Westholm et al, Genomewide analysis of *Drosophila* circular RNAs reveals
their structural and sequence properties and age-dependent neural accumulation

Supplementary Data

Supplementary Figures

Figure S1. Fraction of circular RNAs with GT/AG vs read support, Related to Figure 1.

(A) Analysis of all out-of-order mapped candidate junctions demonstrates that most of the highest-expressed loci were flanked by consensus splice sites. The highest-expressed loci not flanked by splice sites were predominantly composed of rRNA and chrU repetitive loci. (B) Enrichment analysis of splice site consensus sequences amongst out-of-order junction-spanning reads (not including rRNA and chrU loci), shows that the majority of loci with >100 reads were flanked by consensus splice sites.

Figure S2. Classes of out-of-order junction patterns not found at known splice sites, Related to Figure 1.

(A) Examples of non-specific out-of-order junction patterns. Instead of the specific patterns of out-of-order junction-spanning reads found at splice sites, some genes had a population of heterogenously mapping reads with highly non-specific out-of-order junctional patterns. Shown are the top five loci, which collectively generated 3075 out-of-order junctional reads. (B) Examples of specific spliced circles from unannotated splices in intronic or intergenic regions. These loci generated specific out-of-order reads, flanked by canonical GT/AG splice sequences, that were not part of annotated exons.

Figure S3. Circular RNAs are flanked by long introns, independently of the Drosophila bias for long first introns, Related to Figure 3.

We analyzed flanking intron lengths of circular RNAs involving second exons, compared against all other circular RNAs. This analysis shows that many circular RNAs do not involve second exons. Moreover, amongst circular RNAs not involving second exons, there is a statistically significant trend for progressively longer upstream and downstream flanking introns amongst higher-expressed circular RNAs.

Figure S4. Sequence motifs enriched from intronic regions flanking *Drosophila* and mammalian circular RNAs, Related to Figure 3.

(A) Motifs enriched in 500nt windows flanking mammalian circular RNAs. For comparison, shown as the top is the ALU motif described by Jeck (2012) to be enriched in the intronic flanks of circular RNAs. Our de novo MEME analysis identified several motifs that overlap substrings within the ALU sequence. (B) Motifs enriched in 500 nt windows flanking Drosophila circular RNAs annotated in this study, along with motifs identified in a set of control intronic regions flanking non-circular RNAs. We identified only simple repeats and consensus splice sequences, and these were not different between circular and non-circular loci.

Figure S5. Conservation and duplexing properties of circular RNAs, Related to Figure 3.

(A) No overall differences were observed between the PhastCons conservation profiles in the vicinity of intron/exon boundaries of the starts and ends of circular RNAs vs. control non-circularizing exons. (B) Analysis of the amount of duplexing between the intronic regions flanking the circularizing exons, assayed for various sizes of flanking intronic windows. We observed a modest degree of greater duplexing between flanking introns of circular RNAs compared to control exons, which was most prominent when considering shorter window sizes (20nt and 50nt). (C) Upon performing the duplexing analysis by segmenting the circular RNAs by level, however, we did not observe any correlations between the degree of flanking structure and circle accumulation, suggesting that flanking complementarity may not be a primary determinant for circularization. (D) When performing the flanking duplex analysis by binning circular RNAs according to G:C content in the flanking 20 nt intronic regions, it appears that circular RNAs with higher G:C content are associated with greater duplex structure than control exon flanks. (E) However, when perfoming the flanking duplex analysis by segregating according to both flanking G:C content and circular RNA accumulation, increased G:C content/flanking duplex was not correlated with greater circular RNA levels. (F) Analysis of the amount of duplexing between the exonic termini that become juxtaposed in circular RNAs shows there is slightly less pairing amongst circular RNAs than for controls RNAs, especially when considering the terminal 20-50 nts of the exons.

Figure S6. Elevated numbers of loci generate abundant circular RNAs in heads, Related to Figure 6.

The X-axis is presented in log[10] scale, and the dotted line marks genes for which back-splicing generates $\geq$10% of total spliced reads.

Figure S7. Age-dependent accumulation of circular RNAs compared to their host mRNAs in Drosophila heads, Related to Figure 7.

The X-axes represent host gene RPKM compared between the indicated age comparisons, and the Y-axes represent circular RNA junction spanning reads normalized to raw library size. The top graphs are for female heads and the bottom graphs are for male heads; the left graphs are for 4 days vs. 1 day and the right graphs are for 20 days vs. 1 day. There is a mild shift towards higher circular RNA accumulation in the 4 day female head data (p<2E-16) but not in 4 day male data (p=0.21), but there is a substantial shift for increased circular RNA accumulation relative to host mRNAs in both sexes by 20 days (p<2E-16 both sexes).

Supplementary Tables

Table S1. *Drosophila* total RNA-seq and mRNA-seq datasets analyzed in this study, Related to Figure 1.

These tables summarize NCBI Short Read Archive (SRA) and Gene Expression Omnibus (GEO) IDs, mapping statistics, circle numbers for *D. melanogaster*, *D. yakuba*, and *D. virilis* total RNA-seq and/or mRNA-seq datasets analyzed in this study.

Table S2. Annotation of circular RNAs in three *Drosophila* species, Related to Figure 1.

These tables summarize coordinates and expression levels of circular RNAs annotated from *D. melanogaster*, *D. yakuba*, and *D. virilis.* Additional detailed information is provided on the associated genes in *D. melanogaster*, whose genome is better-annotated than the other species.

Table S3. Genomewide comparisons of circular reads, Related to Figures 1 and 7.

