"Putting the Pieces Together: Exons and piRNAs"

A Dissertation Presented

By

Christian K. Roy

Submitted to the Faculty of the University of the
Massachusetts Graduate School of Biomedical Sciences,
Worcester
in partial filfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

MAY 21st 2014

BIOCHEMISTRY

"Putting the Pieces Together: Exons and piRNAs"

A Dissertation Presented

By

Christian K. Roy

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

---

Melissa J. Moore, Co-Thesis Advisor

Phillip D. Zamore, Co-Thesis Advisor

Scot Wolfe, Member of Committee

Job Dekker, Member of Committee

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

---

Zhiping Weng, 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.

Anthony Carruthers, Ph.D.,

Dean of the Graduate School of Biomedical Sciences

Biochemistry and Molecular Pharmacology

MAY 21st 2014

UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL

Abstract

BIOCHEMISTRY AND MOLECULAR PHARMACOLOGY

Doctor of Philosophy

"Putting the Pieces Together: Exons and piRNAs"

by Christian K. ROY

Analysis of gene expression has undergone a technical revolution. What was impossible 6 years ago is now routine. The application of high-throughput DNA sequencing machines capable of generating hundreds of millions of reads allows, indeed forces, a major revision in how we think about the genome's functional output–the transcriptome.

This thesis examines the history of DNA sequencing, the application of sequencing to the measurement of gene expression, and some of the more interesting features of transcripts, including Alternative Splicing and mammalian piRNA biogenesis. Examination of these topics is framed around development of a novel RNA-templated DNA:DNA ligation assay (SeqZip). SeqZip allows for a more efficient analysis of the abundant and function long RNA molecules present in the transcriptome.

The thesis closes with a discussion on future applications of the SeqZip methodology, when major advancements in high-throughput sequencing are a certainty. Finally, challenges facing biomedical researches in the new climate of extremely large and rich datasets are discussed.

Acknowledgements

First I would like to acknowledge and thank Laura Geagan and Rebecca Sendak at Genzyme. They were my supervisors while I was a research associate there, and assisted and encouraged my transition back to graduate school. Without the confidence they instilled in my abilities as a young scientist in, I doubt I would have ever signed up for more school.

Next I'd like to thank Melissa. During my 1st year retreat at Wood's Hole I first learned that Melissa is a fantastic communicator of interesting and important science. When she brought out her rope representing the unspliced pre-mRNA of dystrophin—a rope that reached to the back of a rather large auditorium—and then dramatically held up a No.2 pencil representing the final mRNA product, both to scale, I knew the that I wanted to do my graduate research in her lab. I have never once doubted the decision to join Melissa's lab, and have learned so much from the broad and interconnected approach she takes to important scientific questions. Thank you so much for teaching me to always consider the big picture, go for the answer, and to just ask when I need help.

Soon after joining Melissa's lab, and a project going well, it was proposed to me that I be a joint student between Melissa and Phil. It was not difficult to not jump at the opportunity to be advised by two Howard Hughes Investigators, and I also haven't regretted the decision. Over the past few years, I have been continually amazed at the depth of Phil's knowledge, in scientific and general topics. He is a careful, meticulous, quantitative, and calculating mentor. While I feel that I clicked 'on the level' with Melissa, interacting with Phil forced me to think and act outside my comfort zone, something I always tell myself is a critical aspect of change and growth. Thank you Phil for everything I've learned.

My committee has also been very supportive throughput my PhD. I hardly believed the ease with which I passed my qualifying exam, and took it as a big confidence boost. The following years of TRAC meetings confirmed that I was not thinking way off-base. The one-on-one meetings just prior to my QE were especially helpful. Thanks to Scot, Job, and especially Zhiping for all the guidance.

- Lab Members; Aaron; Alper; Amrit
- Eric and Erin
- Collaborators
- Dave Weaver
- Muro

- Graveley
- Heinrich
- Anna
- Ogo
- Family
- Jul Owen

Contents

List of Figures

List of Tables

Todo list

List of Abreviations

AS	Alternative Splicing
DNA	Deoxyribonucleic acid
ssDNA	Single-stranded DNA
RNA	Ribonucleic acid
ssRNA	Single-stranded RNA
ATP	Adenosine triphosphate
NAD	Nicotinamide adenine dinucleotide
ChIP-Seq	Chromatin Immunoprecipitation followed by sequencing
EST	Expressed Sequence Tag
HTS	High-throughput sequencing (see also NGS)
NGS	Next-generation sequencing
CAP	$5'$ 7meG structure attached to mRNAs
nt	A nucleotide of either DNA or RNA
bp	A base pair of DNA
SRE	Splicing Regulatory Element
IRE	Intron Recognition Element
CNS	Central Nervous System
TSS	{Transcription or Translation} Start Site
TTS	{Transcription or Translation} Termination Site
SAGE	Serial Analysis of Gene Expression
hnRNP	heterogeneous nuclear ribonucleoprotein
FISH	Fluorescence in situ hybridization
GFP	Green Fluorescencent Protein

List of Symbols

5′	The 5 prime end of a DNA or RNA molecule
3′	The 5 prime end of a DNA or RNA molecule
μ	Micro. A value of $1x10^{-6}$ standard units
ρ	Pearson product-moment correlation coefficient

Definitions

RNA-Seq

A technology wherein RNA is fragmented, converted to DNA, and analyzed on a high-throughput sequencing instrument

A 'Read'

The sequence of nucleotides produced from each spot on a high-throughput sequencing machine

Insert

The RNA molecule captured between two cloning sequences in a high-throughput sequencing library preparation workflow

Read length

The number of nucleotides for each given "read"

Read depth

The number of reads obtained from each high-throughput sequencing analysis

Coverage

A measure of the number of times each nt of a genome is sequenced. E.g. 100 million reads of a 10 million nt genome = 10X coverage, assuming uniform distribution of the "reads"

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 et al. [1995]).

Scaffold or contig

A draft sequence of nucleotides, meant to represent the actual biological

sequence as closely as possible, examples include unassembled fragments of chromosomes or fragments of mRNA transcripts.

Argonaute

Protein(s) belonging to a group containing a Piwi (P-element induced wimpy testes) domain, that bind nucleic acids and participate in many target-guided processes, including RNA Interference, and RNA-indicuded transcript/gene silencing.

Ligamer

A DNA oligonucleotide containing two distinct regions of complementarity to a 5′ and 3′ section of RNA. Each region is normalized for ^{Tm} such that the length of each section is ~15–30 nt. These two regions are connected by a short sequence of the designer's choice, usually >5 nt in length. Each ligamers overall length is ~45–60 nt. See figures 2.1 and **??**.

I would like to dedicate this Doctoral dissertation to my grandfather, George Knauf. My grandfather passed away on September 23rd, 2011, just one week shy of his 82nd birthday. I find it difficult to articulate how much I miss him. He spoke carefully and never without purpose or conviction. While I hear from others that he was proud of me, he rarely, if ever, betrayed that type of emotion directly. It is my goal to build as solid a life as he, founded on hard work, playing the long game, responsibility, and maintaining friendships. These are just a few of the personality traits that I observed and try to emulate. The fact that he passed before he could meet our son Owen is one of my biggest regrets. Of all the possessions he left behind, it is the memory of our time together that I will cherish the most. Rest in peace, Grump. I did it.

Preface

The work reported in this dissertation has been published in the following articles. Chapter 4 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 3 are included in a currently submitted manuscript.

