Resource:

Redefining the piRNA-Producing Loci of the Mouse Testis as Genes

Xin Zhiguo Li,[1*] Xianjun Dong,[2*] Jie Wang,[2] Christian K. Roy,[1] Bo W. Han,[1] Jia Xu,[2] Melissa J. Moore,[1] Phillip D. Zamore,[1] and Zhiping Weng[2]

[1]Howard Hughes Medical Institute and Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA
[2]Program in Bioinformatics and Integrative Biology
University of Massachusetts Medical School, 364 Plantation Street, Worcester, MA 01605, USA

*These authors contributed equally to this work.

*Correspondence: zhiping.weng@umassmed.edu (Z.W.),
phillip.zamore@umassmed.edu (P.D.Z.)

Keywords: piRNAs; giant exons; bidirectional promoter; transposons; CpG

SUMMARY

Piwi-interacting RNAs (piRNAs) derive from 1–100 kbp genomic loci whose transcripts yield small silencing RNAs loaded into PIWI proteins. What genomic features distinguish piRNA precursor transcripts from mRNAs and non-coding RNAs are unknown. Here, we define piRNA precursor transcripts using a diverse set of high-throughput sequencing techniques. We find that piRNA precursors are transcribed by RNA polymerase II from both intergenic loci and protein-coding genes. The promoter and exon/intron architecture and transposon composition of the intergenic piRNA loci set them apart from genes producing other non-coding RNAs or mRNAs. In tissues without piRNAs, intergenic piRNA loci are transcriptionally silenced and have less RNA polymerase II at their promoters than active genes, but more than inactive genes, suggesting a "paused polymerase" silencing mechanism. Our data redefine piRNA clusters as genes producing long transcripts with identifiable promoters, introns, and polyadenylation signals, enabling the study of their transcriptional and post-transcriptional regulation.

INTRODUCTION

PIWI-interacting RNAs (piRNAs) silence transposons and repetitive sequences, protecting the genome from insertional mutagenesis, double-stranded breaks, and accumulation of toxic RNAs. piRNAs were discovered in *Drosophila melanogaster* (Aravin et al., 2001), just after small interfering RNAs were discovered in plants (siRNAs; Hamilton and Baulcombe, 1999) and flies (Hammond et al., 2000; Zamore et al., 2000), but were not recognized as a distinct class of small silencing RNAs until five years later (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006a; Grivna et al., 2006b; Lau et al., 2006; Vagin et al., 2006). piRNAs bind PIWI proteins, a sub-group of the Argonaute protein family whose members use small RNA guides, including siRNAs and microRNAs, to mediate transcriptional and post-transcriptional silencing. In most animals, piRNAs are predominantly expressed in germ cells.

In worms, piRNAs silence transposons and provide genome-wide surveillance for foreign sequences (Batista et al., 2008; Das et al., 2008; Lee et al., 2012; Shirayama et al., 2012). In planaria, piRNA pathway components are required for regeneration (Shibata et al., 2010), and piRNAs have been implicated in memory in *Aplysia* (Rajasethupathy et al., 2012). Mutant mice defective for piRNA pathway proteins are invariably male sterile, because of defects in spermatogenesis. Although some mouse piRNAs appear to play a role in silencing LINE1 transposons in the testis (Xu et al., 2008; De Fazio et al., 2011; Reuter et al., 2011) and imprinting (Watanabe et al., 2011), the broader function of mouse piRNAs during spermatogenesis remains unknown.

piRNAs are believed to derive from single-stranded RNA precursors because their biogenesis does not require Dicer, the double-stranded RNA-specific ribonuclease that produces siRNAs and microRNAs (Vagin et al., 2006;

Houwing et al., 2007). piRNA biogenesis has been proposed to begin with fragmentation of piRNA precursor transcripts into shorter intermediates. In the current model, these fragments bear a 5′ monophosphate, which facilitates their binding to PIWI proteins (Vourekas et al., 2012). After PIWI binds to a piRNA intermediate, a nuclease is thought to trim the 3′ end of the piRNA to a length characteristic of the particular PIWI protein (Kawaoka et al., 2011). Finally, the *S*-adenosylmethionine–dependent methyltransferase, Hen1, modifies the 2′ hydroxyl of the 3′ end of the piRNA (Horwich et al., 2007; Saito et al., 2007; Kamminga et al., 2010). Numerous other proteins, including Tudor-, helicase-, and nuclease-domain proteins, participate in piRNA biogenesis, but no specific molecular function has been established for any piRNA biogenesis factor except Hen1.

In flies and mammals, piRNAs map to large blocks of genomic sequence called clusters (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006a; Lau et al., 2006; Aravin et al., 2007). This organization suggests that piRNAs derive from long primary transcripts via multiple RNA processing steps. Supporting this view, a P-element transposon inserted into the 5′ end of the *flamenco* piRNA cluster in flies reduces the production of *flamenco* piRNAs 168 kbp downstream from the insertion site, suggesting that the P-element disrupts transcription of the entire locus (Brennecke et al., 2007). The heterochromatin protein Rhino binds the DNA of *Drosophila* clusters that are convergently transcribed; loss of Rhino blocks transcript accumulation from such dual-strand clusters, consistent with the idea that clusters correspond to transcribed, piRNA-producing genes (Klattenhoff et al., 2009). Analysis of piRNAs from a cluster into which the gene encoding green fluorescent protein was inserted by piggyBac transgenesis of immortalized, cultured silk moth germline cells provides evidence that piRNA clusters generate long, contiguous transcripts from which introns are excised

(Kawaoka et al., 2012). In contrast, each piRNA in *Caenorhabditis elegans* is transcribed separately (Ruby et al., 2006; Batista et al., 2008; Gu et al., 2012), suggesting that piRNA clusters in other animals might produce many shorter transcripts rather than a single long RNA.

Precisely defining the primary transcripts of the clusters cannot be achieved using piRNA data alone because piRNAs are too short to perform de novo transcript assembly. Yet even when strand-specific high-throughput sequencing of RNA (RNA-Seq) data was combined with piRNA data, at least one major class of mouse piRNA clusters, the "intergenic piRNA hotspots," appeared to "lack any defined boundaries" (Vourekas et al., 2012).

An alternative view is that defining the precise boundaries of the primary transcripts from piRNA-producing loci will require combining many distinct data types and multiple computational approaches to detect RNA splicing events and to achieve the resolution required to establish the absolute boundaries of resulting piRNA precursor transcripts. Here, we define piRNA precursor transcripts using a diverse set of high-throughput sequencing techniques. We find that piRNA precursors are transcribed by RNA polymerase II (pol II) from both intergenic loci and protein-coding genes. The promoter and exon/intron architecture and transposon composition of the intergenic piRNA loci set them apart from genes producing other non-coding RNAs or mRNAs. Our data redefine piRNA clusters as genes producing long transcripts with identifiable promoters, introns, and polyadenylation signals, enabling the future study of their transcriptional and post-transcriptional regulation.

RESULTS

Defining the piRNA-Producing Transcripts in Mouse Testis

To define the structure of piRNA-producing primary transcripts, we assembled the transcripts expressed in the testes of adult mice using strand-specific, paired-end RNA-Seq (Figure 1). We combined reference-based assembly (Cufflinks; Trapnell et al., 2012) with de novo assembly (Trinity ; Grabherr et al., 2011) to gain the high sensitivity rendered by reference-based assemblers and leveraging the ability of de novo assemblers to detect novel transcripts. We first mapped reads to the mouse genome using TopHat (Trapnell et al., 2009), then performed de novo transcriptome assembly using Trinity (Grabherr et al., 2011) to identify novel exon-exon junctions. Next, we used all the mapped reads, including the reads mapping to novel exon-exon junctions, to perform reference-based assembly (Cufflinks; Trapnell et al., 2010).

To identify those transcripts that produce piRNAs, we sequenced piRNAs from six developmental stages (10.5 dpp, 12.5 dpp, 14.5 dpp, 17.5 dpp, 20.5 dpp, and adult) of mouse testes (Li, et. al, accompanying manuscript) and then mapped them to the transcripts we assembled. To qualify as a piRNA-producing transcript, an assembled RNA was required to produce either a sufficiently high piRNA abundance (>100 ppm; parts per million uniquely mapped reads) or density (>100 rpkm; reads per kilobase of transcript per million uniquely mapped reads). These criteria retained both long transcripts producing an abundance of piRNAs and short transcripts generating many piRNAs per unit length.

To refine the termini of each piRNA-producing transcript, we supplemented the RNA-Seq data with high-throughput sequencing of the 5′ ends of RNAs bearing a N(5′)ppp(5′)N cap structure (cap analysis of gene expression;

CAGE) and of the 3′ ends of transcripts preceding the poly(A) tail (polyadenylation site sequencing; PAS-Seq). Conventional 5′ and 3′ RACE analysis of piRNA-producing transcripts confirmed the ends of 16 loci (data not shown). To provide independent confirmation of the 5′ end of each piRNA-producing transcript, we also determined the locations of histone H3 bearing trimethylated lysine-4 (H3K4me3), a histone modification highly enriched at transcription start sites (Guenther et al., 2007).

The assembled piRNA-producing transcripts almost certainly correspond to continuous RNAs in vivo, because the CAGE library used to annotate transcript 5′ ends was constructed after two rounds of poly(A) selection. Thus, the RNA molecules in our CAGE library likely derived from complete transcripts extending from the 5′ cap to the poly(A) tail. Indeed, the abundance of CAGE reads at the transcription start site correlated well with the abundance of PAS-Seq reads at the cleavage and polyadenylation sites for both piRNA-producing transcripts ($\rho = 0.46$, $p < 2.2 \times 10^{-16}$) and mRNAs that do not produce piRNAs ($\rho = 0.42$, $p < 2.2 \times 10^{-16}$), suggesting that complete transcripts were detected in the CAGE and PAS-Seq experiments.

A Transcript-based Set of Definitive piRNA Loci

Our transcriptome assembly effort yielded 466 piRNA-producing transcripts that define 214 genomic loci (Figure S1 and Table S1). Among the ~2.2 million distinct piRNA species and ~8.8 million piRNA reads from the adult mouse testis, the 214 genomic loci account for 92% of all piRNAs. Two distinct types of piRNA-producing transcripts emerged from this process. Two hundred forty-nine transcripts define 114 genic piRNA loci that overlap over >80% of their length with a protein-coding gene annotated in the RefSeq database (Pruitt et al., 2012).

These genic loci account for 22% of piRNAs in the 10.5-dpp mouse testis and 3.0% in the adult testis.

The remaining 217 transcripts define 100 "intergenic" piRNA loci that lie far from actively transcribed annotated genes (the median distance to the nearest annotated gene was 11 kbp compared to 2.0 kbp for the genic piRNAs). The intergenic piRNA loci have little coding potential: the median length of the longest open reading frame in their transcripts was 390 nt, compared to 1,311 nt for the transcripts of the genic piRNA loci. The 100 intergenic piRNA loci account for 92% of piRNAs in the adult mouse testis. Among the intergenic piRNA loci, 18 overlap >80% with annotated non-coding transcripts (ncRNAs); the other 83 loci lie in regions devoid of annotated genes.