These tables summarize the following analyses conducting for individual circularizing loci, from which genomewide assessments were made. (1) Comparison of back-spliced reads in matched *Drosophila melanogaster* total RNA-seq and mRNA-seq libraries. (2) Comparison of forward and back-spliced reads at circularizing splice junctions across the aggregate *Drosophila melanogaster* total RNA-seq data. (3) Comparison of forward and back-spliced reads at circularizing splice junctions across *Drosophila melanogaster* head total RNA-seq data only.

Table S4. Conserved miRNA binding sites within Drosophila circular RNAs, Related to Figure 5.

Shown are the numbers of various types of miRNA binding sites in circular RNAs, binned into different genomic classes. Note that intronic regions are shown for comparison only, and were not considered in the main tabulation of miRNA sites on circular RNAs as we showed that circular RNAs are predominantly spliced. Pan-Drosophilid conservation required that the site be present in 11/12 species in the genomewide alignments.

Table S5. Normalized circular RNA expression across all libraries, Related to Figures 6 and 7.

For each circular RNA locus, we tabulate the levels of circular RNA junction reads in each individual library, normalized per million raw reads in each dataset. The second tab summarizes information on those circular RNAs with significantly increased accumulation relative to host mRNAs during head aging.

Table S6. Functional and expression domain enrichments amongst circular RNAs, Related to Figure 6.

These tables summarize Gene Ontology, FlyAtlas, In situ and modENCODE cluster enrichment analyses performed on "all circles", on circles annotated from only "0-2 hr embryos", and from circles annotated only from "S2 cell" datasets, as noted on each tab.

Table S7. Primers used for experimental validations of circular RNAs, Related to Figure 2.

Supplementary Experimental Procedures

Annotation of *Drosophila melanogaster* circular RNAs

We identified circular RNAs from *Drosophila melanogaster* 100nt-PE total RNA-seq using a custom computational pipeline that uses the STAR read aligner (Dobin et al., 2013). Reads were aligned using the following parameters to identify chimeric transcripts: --chimSegmentMin 20 --chimScoreMin 1 --alignIntronMax 100000 --outFilterMismatchNmax 4 --alignTranscriptsPerReadNmax 10000 --outFilterMultimapNmax 2. Thus, at most 3 mismatches were tolerated for each read pair, and only unique mappers were used. The putative chimeric junction reads were then filtered to only include cases where one read in a pair spanned a junction with the splice acceptor on the same chromosome and strand as the splice donor, at most 100,000 bp upstream. In addition, mapping of the other read in the pair had to be consistent with circular RNA formation, i.e. between the splice donor and acceptor, and on the same strand. The resulting set of junction-spanning reads were then collapsed into a set putative circularization junctions. In the subsequent analysis only circular junctions matching GT-AG splice sites, not on chrU or chrUextra, filtered for "internal CDS" events, and supported by at least 10 reads were considered. The scripts used for annotating the circles are available at https://github.com/orzechoj/circRNA_finder.git.

Conservation of circular RNAs in other *Drosophila* species

We utilized total RNA-seq data from *D. yakuba* and *D. virilis* heads that will be described in detail elsewhere (P.S., S.S., E.C.L., in preparation). These data have been submitted to the NCBI Gene Expression Omnibus under GEO-IDs summarized in Table S1. We recognized two main issues that complicated the direct usage of the *D. melanogaster* pipeline on the other species. First, neither the assembly nor gene annotations of the other genomes are as complete as in *D. melanogaster*. Second, the *D. melanogaster* data were paired end 100 nt reads, whereas the other data were paired end 75 nt reads. Therefore, we recovered disproportionally fewer back-spliced reads when using the same mapping requirements in the other species, above and beyond the fact that we had eighteen head datasets in *D. melanogaster* and only one each in the other species.

To facilitate a fairer comparison of these data, we sought to recover additional circular RNAs from the species data. We were unable to do so effectively by relaxing the mapping stringency of the STAR aligner, since it is not suited for distinguishing potential multi-mapping of split reads. Instead, we supplemented our recovery of confident back-spliced reads in the other species by mapping directly to an index of all possible intra-gene back-spliced junctions. To do so, we filtered FASTQ to identify genome-aligning reads, and removing reads that match contiguously as well as across annotated splice junctions. The remaining unmapped reads were aligned to the back-splice reference sequence using Bowtie2, requiring that these reads spanned the back-splice junction by at east 15nt on each side. We confirmed that this relaxed cutoff identified bona fide back-splicing events with high stringency, because only 1-2% of mate-pair reads mapped inconsistently (i.e., that mapped outside the inferred circle, see below). To focus on confident alignments, we filtered the data to remove reads that aligned with any mismatches or in-dels, or that had an inconsistently mapped mate-pair.

To assess the extent to which *Drosophila* genes have conserved propensity to generate circular RNAs, we filtered the *D. yakuba* and *D. virilis* head data for circles supported by at least 2 back-spliced reads, and compared them to *D. melanogaster* circles supported by at least 10 head back-spliced reads. We associated these circles (*Dmel*=2147 circles, *Dyak*=1436 circles and *Dvir*=1934 circles) to their respective gene models using Flybase releases: *Dmel* r5.51, *Dyak* r1.2 and *Dvir* r1.2. Annotation was done with the R package Genomic Ranges using the findOverlaps function with the 'type' parameter set to any and ignore.strand = false (so any strand-specific overlap between circle and gene was acceptable, however, each gene was counted once). We then applied a filter to identify parent genes of the circular RNAs in each species that are associated with 1:1:1 orthologs in the three species, using the OrthoDB7.FlyBase.txt downloaded from http://cegg.unige.ch/orthodb7. From this, we tabulated the number of genes that generated circular RNAs in one, two or three of the *Drosophila* species.