Chapter 1

Introduction

1.1 Fixed Genomes and Flexible Genes

Exodus tells of the liberation of the Israelites from Egyptian slavery. Humble and reluctant Moses, their divine-appointed leader, attempts to force the Pharaoh Ramses to release the Israelites through inflicting 10 plagues. Pharaoh is stalwart and stubborn as water turns to blood, streets are flooded with frogs, lice, and flies. As livestock falls dead from disease, people and animals both are covered in boils, and the land burns in storms of fire, Pharaoh does not bend.

The 8th plague was a swarm of Locusts, described in Exodus 10: 14–15:

> [14] *And the locusts went up over all the land of Egypt, and rested in all the coasts of Egypt: very grievous were they; before them there were no such locusts as they, neither after them shall be such.*
> [15] *For they covered the face of the whole earth, so that the land was darkened; and they did eat every herb of the land, and all the fruit of the trees which the hail had left: and there remained not any green thing in the trees, or in the herbs of the field, through all the land of Egypt.*

The desolation of a locust plague was still not enough to persuade Ramses. Nor three days of darkness. Only the death of all first-born Egyptians, included Ramses own son, was enough to persuade Pharaoh to liberate the Israelites.

Locust swarms are not biblical fantasy. It is perhaps the most *believable* of the 10 plagues. Today the United Nations' (UN) Food and Agriculture division maintains a Locust watch website that provides weekly updates on potential

swarms in northern Africa and the Middle East. Locusts have long been, and continue to be, a powerful and feared force of Nature.

Unlike fire and brimstone, locusts are something that can be observed and studied. What triggers a swam? We know that the desert locust, *Schistocerca gregaria*, is the one of 10 locust species that swarm and cause massive crop damage. *Schistocerca gregaria* are in the insect Order Orthoptera, along with crickets and katydids. Orthoptern members make sound known as *stridulation* by vigorously rubbing their wings, making for a noisy cloud of devastation.

A Desert Locust only weighs 0.05–0.07 ounces, and is less than 2.5 inches long. They consume their own body weight in vegetation per day. One swarm of the infamous, and now curiously extinct, Rocky Mountain locust contained 12.2 trillion insects. Its estimated total weight was 27.5 million tons. The swarm covered almost 200 square miles (2/3 the size of Manhattan), and could travel 60 miles in a day. A locust swarm is truly a modern biblical plague.

By definition swarms are temporary; the movement, en masse, from one location to another. Where do 12.2 trillion locusts go when not swarming? Does anyone care if their crops aren't under assault? It seemed no one cared enough, until about 1921, when an important realization was made.

The power and destruction *Schistocerca gregaria* can inflict makes it difficult to believe that they are nothing more than common grasshoppers. Nothing more than grasshoppers not just by analogy, but by actual *Taxonomy*. "Desert locusts" are actually the *gregarious* form of *Schistocerca gregaria* (See Figure 1.1), while the more familiar and docile looking "grasshopper" is the *solitary form*. How does such a dichotomy to exist within the same organism, indeed the same *genome*?

Schistocerca gregaria are *polyphenic*, meaning that they have multiple (poly) physical forms (phenotypes). Polyphenism is a general feature among insects. Phenotypes are often extremely different. For example, pea aphids (*Acyrthosiphon pisum*), which usually exist in an asexually reproducing, wingless female form, respond to reduced food supply overcrowding by producing winged offspring. Winged organisms travel to new sources of food and revert back to the asexually reproducing form [Purandare et al., 2014, Shingleton et al., 2003]. In the case of *Schistocerca gregaria*, the gregarious form is smaller and more brightly colored compared to its solitary cousins. This transformation can happen in as little as two hours. What is the underlying cause of this transformation?

In 2009, Anstey et al. [2009] reported that in just two hours of forced crowding, *Schistocerca gregaria* displayed elevated levels of the neurotransmitter serotonin in the ganglia (brain). These levels were strongly correlated gregarious

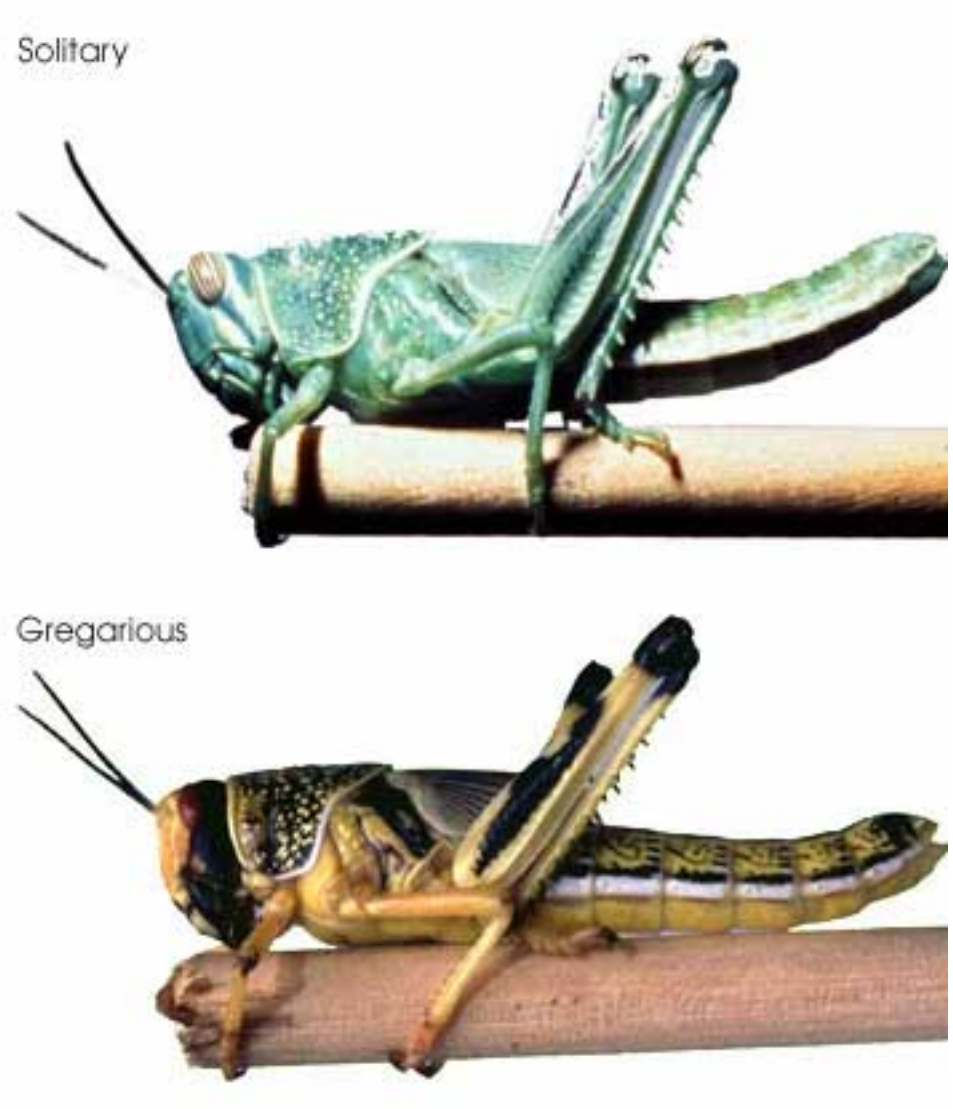


FIGURE 1.1: The Solitary and Gregarious forms of *Schistocerca gregaria*

The two phenotypic forms of *Schistocerca gregaria* appear very different. The Solitary form is green and generally larger, while its gregarious form is more brightly colored, smaller, and capable of swarming in vast numbers, destroying crops and vegetation. Photo from Wikicommons.

form indicators. Serotonin is an important molecule in regulating neuronal junctions and wiring in the brain [Hoeffer et al., 2003]. Through the integration of environmental and social cues, the brain of the insect was being re-wired, resulting in tremendous changes in behavior and phenotype. These changes prepare the organism to deal with a different world in order for the species to survive, but to the detriment of surrounding agriculture.

