Supplementary Information

PingPongPro: a tool for the detection of piRNA-mediated transposon silencing in small RNA-Seq data

Sebastian Uhrig, Holger Klein

Supplementary Methods

Naïve algorithm for the detection of ping-pong cycle activity

Naïve approches for the detection of ping-pong cycle activity are based on raw read counts. Regions which are suppressed through the ping-pong cycle are marked by a high number of reads which overlap with one or more reads on the opposite strand by ten nucleotides at the 5' ends (Fig. S1). If the number of reads with this characteristic overlap is significantly higher than can be explained by random chance, then a region is considered to be suppressed through the ping-pong cycle. Statistical significance is determined by means of a Z-test: The probability to observe a certain number of overlapping reads can be modeled with the help of a normal distribution. Reads which overlap by lengths other than ten nucleotides (e.g., 1-9 and 11-19 nucleotides) are used to estimate the mean and variance of the normal distribution. If the number of reads with an overlap of ten nucleotides exceeds the estimated mean by more than a user-defined cutoff, a region is thought to be ping-pong-controlled.

PingPongPro's algorithm

The detection algorithm of PingPongPro is based on the naïve algorithm: PingPongPro counts the number of reads with an overlap of ten nucleotides and compares it against the number of reads with overlaps of arbitrary lengths. Again, statistical significance is determined by means of a Z-test.

PingPongPro enhances the naïve algorithm by a preliminary step (Fig. S2), where reads are assigned a weight before being counted. PingPongPro first tries to segregate reads into one of two classes: those that overlap by ten nucleotides because they are ping-pong-derived ("ping-pong signatures") and those that overlap by ten nucleotides due to mere coincidence ("random signatures"). The weight directly reflects the probability that the signature is ping-pong-derived. Reads with low probability are assigned a weight close to 0 (and thus are not counted fully) and reads with high probability are assigned a weight close to 1. PingPongPro evaluates several covariates for the prediction. The correlation of these covariates with ping-pong-mediated signatures is determined dynamically from the given dataset. The covariates are based on intrinsic properties of ping-pong signatures:

- The probability that a signature is ping-pong-derived rises with increasing number of reads which constitute the signature. A few reads may overlap by ten nucleotides due to coincidence and thus form a random signature. With increasing number of reads, such an event becomes exponentially less likely: Mathematically, the probability that n reads located in a window of b base pairs overlap by a predefined length can be calculated as $b^{-(n-2)}$.
- The probability of observing random signatures rises with increasing coverage. In order to avoid false classification of signatures in regions with deep coverage, PingPongPro takes the local coverage into account. If the signature under consideration consists of more reads than the average coverage in the vicinity, PingPongPro deems the signature more likely to be ping-pong-derived.
- piRNAs frequently have uracil as the first base at the 5' end [7]. Due to base complementarity, ping-pong signatures often have adenine at position ten on one of the strands, hence. Fig. S6 C demonstrates this base bias in the sample SRR298567, where over 70 % of the reads have adenine at the tenth position. If a signature has guanine, cytosine or uracil as the tenth base, the chances are reduced that it is ping-pong-derived.

PingPongPro scans the entire genome for signatures and assigns each of them to one of 4,000 bins:

- Signatures are first segregated by their number of reads. PingPongPro finds the signature with the highest read count and divides the range between zero and the logarithm of this (highest) read count into 1,000 equally sized bins. (This number of bins is high enough to yield sufficiently fine-grained partitioning of signatures. A logarithmic scale is used to ensure that all bins contain enough signatures to estimate the distribution reliably. Particularly bins for high read counts would contain few signatures, if the boundaries of the bins were defined on a linear scale.) All signatures are then assigned to one of these bins according to their logarithmic read count.
- Next, each bin is subdivided into two bins - one for signatures with a read count higher than 0.2x the average coverage within a window of nine base pairs up- and downstream and one for signatures with a lower read count.
- Lastly, each of these bins is further subdivided into two bins - one for signatures with adenine at position ten and one for signatures with other bases.

In summary, every bin represents a triplet of features: absolute read count, read count relative to the local coverage, and adenine bias. The signatures assigned to a bin match the criteria of the respective bin. After separation, bins representing typical properties of ping-pong signatures are enriched with signatures with an overlap of ten nucleotides compared to other bins (Fig. S6). Vice versa, bins representing properties of random signatures mostly contain random signatures. The precise percentages are unknown, however.

The percentages can be estimated with the help of reads which overlap by arbitrary lengths ("arbitrary signatures"). For each overlap between 1-9 and 11-19 nucleotides, PingPongPro applies the same separation method described previously: Whenever a signature with the given overlap is found, the signature is sorted into the appropriate bin according to its properties. Since all arbitrary signatures are random

events, these bins contain nothing but random signatures. Eventually, PingPongPro has generated 18 x 4,000 bins containing only random signatures (with arbitrary overlaps) and 1 x 4,000 bins containing a mixture of random signatures and ping-pong signatures (with an overlap of ten nucleotides). The bins with arbitrary signatures are used to estimate the percentage of ping-pong signatures in the bins with a mixture of random and ping-pong signatures. The estimation procedure employs a normal distribution with mean and standard deviation determined from the bins with arbitrary signatures.

