Supplemental Materials:

Panoramic transcriptome analysis and functional screening of long non-coding RNAs in mouse spermatogenesis

Kai Li,[1,3] Jiayue Xu,[1,3] Yanyun Luo,[1,3] Dingfeng Zou,[1] Ruiqin Han,[1] Shunshun Zhong,[1] Qing Zhao,[1] Xinyu Mang,[1] Mengzhen Li,[1] Yanmin Si,[1] Yan Lu,[1] Pengyu Li,[1] Cheng Jin,[1] Zhipeng Wang,[1] Fang Wang,[1] Shiying Miao,[1] Bo Wen,[2] Linfang Wang,[1] Yanni Ma,[1,*] Jia Yu,[1,*] and Wei Song[1,*]

[1] Department of Biochemistry and Molecular Biology, State Key Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and School of Basic Medicine, Peking Union Medical College, Beijing 100005, China; [2] Key Laboratory of Metabolism and Molecular Medicine, Ministry of Education; Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China

[3]These authors contributed equally to this work.

*Corresponding authors: songwei@ibms.pumc.edu.cn; j-yu@ibms.pumc.edu.cn; yanni_ma@126.com;

Contents:

Supplemental Figure S1

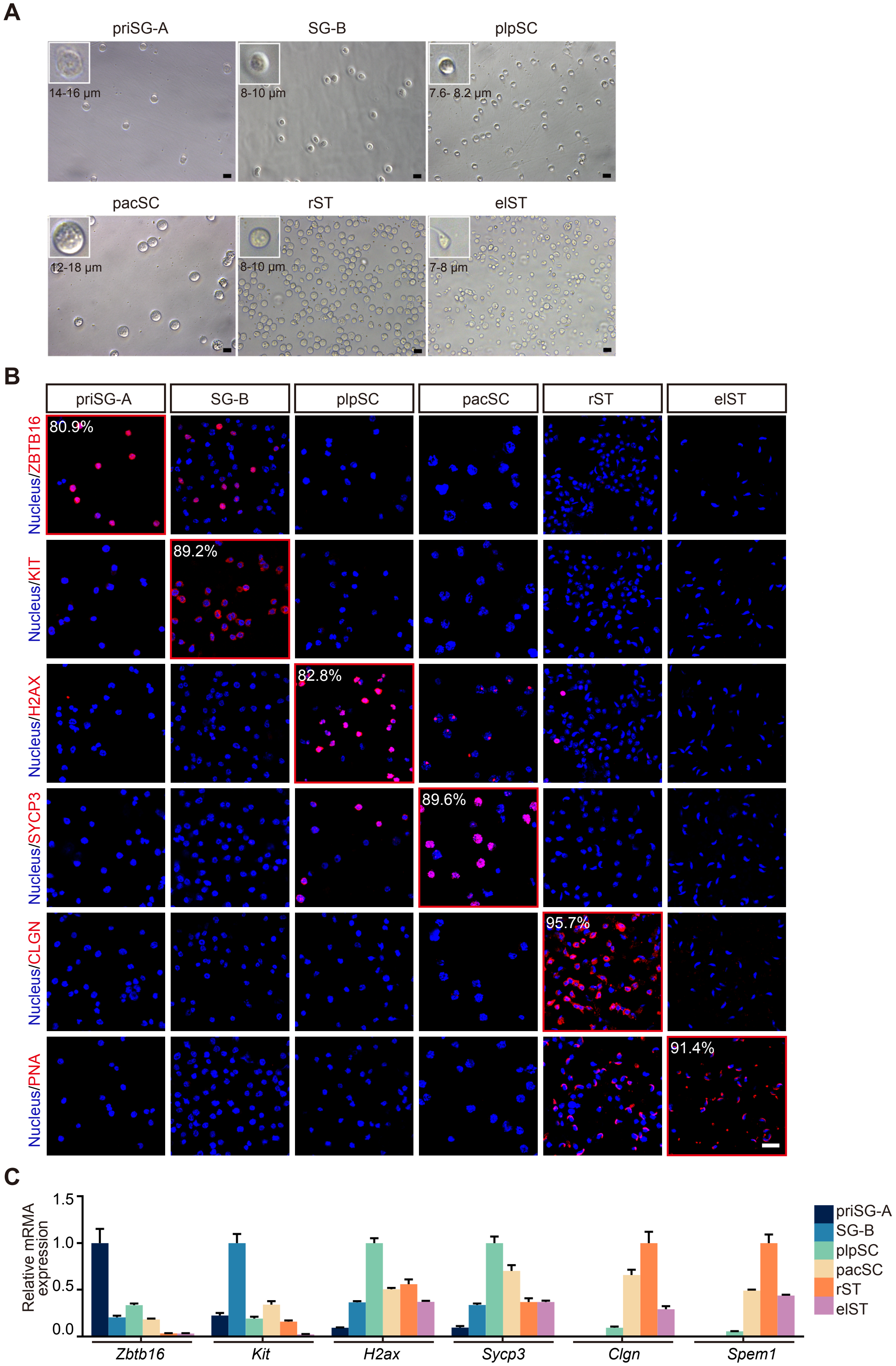

Figure S1. Isolation and identification of six germ cell types from mouse testis. (*A*) Morphology and size detection of six distinct germ cell types under a phase contrast microscope. Scale bar, 10 μm. (*B*) Immunostaining of each isolated germ cell type for marker genes (*red*): ZBTB16, KIT, H2AX, SYCP3, CLGN and PNA. Nuclei were stained with DAPI (*blue*). Six random fields under confocal microscopy were selected to calculate the marker gene positive rate of each germ cell type. Scale bar, 20 μm. (*C*) Relative expression level of marker genes in each germ cell type as detected by qRT-PCR. Data represent the mean ± SEM for three biological replicates.