In an extremely interesting article, David Dobbs compares the two forms of *Schistocerca gregaria* to that of Dr Jekyll and Mr. Hyde, the principle characters of Robert Louis Stevenson novella. For Dr Jekyll in fiction, and for *Schistocerca gregaria* in reality, the power to morph into multiple forms demonstrates the incredible power of a fixed genome yet plastic gene expression. It is often said that something is "in the genes." This thesis will illustrate that another oft-head idiom is more important: "it's how you use them."

Before we could understand how genes were used, we needed to identify them, and understand what they were.

1.2 Nucleic Acid Sequencing

1.2.1 DNA Sequencing History

Soon after it was realized that DNA is the source of genetic information in all living organisms [Watson and Crick, 1953], and the *pretty* and *elegant* arrangement of complementary, antiparrallel, DNA strands was known [Watson et al., 2012], the ability to determine specific arrangements of nucleotide bases (i.e. to sequence) 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 to be able to determine the specific arrangement of nucleotides in the polymer—to sequence it. By 1977, two completely different methods developed by Sanger [Sanger and Coulson, 1975, Sanger et al., 1977] and Maxam-Gilbert [Maxam and Gilbert, 1992] 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 piece of DNA (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 cataloged into knowledge.

During the late 1970's and throughout the 1980's, DNA sequences were typically communicated in important publications [Bell et al., 1980, Sanger et al.,

1978]. The birth of the Internet in the 1990's made essential publically-funded repositories for sequence information easily available [Benson et al., 2011]. However, it was the human genome project [Lander, 2011, Venter et al., 2001], that provided the important activation energy that brought DNA sequencing from a hard-to-perform, but necessary, analysis, to an organized large-scale effort of assembling the complete genetic material complex genomes. An often criticized, but undeniably disrupting force in the human genome project was the competing efforts of the privately-owned company Celera [Venter, 2007]. Taking a higher-throughput and centralized approach to determining the sequence of the human genome, Celera fundamentally changed the landscape of genome assembly. 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 [Staden, 1979], sequenced pairwise [Roach et al., 1995], and combined with sequence scaffolds made available by the publicly-funded project, Celera was able to assemble high-quality genomic sequences very quickly. Arguably, this was the first deep sequencing effort, and changed the landscape of molecular and biochemical research, coincident with the beginning of a new millennium.

1.2.2 History of High-throughput Sequencing

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 and clonally amplified, typically using bacteria. Given that the human genome [Consortium, 2004] comprises >3 billion bp, and each Sanger reaction provides 800 nt of quality sequence, at least 4 million individual reactions are needed to determine the sequence of the human genome. This also assumes all of reads are of sufficient quality, length, and do not overlap by even 1 nt. No overlap is out of the question, as overlaps are essential for assembling individual sequences via their overlaps, allowing assembly through repeated sequences.

Even the best practical improvements to work-flows could not bring Sanger DNA sequencing in-line with aspirations of analyzing genomes of multiple species and/or organisms. The early 2000's saw multiple efforts to improve the scale of DNA sequencing, first using MPSS [Brenner et al., 2000], but perhaps more importantly, by Pyro- [Ronaghi et al., 1998] and Polony sequencing [Shendure et al., 2005]. Both pyro- and polonysequencing utilize emulsion PCR [Nakano

et al., 2003] for clonal amplification prior to sequencing, removing the bottleneck of bacterial cloning. In contrast to Sanger sequencing, where the signal is from fluorescence of the last incorporated chain-terminating nucleotide, pyrosequencing visualizes light given off by luciferase reacting with pyrophosphate (PPi), a by-product of nucleotide incorporation. This approach was later commercialized by 454 technologies. Polony sequencing involves a more complicated sequencing-by-ligation method, eventually commercialized by Applied Biosystems and branded as SOLiD sequencing. While both of these technologies provided valuable, high-throughput sequences, neither has been as successful as the approach commercialized by Solexa, now known as Illumina.

Illumina-produced sequencers use a sequencing-by-synthesis approach. After clonal amplification of DNA on a slide surface [Bentley et al., 2008], fluorescent nucleotides are visualized as they are incorporated into the growing DNA strand. Since 2006, iterations of the Illumina platform (eg. GE, GE-II(x), Hi-Seq, Hi-Seq 2500, Hi X) have demonstrated a steady and impressive increases in both read depth and length. On February 15th 2012, Illumina announced on its Basespace blog, that they had sequenced a HapMap sample at 40X coverage, using the HiSeq 2500 platform and paired-end 100 nt reads in a single run. On Janurary 14th, 2014, Illumina announced its HiSeq X system, the first platform to truly attain produce the $1,000 genome. These machines can demonstrate that sequencing genomes is no longer the monumental endeavor it once was, and that completely new experimental possibilities are a reality for life science research.

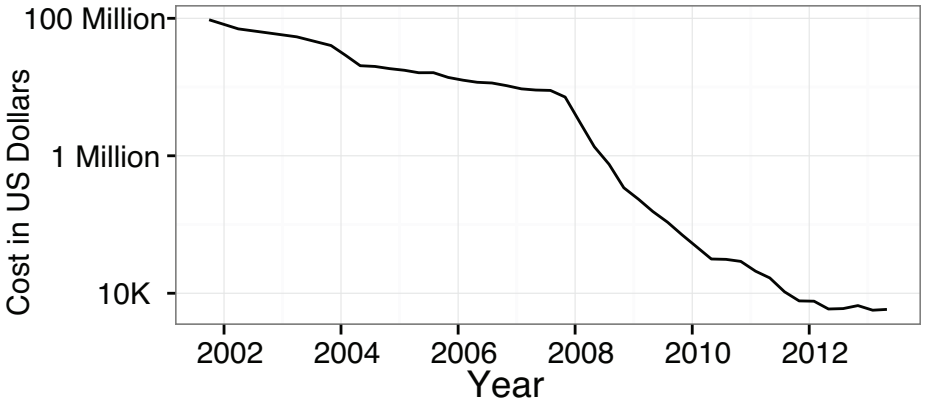


FIGURE 1.2: Cost of sequencing the human genome over time

The costs of sequencing the human genome has decreased on a log scale over a roughly 10 year period thanks to major improvements in high-throughput sequencing. Data from Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 2013-09-03).

1.2.3 Deep-sequencing of RNA

The first widely-accepted large scale method used to measured gene expression was Serial Analysis of Gene Expression (SAGE) [Velculescu et al., 1995]. While the importance of microarrays in the measurement of gene expression via cannot be overstated [Marioni et al., 2008, Shendure and Ji, 2008], the limited ability of microarrays to investigate novel sequences combined with their analogue signal, makes their relevance to this section off-topic. Yet, SAGE, (like the before mentioned MPSS technique) produces a digital output of gene expression using a cleaver procedure of restriction endonucleolytic cleavage of cDNA molecules. Cleaved product sticky ends are ligated using sticky ends left from digestion and concatenated together to form longer DNA fragments. Fragments are cloned into a vector, amplified, and Sanger sequenced. Using sequences incorporated during concatenation, the number of sequenced "fragments" that align to a given gene is related to the abundance of the original RNA molecule. While SAGE was a clever molecular trick allowing researches to dip into the 5-log range of expression typically seen in mRNA expression, it is still limited by Sanger sequencing read length and depth.