For the intergenic piRNA loci, the density of RNA-Seq reads correlated well with the piRNA density of the same locus ($\rho = 0.58$, $p = 3.5 \times 10^{-12}$; Figure S2A). In the accompanying manuscript (Li, et al.), we report that the transcription factor A-MYB initiates transcription of 97 of the 100 intergenic piRNA loci. The *A-Myb* mutant *Mybl1*^{repro9} disrupts transcription of these intergenic piRNA loci. Reinforcing the view that the transcripts defined here are the bona fide precursors of piRNAs in the post-natal mouse testis, production of piRNAs from the 97 loci is lost in the *A-Myb* mutant.

Several previous studies have defined piRNA clusters based solely on small RNA sequencing data (Girard et al., 2006; Lau et al., 2006; Aravin et al., 2007). Our approach differs from earlier studies in (1) using RNA-Seq data, whose greater read length facilitates the identification of introns, allowing us to define the architecture of piRNA precursor transcripts, and (2) using CAGE, PAS-Seq, and H3K4me3 ChIP-Seq data to refine the 5′ and 3′ ends of the piRNA transcripts. Consequently, the piRNA loci presented here account for more piRNAs using fewer genomic base pairs than those previously defined (Figure

S2B; Lau et al., 2006; Girard et al., 2006). Our piRNA-producing loci include 41 piRNA loci that escaped previous detection (Girard et al., 2006; Lau et al., 2006; Aravin et al., 2007), 37 of which contain introns.

piRNA Transcripts are Canonical RNA Pol II Transcripts bearing 5′ Caps and 3′ Poly(A) Tails

The presence of 5′ caps and poly(A) tails (Figure 2) suggests that piRNA transcripts are produced by RNA polymerase II. Indeed, piRNA transcripts bear a consensus AAUAAA element or its variants (Hu et al., 2005) upstream of their polyadenylation sites, just like mRNAs (Table S2). Furthermore, histone H3K4me3, which marks RNA pol II transcription start sites (Guenther et al., 2007), is bound to the genomic DNA immediately upstream of the transcription start site of each piRNA locus (Figure 2). Finally, ChIP-Seq using anti-RNA pol II antibodies showed polymerase occupancy across the entire gene body (Figure 2); little enrichment was seen for RNA pol III (Figure S2C; Kutter et al., 2011). We conclude that, like mRNAs and ncRNAs, piRNA transcripts are conventional RNA pol II transcripts bearing N(5′)ppp(5′)N 5' caps and 3′ poly(A) tails.

Many piRNAs Derive from Spliced Transcripts

Of the 217 intergenic piRNA transcripts, ~59% (127) contained introns, compared to ~98% (243 of 249) for genic piRNA transcripts (Figures 3A). Conventional RT-PCR analysis of intergenic piRNA transcripts and the corresponding genomic DNA confirmed that introns were removed from the primary transcripts (data not shown). Of the piRNA-producing genes with introns, 98% of the introns (1,025 of 1,046 for genic loci; 302 out 313 for intergenic loci) contained canonical GT-AG splice sites; similarly 98% of introns in both protein-coding and noncoding genes expressed in mouse testis (125,465 of 127,956 mRNA loci; 1,649 of 1,769 ncRNA loci) contained canonical GT-AG splice sites (Table S1).

piRNAs are proposed to derive from spliced transcripts (Beyret et al., 2012). Our data indicate that for both intergenic and genic piRNA loci, piRNAs were disproportionately produced after intron removal: 99.9% of piRNAs mapped to exons, compared to 0.112% mapping to introns. Correcting for the greater total length of exons in piRNA-producing primary transcripts, exon-mapping piRNAs were enriched by 1,668-fold compared to intron-mapping piRNAs. We detected piRNAs that failed to map to the genome, but instead mapped to the exon-exon junctions defined by our RNA-Seq-based transcript assembly, further supporting the idea that piRNAs are produced after intron removal and exon-exon joining (Figure 3B). For the 97 intergenic piRNA loci whose transcription requires A-MYB, exon-exon junction mapping piRNAs were depleted in *A-Myb* mutant testes (Figure 3B). Furthermore, the density of piRNAs falls off sharply after the 3′ end of the transcript, i.e., the site of polyadenylation (Figure 3C). The median piRNA density for the exons of genic piRNA precursor transcripts was 4.20 rpkm, but 0.00 rpkm for the 100 nt after the polyadenylation site; the median piRNA density for the exons of intergenic piRNA precursor transcripts was 200 rpkm, but 1.25 rpkm for the 100 nt after the polyadenylation site. We conclude that piRNAs produced after their precursor transcripts are fully processed—capped, spliced and polyadenylated.

Intergenic piRNA Transcripts Have Long Exons and Few Introns

To compare the architecture of piRNA transcripts with other RNA pol II transcripts, we identified the mRNAs and ncRNAs expressed during post-natal spermatogenesis. Our RNA-Seq data define 24,445 transcripts corresponding to RefSeq protein-coding genes, including 17,410 annotated and 7,035 new isoforms.

The data also defined 856 transcripts corresponding to RefSeq ncRNA loci, including 724 known and 132 new isoforms.

To begin to understand how piRNA-producing loci differ from other types of genes, we compared intergenic and genic piRNA loci with genes that do not produce piRNAs. On average, piRNA-producing primary and mature transcripts were longer than typical ncRNAs and mRNAs. Intergenic piRNA transcripts (median length: primary transcripts, 18,009 nt; mature transcripts, 8,699 nt) were longer than ncRNAs (median length: primary transcripts 7,305 nt, $p = 8.8 \times 10^{-15}$; mature transcripts 1,192 nt, $p < 2.2 \times 10^{-16}$; Figure 3A). Similarly, genic piRNA transcripts (median length: primary transcripts, 31,303 nt, mature transcripts, 5,275 nt) were longer than mRNA transcripts (median length: primary transcripts, 23,808 nt, $p = 0.00016$; mature transcripts, 2,715 nt, $p < 2.2 \times 10^{-16}$).

Intergenic piRNA transcripts have long exons, partly because they have few introns. Their exons spanned a broad length range, with 142 exons (24%) longer than 5 kb (Figure 3A). Only 0.56% of ncRNA and 3.3% mRNA exons were >5 kb. In contrast to the broad size distribution of intergenic piRNA exons, the exons of genic piRNA transcripts, mRNAs, and ncRNAs, all showed narrow length distributions. For mRNAs expressed in the mouse testis, the median exon length, 134 bp, was similar to that for vertebrate mRNAs generally (Gudlaugsdottir et al., 2007). Intron size was similar for the four types of transcripts, but intergenic piRNA transcripts had fewer introns. The median number of introns was one for intergenic piRNA transcripts, compared to two for ncRNAs, seven for genic piRNA transcripts, and seven for mRNAs. Forty-one percent (90 of 217) of intergenic piRNA transcripts lacked introns, compared to 21% of ncRNAs ($p = 2.8 \times 10^{-9}$) and just 3.2 % of mRNAs ($p < 2.2 \times 10^{-16}$; Figure 3A).

A Distinct Promoter Architecture for Intergenic piRNA Loci

The promoters of intergenic piRNA-producing loci differ from those of other types of genes transcribed by RNA pol II. Two types of promoters have been described for protein-coding genes: high CpG and low CpG promoters (Saxonov et al., 2006; Lenhard et al., 2012). We compared the promoters of four types of loci expressed in the mouse testis: 106 promoters of intergenic piRNA loci; 151 promoters of genic piRNA loci; 12,908 promoters of protein-coding genes; and 444 promoters of ncRNAs. Both mRNAs and ncRNAs contain low and high CpG promoters, as reported for human promoters expressed in diverse tissues and cell types (Saxonov et al., 2006). Uniquely among these four types of RNA pol II transcribed genes, intergenic piRNA loci predominantly had low CpG promoters: the median normalized CpG content (observed CpG/expected CpG) was 0.24, and 86 of 106 (81%) promoters were low CpG (normalized CpG content ≤ 0.4). In contrast, the median normalized CpG content was 0.70 for the promoters of genic piRNA loci, and 134 of the 151 (89%) promoters were high CpG.

Transcription of most intergenic piRNA loci is induced at the pachytene stage of spermatogenesis by the transcription factor A-MYB, whereas genic piRNAs are first expressed before the pachytene stage and are not regulated by A-MYB (Li et al., accompanying manuscript). Is low CpG content generally associated with A-MYB–regulated transcripts? We examined the promoters of A-MYB–regulated mRNAs, whose transcript levels decreased significantly by the mid-pachytene stage in an *A-Myb* mutant compared to their heterozygous littermates (Bolcun-Filas et al., 2011; Li et al., accompany manuscript). Although A-MYB–regulated mRNAs have a higher percentage of low CpG promoters (24%) compared to mRNAs generally (13%), the majority of the A-MYB–

regulated mRNAs are transcribed from high CpG promoters (Figure 4A). Thus, the low CpG density of intergenic piRNA promoters is not simply a property of their regulation by A-MYB.

Low CpG promoters typically contain a TATA box ~30 bp upstream of their transcription start sites and are associated with tissue-specific expression (Smale and Kadonaga, 2003), whereas high CpG promoters usually lack an identifiable TATA sequence and tend to be ubiquitously expressed. Nonetheless, the promoters of many genes expressed in the testis lack the TATA box (Hofmann et al., 2008) probably due to the use of alternative transcriptional initiation machinery in the testis (Kimmins et al., 2004). Our data indicate that 6.0% of high CpG promoters of genic piRNA loci contain a TATA box, significantly ($p = 0.00039$) higher than for protein-coding genes, 1.6% and 2.7% for high CpG and low CpG promoters respectively. Moreover, 9.3% of low CpG intergenic piRNA promoters contain a TATA box, also significant higher than for protein-coding genes ($p = 0.0031$). Since transcription of most testis-expressed genes is TATA-independent, the high frequency of TATA sequences at piRNA loci raises the possibility that these promoters require a TATA box to be active in another tissue.

Fifteen pairs of intergenic piRNA loci are divergently transcribed, with transcripts from both loci processed into piRNAs. The promoters of six other intergenic piRNA loci are shared with divergently transcribed mRNAs. For genic piRNA loci, 13 promoters are divergently transcribed, paired with 10 mRNA loci that do not produce piRNAs and 3 ncRNA loci. Thus, the frequency of divergent transcription for intergenic piRNA loci (34%, 36 of 106) is higher than that for genic piRNA loci (8.6%, 13 of 151, $p = 8.1 \times 10^{-7}$) and mRNA promoters (mRNA←→mRNA, 13.6%, 1,631 of 12,908; mRNA←→ncRNA, 0.98%, 127 of 12,908; $p = 3.4 \times 10^{-9}$; Figure 4B).

