Developmental Cell

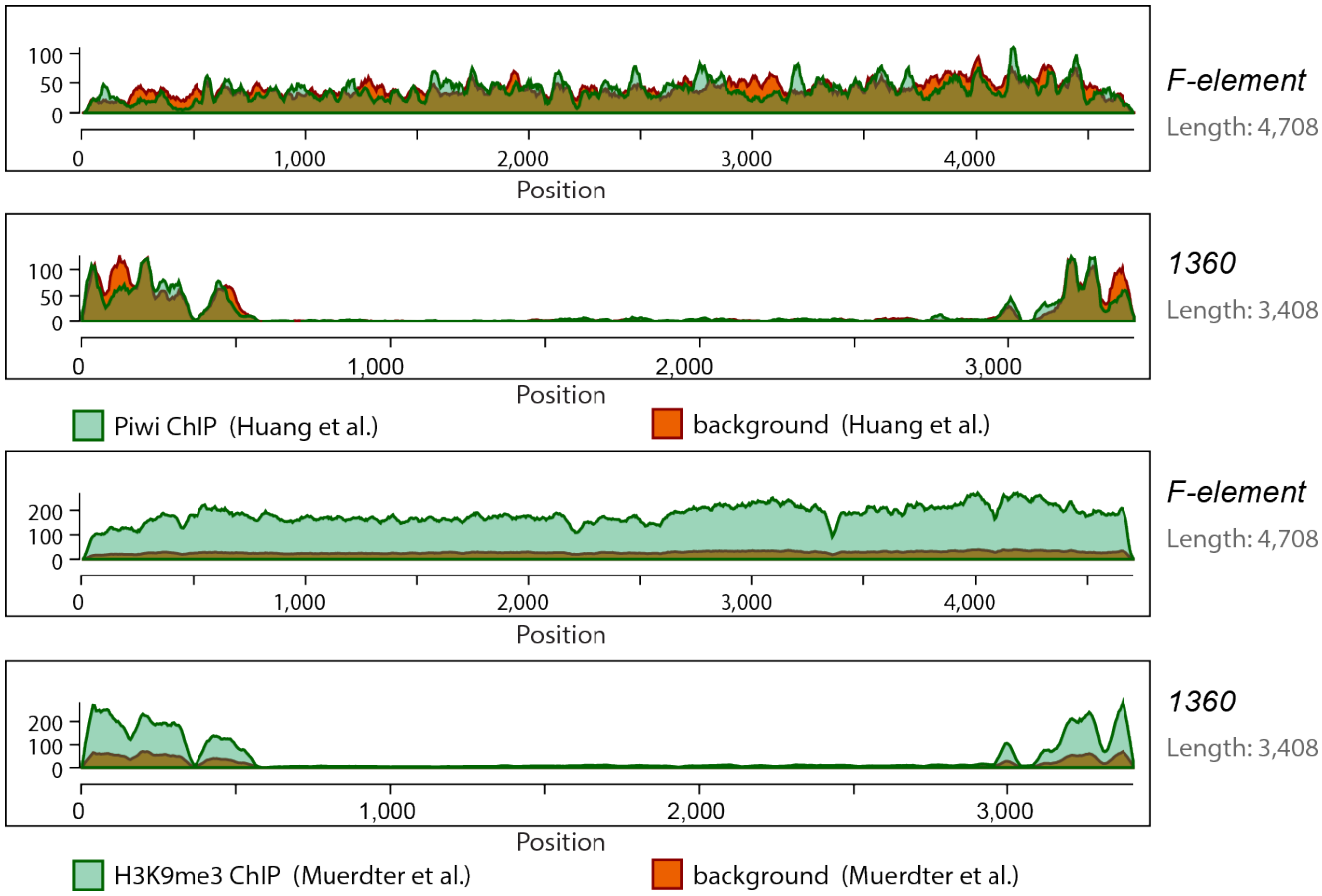
Supplemental Information

Pitfalls of Mapping High-Throughput Sequencing Data to Repetitive Sequences: Piwi's Genomic Targets Still Not Identified