Supplemental Figure S2

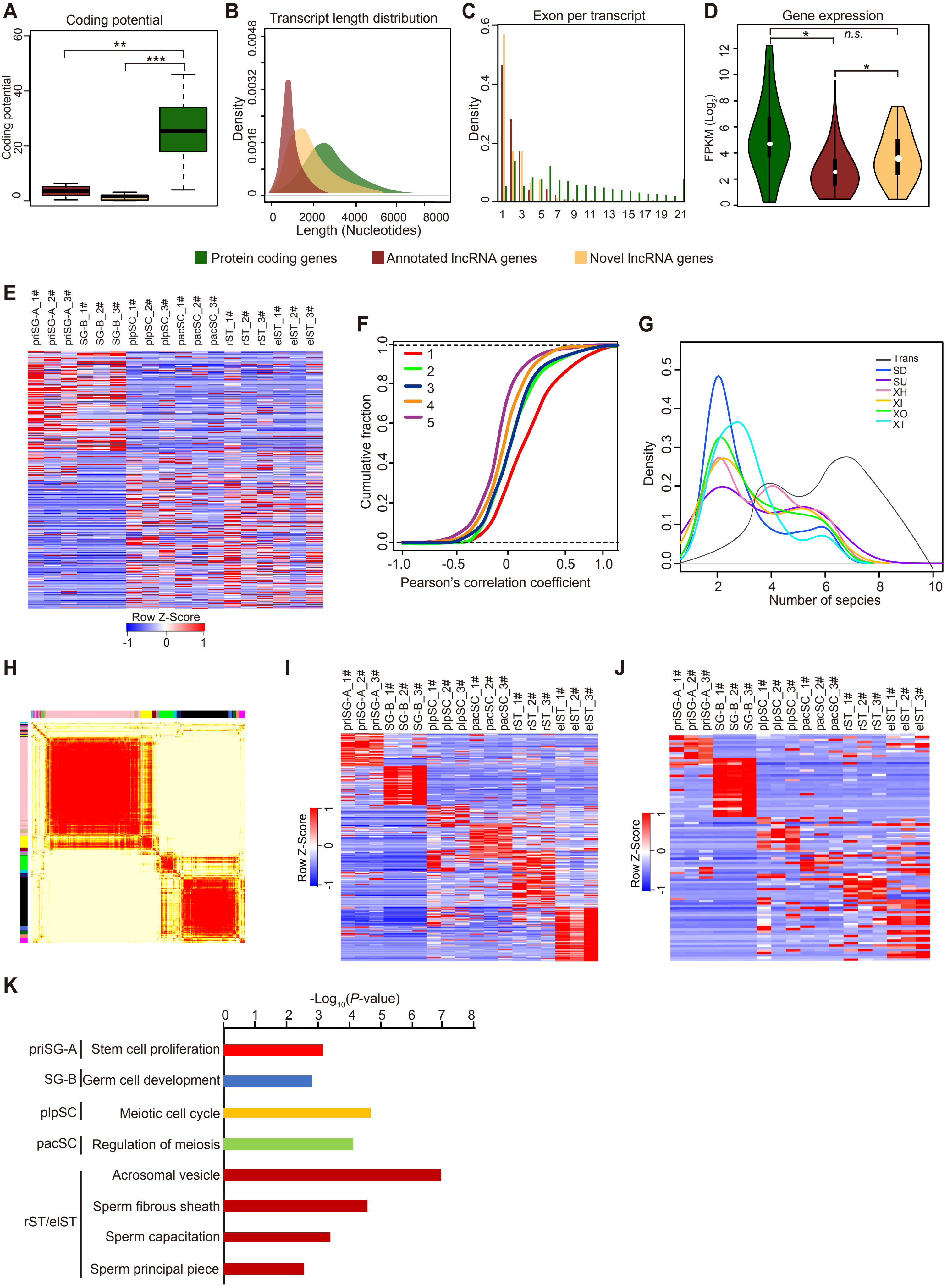

Figure S2. Related to Fig. 1. Global lncRNA expression dynamics during mouse spermatogenesis. (*A*) Protein-coding potential of protein-coding genes, annotated and novel lncRNA genes. (**) *P*<0.01; (***) *P*<0.001, Student's *t*-test. (*B*) Transcript length of protein-coding genes, annotated and novel lncRNA genes. (*C*) Number of exons per transcript of protein-coding genes, annotated and novel lncRNA genes. (*D*) Expression level of protein-coding genes, annotated and novel lncRNA genes. (*) *P*<0.05; (*n.s.*) *P*>0.05, Student's *t*-test. (*E*) Heatmap showing novel lncRNAs genes identified during spermatogenesis. (*F*) Pairwise Pearson's correlation analysis between lncRNAs and their nearby genes. "1" refers to the closest protein-coding gene; "2-4" refers the other distal protein-coding genes. (*G*) Conservation analysis of the different biotypes of lncRNAs in different species. Biotypes of lncRNAs according to their genomic locations relative to the neighboring protein-coding genes: antisense lncRNA/mRNA gene pairs in the head-to-head position were designated divergent ("XH"); antisense lncRNA/mRNA gene pairs in the tail-to-tail position were designated convergent ("XT"); the gene body of an antisense lncRNA can be located within a protein-coding gene ("XI") or can completely encompass a protein-coding gene ("XO"); the lncRNAs transcribed in the same direction can be located downstream ("SD") or upstream ("SU") of the neighboring coding genes. (*H*) Co-expression network analysis of lncRNAs and mRNAs during spermatogenesis using WGCNA. (*I*) Heatmap showing 271 differentially expressed novel lncRNAs during spermatogenesis. (*J*) Heatmap showing 94 germ cell signature novel lncRNAs. (*K*) Enriched GO terms for the germ cell developmental signature lncRNAs based on their neighboring mRNAs.