Mate-pair consistency analysis

We assessed the frequency with which back-splice spanning reads are mated to reads that are inconsistent with the circular RNA interpretation. However, STAR does not report alignments for mates that align to separate chromosome, or on the same chromosome but entirely outside of the back-spliced alignment. To assess the degree of mate-pair inconsistency, we aligned 18 *D. melanogaster* head libraries using STAR, by mapping the 1st and 2nd read pairs to the genome independently.

The chimeric-mapping BAM output files from STAR were filtered to identify reads that span identified circle junctions. Next, using the read name as an identifier, we sought reported alignments for the mate of back-splice spanning reads in the genome-mapped BAM output files from STAR. For each read spanning a back-splice junction, we sought its mate, and evaluated the mate's consistency with the respective junction. If the mate was not unmapped, then it could be (1) mapped across the same back-splice, (2) mapped entirely within the boundaries of the circle, or (3) partially or entirely outside the boundaries of the circle. We considered all reads in categories 1 and 2 to be consistent with the circle. Since reads that span a splice junction by only a few nucleotides will be flanked by a short sequence that cannot be mapped unambiguously, we considered mates in category 3 to be consistent if the read mapped to the same chromosome and strand as the circle, and fewer than 15 nucleotides aligning outside the boundaries delimited by the back-splice junction. All mapped mates that were not classified as consistent by these criteria were considered inconsistent. A similar classification approach was used to evaluate direct-mapping outputs obtained with the *D. yakuba* and *D. virilis* data.

Assessment of secondary structures between flanking intronic regions of circular RNAs

We assessed further if there might be enrichment of secondary structures formed between the introns flanking circularizing exons. We used RNAduplex (Lorenz et al., 2011) to compare the extent of pairing between intron pairs that flanked circles compared to control exon pairs. We compared windows of 20bp, 50bp, 100bp, 200bp, and 500bp, and observed only very mild differences between these sets (**Figure S4B)**. However, there appeared to be an overall trend for modestly greater duplexing between intronic regions flanking circular RNAs compared to control, especially when considering shorter window sizes (20 and 50 bp). However, such trends did not correlate at all with circular RNA accumulation. That is, there are specific bins of circular RNAs for which there is statistically greater duplexing between flanking introns of particular length windows, but invariably these do not encompass the higher-expressed sets of circles (**Figure S4C)**. Moreover, there is not progressive trend between adjacent bins of circle levels.

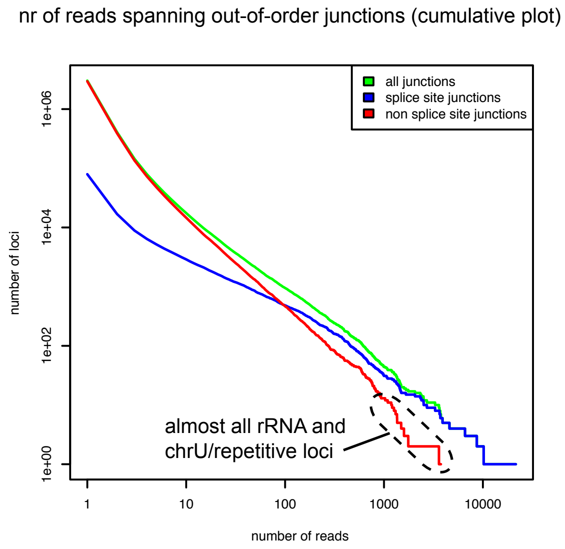
As G:C pairing has disproportionate influence on predicted pairing, we redid this analysis by binning circles and control sets for G:C content in the flanking 20 bp. This analysis appeared to show that circularizing exons had slightly greater pairing when controlling for G:C content (**Figure S4D)**. However, when we co-stratified circles for flanking G:C content and their level of accumulation, we again did not observe any correlation in which higher-expressed circles might be preferentially associated with greater pairing (**Figure S4E)**. Therefore, the potential trend for increased pairing between local intronic regions that flank circles does not appear to facilitate the process of circularization. This is in contrast to other features that exhibit strong progressive correlation with circle accumulation, such as exon position and especially length of flanking upstream and downstream introns.

miRNA target site density

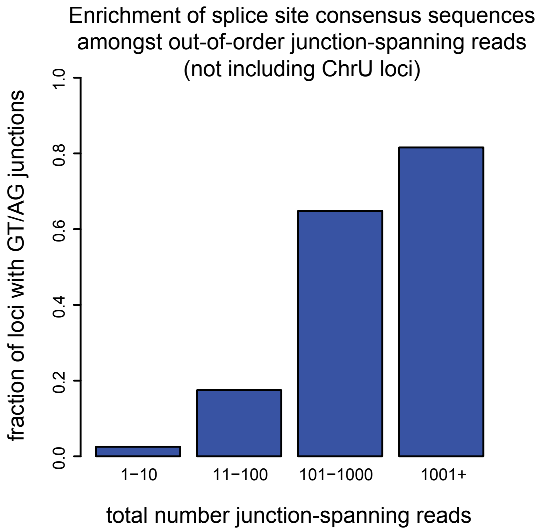
Whole genome multiple alignments (.maf) were downloaded from UCSC and scanned to identify all instances of conserved 7mers. Since selective forces operating on coding and non-coding sequences are quite varied, and back-splice events are frequently detected at internal exons of protein coding genes, we focused on the portion of circles that overlap protein coding sequence. We considered a 7mer that was aligned (without any gaps or mismatches in 11 out of 12 genomes), to be conserved. From these 7mers, 94 independent sequences (i.e. with miRNA seed families each counted only once) corresponding to 7m8 target sites of conserved miRNAs (miRNAs conserved from *D. melanogaster* to either *D. grimshawi* or *D. virilis*) were selected. For comparison, we also examined frequency of the antisense of these 94 target sites, as well as a set of 94 di-nt matched control 7mers, and 118 target sites of the star-strand of conserved miRNAs.