For mRNA loci, divergently transcribed promoters are generally high CpG (Engstrom et al., 2006; Yang et al., 2007; Yang et al., 2008). For promoters producing divergently transcribed mRNAs (mRNA←→mRNA) 98% (1,604 of 1,631) were high CpG, while only 1.6% were low CpG, a significant difference ($p < 2.2 \times 10^{-16}$). For divergently transcribed promoters producing an mRNA and a ncRNA, 93% (118 of 127) were high CpG. In contrast, nearly all (14 of 15) intergenic piRNA genes with promoters driving transcription in both directions were low CpG. Intriguingly, all of the 18 bidirectional promoters that pair a piRNA locus with a gene producing an mRNA (15) or a ncRNA (3) were high CpG. Thus, divergently transcribed, low CpG promoters seem to be a unique feature associated with producing a pair of piRNA precursor transcripts.

Alternative Transcription of piRNA Loci

Approximately 70% of piRNA-producing loci generated multiple isoforms (82 of 114 genic, 70 of 100 intergenic). Alternative splicing, alternative transcriptional initiation sites, and alternative polyadenylation all contributed to the observed diversity (Figure 4C). Genic piRNA loci had a significantly higher percentage of alternative promoters than intergenic piRNA loci (29% vs. 5.9%; $p = 2.4 \times 10^{-5}$) or protein-coding genes (17%; $p = 0.0015$). Among the genic piRNA loci with alternative promoters, 33% contain both high CpG promoters and low CpG promoters, compared to 16% for mRNA loci with alternative promoters ($p = 0.00032$), suggesting that these piRNA loci can be transcribed from different promoters in different cell types, times of development, or environmental conditions. Indeed, two isoforms of *pi-Wdfy3* are transcribed from different promoters in the testis (Figure S2D). The short isoform is transcribed from a low CpG promoter bound by A-MYB, while the long isoform is transcribed from a high CpG with no detectable A-MYB binding. piRNAs derive from the genomic region corresponding to the short transcript. In an *A-Myb* mutant, piRNA from the *pi-Wdfy3* loci are reduced 165-fold (31.6 rpkm in heterozygotes versus 0.191 in mutants). The abundance of long isoform transcripts did not decrease (2.08 rpkm in heterozygotes versus 3.86 rpkm in mutant), while the small isoforms became undetectable (10.0 rpkm versus 0 rpkm), consistent with the idea that the short, but not the long, *Wdfy3* transcript is a piRNA precursor RNA. Little is known about the regulation of genic piRNAs, because only seven genic piRNA loci are A-MYB-dependent (Li, et al., accompanying manuscript). We do not yet know if such differential regulation of transcript isoforms occurs more generally for genic piRNA loci.

Many piRNA Precursor Transcripts Contain Transposon Fragments

In fruit flies and zebrafish (Brennecke et al., 2007; Houwing et al., 2007), piRNAs derive mainly from transposon sequences. The RepeatMasker algorithm annotates 40% of the mouse genome as transposons: DNA transposons, SINEs, LINEs, and long terminal repeat (LTR) retrotransposons. In the adult mouse testis, 19% of all piRNAs mapped to transposons. The majority (91.3%) of these piRNAs map to only one location in the mouse genome, suggesting that they derive from ancient transposons that have diverged from other genomic copies of the same transposon family.

Although only 19% of mouse piRNAs map to transposon sequences, it remains possible that transposon content still distinguishes piRNA precursor transcripts from mRNAs and ncRNAs. To begin to answer this question, we computed the distance from the start site of each transcript to the nearest transposon in the genome. We found that intergenic piRNA precursor transcripts are significantly closer to LINE and LTR retrotransposons than are genic piRNA genes, mRNAs, or ncRNAs (Figure 5A, $p \leqslant 2.8 \times 10^{-6}$ for LINE; $p \leqslant 2.9 \times 10^{-5}$ for LTR).

We next computed the percentages of exonic and intronic nucleotides that are annotated as part of a transposon for each locus. Among the four types of genes transcribed by RNA pol II in the mouse testis, the median percentage of exonic nucleotides of intergenic piRNA precursor transcripts that corresponded to sense transposon sequences was 8.2% and to antisense transposon sequences was 9.4%, while genic piRNA genes, protein-coding genes, and non-coding genes contained much lower percentages of exonic nucleotides corresponding to transposon sequences (median = 0%; $p < 2.2 \times 10^{-16}$; Figure 5B). Among the 217 intergenic piRNA precursor transcripts, 72 contain DNA transposons in their exonic regions (sense or antisense), 164 contain SINEs, 148 contain LINEs, and 170 contain LTR retrotransposons. Indeed, for the transposon-mapping piRNAs that map uniquely to the mouse genome (17.3% of all adult testis piRNAs), 93.4% are in intergenic piRNA genes, suggesting that intergenic piRNA genes harbor the fossils of ancient transposons. The intronic regions of all four types of transcripts contain transposons, with 8.2–11% of intronic nucleotides sense to transposons and 14–19% of intronic nucleotides antisense to transposons, i.e., the intronic sequences were more often antisense to transposons (Figure 5B).

To test the origins of the transposon-mapping piRNAs, we focused on the youngest transposons in the mouse genome, defined as the genomic regions that the BLAST algorithm matched to the consensus sequences of the 272 transposon families annotated in Repbase (Buisine et al., 2008). We analyzed the temporal expression pattern of the piRNAs that map to these transposons, across development in wild-type and *A-Myb* mutant testes (Figure S3). piRNAs from 20 transposon families were first expressed at the pachytene stage of meiosis and were lost in the *A-Myb* mutant (Figure 5C and 5D). Fragments from these 20 transposon families were present in the exons of intergenic piRNA loci (61 sites), but not in genic piRNA loci. In contrast, *A-Myb*–independent, transposon-mapping piRNAs were detected in the exons of both genic and intergenic piRNA loci. We conclude that intergenic piRNA loci are the source of the piRNAs that map to at least 20 transposon families.

During the pre-pachytene stage of spermatogenesis, piRNAs are amplified by a cycle of PIWI-protein–catalyzed cleavage called Ping-Pong amplification (Brennecke et al., 2007; Gunawardane et al., 2007). Each Pong-Pong cycle is thought to involve two distinct transcripts: a primary piRNA transcript and a transcript antisense to the primary piRNA transcript, possibly from a disperse transposon copy. While we can detect significant Ping-Pong amplification for transposon-mapping piRNAs in the testis at the pre-pachytene stage (Z score = 29), we did not detect significant Ping-Pong in the adult testis (Z score = 0.81; a Z score < 1.96 corresponds to $p > 0.05$), consistent with the idea that most transposon-mapping piRNAs in the adult testis derive from precursor RNAs transcribed from the intergenic piRNA loci.

piRNA Loci are Transcriptionally Silenced in Embryonic Stem Cells

In mice, piRNAs can be readily detected in the ovary (Ro et al., 2007) and the testis, but have not been found outside the germline. Among publicly available small RNA libraries from the testis and 36 somatic tissues and cell types (Kuchen et al., 2010), only for the testis did we reliably detect small RNAs (>23 nt) mapping to the 214 piRNA-producing loci defined here (Figure 6A) Most piRNAs begin with uracil (68% in the adult testis sample). However, in none of the 36 somatic libraries was the first nucleotide of the >23 nt RNAs more likely to begin with uracil than expected by chance (Figure 6A), consistent with the idea that piRNAs are restricted to the germline.

We also examined embryonic stem (ES) cells for piRNA production, because mRNA encoding the PIWI protein MILI is readily detectable in these cells (Figure S5; Wang et al., 2011). We were unable to detect small RNAs >23 nt in mouse ES cells (data not shown). Thus, despite the presence of at least one PIWI protein, mature piRNAs are either very rare or altogether absent in ES cells.

Indeed, our analysis of publicly available RNA-Seq data from ES cells (Ma et al., 2011) showed no accumulation of transcripts from intergenic piRNA loci (Figure 6B) compared to testis. In the accompanying manuscript, we showed that at pachytene stage of spermatogenesis, A-MYB initiates transcription of most intergenic piRNA loci (Li et al., accompanying manuscript). The absence of A-MYB in ES cells helps explain why transcripts from intergenic piRNA loci are not detected in these cells. In ES cells, RNA pol II occupancy (Young et al., 2011) across the whole intergenic piRNA loci was significantly lower than that in the testis ($p < 2.2 \times 10^{-16}$; Figure 6B), indicating that intergenic piRNA loci are transcriptionally silenced.

Although RNA pol II occupancy was low across the whole intergenic piRNA loci, RNA pol II occupancy at their transcriptional start sites was significantly greater than at the transcriptional start sites of mRNAs not expressed in ES cells ($p = 6.1 \times 10^{-8}$, Figure 6C). That is, intergenic piRNA loci appear not to be expressed in ES cells, yet RNA pol II can be readily detected at their promoters. More RNA pol II at a gene’s promoter compared to its transcribed region is the hallmark of polymerase pausing (Maderious and Chen-Kiang, 1984; Spencer and Groudine, 1990; Adelman and Lis, 2012). Paused polymerases occur when an early elongation complex appears to be transiently halted prior to productive RNA synthesis (Gilchrist et al., 2010), i.e., the rate of transcriptional initiation is much faster than the rate of elongation. In metazoans, promoter-proximal pausing is a widespread mechanism of transcriptional control (Min et al., 2011).

To test whether RNA pol II detected on the promoters of the intergenic piRNA loci reflects paused polymerase, we analyzed publicly available high-throughput nuclear global run-on data from ES cells (GRO-Seq; Min et al., 2011). GRO-Seq detects nascent transcripts from RNA pol II. Indeed, we found nascent RNA transcripts at intergenic piRNA loci, and 82 of the 179 intergenic piRNA transcripts for which RNA pol II was detected had a pausing index >2 (pausing index is the ratio of transcript abundance surrounding the transcription start site to that for the entire gene; Figure 6D). Our data suggest that RNA pol II pausing helps prevent transcription of intergenic piRNA loci in ES cells.

Intergenic piRNA Loci are Transcriptionally Silenced in Somatic Tissues

ES cells have an unusual open chromatin structure, and more genomic regions are transcribed in ES cells compared to differentiated cells. As ES cells differentiate, fewer regions are transcribed (Efroni et al., 2008). We asked whether the transcriptional silencing mechanism we observed for intergenic piRNA loci in ES cells might also function in somatic cells that lack piRNAs. We analyzed ChIP-Seq and RNA-Seq data from mouse ENCODE (Shen et al., 2012). RNA-Seq and pol II ChIP-Seq data indicate that intergenic piRNA loci are transcriptionally silenced in six tissues for which the most complete set of ENCODE data are publicly available (Figure 7, data from another 15 tissues are shown in Figure S5). For these somatic tissues, RNA pol II was significantly more enriched at the promoter regions of intergenic piRNA loci compared to inactive genes ($2.2 \times 10^{-16} \le p$ value $\le 1.7 \times 10^{-7}$), suggesting polymerase pausing may act to further transcriptionally silence intergenic piRNA loci.

