Supplementary Texts and Figures

Long-read direct RNA sequencing by 5'-Cap capturing reveals the impact of *Piwi* on the widespread exonization of transposable elements in locusts

Feng Jiang[1, 2, *], Jie Zhang[1, *], Qing Liu[3], Xiang Liu[2], Huimin Wang[1], Jing He[2], Le Kang[1, 2, 4, #]

1 Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China

2 State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

3 Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China.

4 College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China.

*These authors contributed equally to this study.

Corresponding authors:

Le Kang, Ph.D. and Professor
Institute of Zoology, Chinese Academy of Sciences
Beijing 100101, China
Tel: 86-10-6480-7219
Fax: 86-10-6480-7099
E-mail: lkang@ioz.ac.cn

Supplementary Texts

Direct RNA sequencing of locust transcripts

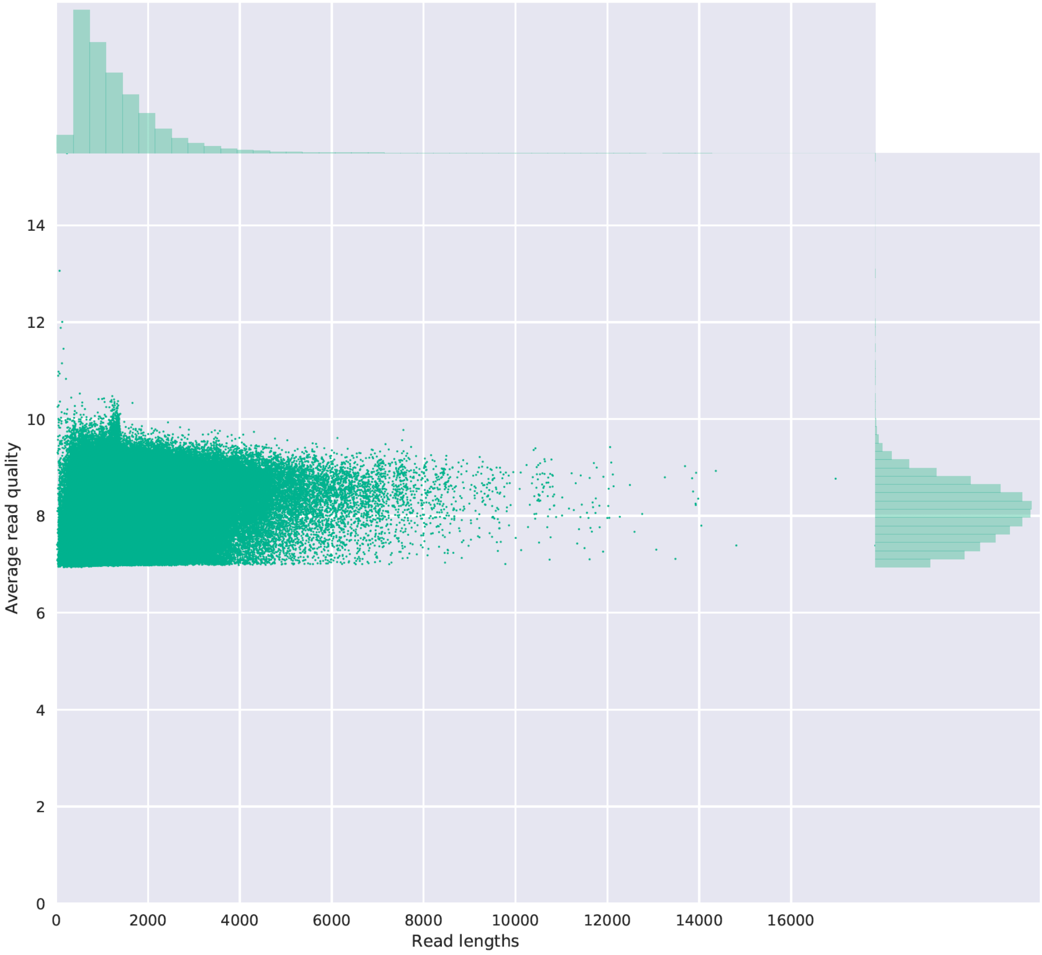
To assess the performance of direct RNA sequencing in the locusts, we sequenced a low-input RNA library (100 ng RNA, polyAControl-1), a standard-input RNA library (500 ng RNA, polyAControl-3) and a high-input RNA library (700 ng RNA, polyAControl-2) from poly-A enriched RNAs on a GridION X5 system using R9.4 flowcells. We used Albacore to perform base calling, and only the high-quality reads with a min_qscore_1d cutoff of 7.0 were retained for further analysis. After quality filtering, we obtained 119,574 reads with an N_{50} of 1,588 bases for the low-input library, 188,859 reads with an N_{50} of 1,850 bases for the standard-input library and 555,431 reads with an N_{50} of 1,632 bases for the high-input library. These results suggested that increasing the RNA input would markedly improve the sequencing data yield. Therefore, we used 700 ng of RNA input for further direct RNA sequencing experiments. The read length in the three libraries varied over a wide range from 0.01 Kb to 26.03 Kb, and the mean length of reads was 1.32 Kb with a standard deviation of 0.98 Kb (Supplementary Figure 1). The length distribution of sequencing reads has similar shapes. Despite their differences in data amount production, a consistent transcriptome signature was observed in the three libraries (Supplementary Figure 2A). Moreover, compared with the coding region obtained from the locust official gene set, the direct RNA sequencing reads showed a shift toward a longer length, implying a considerable portion of untranslated regions in the direct RNA sequencing reads. The sequencing errors in the direct RNA sequencing reads were corrected using Illumina reads with LoRDEC [1], and the error-corrected reads were aligned to the locust genome using GMAP [2]. As a representative example, a smoothscatter plot of sequence identity and read length for the high-input RNA library is shown in Figure 2B. The percentage values of direct RNA sequencing reads aligned to the locust genome were 99.44% (118,912/119,574), 99.55% (188,013/188,859) and 99.60% (553,200/555,431). Although the massive intron size expansion results in large transcriptional units in the locusts [3], the complete gene structure and isoforms could be inferred by comparing the genomic sequence with their corresponding direct RNA sequencing reads (Supplementary Figure 2C). The error rates of the aligned direct RNA sequencing reads were evaluated based on the locust genome. The sequence identity showed a similar distribution in the three libraries (Supplementary Figure 2D). Averaged across the three libraries, the sequence identity of the aligned reads is 92.68%; thus, the average error rate for the error-corrected reads is ~8%. The number of detected protein-coding genes increases as the direct RNA sequencing reads increase (Supplementary Figure 2E). Due to the high sequencing depth, 13,634 (representing 77.52% of the protein-coding genes in the official gene set) protein-coding genes could be detected in the Illumina datasets. Among these 13,634 protein-coding genes, 87.82% (11,974/13,634) of them could be detected in the combined datasets (number of sequencing reads: 863,864) of the three libraries. These data suggested that direct RNA sequencing demonstrates an inherent capacity to accurately sequence RNA transcripts for the locust transcriptome.