Supplemental Figure S3

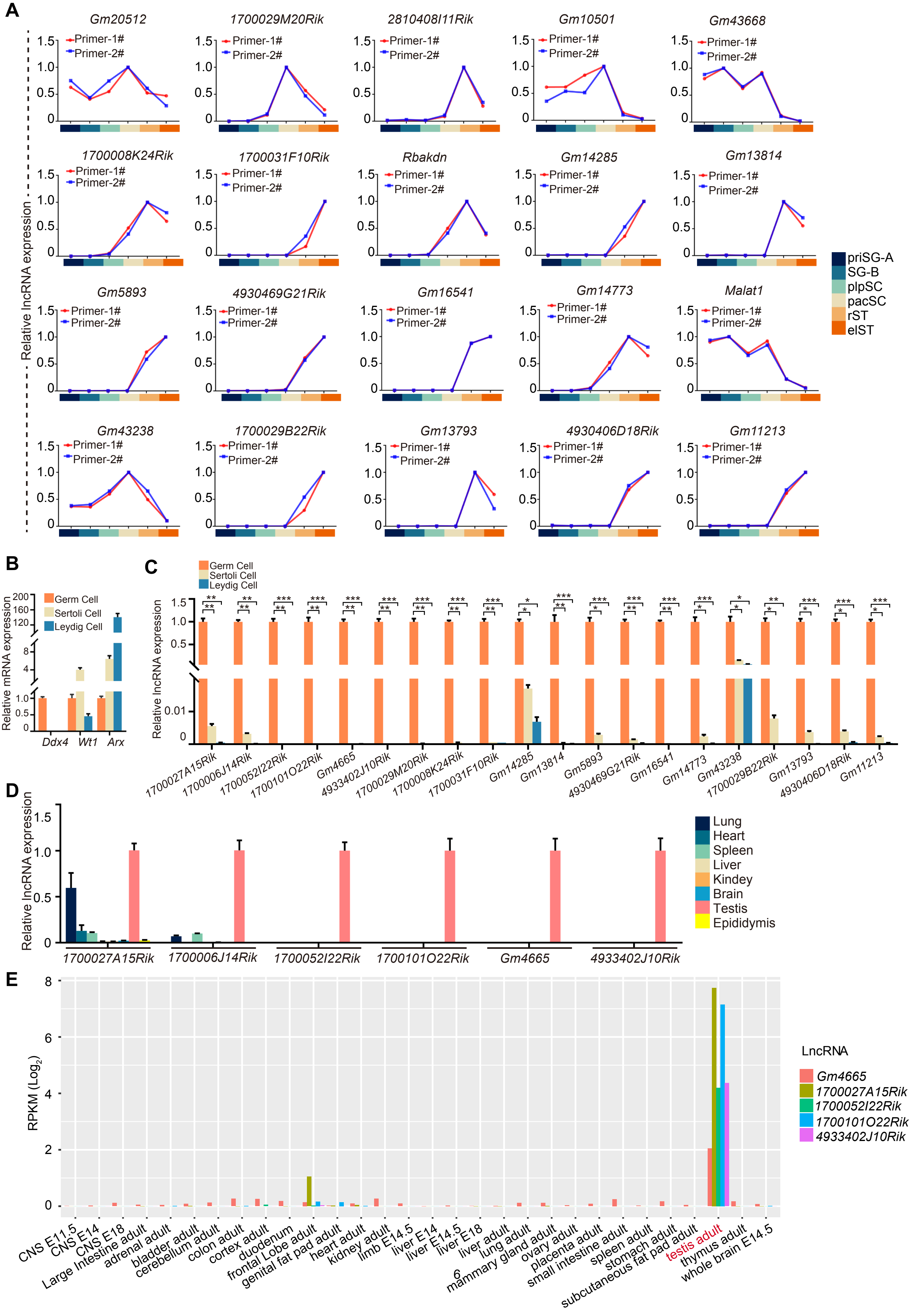

Figure S3. Related to Fig. 2. Characterization of lncRNA candidates in mouse spermatogenesis. (*A*) Expression validation of 27 candidate lncRNAs in each germ cell types by qRT-PCR. Two independent pairs of primers were used for each lncRNA (primer-1# and primer-2#). *Gm43263* was not detected in germ cells. Data represent mean ± SEM for three biological replicates. (*B*) Relative expression level of marker genes in Germ cells, Sertoli cells, and Leydig cells as dectected by qRT-PCR. Data represent mean ± SEM for three biological replicates. (*C*) Germ cell-specific expression screening of 20 lncRNA candidates by qRT-PCR in Germ cells, Sertoli cells and Leydig cells. Data represent the mean ± SEM for three biological replicates. (*) *P*<0.05; (**) *P*<0.01; (***) *P*<0.001, Student's *t*-test. (*D*) Testis-specific expression validation of six candidate lncRNAs in multiple tissues by qRT-PCR. Data represent the mean ± SEM for three biological replicates. (*E*) Expression pattern analysis of candidate lncRNAs using transcriptome data generated by the Mouse ENCODE Project.

Supplemental Figure S4

A

Candidate lncRNAs	Chromosome	Location	Full length (nt)	Poly(A)
1700027A15Rik	1	72,984,844-73,025,507	404	+
1700006J14Rik	10	120,364,157-120,384,336	1128	+
1700052I22Rik	12	80,923,432-80,924,449	428	+
1700101O22Rik	12	7,372,039-7,380,330	335	+
Gm4665	9	115,180,404-115,186,058	745	+
4933402J10Rik	5	59,963,119-59,986,004	1219	+

B

Candidate lncRNAs	C/NC	Coding potential score
1700027A15Rik	Noncoding	0.144823
1700006J14Rik	Noncoding	0.189091
1700052I22Rik	Noncoding	0.0220451
1700101O22Rik	Noncoding	0.0265628
Gm4665	Noncoding	0.0522966
4933402J10Rik	Noncoding	0.40559
Hotair	Noncoding	0.184882
Actb	coding	1

