package.

miRNA discovery

Analysis of small RNA reads was performed using miR-Intess software tuned for performance on *Drosophila* (Lau et al. 2009; Berezikov et al. 2010). Adapter sequences were removed from raw data and reads were mapped to the *D. melanogaster* genome (dm3 assembly, downloaded from UCSC Genome Browser web site) using blast software. Reads that mapped perfectly to the genome in at least 18 first bases, were considered in further analysis. Mapped loci were annotated using information from Ensembl database (v.58) and miRBase database (v.15) and classified into miRNA, tRNA, rRNA, snoRNA, siRNA, senseRNA, repeats and intergenic regions according to annotations.

Non-repetitive loci were analysed for the presence of hairpin structures using RNAshapes software (Steffen et al. 2006). For the identified hairpins a number of parameters were calculated, including abundance and 5’ variability of the reads mapped

to the hairpin as well as their position relative to the stem and the loop, number of unpaired bases, size of the bulges and Drosha/Dicer overhangs, randfold value and number of antisense reads. Based on combinations of these parameters, hairpins were assigned to various confidence levels and then subjected to manual inspection and curation for assignment as confident novel miRNAs or candidate miRNA loci.

The loci judged as confident for assignment of miRbase gene names fell into two general categories. First, we considered confident those loci with dominant mature/star reads exhibiting 3' overhangs as duplexes, with no larger than 4 bp internal loops or asymmetric bulges. Bearing in mind that some confident loci exhibit alternative processing for which there can be an abundant isomiR (e.g. main Figures 4 and 5), we required $\geq$10 mature strand reads (including up to one 5' isomiR) to constitute >2/3 of reads mapped to the hairpin arm, and $\geq$2 star reads. A second category of loci might not have star reads in the sample data, if the duplex is subject to strongly asymmetric strand selection. Since loci lacking star reads cannot confidently be inferred to have been generated by RNAse III cleavage, we required in these cases that there be at least 3 reads in a wild-type AGO1-IP library, along with proviso that the given species (along with up to one 5' isomiR) constituted >2/3 of reads mapped to the hairpin arm. We considered one or two reads in a wild-type AGO1-IP library, or exclusive AGO1-IP reads from mutant or knockdown samples (often signifying endo-siRNAs), to be insufficient evidence to support annotation in miRbase. All of the loci were manually vetted to meet confident criteria, and the strong majority of annotated loci had both star reads as well as AGO1-IP reads. Loci that met some, but not all confidence criteria were provisionally annotated as "candidates". The analysis of candidate *sblock109902* in Figure 8 was drawn using the R package. All of the analyses of confident and candidate miRNAs can be browsed in the "Supplementary Analyses available online, http://www.macgenome.org/pub/lai_mirna/main.html"

To find homologous miRNAs in genomes of other species, miRNA hairpin sequences were blasted against 12 available *Drosophila* genomes, as well as *A. mellifera* and *A. gambiae* genomes. Where available, BLASTZ_NET aligned regions were also retrieved from Ensembl or UCSC Genome Browser. All hits that contained conserved seed sequence (nt 2-8 of the most abundant mature sequence) were checked for the presence of hairpin structures using RNAshapes software, and similarity between hairpins was calculated using RNAforester software (Hochsmann et al. 2003).

Hairpins with the highest RNAforester score above the threshold of 0.3 were assigned as most likely orthologs in a particular species.

Analysis of untemplated additions

We mapped all sequencing reads to the fly genome with a previously described prefix matching algorithm (Ameres et al. 2010), which allowed a 3′ overhang of any number of mismatches on the reads. We counted reads with genome-matching (i.e., no 3′ overhang, the entirety of the read was able to map to the genome), or prefix matching (i.e., with a 3′ overhang, also called tailing) for each miRNA, as well as for all annotated non-coding RNAs (ncRNAs). We binned the 3′ overhang according to their sequences: homo-A, -C, -G, -T or X (X means mixed ACGT). We pooled multiple datasets, normalizing each dataset by sequencing depth. We performed a t-test or a KS test for whether one arm of the pre-miRNA hairpin had more tailed sequences than the other arm. We identified miRNAs that were consistently adenylated (or uridylated) as follows. For each miRNA, we identify the percentages of datasets in which it had higher than 1%, 5%, and 10% adenylation. Then we ranked all miRNAs by their percentages of datasets for each cutoff (1%, 5% or 10%). The miRNAs that were in the top 20 for all three cutoffs were retained.

Identification of candidate RNA editing events

The Illumina reads were stripped of 3' linkers and the resulting sequences were mapped to the release 5.26 of *Drosophila melanogaster* genome, with the exception of the Uextra portion. The alignment of short reads to the genome was performed with Bowtie package v.0.12.3 allowing up to two mismatches per read (-v 2 --best) (Langmead et al. 2009). Only single valid alignment per read was reported. The resulting read pileups were processed with several filters to reveal potential RNA editing events. The filtration criteria were designed to narrow down candidate space to high-confidence ones, avoid SNPs and minimize the impact of possible sequencing errors. The criteria used are as follows: (1) Average base sequencing quality score for a given position is greater than 20. (2) All candidate positions satisfy the neigbourhood quality score criteria (NQS20/15). (3) Modification is neither first nor the last two bases of a read. (4) Base coverage is greater than 15. (5) Frequency of the most abundant variant base is within 10-85% interval of total coverage at mismatched position. In the absence of a prevalent variant the candidate position is discarded. Candidates that passed this multi-level

filtration procedure are listed in Table S6.