The estimated fraction of ping-pong signatures of a bin is the weight that PingPongPro assigns to reads belonging to signatures in this bin. Like so, signatures which have a high chance of being random are given a low weight, whereas signatures with a high chance of being ping-pong-driven are given a high weight. After weights have been assigned to all reads, PingPongPro applies the naïve algorithm to identify regions suppressed through the ping-pong cycle. Finally, the Z-scores are corrected for multiple testing using the FDR method after Benjamini and Hochberg [1].

Benchmarking of PingPongPro vs. the algorithm based on raw counts

In order to compare PingPongPro's performance against that of the algorithm based on raw counts, which is equivalent to running PingPongPro without weighting, we ran both methods on several small RNA-Seq samples from the testes/ovaries of various organisms (*C. elegans, D. rerio, D. melanogaster, H. sapiens, M. musculus*).

Adapters were removed with the help of cutadapt 1.16 [4] with default parameters. Reads were mapped to the ce11/danRer11/dm6/hg38/mm10 assembly of the respective reference genome using bowtie 2.2.1 [3] with sensitive local alignment settings and reporting of only the best alignments in case of multi-mapping reads. Table S1 lists the number of reads and mapping rates for each sample.

To benchmark the performance of PingPongPro, we screened each sample for reads which overlap with at least one read on the opposite strand by 10 nt. Such pairs of reads are candidates for ping-pong signatures. All candidate ping-pong signatures which were within a distance of 1 kbp to the next signature were merged into a region. These regions are candidates for regions being suppressed by the ping-pong cycle. Only regions containing at least two candidate ping-pong signatures were kept for further analysis. To each of the candidate regions, we applied PingPongPro's weighted algorithm and the unweighted algorithm to classify them as either ping-pong-suppressed or not. Since the samples chosen for the benchmark are expected to exhibit ping-pong cycle activity, the fraction of regions classified as ping-pong-suppressed can be regarded as a measure of the sensitivity. In an attempt to measure the specificity of both algorithms, we repeated this procedure with an overlap of 13 nt. Assuming there is no underlying biological mechanism which produces reads overlapping by 13 nt, all predictions made by the algorithms using the wrong overlap can be considered false positives. The fraction of regions classified as ping-pong-suppressed given the wrong overlap can be interpreted as a measure of the specificity, hence. Table S2 shows the total number of tested regions and the fraction of regions identified as ping-pong-suppressed at a false-discovery rate of 1 % using both algorithms and overlaps.

Sensitivity and specificity were further characterized with the help of receiver operating characteristic (ROC) curves (Fig. S5). Ideally, the ROC curves would be generated on a set of regions with known ping-pong status. However, this information is unavailable and can only be approximated via indirect measurements. We therefore used different overlaps (10 nt and 13 nt) to define a set of true and false positives for the calculation of the true and false positive rates of the ROC curves. As described above, all candidate regions given an overlap of 10 nt were considered true positives, whereas all candidate regions given an overlap of 13 nt were considered false positives. The ROC curves illustrate the sensitivity-specificity trade-off when using different p-value cutoffs to call regions as being suppressed through the ping-pong cycle. In view of the uncertainty of the real fraction of true positives, the curves cannot be interpreted in absolute terms. Still, they reflect the improved sensitivity-specificity trade-off of PingPongPro.

Supplementary Tables

organism	SRA accession	reads	uniquely mapped	multi-mappers	unmapped	PingPongPro runtime (seconds)	PingPongPro memory usage (MB)
C. elegans	SRR087428	8292578	71.31%	8.08%	20.61%	51	21
C. elegans	SRR1166535	24711508	69.82%	9.75%	20.43%	232	21
C. elegans	SRR1820954	19129826	61.83%	13.42%	24.75%	113	34
C. elegans	SRR513311	12187012	63.08%	11.45%	25.47%	81	35
D. rerio	SRR298567	27659974	15.18%	53.72%	31.10%	198	279
D. rerio	SRR298568	27114461	4.09%	36.57%	59.34%	96	79
D. rerio	SRR363985	8604155	18.95%	72.24%	8.81%	75	110
D. rerio	SRR578904	23276412	12.37%	78.83%	8.81%	232	293
D. rerio	SRR578913	26882577	11.89%	76.67%	11.44%	237	252
D. rerio	SRR578922	25233936	19.54%	75.20%	5.26%	249	277
D. melanogaster	SRR010960	3302340	11.95%	68.91%	19.15%	24	46
D. melanogaster	SRR1187947	36313264	21.32%	59.45%	19.24%	396	693
D. melanogaster	SRR1435859	22721233	17.75%	46.47%	35.78%	231	54
D. melanogaster	SRR916073	2268550	13.47%	62.25%	24.28%	10	148
H. sapiens	ERR328151	15118197	40.09%	58.35%	1.56%	97	141
H. sapiens	SRR835324	22907658	31.92%	50.15%	17.93%	173	219
H. sapiens	SRR835325	25296024	33.66%	51.59%	14.74%	200	224
H. sapiens	SRR950451	6845544	21.66%	35.88%	42.46%	40	59
M. musculus	SRR1146664	7292666	40.14%	25.55%	34.30%	60	35
M. musculus	SRR1509747	61726722	9.38%	81.84%	8.78%	464	278
M. musculus	SRR1769730	7627862	57.53%	13.63%	28.84%	50	27
M. musculus	SRR636661	21719765	39.94%	11.01%	49.05%	201	43