C

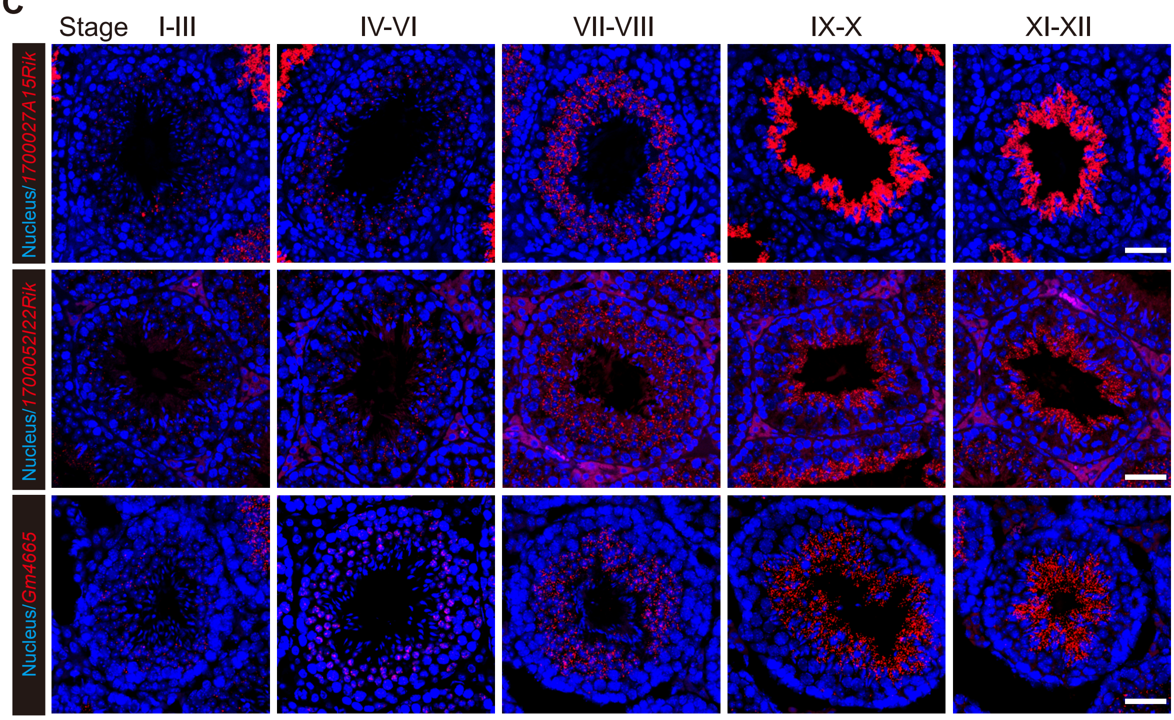

Figure S4. Coding potential analysis and subcellular localization of candidate lncRNAs. (*A*) Full length of six candidate lncRNAs determined by 5’ and 3’-rapid amplification of cDNA ends (RACE) assay. (*B*) Coding potential analysis of six candidate lncRNAs by Coding Potential Calculator (CPC 2.0) program. *Hotair and Actb* were respectively used as non-coding or coding RNA control. (*C*) Subcellular localization of candidate lncRNAs (*red*) at different stages of spermatogenesis via RNAscope analysis. Nuclei were stained with DAPI (*blue*). Scale bar, 50 μm.

Supplemental Figure S5

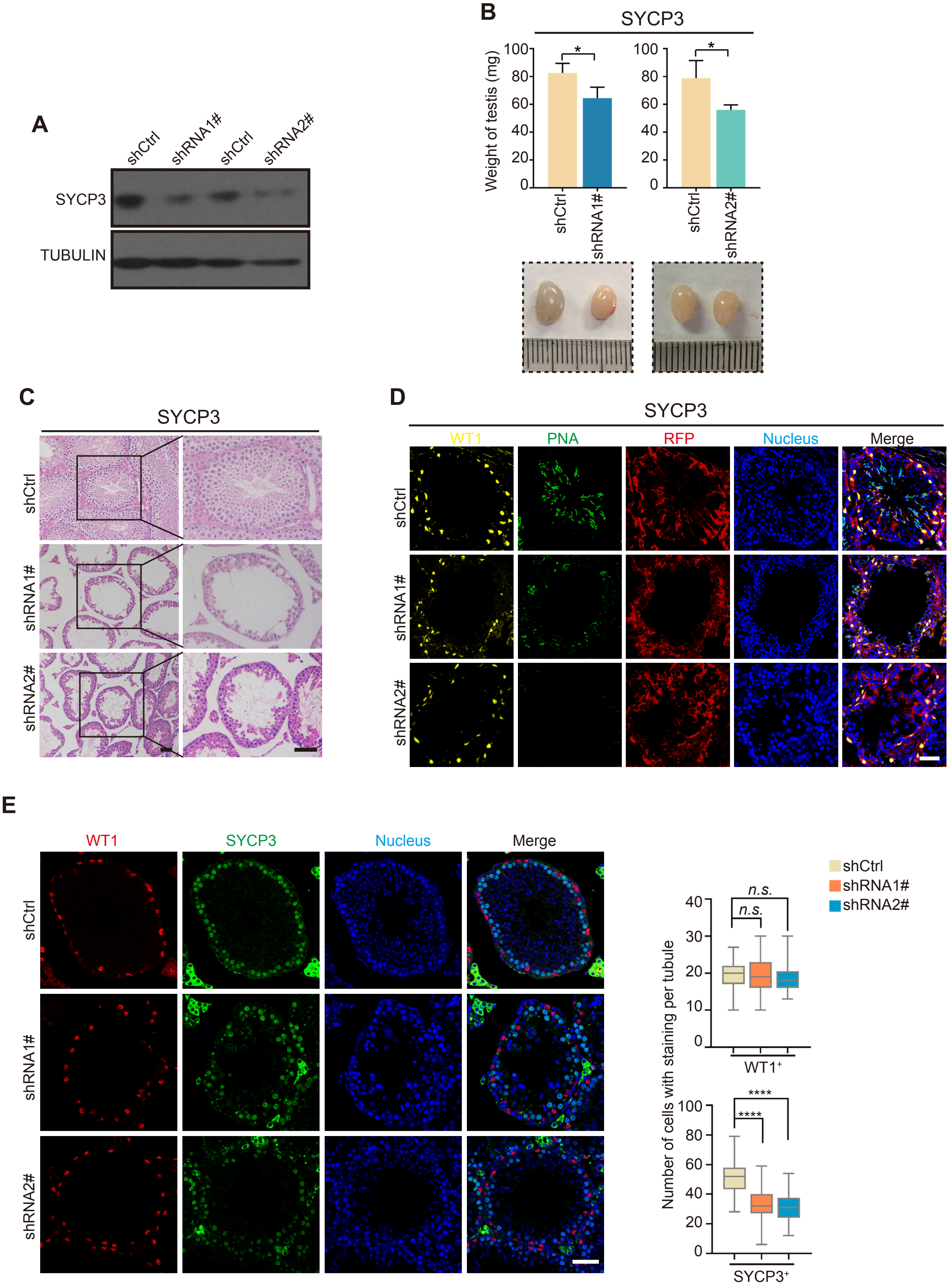

Figure S5. Knockdown of SYCP3 leading to defects in spermatogenesis by *in vivo* functional screening strategy. (*A*) Western blot analysis of SYCP3 knockdown *in vivo* by AAV9-shRNA-RFP. TUBULIN was used as loading controls. (*B*) Testis morphology and weight from the control (shCtrl) and SYCP3 knockdown (shRNA1#, shRNA2#) mice. *Top*: the average weight of testes; *bottom*: representative images of testes. Data represent mean ± SEM (*n*=3). (*) *P*<0.05 compared with shCtrl, Student's *t*-test. (*C*) H&E staining of testis sections from the control (shCtrl) and SYCP3 knockdown (shRNA1#, shRNA2#) mice. Scale bar, 50 μm. (*D*) Immunostaining of testis sections from the control (shCtrl) and SYCP3 knockdown (shRNA1#, shRNA2#) mice for WT1 (*yellow*), PNA (*green*), and RFP (*red*). Nuclei were stained with DAPI (*blue*). Scale bar, 50 μm. (*E*) Immunostaining of testis sections from the control (shCtrl) and SYCP3 knockdown (shRNA1#, shRNA2#) mice for WT1 (*red*), SYCP3 (*green*). Nuclei were stained with DAPI (*blue*). *Left*: representative images; *right*: number of cells with staining per tubule. Data represent the mean ± SEM of at least 100 seminiferous tubules from 3 mice; (****) *P*<0.0001; (*n.s.*) *P*>0.05 compared with shCtrl, Student's *t*-test. Scale bar, 50 μm.