To compare miRNA target density in circular RNAs compared to linear portions of the genome, we partitioned the genome into segments depending on whether it overlaps an annotated circle. Regions overlapping circles were further stratified by the number of reads detected for each circle, such that each bin covered roughly equal genomic space. We included circles that did not meet our confidence threshold, with 1-9 reads, as a separate bin. We compared CDS regions overlapping circles with CDS regions that don’t overlap circles in the remainder of the genome, linear portions of CDS from genes that circularize, and the canonical site of miRNA targeting in 3' UTRs.

For each of these categories, we estimated the conserved site density (# sites/kb/7mer) by dividing the total number of conserved 7mer ***n*** by the length of the region examined. We treated ***n*** as an observation of the Poisson-process, and used an exact 95% Poisson confidence intervals were used to assign uncertainty to these estimates.

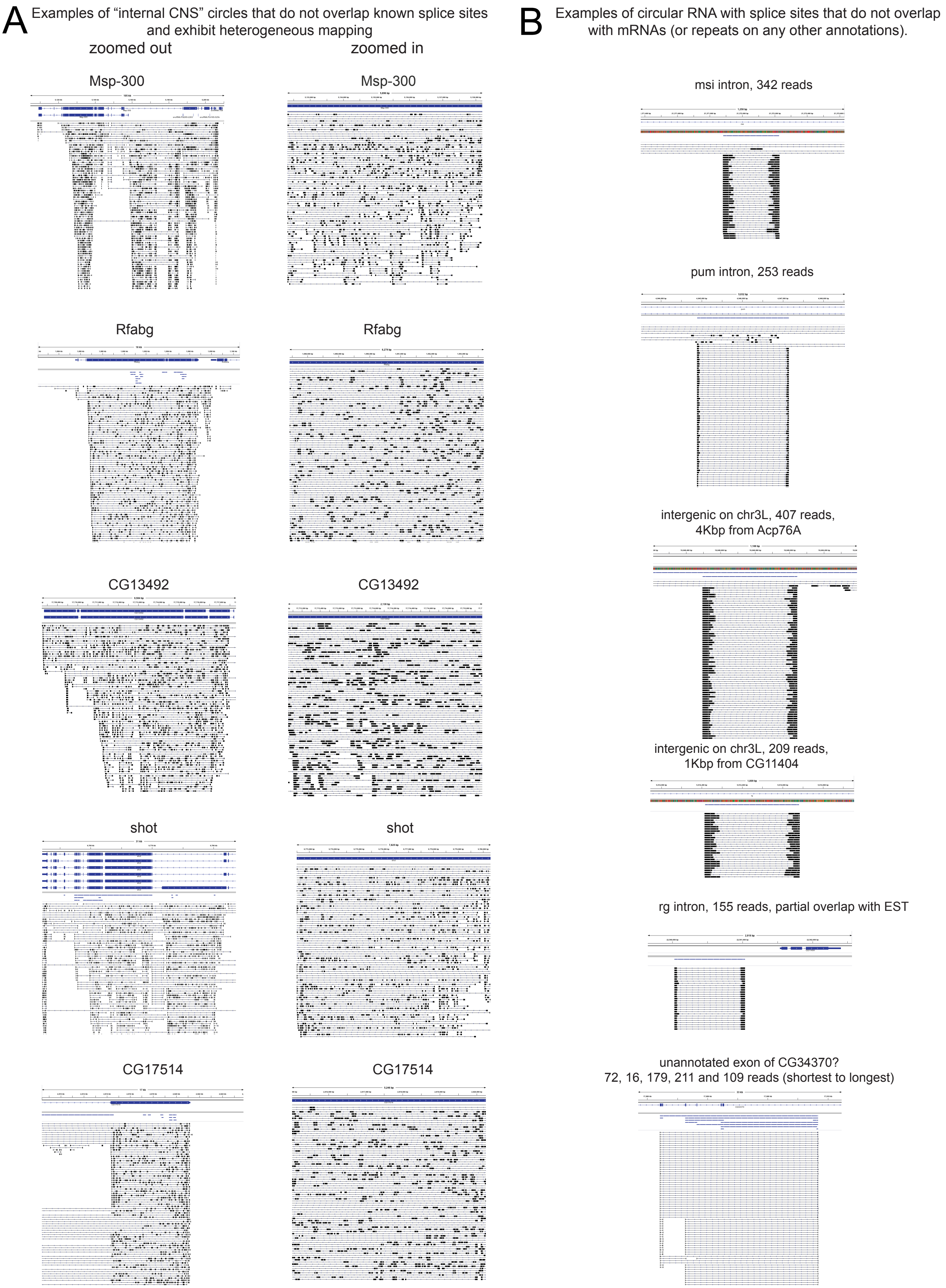


nr of reads spanning out-of-order junctions (cumulative plot)



Enrichment of splice site consensus sequences amongst out-of-order junction-spanning reads (not including ChrU loci)