In other tissues where intergenic piRNA loci are transcriptionally silenced, we did not detect enrichment of trimethylated lysine 27 on histone H3 (H3K27me3, a repressive modification associated with Polycomb complexes (Figure S5; Barski et al., 2007). Considering that Polycomb-mediated H3K27me3 effectively silences high CpG promoter (Deaton and Bird, 2011), the intergenic piRNA loci are probably silenced by a Polycomb-independent repression mechanism in somatic tissues.

We also examined other histone modifications associated with active transcription (Figure 7 and S5), including H3K4me3 (a mark of actively transcribed promoters ; Guenther et al., 2007), trimethylated lysine 36 on histone H3 (H3K36me3, a mark of actively transcribed gene bodies; Guenther et al., 2007), mono-methylated lysine 4 on histone H3 (H3K4me1, a mark of regulatory elements associated with enhancers and enriched in the promoter regions as well; Koch et al., 2007), and acetylated lysine 27 on histone H3 (H3K27ac, a mark associated with active regulatory regions and which may distinguish active enhancers and promoters from their inactive counterparts; Heintzman et al., 2009). The expression of intergenic piRNA loci follow the patterns of active histone marks predicted by their testis-specific expression: active histone marks were high in the testis and low in somatic cells (Figure 7 and S5). In contrast, active histone marks were high across all cell types for housekeeping genes. The expression of intergenic piRNA loci therefore follows the canonical patterns of epigenetic modification found for protein-coding genes with highly restricted expression patterns (Figure 7 and S5).

Finally, we analyzed the mouse ENCODE DNase I hypersensitivity data for 23 somatic organs and cell types, including ES cells (Figure S5; Stamatoyannopoulos et al., 2012). The promoters of intergenic piRNA loci were less accessible to DNase I than the promoters of housekeeping genes, yet significantly more accessible than the promoters of inactive genes ($2.3 \times 10^{-26} < p < 1.5 \times 10^{-8}$). The pattern of chromatin accessibility mostly mirrored RNA pol II occupancy, suggesting that the largely inaccessible chromatin at these loci causes decreased RNA pol II binding in somatic cells.

DISCUSSION

piRNAs are the most recently discovered family of small silencing RNAs; how they are made and function remain unknown. Despite the strong consensus that piRNAs in mice are processed from long precursor transcripts originating from large regions of the genome, no piRNA loci or their primary transcripts had heretofore been precisely defined in any animal. We used RNA-Seq, CAGE, PAS, and ChIP-Seq of RNA pol II and H3K4me3 experiments to define the piRNA-producing loci and their transcripts in the post-partum mouse testis. Our analysis identifies 466 piRNA precursor transcripts produced from 214 loci. Among them, 114 loci (249 transcripts) overlap with protein-coding genes, while the rest lie far from any known genes. In total, the 214 loci constitute just 0.33% of the mouse genome, yet account for 95% of piRNAs in adult mouse testis. This definite set of piRNA loci and primary piRNA transcripts should provide an invaluable resource to study piRNA biogenesis and function.

The primary transcripts of the piRNA loci we define possess all of the features of standard RNA pol II transcripts—they have 5′ caps and 3′ poly(A) tails, and their promoters are occupied by RNA pol II and enriched in H3K4me3. Many primary piRNA transcripts contain introns, which are removed prior to piRNA production. Whereas few piRNAs map to introns, many map to spliced exon-exon junctions. For those piRNA loci first transcribed at the pachytene stage of spermatogenesis, production of these junction piRNAs requires the transcription factor A-MYB, confirming that those piRNAs are produced from processed transcripts.

Despite sharing many features with protein-coding genes, intergenic piRNA loci stand out as a unique class of transcriptional units. Almost half contain no introns. Intriguingly, they possess a distinct promoter architecture. The promoters of ubiquitously expressed protein-coding genes tend to be high CpG, while the promoters of tissue-specific protein-coding genes tend to be low CpG. Promoters that drive divergently transcribed genes tend to be high CpG, and these genes tend to be ubiquitously expressed. The promoters of intergenic piRNA loci violate these seemingly self-consistent patterns: although they are transcribed exclusively in the germline, more than one-third of them are regulated by bidirectional promoters, three times higher than protein-coding genes, yet almost all of these bidirectional promoters are low CpG. The unique promoter architecture of intergenic piRNA loci raises questions about how they are transcriptionally regulated.

We analyzed RNA-Seq and ChIP-Seq data in a variety of tissue types to begin to understand why intergenic piRNA loci are not expressed in somatic cells. In the testis, these loci are strongly bound by RNA pol II, enriched in histone modifications known to be associated with genes actively transcribed by RNA pol II: H3K4me3, H3K4me1, H3K27ac at the promoter and H3K36me3 in the gene body. The loci were depleted of the repressive histone modification trimethylated lysine 27 on histone H3 (H3K27me3). In somatic tissues, these loci show markedly decreased RNA pol II occupancy as well as all of these active histone marks, but show no detectable increase in the repressive mark H3K27me3. Thus, a lack of transcriptional initiation is the main cause of the inactivity of the piRNA-producing loci in the soma. However, RNA pol II occupancy of intergenic piRNA loci is still significantly higher at their promoters than their transcribed regions, suggesting that RNA pol II pausing plays a secondary role in repressing their transcription.

What causes the decrease in RNA pol II binding at intergenic piRNA loci? DNase I hypersensitivity data show that the promoters of these loci are in a largely closed state in somatic cells. Thus, intergenic piRNA loci are silenced in somatic cells by the establishment of inaccessible chromatin. Because we did not observe any change in the repressive histone modification H3K27me3, we hypothesize that this may reflect increased DNA methylation, as DNA methylation is thought to limit chromatin accessibility (Thurman et al., 2012).

EXPERIMENTAL PROCEDURES

Animals

Mice were maintained and used according to the guidelines of the Institutional Animal Care and Use Committee of the University of Massachusetts Medical School.

RNA Sequencing

Small RNA libraries were constructed and sequenced as described (Ghildiyal et al., 2008; Seitz et al., 2008) except that 18–35 nt RNA was isolated and β-elimination and 2S rRNA depletion were omitted. Sequencing (36 or 76 nt reads) was performed using either a Genome Analyzer GAII or HiSeq 2000 (50 nt; Illumina). RNA-Seq libraries (Zhao Zhang and P.D.Z., manuscript submitted) were performed using Illumina protocols. PAS-Seq library was prepared as described (Shepard et al., 2011) using total RNA from adult testes and then sequenced using the 100-nt read protocol on a HiSeq 2000. Cap-analysis of gene expression (CAGE) was as described (Yang et al., 2011), except that the library was sequenced using the 100-nt read protocol on a HiSeq 2000.

Chromatin Immunoprecipitation (ChIP)

ChIP was performed as described (Chen et al., 2008), except that testes were macerated on ice and then fixed with 1.5% (w/v) formaldehyde for 20 min. Samples were then crushed further using a Dounce homogenizer (Kimble-Chase). ChIP-Seq libraries for anti-RNA pol II (Santa Cruz Biotechnology) and H3K4me3 antibody (Abcam) and control input DNA were prepared following the Illumina ChIP-Seq protocol and were sequenced the 50-nt read protocol on a HiSeq 2000.

ACCESSION NUMBERS

All sequence data is available through the NCBI Short Read Archive (www.ncbi.nlm.nih.gov/sites/sra) using accession number SRA061530.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, 5 figures, and two tables.

ACKNOWLEDGEMENTS

We thank Cindy Tipping and Gwen Farley for technical assistance; Zhong Wang for sharing the E14 ES cells; Xi Chen for advice on ES cell culture; Xiaotu Ma, Oliver Rando and Benjamin Carone for advice on ChIP; Hao Wu, Raga Krishnakumar, and Nasun Hah for discussions; and members of our laboratories for advice, suggestions, and critical comments on the manuscript. This work was supported in part by National Institutes of Health grants HG007000 to Z.W. and GM62862 and GM65236 to P.D.Z. X.Z.L. was supported by the Lalor and Jane Coffin Childs foundations.

REFERENCES

Adelman, K., and Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet *13*, 720-731.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T., Chien, M., Russo, J. J., Ju, J., Sheridan, R., Sander, C., Zavolan, M., and Tuschl, T. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203-207.

Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744-747.

Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., and Gvozdev, V. A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. Curr Biol *11*, 1017-1027.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., Chaves, D. A., Gu, W., Vasale, J. J., Duan, S., Conte, D. J., Luo, S., Schroth, G. P., Carrington, J. C., Bartel, D. P., and Mello, C. C. (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. Mol Cell *31*, 67-78.

Beyret, E., Liu, N., and Lin, H. (2012). piRNA biogenesis during adult spermatogenesis in mice is independent of the Ping-Pong mechanism. Cell Res *22*, 1429-1439.

Bolcun-Filas, E., Bannister, L. A., Barash, A., Schimenti, K. J., Hartford, S. A., Eppig, J. J., Handel, M. A., Shen, L., and Schimenti, J. C. (2011). A-MYB (MYBL1) transcription factor is a master regulator of male meiosis. Development *138*, 3319-3330.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell *128*, 1089-1103.

Buisine, N., Quesneville, H., and Colot, V. (2008). Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics *91*, 467-475.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L., and Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106-1117.

Das, P. P., Bagijn, M. P., Goldstein, L. D., Woolford, J. R., Lehrbach, N. J., Sapetschnig, A., Buhecha, H. R., Gilchrist, M. J., Howe, K. L., Stark, R., Matthews, N., Berezikov, E., Ketting, R. F., Tavare, S., and Miska, E. A. (2008). Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. Mol Cell *31*, 79-90.

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P. N., Enright, A. J., and O'Carroll, D. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature *480*, 259-263.

Deaton, A. M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev *25*, 1010-1022.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D. J., Dash, C., Bazett-Jones, D. P., Le Grice, S., McKay, R. D., Buetow, K. H., Gingeras, T. R., Misteli, T., and Meshorer, E. (2008). Global transcription in pluripotent embryonic stem cells. Cell Stem Cell *2*, 437-447.

Engstrom, P. G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S. L., Yang, L., Kunarso, G., Ng, E. L., Batalov, S., Wahlestedt, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Wells, C., Bajic, V. B., Orlando, V., Reid, J. F., Lenhard, B., and Lipovich, L. (2006). Complex Loci in human and mouse genomes. PLoS Genet *2*, e47.

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E. L., Zapp, M. L., Weng, Z., and Zamore, P. D. (2008). Endogenous siRNAs Derived from Transposons and mRNAs in *Drosophila* Somatic Cells. Science *320*, 1077-1081.

Gilchrist, D. A., Dos Santos, G., Fargo, D. C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell *143*, 540-551.

Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199-202.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644-652.

Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006a). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev *20*, 1709-1714.

Grivna, S. T., Pyhtila, B., and Lin, H. (2006b). MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc Natl Acad Sci U S A *103*, 13415-13420.

Gu, W., Lee, H. C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D., and Mello, C. C. (2012). CapSeq and CIP-TAP map 5′ ends of Pol II transcripts and reveal capped-small RNAs as C. elegans piRNA precursors. Cell in press.