The Solexa/Illumina platform relies on clonal amplification of a single template directly on a slide surface. Imaging these spots with sensitive digital cameras after sequential addition of fluorescent nucleotides (*sequencing by synthesis*) turned out to be the right mix for a "second generation" HTS platform. Not long after the Solexa/Illumina platform achieved read lengths of sufficient length of depth necessary to measure gene expression were the first RNA-Seq papers published [Lister et al., 2008, Mortazavi et al., 2008, Nagalakshmi et al., 2008]. These papers gave a powerful glimpse into the future of molecular biology. Indeed, in the years since, analysis by RNA-Seq has quickly overtaken 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 digital quantification of RNA expression across physiologically-relevant ranges [Blencowe et al., 2009]. While simultaneously measuring gene expression, the data can be used for novel sequence discovery, measuring RNA-editing [Li et al., 2011], and transcript assembly [Trapnell et al., 2010]. By modifying the basic protocol or performing additional biochemical steps, RNA-Seq can be used to investigate many aspects of RNA biology (see 1.3 and [Mutz et al., 2013]).

RNA processing begins the moment the nascent RNA is exposed from the polymerase exit channel. Many methodologies have been developed that enrich RNA-Seq libraries for particular types of RNA. For example, measurement of nascent transcripts can be performed via GRO-Seq [Core et al., 2008], and the

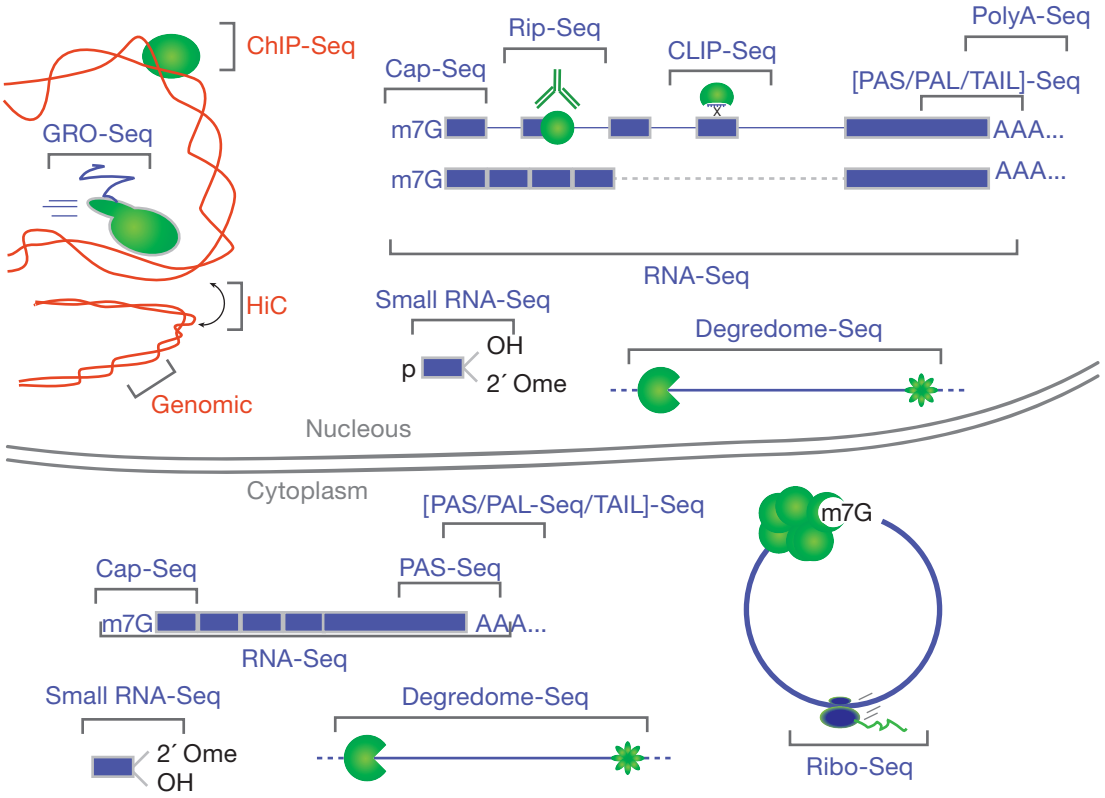


FIGURE 1.3: Methods for High-throughput sequencing of RNA

In the short years since the first report of RNA-Seq, many variations have been reported. The figure above provides an incomplete graphical illustration of some of these variations. A more complete list of *Seq applications is maintained on this blog.

extremely complicated process of RNA turnover (referring to the rates at which RNAs both are produced and degraded) has been examined [Ghosh and Jacobson, 2010, Tani et al., 2012]. RNA::Protein interactions can be measured with or without cross-linking the protein to the RNA, via CLIP or RIP, respectively. Once an RNA has been fully transcribed, known processing steps such as 5′ 7meG (CAP) formation and poly(A) tail formation can be measured using any of the Cap-Seq/CAGE methodologies [Shiraki et al., 2003], or PAS/TAIL/-PAL [Chang et al., 2014, Shepard et al., 2011, Subtelny et al., 2014]. 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 of traditional mRNAs.

RNA-Seq and its associated flavors are also traditionally associated quantifying RNA obtained from *many* tissue culture cells bulk pieces of a particular tissues.

Recently, efforts to measure the RNA expression occurring in individual cells has gained attention [Shapiro et al., 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 Heisenberg's uncertainty principle [Kennard, 1927] and is often confused with the more appropriate "Observer effect" [Riley and Steitz, 2013]. Leaving that issue aside, measuring the unique transcriptome of a given cell among cells of a common tissue is surly an exciting and informative endeavor [Marinov et al., 2013, Shalek et al., 2013, Wills et al., 2013]. Compared to DNA, the diversity of RNA synthesis within living cells is potentially much more complicated [Shendure and Aiden, 2012], 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].

1.3 Nucleic Acid Splicing

1.3.1 Alternative Splicing

Insert information on RNASeq looking for splicing!

[**?**]

[Barbosa-Morais et al., 2012, Merkin et al., 2012]

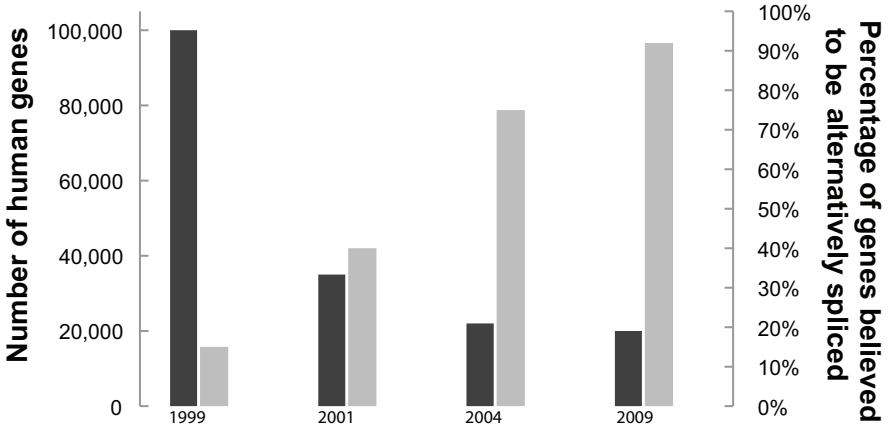


FIGURE 1.4: Dark Grey - Estimates of number of human genes; Light Grey - Estimates of what percent of genes undergo some form of alternative splicing.