Georgi K. Marinov, Jie Wang, Dominik Handler, Barbara J. Wold, Zhiping Weng, Gregory J. Hannon, Alexei A. Aravin, Phillip D. Zamore, Julius Brennecke, and Katalin Fejes Toth

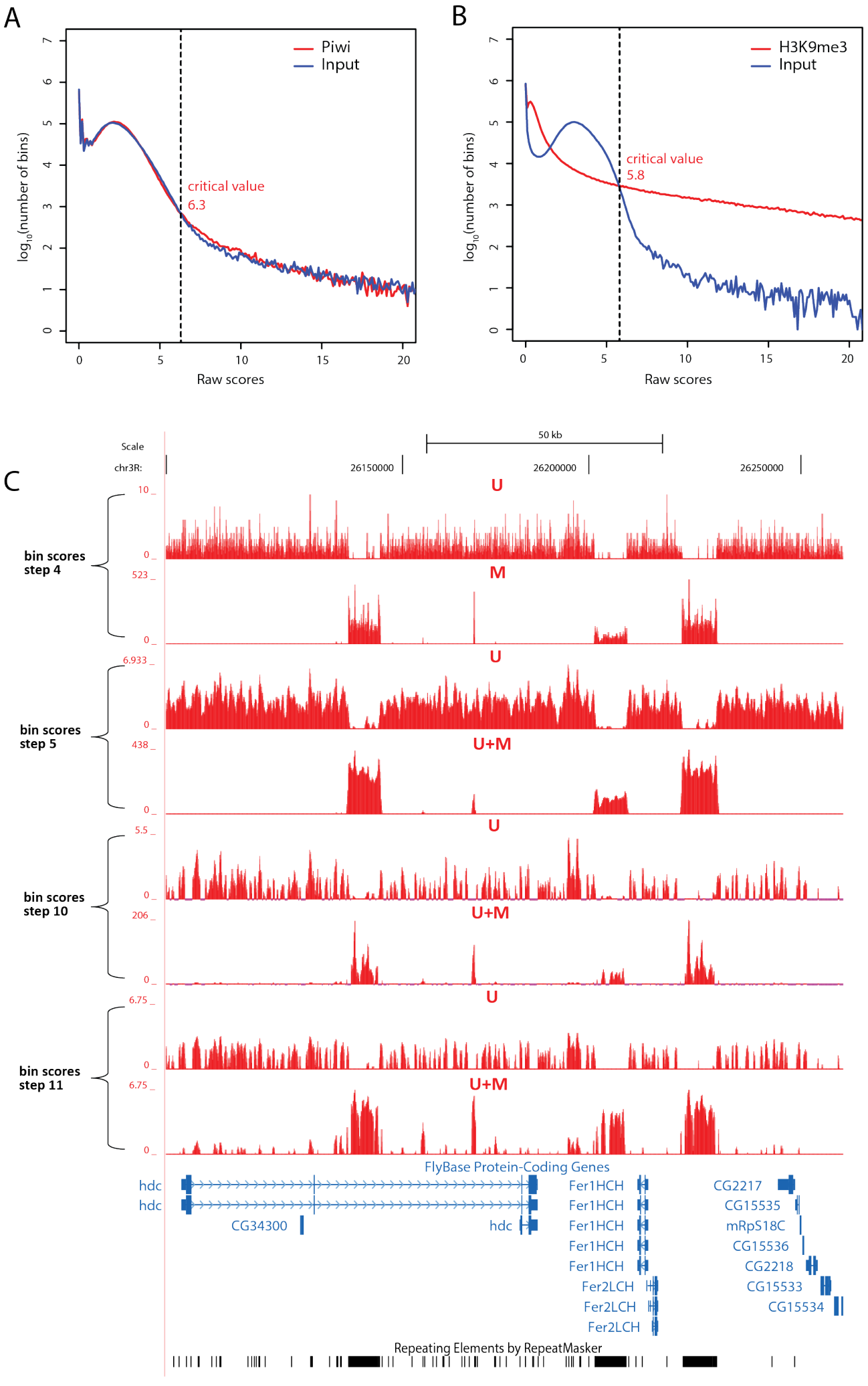
Pitfalls of mapping high throughput sequencing data to repetitive sequences: Piwi's genomic targets still not identified. Supplementary Materials

