

SUPPLEMENTAL METHODS

Annotation of RNA and DNA reads

Annotated unique RNAs of *Drosophila* was downloaded from FlyBase (Drysedale and FlyBase 2008) and the repeat sequences from RepeatMasker track in the UCSC genome browser (Jurka et al. 2005). RNA reads were annotated to unique and repeat RNAs using IntersectBed tool with the parameter *-f 1.0*, *-split* and *-s*. The *Drosophila* genome was scanned with EMBOSS (Rice et al. 2000) based on AluI restriction sites from REBASE (Roberts et al. 2015) to generate AluI DNA bins. IntersectBed tool was used to connect AluI DNA bins using the parameter *-f 1.0*.

Assigning RNA-DNA interactions

Multiple read pairs that have the same mapped RNA and DNA loci associated with the same PCR primer sequences were considered PCR duplicated, and thus were counted only once. The following equation was used to compute RNA-DNA interactions in each AluI DNA bin:

$$Num_{(interaction)} = \sum_{i=1}^n (R_i * D_i)$$

Where n is the number of assigned read pairs. For each read pair i , R_i is the contribution score of this read to the RNA part and D_i is the contribution score of this read to the interacting AluI DNA bin. $R_i = 1$, if the RNA part of the read pair is uniquely mapped, or equals to a fraction based on the weighted score.

Construction of non-specific background using mixed GRID-seq libraries

A mixed GRID-seq library from human MDA-MB-231 and *Drosophila* S2 cells was used to construct the non-specific RNA-DNA interaction profile. RNA reads, which were only mapped to the human genome (hg38) using Bowtie (Langmead et al. 2009) with the parameter *-n 0*, were kept. Their mated DNA reads were processed using ShortStack, as described above. The human RNA signals within each 1kb DNA bin were normalized to one million, which was further smoothed by a moving window that includes 5 upstream and 5 downstream bins. The final coverage of the 1kb DNA bin i is:

$$Cov_i = \frac{1}{11} \sum_{i-5}^{i+5} \left(\frac{10^6 * \sum_m Read_{im}}{N} \right)$$

where m is the number of reads mapped to the 1k DNA bin i and N is the total read number mapped to the *Drosophila* genome.

To make DNA binding scores comparable between Alu I binned versus 1kb binned genome, RNA binding signals in each 1k DNA bin were converted to RNA binding signals in each AluI DNA bin by first dividing each 1kb DNA bin into 1000 1bp bins to calculate the signal in each small bin based on $\frac{Cov_{bin}}{1000}$. IntersectBed tool with the parameter *-f 1.0* was used to compute signals in each AluI DNA bin by summing the signals from all 1bp bins in the fragment. Finally, signals in AluI DNA bins were all normalized to signals *per* 1kb:

$$Score_{AluI\ DNA\ bin\ j\ in\ mixed\ library} = \frac{(10^3 * \sum_{m=1}^{Len_j} Cov_m)}{Len_j}$$

where Len_j is the length of the AluI DNA bin j .

Filtering singular and background to identify specific RNA-DNA interactions

To support specific RNA-DNA interactions, we required at least two RNA-DNA mates for each RNA transcript in a given AluI DNA bin. We also developed two background models to simulate Poisson distribution of RNA binding signals on DNA. The first was based on uniform distribution of individual RNAs on DNA, based on which we estimated the background score:

$$Score_{interaction\ i-j} = \frac{(Len_j * N_i)}{Len_g}$$

where the Len_g is the length of the genome, Len_j is the length of AluI DNA bin j and N_i is total signals for RNA i . This score was further normalized according to length in each AluI DNA bin, and the resulting RPK (reads *per* kilobase) was used as the λ_{BG_i} value to obtain the Poisson distribution of this RNA-engaged genomic interactions and to calculate the p-value for such interactions. The ratio of $Num_{interaction\ i-j}$ over $Score_{interaction\ i-j}$ was reported as the fold_change (FC) above the background. We also developed a second background model with that data deduced with RNA signals from the mixed library. We first calculated the non-specific interaction score ($Score_{AluI\ DNA\ bin\ j\ in\ mix\ library}$) based on human-derived RNA binding signals, and then used this score as the λ_{BG_mix} value to obtain the Poisson distribution of human RNA-engaged genomic interactions and to calculate the p-value for such interactions. The ratio of length and sequencing depth normalized $Num_{interaction\ i-j}$ (in RPKM) over λ_{BG_mix} was reported as the fold_change (FC). RNA-DNA interactions that met the requirement of p-value <0.05 and the fold_change (FC) >2 based on both background models were considered specific and thus retained for further analysis.

