# NOTES TO Self

1. Check the text of the MolCel2013 paper against the DVD backup. The main text is from the MolCel website, as are the main figures. Supplemental text and figures need to be reconciled with your DVD backup.
   1. Add the piRNA references to this document
      1. This will take awhile.
2. Write another section tomorrow morning.
3. Phrases to FIND and Replace for symbols ETC.
   1. 5prime and 3prime = I want the symbols!

# COVER PAGE

# TITLE PAGE

Possible titles:

* Investigation of long RNAs using a novel ligation-based approach
* Investigation of long RNAs using RNA templated DNA ligation
* Long RNAs: biology, technology, and perspective
* Examination of dynamic long RNAs

A Dissertation Presented

By

Christian Knauf Roy

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSPHY

(MONTH, DAY, YEAR)

BIOCHEMISTRY

# SIGNATURE PAGE

A Dissertation Presented

By

Christian Knauf Roy

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

(Signature) (Name Typed),

Melissa J. Moore (Signature) (Name Typed),

(Signature) (Name Typed),

Phillip D. Zamore (Signature) (Name Typed),

Scot Wolfe

(Signature) (Name Typed),

Job Dekker

(Signature) (Name Typed),

Brenton R. Graveley

(Signature) (Name Typed),

Dissertation Committee

Zhiping Weng

(Signature) (Name Typed),

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the

Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school. (Signature)

Anthony Carruthers, Ph.D., Dean of the Graduate School of Biomedical Sciences Program (Typed)

Month, Day and Year (Typed)

# Front Matter

## Dedication

## Acknowledgements

Talk about going to Grad School

Advisors

MJM

PDZ

Committee members

Zhiping

Scot

Job

Lab Members

Aaron

Alper

Amrit

Eric and Erin

Collaborators

Dave Weaver

Muro

Graveley

Heinrich

Anna

Ogo

Family

Table of Contents

COVER PAGE i

TITLE PAGE 1

SIGNATURE PAGE 2

Front Matter 3

Dedication 4

Acknowledgements 5

Abstract 7

List of Tables 8

List of Figures 10

List of Symbols, Abbreviations or Nomenclature (optional) 11

Preface 13

Body Matter 14

Introduction 15

Research Chapters 22

Perspective/Final Summary & Conclusions 43

Back Matter 44

Appendices 45

Biblography 46

## Abstract

## List of Tables

## List of Figures

* RNA Sequencing technology figure
* SeqZip paper figures
* piRNA expression in mice
* MolCel paper figures

[Figure 0‑1 piRNA Precursors are RNA Pol II Transcripts 32](#_Toc364761512)

[Figure‑0‑2. The Major piRNA-Producing Genes of the Post-Partum Mouse Testis 33](#_Toc364761513)

[Figure 0‑3 Three Classes of piRNA-Generating Loci 36](#_Toc364761514)

[Figure 0‑4 Pre-pachytene piRNAs Persist in Pachytene Spermatocytes 37](#_Toc364761515)

[Figure 0‑5 Examples of Pachytene piRNA Genes 39](#_Toc364761516)

[Figure 0‑6 Examples of Pre-Pachytene piRNA Genes 39](#_Toc364761517)

[Figure 0‑7. A-MYB Binds the Promoters of Pachytene piRNA Genes 41](#_Toc364761518)

[Figure 0‑8 ChIP-qPCR Confirms ChIP-seq Data 42](#_Toc364761519)

[Figure 0‑9 Change in piRNA Expression in A-Myb, *Spo11*, *Mili, Tdrd6*, and *Tdrd9* Mutants 44](#_Toc364761520)

[Figure 0‑10 Examples of the Effect of the A-Myb Mutant on piRNA Expression 45](#_Toc364761521)

[Figure 0‑11 Pachytene piRNA Precursor Abundance in *A-Myb*, *Miwi*, and *Trip13* mutants 46](#_Toc364761522)

[Figure 0‑12 A-MYB Regulates Expression of mRNAs Encoding piRNA Pathway Proteins 48](#_Toc364761523)

[Figure 0‑13. A-Myb mutants, but Not Miwi Mutants, Change the Expression of RNA Silencing Pathway Genes 49](#_Toc364761524)

[Figure 0‑14 Feed-Forward Regulation of piRNA Biogenesis by A-MYB is Conserved in Rooster 53](#_Toc364761525)

[Figure 0‑15. Genomic Locations of piRNA Clusters in the Rooster (*Gallus gallus)* Testis. 54](#_Toc364761526)

## List of Symbols, Abbreviations or Nomenclature (optional)

### Abbreviations

|  |  |
| --- | --- |
| AS | Alternative Splicing |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| ChIP-Seq | Chromatin Immunoprecipitation followed by sequencing |
| HTS | High-throughput sequencing (see also *NGS*) |
| NGS | Next-generation sequencing |
| nt | A nucleotide of either DNA or RNA |
| bp | A basepair of DNA |
|  |  |

### Symbols

|  |  |
| --- | --- |
| 5´ |  |
| 3´ |  |
| μ |  |
| $ | Reference |
| % | Needed definition |
| # | Figure or Table |

### Definitions

|  |  |
| --- | --- |
| RNA-Seq |  |
| A ‘Read’ |  |
| Insert |  |
| Read length |  |
| Read depth |  |
| Coverage |  |
| Paired-end | When both sides of a DNA insert or template are sequenced, utilizing the original length of DNA between the reads to facilitate mapping (Roach, Boysen, Wang, & Hood, 1995) |
| Scaffold |  |
| contig |  |

## Preface

The work reported in this dissertation has been published in the following articles.

Chapter III has been published previously as Li, X. Z. Z., Roy, C. K. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W. W., … Zamore, P. D. D. (2013). An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. Molecular Cell, 50(1), 1–15. doi:10.1016/j.molcel.2013.02.016

Some contents of Chapter I are included in an accepted for publication:

# Body Matter

Introduction

#### Motivation

High school Biology class. That is where I first learned about how 'genes' - this things which I always heard everything was 'in' - actually worked.  There was a picture of a squiggly line called mRNA. This line was broken into boxes marked exons.  Transcribed mRNA went on to code for protein, and along with good part water, is what I was made of.  Made sense.  A subtle note on the page drew my eye. It pointed to one of the exon boxes and said 'alternative splicing.'  What's this?  The arrangement of the exons is not set in stone? It was around this time that the human genome was being finalized, we had something like 3.9 billion base pairs, and this little box only contained ~150 bases.  The implication struck me immediately.  If these boxes could be alternatively arranged, then the possibilities of unique mRNA molecules is staggering!  Far more then the 20K protein coding genes we are now known to possess.  Something that also made sense - this MUST be the reason why humans are so much more advanced then everything else - our mRNAs are alternatively spliced ( I didn't realize that even X *a very simple organism* also alternatively splices its genes.

That was ~10 years  before I found myself development a methodology that would greatly assist in the routine measurement of mRNAs that are alternatively spliced.  Strange to think that I was amazed with the biological process I would go on to study.  I was similaly amazed to learn about sanger sequencing - it seemed so elegant to me.  Another topic that truely interests me is HTS.

#### Sequencing history

Soon after it was realized that DNA is the source of genetic information in all living organisms(Watson & Crick, 1953), and the ‘pretty’ and ‘elegant’ arrangement of complementary, antiparallel DNA strands was known (Watson, Gann, & Witkowski, 2012), the ability to determine the specific arrangement, or ‘sequence’ of nucleotide bases in a given length of DNA was seen as a critical missing piece of technology. It took 25 years after the nature of DNA’s architecture for the ability to ‘[sequence’](#_Definitions) it to become a reality. By 1977, two very different methods reported by Sanger (Sanger & Coulson, 1975; Sanger, Nicklen, & Coulson, 1977) and Maxam-Gilbert (Maxam & Gilbert, 1977) were reported. These sequencing technologies, from then on referred to eponymously as ‘Sanger’ or ‘Maxam-Gilbert’ sequencing, were used to determine the specific order of a small DNA piece (200­­­­–300 nt). Sanger sequencing soon dominated most sequencing reactions, likely due to the conceptually more intuitive nature of the technology, and over the past 35 years, DNA sequences have been slowly cloned, sequenced, analyzed, and dutifully catalogued into knowledge.

During late 70’s and throughout the 80’s, DNA sequences were typically communicated in important publications (Bell et al., 1980). The birth of the internet in the 1990’s made essential publically-funded repositories for sequence information easily available (Benson, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2011). However, it was the human genome project (Eric S Lander, 2011), that provided the important activation energy which brought DNA sequencing from a hard-to-perform but necessary piece of analysis, to an organized, large-scale effort to assemble complete source genetic material for relatively simple (Sanger et al., 1978) to complex (E S Lander et al., 2001; J C Venter et al., 2001) genomes. An often criticized, but undeniably disrupting force, in the human genome project was the competing efforts of the privately-owned company Celera (J. Craig Venter, 2008). Taking a higher-throughput and centralized approach to determining the sequence of the human genome, Celera fundamentally changed to approach to a large-scale sequencing project. Instead of assigning specific sections of the genome to be worked out by individual labs, Celera parallelized the effort, by collecting many of the best “high-throughput” Sanger-sequencing devices from Agilent (ABI 3700 DNA Analyzer). Using a shotgun approach (R.Staden, 1979) sequenced pairwise (Roach et al., 1995) combined with sequence [scaffolds](#_Definitions) made available by the publically-funded project, Celera was able to create high-quality genomic sequences very quickly. Arguable, this was the first [deep sequencing](#_Definitions) effort, and changed the landscape of molecular and biochemical research, coincident with the beginning of a new millennium.

#### Deep sequencing history

Sequencing DNA by Sanger’s technology remains a valuable and critical tool in every biological scientist’s arsenal. However, the technology has a practical throughput limit. Each DNA molecule to be sequenced must be isolated & clonally amplified, typically using bacteria to do the heavy lifting. Given that the human genome (Consortium, 2005) comprises > 3 billion nt (on just one strand), and that each Sanger reaction will provide ~800 nt of quality sequence, we need at least ~4 Million individual reactions to determine the sequence of the human genome, assuming that all of our reads are of sufficient quality, length, and do not overlap by even 1 nt. Even the best practical improvements to workflows could not bring the Sanger approach to DNA sequencing in-line with aspirations of analyzing the genomes of many different species or individual organisms.

In the early 2000’s the first efforts to change the approach to DNA sequencing, first using MPSS (Brenner et al., 2000), but perhaps more importantly, and disruptively from a technology perspective, by Pyrosequencing (Ronaghi, Uhlén, & Nyrén, 1998) and Polony sequencing (Shendure et al., 2005). Both methods utilized emulsion PCR (Nakano et al., 2003) for clonal amplification prior to sequencing, removing the bottleneck of bacterial cloning in traditional workflows. In contrast to Sanger sequencing, where the final read out is the fluorescence of the last incorporated chain-terminating nucleotide, Pyrosequencing visualizes light given off by luciferase as it reactions with ATP generated from the pyrophosphate (PPi) by-product of nucleotide addition to the growing chain. Pyrosequencing was commercialized by 454 technologies. Polony sequencing involves a more complicated sequencing-by-ligation method, eventually commercialized by Applied Biosystems as the SOLiD platform. While both of these technologies provided valuable, high-throughput sequences, neither has been as successful as the approach commercialized by Solexa (eventually purchased and now known as Illumina).

Illumina uses a sequencing-by-synthesis approach using fluorescent nucleotides after clonal amplification of DNA on a slide surface. Since 2006, iterations of the Illumina platform (eg. GE, GE-II(x), Hi-Seq, Hi-Seq 2500) have demonstrated a steady and impressive increases in both [read depth and length](#_Definitions) [/#Figure]. On Februrary 15th 2012, Illumina announced on its [Basespace blog](http://blog.basespace.illumina.com/2012/02/15/genome-in-a-day/), that they had sequenced a HapMap [sample](http://ccr.coriell.org/Sections/Search/Sample_Detail.aspx?Ref=GM18507) at 40X [coverage](#_Definitions), using the HiSeq 2500 platform and paired-end 100 nt reads. This announcement demonstrated that in a single analysis attempt (but certainly not the day claimed by the title), analysis and assembly of a human genome is no longer the monumental endeavor it once was, and that completely new experimental possibilities are a reality for life science research.

#### Methodologies of Deep-sequencing of RNA

Just found the mother load of literature reviews http://blog.sbgenomics.com/history-of-rna-seq/

The first widely-accepted method for measuring gene expression via sequencing by proxy of cDNA molecules was Serial Analysis of Gene Expression (SAGE)(Velculescu, Zhang, Vogelstein, & Kinzler, 1995). While the importance of microarrays in the measurement of gene expression via cannot be overstated (Marioni, Mason, Mane, Stephens, & Gilad, 2008; Shendure, 2008) the technologies limited ability of novel sequence, and analogue nature of the signal, makes their relevance to this section somewhat off-topic. However, Sage, (similar to the previously discussed MPSS technique) produces a digital output of gene expression using a cleaver procedure of cleaving cDNA molecules using restriction endonucleases that leaves a ‘sticky end’. After cleavage, these molecules ligated and concatenated together to form longer DNA fragments. These fragments are cloned into a vector, amplified, and Sanger sequenced. Using known sequences incorporated during concatenation, the number of sequenced ‘fragments’ that align to a given gene is related to the abundance of the original mRNA molecule. While SAGE was a cleaver molecular trick allowing researches to dip into the 5-log range of expression typically seen in mRNA expression, it is still limited by read lengths and practical read depth of Sanger sequencing.

Not long after the Solexa/Illumina platform produced read lengths of sufficient length of depth to consider measuring gene expression were the first [RNA-Seq](#_Abbreviations) papers published (Lister et al., 2008; Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Nagalakshmi et al., 2008). These papers gave a powerful glimpse into the feature of molecular biology. Indeed, in the years since, analysis by RNA-Seq is quickly overtaking other forms of gene expression analysis, as demonstrated by the number of accessions deposited in GEO per year (Barrett et al., 2013). RNA-Seq allows for quantitative, digital measurement of RNA expression across a physiologically-relevant range (Blencowe, Ahmad, & Lee, 2009). While simultaneously measuring gene expression, the data can be used for novel sequence discovery, measuring RNA-editing (Li et al., 2011), transcript assembly (Trapnell et al., 2010), and by modifying the protocol or performing additional biochemical steps, can be used to investigate many aspects of RNA biology; Also #Figure of my own design).

Starting from the moment the RNA leaves the exit channel of the polymerase, the biology of RNA processing begins to occur. Many methodologies have been developed that enrich RNA-Seq libraries for a specific type of RNA molecule. For example, measurement of nascent RNA can be performed via GRO-Seq (Core, Waterfall, & Lis, 2008). Measuring the extremely complicated process of RNA turnover (referring to the rate at which given RNAs are produced and degraded) (Ghosh & Jacobson, 2010), can be done using XXX-Seq after incorporation of XX nucleotides or a biochemical handle such as biotin. RNA:protein interactions can be measured with or without crosslinking the protein to the RNA, via CLIP or RIP, respectively. Once an RNA has been fully transcribed, known processing steps such as Cap formation and poly(A) tail formation can be measured using any of the Cap-Seq/CAGE methodologies (Shiraki et al., 2003), or PAS-Seq (Shepard et al., 2011). With appropriate size-selection steps, small RNAs (Ghildiyal et al., 2008) can also be captured into a sequencing library. Finally, traditional RNA-Seq, can effectively capture fragments of all of the above mentioned libraries, even though it is mainly associated with measurement or analysis of traditional mRNAs.

RNA-Seq and its associated flavors are also traditionally associated with measuring gene expression in tissue culture cells, or RNA extracted from particular tissues. Recently, efforts to measure the RNA expression occurring in individual cells has gained attention (Shapiro, Biezuner, & Linnarsson, 2013). Perhaps the most interesting concept when thinking about measurement of gene expression in a single cell is the ‘biological uncertainty principle,’ wherein it is possible to either know, or change (but not both) the RNA composition of a single cell. The name borrows from Heisenburg’s uncertainty principle (Kennard, 1927) and is often confused with the more appropriate ‘observed effect’ (Riley & Steitz, 2013). Leaving that issue aside, measuring how unique a given cell, within a tissue of similar, but obviously unique cells, is an exciting and informative endeavor (Shalek et al., 2013; Wills et al., 2013). Compared to the plastic nature of DNA (Shendure & Aiden, 2012), the diversity of RNA synthesis within living cells is potentially much more complicated, and the ability to accurately measure RNA dynamics should allow us to make much more informative observations concerning biology then is currently possible (Djebali et al., 2012).

#### RNA Biology

##### RNA Expression

* ENCODE

The encode project revealed that most of the genome is transcribed into RNA.  This was done in cancerous cell lines, and while it revealed the potential for transcription, it did not reveal much biology beyond a cells in culture simply perpetuating their existence.

* Integration of different datasets for a more complete transcriptional picture

##### Splicing

\_\_Do I need much of a history of splicing?\_\_

##### RNA-Processing

* Transcription
* mRNP formation (MJM’s 2009 Cell review)
* nuclear export
  + ALREX
* Interactions of mRNP’s with the ribosome?