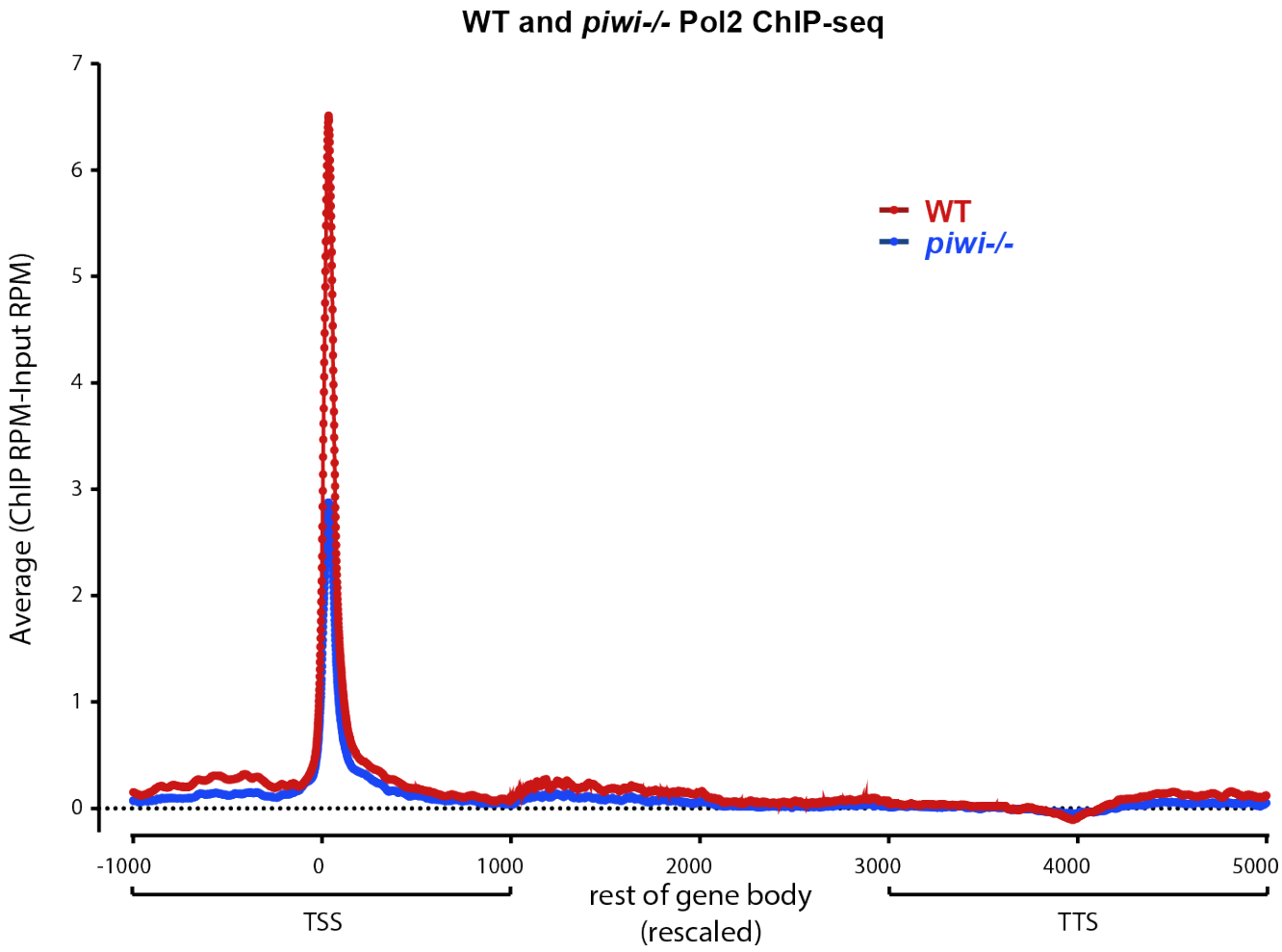
Supplementary Figures



Supplementary Figure 1: (Related to Figure 1) Absence of enrichment in the Piwi ChIP-seq dataset (from Huang et al. 2013) and high enrichment of H3K9me3 (from Muerdter et al. 2013) over consensus transposons. Shown is the coverage of Piwi and H3K9me3 ChIP and control datasets over the *F*-element and the *1360* transposon.

Supplementary Figure 2 *(following page)*: (Related to Figure 3) Reproduction and output of the Huang et al. pipeline. (A) and (B) Raw score distribution and "critical value" determination of ChIP-seq data and input data: (A) Piwi ChIP-seq and background (input) data from Huang et al. 2013; (B) H3K9me3 ChIP-seq and background data from Muerdter et al. 2013. The scores, which are smaller than the critical value, are considered as noises. (C) Piwi ChIP scores at different steps of the Huang et al. data processing pipeline. The tracks show the scores after each indicated step of the pipeline as described in the Supplementary Methods.


A
Piwi
Input
critical value
6.3
log10(number of bins)
Raw scores
B
H3K9me3
Input
critical value
5.8
log10(number of bins)
Raw scores
C
Scale
chr3R:
50 kb
26150000
26200000
26250000
bin scores step 4
bin scores step 5
bin scores step 10
bin scores step 11
U
M
U+M
FlyBase Protein-Coding Genes
hdc
CG34300
Fer1HCH
Fer2LCH
CG2217
CG15535
mRpS18C
CG15536
CG2218
CG15533
CG15534
Repeating Elements by RepeatMasker



Supplementary Figure 3: (Related to Figure 4) Metagene profiles of Pol II ChIP-seq coverage in WT and *piwi*-/- flies.

Supplementary Methods