RNA adaptor sequences

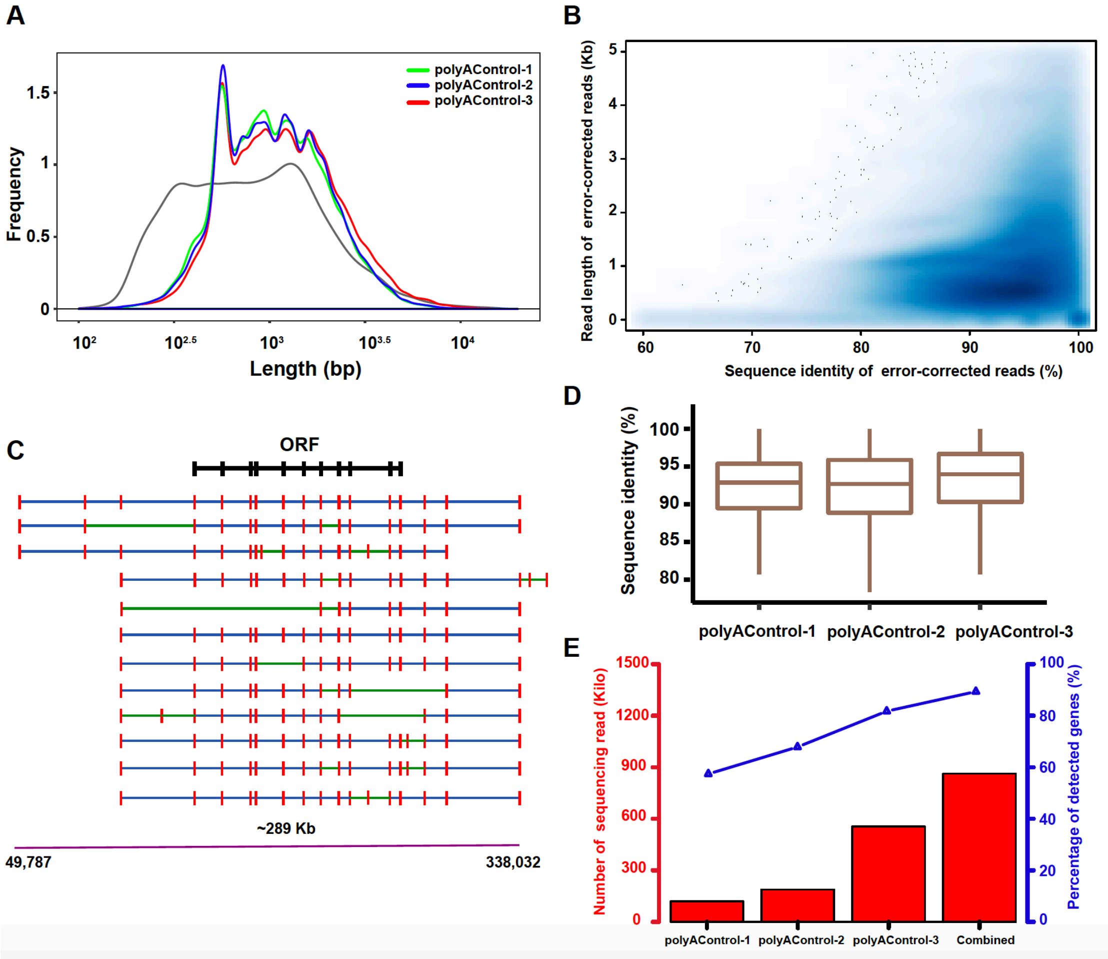
Short RNA adaptor (16 bp):
AGGCACGGGCTATGAG