Supplemental Figure S6

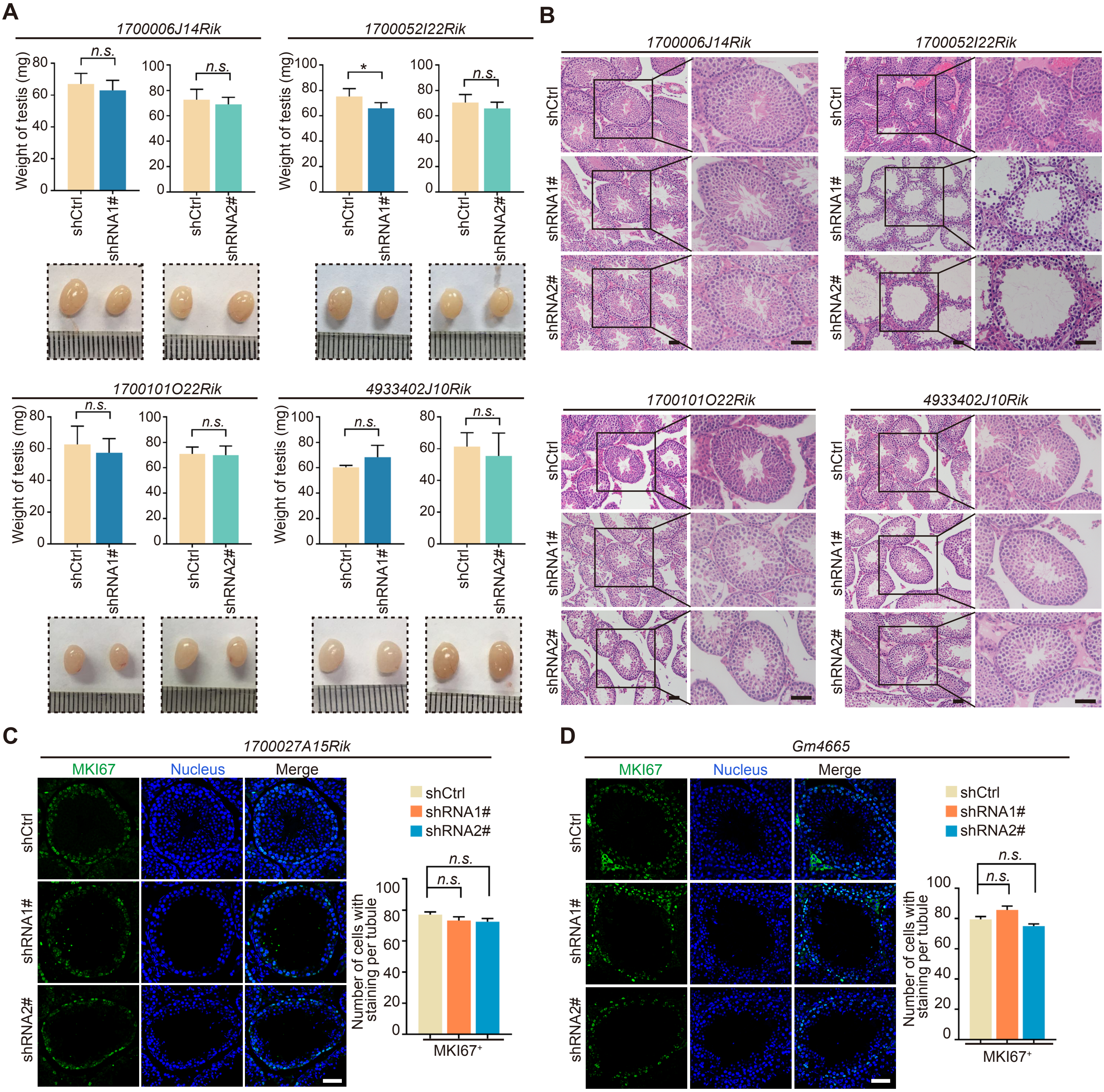

Figure S6. Related to Fig. 3. Functional screening of candidate lncRNAs in mouse spermatogenesis by *in vivo* functional screening strategy. (*A*) Testis morphology and weight from the shCtrl, shRNA1# and shRNA2# mice. *Top*: the average weight of testes; *bottom*: representative images of testes. Data represent mean ± SEM (*n*=3). (*) *P*<0.05, (*n.s.*) *P*>0.05 compared with shCtrl, Student's *t*-test. (*B*) H&E staining of testis sections from the shCtrl, shRNA1# and shRNA2# mice. Scale bar, 50 μm. (*C, D*) Immunostaining of testis sections from the control (shCtrl), *1700027A15Rik* or *Gm4665* knockdown (shRNA1#, shRNA2#) mice for MKI67 (*green*). Nuclei were stained with DAPI (*blue*). *Left*: representative images; *right*: number of cells with staining per tubule. Data represent mean ± SEM of at least 100 seminiferous tubules from 3 mice. (*n.s.*) *p*>0.05 compared with shCtrl, Student's *t*-test. Scale bar, 50 μm.