Descriptions of the eight Supplementary Tables

Table S1. Statistics of the *Drosophila* small RNA libraries analyzed in this study. The first worksheet summarizes all of the datasets analyzed in this study, including the library descriptions, GEO/SRA/modENCODE ID numbers, and the statistics for linker clipping and mapping. The second worksheet segregates those datasets reported in this study for the first time.

Table S2. Aggregate read counts of reads matching miRBase 15 annotations. This table summarizes the aggregate miRNA/star/loop/5'moR/3' moR reads mapped to known *D. melanoga*ster miRNA loci.

Table S3. Ovary, head and S2 reads matching miRBase 15 annotations. We aggregated all the small RNA data available from ovaries, heads, and S2 cells as tissue/cell sources (see Table S1), and tabulated their absolute counts of miRNA and star reads (counting all isomir sequences $\pm$4 nt of the canonical read). The second worksheet compares those miRNAs contributing >1% of the total miRNA pool in each tissue/cell. The third worksheet compares overlaps in miRNA content across lower levels of co-expression. These data were used to generate Figure 1.

Table S4. 5' variability of miRNAs derived from mirtrons and canonical miRNA loci. This table summarizes 5' isomiR sequences across known miRNA loci. The canonical start positions are highlighted in red, and 5' isomiRs either 5' or 3' to the canonical start are noted. The first worksheet summarizes canonical miRNA loci, sorted by miRNA and miRNA* content. The second worksheet summarizes mirtron loci, sorted by 5p and 3p content. The data were used to generate Figure 3.

Table S5. Statistics of antisense, novel, and candidate miRNA loci annotated in this study. The first worksheet summarizes antisense miRNA loci, including novel confident loci and a set of candidates that have limited but compelling validation evidence (such as star species and/or AGO1-IP reads). The second worksheet summarizes confident novel genomic locations of miRNA production. The third worksheet summarizes candidate miRNA loci, with limited but compelling validation evidence.

Table S6. Reads mapping to a transitional miRNA locus, sblock109902, located in the CG15102 3'UTR. This region generates a highly heterogeneous set of short RNA reads, with diverse 5' ends and lengths. The reads highlighted in green exhibit strong evidence for constituting a miRNA/miRNA* duplex with detectable AGO1 incorporation. These data were used to generate Figure 8.

Table S7. 3' untemplated additions to mature miRNA reads. The first worksheet summarizes all categories of 3' untemplated additions on miRBase 15 loci in each of the small RNA datasets analyzed, along with t-test and KS test of whether the addition rate onto the aggregate 5p and 3p species were statistically different. The second worksheet summarizes the trends across all libraries, normalized by library. These data were used to generate Figure 9.

Table S8. Evidence for candidate miRNA editing events. This worksheet summarizes miRNAs and star species exhibiting candidate evidence for A-I editing, and the libraries for which the level of evidence passed the filtering steps outlined in the methods.

Descriptions of the Supplementary HTML documents

Detailed analysis of all the read mappings to miRNAs and other candidate loci is provided in a series of Supplementary HTML documents that can be accessed http://www.macgenome.org/pub/lai_mirna/main.html". A series of hierarchical websites provide the following analyses:

• Mappings to extant miRBase loci. The main page summarizes the total numbers of 5'moR/miR-5p/loop/miR-3p/3'moR reads for each locus, and calls the mature/star strands based on their relative abundance. Reads were placed into these categories based on their limits being within ± 4 nt of the most abundant read of that class; all other reads not matching this definition were tabulated in the "other" category. Clicking on any locus opens up a summary of the reads mapped in each library to the hairpin, along with graphics for secondary structure and read density along the hairpin.

• Mappings to miRBase loci with confident or candidate antisense processing. All of the antisense read mappings are available in the main miRBase page; however, we segregated the notable antisense miRBase loci to permit convenient browsing.

• Mappings to novel miRNA loci annotated in this study. The main page summarizes the total numbers of reads mapped to each locus, as well as to an extended locus ± 50 bp, the number of mappings to the antisense strand, reads mapped to the antisense strand, and read counts from AGO1-IP libraries from wild-type samples, AGO1-IP libraries from mutants or knockdown samples, AGO2/beta-eliminated/oxidized libraries, and Piwi-class libraries. Clicking on any locus opens up a summary of the reads mapped in each library to the hairpin. A link is provided from each individual locus page to the corresponding region in the UCSC Genome Browser.