Long RNA adaptor (50 bp):
ATTGCCAGTGGTGTGTGTCATAAATAGCGCGCAGTTTATCAAAGCAGGAC

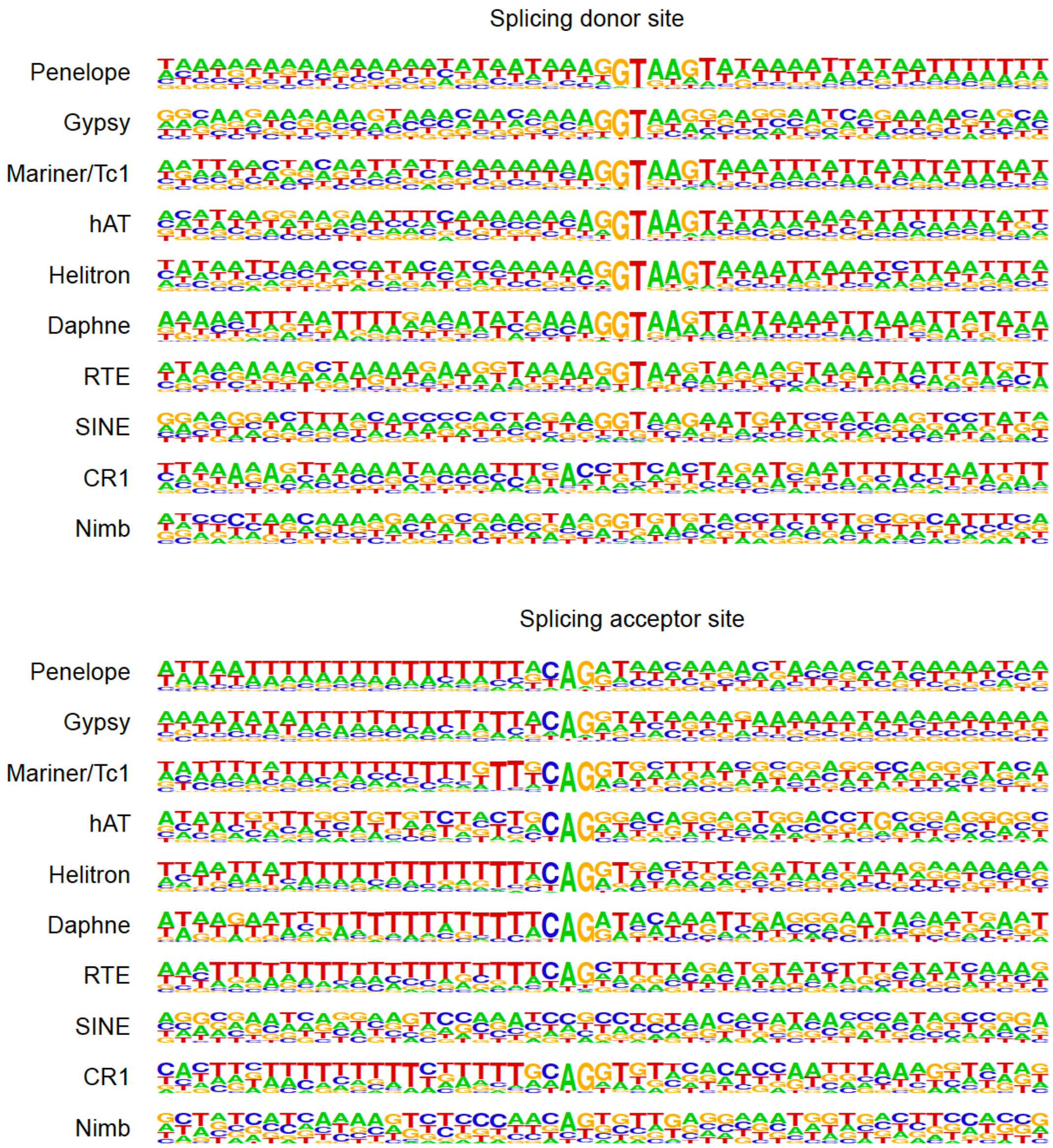
Supplementary Figures



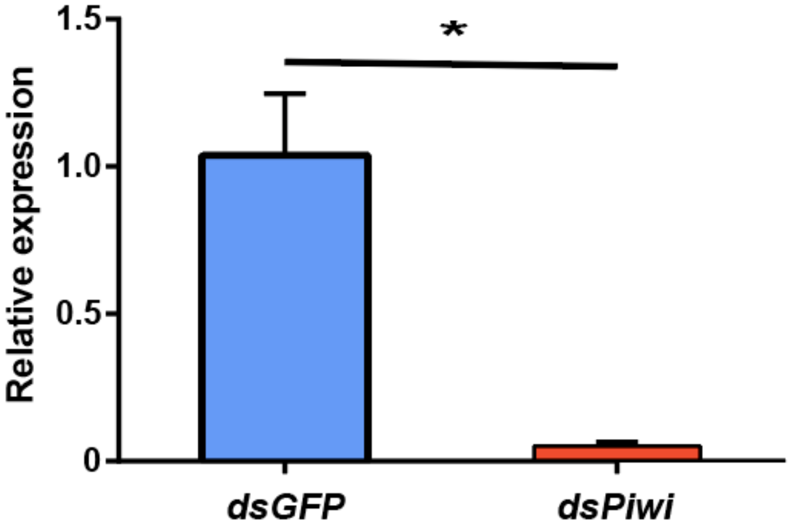
Supplementary Figure 1. Distribution of the length and average basecall quality score of the pass reads with a min_qscore_1d cutoff of 7.0. This graph was generated by NanoPlot version 1.0.0.



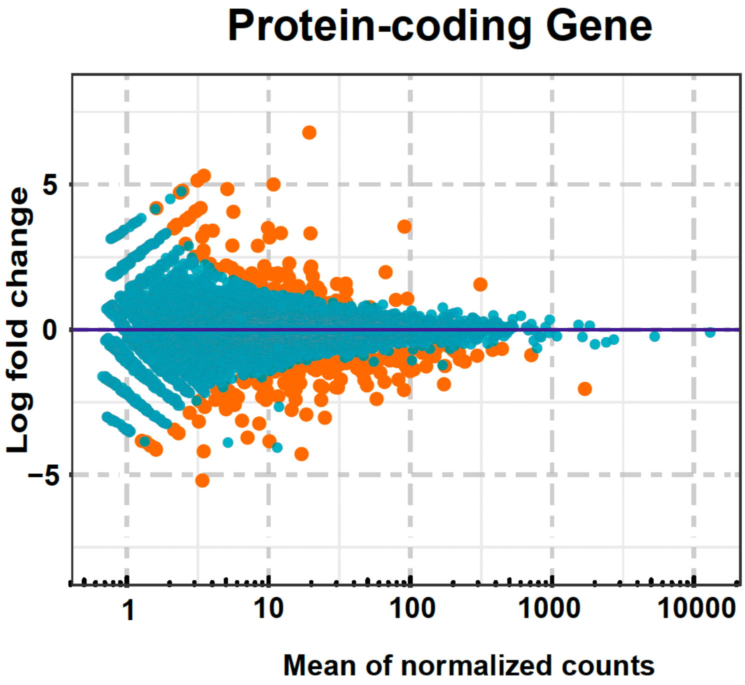
Supplementary Figure 2. Summary of the Nanopore direct RNA sequencing reads. (A) Distribution of the read length of direct RNA sequencing reads and open reading frames (in brown) obtained from the official gene set of the locust genome. (B) Distribution of the read length and sequence identity of error-corrected Nanopore direct RNA sequencing reads. Reads longer than 5 Kb are not shown. Kb, kilobases. (C) An example diagram shows the isoform diversity in a gene whose gene structure covers a ~289-Kb genomic region. (D) Distribution of the sequence identity of error-corrected direct RNA sequencing reads. Sequence identities of the aligned direct RNA sequencing reads were evaluated based on the alignment against the locust genome. (E) Relationships between the number of direct RNA sequencing reads and percentage of detected genes in the 13,634 protein-coding genes that are detected in the Illumina datasets.