Gudlaugsdottir, S., Boswell, D. R., Wood, G. R., and Ma, J. (2007). Exon size distribution and the origin of introns. Genetica *131*, 299-306.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M. C. (2007). A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5′ End Formation *in Drosophi*la. Science *315*, 1587-1590.

Hamilton, A. J., and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. Science *286*, 950-952.

Hammond, S. M., Bernstein, E., Beach, D., and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. Nature *404*, 293-296.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Hofmann, O., Caballero, O. L., Stevenson, B. J., Chen, Y. T., Cohen, T., Chua, R., Maher, C. A., Panji, S., Schaefer, U., Kruger, A., Lehvaslaiho, M., Carninci, P., Hayashizaki, Y., Jongeneel, C. V., Simpson, A. J., Old, L. J., and Hide, W. (2008). Genome-wide analysis of cancer/testis gene expression. Proc Natl Acad Sci U S A *105*, 20422-20427.

Horwich, M. D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., and Zamore, P. D. (2007). The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. Curr Biol *17*, 1265-1272.

Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D. V., Blaser, H., Raz, E., Moens, C. B., Plasterk, R. H., Hannon, G. J., Draper, B. W., and Ketting, R. F. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell *129*, 69-82.

Hu, J., Lutz, C. S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485-1493.

Kamminga, L. M., Luteijn, M. J., den Broeder, M. J., Redl, S., Kaaij, L. J., Roovers, E. F., Ladurner, P., Berezikov, E., and Ketting, R. F. (2010). Hen1 is required for oocyte development and piRNA stability in zebrafish. EMBO J *29*, 3688-3700.

Kawaoka, S., Izumi, N., Katsuma, S., and Tomari, Y. (2011). 3′ end formation of PIWI-interacting RNAs in vitro. Mol Cell *43*, 1015-1022.

Kawaoka, S., Mitsutake, H., Kiuchi, T., Kobayashi, M., Yoshikawa, M., Suzuki, Y., Sugano, S., Shimada, T., Kobayashi, J., Tomari, Y., and Katsuma, S. (2012). A role for transcription from a piRNA cluster in de novo piRNA production. RNA *18*, 265-273.

Kimmins, S., Kotaja, N., Davidson, I., and Sassone-Corsi, P. (2004). Testis-specific transcription mechanisms promoting male germ-cell differentiation. Reproduction *128*, 5-12.

Klattenhoff, C., Xi, H., Li, C., Lee, S., Xu, J., Khurana, J. S., Zhang, F., Schultz, N., Koppetsch, B. S., Nowosielska, A., Seitz, H., Zamore, P. D., Weng, Z., and Theurkauf, W. E. (2009). The *Drosophila* HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. Cell *138*, 1137-1149.

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaoz, U., Clelland, G. K., Wilcox, S., Beare, D. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Weng, Z., Birney, E., Carter, N. P., Vetrie, D., and Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res *17*, 691-707.

Kuchen, S., Resch, W., Yamane, A., Kuo, N., Li, Z., Chakraborty, T., Wei, L., Laurence, A., Yasuda, T., Peng, S., Hu-Li, J., Lu, K., Dubois, W., Kitamura, Y., Charles, N., Sun, H. W., Muljo, S., Schwartzberg, P. L., Paul, W. E., O'Shea, J., Rajewsky, K., and Casellas, R. (2010). Regulation of MicroRNA Expression and Abundance during Lymphopoiesis. Immunity *32*, 828-839.

Kutter, C., Brown, G. D., Goncalves, A., Wilson, M. D., Watt, S., Brazma, A., White, R. J., and Odom, D. T. (2011). Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nat Genet *43*, 948-955.

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., and Kingston, R. E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363-367.

Lee, H. C., Gu, W., Shirayama, M., Youngman, E., Conte, D. J., and Mello, C. C. (2012). *C. elegans* piRNAs Mediate the Genome-wide Surveillance of Germline Transcripts. Cell *150*, 78-87.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet *13*, 233-245.

Ma, Z., Swigut, T., Valouev, A., Rada-Iglesias, A., and Wysocka, J. (2011). Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. Nat Struct Mol Biol *18*, 120-127.

Maderious, A., and Chen-Kiang, S. (1984). Pausing and premature termination of human RNA polymerase II during transcription of adenovirus in vivo and in vitro. Proc Natl Acad Sci U S A *81*, 5931-5935.

Min, I. M., Waterfall, J. J., Core, L. J., Munroe, R. J., Schimenti, J., and Lis, J. T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev *25*, 742-754.

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res *40*, D130-5.

Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., Tuschl, T., and Kandel, E. R. (2012). A Role for Neuronal piRNAs in the Epigenetic Control of Memory-Related Synaptic Plasticity. Cell *149*, 693-707.

Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., Antony, C., Sachidanandam, R., and Pillai, R. S. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature *480*, 264-267.

Ro, S., Song, R., Park, C., Zheng, H., Sanders, K. M., and Yan, W. (2007). Cloning and expression profiling of small RNAs expressed in the mouse ovary. RNA *13*, 2366-2380.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell *127*, 1193-1207.

Saito, K., Sakaguchi, Y., Suzuki, T., Suzuki, T., Siomi, H., and Siomi, M. C. (2007). Pimet, the *Drosophila* homolog of HEN1, mediates 2□}□□-O-methylation of Piwi-interacting RNAs at their 3′ ends. Genes Dev *21*, 1603-1608.

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Seitz, H., Ghildiyal, M., and Zamore, P. D. (2008). Argonaute loading improves the 5' precision of both microRNAs and their miRNA* strands in flies. Curr Biol *18*, 147-151.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.

Shepard, P. J., Choi, E. A., Lu, J., Flanagan, L. A., Hertel, K. J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA *17*, 761-772.

Shibata, N., Rouhana, L., and Agata, K. (2010). Cellular and molecular dissection of pluripotent adult somatic stem cells in planarians. Dev Growth Differ *52*, 27-41.

Shirayama, M., Seth, M., Lee, H. C., Gu, W., Ishidate, T., Conte, D. J., and Mello, C. C. (2012). piRNAs Initiate an Epigenetic Memory of Nonself RNA in the *C. elegans* Germline. Cell *150*, 65-77.

Smale, S. T., and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. Annu Rev Biochem *72*, 449-479.

Spencer, C. A., and Groudine, M. (1990). Transcription elongation and eukaryotic gene regulation. Oncogene *5*, 777-785.

Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., Giste, E., Johnson, A., Zhang, M., Balasundaram, G., Byron, R., Roach, V., Sabo, P. J., Sandstrom, R., Stehling, A. S., Thurman, R. E., Weissman, S. M., Cayting, P., Hariharan, M., Lian, J., Cheng, Y., Landt, S. G., Ma, Z., Wold, B. J., Dekker, J., Crawford, G. E., Keller, C. A., Wu, W., Morrissey, C., Kumar, S. A., Mishra, T., Jain, D., Byrska-Bishop, M., Blankenberg, D., Lajoie1, B. R., Jain, G., Sanyal, A., Chen, K. B., Denas, O., Taylor, J., Blobel, G. A., Weiss, M. J., Pimkin, M., Deng, W., Marinov, G. K., Williams, B. A., Fisher-Aylor, K. I., Desalvo, G., Kiralusha, A., Trout, D., Amrhein, H., Mortazavi, A., Edsall, L., McCleary, D., Kuan, S., Shen, Y., Yue, F., Ye, Z., Davis, C. A., Zaleski, C., Jha, S., Xue, C., Dobin, A., Lin, W., Fastuca, M., Wang, H., Guigo, R., Djebali, S., Lagarde, J., Ryba, T., Sasaki, T., Malladi, V. S., Cline, M. S., Kirkup, V. M., Learned, K., Rosenbloom, K. R., Kent, W. J., Feingold, E. A., Good, P. J., Pazin, M., Lowdon, R. F., and Adams, L. B. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol *13*, 418.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P. D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science *313*, 320-324.

Vourekas, A., Zheng, Q., Alexiou, P., Maragkakis, M., Kirino, Y., Gregory, B. D., and Mourelatos, Z. (2012). Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat Struct Mol Biol *19*, 773-781.

Wang, Q. E., Han, C., Milum, K., and Wani, A. A. (2011). Stem cell protein Piwil2 modulates chromatin modifications upon cisplatin treatment. Mutat Res *708*, 59-68.

Watanabe, T., Tomizawa, S., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Iida, N., Hoki, Y., Murphy, P. J., Toyoda, A., Gotoh, K., Hiura, H., Arima, T., Fujiyama, A., Sado, T., Shibata, T., Nakano, T., Lin, H., Ichiyanagi, K., Soloway, P. D., and Sasaki, H. (2011). Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. Science *332*, 848-852.

Xu, M., You, Y., Hunsicker, P., Hori, T., Small, C., Griswold, M. D., and Hecht, N. B. (2008). Mice deficient for a small cluster of Piwi-interacting RNAs implicate Piwi-interacting RNAs in transposon control. Biol Reprod *79*, 51-57.

Yang, M. Q., Koehly, L. M., and Elnitski, L. L. (2007). Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. PLoS Comput Biol *3*, e72.

Yang, M. Q., Taylor, J., and Elnitski, L. (2008). Comparative analyses of bidirectional promoters in vertebrates. BMC Bioinformatics *9 Suppl 6*, S9.

Yang, Z., Bruno, D. P., Martens, C. A., Porcella, S. F., and Moss, B. (2011). Genome-wide analysis of the 5' and 3' ends of vaccinia virus early mRNAs delineates regulatory sequences of annotated and anomalous transcripts. J Virol *85*, 5897-5909.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Res *39*, 7415-7427.

Zamore, P. D., Tuschl, T., Sharp, P. A., and Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25-33.

FIGURE LEGENDS

Figure 1. Assembly of the Mouse Testis Transcriptome

Round-corner rectangles are input or output data, rectangles are processes, and diamonds represent decision points.

See also Figure S1 and Table S1.

Figure 2. piRNA Producing Transcripts are Conventional RNA pol II Transcripts

Aggregation charts for genic and intergenic piRNA-producing transcripts and for mRNAs and ncRNAs. Oxidized small RNA sequencing was used to detect piRNAs; RNA-Seq of poly(A)+ selected RNA (poly(A)$^+$ RNA-Seq; GSE36025) and RNA-Seq of RNA depleted of rRNA (−rRNA RNA-Seq) was used to measure transcript abundance; cap analysis of gene expression (CAGE) and polyadenylation site sequencing (PAS) were used to define the ends of the transcripts. ChIP-Seq for H3K4me3 and RNA pol II were used to corroborate sites of transcriptional initiation. Dotted lines show the transcriptional start site (Start) and the site of cleavage and polyadenylation (End).

See also Table S2.

Figure 3. piRNA-Producing Transcripts are Composed of Exons and Introns, Like mRNAs

(A) Box plots showing transcript length distributions for genic and intergenic piRNA loci and for genes producing mRNAs and ncRNAs.