Table S1. Mapping statistics of bowtie2 and runtimes of PingPongPro on an Intel Xeon E7-8837 CPU at 2.67GHz.

organism	SRA accession	regions using 10 nt overlap (positive/tested)		regions using wrong overlap (positive/tested)	
		PingPongPro	raw counts	PingPongPro	raw counts
C. elegans	SRR087428	18/39 (46%)	13/39 (33%)	2/34 (6%)	7/34 (21%)
C. elegans	SRR1166535	57/182 (31%)	58/182 (32%)	5/160 (3%)	31/160 (19%)
C. elegans	SRR1820954	69/138 (50%)	64/138 (46%)	0/131 (0%)	34/131 (26%)
C. elegans	SRR513311	28/99 (28%)	30/99 (30%)	2/97 (2%)	29/97 (30%)
D. rerio	SRR298567	14485/14746 (98%)	11474/14746 (78%)	143/8257 (2%)	1308/8257 (16%)
D. rerio	SRR298568	2154/2223 (97%)	1807/2223 (81%)	75/465 (16%)	138/465 (30%)
D. rerio	SRR363985	4061/4296 (95%)	3108/4296 (72%)	195/1494 (13%)	387/1494 (26%)
D. rerio	SRR578904	14305/14850 (96%)	11953/14850 (80%)	528/4989 (11%)	1005/4989 (20%)
D. rerio	SRR578913	13976/14455 (97%)	12070/14455 (84%)	540/4414 (12%)	886/4414 (20%)
D. rerio	SRR578922	13616/14139 (96%)	10284/14139 (73%)	280/7645 (4%)	1092/7645 (14%)
D. melanogaster	SRR010960	1994/2012 (99%)	1840/2012 (91%)	69/663 (10%)	200/663 (30%)
D. melanogaster	SRR1187947	3885/4170 (93%)	2797/4170 (67%)	8/4205 (0%)	371/4205 (9%)
D. melanogaster	SRR1435859	965/1022 (94%)	695/1022 (68%)	3/725 (0%)	234/725 (32%)
D. melanogaster	SRR916073	0/40 (0%)	17/40 (43%)	0/39 (0%)	23/39 (59%)
H. sapiens	ERR328151	1048/1106 (95%)	676/1106 (61%)	61/635 (10%)	186/635 (29%)
H. sapiens	SRR835324	2051/2081 (99%)	1566/2081 (75%)	15/1214 (1%)	359/1214 (30%)
H. sapiens	SRR835325	2449/2494 (98%)	1892/2494 (76%)	13/1475 (1%)	440/1475 (30%)
H. sapiens	SRR950451	359/365 (98%)	282/365 (77%)	10/163 (6%)	57/163 (35%)
M. musculus	SRR1146664	154/156 (99%)	138/156 (88%)	11/27 (41%)	17/27 (63%)
M. musculus	SRR1509747	38388/39529 (97%)	37411/39529 (95%)	490/1295 (38%)	557/1295 (43%)
M. musculus	SRR1769730	12/16 (75%)	11/16 (69%)	0/3 (0%)	0/3 (0%)
M. musculus	SRR636661	45/49 (92%)	36/49 (73%)	1/11 (9%)	2/11 (18%)

Table S2. Sensitivity/specificity benchmark of PingPongPro's algorithm versus the algorithm based on raw counts only.

To assess PingPongPro's accuracy, we ran the program on small RNA-Seq datasets from testes/ovaries of various model organisms. The datasets were obtained from the NCBI Sequence Read Archive via the given accession numbers. In each case, we compared the results produced by PingPongPro's multi-factor model (columns headed "PingPongPro") against those produced by the unweighted method, which only relies on raw read counts (columns headed "raw counts"). We instructed the algorithms to test all 1 kb regions which contain at least two putative ping-pong signatures.
Sensitivity was measured as the percentage of regions that are identified as being suppressed through the ping-pong cycle at an FDR < 0.01 (column "regions using 10 nt overlap (positive/tested)"). Not all of the regions are expected to show signs of ping-pong cycle activity, because they were defined solely on the basis of whether they contain overlapping reads – which may just be random events. But in general the percentage should be high, because the datasets were chosen with regard to high expected ping-pong cycle activity.
Specificity was measured as the percentage of regions identified as suppressed, when the algorithms were run with an incorrect value for the parameter which defines the overlap length (column "regions using wrong overlap (positive/tested)"). Instead of the correct value of 10 nt, we instructed the algorithms to assume a ping-pong overlap of 13 nt. Ideally, the number of identified regions should then be zero (perfect specificity).