Except for where specifically specified otherwise, all data processing was carried out using custom-written python scripts. The dm3/BDGP assembly, release 5 version of the *Drosophila melanogaster* genome was used.

ChIP-seq data processing

Sequencing reads (36bp in data from Huang et al. 2013; paired 75bp reads in data from Muerdter et al. 2013, also analyzed as single-end reads trimmed down to 36bp length, with essentially identical results; mixed read lengths trimmed down to 36bp in modENCODE data) were mapped against the genome using Bowtie 0.12.7 (Langmead et al., 2009) with the following settings: `''-v 2 -k 2 -m 1 --best --strata''` for unique 36bp alignments, `''-v 3 -k 2 -m 1 --best --strata''` for unique 2x75bp alignments, and `''-v 0 -a --best --strata''` for alignments in which multireads were retained. The `-X 1000` option was applied and only concordant read pairs were retained for 2x75bp H3K9me3 data.

Three different types of signal tracks were then generated.

1. Unique tracks retaining uniquely mapping reads only, normalized to RPMs (**R**eads **P**er **M**illion mapped reads) according to the following formula:

$$S_{c,i} = \frac{|R_{c,i}|}{\frac{|R|}{10^6}} \quad (1)$$

   Where $S_{c,i}$ is the signal score for position i on chromosome c, $|R|$ is the total number of mapped reads, and $|R_{c,i}|$ is the number of reads covering position i on chromosome c.

2. Tracks normalized for read multiplicity based on all alignable reads, where the normalization to RPMs is carried out as follows:

$$S_{c,i} = \frac{\sum_{R \in R_{c,i}} \frac{1}{NH_R}}{\frac{|R|}{10^6}} \quad (2)$$

   Where NH_R is the number of locations in the genome a read maps to.