(B) Above, aggregation plots of genomic coordinates surrounding the 5′ and 3′ splice sites of introns contained within the most highly expressed genic (2,113 genic splice junctions) and intergenic (383 splice junctions) piRNA-producing transcripts in the adult mouse testis. Below, the signal was calculated for non-genome matching reads mapped to exon-exon junction sequences. The data are from *A-Myb* mutants and their heterozygous littermates.

(C) The density of uniquely mapping piRNAs in the exons, introns, and 100 nt after the 3′ end (i.e., the polyadenylation site, PAS) of genic and intergenic piRNA-producing transcripts.

Figure 4. The Promoter Architecture of piRNA-Producing Genes

(A) Box plots presenting the distribution of normalized CpG content for the promoters of transcript groups, as well as for *A-Myb*–regulated transcripts.

(B) Strip chart showing the normalized CpG content for the promoters of divergently transcribed loci.

(C) Venn diagrams showing the three types of alternative RNA processing events for genic and intergenic piRNA loci. White, alternative polyadenylation; black, alternative splicing of internal exons; gray, alternative transcriptional start sites.

See also Figure S2.

Figure 5. Intergenic piRNA Loci Contain Transposon Fragments

(A) Box plots showing distributions of the distance between the start site of a transcript and the nearest transposon annotated by RepeatMasker for genic piRNA precursor transcripts, mRNAs, intergenic piRNA precursor transcripts, and ncRNAs, separated by transposon type.

(B) Box plots showing distribution by transcript class of exonic and intronic region sequence correspondence to known transposon sequences.

(C) Box plot showing distribution of piRNAs mapping to transposons families over five developmental time points after birth (dpp).

(D) Box plot showing the change in piRNAs mapping to transposons at 14.5 and 17.5 dpp in *A-Myb* and *Miwi* mutant testis, compared to their heterozygous littermates.

See also Figure S3.

Figure 6. piRNAs are Not Detected outside the Testis

(A) Heat map showing the density of small RNAs >23 nt for individual transcripts for genic and intergenic piRNA-producing loci and for housekeeping genes (Kuchen et al., 2010). Representative percentages of 5′ uracil are shown; Table S1 provides percentages for all samples with sufficient small RNAs >23 nt.

(B) Box plots showing piRNA transcript abundance and RNA pol II occupancy (rpkm) of genic and intergenic loci from ES cells (Ma et al., 2011; Young et al., 2011).

(C) The RNA pol II occupancy (in rpkm) was calculated for a region beginning 250 bp before and ending 250 bp after the transcriptional start sites of intergenic piRNA loci, housekeeping genes, and inactive genes (<1 rpkm for all seven tissues shown in Figure 7, including ES cells). Panels (C) and (D) contain some value less than zero because of background subtraction.

(D) The pausing index was calculated for 179 intergenic piRNA transcripts for which RNA pol II was detected using publicly available ES cell data (Min et al., 2011). Pausing index was calculated as ratio of transcript abundance surrounding the transcriptional start site to that for the entire gene.

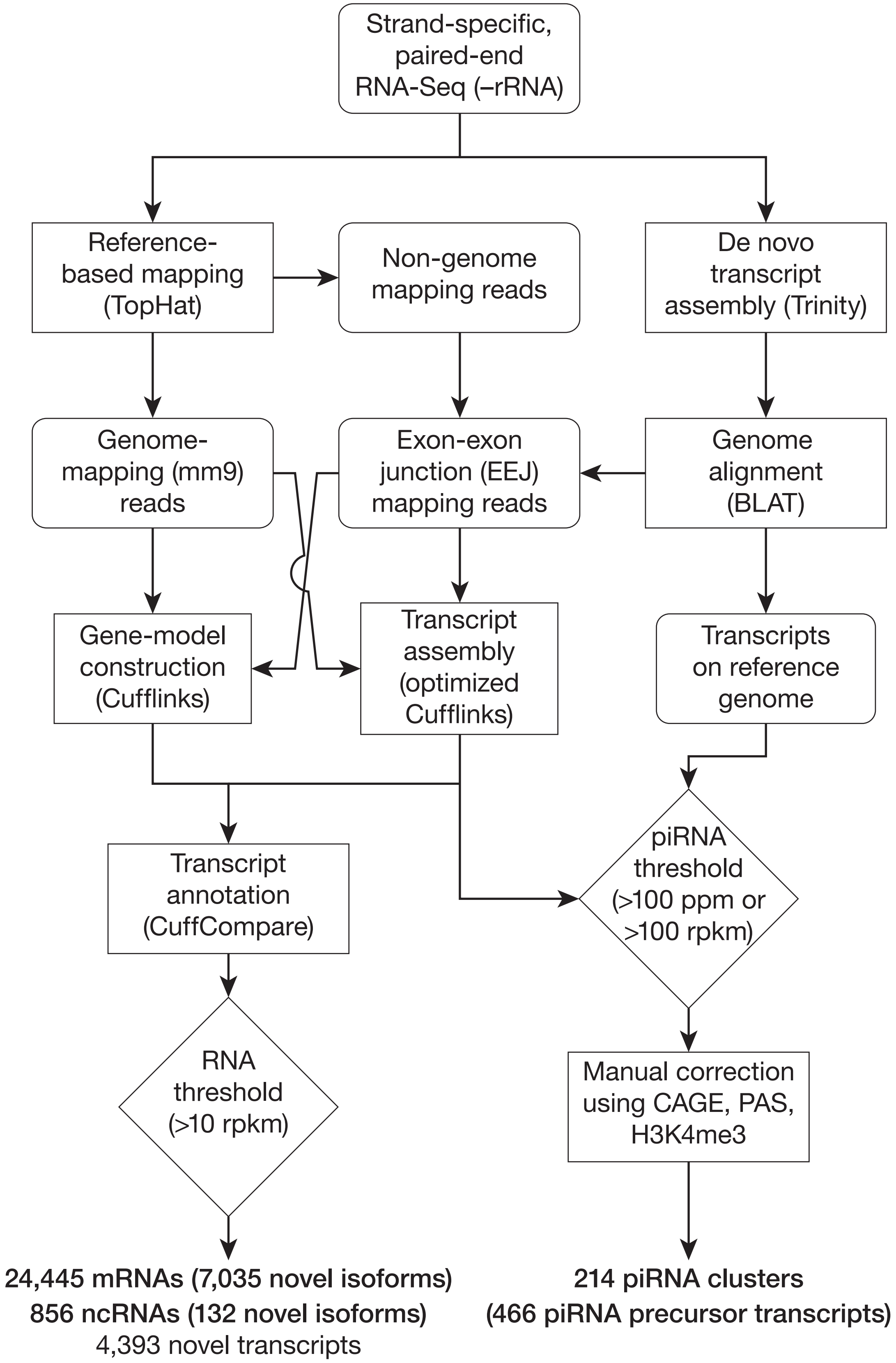
See also Figure S4.

Figure 7. Intergenic piRNA are Transcriptionally Silenced outside Testis

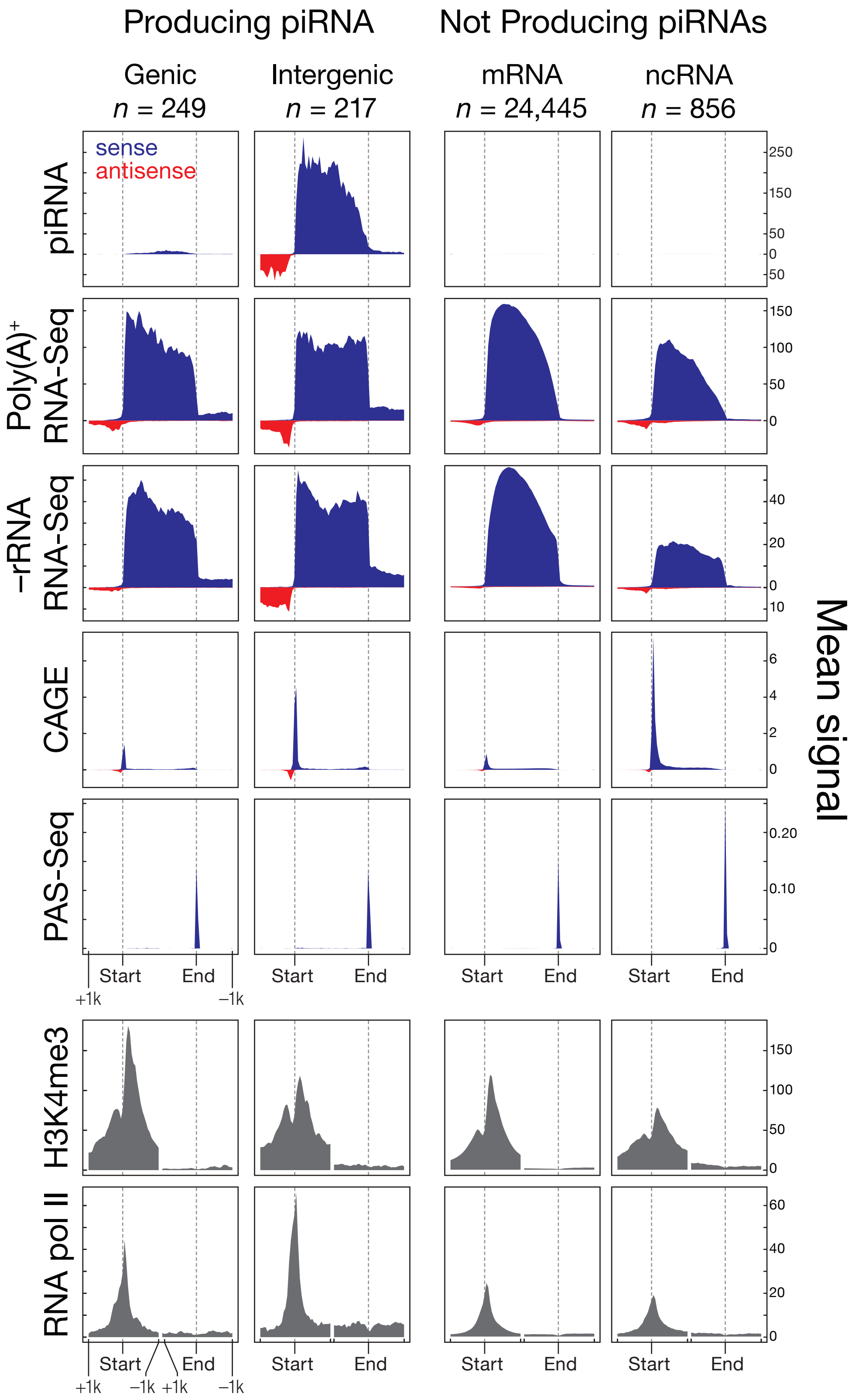
For four classes of transcripts (intergenic and genic piRNA loci, housekeeping genes, and inactive genes) were analyzed using divers, publicly available high throughput data.

See also Figure S5.