Supplementary Figures

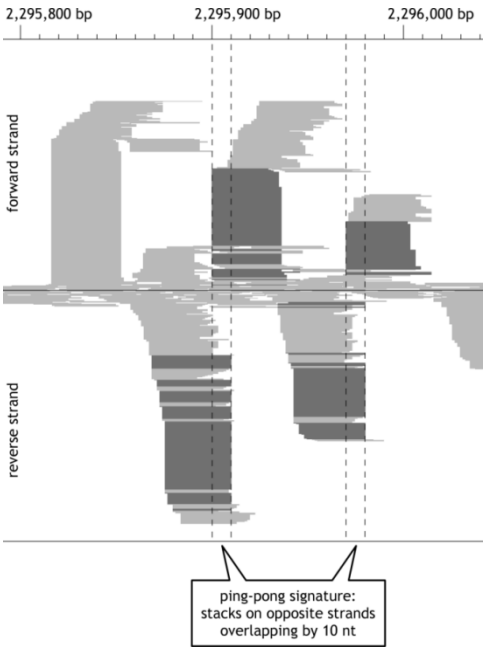


Figure S1. Coverage profile of small RNA-Seq data from *D. melanogaster* (available at the NCBI Sequence Read Archive via the accession number SRR499784 [5]) in the transposon *copia* showing two examples of ping-pong signatures.

The ping-pong cycle produces short RNA molecules which manifest in small RNA-Seq data as reads which share the same 5' position ("stacks" of reads) and which overlap with a stack of reads on the opposite strand by 10 nt. A pair of stacks on the forward and reverse strands make up a "ping-pong signature". The figure shows three notable stacks – three on the forward strand and two on the reverse strand. The stacks highlighted in dark shades are likely a product of the ping-pong cycle, because they overlap with a stack on the opposite strand by 10 nt (indicated by dashed lines). In contrast, the stack with the lowest coordinate is not accompanied by a stack on the reverse strand and presumably does not result from ping-pong cycle activity, hence.

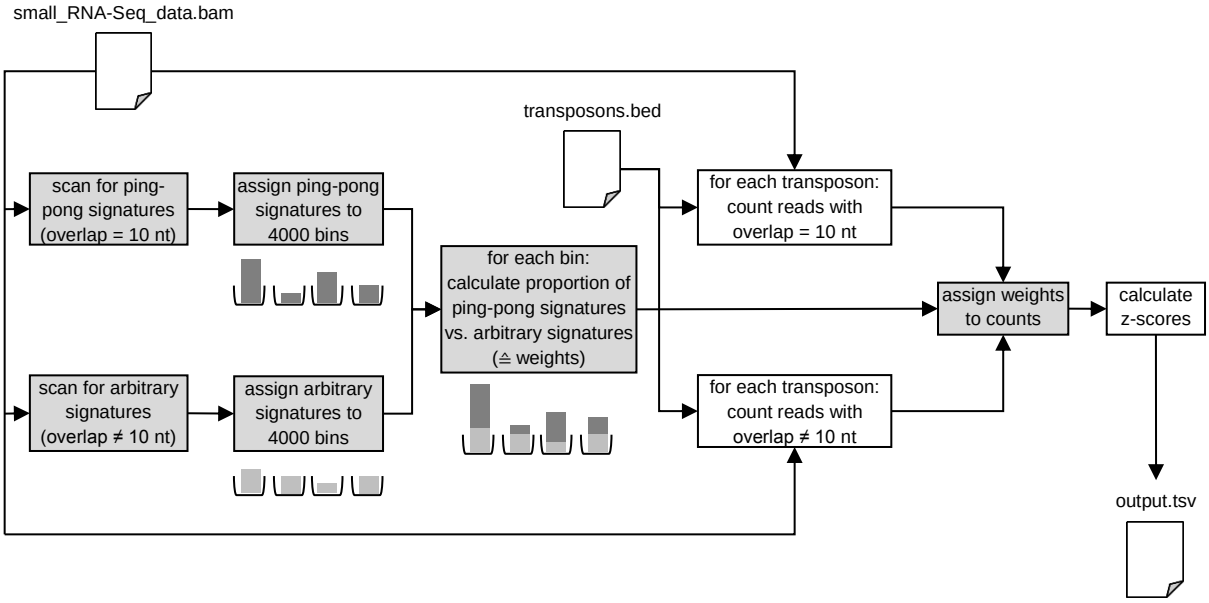


Figure S2. Flowchart of PingPongPro's algorithm.

PingPongPro takes two input files: small RNA-Seq data in SAM/BAM format and a list of coordinates of transposons to be checked for ping-pong cycle activity in BED/TSV/CSV/GFF/GTF format. PingPongPro adds a number of steps (grey boxes) to the naïve algorithm, which only considers raw counts (white boxes). These additional steps are necessary to calculate weights and assign them to read counts.