Data normalization for comparison between GRID-seq libraries and different RNAs within the same libraries

The interaction RNA-DNA score for each RNA is affected by the sequencing depth in different GRID-seq libraries, the length of each AluI DNA bin, and the length of each RNA. To enable comparison among different libraries and different RNAs within the same libraries, we normalized these variables according to:

$$RRPKM_{ij} = \frac{(Num_{ij} * 10^6 * 10^3 * 10^3)}{(TotalReadCounts * Len_{Dj} * Len_{Ri})}$$

where Num_{ij} is the interaction score of RNA i with Alu I DNA bin j , $TotalReadCounts$ is the sequencing depth of individual libraries, Len_{Dj} is the length of AluI DNA bin j and Len_{Ri} is the length of RNA i .

Comparison between GRID-seq and ChAR-seq datasets

ChAR-seq raw data were downloaded from the GEO database (Supplemental Table S2). PCR duplicates were removed using Clumpify with default parameters. All five independent ChAR-seq libraries were combined followed by adapter trimming and filtering low-quality reads with Trimmomatic (Bolger et al. 2014) using the parameters *MINLEN: 36, LEADING:3 TRAILING:3* and *SLIDINGWINDOW: 4:15*. Filtered ChAR-Seq reads were split into paired RNA and DNA reads, and then processed as described above for the GRID-seq data. FeatureCounts (Liao et al. 2014) was used to calculate reads in 1kb DNA bins and then converted to RPKM using edgeR

package (Robinson et al. 2010) in R language. For validation for assigning multi-mapped reads, we only considered 1kb DNA bins that contain newly assigned multi-mapped reads from the GRID-seq dataset.

Comparison of RNA-DNA interactions in relationship with chromatin marks

Public ChIP-seq data from S2 cells were downloaded from the SRA database (Supplemental Table S2). Fastq-dump were used to covert raw SRA data to Fastq. Quality control and data processing were similar to the procedures for processing the GRID-seq data with parameters adjusted to *MINLEN: 36* and *SLIDINGWINDOW: 4:20* according to the read length. Filtered reads were mapped to the reference *Drosophila* genome (dm6) using STAR (Dobin et al. 2013) with the parameters: *--outFilterScoreMinOverLread 0.1, --outFilterMatchNminOverLread 0.1, --alignIntronMax 1* and *--alignEndsType EndToEnd*. The wig files of each chromatin mark ChIP-seq dataset was obtained through STAR using the parameters: *--outWigType wiggle read1_5p, --outWigStrand Unstranded* and *--outWigNormRPM*. Enriched peaks were detected by MACS2 (Zhang et al. 2008) with input data as control. Top 500 peaks were used for further analysis.

Identification of repeat RNA on constitutive heterochromatin

The ChIP-seq data for constitutive heterochromatin markers (H3K9me3 and Su(var)205) were obtained from the GEO database (Supplemental Table 2) and the RPKM values were calculated on 10kb DNA bins. The Pearson correlation score was calculated on each DNA bin containing binding signals from repeat RNAs. Repeat sequences, except rRNAs, were defined as CHARRs if the correlation score with H3K9me3 or Su(var)205 is >0.3 .

Define cis or trans interacting repeat-derived RNAs

As illustrated in Supplemental Fig. S5C, we link the RNA part (yellow) and the DNA part (blue) from GRID-seq reads. Using *gypsy4_I-int* as an example, we first assign specific GRID-seq reads to the genome to obtain the mapped RNA part (which corresponds to its origin of transcription) and then the DNA part (which corresponds to the RNA bound genomic region). In general, we refer *cis*-acting RNAs as those interacting with DNA in the same chromosomes (note that some may result from their actions in *trans*) and *trans*-acting RNAs as those interacting with DNA in different chromosomes. In the *gypsy4_I-int* case, its RNA part was mapped to Chr2L:21545817-21545837 (*gypsy4_I-int* annotation region), but its DNA parts could be mapped to Chr2R:5238402-5238421 and ChrX:16415244-16415263, the former corresponding to *cis*-interaction and the latter *trans*-interaction. We normalize the total *cis*-reads and *trans*-reads to RPK (reads *per* kilobase) to determine whether a given CHARR prefers *cis*- over *trans*-interactions.

Analysis of CHARR-DNA interactions relative to Hi-C defined compartments

Public Hi-C data from S2 cells were downloaded (Supplemental Table S2) and processed with Trimmomatic using default settings plus trim tool in Homer using the parameter *-3 AAGCTT*. Trimmed reads with length of ≥ 38 nt were mapped to the *Drosophila* genome using end-to-end alignment model provided by Bowtie2 (Langmead and Salzberg 2012). We discarded potential PCR duplicates as well as reads with no useful information, including (1) read pairs separated $<1.5\times$ of the sequenced insert fragment length, (2) reads from 10kb regions containing $>5\times$ of the average coverage, (3) read pairs lacking restriction sites at the 3' end of either read within the estimated fragment length, (4) reads with their ends resulting from self-

ligation with adjacent restriction fragments. We then used the filtered dataset to perform PCA analysis with HOMER, using runHiCpca.pl with the parameters *-res 10kb* and *-superRes 20kb* and compartments analysis using runHiCpca.pl and findHiCCompartments.pl. Finally, CHARR-DNA interactions signals and Hi-C compartments were intersected with IntersectBed from BedTools. The length of Hi-C compartment A and B was normalized to 1Mb.