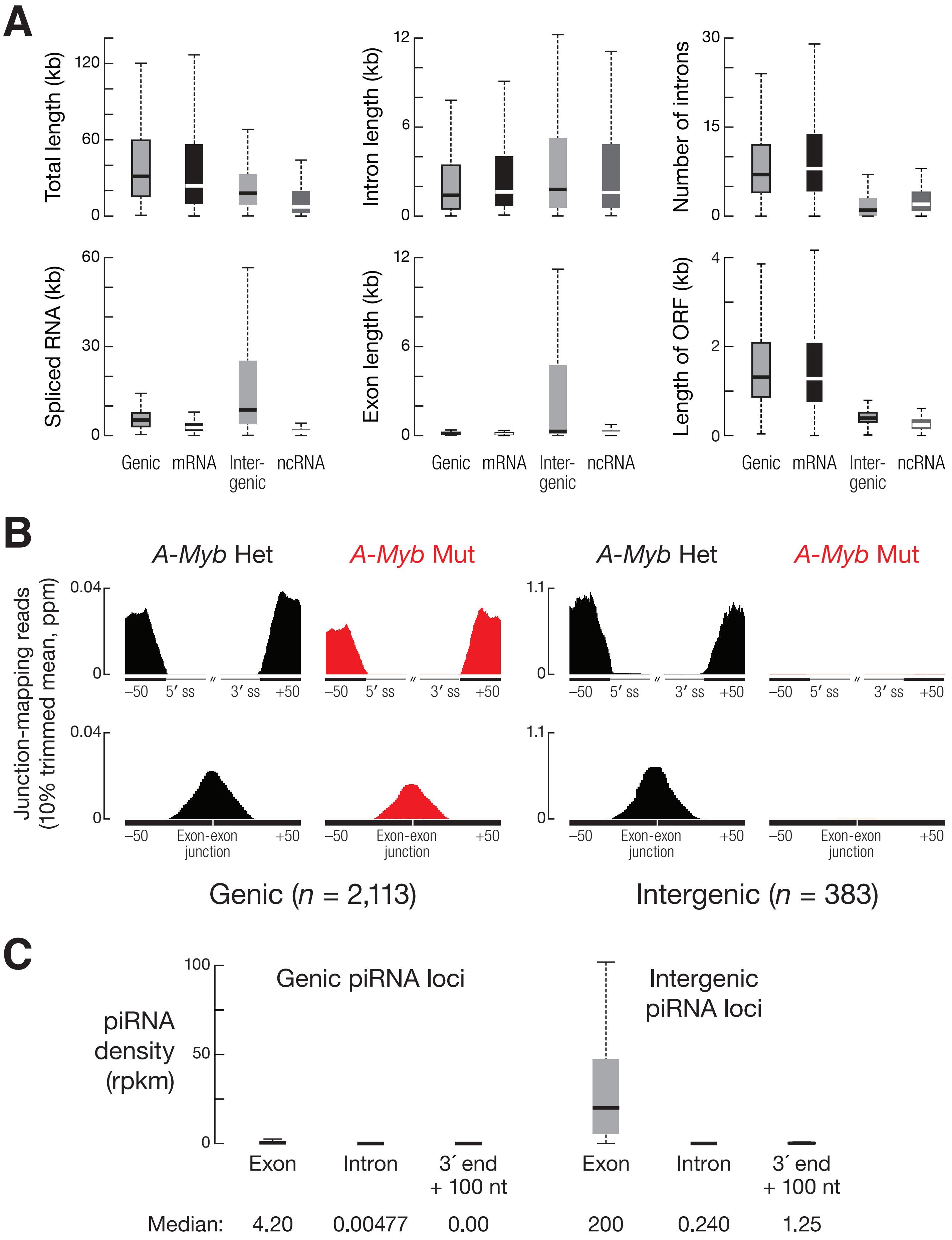
Li et al., Figure 1


Strand-specific, paired-end RNA-Seq (–rRNA)
Reference-based mapping (TopHat)
Non-genome mapping reads
De novo transcript assembly (Trinity)
Genome-mapping (mm9) reads
Exon-exon junction (EEJ) mapping reads
Genome alignment (BLAT)
Gene-model construction (Cufflinks)
Transcript assembly (optimized Cufflinks)
Transcripts on reference genome
Transcript annotation (CuffCompare)
piRNA threshold (>100 ppm or >100 rpkm)
RNA threshold (>10 rpkm)
Manual correction using CAGE, PAS, H3K4me3
24,445 mRNAs (7,035 novel isoforms)
856 ncRNAs (132 novel isoforms)
4,393 novel transcripts
214 piRNA clusters
(466 piRNA precursor transcripts)