##### Alternative Splicing

Mention the pair of papers in science, dec 2012 that discuss evolution using RNA-Seq

===From QE===

## Updated references and this is a very good and useful section of language

Soon after the discovery of introns, it was reasoned that genes could be arranged in different combinations, greatly increasing the coding potential of a genome (Gilbert, 1978). The process of rearranging genes, now known as alternative splicing (AS), has proven to be an integral phase of gene expression in most eukaryotes. In just 15 years, the number of genes estimated to be alternatively spliced has grown considerably. In his Nobel lecture, Phil Sharp stated that: “Approximately, one of every twenty genes is expressed by alternative pathways of RNA splicing in different cell types or growth states” (Sharp, 1994). Not long after the assembly of the first human genome, a number of groups combed through Expressed Sequence Tag (EST) databases to increase that estimate to 35%-59% (Modrek & Lee, 2002). Soon after, analysis using specially designed microarrays resulted in an increased estimate of 74% (Johnson et al., 2003). However, in late 2008, three groups utilizing high-throughput sequencing (HTS) of cDNA (referred to as RNA-Seq) demonstrated that between 86% and 95% of human multi-exon genes are subject to AS (Pan et al., 2006; Sultan et al., 2008; E. T. Wang et al., 2008). Not only did they demonstrate that almost all genes are alternatively spliced, they also showed that AS often occurs in a tissue- and cell type-specific manner. In combination with regulation of transcription itself, the study of AS is critical to our understanding of the connections between the comparably static genomic DNA sequence and the highly flexible and adaptive abilities of organisms.

##### Deciphering a splicing code

A gene is alternatively spliced when, as a result of transcription and processing, there are at least two unique transcripts produced from one genomic sequence. Beyond counting observed isoforms, one major area of effort is to decode sequence regulatory elements (SREs) contained in pre-mRNA that define AS site selection (Z. Wang & Burge, 2008). In contrast to the core splicing signals, we have limited knowledge of the SREs that serve to increase or decrease the strength of a particular splice site, often within a sea of other potential sites. Through a variety of mechanisms, these elements serve as cis-acting sequences and binding sites for trans-acting factors. Some of the best-studied SREs include Exon Splicing Enhancers and Silencers (ESEs and ESSs). Members of the Serine-Arginine (SR) protein family typically bind to ESEs located in an exon, promoting its definition and thereby increasing the probability that the exon will be included in the final transcript (Graveley, 2000; Long & Caceres, 2009). Meanwhile, ESSs serve to squelch inclusion, often through binding trans-acting heterogeneous ribonucleoprotein particles (hnRNPs) (Martinez-Contreras et al., 2007). Therefore, binding of these trans-acting factors to their appropriate SREs can either promote or inhibit interactions between the splicing machinery and the pre-mRNA. The current working hypothesis is that a finely tuned combination of these binding events determines the final exon content of each isoform (House & Lynch, 2008).

Sequence motifs that compose the AS code have been teased out (Barash et al., 2010; Ladd & Cooper, 2002). Additionally, assignment of the binding motifs to tissue-specific trans-acting factors has also progressed (Jin et al., 2003; Licatalosi et al., 2008; Ule, Jensen, Mele, & Darnell, 2005). Many of these binding motifs were identified using combined computational and biochemical approaches. Computational approaches usually involve searching for a comparative enrichment of sequences near splice sites. Biochemical approaches typically include gel shift, SELEX, and cross linking. Many of these approaches are performed in vitro and disregard the importance of cellular context on binding affinities. However, with the increasing accessibility of deep sequencing, many groups are extracting physiologically relevant, high-resolution data from traditional biochemical techniques(Ingolia, Ghaemmaghami, Newman, & Weissman, 2009; Ingolia, Lareau, & Weissman, 2011). Deep-sequencing approaches are also being applied to questions involving mechanisms of AS. In addition to the RNA-Seq experiments, High-Throughput Sequencing [following] Cross-Linking Immunoprecipitation (HTS-CLIP) has confirmed SRE motif data predicted from computational and microarray experiments (Hafner et al., 2010; Licatalosi et al., 2008). Using this approach, researchers can now enrich their samples for sequences that bind trans-acting factors of interest.

##### Tissue-specific RNAs

Explain how alternative splicing may just be tissue-specific splicing, cite Graveley papers?

##### The Isoform problem & methods for large-scale analysis of RNA

As with many areas of basic research, the field of AS relies on large-scale (aka – global, genome-wide, high-throughput) techniques. Two of the most widely applied technologies employed for large-scale analysis of gene expression are microarrays and ‘2nd generation’ HTS sequencing. Unfortunately, both of these techniques have fundamental limitations, with the major issues being probe specificity for the former and read length for the latter.

Microarrays rely on hybridization of a target sequence to a known probe averaging 25 to 100 nt in length (Southern, 2001). Therefore, microarrays indicate only the presence of short sequences in the target sample and do not provide adequate linkage information of these sequences. A hypothetical scenario can be used to describe it another way. Say we are investigating a transcript known to display two different regions of AS (Figure 1-A&B). Probes targeting these two regions demonstrated an increase in signal for both AS events. Unfortunately, we could not determine if we observed an increase in unique transcripts, each containing only one region of AS, or an increase in production of a single transcript containing both regions (Calarco et al., 2007a). This binary analysis is the heart of the “connectivity problem.” Microarrays have proven extremely informative and will likely continue to do so in more targeted applications. However, this issue, combined with concerns of cross-hybridization, reproducibility, and a comparably small dynamic range, will likely hasten microarray displacement by RNA-Seq as the preferred method for comprehensive analysis of gene expression (Shendure, 2008).

Many researchers are turning toward 2nd generation HTS methodologies for comprehensive transcriptome analysis. This sequencing approach has significance advantages over microarrays. Specifically, it allows de novo identification of isoforms, over a larger dynamic range, in a quantitative fashion (Mortazavi et al., 2008). Additionally, newly developed techniques enrich samples for low-abundance isoforms, making the complete cataloging of AS events a possibility (Djebali et al., 2008; Salehi-Ashtiani et al., 2008). Unfortunately, the current read-length abilities (depicted in Figure 1-B) of all sequencing platforms do not solve the connectivity problem. Excluding single-molecule HTS read lengths of sufficient length (Shendure et al., 2004), other approaches proposed to solve the connectivity problem include traditional cloning and sequencing or hybridization of query oligos to single-molecule transcripts (Zhu et al., 2003; Calarco et al., 2007a; Emerick et al., 2007). While these approaches can determine exon sequence connectivity, they scale poorly and are not feasible for large-scale applications.

Clearly, AS is an essential regulatory mechanism involved in the control of human gene expression. Its combinatorial nature could potentially answer many questions, such as a physical explanation of what separates us from our closest evolutionary ancestor, the chimpanzee (Calarco et al., 2007b). Additionally, the influence of AS on disease and cancer is slowly coming to light (Tazi et al., 2009). Unfortunately, because of the limitations of methods currently used for the large-scale analysis of isoform expression we fail to obtain the complete picture of AS. One specific missing element of that picture is the prevalence of coordination between different regions of AS separated by large spans of sequence. An efficient, large-scale, single-molecule technique that maintains isoform sequence connectivity is required to complete the complicated picture of AS.

#### Coordination in splicing

==From my qualifying exam==

Fededa, Fagnani, Parra (Deep intron), Peng (exons)

Identification of proximally acting SREs is progressing at a rapid pace. New and traditional biochemical methods, coupled with HTS, will undoubtedly fuel this progress. Unfortunately, a critical component of AS regulation currently neglected by the field is that of SREs acting across a considerable distance. One observation that may lead to the identification of long-range SREs is intramolecular coordination between distal splicing decisions. Figure 1-A shows a model transcript that may exhibit coordinated distal regions of AS. In this model, the 5’ region of AS contains a cassette exon, which may or may not be included. This region is separated from the 3’ region of AS by many thousands of nucleotides. Does the decision to include the cassette exon have an effect on which of the mutually exclusive exons is included? This type of AS regulation may represent a general and pervasive phenomenon.

There is precedence in the literature for genes known to display coordinated regions of AS. One of the clearest examples is mouse fibronectin (see Figure 2). In this gene, inclusion of the alternatively spliced Extra Domain A (EDI or EDA) region promotes splicing from one of three alternative 3’ Splice Site (3’SS) in the type III homologies connecting segment (IIICS) region, resulting in more frequent production of shorter transcripts (Fededa et al., 2005). This effect occurs over six constitutively expressed exons and 800 nt of sequence (5400 nt if introns are considered). The authors of this study also analyzed EST databases, concluding that approximately 25% of human genes contain multiple regions of AS. How many of these regions could show a coordinated effect, similar to that observed in fibronectin? Providing some insight into this question, Fagnani et al used microarrays designed to report on inclusion levels of cassette exons in mammalian central nervous system tissues (Fagnani et al., 2007). The results produced a set of 38 pairs of exons mapping to the same gene that showed a coordinated increase or decrease of inclusion levels. A partial list, focusing on high-confidence events containing potentially coordinated regions of AS separated by more than 1000 nt of final transcript is shown in Table 1. Unfortunately, this and other large-scale studies preclude the complete determination of intramolecular coordination of AS.

#### Long range RNA interactions

Bindewald, Eckart, and Bruce a Shapiro. 2013. “Computational Detection of Abundant Long-range Nucleotide Covariation in Drosophila Genomes.” RNA (New York, N.Y.) (July 25): 1171–1182. doi:10.1261/rna.037630.112. <http://www.ncbi.nlm.nih.gov/pubmed/23887147>.

#### lncRNAs

#### piRNAs

! Importance of mammalian piRNAs to maintain sterility

Mammalian spermatogenesis is a critical biological process that ensures the future of a species. Recently the importance a specific kind of small RNA—piRNAs—on the proper function of spermatogenesis has become clear [$Ref to review].  Even after near a decade since their discovery, the essential mechanisms of piRNA importance to proper mammalian spermatogenesis remain unknown. These mechanisms include: biogenesis, physiological targets, and terminal function of maintain sterility. Even with these extremely important unknowns, the importance of piRNAs on mammalian spermatogenesis is clear—without a functioning piRNA pathway—males are sterile. Studies in humans have also linked SNPs in the Argonaute proteins that bind piRNAs, PIWI, to decreased fertility [$REF – PLOS bio paper].

!Description of the three classes of mammalian piRNAs (#Figure)

Mammalian piRNAs can be divided into three major classes (#Figure of my own design on piRNAs). The first, present before birth, are those of the ‘fetal piRNAs’. These piRNAs tend to be short, bind the PIWI protein MILI2 in mice, and have sequences found in transposable elements [$REF – 2008 Mol cel paper?]. The next class of piRNAs, historically but confusing grouped with the previous class, are called Pre-pachytene piRNAs. Pre-pachytene piRNAs are expressed just before birth, and continue to be expressed in functioning testes, often associated with spermotoXX and spermotXX, precursor cells to mature sperm(?), and tend to map to traditional, and annotated, protein coding genes. Finally, due to their unique sequence in the genome, the genetic origin of a millions of piRNAs belonging to the third class, the pachytene piRNAs, was immediately known.  Pachytene piRNAs are extremely abundant after, not coincidently, the pachytene stage of meiosis I when (descriptor) chromosomes pair up, cross over, and rearrange their genetic material. Extremely abundant means XX fold enrichment of pachytene piRNAs compared to miRNAs, as compared to developmental timepoint just two days earlier. The genomic origins of these piRNAs, while unique in terms of sequence, are often in ‘gene deserts,’ unannotated, and devoid of intronic sequences. This ‘gene architecture’ makes the pachytene piRNA loci some of the most interesting RNA-producing regions of the mammalian genome.

!Known functions of mammalian piRNAs

Morolotos paper

The two nature MIWI and MILI slicer papers form 2011

! Comparisons between fly and mammalian piRNA systems

The transposon-mapping nature of the fetal piRNA class made obvious comparisons to the fly piRNA system nature. In the fly system, primary piRNAs transcribed from discrete loci and fed into an amplification loop between two PIWI proteins PIWI(3) and AGO3 (‘the ping-pong’ cycle) ($REF Brenneki cell paper 2007). It is believed that PIWI proteins loaded with piRNAs bind and silence transposon messages, using the cleaved transposon transcripts, in combination with primary piRNA transcripts, as a substrates in the Ping-Pong cycle ($REF some newer review demonstrating this activity).

! Our annotation of the pachytene piRNAs and discovery of A-MYB

However, annotation of the molecular RNA precursors to these polymers–the piRNA precursor transcripts–was a simple inclusion of large chunks of chromosomal coordinates.  Using 6 different types of high-throughput sequencing datasets, we have accurately defined the genetic structure of 2124 piRNA-generating loci. With these annotations, investigation into the cause of coordinated expression of pachytene piRNAs–a highly abundant and experimentally tractable class of piRNAs–, became possible.  Using transcriptional start sites, a motif for the transcription factor A-MYB was clearly observed within the promoters of most of pachytene piRNA-generating loci.  Additionally, gene expression analysis of mice containing a mutant form of A-MYB revealed that A-MYB also coordinates the expression of many proteins involved, or associated with, proper piRNA biogensis and function.

!Critical importance of integrating many different HTS datasets into mammalian piRNA study

piRNAs lend themselves to study using HTS

Require more datasets to work back to molecular precursors

RNA-Seq does not give precision necessary for annotation of 5prime and 3prime ends.

Cap-seq is too noisey to not have orthogonal dataset for comparison

Precision of 5prime end is critical for proper measurement of proximity to suspected transcription factor binding motifs and experimentally-determined ChIP signal

PAS-Seq challenging due to internal priming sites, difficulty of HTS platforms to read through homopolymers, and dynamic nature of 3prime end processing.

!Current status and future of transcriptome assembly

Reference most recent transcript assembly review.

Most successful methods are guided by known annotations

Unguided typically provide many discontinuous ‘contigs’

Maybe assisted by deeper RNA-Seq

Paired-end is critically important

## Research Chapters

1. CHAPTER I: SeqZip methodology

#### Nature methods paper

##### SeqZip Development

##### Discovery of novel enzyme activity

##### CD45

##### FN1

##### 10 Gene set work?

Can take language from your QE for an introduction

You will have to write up the results.

##### DSCAM

You should tie together the fact that Dscam ALSO has Fibronectin sections!

1. CHAPTER II:HIV and piRNA precursors

#### piRNA precursors

! Mention the TALEN work you completed

! Insert the table you made of the top paired clusters and the talens that were created.

#### HIV Transcript integrity

1. CHAPTER III : MolCel2013

#### INTRODUCTION

P-element induced wimpy testis (PIWI)-interacting RNAs (piRNAs) can be distinguished from other animal small silencing RNAs by their longer length (typically 23–35 nt), 2′-O-methyl-modified 3′ termini, and association with PIWI proteins, a distinct subgroup of Argonaute proteins, the small RNA-guided proteins responsible for RNA interference and related pathways (Kumar and Carmichael, 1998; Aravin and Hannon, 2008; Farazi et al., 2008; Kim et al., 2009; Thomson and Lin, 2009; Cenik and Zamore, 2011). piRNA production does not require Dicer, the double-stranded RNA endonuclease that makes microRNAs (miRNAs) and small interfering RNAs (siRNAs), and piRNAs are thought to derive from single-stranded rather than double-stranded RNA (Vagin et al., 2006; Houwing et al., 2007).

In most bilateral animals, germline piRNAs protect the genome from transposon activation, but also have other functions (Aravin et al., 2001, 2007, 2008; Vagin et al., 2004, 2006; Brennecke et al., 2007; Carmell et al., 2007; Hartig et al., 2007; Kuramochi-Miyagawa et al., 2008; Ashe et al., 2012; Lee et al., 2012; Shirayama et al., 2012). A few days after birth, the majority of piRNAs in the mouse testis are pre-pachytene piRNAs; 25% of these piRNA species map to more than one location in the genome. A second class of piRNAs, typically derived from intergenic regions, has been reported to emerge in the mouse testis 14.5 days postpartum (dpp), when the developing spermatocytes synchronously enter the pachytene phase of meiotic prophase I. These pachytene piRNAs compose >95% of piRNAs in the adult mouse testis. Loss of genes required to make pachytene piRNAs blocks production of mature sperm (Deng and Lin, 2002; Aravin and Hannon, 2008; Reuter et al., 2011; Vourekas et al., 2012). What triggers the accumulation of pachytene piRNAs when spermatocytes enter the pachynema is unknown.

In Caenorhabditis elegans, each piRNA is processed from its own short RNA polymerase II (Pol II) transcript (Gu et al., 2012). In contrast, insect and mouse piRNAs are thought to be processed from long RNAs transcribed from large piRNA loci. Supporting this view, a transposon inserted into the 5′ end of the flamenco piRNA cluster in flies reduces the production of flamenco piRNAs 168 kbp 3′ to the insertion, suggesting that it disrupts transcription of the entire locus (Brennecke et al., 2007). High-throughput sequencing and chromatin immunoprecipitation (ChIP) has been used to define the genomic structure of the piRNA-producing genes of immortalized, cultured silk moth BmN4 cells (Kawaoka et al., 2013). However, for flies and mice, we do not know the structure of piRNA-producing genes, their transcripts, or the nature of the promoters that control their expression.

Instead, piRNA loci have been defined as clusters: regions of the genome with a high density of mapping piRNA sequences (Aravin et al., 2006, 2007; Girard et al., 2006; Grivna et al., 2006a; Lau et al., 2006; Brennecke et al., 2007; Ro et al., 2007). In reality, piRNA-producing loci correspond to discrete transcription units that include both intergenic loci believed to encode no protein (Brennecke et al., 2007, 2008; Vourekas et al., 2012) and protein-coding genes that also produce piRNAs (Aravin et al., 2007; Robine et al., 2009; Saito et al., 2009).