Supplemental Figure S7

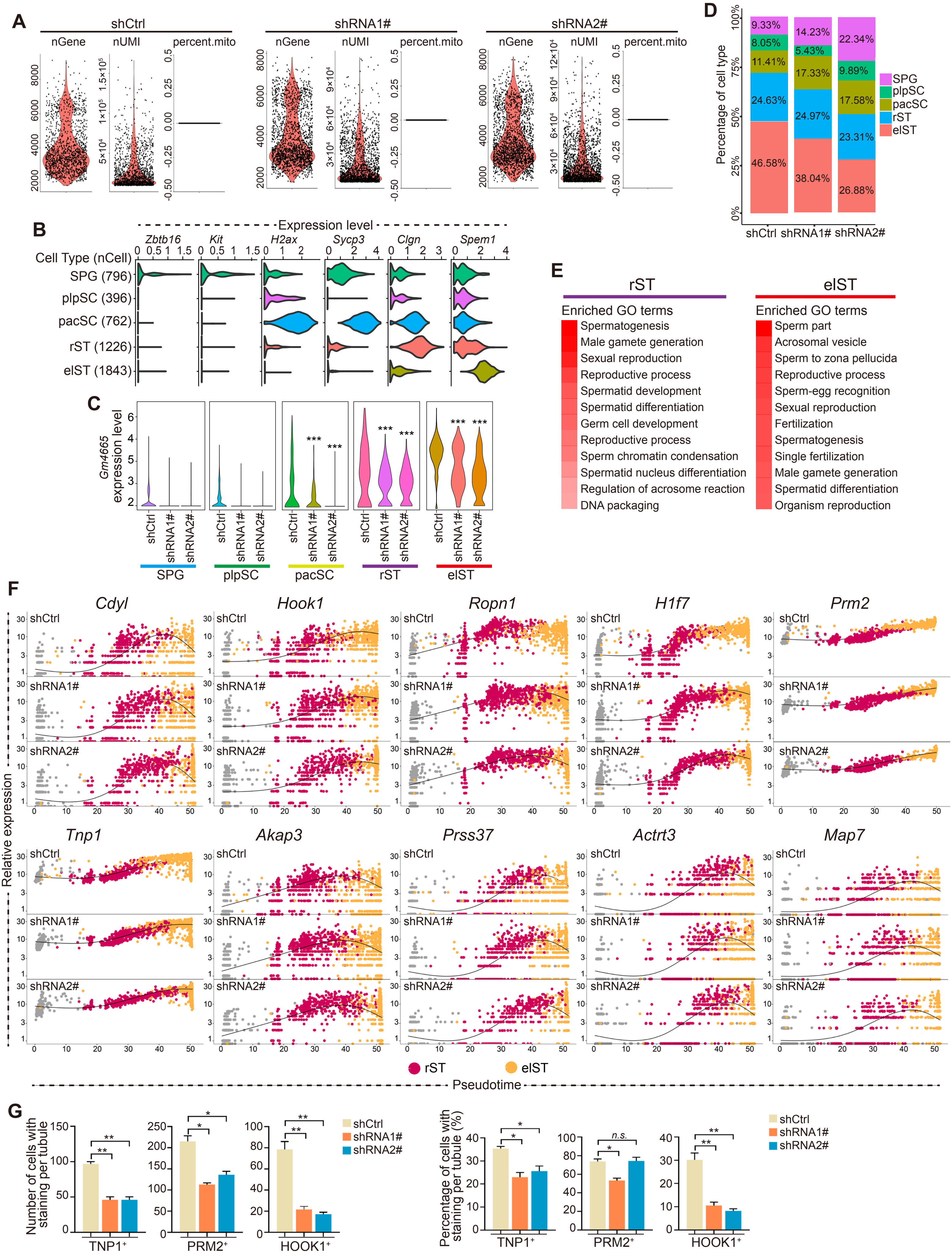

Figure S7. Related to Fig. 5. High-resolution dissection of transcriptional alterations induced by *Gm4665* knockdown in testis. (*A*) Quality control of single-cell RNA-seq dataset. Sequencing metrics of the three datasets generated (shCtrl, shRNA1#, shRNA2#). (*B*) Violin plot displaying the expression level of marker genes per cell of five distinct cell populations. (*C*) Violin plot displaying expression level of *Gm4665* in indicated cell populations between control and shRNA1#shRNA2# groups. *P*-values were calculated using Wilcoxon Rank-Sum test. (***) *P*<0.001 compared with shCtrl. (*D*) Proportion of each germ cell population in control and shRNA1#shRNA2# groups. (*E*) Enriched GO terms for the differentially expressed genes induced by *Gm4665* knockdown in rST (*left*) and elST (*right*) populations. (*F*) The expression dynamics of functional spermatogenic genes upon *Gm4665* knockdown in a pseudotime manner. (*G*) Number and percentage of cells with staining for TNP1, PRM2 and HOOK1 per tubule. Data represent the mean ± SEM of at least 50 seminiferous tubules from 3 mice. (*) *P*<0.05; (**) *P*<0.01; (*n.s.*) *P*>0.05 compared with shCtrl, Student’s *t*-test.