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. Phillip Sharp, Co-Nobel-prize winner for the discovery of splicing, stated that: "Approximately, one of every twenty genes is expressed by alternative pathways of RNA splicing in different cell types or growth states" Sharp [2014]. 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 and 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 used RNA-Seq to demonstrate that between 86% and 95% of human multi-exon genes are subject to AS [Pan et al., 2008, Sultan et al., 2008, 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.

1.3.2 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 [Wang et al., 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 and 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 and Lynch, 2008].

Sequence motifs that compose the AS code have been teased out [Barash et al., 2010, Ladd and 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 et al., 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 et al., 2009, 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.

1.3.3 The Isoform Problem

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. 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 (See Figure 1.5). 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

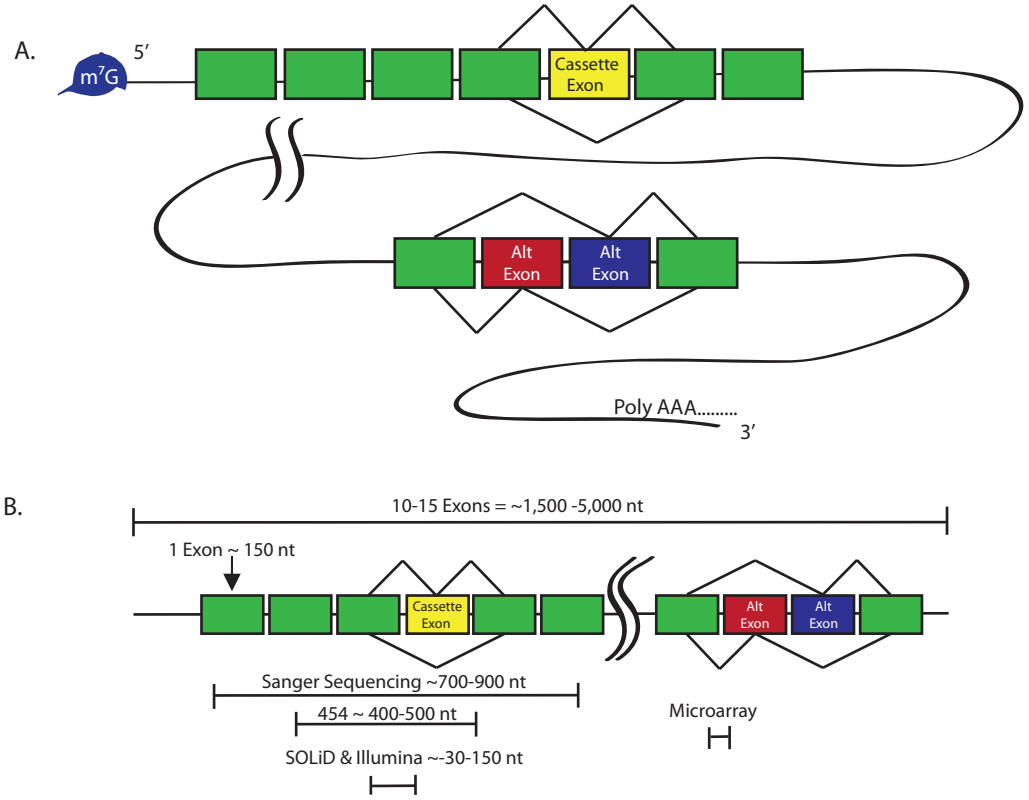


FIGURE 1.5: HTS read lengths are not sufficient to maintain AS connectivity

A) Long RNAs may have multiple sites of AS, separated by 1000's of nt; B) Most mRNAs have 10 exons of 150 nt each. Some have many more (and longer) exons. Read lengths of current sequencing technologies do not maintain connectivity between distant sites.

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, has hastened the displacement of microarray by RNA-Seq as the preferred method for comprehensive analysis of gene expression [Shendure and Ji, 2008].

Many researchers are turning toward 2nd generation HTS methodologies for comprehensive transcriptome analysis. HTS allows for *de novo* identification of isoforms, over a larger dynamic range, in a quantitative fashion [Mortazavi et al., 2008]. Additionally, techniques exist to 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 1.5) 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 [Calarco et al., 2007a, Emerick et al., 2007, Zhu et al., 2003]. 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.

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 (>800 nt). One observation that may lead to the identification of long-range SREs is intramolecular coordination between distal splicing decisions. In Figure 1.5 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.

1.3.4 Coordination in splicing

The provocative “Miller Spread” showing spliceosomes associated with RNA transcripts likely ignited the first thought that transcription and splicing are intracately linked [Osheim et al., 1985]. Twelve years later, the first observation that polymerase speed can affect downstream splicing decisions was reported

[Cramer et al., 1997], spawning a new field of research into linked transcription and splicing. But how linked are these two processes?

How would linkage between transcription and splicing manifest? One of the clearest examples is mouse Fibronectin *Fn1* (Figure 1.6) [Schwarzbauer et al., 1983, White and Muro, 2011]. 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 homology 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). [Fededa et al., 2005] 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., 2007] 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.

Need to transition better here. Logic isn't great!

There have been a few studies that investigate forms of splicing coordination between adjacent exons present in mRNA. The vertebrate genes 4.1B and 4.1R, members of the protein 4.1 family encoding for cytoskeletal adaptor proteins, both undergo splicing of upstream 5′ first exons to distal 3′ second exons, skipping a stronger proximal 3′ second exon [Parra et al., 2012, 2008]. This is accomplished through "intrasplicing" involving an intronic sequence element (the "intraexon") only present when transcription begins at the upstream 5′ exon, allowing the exon to ligate do the weaker distal 3′ second exon via an intermediate splicing event. Importantly, this type of splicing would be similar, but different from recursive splicing seen in drosophila [Burnette et al., 2005]. Another example of the importance of intron sequence elements on AS is observed in the equine β-casin gene, where the authors propose a model involving an intronic splicing enhancer bound to the exit channel of the elongating polymerase, promoting inclusion of downstream cassette exons [Lenasi et al., 2006]. Taking a more genome-wide approach Peng et al. examined human and mouse EST data looking for correlations between adjacent AS cassette exons [Peng et al., 2008]. The authors note that positively correlated pairs of adjacent cassette exons typically resemble constitutive exons in splice strength, whereas negatively, or weakly correlated pairs are likely to be newly emerging exons, whose strength of splicing has not evolved enough to be constitutively included.