Westholm et al
Supplementary Figure 1



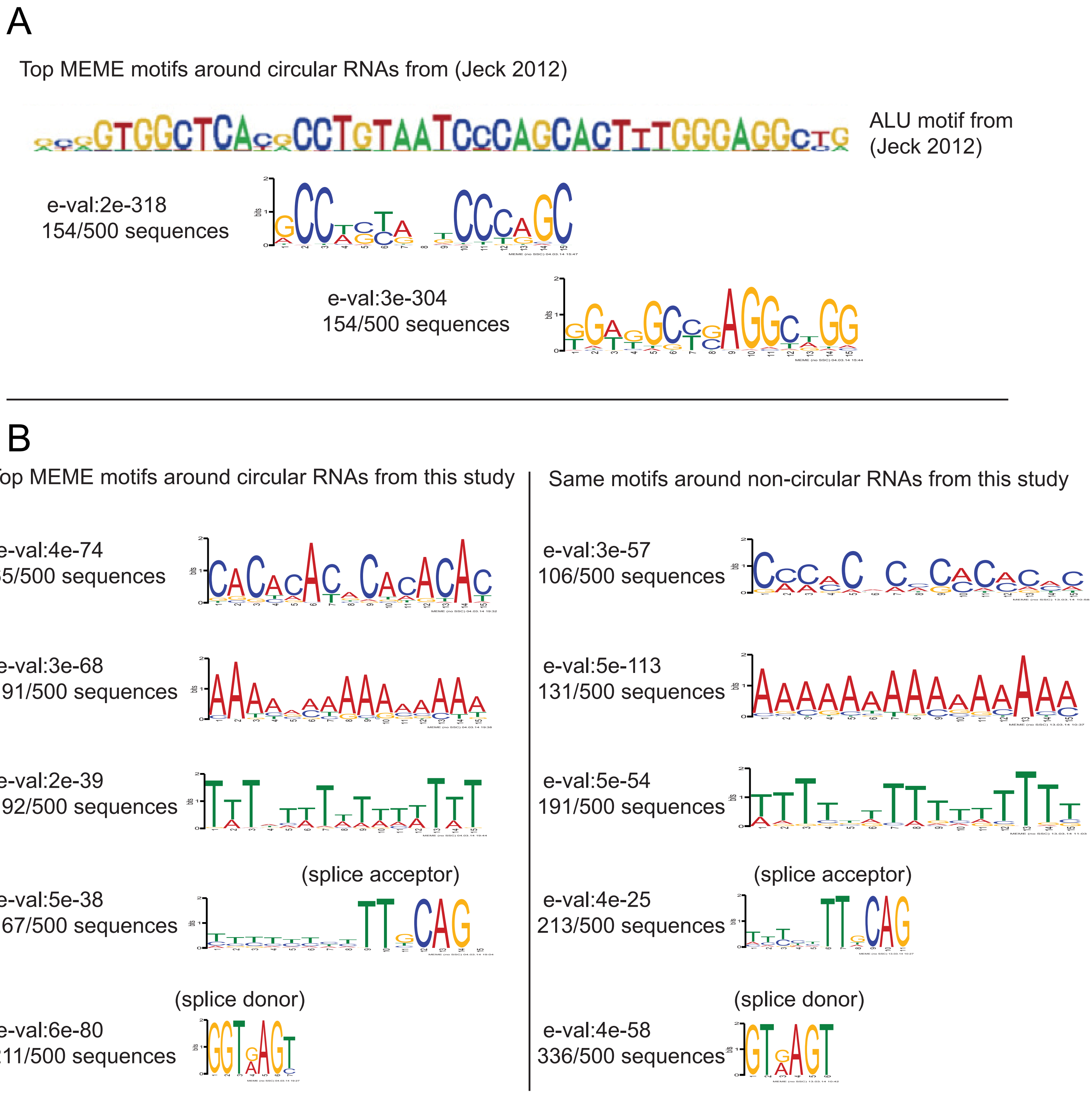
A Examples of “internal CNS” circles that do not overlap known splice sites and exhibit heterogeneous mapping

zoomed out
zoomed in
Msp-300
Msp-300
Rfabg
Rfabg
CG13492
CG13492
shot
shot
CG17514
CG17514

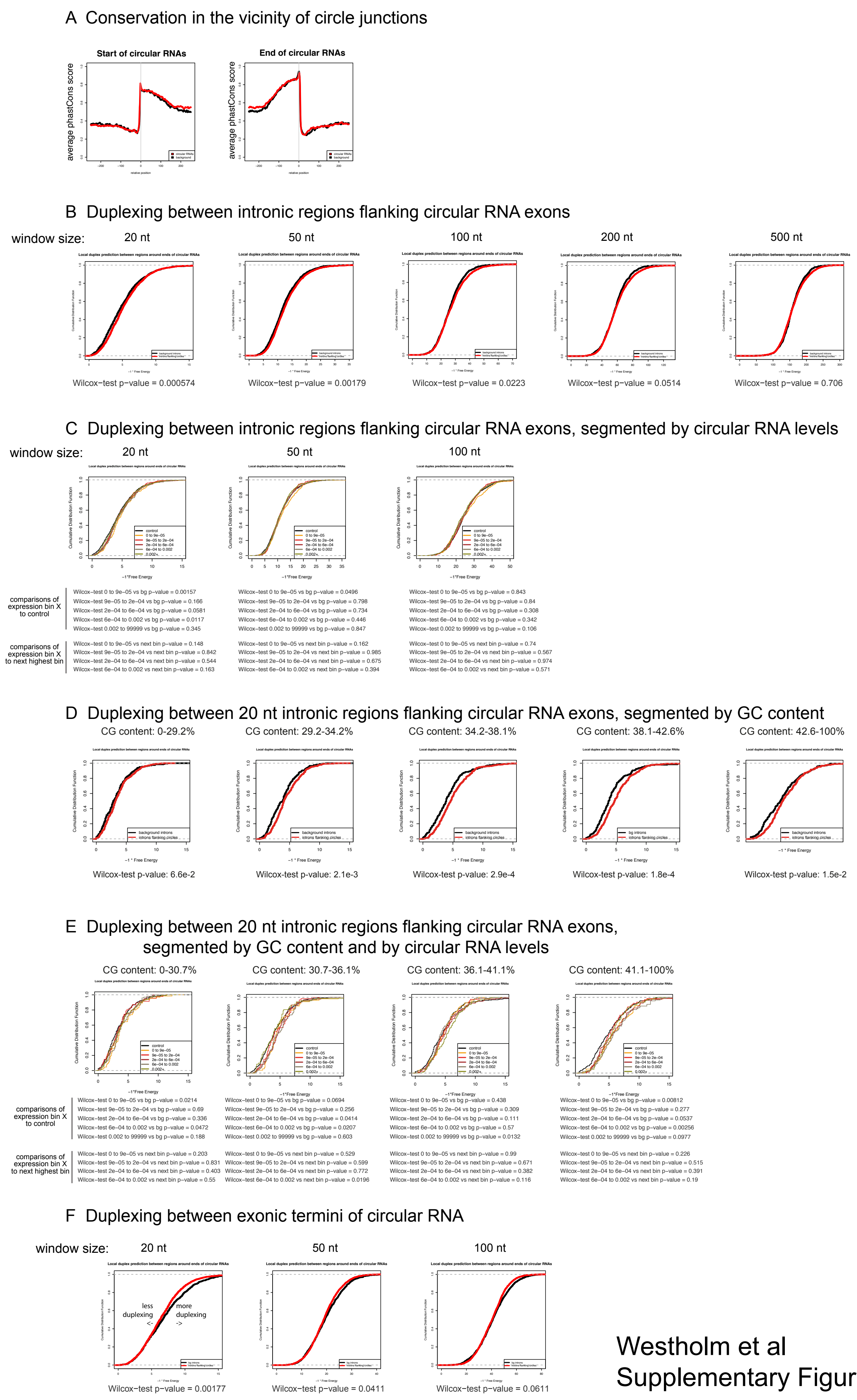
B Examples of circular RNA with splice sites that do not overlap with mRNAs (or repeats on any other annotations).