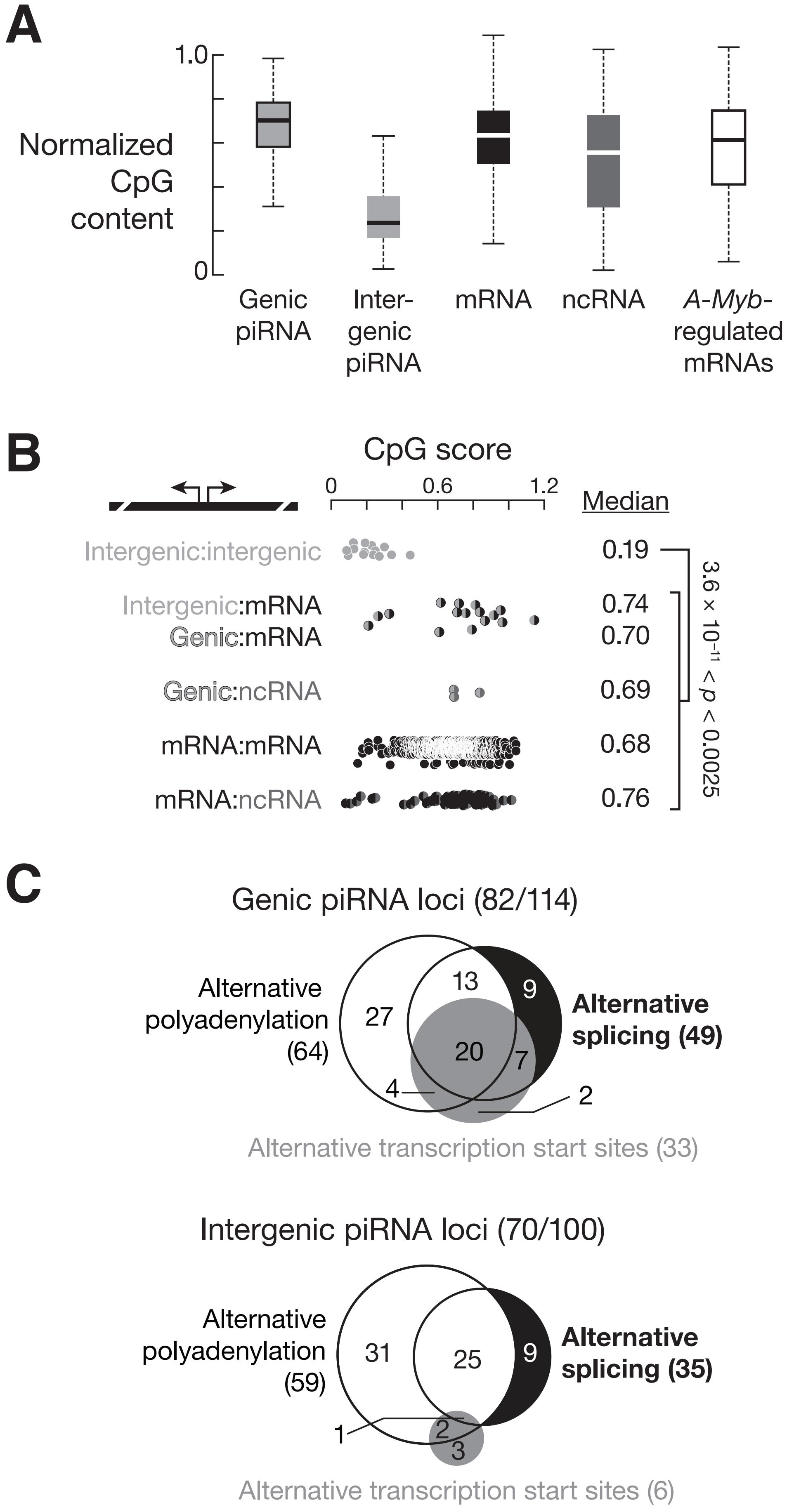
Li et al., Figure 2



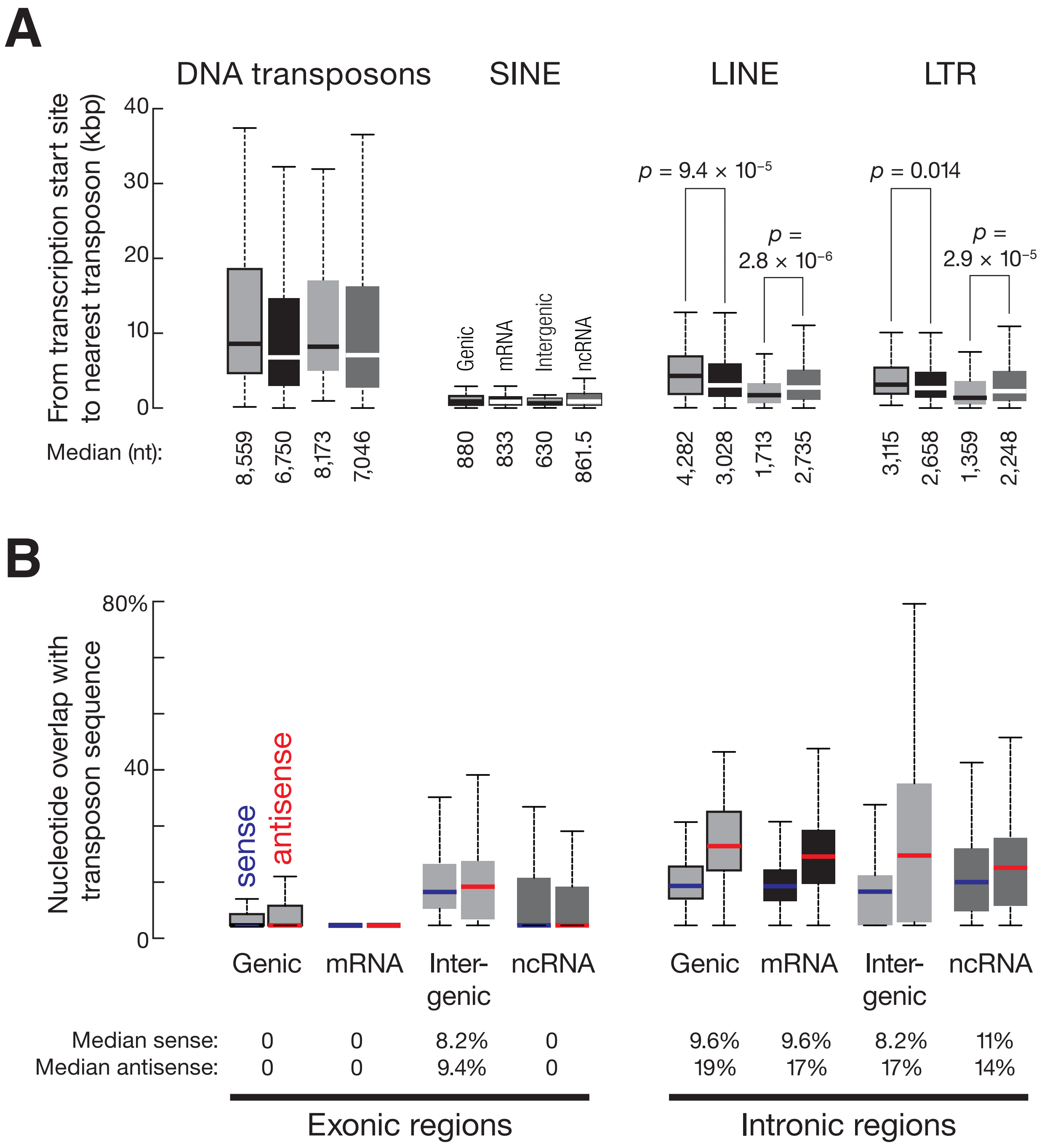
Li et al., Figure 3

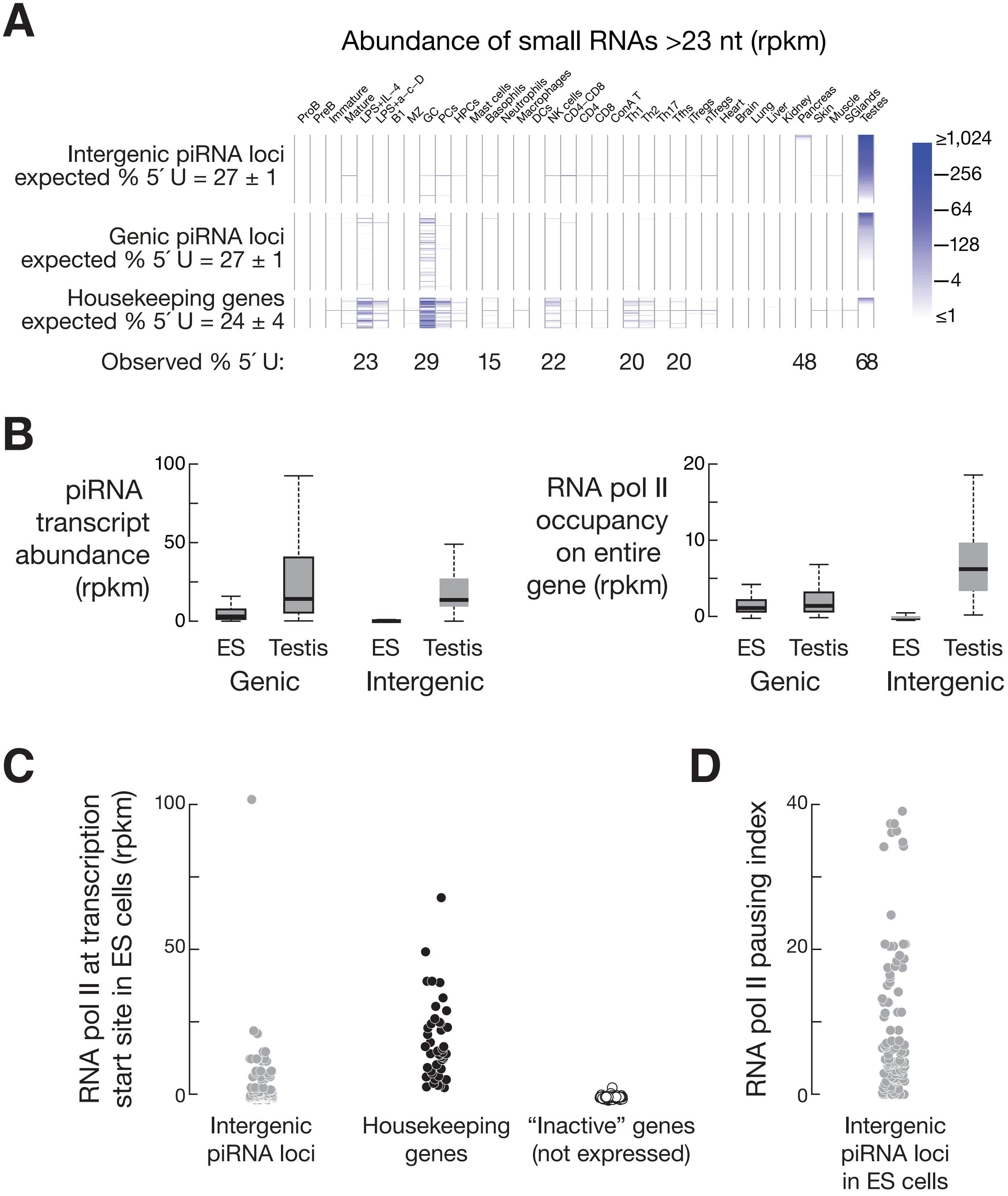


Li et al., Figure 4

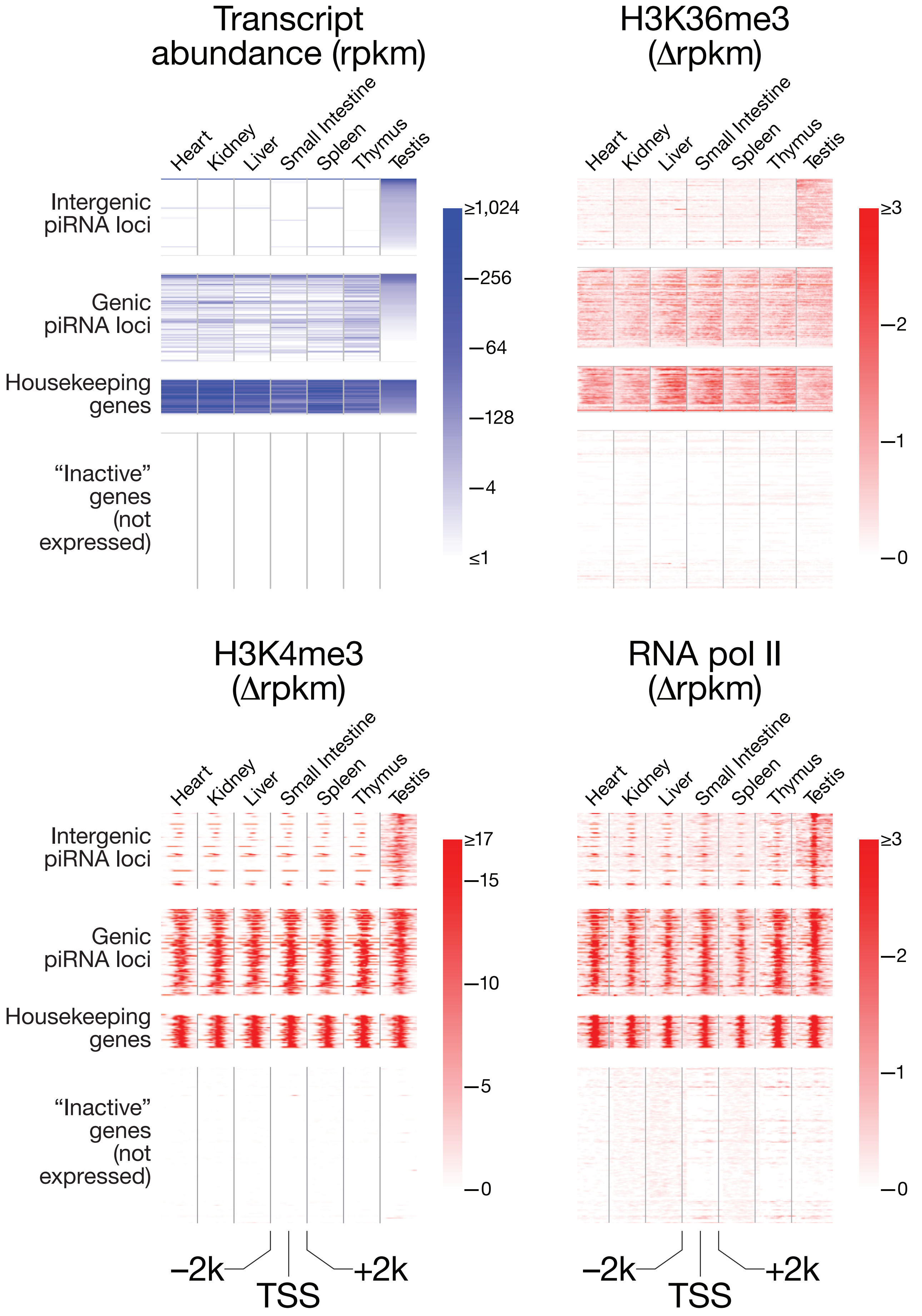


Li et al., Figure 5




A
Abundance of small RNAs >23 nt (rpkm)
ProB
PreB
Immature
Mature
LPS+IL–4
LPS+a–c–D
B1
MZ
GC
PCs
HPCs
Mast cells
Basophils
Neutrophils
Macrophages
DCs
NK cells
CD4–CD8
CD4
CD8
ConA T
Th1
Th2
Th17
Tfhs
iTregs
nTregs
Heart
Brain
Lung
Liver
Kidney
Pancreas
Skin
Muscle
SGlands
Testes
Intergenic piRNA loci
expected % 5´ U = 27 ± 1
Genic piRNA loci
expected % 5´ U = 27 ± 1
Housekeeping genes
expected % 5´ U = 24 ± 4
≥1,024
256
64
128
4
≤1
Observed % 5´ U:
23
29
15
22
20
20
48
68
B
piRNA transcript abundance (rpkm)
100
50
0
ES
Testis
ES
Testis
Genic
Intergenic
RNA pol II occupancy on entire gene (rpkm)
20
10
0
ES
Testis
ES
Testis
Genic
Intergenic
C
RNA pol II at transcription start site in ES cells (rpkm)
100
50
0
Intergenic piRNA loci
Housekeeping genes
"Inactive" genes (not expressed)
D
RNA pol II pausing index
40
20
0
Intergenic piRNA loci in ES cells

Li et al., Figure 7


Transcript abundance (rpkm)
Heart
Kidney
Liver
Small Intestine
Spleen
Thymus
Testis
Intergenic piRNA loci
Genic piRNA loci
Housekeeping genes
“Inactive” genes (not expressed)
≥1,024
—256
—64
—128
—4
≤1
H3K36me3 (Δrpkm)
Heart
Kidney
Liver
Small Intestine
Spleen
Thymus
Testis
≥3
—2
—1
—0
H3K4me3 (Δrpkm)
Heart
Kidney
Liver
Small Intestine
Spleen
Thymus
Testis
Intergenic piRNA loci
Genic piRNA loci
Housekeeping genes
“Inactive” genes (not expressed)
≥17
—15
—10
—5
—0
–2k
TSS
+2k
RNA pol II (Δrpkm)
Heart
Kidney
Liver
Small Intestine
Spleen
Thymus
Testis
≥3
—2
—1
—0
–2k
TSS
+2k