We used high-throughput sequencing data to define the genes and transcripts that produce piRNAs in the juvenile and adult mouse testis. Using these data, we identified the factor that initiates transcription of pachytene piRNA genes: A-MYB (MYBL1), a spermatocyte protein that serves as a master regulator of genes encoding proteins required for cell-cycle progression through the pachytene stage of meiosis (Trauth et al., 1994; Bolcun-Filas et al., 2011). A-MYB also initiates transcription of the genes encoding many piRNA biogenesis factors. The combined action of A-MYB at the promoters of genes producing pachytene piRNA precursor transcripts and genes encoding piRNA biogenesis proteins creates a coherent feedforward loop that triggers a >6,000-fold increase in pachytene piRNA abundance during the ∼5 days between the early and late phases of the pachytene stage of male meiosis. A-MYB also promotes its own transcription through a positive feedback loop. The A-MYB-regulated feedforward loop is evolutionarily conserved: A-MYB is bound to the promoters of both piRNA clusters and PIWIL1, TDRD1, and TDRD3 in the rooster (Gallus gallus) testis.

#### RESULTS

##### Defining piRNA-Producing Transcripts in the Mouse Testis

To define the structure of piRNA-producing loci in the testis of wild-type adult mice, we assembled the transcripts detected by three biological replicates of strand-specific, paired-end, rRNA-depleted, total RNA sequencing (RNA-seq; Figure 1A). We mapped reads to the mouse genome using TopHat (Trapnell et al., 2009) and performed de novo transcriptome assembly using Trinity (Grabherr et al., 2011) to identify unannotated exon-exon junctions. We used all mapped reads, including reads corresponding to unannotated exon-exon junctions, to perform reference-based transcript assembly (Cufflinks; Trapnell et al., 2010).

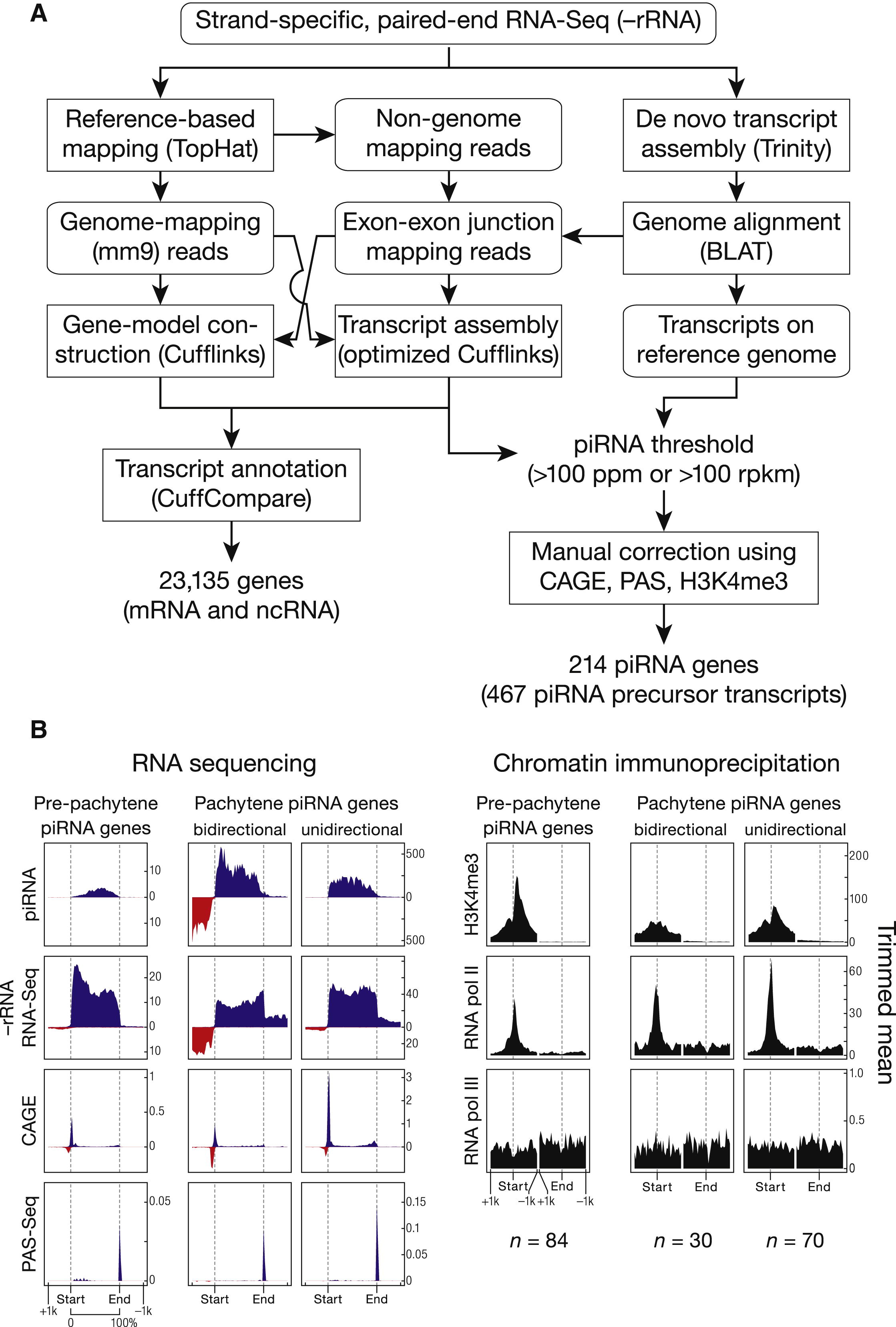
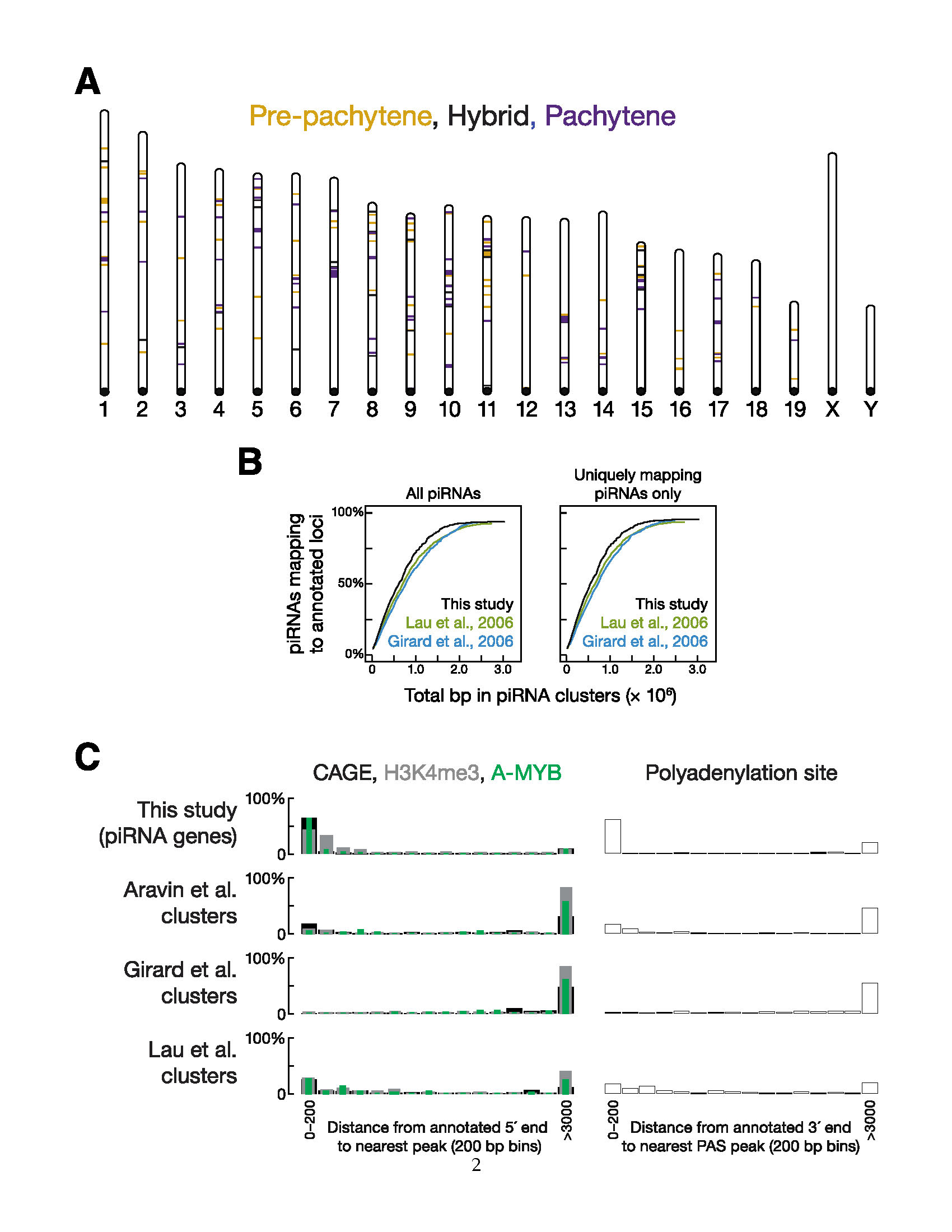


Figure ‑ piRNA Precursors are RNA Pol II Transcripts

(A) Strategy to assemble the mouse testis transcriptome. Rectangles with rounded corners, input or output data; rectangles, processes. Decisions are shown without boxing.(B) Aggregated data for piRNA-producing transcripts (5% trimmed mean). Oxidized small RNA (>23 nt) sequencing data were used to detect piRNAs; transcript abundance was measured using total RNA depleted of rRNA (RNA-seq). RNA Pol III data were from SRA001030. Dotted lines show the transcriptional start site (Start) and site of polyadenylation (End). See also Figure S1 and Table S1.



Figure‑‑. The Major piRNA-Producing Genes of the Post-Partum Mouse Testis

(A) Positions of the 214 major piRNA-producing genes on the 19 autosomes of mice. We detected no loci on the X or Y chromosomes. (B) Cumulative distributions for all piRNAs and for uniquely mapping piRNAs comparing the piRNA loci defined by our methods and by previous approaches (Girard et al., 2006; Lau et al., 2006). (C) Histogram of distances (in 200 bp bins) from the annotated 5′ or 3′ end of a piRNA gene (this study) or cluster to the nearest peak of reads from high-throughput sequencing for transcript 5′ (CAGE-seq) or 3′ (PAS-seq) ends, transcription start sites (H3K4me3) or A-MYB binding.

To identify the transcripts that produce piRNAs, we sequenced piRNAs from six developmental stages of mouse testes (10.5 dpp, 12.5 dpp, 14.5 dpp, 17.5 dpp, 20.5 dpp, and adult) and mapped them to the assembled transcripts. The first round of spermatogenesis proceeds synchronously among the tubules of the testis: mouse testes at 10.5 dpp advance no further than the zygotene stage (staging according to Nebel et al., 1961); 12.5 dpp to the early pachytene; 14.5 dpp to the middle pachytene; 17.5 to the late pachytene; and 20.5 dpp to the round spermatid stage. For each stage, we prepared two sequencing libraries: one comprising all small RNAs and one in which oxidation was used to enrich for piRNAs by virtue of their 2′-O-methyl-modified 3′ termini (Ghildiyal et al., 2008).

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 of 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 an N(5′)ppp(5′)N cap structure (cap analysis of gene expression; CAGE) and the 3′ ends of transcripts preceding the poly(A) tail (polyadenylation site sequencing; PAS-seq). The assembled piRNA-producing transcripts likely 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 the library derive from complete transcripts extending from the 5′ cap to the poly(A) tail (Figure 1B). Conventional 5′ and 3′ RACE (rapid amplification of cDNA ends) analysis of piRNA-producing transcripts confirmed the ends of 16 loci (data not shown). To provide additional 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 associated with RNA Pol II transcription start sites (Guenther et al., 2007).

##### piRNA Precursor RNAs are Canonical RNA Pol II Transcripts

The presence of 5′ caps and poly(A) tails and the binding of histone H3K4me3 to the genomic DNA immediately upstream of the transcription start site of each piRNA locus suggest that piRNA transcripts are produced by RNA pol II (Figure 1B). Moreover, using antibodies to RNA pol II but not RNA pol III, ChIP-seq showed a peak at the transcription start site as well as polymerase occupancy across the entire piRNA gene (Figure 1B; Kutter et al., 2011). We conclude that piRNA transcripts are conventional RNA pol II transcripts bearing 5′ caps and 3′ poly(A) tails.

##### A Transcript-based Set of piRNA Loci

Our transcriptome assembly yielded 467 piRNA-producing transcripts that define 214 genomic loci (Figure S1A 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 95% of all piRNAs.

Previous studies 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 in that it (1) uses RNA-seq data, whose greater read length facilitates the identification of introns, allowing us to define the architecture of piRNA precursor transcripts and (2) uses 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 (Figures S1B and S1C; 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. The 41 loci account for 2% of piRNAs at 10.5 dpp and 0.36% in the adult testis.

##### Three Classes of piRNAs During Post-Natal Spermatogenesis

Mice produce three PIWI proteins: MIWI2 (PIWIL4), which binds piRNAs in perinatal testis (Carmell et al., 2007; Aravin et al., 2008); MILI (PIWIL2), which binds piRNAs at least until the round spermatid stage of spermatogenesis (Kuramochi-Miyagawa et al., 2004; Aravin et al., 2006, 2007); and MIWI (PIWILl), which is first produced during the pachytene stage of meiosis (Deng and Lin, 2002). From 10.5 to 20.5 dpp, piRNA abundance increases and longer piRNAs appear, reflecting a switch from MILI-bound piRNAs, which have a 26–27 nt modal length (Montgomery et al., 1998; Aravin et al., 2006, 2008; Robine et al., 2009), to MIWI-bound piRNAs, which have a 30 nt modal length (Figure S2A; Reuter et al., 2009; Robine et al., 2009). This switch occurs at the pachytene phase of meiosis. MILI-bound pre-pachytene piRNAs predominate before the onset of pachynema; at the pachytene and round spermatid stages, most piRNAs are MIWI-bound pachytene piRNAs.

We used hierarchical clustering to analyze the change in piRNA abundance from 10.5 to 20.5 dpp for the 214 genes defined by our data (Figures 2A and S2A and Table S2). Three types of piRNA-producing genes were identified according to when their piRNAs first accumulate and how their expression changes during spermatogenesis: 84 pre-pachytene, 100 pachytene, and 30 hybrid loci. At 10.5 dpp, the earliest time we evaluated, 84 genes dominate piRNA production (median piRNA abundance per gene = 16 rpkm; Figure 2B). Nearly all (81 out of 84) were congruent with protein-coding genes. The 84 pre-pachytene piRNA genes account for 13% of piRNAs at 10.5 dpp, but only 0.31% of piRNAs in the adult testis. Of the pre-pachytene piRNAs accounted for by the 84 loci, 15% derive from 31 piRNA-producing genes that, to our knowledge, have not previously been described.

.

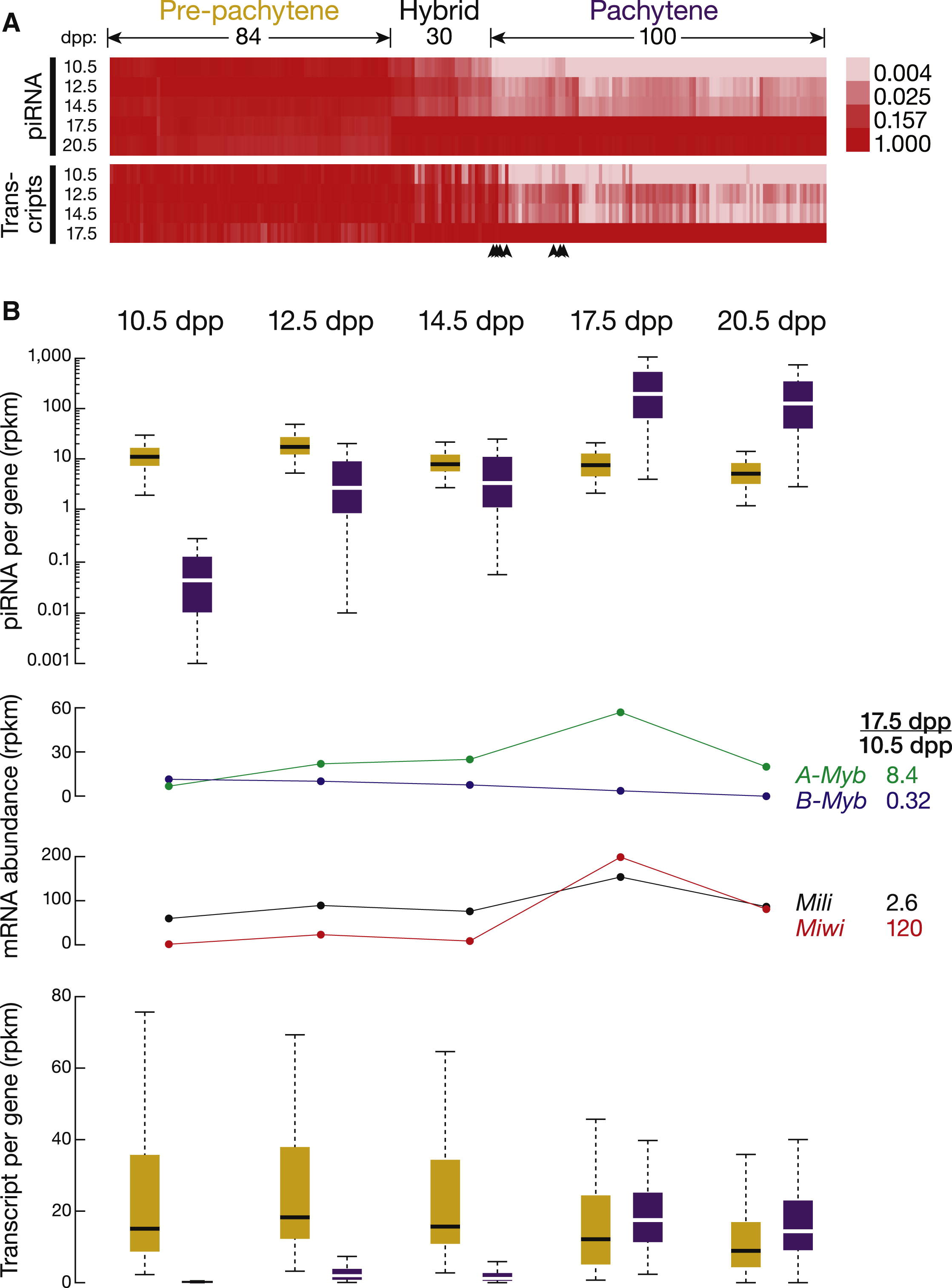


Figure ‑ Three Classes of piRNA-Generating Loci

(A) Normalized piRNA density (rpkm) for each piRNA-producing gene is shown as a heatmap across the developmental stages. Hierarchical clustering divided the genes into three classes. Arrowheads mark seven pachytene piRNA genes that were not classified as pachytene according to the change in the abundance of their precursor RNAs from 10.5 to 17.5 dpp.(B) Top: box plots present piRNA density per gene as spermatogenesis progresses (here and elsewhere, pre-pachytene in yellow and pachytene in purple). Middle: expression of A-Myb, B-Myb, Mili, and Miwi was measured by RNA-seq. Bottom: box plots present piRNA precursor expression per gene, measured by RNA-seq, from 10.5 to 20.5 dpp. See also Figure S2 and Table S2.