msi intron, 342 reads
pum intron, 253 reads
intergenic on chr3L, 407 reads,
4Kbp from Acp76A
intergenic on chr3L, 209 reads,
1Kbp from CG11404
rg intron, 155 reads, partial overlap with EST
unannotated exon of CG34370?
72, 16, 179, 211 and 109 reads (shortest to longest)

Westholm et al
Supplementary Figure 2



Westholm et al
Supplementary Figure 3



A Conservation in the vicinity of circle junctions

B Duplexing between intronic regions flanking circular RNA exons

Wilcox−test p−value = 0.000574

Wilcox−test p−value = 0.00179

Wilcox−test p−value = 0.0223

Wilcox−test p−value = 0.0514

Wilcox−test p−value = 0.706

C Duplexing between intronic regions flanking circular RNA exons, segmented by circular RNA levels

comparisons of expression bin X to control

20 nt:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.00157
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.166
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.0581
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.0117
Wilcox−test 0.002 to 99999 vs bg p−value = 0.345

50 nt:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.0496
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.798
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.734
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.446
Wilcox−test 0.002 to 99999 vs bg p−value = 0.847

100 nt:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.843
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.84
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.308
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.342
Wilcox−test 0.002 to 99999 vs bg p−value = 0.106

comparisons of expression bin X to next highest bin

20 nt:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.148
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.842
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.544
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.163

50 nt:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.162
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.985
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.675
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.394

100 nt:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.74
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.567
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.974
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.571

D Duplexing between 20 nt intronic regions flanking circular RNA exons, segmented by GC content

Wilcox-test p-value: 6.6e-2

Wilcox-test p-value: 2.1e-3

Wilcox-test p-value: 2.9e-4

Wilcox-test p-value: 1.8e-4

Wilcox-test p-value: 1.5e-2

E Duplexing between 20 nt intronic regions flanking circular RNA exons, segmented by GC content and by circular RNA levels

comparisons of expression bin X to control

CG content: 0-30.7%:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.0214
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.69
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.336
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.0472
Wilcox−test 0.002 to 99999 vs bg p−value = 0.188

CG content: 30.7-36.1%:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.0694
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.256
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.0414
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.0207
Wilcox−test 0.002 to 99999 vs bg p−value = 0.603

CG content: 36.1-41.1%:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.438
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.309
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.111
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.57
Wilcox−test 0.002 to 99999 vs bg p−value = 0.0132

CG content: 41.1-100%:

Wilcox−test 0 to 9e−05 vs bg p−value = 0.00812
Wilcox−test 9e−05 to 2e−04 vs bg p−value = 0.277
Wilcox−test 2e−04 to 6e−04 vs bg p−value = 0.0537
Wilcox−test 6e−04 to 0.002 vs bg p−value = 0.00256
Wilcox−test 0.002 to 99999 vs bg p−value = 0.0977

comparisons of expression bin X to next highest bin

CG content: 0-30.7%:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.203
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.831
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.403
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.55

CG content: 30.7-36.1%:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.529
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.599
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.772
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.0196

CG content: 36.1-41.1%:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.99
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.671
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.382
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.116

CG content: 41.1-100%:

Wilcox−test 0 to 9e−05 vs next bin p−value = 0.226
Wilcox−test 9e−05 to 2e−04 vs next bin p−value = 0.515
Wilcox−test 2e−04 to 6e−04 vs next bin p−value = 0.391
Wilcox−test 6e−04 to 0.002 vs next bin p−value = 0.19