Processing GRO-seq, small RNA-seq and AGO2 RIP data from S2 cells

All data were downloaded from public databases (Supplemental Table S2). GRO-seq data were processed as with the ChIP-seq data for chromatin marks. Annotated unique and repeat RNA species were used to calculate individual transcription scores by using FeatureCounts and EdgeR.

Small RNA-seq data were similarly processed as above using adjusted Trimmomatic parameters: *SLIDINGWINDOW:4:24* and *MINLEN:18*. SortMeRNA (Kopylova et al. 2012) were used to filter ribosomal RNAs with default setting. We also used ShortStack to map both uniquely and multi-mapped reads to obtain the RPM value for each CHARR.

Reads for the downloaded AGO2 RIP data were blasted to identify repeat RNA-derived sequences using the parameters *-outfmt 6* and *-word_size 7*. Reads will be kept for further analysis if the mismatch number is under 3 and the mapping region of the reads equals to the whole length of the reads. We then counted the reads number for each repeat sequence.

ChIP-seq library construction and data analysis

DNA libraries were constructed using the NEBNext® Ultra™ II DNA Library Prep kit (NEB, USA) following manufacturer's recommendations. After end repair, 5' phosphorylation, and dA-tailing of purified DNA fragments, NEBNext adaptors with hairpin loop structure were ligated and library fragments were purified with SPRIselect sample purification beads (NEB, USA). After ensuring the quality of libraries on the Agilent Bioanalyzer 2100 system, individual libraries were sequenced on the Illumina HiSeq X Ten platform to generate 150 bp paired-end.

Adapters and low-quality reads were filtered to obtain clean reads and clean reads mapped to the reference *Drosophila* genome (dm6) using Bowtie2. The wig files of each chromatin mark ChIP-seq dataset was obtained by using bamCoverage tools from Deeptools (Ramirez et al. 2016) with the parameters: *--binSize 1000*, *--normalizeTo1x 142573017* and *--ignoreForNormalization chrM*. Enriched peaks were detected by MACS2 with input data as control using the parameters *-f BAMPE --nomodel --keep-dup all --broad*.

Immunofluorescence, Western blotting, RT-qPCR and Northern blotting

For immunostaining, S2 cells were washed in 1× PBS and fixed in 4% paraformaldehyde (pH-7.2) for 10 min at room temperature. After washing four times, cells were permeabilized with 0.1% Triton X-100 in 1× PBS for 5 min at room temperature. Permeabilized cells were then incubated in 1% normal goat serum in 1× PBST for 30 min at room temperature, and then with the primary antibody (α -rabbit H3K9me3 1:500) and secondary antibodies (1:400) both in blocking buffer (3% BSA, 1% goat serum in PBST), each for 1h at room temperature. Cells were washed three times at room temperature, each for 5 min with 1× PBST. After mounting on coverslip with DAPI, cells were examined under Zeiss LSM-700 confocal laser scanning microscope.

Total cell lysate containing 15-25 µg protein from S2 cell cultures were fractionated by SDS-PAGE, immunoblotted and probed with specific antibodies. After incubation with peroxidase-conjugated secondary antibodies (1:5000; Abcam), blots were developed with Supersignal West Pico Chemiluminescent Substrate (Pierce) and exposed to film (SAGECREATION, MiNiChemi). Signal intensity was quantified using ImageJ.

For RNA quantification, total RNA was extracted, and genomic DNA was removed with DNase I (Roche, 04716728001). First-strand cDNA was generated with the SuperScript III First-Strand using random hexamers. The expression levels of RNAs were quantified on Rotor-Gene Q (QIAGEN) and normalized against GAPDH mRNA. PCR primers sequences were listed in Supplemental Table S3.

For Northern blotting analysis, ~30 µg of total RNA isolated with TRIzol was loaded into each lane of agarose gel and blotted onto membrane with Chemiluminescent Nucleic Acid Detection Module (Thermo Fisher) according to manufacturer's instruction. Synthetic biotin-label RNA probes were used for hybridization. The primers used were listed in Supplemental Table S3.

REFERENCES

- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.
- Drysdale R, FlyBase C. 2008. FlyBase : a database for the Drosophila research community. *Methods Mol Biol* **420**: 45-59.
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**: 462-467.
- Kopylova E, Noe L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**: 3211-3217.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.
- Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-165.
- Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-277.
- Roberts RJ, Vincze T, Posfai J, Macelis D. 2015. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **43**: D298-299.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.