Supplemental Figure S8

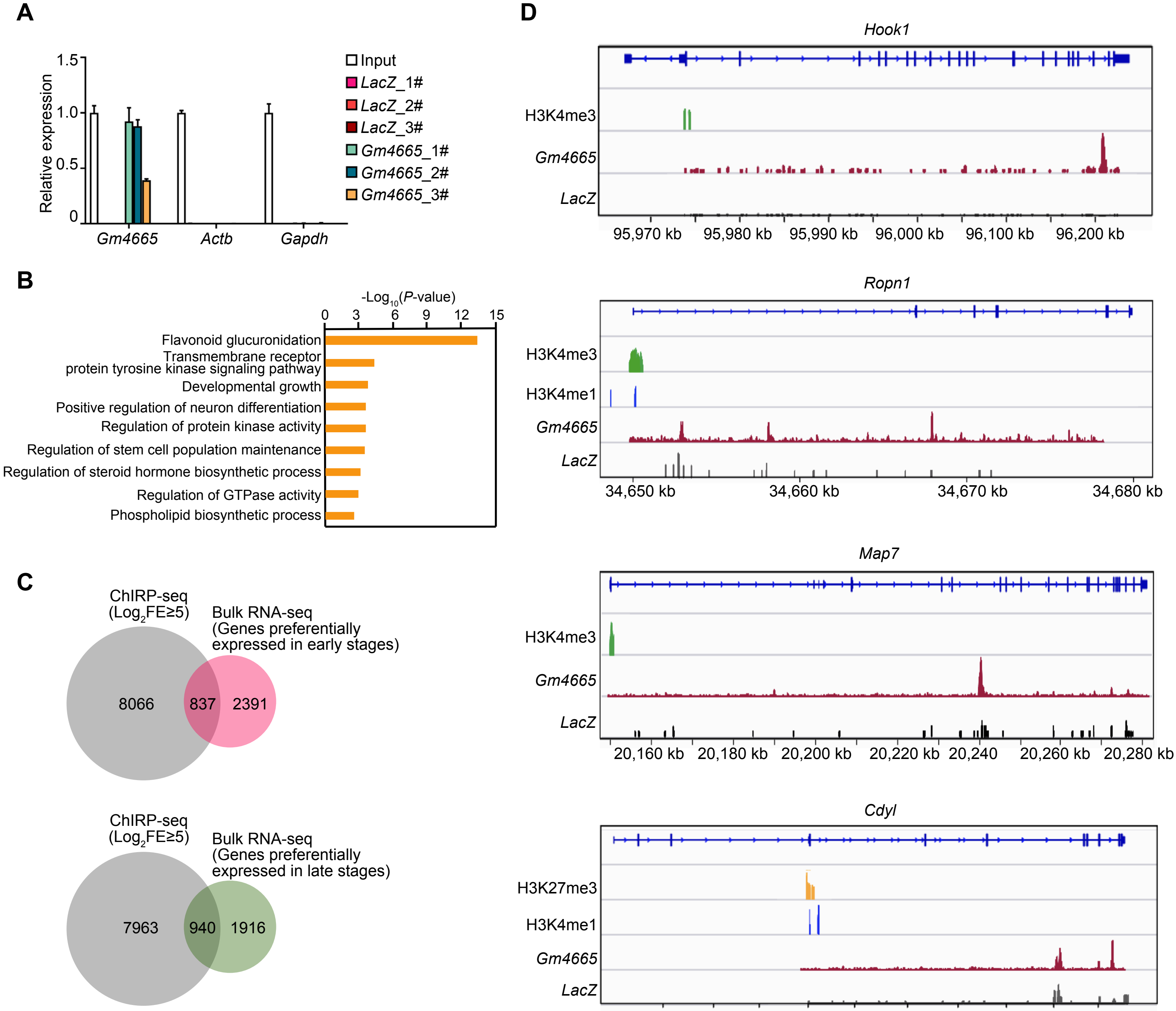

Figure S8. Related to Fig. 6. ChIRP-seq analysis of chromatin interactions between *Gm4665* and functional spermatogenic genes. (*A*) qRT-PCR validation of the biotinylated probes specifically targeting the endogenous *Gm4665* but not *LacZ* control*.* *Actb* and *Gapdh* mRNA were used as negative control. Data represent the mean ± SEM. (*B*) Enriched GO terms for the genes whose enhancer was bound by *Gm4665* and also marked with the histone modification marker. (*C*) Venn diagram showing the overlap of *Gm4665* ChIRP-seq genomic regions with genes preferentially expressed in early stages (*top*) and late stages (*bottom*) of spermatogenesis. FE: Fold Enrichment. (*D*) Graphical representation of *Gm4665* ChIRP-seq peaks as well as the histone modification ChIP-seq peaks among functional spermatogenic genes loci.

Supplemental Tables (.xlsx)

Supplemental Table S1. Quality control of sequencing data, including RNA-seq, Single-cell RNA-seq and ChIRP-seq.

Supplemental Table S2. Gene list: lncRNAs and mRNAs expressed in six distinct germ cell types during mouse spermatogenesis.

Supplemental Table S3. LncRNAs of six locus biotypes classification.

Supplemental Table S4. Differentially expressed genes (DEGs) upon *Gm4665* knockdown in rST and elST populations at the single-cell level.

Supplemental Table S5. Analysis of the chromatin fragments binding to *Gm4665* by ChIRP-seq.

Supplemental Table S6. Gene list: 134 candidate target genes mediated by *Gm4665*.

Supplemental Table S7. Primer sequences of qRT-PCR.

Supplemental Table S8. shRNA sequences used for knockdown of candidate lncRNAs and mRNA.

Supplemental Methods

Immunostaining

Male germ cells isolated from wild-type mouse testis were smeared on slides and fixed with 4% paraformaldehyde (PFA) in PBS. For testis tissue analyses, the testis were dissected from mice injected with AAV9-RFP after four weeks of injection, fixed with 4% PFA in PBS, and then incubated in sucrose-PBS solution (10% sucrose for 1 hr, 20% sucrose for 1 hr, and 30% sucrose overnight) at 4°C. Cryosections or cells were washed in PBS and permeabilized with PBS containing 0.5% Triton X-100. The slides were blocked with 5% BSA and then incubated with primary antibody at 4°C overnight. AlexaFluor™ conjugated secondary antibody (Invitrogen) was added for 1 hr at room temperature. Finally, the slides were mounted with SlowFade™ Gold Antifade Mountant with DAPI (Invitrogen) and observed under LSM 780 confocal microscope (Zeiss) for fluorescent signal analysis. Primary antibodies used in immunostaining were as follows: anti-ZBTB16 (Santa cruz, sc-28319), anti-HOOK1 (santa cruz, sc-398233), anti-KIT (Cell Signalling Technology, 3074), Anti- H2AX (Abcam, ab2893), anti-SYCP3 (Abcam, ab97672), anti-WT1 (Abcam, ab89901), anti-CLGN (Abcam, ab172477), Alexa Fluor™ 594 Conjugated-Lectin PNA (Invitrogen, L32459), anti-TNP1 (Proteintech, 17178-1-AP), anti-PRM2 (Proteintech, 14500-1-AP). Intensity and number of cells with staining were calculated using image analysis software of ZEISS ZEN lite 3.0; percentage of cells with staining per tubule was the ration of positive cell number/ nucleus number.

Northern blotting

Northern blotting was performed using the DIG Northern Starter Kit (Roche). Briefly, 10~20 μg of total RNA was denatured with RNA loading dye with formaldehyde at 65°C for 10 min and separated on a 1.2% agarose gel with formaldehyde for at least 2 hr until the RNA fragments were well separated. The RNA was then transferred to a nylon membrane (GE Healthcare) by capillary blotting with 20×SSC buffer for at least 6 hr or overnight and cross-linked with UV light. The membrane was subjected to a prehybridization step by incubation in DIG Easy Hyb Granules for 30 min at 68°C with gentle agitation. Denatured antisense DIG-labeled RNA probe (100~400 ng/ml) was added to prewarmed DIG Easy Hyb Granules and mixed well. The membrane was incubated at 68°C overnight with a probe/hybridization mixture. After hybridization and stringency washes with washing buffer, the membrane was incubated for 30 min in 50~100 ml blocking buffer. Finally, the membrane was exposed to CDP-star, a ready-to-use solution to detect chemiluminescent signals with X-ray film.

