Critical roles of long noncoding RNAs in *Drosophila* spermatogenesis

Kejia Wen[1, #], Lijuan Yang[1, 2, #], Tuanlin Xiong[1, #], Chao Di[1], Danhui Ma[1], Menghua Wu[1], Zhaoyu Xue[1], Xuedi Zhang[1], Li Long[1], Weimin Zhang[1], Jiaying Zhang[1], Xiaolin Bi[3], Junbiao Dai[1, 4], Qiangfeng Zhang[1, 4, 5], Zhi John Lu[1, 4], and Guanjun Gao[1, 4, *]

[1] MOE Key Laboratory of Bioinformatics, Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China

[2] College of Animal Science, Tarim University, Xinjiang, 843300, China

[3] Institute of Cancer Stem Cell, Dalian Medical University, Dalian, 116044, China

[4] Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

[5] Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, China

[#]Equal contribution

[*]Corresponding author:

Guanjun Gao, School of Life Sciences, Tsinghua University, Yuanmingyuan Road 1, Beijing 100084, China. E-mail: gaogu@mail.tsinghua.edu.cn

Supplemental Methods

Fly culture and stocks

Flies were maintained on standard maize/sucrose/yeast medium at 25°C. w1118 was used as the wild type strain. The Ligase4 mutant, HIS2AV-RFP, Protamine B-GFP, and all balancer fly stocks were obtained from the Bloomington Stock Center (Bloomington, IN, USA).

RNA isolation and qRT-PCR

Total RNA from wild type fly testes, ovaries, 0-2 h embryos, and lncRNA KO mutant fly testes were isolated using TRIzol (Invitrogen). The PrimeScript™ II First Strand cDNA Synthesis Kit (Takara) was used for reverse transcription with random primers. qRT-PCR analyses of tissue-specific lncRNA expression and gene expression changes in lncRNA KO mutants were performed on a StepOne Plus (ABI) using SYBR Premix Ex Taq™ II (Takara) in 20 μl total volume. All reactions were performed in triplicate. The housekeeping gene *Rpl32* was used as an internal control. The testis-specific expressed gene *Mst35Ba* (also known as Protamine A) was used as a testis-specific control. We chose *X8C*, a chromosome X-linked intergenic region that is transcriptionally silent, as a negative control. Primer sequences for these three controls were as follows:

Rpl32 forward primer: 5’ AAGATGACCATCCGCCCAGCATAC 3’;

Rpl32 reverse primer: 5’ ACGCACTCTGTTGTCGATACCCTTG 3’;

Mst35Ba forward primer: 5’ CATCCAATAAGGAGAGCACCTCA 3’;

Mst35Ba reverse primer: 5’ CTTTCTATTCTCCGAGAGCCTG 3’;

X8C forward primer: 5’ TTCACTTGTCCGACGTTGTT 3’ and

X8C reverse primer: 5’ CGTTTAGATCTGCGGCAATT 3’.

Whole-mount RNA *in situ* hybridization

DNA templates for probe synthesis were the RT-PCR products (400 bp-1 kb) generated from total RNA of *Drosophila* testes. Digoxigenin (dig)-labeled antisense RNA probes were *in vitro* transcribed from the DNA templates using a T7 RNA polymerase promoter and the Roche Dig-RNA labeling mix. Testes were dissected from young adult flies (0-2 days old), fixed, and processed for hybridization. *In situ* hybridization was performed as described (Morris et al., 2009). Labeling was detected using alkaline phosphatase-conjugated anti-Dig antibodies. An RFP probe was used as negative control, and the Cyclin B probe was used as the positive control.

Donor plasmid construction and gRNA design

The donor plasmid include an RFP marker, two FRT sites for FLP-mediated integration, and two fragments homologous to the 5’ and 3’ ends of the target lncRNA to support HR. To construct the donor plasmid, a synthetic DNA fragment containing two FRT and 3P3-RFP sequences was cloned into vector pUAST. Subsequently, the two homologous fragments (1-1.5 kb each) were inserted flanking the RFP cassette to obtain the final donor construct. To efficiently generate DNA double-strand breaks inside the lncRNA loci, two gRNAs were designed inside each lncRNA target (Supplemental Table S3). The target sequence template is 5’GGA/G-N17/18-NGG-3’.