3. Tracks generated using all alignments without normalization for multiplicity, i.e. treating each individual alignment A as if it is a uniquely mappable read:

$$S_{c,i} = \frac{|A_{c,i}|}{\frac{|A|}{10^6}} \quad (3)$$

Reproduction of the processing pipeline used by Huang et al. 2013

Huang et al. 2013 used the following pipeline to process their data (described in Yin et al. 2011):

1. Identical sequencing reads were merged into single sequences. Of note, this was done before alignment and apparently (contrary to established practices and submission guidelines for high-throughput sequencing data) the collapsed rather than the raw reads were submitted to the Short Read Archive.

2. The collapsed reads were mapped to the genome using SOAP. An extremely loose recursive alignment policy was applied, allowing for up to 5 mismatches and 4 indels for the 36bp reads used.

3. SOAP results were filtered by imposing the following requirement on alignments:

$$|A| \geq |MM| + |ID| + 23 \quad (4)$$

   where $|A|$ is the length of the alignment, $|MM|$ is the number of mismatches and $|ID|$ is the number of indels.

   The 5' end of each alignment was recorded. No normalization for mapping multiplicity was performed.

4. Chromosomes were split into 50bp bins and each read contributed to 10 bins according to a scoring matrix, which varies for different samples based on gel electrophoresis images that are not publicly available. We used the published matrix from Yin et al. 2011:

Bin index	Tag weight
0	1.000
1	1.000
2	0.988
3	0.958
4	0.916
5	0.844
6	0.747
7	0.628
8	0.502
9	0.387

5. Scores for each 50bp bin were normalized to the total number of alignments, i.e. analogous to Equation 3:

6. The same steps were carried out for both ChIP and input

7. Next, a "critical values" was calculated, "beyond which the corresponding bin numbers in an experimental dataset are always more than those in the control dataset, was determined for each experimental/control dataset pair" (Fig. S2)

8. A "normalizer" score was calculated as the mean score for the bins whose values are lower than the critical value in the whole genome: A normalizer was further determined for each experimental/control dataset pair in a way that the correlation coefficient between these two datasets for values lower than the critical value are maximized when the scores of the experimental dataset are multiplied by this normalizer."

9. The score of each ChIP sample was normalized by the ratio of the mean score background and ChIP in the bins that are lower than the critical value.

10. The ChIP score was further normalized by subtracting the background; however, negative values were given a score of 0:

$$S_{N_i} = max((S_{ChIP_i} - S_{input_i}), 0) \quad (5)$$

11. Final score (S_F) profiles were calculated as:

$$S_{F_i} = max\left(0, log_2\left(\frac{S_{N_i}}{TM(S_N)}\right)\right) \quad (6)$$

Where TM is the trimmed mean.

The bin scores at each step of the pipeline for the region presented in Fig. 1A are shown in Fig. S2.

Analysis of RepeatMasker-annotated repeat element coverage

The RepeatMasker repeat element annotation downloaded from UCSC (Kent et al. 2002) was used for all repeat analysis. An RPM score was calculated for each repeat using the following formula:

$$RPM_{RE} = \frac{\sum_{R \in RE} \frac{1}{NH_R}}{\frac{|R|}{10^6}} \quad (7)$$

Analysis of consensus-sequence repeat element coverage

Consensus repetitive elements for *Drosophila melanogaster* were downloaded from FlyBase (Marygold et al. 2013). Reads were trimmed down to 36bp as this was the read length of the Piwi ChIP-seq data from Huang et al. 2013. Reads were then aligned against the Flybase repetitive element consensus sequences using Bowtie 0.12.7 (Langmead et al., 2009) with the following settings: `''-v 3 -a --best --strata''`, i.e. allowing for up to 3 mismatches, and unlimited number of locations a read can map to. Read counts were calculated for each repetitive element and normalized to RPM against the total number of reads ($|R \in G|$) aligning to the whole genome (with unlimited number of locations a read can map to) as follows:

$$RPM_{RE_c} = \frac{|R \in RE_c|}{\frac{|R \in G|}{10^6}} \quad (8)$$

where RE_c refers to the consensus repetitive element. For the H3K9me3 dataset from Muerdter et al. 2013, 1x36bp reads were used in these analyses.

Supplementary References

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**(6):996-1006.

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**(3):R25.