F Duplexing between exonic termini of circular RNA

Wilcox−test p−value = 0.00177

Wilcox−test p−value = 0.0411

Wilcox−test p−value = 0.0611

Westholm et al
Supplementary Figure 4

5' bias vs upstream intron length

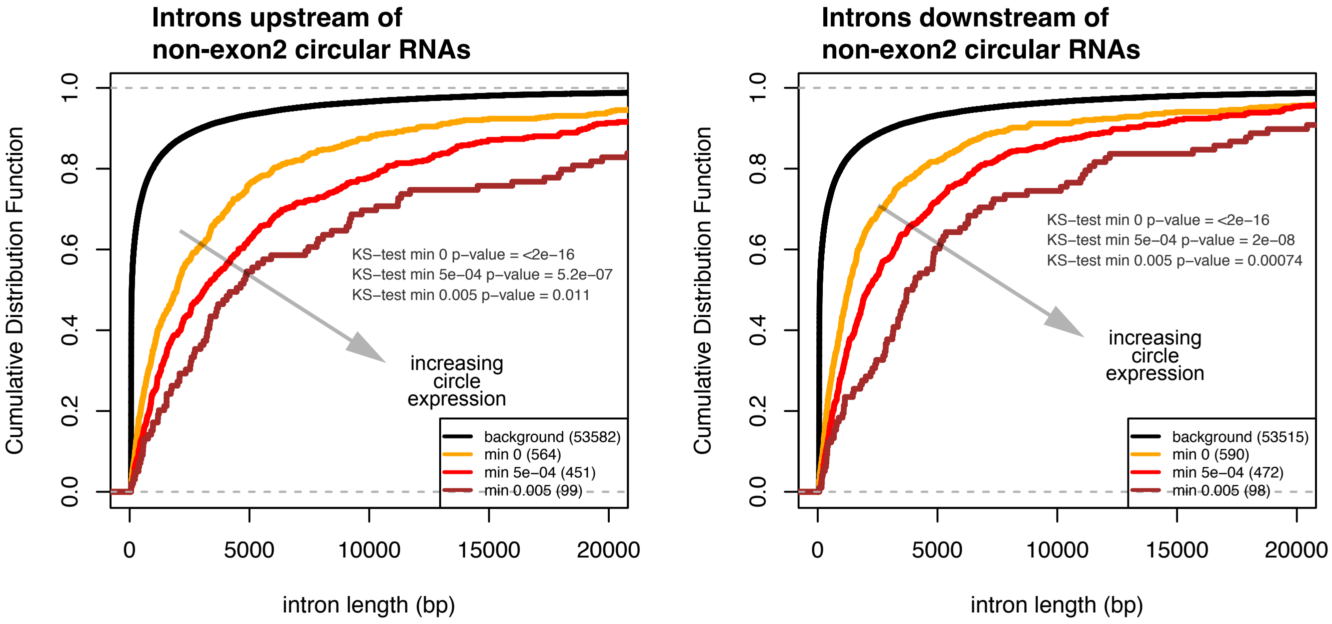
	exon2 circle**	non-exon2 circle**
long us intron*	604	1070
short us intron*	80	206

5' bias vs downstream intron length

	exon2 circle**	non-exon2 circle**
long ds intron*	599	1020
short ds intron*	85	256