Modified cloning strategy for programmable donor plasmid construction

For programmable construction of the donor plasmid, we performed a two-step Golden Gate assembly to generate the lncRNA knockout homologous donor library. First, the three parts (left homology arm, 3P3-RFP, and right homology arm) were assembled in the HCKan-P, HCKan-O, and HCKan-T acceptor vectors, respectively. The 5' primer adaptor agcgtgGGTCTCgGGCT and 3' primer adaptor gtgctgGGTCTCaCATC were used for left homology arm amplification; the 5' primer adaptor agcgtgGGTCTCaGATG and 3' primer adaptor gtgctgGGTCTCgGCTA were used for 3P3-RFP amplification; and the 5' primer adaptor agcgtgGGTCTCtTAGC and 3' primer adaptor gtgctgGGTCTCgGAGG were added for right homology arm amplification. These three parts could be efficiently assembled in the corresponding vectors by following the standard Golden Gate protocol (Engler et al. 2008). Another acceptor vector, POT1-RFP, was used to receiving the three parts in the correct order, yielding a homologous donor plasmid for knockout of each lncRNA.

Cas9/gRNA-mediated lncRNA deletion screen

lncRNA knockout flies were identified by RFP expression in the eye when viewed under a fluorescence stereomicroscope. Homozygous lncRNA knockout flies were established by subsequent crossing with balancer flies. For molecular verification, genomic DNA was extracted from heterozygous and homozygous lncRNA mutant flies, followed by PCR using primers 100-300 base pairs upstream and downstream of the lncRNA loci.

***In cis* and *in trans* rescue of lncRNA knockout flies**

For rescue plasmid construction, the promoter and cDNA of the targeted lncRNA were amplified and inserted into vector pUAST-attB250. For *in cis* lncRNA rescue, purified plasmids (0.8 μg/μl) were injected into embryos of the corresponding lncRNA knockout flies carrying phiC31 recombinase. For transgenic rescue, purified plasmids (0.8 μg/μl) were injected into the embryos of flies carrying attP on a chromosome other than the one harboring the lncRNA. lncRNA rescue flies were identified by the red eye phenotype.

Off-target analysis

To investigate the possibility of off-target effects from our optimized CRISPR system, we searched the fly genome for potential off-target matches to the gRNA sequence of each lncRNA. For sequencing analysis, genomic DNA from flies of each lncRNA knockout mutant was used as the template, and the potential off-target sites were PCR-amplified by specific primers flanking the putative cleavage site.

Testis imaging, phalloidin staining, and immunohistochemistry

For imaging, testes of 4-7 days old adult males of lncRNA knockout stocks were dissected in PBS buffer. For live squash, dissected testes were transferred to a small drop of PBS buffer on a slide. To release the contents of the testes, a cover slip was placed on the microscope slide and gently pressed down. The testes were stained with DAPI in PBS plus 0.1% Triton X-100 for 5 min, and mounted in PBS. Images were acquired on a Zeiss Imager Z2.

For phalloidin staining, ICs were stained using FITC-phalloidin as described in Fabrizio et al. (Fabrizio et al., 1998).

For immunohistochemistry, testes from lncRNA mutant and wild type flies were dissected in chilled PBS. Fixation was performed in 4% paraformaldehyde in PBS containing 0.1% Triton X-100 (PBST) for 25 min at room temperature. The fixed testes were incubated overnight at 4°C with primary antibody diluted in PBST, followed by incubation for 2 h at room temperature in secondary antibody diluted in PBST. The immunostained testes were mounted in Vectashield medium containing DAPI. Antibodies were rabbit anti-myosin VI (1:200) and goat anti-rabbit IgG/RBITC (1:200).