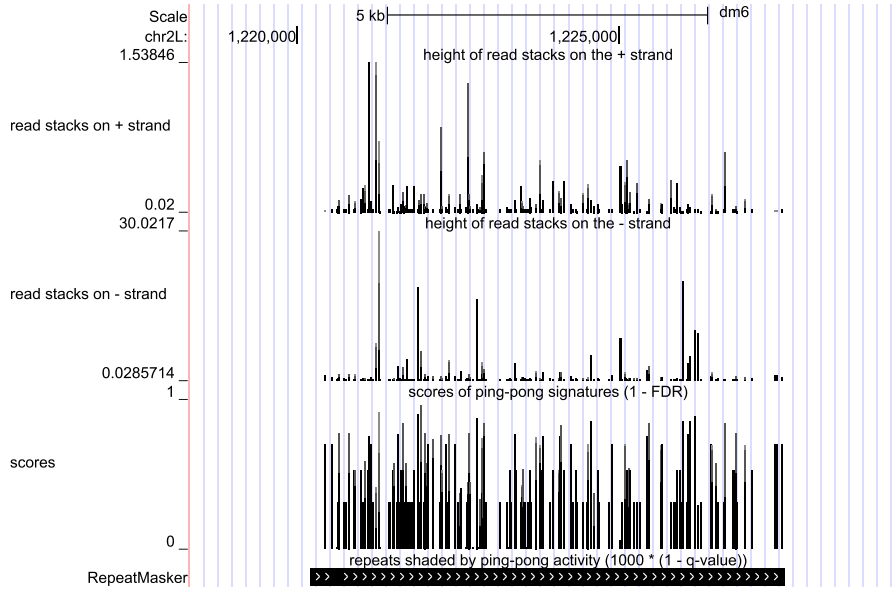


Figure S3. Visualization of ping-pong signatures in the UCSC Genome Browser [2].

PingPongPro is able to generate output in formats suitable for visualization in genome browsers like the UCSC Genome Browser and the Integrative Genomics Viewer [6]. This figure is a screenshot of the UCSC Genome Browser. It is based on small RNA-Seq data from *D. melanogaster* (SRA accession number SRR010960). It shows ping-pong signatures within the transposable element *BLOOD_I-int* of the *Gypsy* family with the genomic coordinate chr2L:1220184-1227592. The topmost row ("read stacks on + strand") shows the heights of stacks on the forward strand; the second row ("read stacks on - strand") shows the heights of stacks on the reverse strand. Most of the stacks have fractional heights, because they are mostly made up of multi-mapping reads. The third row ("scores") shows the empirical probabilities that PongPongPro calculated for the ping-pong signatures. The fourth row ("RepeatMasker") shows the location of repetitive elements as identified by RepeatMasker. The elements are shaded according to their Z-score: Regions with high Z-scores are darker and regions with low Z-scores are brighter. The shown region is covered with ping-pong signatures from start to end, which is an indication that it is a transposable element which is thoroughly suppressed through the ping-pong cycle.