* long intron are 500bp or longer

** exon2 circles are those circular RNAs where the circular junction is at the acceptor site of the first intron

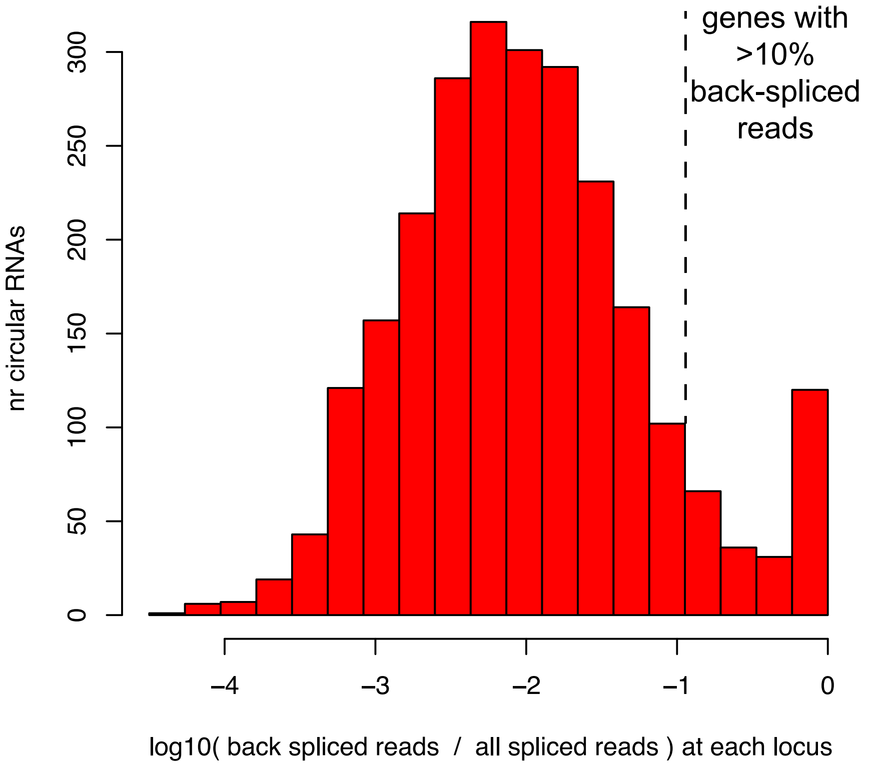


Introns upstream of non-exon2 circular RNAs

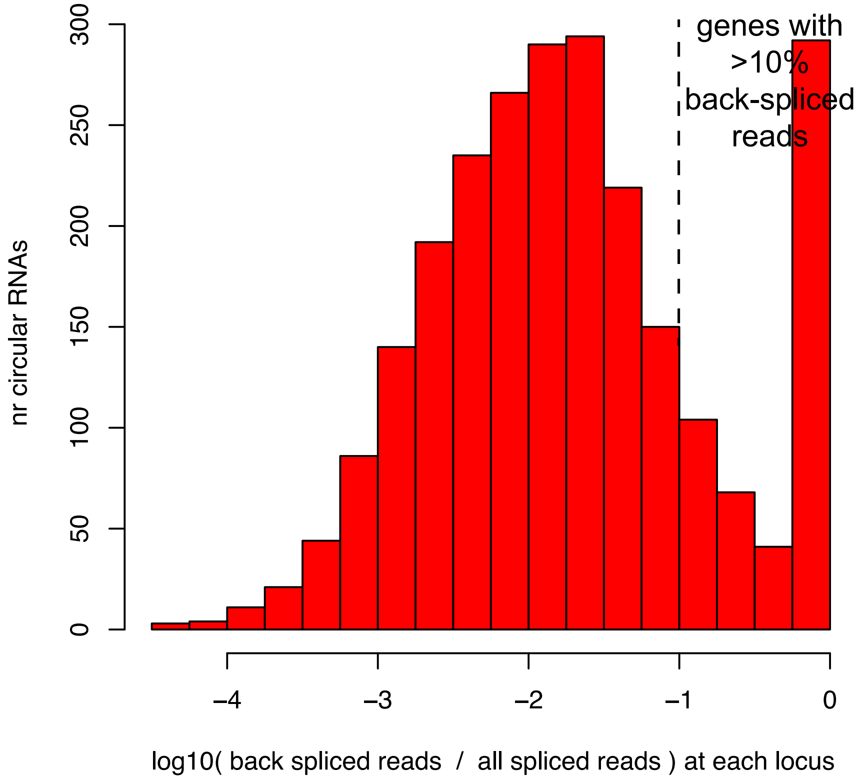
Introns downstream of non-exon2 circular RNAs

Westholm et al
Supplementary Figure 5

fraction back spliced reads out of all spliced reads : all data

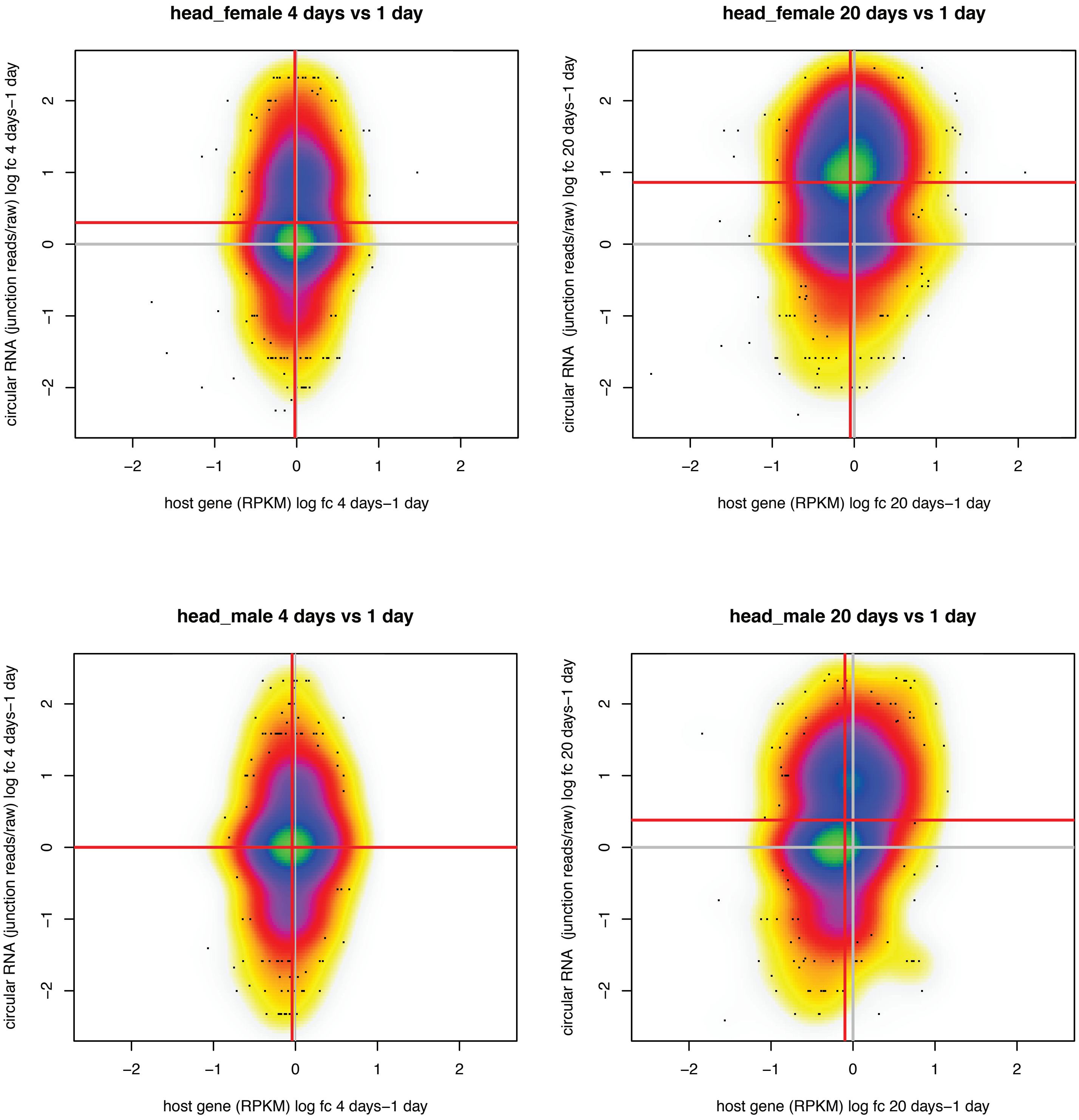


fraction back spliced reads out of all spliced reads: head data



Westholm et al
Supplementary Figure 6

Circular RNA expression in head time course data, segregated by male and female data.



Westholm et al
Supplementary Figure 7