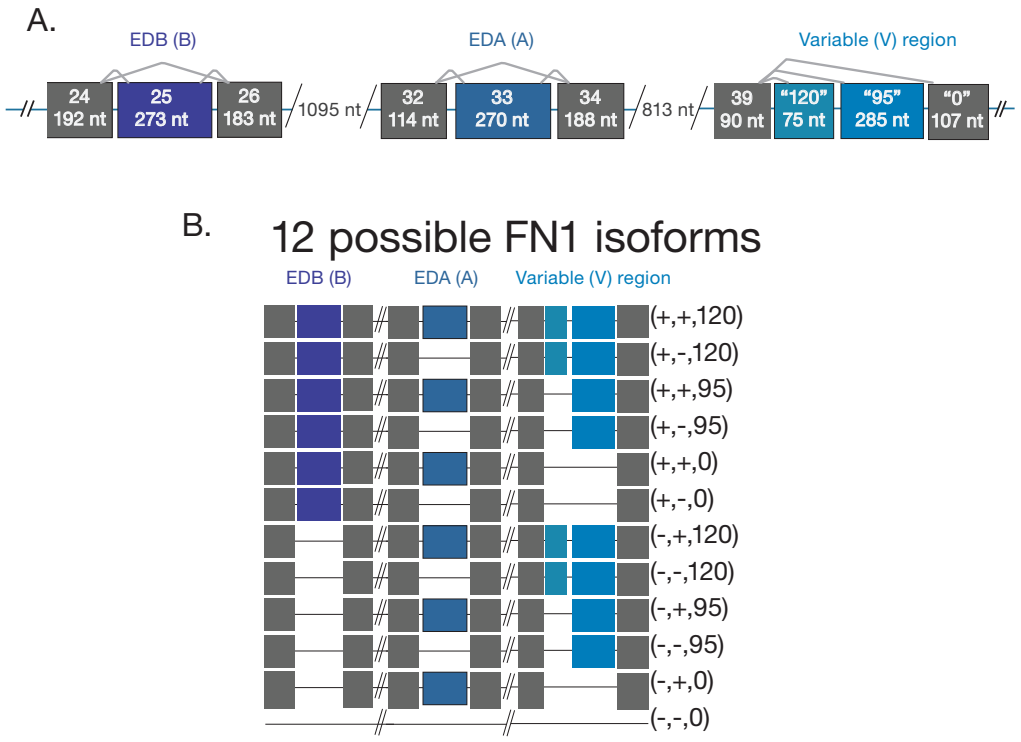
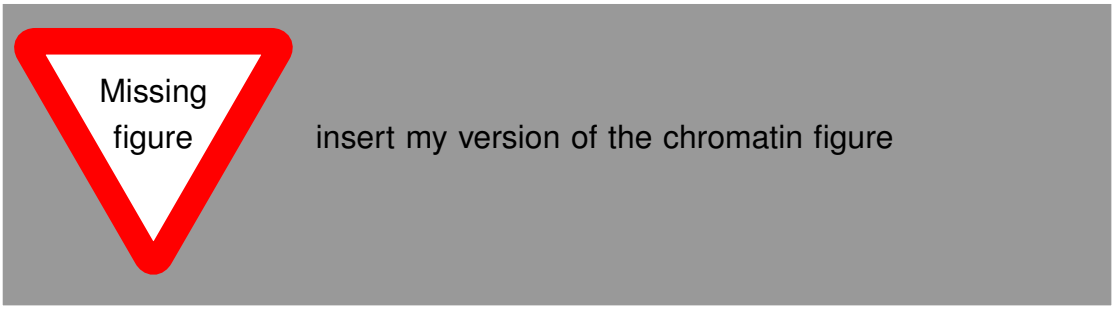


FIGURE 1.6: Mouse *Fn1* contains multiple sites of Alternative Splicing

A) There are three highly-studied regions of AS in mouse *Fn1*: The cassette exons EDB and EDA, and the Variable(V)-region (AKA the IIICS) exon, which displays multiple 3′ splice sites. Each of these sites is separated by multiple constitutive exons.; B) Considering simplistic splicing of these three exons, there are 12 different isoforms of mouse *Fn1*.



The last, most current, and thorough study of intra-gene splicing coordination involves the *Caenorhabditis elegans* gene *slo1* [Glauser et al., 2011, Johnson et al., 2011]. *slo1* is the *Caenorhabditis elegans* orthologue of the human BK channel gene *Kcnma*, also known to undergo extensive alternative splicing

[Nilsen and Graveley, 2010] via 13 cassette exons, potentially coding for over 1,000 different isoforms. *Kcnma* is highly developmentally, spatially, and tissue regulated. It is involved in a diverse range of cellular processes, including hearing, circadian rhythms, urinary function, and vasoregulation [Fodor and Aldrich, 2009]. While the gene is highly conserved, as organism complexity grows, so does the apparent transcriptional diversity of this gene. In worms, *slo1* can produce up to 12 different isoforms. Glauser et al. used QPCR to demonstrate individual, AS region inclusion frequencies do not correspond to complete isoform frequencies, when measured via a TaqMan probe approach. They go on to describe a interdependent-splicing model that best fits the data, and support interdependence via mutations at one sight altering splicing at both upstream and downstream sites of AS, separated by atleast one other splicing event. After measuring the biophysical properties of the isoforms [Johnson et al., 2011], they conclude that coordinated AS is critical for proper BK channel function in vivo. It is interesting to note that this study also identified an intronic sequence element that displayed some type of coordinated, or co-regulated effect on AS.

1.3.5 Many isoforms per gene

It is easy to think of AS as a binary process. Isoform A or B is produced based upon picking either exon A or B. What quickly becomes evident, and is far too real for researchers building transcriptome assembly algorithms, is that the combinatorial nature of AS makes it both a power means of generating isoform diversity and a difficult problem to study [Trapnell et al., 2012].

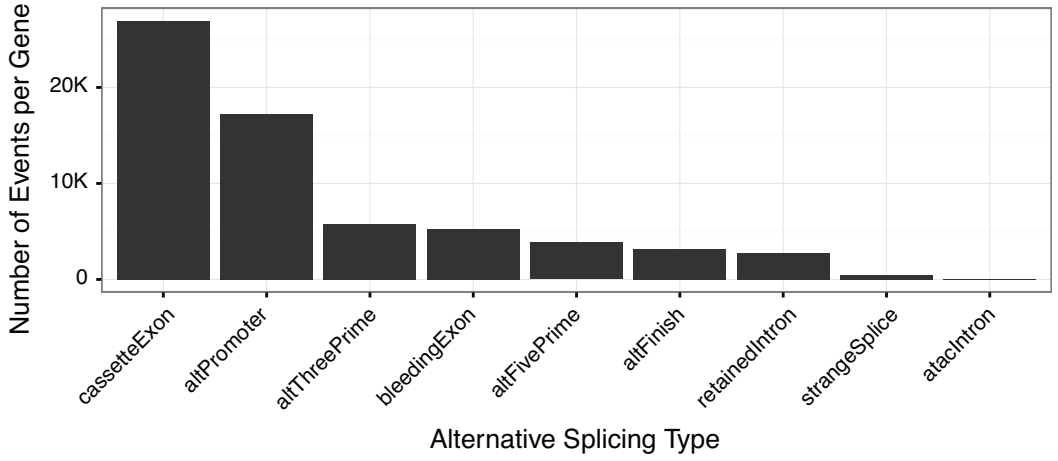


FIGURE 1.7: Number of hg19 Alternative Event types per gene

Alternative Event types per gene. RefSeq on 2014-03-24

One of the most recent attempts to investigate the breath of combinations produced by AS is the already mentioned ENCODE project [Djebali et al., 2012]. ENCODE performed extremely in depth analysis of 15 cell lines, and find that each isoform produces 10 transcripts per gene, with a broad distribution in terms of isoforms expressed per sample.