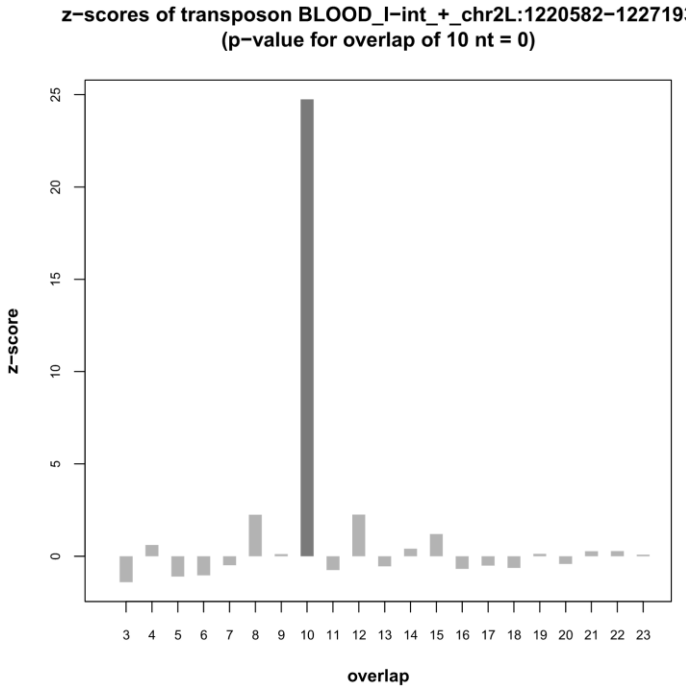
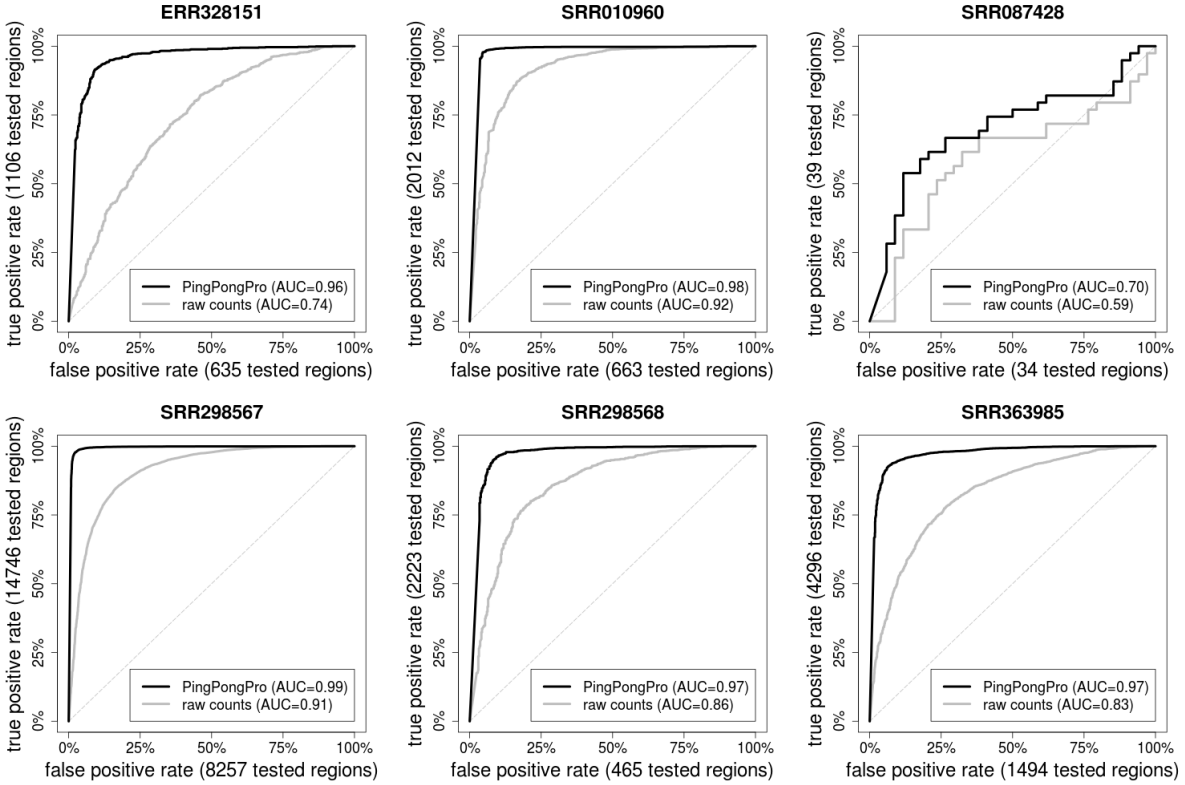
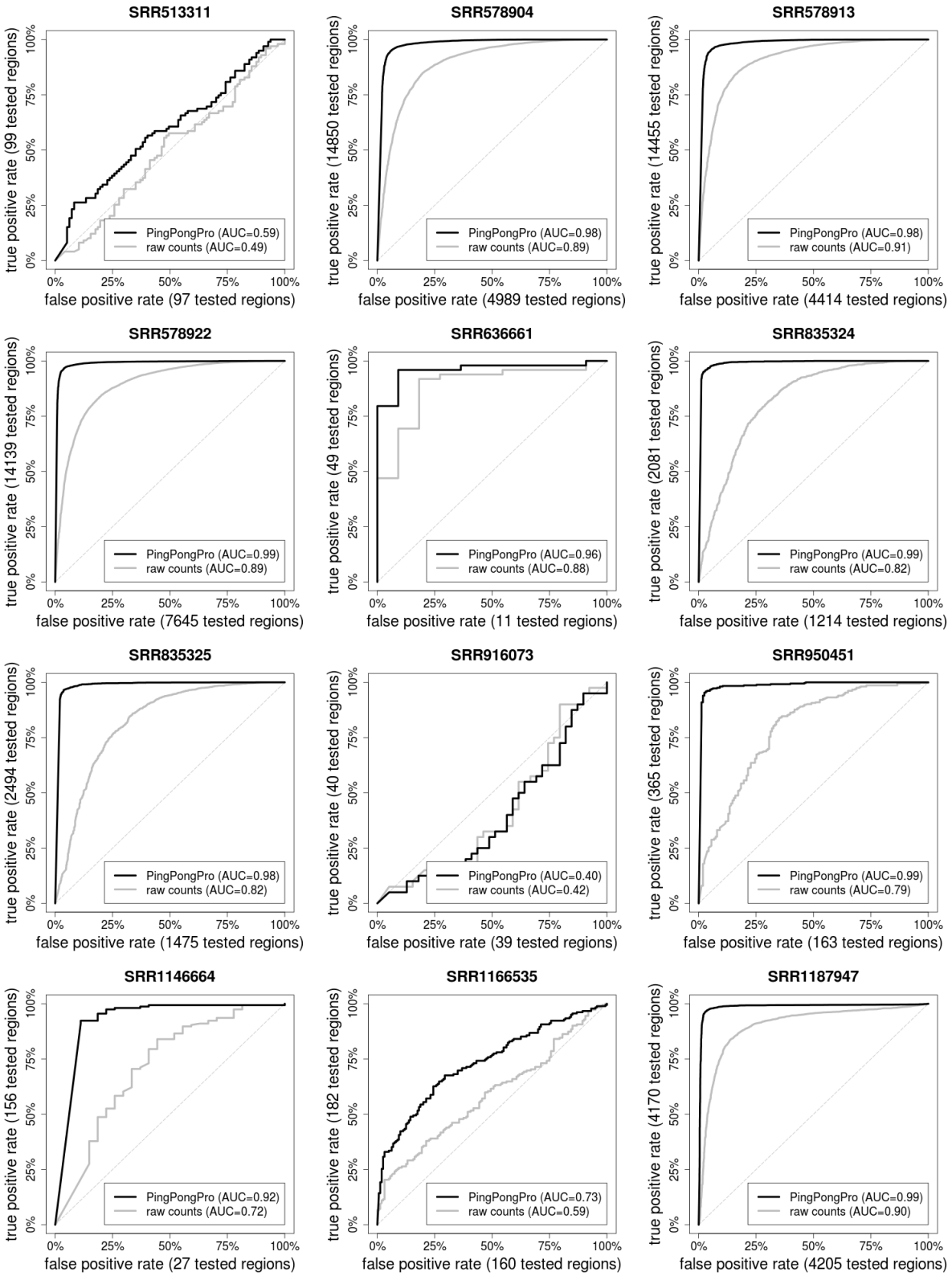