RNA-seq analysis

Testes of 3-5 days old adult males from *CR42858*$^{-/-}$, *CR44585*$^{-/-}$, *lncRNA:TS1*$^{-/-}$, *lncRNA:TS2*$^{-/-}$, *CR45542*$^{-/-}$ and wild type stocks were dissected in PBS buffer. For each of lncRNA knockout mutants and wild type, two batches of testes were dissected independently as biological replicates. Total RNA from testes were isolated using TRIzol separately. To remove contaminating DNA, each of total RNA extracts was first treated with RNase-free DNase I (TaKaRa, Dalian, China), and RNAs with 3' poly(A) tails were concentrated by capture on magnetic oligo (dT) beads. The enriched poly(A) RNA was fragmented to ~200 bp in fragmentation buffer, and the resultant fragments were subjected to reverse transcription by priming with random hexamers to synthesize first-strand cDNA. The second strand was generated using the SuperScript Double-Stranded cDNA Synthesis kit (Invitrogen, CA, USA) and purified using magnetic beads. Following end repair, a single adenine base was added to the 3' end of each cDNA. Sequencing adaptors were ligated to the fragments, and the sample was subjected to agarose gel electrophoresis to select the range of fragments suitable for PCR

amplification. Sample quality was verified on an Agilent 2100 Bioanalyzer, and RNA was quantitated on an ABI StepOnePlus Real-Time PCR System. Library products were sequenced on an Illumina HiSeq 2000 at the BGI (Shenzhen, China; http://www.genomics.cn/index). At the data-processing step, raw reads were filtered by removal of adaptor sequences, low-quality reads in which > 50% of bases were low-quality (≤ 5), and reads in which the percentage of unknown bases (N) was greater than 10%. The cleaned data were aligned to the reference sequence using SOAPaligner/SOAP2 **(**http://soap.genomics.org.cn/soapaligner.html**)**. For the purpose of alignment, a maximum of two mismatches was permitted. The *D. melanogaster* genome and gene datasets were downloaded from FlyBase (r6.02), which was used as reference. To assess sequencing saturation, the number of genes identified was plotted against the number of cleaned reads to determine when no further genes could be detected by adding reads, implying full saturation. To evaluate the quality of the RNA-seq dataset, the distribution of gene coverage in each sample was analyzed. Gene expression levels were calculated using the Reads Per Kb per Million reads (RPKM) method, as described by Mortazavi et al. (Mortazavi et al., 2008). Furthermore, R package NOISeq (version 2.14.1) was utilized to identify differentially expressed genes between two samples with biological replicates (Tarazona et al., 2008). In this approach, the P-value corresponds to differential gene expression test, and the FDR (False Discovery Rate) is a method for determining the P-value threshold for multiple tests. We used FDR ≤ 0.001 and |log2Ratio ≥ 1| as the threshold of significance for differences in gene expression.

Evolutionary age and sequence conservation analysis of lncRNAs

PhastCons scores represent probabilities of negative selection (range between 0 and 1) at single-nucleotide resolution. The smaller the divergence of a DNA segment across species, the more likely it is that the segment belongs to a conserved element maintained by negative selection (Siepel et al. 2005). We parsed the PhastCons wigFix file of *D. melanogaster* and obtained per-site scores after unifying coordinate information from the 15-way genome-wide base-by-base scores from the UCSC Genome Browser (http://genome.ucsc.edu/), and then computed the overall distribution of PhastCons scores. In addition to lncRNAs, as comparative controls we also randomly sampled and computed PhastCons score distribution for groups of focused genomic features or regions that covered protein-coding regions (also including testis-specific CDS), intergenic regions, and so on. The sample size represented at least $1/10^{th}$ of the total number of nucleotide sites in each dataset of genome features. Statistical processing and graphing of score data was performed in R (R Core Team 2016).

References

Engler C, Kandzia R, Marillonnet S. 2008. A one pot, one step, precision cloning method with high throughput capability. *PloS One* **3**: e3647.

Fabrizio JJ, Hime G, Lemmon SK, Bazinet C. 1998. Genetic dissection of sperm individualization in Drosophila melanogaster. *Development* **125**: 1833-1843.

Morris CA, Benson E, White-Cooper H. 2009. Determination of gene expression patterns using in situ hybridization to Drosophila testes. *Nat Protoc* **4**: 1807-1819.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.

R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-1050.

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. 2011. Differential expression in RNA-seq: a matter of depth. *Genome Res* **21**: 2213-2223.