The ENCODE project clearly demonstrated that most human genes can under AS in many more ways than previously appreciated. Most genes could be considered as undergoing "complex" AS, with numerous forms of AS (See figure 1.7). Despite the prevalence of complex alternative spliced genes, just a few genes are used as examples to illustrate numerical possibilities and biological significance. For example the human immune system relies heavily on AS to be plastic toward antigen recognition and response [Lynch, 2004]. Modulation of extracellular signaling proteins such as *CD44* and cellular adhesion protein *CD45* have been well-studied [Zikherman and Weiss, 2008, **?**]. Alternative splicing in humans, however, does not seem to produce the number of unique possible combinations as AS of genes in simpler organisms, such as fruit flies, perhaps due to specialization of genes, or different genes that work in combination or complexes, as oppose to utilizing unique gene isoforms [Park and Graveley, 2007]. For example, the fruit fly gene muscle myosin heavy chain (*Mhc*) can produce up to 480 different isoforms through AS of 17 different cassette exons [Bernstein et al., 1983].

TABLE 1.1: Fly genes with >2,000 assembled transcripts according to [Brown et al., 2014].

Gene Name	# Introns	# Transcripts	# Proteins
Mhc	60	2040	511
slo	49	2070	279
ps	30	2099	27
rg	45	2178	23
shot	60	2478	886
scrib	53	2555	259
heph	75	2876	52
CG42748	26	2876	51
rdgA	35	3003	89
Mbs	39	3080	119
CaMKI	41	3992	7
par-1	48	4410	142
GluClalpha	27	4945	188
Sap47	24	5011	49
Patronin	50	5615	590
CG17838	37	8333	147
unc-13	52	8391	279
A2bp1	29	9055	58
Imp	33	9131	12
pan	38	9432	72
Sh	40	15995	66
gish	48	18972	142

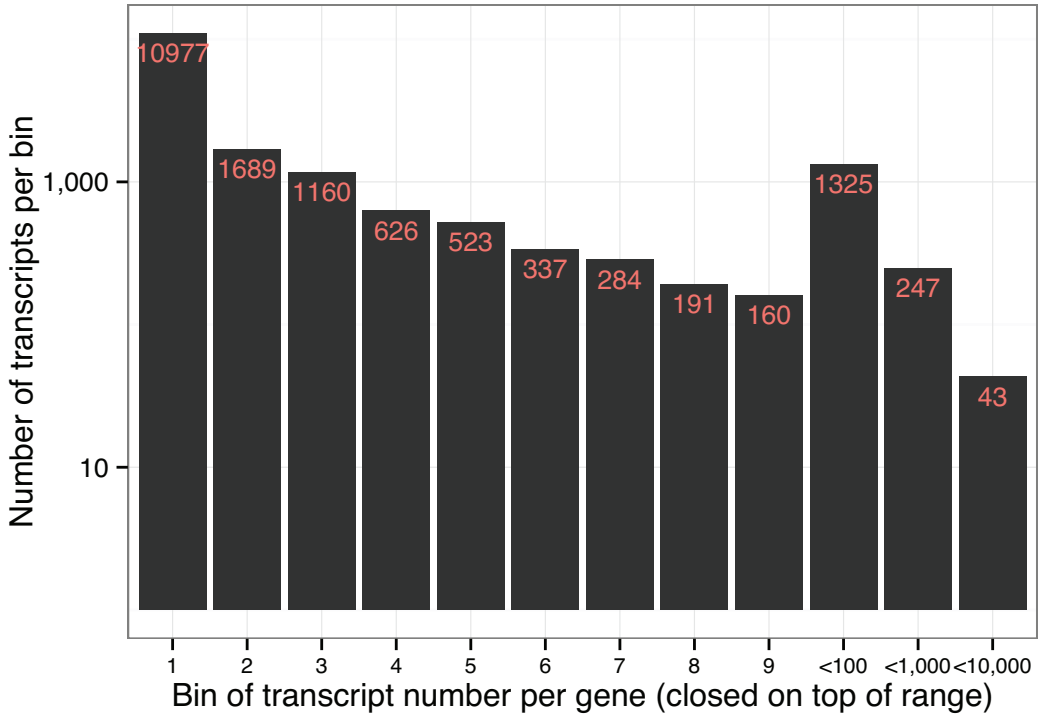


FIGURE 1.8: Number of transcripts per *Drosophila melanogaster* gene

Data from [Brown et al., 2014], Supplemental Table 3. Number of transcript per bin, with bin sizes "closed" on the upper part of range.

1.3.6 *Drosophila melanogaster Dscam1*

Unquestionably, the gene most frequently used to demonstrate the combinatorial power of AS is fly *Dscam1*. The "architecture" of *Dscam1* is rather unique among other organisms, but as we saw in Section 1.3.5, contain some genes that generate tremendous isoform diversity from a single genetic locus [Brown et al., 2014]. The basic structure of *Dscam1* is shown in Figure 1.9.

Human *Dscam*, for which *Dscam1* was named, was identified while looking for genes on chromosome 21, specifically band 21q22, where extra copies expressed in Down syndrome patients, a trisomy 21 disorder, may be causative for disease [**?**]. *Dscam* (Down Syndrome Cellular Adhesion Molecule) was named according to this association, and its membership in the immunoglobulin super family of proteins with extracellular adhesion functions. Human *Dscam* does undergo some alternatively splicing and broadly expressed in the developing nervous system. Yet, it does not contain the same architecture of cassette exon banks as *Dscam1*.

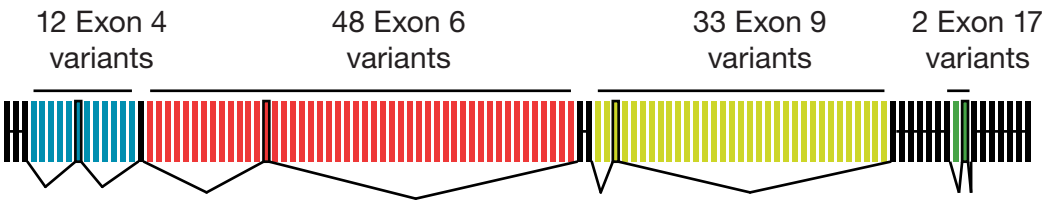


FIGURE 1.9: The architecture of the *Drosophila melanogaster* gene *Dscam1*

Dscam1 has three *clusters* or "banks" of alternative cassette exons that are splicing out in a mutually-exclusive manner. The first bank, "Exon 4", contains 12 different variants, of which one one is ever included into the mRNA. Similarly, banks 6 & 9 each contain 48 and 33 different variants, respectively. These three banks code for extracellular IgG domains, while the final region of AS, exon 17, encodes two different trans-membrane domains, again of which only one is included in the final mRNA.

Complex AS of*Dscam1* was first noticed by the Zipursky lab in 2000 [Schmucker et al., 2000]. While looking for proteins associated with *dock* and *pak*, two proteins important for neuronal growth cone guidance, they biochemically co-purified DSCAM1. Sequencing of *Dscam1* clones revealed that virtually all contained different combinations of exons 4,6, and 9. In fact, these three exons are chosen from three clusters of mutually-exclusive cassette exons, containing 12, 48, and 33 different options each(see Figure 1.9). The initial report kicked off an exciting period of research into *Dscam1* structure and function. The functional significance of *Dscam1* AS was a major goal of multiple labs.

Before the highlights of *Dscam1* research are reviewed, it is illustrative to discuss some basic *Drosophila melanogaster* anatomy. There are 4 main regions where *Dscam1* expression has been highly-studied. These four biologically important roles are shown in Figure 1.10.

- Hemocyte cells of the immune system
- Larva Class IV da Neurons
- Pupal Mushroom-body neurons in the developing brain
- Tetrad synapses of the eye

First, *Dscam1* expression in hemocyte cells of the immune system is important for recognition of foreign antigens [Watson et al., 2005].