Figure S4. Graphical representation of the Z-test of the transposon shown in Suppl. Figure S1.

PingPongPro can visualize the Z-scores of regions inspected for ping-pong cycle activity in the form of a bar plot. Every bar represents the Z-score of the weighted read counts for the overlap denoted on the horizontal axis. The Z-score for the ping-pong-characteristic overlap of 10 nt is highlighted in dark grey. The (weighted) number of reads with an overlap of 10 nt exceeds the (weighted) number of reads with other overlaps with high statistical significance.




SRR513311
true positive rate (99 tested regions)
false positive rate (97 tested regions)
PingPongPro (AUC=0.59)
raw counts (AUC=0.49)
SRR578904
true positive rate (14850 tested regions)
false positive rate (4989 tested regions)
PingPongPro (AUC=0.98)
raw counts (AUC=0.89)
SRR578913
true positive rate (14455 tested regions)
false positive rate (4414 tested regions)
PingPongPro (AUC=0.98)
raw counts (AUC=0.91)
SRR578922
true positive rate (14139 tested regions)
false positive rate (7645 tested regions)
PingPongPro (AUC=0.99)
raw counts (AUC=0.89)
SRR636661
true positive rate (49 tested regions)
false positive rate (11 tested regions)
PingPongPro (AUC=0.96)
raw counts (AUC=0.88)
SRR835324
true positive rate (2081 tested regions)
false positive rate (1214 tested regions)
PingPongPro (AUC=0.99)
raw counts (AUC=0.82)
SRR835325
true positive rate (2494 tested regions)
false positive rate (1475 tested regions)
PingPongPro (AUC=0.98)
raw counts (AUC=0.82)
SRR916073
true positive rate (40 tested regions)
false positive rate (39 tested regions)
PingPongPro (AUC=0.40)
raw counts (AUC=0.42)
SRR950451
true positive rate (365 tested regions)
false positive rate (163 tested regions)
PingPongPro (AUC=0.99)
raw counts (AUC=0.79)
SRR1146664
true positive rate (156 tested regions)
false positive rate (27 tested regions)
PingPongPro (AUC=0.92)
raw counts (AUC=0.72)
SRR1166535
true positive rate (182 tested regions)
false positive rate (160 tested regions)
PingPongPro (AUC=0.73)
raw counts (AUC=0.59)
SRR1187947
true positive rate (4170 tested regions)
false positive rate (4205 tested regions)
PingPongPro (AUC=0.99)
raw counts (AUC=0.90)