Rapid amplification of cDNA ends (RACE)

To identify the full length of lncRNAs, 5’ and 3’ RACE was performed using the SMARTer® RACE 5’/3’ Kit (TaKaRa) according to the manufacturer’s instructions. RACE-Ready cDNAs were synthesized from total RNA. RACE-PCR products were separated on 1%~2% agarose gels. The bands were excised and the extracted DNA (Gel Extraction Kit, CWBIO) was cloned into T vectors and sequenced bidirectionally using M13 forward and reverse primers. At least two colonies were sequenced for each clone product.

RNAscope in situ hybridization

Briefly, 10 μm cryosections were mounted on glass slides (SuperFrost, Fisher Scientific) and washed in PBS, followed by target retrieval by boiling slides in RNAscope Target Retrieval Reagent for 15 min while maintaining the temperature between 98°C and 102°C, then washed twice for 2 min in distilled water. Slides underwent protease treatment for 30 min at 40°C in an ACD HybEZ hybridization oven. Probes were then hybridized for 2 hr at 42°C in a HybEZ oven followed by RNAscope signal amplification. The slides were mounted with SlowFade™ Gold Antifade Mountant with DAPI (Invitrogen) and images were collected on an LSM 780 confocal microscope (Zeiss).

Adeno-associated virus (AAV9) production

For lncRNA knockdown, shRNA sequences were designed according to online BLOCK-iT RNAi tools (Thermo Fisher Scientific). AAV9 production was carried out following standard procedures as described previously. In brief, an AAV vector plasmid (pAAV-H1-RFP), an adenovirus helper plasmid (pAAV29), and an AAV helper plasmid (pAAV-RC) were transiently transfected into AAV-293 cells. Cell culture media and transfected cells were harvested separately after 72 h. Added 40% PEG (Sigma-Aldrich) in 2.5 N NaCl to the supernatant (final concentration: 8%) and mixed well, and then the cell pellet was suspended in 5 ml of lysis buffer (50 mM Tris-Cl, 150 mM NaCl and 2 mM $MgCl_2$) and stored at 4°C. Virus was purified by iodixanol gradient ultracentrifugation. The extracted virus was then transferred to an Amicon Ultra-15 concentration unit (Millipore) followed by a 10-min centrifugation at 3000 g followed by

4 rounds of repeated washing and spinning. Finally, virus was aliquoted and stored at −80°C. Viral titer was determined by qRT-PCR. The viral titer was 1×10^{13}/mL unless otherwise indicated.

Bulk RNA-seq

For RNA-seq, ribosomal RNA depleted and strand-specific libraries were constructed with the Ribo-zero gold (Epicentre) and TruSeq Stranded Total RNA Sample Prep kit (Illumina), and the libraries were sequenced using the Illumina HiSeq system.

Alignment of sequencing reads and gene expression analysis

All reads were trimmed down to 150 bp and mapped to the Gencode lncRNA catalog of the mouse genome (version mm10) using HISAT2 with default settings, and only uniquely mapped reads were kept (Pertea et al. 2016). Gene expression levels were expected of fragments per kilobase of transcript sequence per millions of base pairs mapped (FPKM) using HTseq (Anders et al. 2015). To ensure the accuracy of estimated FPKM values and removal of the auxiliary data, only genes with FPKM>0.1 in at least one sample were analyzed. The DEseq package was used to evaluate the statistical significance of differentially expressed genes for different single-cell groups based on the negative binomial distribution.

Co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) was performed according to previously described protocol (Casero et al. 2015). Expression associations of lncRNA genes and the adjacent protein-coding genes were built by Circle tools

(Krzywinski et al. 2009; R Core Team 2017). The dynamic expression of each module was performed using STEM (Ernst and Bar-Joseph 2006).

ChIRP-seq analysis

For each sample, reads from target and control lanes were aligned to reference genome mm10 using Bowtie, separately, and perbase coverage was normalized (Langmead et al. 2009). For each base pair of the genome, true coverage of this base in this sample was defined as the minimum coverage of the even lane and odd lane and a SAM file was generated for peak calling. Peaks of each sample were then called using MACS against its corresponding input with p-value cutoff 1×10^{-5} (Zhang et al. 2008). We filtered for peaks that share the same shape in the raw data from the two independent experiments and the following criteria: thresholds of average coverage >1.5, Pearson’s correlation >0.3, and fold enrichment against input >2 and annotated peaks by Homer. We then used Chip-seq reads in GSE49621, ChIP-seq reads were aligned to mm10 genome assembly using Bowtie (Langmead et al. 2009; Hammoud et al. 2014). Peak calling was done using MACS with the following configurations: -q 1e-5 --llocal=10000 --bw 300 -m 5,30 --slocal=1000 -B --keep-dup=1. Overlap region between ChIRP-seq reads and ChIP-seq reads were calculate using bedtools intersect and peaks were visualizated by IGV tools. Gene Ontology (GO) analysis was performed using Matescape (http://metascape.org/).

References

Anders S, Pyl PT, Huber W. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.doi:10.1093/bioinformatics/btu638.

Casero D, Sandoval S, Seet CS, Scholes J, Zhu YH, Ha VL, Luong A, Parekh C, Crooks GM. 2015. Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nat Immunol* **16**: 1282-1291.doi:10.1038/ni.3299.

Ernst J, Bar-Joseph Z. 2006. STEM: a tool for the analysis of short time series gene expression data. *Bmc Bioinformatics* **7**.doi:10.1186/1471-2105-7-191.

Hammoud SS, Low DH, Yi C, Carrell DT, Guccione E, Cairns BR. 2014. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell stem cell* **15**: 239-253.doi:10.1016/j.stem.2014.04.006.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome research* **19**: 1639-1645.doi:10.1101/gr.092759.109.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**: R25.doi:10.1186/gb-2009-10-3-r25.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* **11**: 1650-1667.doi:10.1038/nprot.2016.095.

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**: R137.doi:10.1186/gb-2008-9-9-r137.