• Mappings to candidate miRNA loci annotated in this study. The main page summarizes the total numbers of reads mapped to each locus, as well as to an extended locus ± 50 bp, the number of mappings to the antisense strand, reads mapped to the antisense strand, and read counts from AGO1-IP libraries from wild-type samples, AGO1-IP libraries from mutants or knockdown samples, AGO2/beta-eliminated/oxidized libraries, and Piwi-class libraries. Clicking on any locus opens up a summary of the reads mapped

in each library to the hairpin. A link is provided from each individual locus page to the corresponding region in the UCSC Genome Browser.

• Mappings to hairpin candidates reported in Ruby and colleagues (Ruby et al. 2007). We summarize here read mappings to 243 high-scoring conserved hairpins that were not validated by small RNA sequencing in that study. The main page summarizes the locus coordinates, overlapping gene annotations or hairpin designations (where relevant), min/max BLAST hits for mapped reads, normalized read numbers, reads mapped to the antisense strand, and read counts from AGO1-IP libraries from wild-type samples, AGO1-IP libraries from mutants or knockdown samples, AGO2/beta-eliminated/oxidized libraries, and Piwi-class libraries.

• Mappings to hairpin candidates reported in Stark and colleagues (Stark et al. 2007). We summarize here read mappings to 61 high-scoring conserved hairpins that were not validated by small RNA sequencing in that study. The main page summarizes the locus coordinates, overlapping gene annotations or hairpin designations (where relevant), min/max BLAST hits for mapped reads, normalized read numbers, reads mapped to the antisense strand, and read counts from AGO1-IP libraries from wild-type samples, AGO1-IP libraries from mutants or knockdown samples, AGO2/beta-eliminated/oxidized libraries, and Piwi-class libraries.

Supplementary References

Ameres, S.L., Horwich, M.D., Hung, J.H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P.D. 2010. Target RNA-directed trimming and tailing of small silencing RNAs. *Science* **328**(5985): 1534-1539.

Berezikov, E., Liu, N., Flynt, A.S., Hodges, E., Rooks, M., Hannon, G.J., and Lai, E.C. 2010. Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nat Genet* **42**(1): 6-9.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* **128**(6): 1089-1103.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G.J. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**(5906): 1387-1392.

Chung, W.J., Okamura, K., Martin, R., and Lai, E.C. 2008. Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Current Biology* **18**: 795-802.

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

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., and Hannon, G.J. 2009. Hierarchical rules for Argonaute loading in Drosophila. *Mol Cell* **36**(3): 445-456.

Flynt, A.S., Liu, N., and Lai, E.C. 2009. Dicing of viral replication intermediates during silencing of latent Drosophila viruses. *Proc Natl Acad Sci U S A* **106**: 5270-5275.

Ghildiyal, M., Seitz, H., Horwich, M.D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E.L., Zapp, M.L., Weng, Z. et al. 2008. Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. *Science* **320**: 1077-1081.

Ghildiyal, M., Xu, J., Seitz, H., Weng, Z., and Zamore, P.D. 2010. Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *Rna* **16**(1): 43-56.

Hartig, J.V., Esslinger, S., Bottcher, R., Saito, K., and Forstemann, K. 2009. Endo-siRNAs depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. *EMBO J* **28**: 2932-2944.

Hochsmann, M., Toller, T., Giegerich, R., and Kurtz, S. 2003. Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf* **2**: 159-168.

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T., Siomi, M.C., and Siomi, H. 2008. Drosophila endogenous small RNAs bind to Argonaute2 in somatic cells. *Nature* **453**: 793-797.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Lau, N., Robine, N., Martin, R., Chung, W.J., Niki, Y., Berezikov, E., and Lai, E.C. 2009. Abundant primary piRNAs, endo-siRNAs and microRNAs in a Drosophila ovary cell line. *Genome Res* **19**(10): 1776-1785.

Li, C., Vagin, V.V., Lee, S., Xu, J., Ma, S., Xi, H., Seitz, H., Horwich, M.D., Syrzycka, M., Honda, B.M. et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**(3): 509-521.

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R.W., Wang, S.M., and Wu, C.I. 2008. The birth and death of microRNA genes in Drosophila. *Nat Genet* **40**(3): 351-355.

Malone, C.D. and Hannon, G.J. 2009. Small RNAs as guardians of the genome. *Cell* **136**(4): 656-668.

Marques, J.T., Kim, K., Wu, P.H., Alleyne, T.M., Jafari, N., and Carthew, R.W. 2010. Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. *Nat Struct Mol Biol* **17**(1): 24-30.

Okamura, K., Liu, N., and Lai, E.C. 2009. Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. *Mol Cell* **36**(3): 431-444.

Okamura, K., Robine, N., Liu, Y., Liu, Q., and Lai, E.C. submitted. R2D2 organizes small regulatory RNA pathways in Drosophila.

Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P., and Lai, E.C. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* **17**: 1850-1864.

Shi, W., Hendrix, D., Levine, M., and Haley, B. 2009. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat Struct Mol Biol* **16**(2): 183-189.

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G.J., and Kellis, M. 2007. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res* **17**: 1865-1879.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22**(4): 500-503.