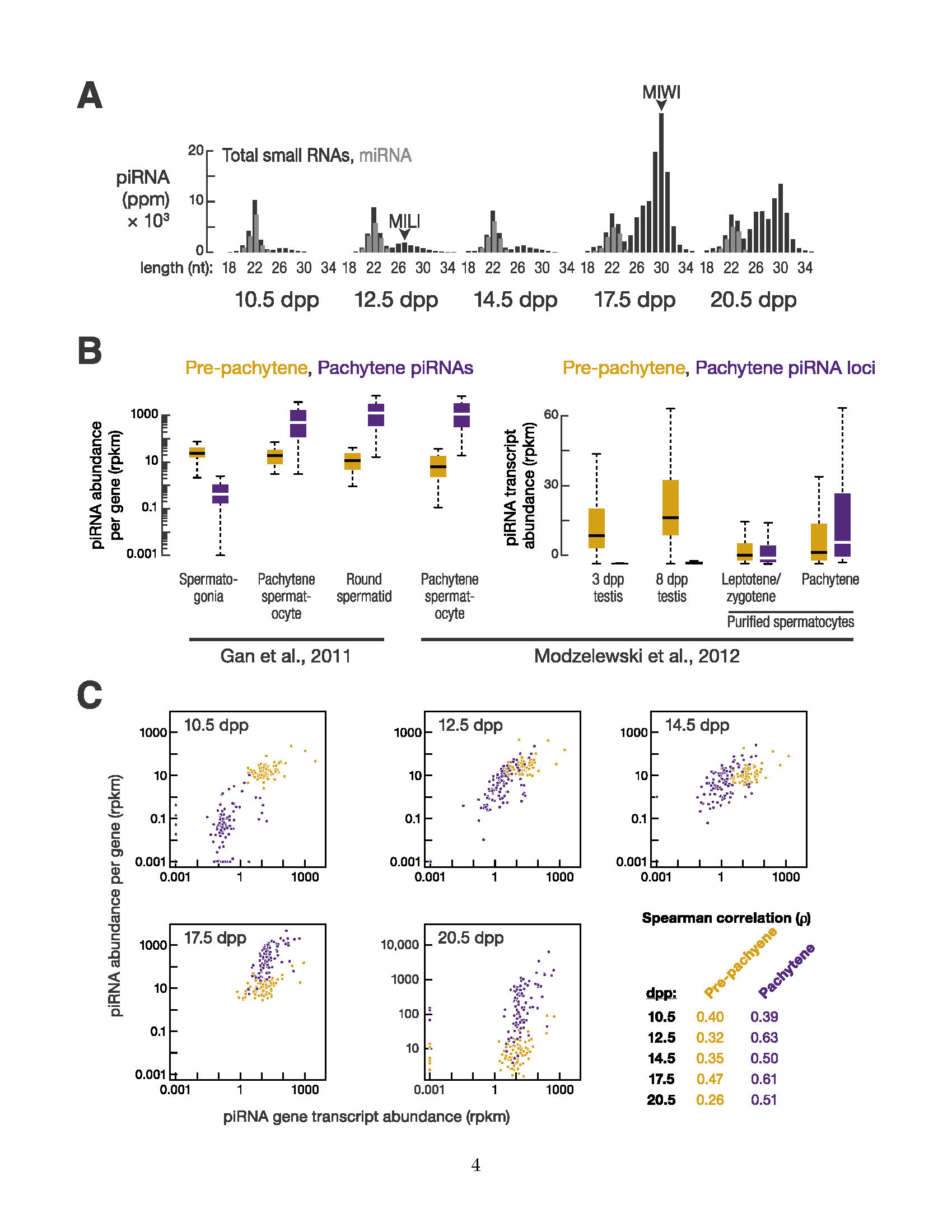


Figure ‑ Pre-pachytene piRNAs Persist in Pachytene Spermatocytes

(A) As shown previously by others using lower temporal resolution, the modal length of piRNAs increases as spermatogenesis proceeds to more advanced stages. (B) Total piRNA rpkm abundance and piRNA transcript abundance per locus by class, from purified spermatogonia, spermatocytes, round spermatids, and 3 dpp and 8 dpp testis (Gan et al., 2011; Modzelewski et al., 2012). (C) Correlation between piRNA abundance per locus and piRNA precursor transcription from 10.5 to 20.5 dpp. Throughout the Supplemental Figures, gold indicates pre-pachytene and purple indicates pachytene piRNA loci.

A parallel analysis of piRNA precursor transcription using RNA-seq (>100 nt) corroborated the classification based on piRNA abundance; of the 100 piRNA genes classified as pachytene based on the developmental expression profile of their piRNAs, 93 were grouped as pachytene according to the developmental expression profile of their transcripts. Of these 93, 89 are intergenic. All 84 piRNA genes designated pre-pachytene using piRNA data were classified as pre-pachytene according to their transcript abundance.

Despite their name, pre-pachytene piRNAs were readily detected in >90% and ∼95% pure pachytene spermatocytes, as well as round spermatids (Figure S2B; Gan et al., 2011; Modzelewski et al., 2012). Transcript abundance from the 84 pre-pachytene loci was high at 3 dpp (median abundance = 11 rpkm), higher by 8 dpp (18 rpkm), and lower in purified leptotene/zygotene spermatocytes (3.3 rpkm; Figure S2B). Yet piRNA precursor transcripts were readily detectable in purified pachytene spermatocytes at a level (4.6 rpkm) comparable to that in purified leptotene/zygotene spermatocytes (Figure S2B; Gan et al., 2011; Modzelewski et al., 2012). From 10.5 to 20.5 dpp, the steady-state level of pre-pachytene piRNA precursor transcripts remained constant (Figure 2B).

Finally, the abundance of pre-pachytene piRNA precursor transcripts was better correlated with pre-pachytene piRNA abundance at 17.5 dpp (ρ = 0.47), when pachytene spermatocytes compose a larger fraction of the testis, than at 10.5, 12.5, or 14.5 dpp (0.32 ≤ ρ ≤ 0.40; Figure S2C). Our data suggest that the pre-pachytene loci continue to be transcribed and processed into piRNAs long after spermatocytes enter the pachytene stage of meiosis. Thus, the name pre-pachytene piRNA is a misnomer that should be retained only for historical reasons.

Hierarchical clustering identified 100 pachytene genes whose piRNAs emerge at 12.5 dpp, 2 days earlier than previously reported (Girard et al., 2006). Nearly all the pachytene genes are intergenic (93 out of 100). piRNA expression from pachytene piRNA genes peaks at 17.5 dpp (Figure 2B). Overall, the median abundance of piRNAs from these 100 loci increased >6,000-fold from 10.5 to 17.5 dpp. Transcripts from pachytene genes were low at 10.5 dpp (median abundance = 0.15 rpkm) and increased 116-fold from 10.5 to 17.5 dpp. From 10.5 to 20.5 dpp, the dynamics of pachytene piRNA abundance from each piRNA gene correlated with the increase in abundance of its precursor transcripts (0.39 ≤ ρ ≤ 0.63; p value ≤ 7.3 × 10−5; Figure S2C). The 100 pachytene genes account for 92% of piRNAs in the adult testis, making it unlikely that biologically functional pachytene piRNAs originate from thousands of genomic loci (Gan et al., 2011). Figures 3 and S3 provide examples of pachytene and pre-pachytene piRNA genes defined by our data..

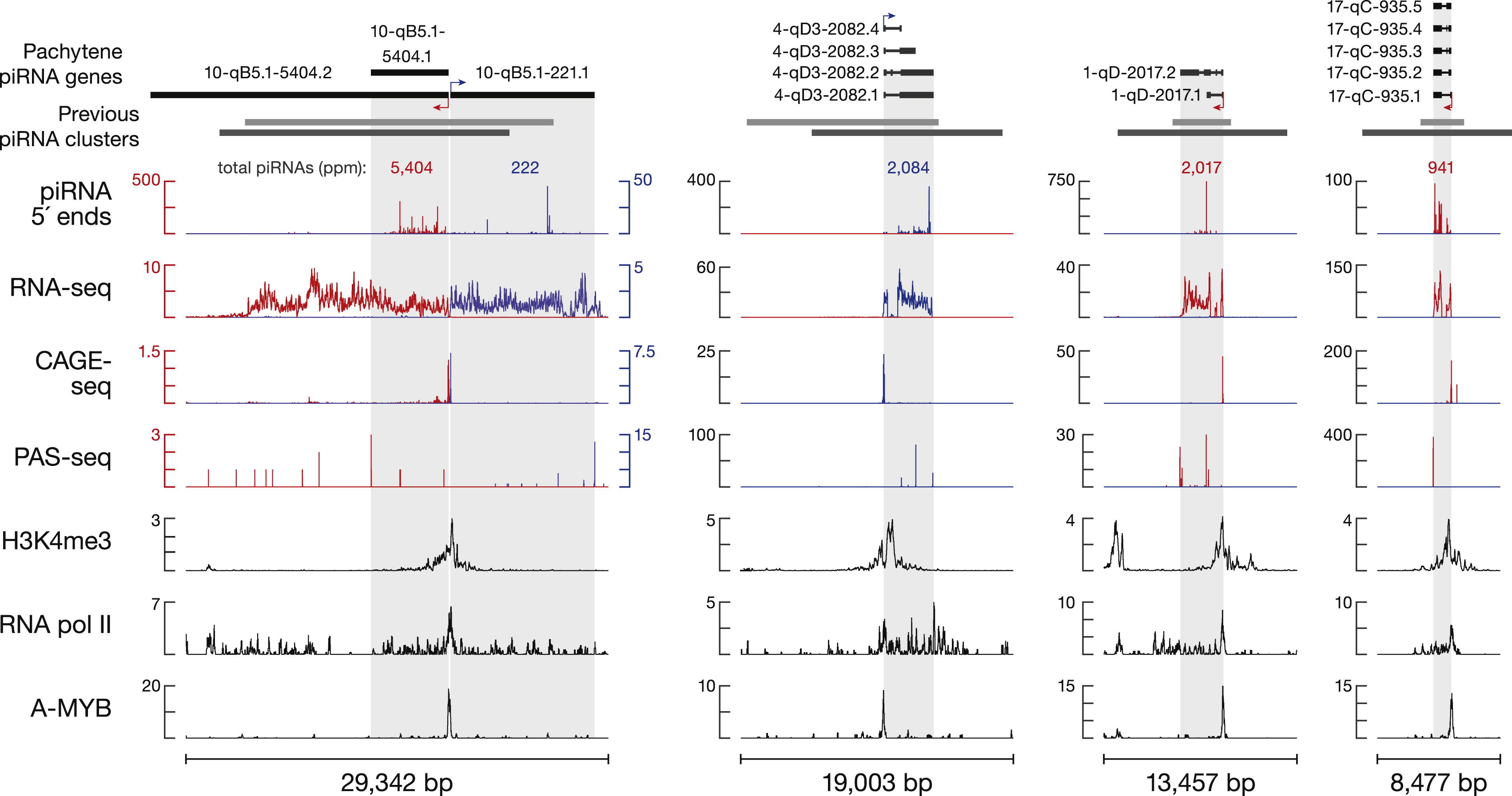


Figure ‑ Examples of Pachytene piRNA Genes

Previous cluster boundaries are from Lau et al. (2006; gray) and Girard et al. (2006; dark gray).

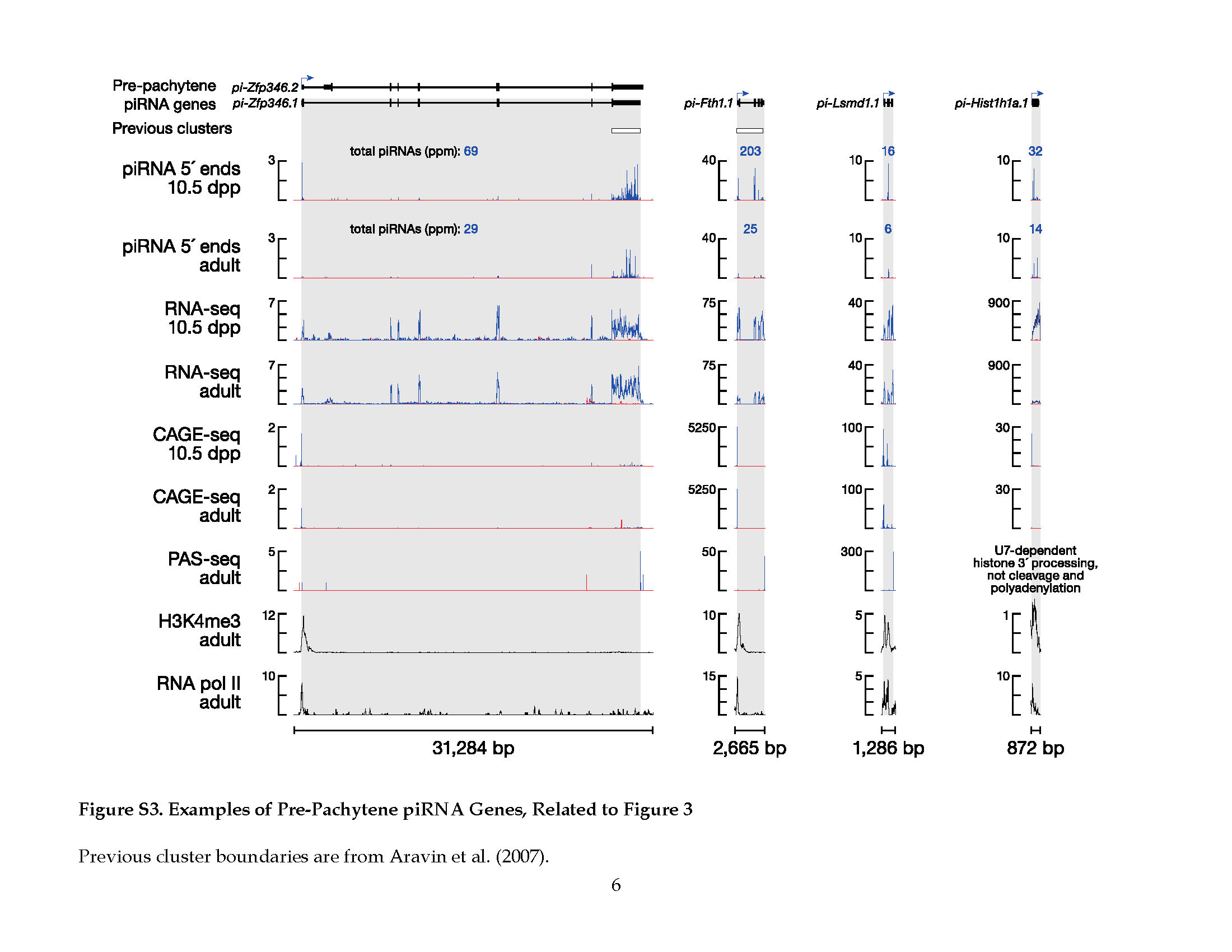


Figure ‑ Examples of Pre-Pachytene piRNA Genes

Previous cluster boundaries are from Lau et al. (2006; gray) and Girard et al. (2006; dark gray).

Hierarchical clustering detected a third class, hybrid piRNAs, which derives from 30 genes with characteristics of both pre-pachytene and pachytene piRNA loci. Like pre-pachytene, hybrid piRNAs were detected at 10.5 dpp (median abundance = 3.7 rpkm) and in purified spermatogonia (Gan et al., 2011). Like pachytene piRNAs, hybrid piRNA abundance increased during the pachytene stage of meiosis, but the increase was delayed until late (17.5 dpp) rather than early pachynema (14.5 dpp). Overall, piRNAs from hybrid genes increased >10-fold from 14.5 to 17.5 dpp. The median abundance of piRNAs from hybrid piRNA genes ranged from 90–120 rpkm in purified pachytene spermatocytes, >20-fold greater than their median abundance in spermatogonia (Gan et al., 2011; Modzelewski et al., 2012). Moreover, hybrid piRNA precursor transcripts were readily detected in purified pachytene spermatocytes (median abundance = 9.0 rpkm; Modzelewski et al., 2012).