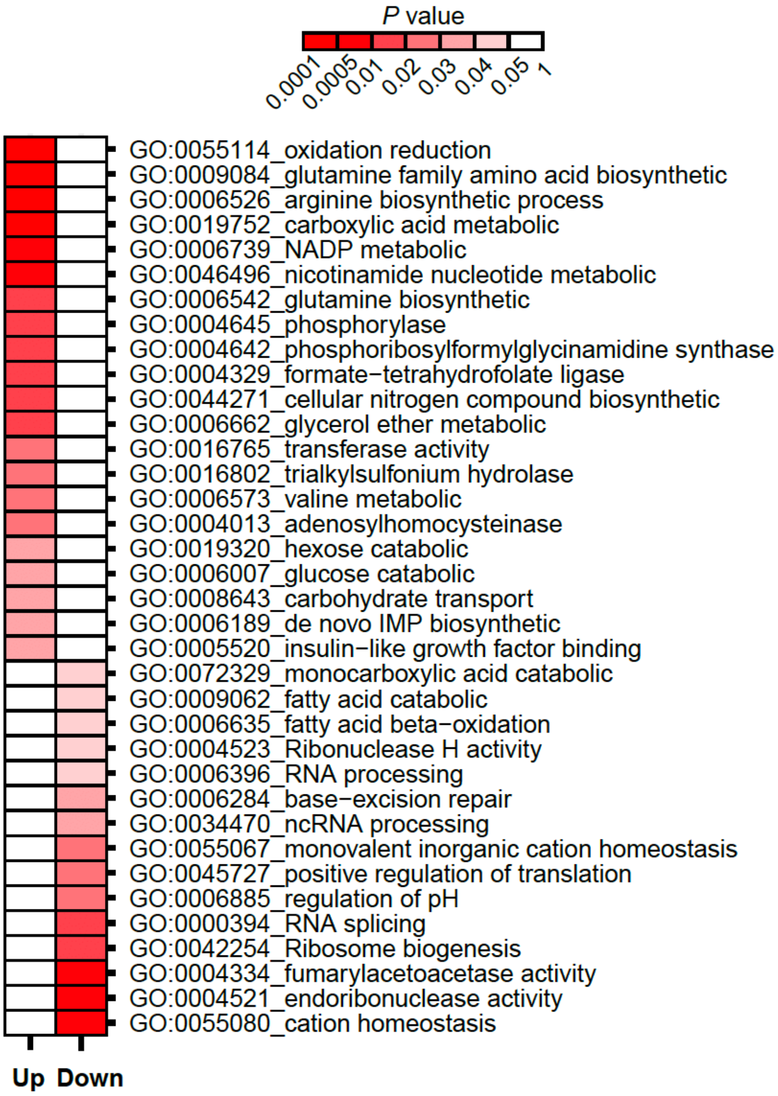
Supplementary Figure 3. Motif enrichment in the flanking region of splicing donor sites and acceptor sites in the top 10 TE families ranked by transcriptomic coverage. The motif enrichment was performed using the findMotifs program from HOMER package.



Supplementary Figure 4. Examination of *Piwi* RNAi efficiency. The data are shown as mean ± SEM (n=6), *$P < 0.05$. Ribosomal protein RP49 was used as endogenous control.



Supplementary Figure 5. MA plot comparing gene expression in *dsPiwi* and *dsGFP* samples. The significantly differential protein-coding genes inferred from DEseq2 package were marked in orange.



Supplementary Figure 6. Heatmap illustration of gene ontology enrichment analysis. Up and down represent the up-regulated and down-regulated gene ontologies in *dsPiwi* samples, respectively.

References:

1. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics 2014; 30:3506-14.
2. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005; 21:1859-75.
3. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, et al. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun 2014; 5:2957.