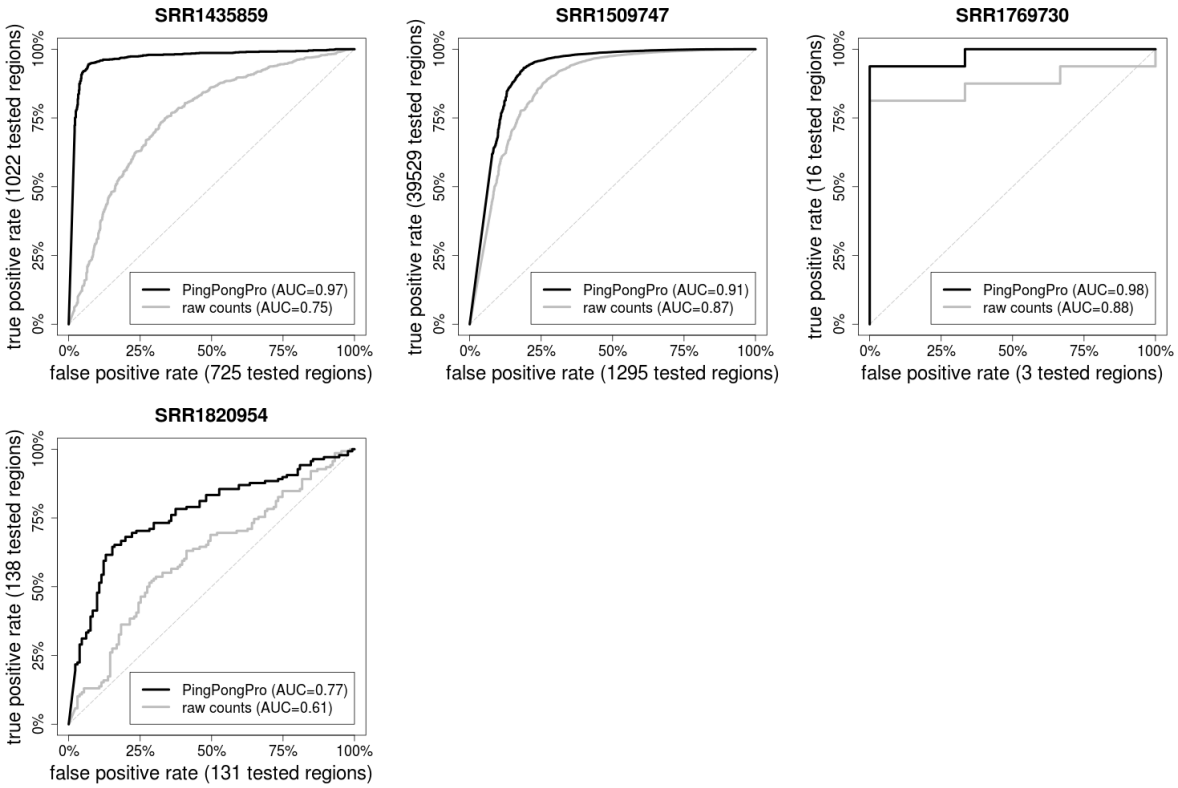


Figure S5. ROC curves of PingPongPro's algorithm versus the algorithm based solely on raw counts.

Sensitivity (true positive rate) and specificity (false positive rate) were measured as described in the supplementary methods section. The plots were generated on all datasets of Suppl. Table S2. The dashed line indicates the accuracy of an algorithm that makes random decisions on whether a region is suppressed through the ping-pong cycle or not. The plots demonstrate that PingPongPro consistently excels the method not using weights.

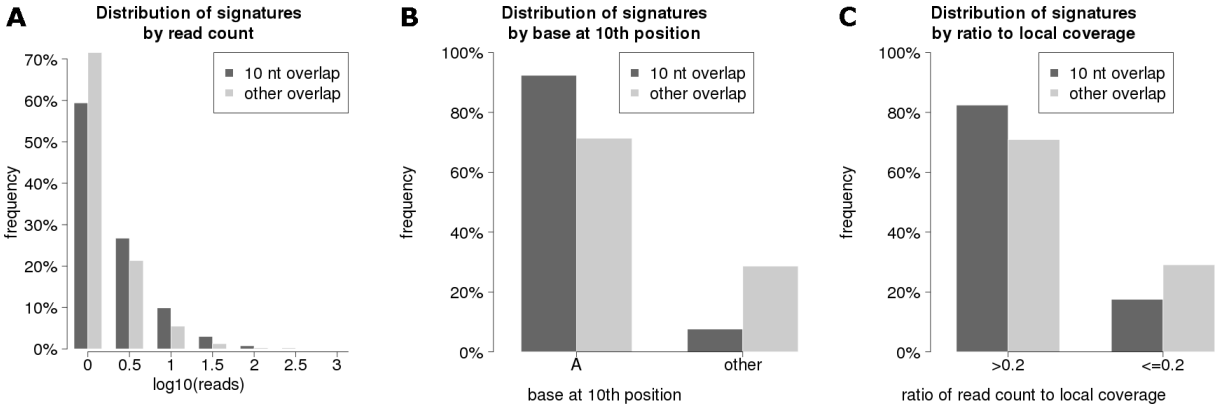


Figure S6. Distribution of ping-pong and arbitrary signatures with respect to the covariates of PingPongPro's weighted algorithm.

The correlation of PingPongPro's covariates and ping-pong signatures is reflected in shifted distributions compared to arbitrary signatures. PingPongPro exploits these shifts to assign different weights to reads. The plots were generated based on the sample SRR298567. A: Ping-pong signatures tend to have more supporting reads than arbitrary signatures. B: Adenine is enriched at the tenth position of reads of ping-pong signatures compared to arbitrary signatures. C: The number of reads of ping-pong signatures exceeds 0.2x the mean local coverage in 82 % of the signatures (versus 71 % for arbitrary signatures).

Supplementary References

[1] Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful. Approach to Multiple Testing. *Journal of the Royal Statistical Society* 1995, Series B (Methodological), 57(1):289-300.

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

[3] Langmead B, Salzberg S: Fast gapped-read alignment with Bowtie2. *Nature Methods* 2012, 9:357-359.

[4] Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011, 17(1):10-12.

[5] Preall JB, Czech B, Guzzardo PM, Muerdter F, Hannon GJ: shutdown is a component of the Drosophila piRNA biogenesis machinery. RNA 2012, 18(8):1446-57.

[6] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative Genomics Viewer. *Nature Biotechnology* 2011, 29, 24–26.

[7] Siomi MC, Sato K, Pezic D, Aravin AA: PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews Molecular Cell Biology* 2011, 12(4):246-58.