##### A-Myb Regulates Pachytene piRNA Precursor Transcription

The coordinated increase in pachytene piRNA precursor transcripts suggests their regulation by a common transcription factor or factors. Among the 100 pachytene piRNA genes, 15 pairs (30 genes) are divergently transcribed. The 5′ ends of the piRNA precursor RNAs from each pair are close in genomic distance (median = 127 bp), suggesting that a shared promoter lies between the two transcription start sites.

We took advantage of the unique genomic organization of these 15 pairs of divergently transcribed piRNA genes to search for sequence motifs common to their promoters. The MEME algorithm (Bailey and Elkan, 1994) revealed a motif highly enriched in these bidirectional promoters (E = 8.3 × 10−12; Figure 4A). This motif matches the binding site of the Myb family of transcription factors (Figure 4A; Gupta et al., 2007; Newburger and Bulyk, 2009). The Myb motif is not restricted to bidirectional promoters; MEME identified the same motif using the promoters of all pachytene piRNA genes (E = 9.1 × 10−28; Figure 4B).

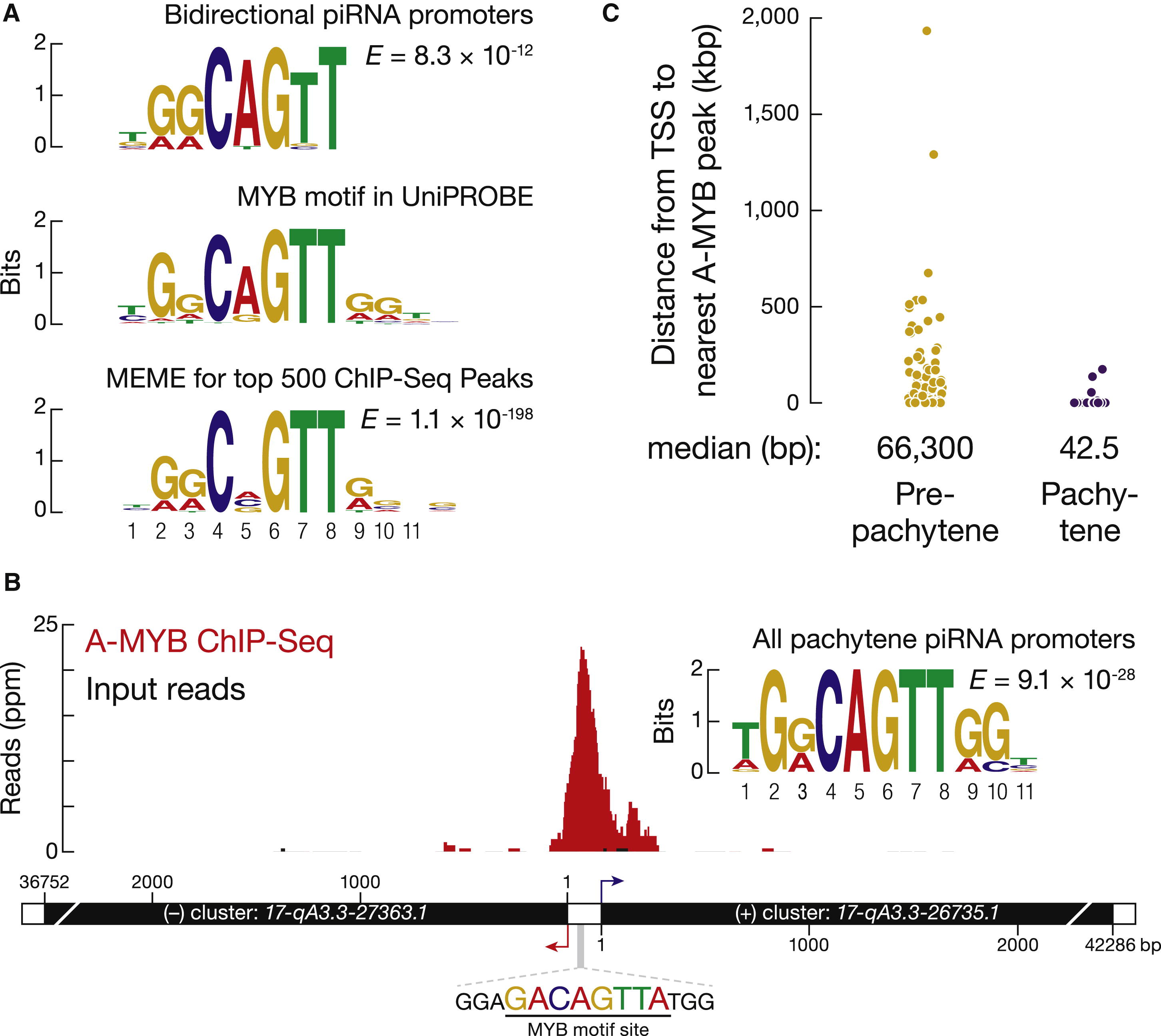


Figure ‑. A-MYB Binds the Promoters of Pachytene piRNA Genes

(A) Top: MEME identified a sequence motif in the bidirectional promoters of the 15 pairs of divergently transcribed pachytene piRNA genes. E value computed by MEME measures the statistical significance of the motif. Middle: Myb motif from the mouse UniPROBE database. Bottom: MEME-reported motif for the top 500 (by peak score) A-MYB ChIP-seq peaks from adult mouse testes.(B) A-MYB ChIP-seq data for the common promoter of the divergently transcribed pachytene piRNA genes 17-qA3.3-27363.1 and 17-qA3.3-26735.1.(C) The distance from the annotated transcription start site (TSS) of each piRNA gene to the nearest A-MYB peak. See also Figure S4.

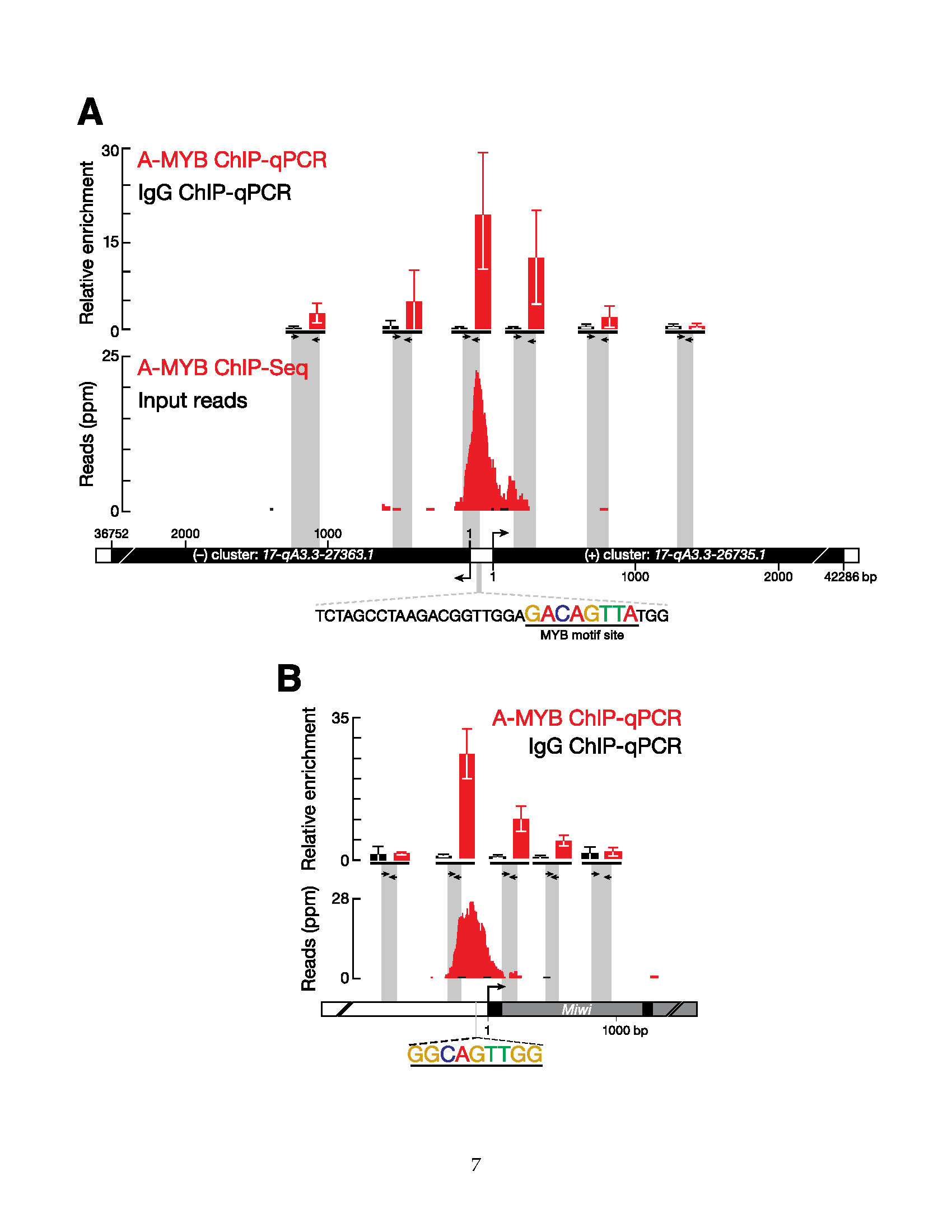


Figure ‑ ChIP-qPCR Confirms ChIP-seq Data

(A) A-MYB binds to the common promoter of divergently transcribed pachytene piRNA loci 17-qA3.3-27363.1 and 17-qA3.3-26735.1. The abundance of DNA fragments at the amplified region relative to a control region (mean ± standard deviation; n = 3) was measured by qPCR (top). The A-MYB ChIP-seq (red) and input (black) data for this pair of genes is presented as in Figure 4B. (B) ChIP-seq and qPCR were as in (A), but for the promoter region of Miwi (Piwil1). Also shown is the RefSeq gene model. Exons, black; introns, gray.

The Myb transcription factor family is conserved among eukaryotes. Like other vertebrates, mice produce three Myb proteins, A-MYB (MYBL1), B-MYB (MYBL2), and C-MYB (MYB), each with a distinct tissue distribution (Mettus et al., 1994; Trauth et al., 1994; Latham et al., 1996; Oh and Reddy, 1999). Testes produce both A- and B-MYB proteins. Multiple lines of evidence implicate A-MYB, rather than B-MYB, as a candidate for regulating pachytene piRNA transcription. First, the expression of A-Myb during spermatogenesis resembles that of pachytene piRNAs: A-Myb transcripts appear at ∼12.5 dpp and peak at 17.5 dpp (Figure 2B; Bolcun-Filas et al., 2011). The expression of A-Myb messenger RNA (mRNA) increases ∼15-fold from 8 dpp to 19 dpp, whereas B-Myb mRNA expression remains constant and low during the same time frame and into adulthood (Horvath et al., 2009). Our RNA-seq data (Figure 2B) corroborate these findings. Indeed, in our RNA-seq analysis of adult testes, A-Myb mRNA was 24-fold more abundant than B-Myb. Second, a testis-specific A-Myb point-mutant allele, Mybl1repro9, which is caused by a cytosine-to-adenine transversion that changes alanine 213 to glutamic acid, leads to meiotic arrest at the pachytene stage with subtle defects in autosome synapsis; A-Myb null mutant mice have defects in multiple tissues, including the testis and the mammary gland (Toscani et al., 1997; Bolcun-Filas et al., 2011). Third, our RNA-seq analysis of A-Myb mutant testes shows that there is no significant change in B-Myb expression in the mutant, compared to the heterozygous controls, at 14.5 or 17.5 dpp. Finally, B-MYB protein is not detectable in pachytene spermatocytes (Horvath et al., 2009).

To assess more directly the role of A-MYB in pachytene piRNA precursor transcription, we used anti-A-MYB antibody to perform ChIP followed by high-throughput sequencing of the A-MYB-bound DNA. The anti-A-MYB antibody is specific for A-MYB, and the peptide used to raise the antibody is not present in B-MYB. The model-based analysis of ChIP-seq (MACS) algorithm (Zhang et al., 2008) reported 3,815 genomic regions with significant A-MYB binding (false discovery rate, FDR < 10−25); we call these regions A-MYB peaks or peaks. Among the 500 peaks with the lowest FDR values, 394 (80%) contained at least one significant site (p < 10−4) for the MYB binding motif (Figure 4A). Figure 4B shows an example of such an A-MYB peak at the bidirectional promoter of the divergently transcribed pair of pachytene piRNA genes 17-qA3.3-27363.1 and 17-qA3.3-26735.1. A-MYB occupancy of this genomic site was confirmed by ChIP and quantitative PCR (ChIP-qPCR) (Figure S4A).

The median distance from the transcription start site to the nearest A-MYB peak was ∼43 bp for the 100 pachytene piRNA genes but >66,000 bp for the 84 pre-pachytene genes (Figure 4C). Our data suggest that during mouse spermatogenesis A-MYB binds to the promoters of both divergently and unidirectionally transcribed pachytene piRNA genes.

To test the idea that A-MYB promotes transcription of pachytene, but not pre-pachytene, piRNA genes, we used RNA-seq to measure the abundance of RNA > 100 nt long from the testes of A-Myb point-mutant (Mybl1repro9) mice and their heterozygous littermates (Figure 5). Pachytene piRNA precursor transcripts—both divergently and unidirectionally transcribed—were significantly depleted in A-Myb mutant testes compared to the heterozygotes: the median decrease was 45-fold at 14.5 dpp (q = 1.1 × 10−13) and 248-fold at 17.5 dpp (q = 3.9 × 10−23). The abundance of pre-pachytene piRNA transcripts was not significantly changed (q ≥ 0.34). The binding of A-MYB to the promoters of pachytene piRNA genes, together with the depletion of pachytene piRNA transcripts in the A-Myb mutant, further supports the view that A-MYB directly regulates transcription of pachytene piRNA genes.

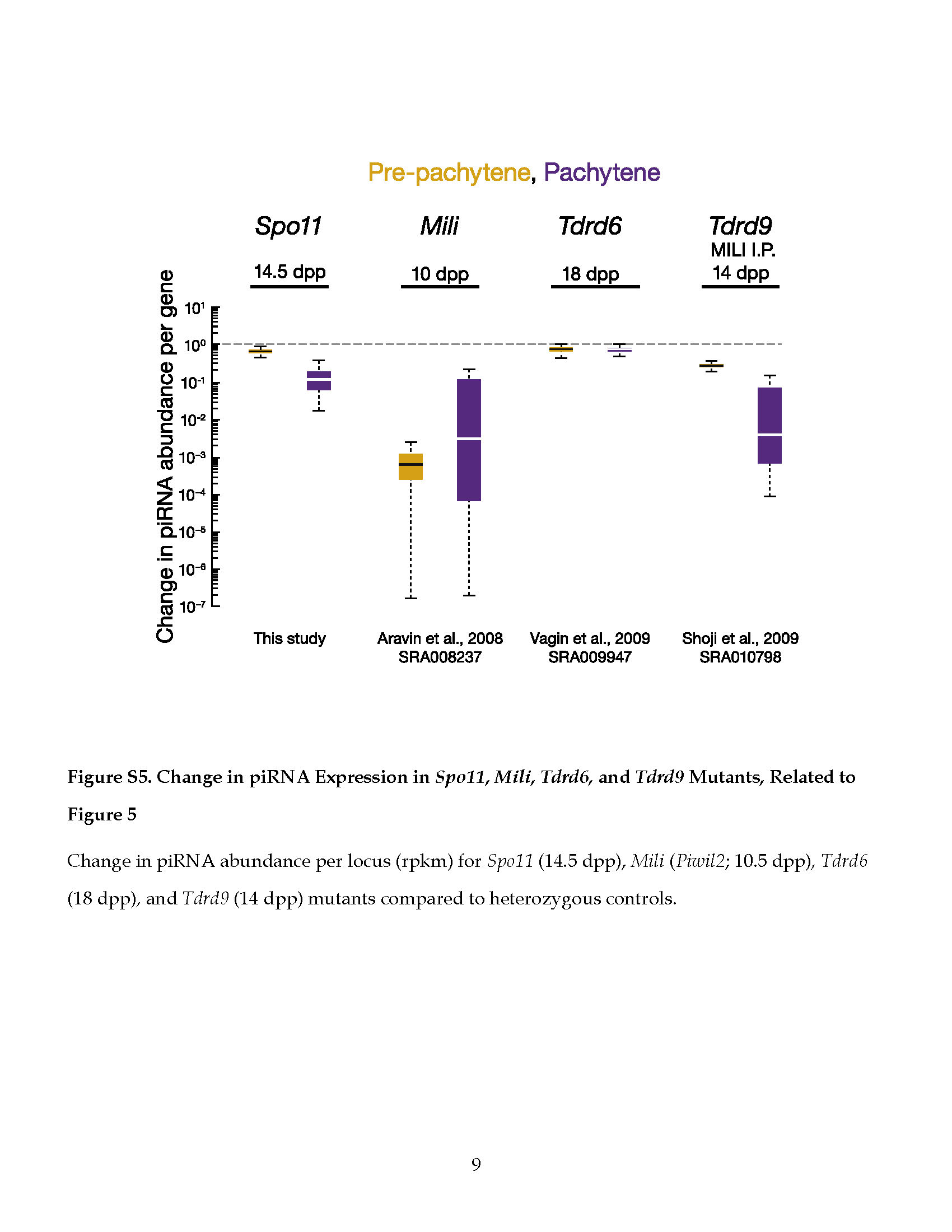
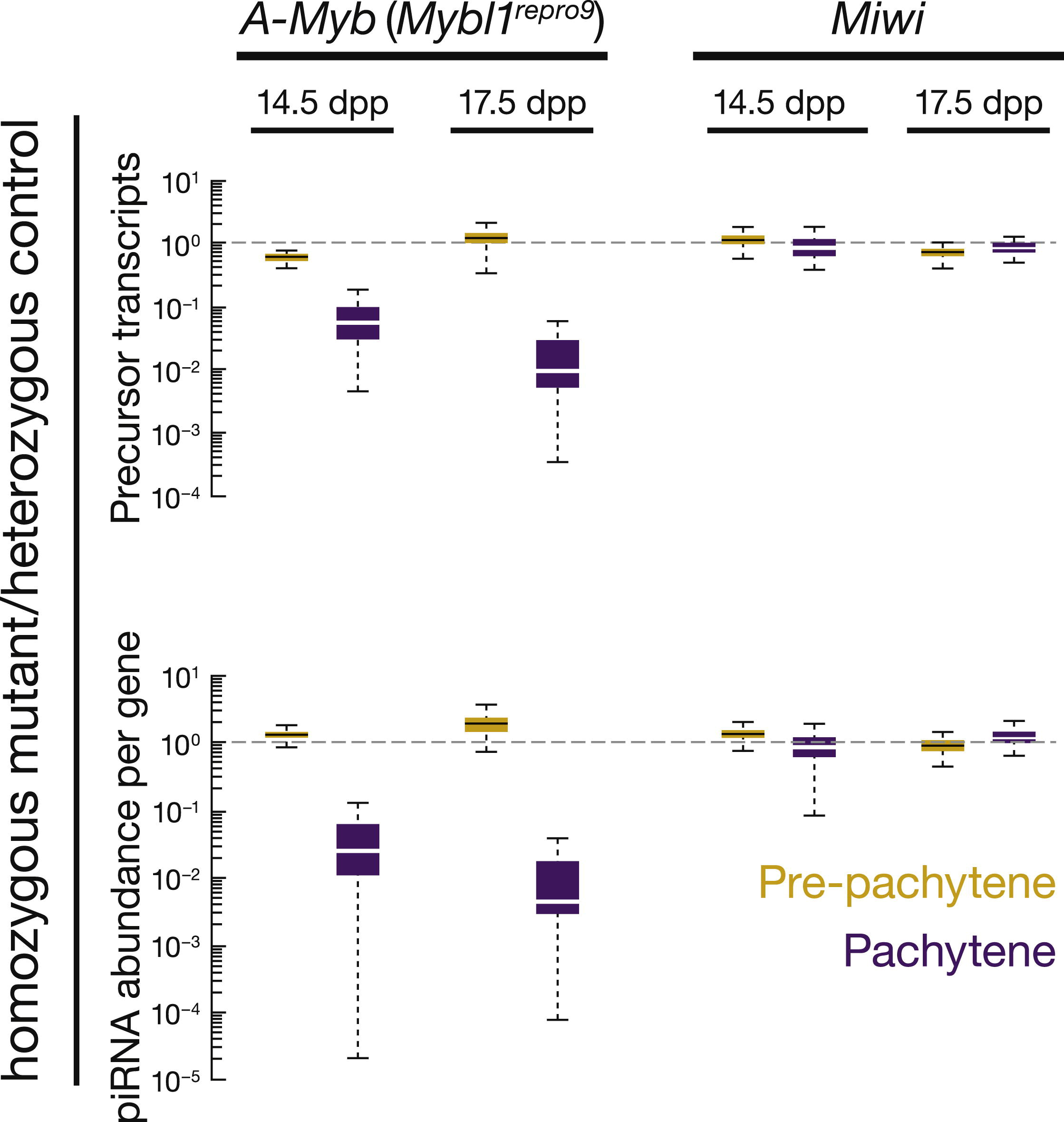


Figure ‑ Change in piRNA Expression in A-Myb, *Spo11*, *Mili, Tdrd6*, and *Tdrd9* Mutants

(Left) The change in transcript or piRNA abundance per gene in A-Myb (n = 3) and Miwi (n = 1) mutants compared to heterozygotes in testes isolated at 14.5 and 17.5 dpp. (Right) Change in piRNA abundance per locus (rpkm) for Spo11 (14.5 dpp), Mili (Piwil2; 10.5 dpp), Tdrd6 (18 dpp), and Tdrd9 (14 dpp) mutants compared to heterozygous controls.

##### A-Myb Regulates Pachytene piRNA Production

To test the consequences of the loss of piRNA precursor transcripts, we measured piRNA abundance in the A-Myb mutant. Like pachytene piRNA precursor transcription, pachytene piRNA abundance significantly decreased in mutant testes. At 14.5 dpp, median piRNA abundance per pachytene gene decreased 87-fold in A-Myb homozygous mutant testes compared to heterozygotes (p < 2.2 × 10−16; Figure 5). By 17.5 dpp, median pachytene piRNA abundance was >9,000 times lower in the A-Myb mutant than the heterozygotes (p < 2.2 × 10−16). In contrast, pre-pachytene piRNA levels were essentially unaltered. Figure 6 presents examples of the effect at 14.5 and 17.5 dpp of the A-Myb mutant on piRNA precursor transcript and mature piRNA abundance for one pre-pachytene and three pachytene piRNA genes.

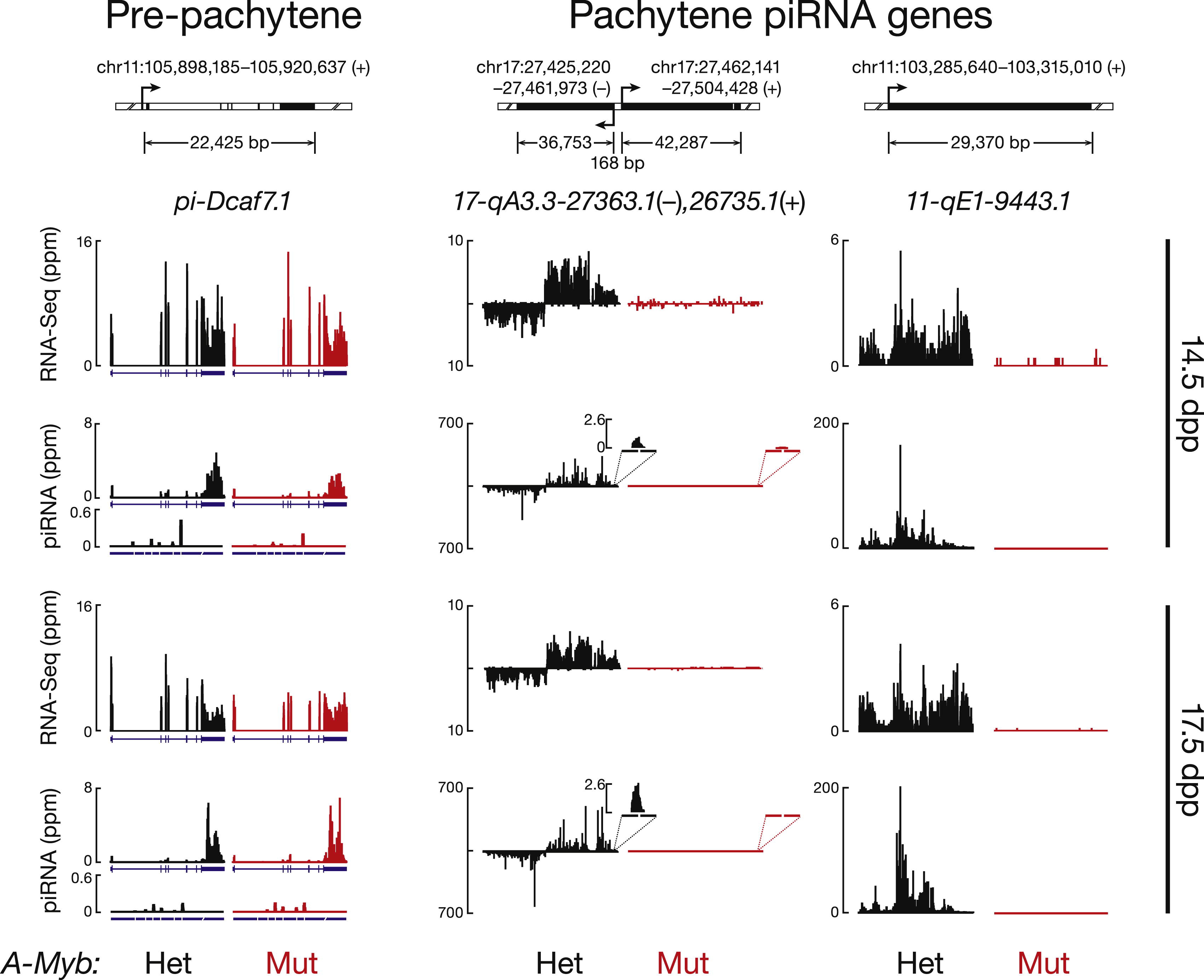


Figure ‑ Examples of the Effect of the A-Myb Mutant on piRNA Expression

Transcript (rpkm) and piRNA (ppm) abundance in heterozygous (Het) and homozygous A-Myb (Mut) point mutant testes is shown for four illustrative examples at 14.5 and 17.5 dpp. Also shown is the abundance of piRNA sequencing reads that map to the exon-exon junctions. Gene 11-qE1-9443 does not have an intron. Exons, blue boxes; splice junctions, gaps; the last exon is compressed and not to scale.

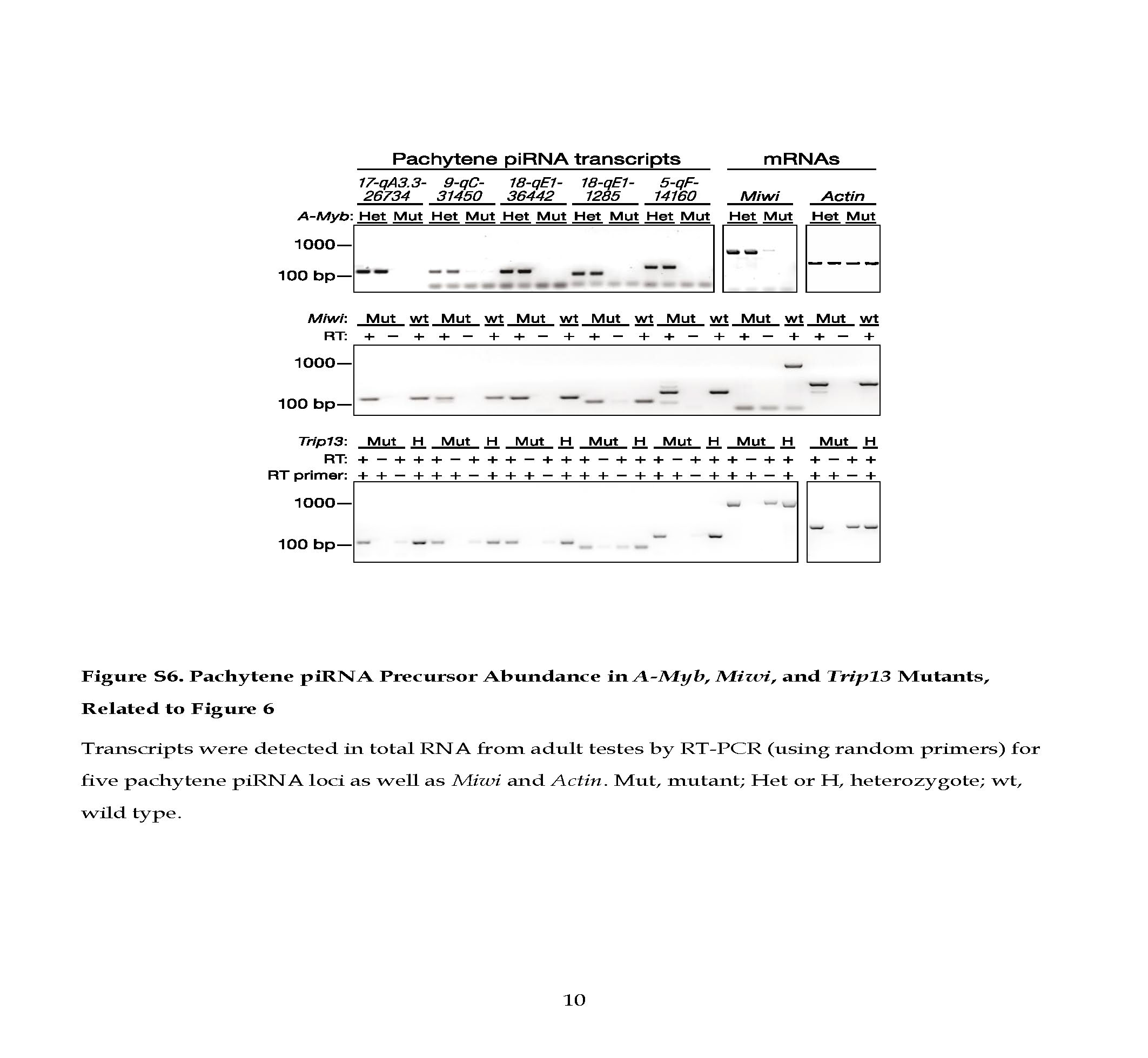


Figure ‑ Pachytene piRNA Precursor Abundance in *A-Myb*, *Miwi*, and *Trip13* mutants

Transcripts were detected in total RNA from adult testes by RT-PCR (using random primers) for five pachytene piRNA loci as well as Miwi and Actin. Mut, mutant; Het or H, heterozygote; wt, wild type.

Our data show that A-MYB binds to the promoters of pachytene piRNA genes; A-Myb, Miwi, and pachytene piRNA precursor transcription begins at 12.5 dpp; and A-Myb mutant spermatocytes reach pachynema with subtle defects in autosome synapsis (Bolcun-Filas et al., 2011). Could pachytene piRNA depletion nonetheless be an indirect consequence of the meiotic arrest caused by the A-Myb mutant? To test this possibility, we sequenced small RNAs from Spo11 mutant testes, which failed to generate double-stranded DNA breaks at the leptotene stage and display a meiotic arrest (Baudat et al., 2000; Romanienko and Camerini-Otero, 2000). The median abundance of piRNAs from pre-pachytene genes did not decrease at 14.5 dpp. By 17.5 dpp, piRNA from pachytene genes decreased just 5.9-fold in the Spo11 mutant testes compared to the heterozygotes (Figure S5). We note that A-MYB protein abundance is reduced in the Spo11 mutant (Bolcun-Filas et al., 2011).

Trip13 is required to complete the repair of double-strand DNA breaks on fully synapsed chromosomes. Trip13 mutants display a meiotic arrest similar to that in A-Myb mutant testes (Li and Schimenti, 2007): pachytene arrest with synapsed chromosomes. To further test whether the loss of pachytene piRNA precursor transcripts in A-Myb mutants reflects a general effect of meiotic arrest, we measured piRNA precursor transcript abundance in Trip13 mutant testes at 17.5 dpp. Unlike A-Myb, piRNA precursor transcripts were readily detectable in the Trip13 mutant (Figure S6). We conclude that the loss of pachytene piRNA precursor transcripts and piRNAs in A-Myb mutant testes is a direct consequence of the requirement for A-MYB to transcribe pachytene piRNA genes and not a general feature of meiotic arrest at the pachytene stage.

##### A-Myb Regulates Expression of piRNA Biogenesis Factors

The A-Myb mutant more strongly affected pachytene piRNA accumulation than it did the steady-state abundance of the corresponding piRNA precursor transcripts (Figure 5); the median decrease in pachytene piRNA abundance was 2-fold greater at 14.5 dpp and 38-fold greater at 17.5 dpp than the decrease in the steady-state abundance of pachytene precursor transcripts (Table S1). These data suggest that A-MYB exerts a layer of control on piRNA accumulation beyond its role in promoting pachytene piRNA precursor transcription.

Miwi has previously been proposed to be a direct target of A-MYB; Miwi mRNA abundance is reduced in A-MYB mutant testes, and ChIP microarray data place A-MYB on the Miwi promoter (Bolcun-Filas et al., 2011). Our RNA-seq data confirm that accumulation of Miwi mRNA requires A-MYB: Miwi mRNA decreased more than 50-fold in testes isolated from A-Myb mutant mice at 14.5 dpp compared to their heterozygous littermates (Figures 7A and S7 and Table S3). Furthermore, our ChIP data confirm that A-MYB binds the Miwi promoter in vivo (Figures 7B, S4B, and S7). Like pachytene piRNAs, Miwi transcripts first appear at 12.5 dpp (Figure 2B), and MIWI protein is first detected in testes at 14.5 dpp (Deng and Lin, 2002). Loss of MIWI arrests spermatogenesis at the round spermatid stage (Deng and Lin, 2002).

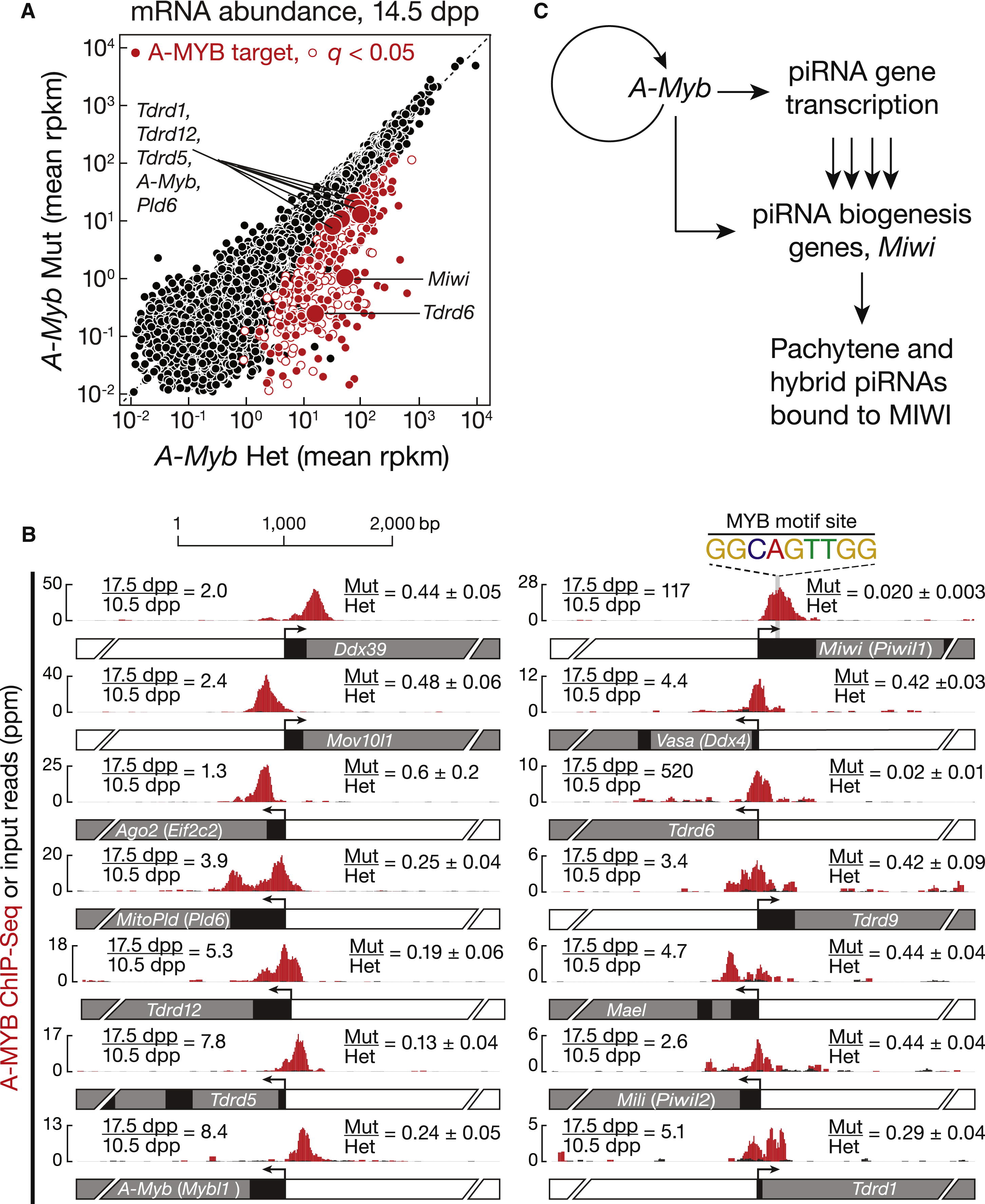


Figure ‑ A-MYB Regulates Expression of mRNAs Encoding piRNA Pathway Proteins

(A) mRNA abundance in A-Myb mutant versus heterozygous testes. The 407 genes with a significant (q < 0.05) change in steady-state mRNA levels are shown as red circles. The 203 with A-MYB peaks within 500 bp of their transcription start site are filled.(B) A-MYB ChIP-seq signal at the transcription start sites of A-Myb and genes implicated in RNA silencing pathways. For each, the figure reports the change in mRNA abundance between 17.5 and 10.5 dpp in wild-type testes and the mean change between A-Myb mutant and heterozygous testes at 14.5 dpp (mean ± SD; n = 3).(C) A model for the regulation of pachytene piRNA biogenesis by A-MYB. See also Figure S7 and Table S3.

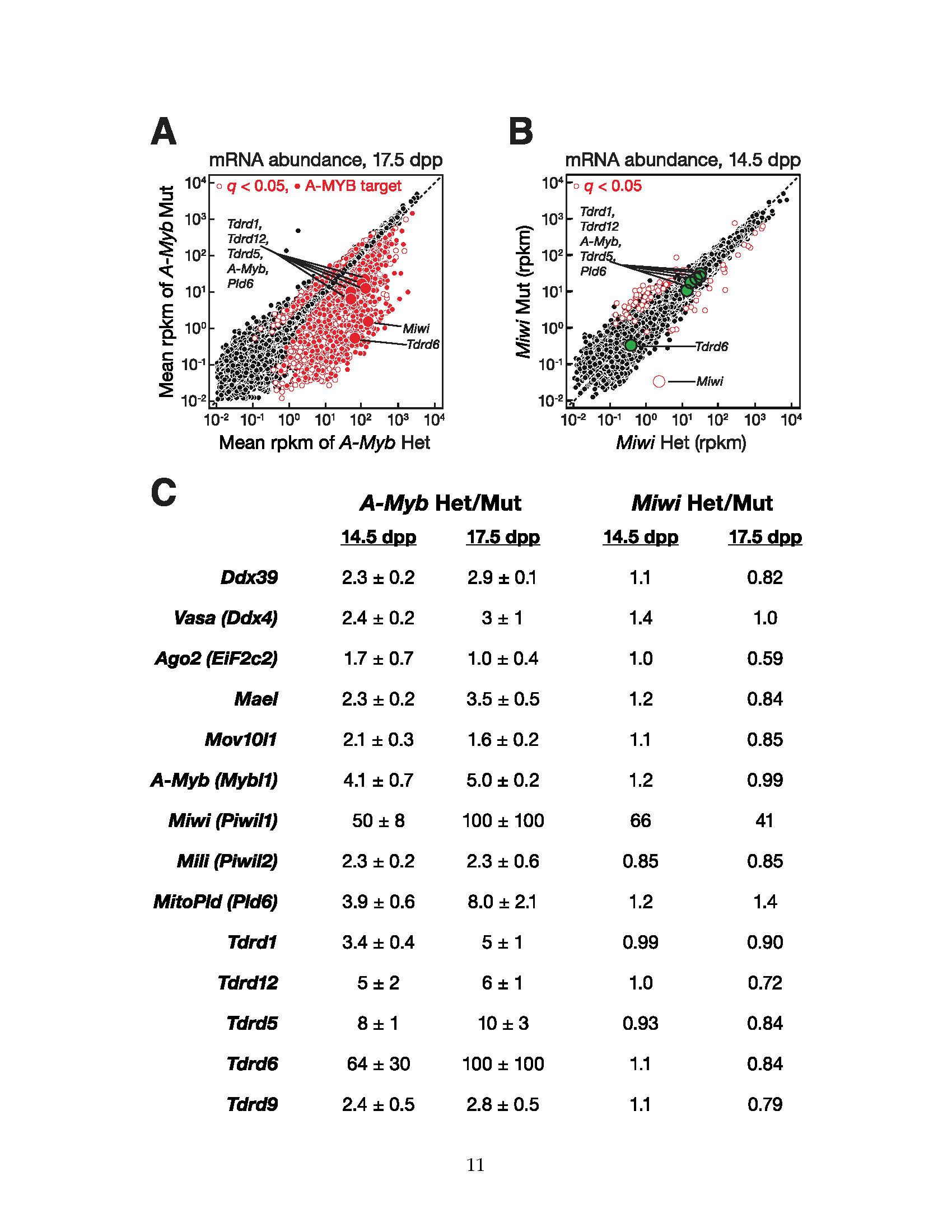


Figure ‑. A-Myb mutants, but Not Miwi Mutants, Change the Expression of RNA Silencing Pathway Genes

A) mRNA abundance in 17.5 dpp A-Myb versus heterozygous testes. The 2,853 genes with a significant (q < 0.051) change in steady-state mRNA abundance are shown as open red circles. Among them, 8721,009 genes also had A-MYB peaks within 500 bp of their transcription start sites. These “A-MYB targets” are marked with filled red circles. (B) Same as (A) but in 14.5 dpp Miwi mutant versus heterozygous testes. The genes encoding proteins implicated in RNA silencing pathways that were labeled in (A) and that showed no change in expression in Miwi mutant testes are highlighted as green filled circles. As expected, Miwi, showed a significant decrease in mRNA abundance in Miwi−/− testes. (C) The change in mRNA abundance (rpkm) in A-Myb and Miwi mutant testes versus heterozygous controls for the RNA silencing genes highlighted in (A) and (B).

A previous study reported that piRNAs fail to accumulate to wild-type levels in Miwi mutant testes (Grivna et al., 2006b). However, our data suggest that the overall change in piRNA abundance caused by loss of MIWI is quite small: RNA-seq detected no change at 14.5 dpp (change in total piRNA abundance = 1.1; n = 2) and only a modest decrease at 17.5 dpp (change in total piRNA abundance = 0.58; n = 1). piRNAs from pachytene loci decreased just 2.7-fold at 14.5 dpp (p = 0.0046) and 3.5-fold at 17.5 dpp (p = 1.8 × 10−6) in Miwi mutant testes (Figure 5). By comparison, pachytene piRNAs declined 87-fold at 14.5 dpp and 9,400-fold at 17.5 dpp in the A-Myb mutant.

Does the loss of MIWI affect piRNA precursor transcription? We measured transcript abundance and piRNA expression in Miwi null mutant testes at 14.5 and 17.5 dpp. In Miwi−/− testes, pachytene piRNA precursor transcripts were present at levels indistinguishable from Miwi heterozygotes (median change = 1.0- to 1.4-fold; q = 1; Figure 5). Thus, loss of MIWI does not explain loss of pachytene piRNA precursor transcripts in A-Myb mutant testes.

In addition to Miwi, ChIP-seq detected A-MYB bound to the promoters of 12 other RNA-silencing-pathway genes (Figure 7B and Table S3). Of these, the mRNA abundance—measured by three biologically independent RNA-seq experiments—of Ago2, Ddx39 (uap56 in flies), Mael, Mili, Mov10l1, Tdrd9, and Vasa did not change significantly at 14.5 dpp in A-Myb mutant testes compared to heterozygotes (q > 0.05); except for Ago2, all decreased significantly in the mutant at 17.5 dpp. In contrast, the abundance of the mRNAs encoding Tudor domain proteins decreased significantly in A-Myb mutant testes: Tdrd6 (64-fold decrease; q = 3.1 × 10−5) and Tdrd5 (7.5-fold decrease; q = 1.0 × 10−5). Tdrd5 is expressed in embryonic testes then decreases around birth (Yabuta et al., 2011). TDRD5 protein reappears at 12 dpp, increasing throughout the pachynema (Smith et al., 2004; Yabuta et al., 2011). Our data indicate that A-MYB activates Tdrd5 transcription at the onset of the pachytene stage of meiosis. Similarly, Tdrd6 mRNA can be detected at the middle pachytene, but not the zygotene stage, and peaks after late pachytene; TDRD6 protein can be detected at 17 dpp and continues to increase until 21 dpp (Vasileva et al., 2009). The findings that TDRD5 and TDRD6 colocalize with MIWI in pachytene spermatocytes (Hosokawa et al., 2007; Vasileva et al., 2009; Yabuta et al., 2011) and that TDRD6 binds MIWI (Chen et al., 2009; Vagin et al., 2009; Vasileva et al., 2009) suggest a role for these Tudor domain proteins in pachytene piRNA production or function. As in Miwi−/− testes, spermatogenesis arrests at the round spermatid stage in Tdrd5−/− and Tdrd6−/− mutant testes (Vasileva et al., 2009; Yabuta et al., 2011). Loss of Tdrd6 expression has little effect on piRNA levels (Figure S3; Vagin et al., 2009), perhaps because the functions of Tudor domain proteins overlap.

Other genes encoding piRNA pathway proteins whose promoters are bound by A-MYB and whose expression decreased significantly in A-Myb mutant testes include MitoPld (Pld6; 3.9-fold decrease; q = 0.0095) and Tdrd12 (5.3-fold decrease; q = 0.0046). MitoPld encodes an endoribonuclease implicated in an early step in piRNA biogenesis in mice and flies (Houwing et al., 2007; Pane et al., 2007; Haase et al., 2010; Huang et al., 2011; Watanabe et al., 2011; Ipsaro et al., 2012; Nishimasu et al., 2012). The function of Tdrd12 is not known, but its fly homologs (Yb, Brother of Yb, and Sister of Yb) are all required for piRNA production (Handler et al., 2011). Tdrd1 decreased 3.4-fold, but with q value = 0.015. Tdrd1 is first expressed in fetal prospermatogonia, then re-expressed in pachytene spermatocytes (Chuma et al., 2006). In Tdrd1 mutant testes, spermatogenesis fails, with no spermatocytes progressing past the round spermatid stage (Chuma et al., 2006). TDRD1 binds MILI and MIWI (Chen et al., 2009; Kojima et al., 2009) and colocalizes with TDRD5 and TDRD6 in the chromatoid body (Hosokawa et al., 2007).

Together, these data support the idea that at the onset of the pachytene phase of meiosis, A-MYB coordinately activates transcription of many genes encoding piRNA pathway proteins.

##### A-MYB and the Pachytene piRNA Regulatory Circuitry

A number of genes encoding known and suspected piRNA pathway proteins are bound and regulated by A-MYB (Figures 7B and S7C). Our data support a model in which A-MYB drives both the transcription of pachytene piRNA genes and the mRNAs encoding genes required for piRNA production including Miwi, MitoPld, and Tdrd9. Regulation by A-MYB of both the sources of pachytene piRNAs and the piRNA biogenesis machinery creates a coherent feedforward loop (Figure 7C). Feedforward loops amplify initiating signals to increase target gene expression. Furthermore, they function as switches that are sensitive to sustained signals; they reject transient signals (Shen-Orr et al., 2002; Osella et al., 2011).

A-MYB also bound to the A-Myb promoter (Figure 7B), and A-Myb transcripts decreased 4.2-fold in testes from an A-Myb point mutant (Mybl1repro9; Figure 7B). The A-Myb mutant fails to produce the high level of A-MYB protein observed in wild-type testes at the late pachytene stage of meiosis (Bolcun-Filas et al., 2011). Instead, A-MYB protein never becomes more abundant than the level achieved in wild-type testes by the beginning of the pachytene stage. While the lower level of A-MYB in the A-Myb mutant may reflect instability of the mutant protein, a simpler explanation is that mutant A-MYB cannot activate A-Myb transcription.

##### Feed-Forward Regulation of piRNA Production is Evolutionarily Conserved

Is A-MYB-mediated, feedforward control a general feature of regulation of piRNA production among vertebrates? To test whether A-MYB control of piRNA precursor transcription is evolutionarily conserved, we used high-throughput sequencing to identify piRNAs in adult rooster testes. Birds and mammals diverged 330 million years ago (Benton and Donoghue, 2007). After removing the sequences of identifiable miRNAs (Burnside et al., 2008) and annotated noncoding RNAs, total small RNA from the adult rooster testis showed peaks at both 23 and 25 nt (Figure 8A). When the RNA was oxidized before being prepared for sequencing, only a single 25 nt peak remained, consistent with the 25 nt small RNAs corresponding to piRNAs containing 2′-O-methyl-modified 3′ termini. These longer, oxidation-resistant species typically began with uracil (62% of species and 65% of reads; Figure 8B), and we detected a significant Ping-Pong amplification signature (Z score = 31; Figure 8C). We conclude that the oxidation-resistant, 24–30 nt long small RNAs correspond to rooster piRNAs. Like piRNAs generally, rooster piRNAs are diverse, with 5,742,529 species present among 81,121,893 genome-mapping reads. Like mouse pachytene piRNAs, 70% of piRNAs from adult rooster testes mapped to unannotated intergenic regions, 19% mapped to transposons, and 14% mapped to protein-coding genes. Of the piRNAs that map to protein-coding genes, >95% derive from introns. Forty-two percent of piRNA species mapped uniquely to the Gallus gallus genome.

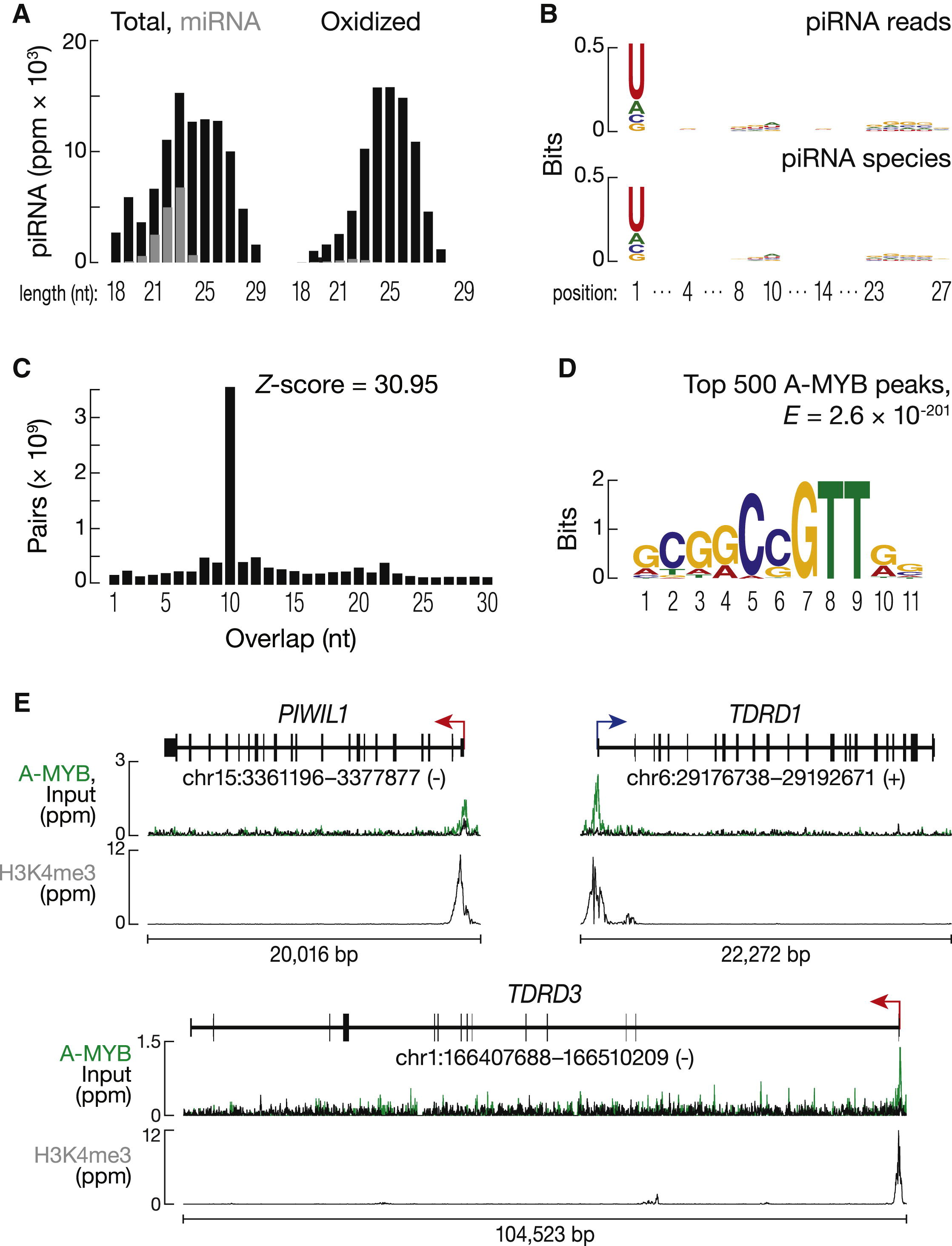


Figure ‑ Feed-Forward Regulation of piRNA Biogenesis by A-MYB is Conserved in Rooster

(A) Length distributions of total rooster testis small RNAs (black) and miRNAs (gray).(B) Sequence logo showing the nucleotide composition of piRNA reads and species.(C) The 5′-5′ overlap between piRNAs from opposite strands was analyzed to determine if rooster piRNAs display Ping-Pong amplification. The number of pairs of piRNA reads at each position is reported. Z score indicates that a significant 10 nt overlap (Ping-Pong) was detected. Z score > 1.96 corresponds to p value < 0.05.(D) MEME-reported motif of the top 500 (by peak score) A-MYB ChIP-seq peaks from adult rooster testes.(E) A-MYB, H3K4me3, and input ChIP-seq signals at the transcription start sites of rooster PIWIL1, TDRD1, and TDRD3. See also Figure S8.

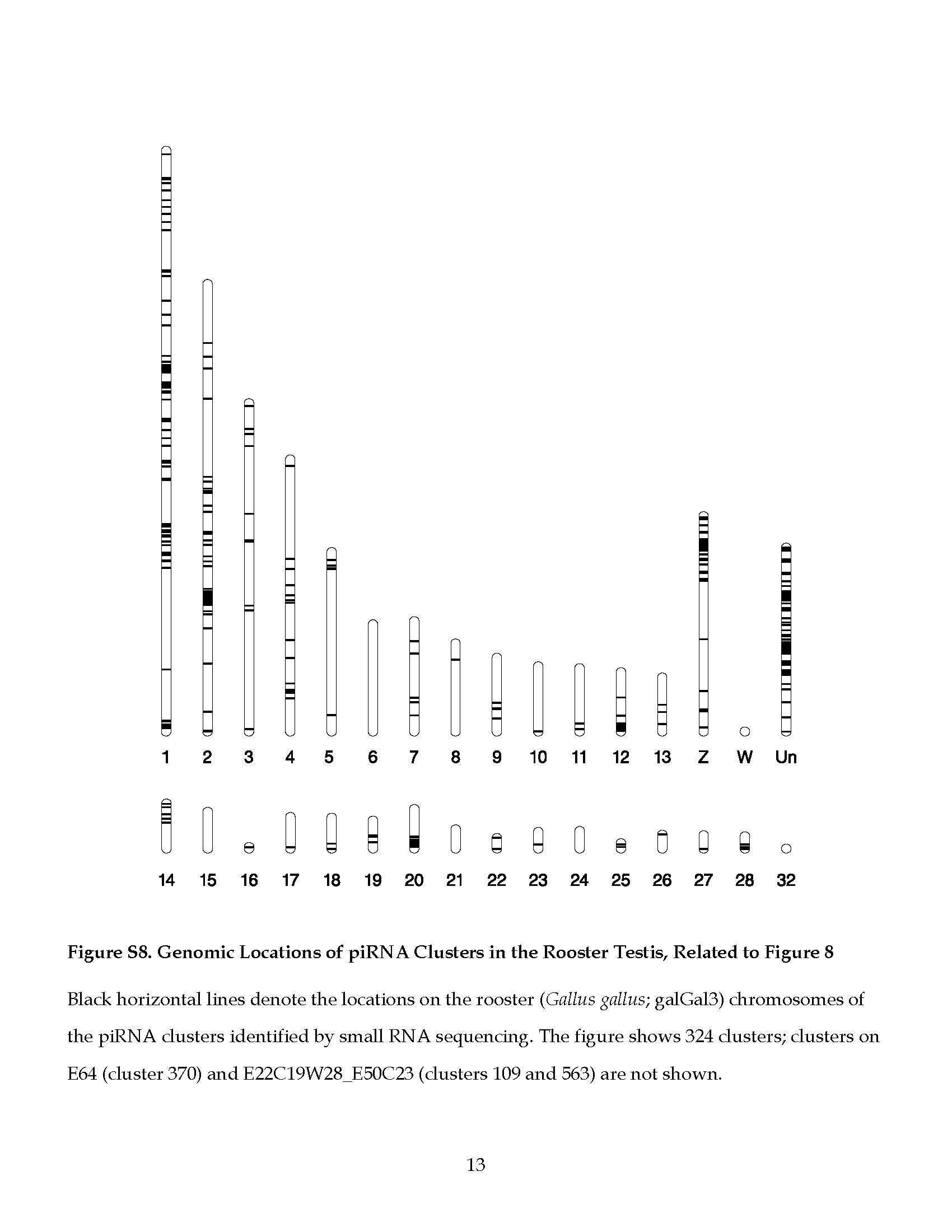


Figure ‑. Genomic Locations of piRNA Clusters in the Rooster (*Gallus gallus)* Testis.

Black horizontal lines denote the locations on the Gallus gallus (galGal3) chromosomes of the piRNA clusters identified by small RNA sequencing. The figure shows 324 clusters; clusters on E64 (cluster 370) and E22C19W28\_E50C23 (clusters 109 and 563) are not shown.

Using 24–30 nt piRNAs from oxidized libraries, we identified 327 rooster piRNA clusters (Figure S8). These account for 76% of all uniquely mapping piRNAs. Of the 327 clusters, 25 overlapped with protein-coding genes. To begin to identify the transcription start sites for the rooster piRNA clusters, we analyzed adult rooster testes by H3K4me3 ChIP-seq. More than 81% (268 out of 327) of the clusters contained a readily detectable H3K4me3 peak within 1 kbp of the piRNA cluster. In contrast, the median distance from a cluster to the nearest transcription start site of an annotated gene was 73 kbp, suggesting that the H3K4me3 peaks reflect the start sites for rooster piRNA precursor transcripts.

Next, we asked where in the genome A-MYB bound in adult rooster testes. A-MYB ChIP-seq identified 5,509 significant peaks (FDR < 10−25). MEME analysis of the top 500 peaks with the lowest FDR values identified a motif (E = 2.6 × 10−201; Figure 8D) similar to that found in the mouse (Figure 4A). A-MYB is the only one of the three chicken MYB genes expressed in adult testis (X.Z.L. and P.D.Z., unpublished data), supporting the view that these peaks correspond to A-MYB binding. The core sequence motif associated with A-MYB binding in mouse differs at one position (CAGTT) from that in rooster (C C/G GTT). This difference between mammalian and chicken MYB proteins has been noted previously (Weston, 1992; Deng et al., 1996).

To determine whether chicken A-MYB might regulate transcription of some piRNA clusters in the testis, we compared the A-MYB peak nearest to each piRNA cluster with the nearest H3K4me3 peak. Of the 327 rooster piRNA clusters, at least 104 were occupied by A-MYB at their promoters, as defined by an overlapping H3K4me3 peak. These 104 clusters account for 31% of uniquely mapping rooster piRNAs.

The chicken genome encodes at least two PIWI proteins: PIWIL1 and PIWIL2. Remarkably, the promoter of Gallus gallus PIWIL1, the homolog of mouse Miwi, contained a prominent A-MYB peak (Figure 8E). TDRD1 and TDRD3 also showed A-MYB peaks (Figure 8E). Thus, as in mice, Gallus gallus A-MYB controls the transcription of both piRNA clusters and genes encoding piRNA pathway proteins. We conclude that A-MYB-mediated feedforward regulation of piRNA production was likely present in the last common ancestor of birds and mammals.

In mice, we found no piRNA-producing genes on the sex chromosomes (Figure S1A), perhaps because mouse sex chromosomes are silenced during the pachytene stage (Li et al., 2009b). Birds use a ZW rather than an XY mechanism for sex determination, so roosters are homogametic (ZZ), allowing the sex chromosomes to remain transcriptionally active in males (Namekawa and Lee, 2009; Schoenmakers et al., 2009). Indeed, we find that 39 of the 327 rooster piRNA clusters are on the Z chromosome, accounting for 12% of uniquely mapping piRNAs (Figure S8). Of the 39 Z chromosome clusters, 18 had an A-MYB peak at their promoter.

#### DISCUSSION

The data presented here provide strong support for the view that piRNAs in mammals begin as long, single-stranded precursors generated by testis-specific, RNA Pol II transcription of individual piRNA genes (see also Vourekas et al., 2012). Transcription by RNA Pol II affords piRNA genes the same rich set of transcriptional controls available to regulate mRNA expression. Our data establish that developmentally regulated transcription of piRNA genes determines when specific classes of piRNAs emerge during spermatogenesis.

During mouse spermatogenesis, transcription of pachytene piRNA genes begins at the onset of the pachytene stage of meiosis; pachytene piRNAs accumulate subsequently. The presence of the MYB binding motif near the transcription start sites of pachytene piRNA genes, the physical binding of A-MYB to those genes, and the loss of pachytene piRNA precursor transcripts and piRNAs in testes from A-Myb mutant mice all argue that A-MYB regulates pachytene piRNA production.

A-MYB also drives increased expression of piRNA pathway genes. Among these, Miwi expression shows the greatest dependence on A-MYB, but A-MYB also drives transcription of genes encoding other proteins in the piRNA pathway, including MitoPld, Mael, and five genes encoding Tudor domain proteins. For example, A-MYB increases expression of Tdrd6 more than 500-fold. Loss of A-MYB function more strongly depletes pachytene piRNAs than loss of MIWI, in part because pachytene piRNAs can still be loaded into MILI in Miwi mutant testes, although MILI-loaded pachytene piRNAs do not suffice to produce functional sperm. In the A-Myb mutant, expression of mRNAs encoding multiple piRNA pathway proteins decreases. We speculate that in wild-type male mice, the increased expression of these mRNAs at the onset of the pachytene stage of meiosis ensures that sufficient piRNA-precursor-processing and MIWI-loading factors are available to cope with the large increase in pachytene piRNA precursor transcription.

We propose that induction of A-MYB during the early pachytene stage of spermatogenesis initiates a feedforward loop that ensures the precisely timed production of these piRNAs. Coherent feedforward loops show delayed kinetics in order to reject background stimuli (Mangan and Alon, 2003). Indeed, we observed a delay from the early to middle pachytene in the accumulation of pachytene piRNAs, despite the continued increase in A-Myb expression (Figure 2A). Pachytene piRNA levels increase 75-fold (median for the 100 genes) from 10.5 to 12.5 dpp, coincident with increased expression of A-Myb. However, from 12.5 to 14.5 dpp, pachytene piRNAs increase only 1.2-fold. Pachytene piRNAs subsequently resume their accumulation, increasing 65-fold from 14.5 to 17.5 dpp. We believe this delay is a consequence of a feedforward loop that ensures the production of pachytene piRNAs only at the pachytene stage of spermatogenesis. Regulation by a feedforward loop also predicts a rapid shutdown of pachytene piRNA pathways at round spermatid stage VIII, when A-MYB protein levels decrease (Horvath et al., 2009). Supporting this idea, the abundance of MIWI decreases sharply by the elongated spermatid stage of spermatogenesis (Deng and Lin, 2002). Testing this proposal is a clear challenge for the future.

In fruit flies and zebrafish (Brennecke et al., 2007; Houwing et al., 2007), most piRNAs map to repetitive regions, whereas in mammals, uniquely mapping intergenic piRNAs predominate in the adult testis. The discovery that 70% of rooster piRNA reads map to intergenic regions suggests that the expansion of intergenic piRNAs controlled by A-MYB feedforward regulation arose before the divergence of birds and mammals. In the future, detailed analysis of piRNA production across avian spermatogenesis should provide insight into the evolutionary origins and functions of pachytene piRNAs, a class of piRNAs thus far only detected in mammals.

In summary, we have shown that mouse piRNA genes are coregulated transcriptionally, establishing that A-MYB coordinately regulates the biogenesis of an entire piRNA class, the pachytene piRNAs. The discovery that a loss-of-function A-Myb mutant, Mybl1repro9, disrupts piRNA precursor transcription in vertebrates provides a tool to understand the transformation of long, single-stranded piRNA precursors into mature piRNAs and to explore the functions and targets of the pachytene piRNAs.

#### EXPERIMENTAL PROCEDURES

###### Mice

Mybl1repro9, Spo11tm1Sky, and Piwil1tm1Hf mice were maintained and used according to the guidelines of the Institutional Animal Care and Use Committee of the University of Massachusetts Medical School and genotyped as described (Baudat et al., 2000; Deng and Lin, 2002; Bolcun-Filas et al., 2011).

###### Sequencing

Small (Ghildiyal et al., 2008; Seitz et al., 2008) and long RNA-seq (Zhang et al., 2012) and analysis (Li et al., 2009a) were as described. Reads that did not map to mouse genome mm9 were mapped to piRNA precursor transcripts to obtain splice junction mapping small RNAs. Total small RNA libraries from different developmental stages and from mutants were normalized to the sum of all miRNA hairpin mapping reads. Oxidized samples were calibrated to the corresponding total small RNA library via the abundance of shared, uniquely mapped piRNA species. piRNA expression data were grouped with Cluster 3.0. Differential gene expression was analyzed with DESeq R (Anders and Huber, 2010); ChIP-seq reads were aligned to the genome using Bowtie version 0.12.7 (Langmead et al., 2009), and peaks were identified using MACS (Zhang et al., 2008).

##### Acknowledgments

We thank K. Chase and K. Schimenti for help collecting tissues; C. Tipping for help with mouse husbandry; P. Johnson and B. Keagle for providing rooster testes; G. Farley for technical assistance; H. Lin for reagents; Xi Chen, Xiaotu Ma, Oliver Rando, and Benjamin Carone for advice on ChIP; and members of our laboratories for critical comments on the manuscript. X.Z.L. was supported by the Lalor Foundation and the Jane Coffin Childs Memorial Fund for Medical Research.

##### Accession Numbers

The Gene Expression Omnibus (GEO) accession number for the RNA-seq, ChIP-seq, and small RNA data reported in this paper is [GSE44690](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44690).

## Perspective/Final Summary & Conclusions

Things to remember about this section:

1. Multiple people have stated that this can be highly speculative
2. This is my chance to show that I am a scientist
3. If I were to die tomorrow, what would be the experiments to conduct

* Discuss long range RNA secondary structure and implecations of regulating alternative splicing (See Li, S., & Breaker, R. R. (2013). Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. Nucleic acids research, 41(5), 3022–31. doi:10.1093/nar/gkt057) and Reg of AS by long RANGE SS folder in Mendeley
* Full length analysis of mRNAs
* Implications for discrination past one's DNA as it is the actual PRODUCT of the DNA and the actual biology (or at least closer to the functional biology) that is going on inside of every person
* Catalogue of every possible mRNA isoform
* Seperate the signal from the noise of transcription
* Most high cited pnas papers are methodology papers [www.pnas.org/reports/most-cited](http://www.pnas.org/reports/most-cited)
* Importantance of bioinformatics for every new and current molecular biogist
  + Talk about my path during PhD
  + Mention BG’s paper New Insights from Existing Sequence Data: Generating Breakthroughs without a Pipette AM Plocik, BR Graveley Molecular cell 49 (4), 605-617
  + Discuss importance of BigData and being able to analyze in parallel, looking for genome wide changes

! Unanswered questions concerning mammalian piRNAs

Interaction w/ ribosomes?

How does cell partition prepachytene transcripts to mRNA or piRNA use?

# Back Matter

## Appendices

### Methods

#### SeqZip

#### Computational Scripts

#### Computational Approaches

##### Git

##### Perl

##### IGV

##### MySQL

#### ImageQuant

#### LOD of Radioactivity

### Oligo Database

## Biblography