During larval development, *Dscam1* is expressed in the da neurons of the larval body wall. Da neurons create a uniform sensory feed, allowing the larva to respond to mechanical stimulus. Morphologicially da neurons resemble an oak tree growing in a sunny field, allowing the larvae to sense as much as possible. In order to maximize coverage of the field, every cell:cell interaction (i.e. every synapse) must be a productive one. Molecularly, this is accomplished via an extracellular handshake between two copies of DSCAM1. If this handskake feels

too familiar, a stable, lasting, and *productive* synapse is actively discouraged until a new and different handshake is felt.

The use of DSCAM1 to discern self from non-self determination is not unique to da nuerons, but is also essential on equally critical, and arguably more complex, nervous systems including the eye and brain. In the developing brain, *Dscam1* is expressed in both axonal projections of neurons extend from their Kenyon cell bodies and bifurcate into two different mushroom body lobs [Zhan et al., 2004].

Celotto and Graveley [2001] investigated developmental regulation of *Dscam1*. They focused on the 12 variants of cluster 4, and observed regulation of exon 4.2, with embryonic transcripts showing little inclusion of this exon, while adult transcripts showed frequent inclusion. Exon 4.8 demonstrated the opposite behavior. Similar regulation of exons with cluster 4 was also observed in a closely related species, *Drosophila yakuba*.

Four years after *Dscam1* complexity was first reported, [Neves et al., 2004] used a specially designed microarray to robustly characterize its molecular diversity. They observed, at some rate, inclusion of virtually all alternative exons with each of the three clusters examined. Additionally, they examined *Dscam1* transcripts obtained from colonies grown from single cells and reported that multiple *Dscam1* transcripts were expressed per cell, estimating the number to be between 7–50 different combinations, depending on the cell type. As discussed above, the use of microarrays to perform this analysis precluded observing any potential coordination between variant exons.

Quickly after Neves et al. [2004] published their results, the Zipursky lab also published a microarray study of *Dscam1* isoforms [Zhan et al., 2004]. They focused their analysis to neurons of the developing Mushroom body (see Figure 1.10). Not only did they also show that most *Dscam1* combinations are likely produced at some level, but the diversity of isoforms is required for bifurcation of neurons into different lobes of the developing mushroom body. These results highlighted a critical function for DSCAM–mediating extracellular interactions via homophillic binding.

Having some functional purpose for *Dscam1* diversity, the next area of research focused on mechanisms of mutually exon selection. Graveley [2005] observed a single “docking site” within the intronic sequence just $5'$ to exon 6.1. This docking site was conserved among 15 insect species examined, from closely-related *Drosophila simulans* to a distantly-related *Tribolium castaneum* (Red flour beetle). Astonishingly, the docking site was complementary to “selector sites” within intronic regions just $5'$ of each of the 48 variant exons. A model is proposed where docking::selector site interactions is required to choosing which of the variant exons is included, while a splicing regulator protein, likely

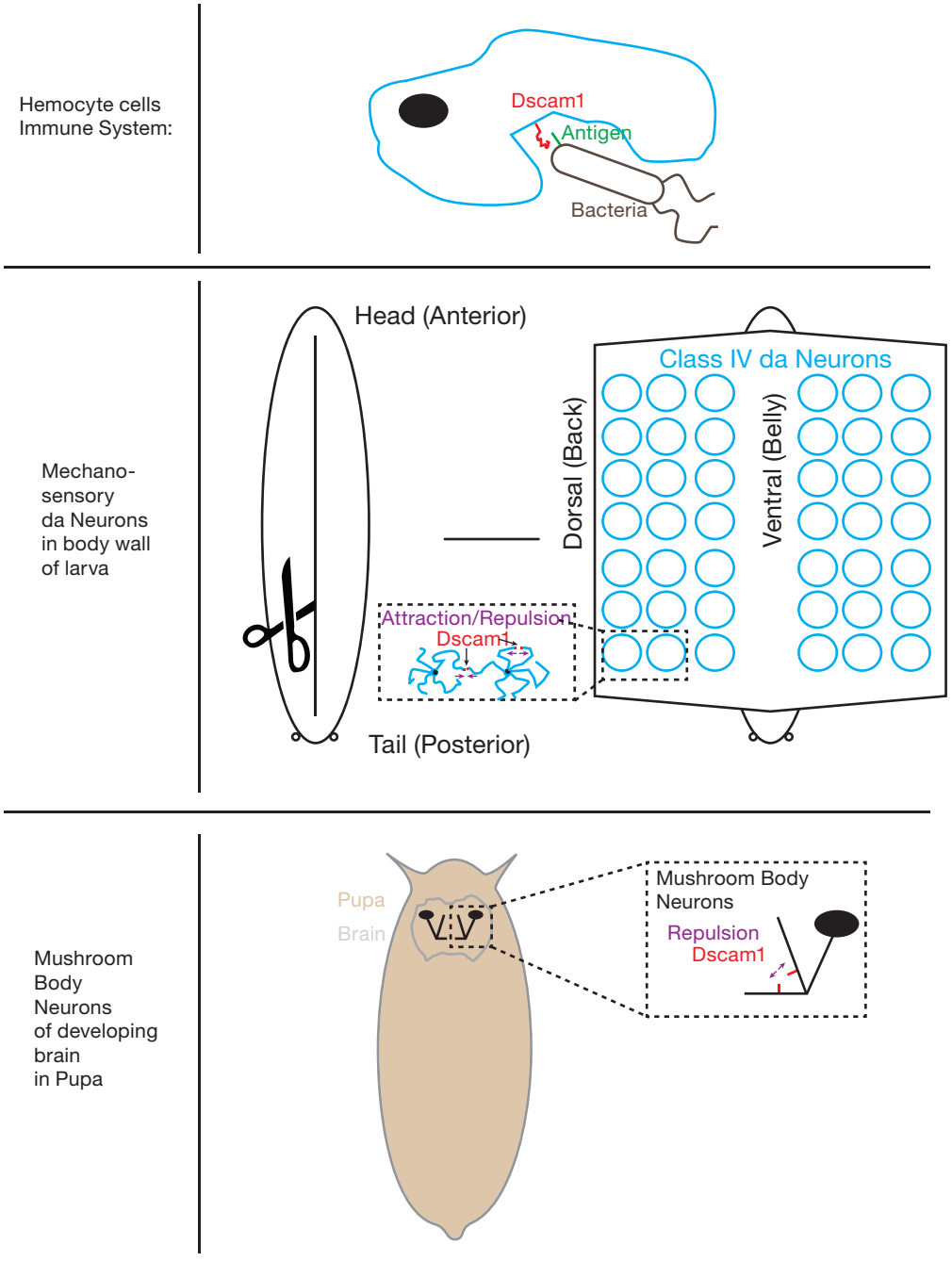


FIGURE 1.10: Important *Dscam1* expression during *Drosophila melanogaster* cycle

Dscam1 has been high-studied in four different regions/cell types. 1) Hemocytes of the immune system, where DSCAM1 is involved in antigen recognition; 2) In Class IV da neurons, which sense mechanical stimulation of the larval body wall; 3) In mushroom body neurons of the pupal developing brain; and 4) (not shown) in Tetrad neurons of the eyes.

and hnRNP due to the repressive nature of the interaction [Graveley, 2000]. Additional mechanisms have been reported for other clusters, including the *iStem* [Kreahling and Graveley, 2005] in cluster 4, and the hnRNP protein hrp36 [Olson et al., 2007].

[Neves et al., 2004] examined *Dscam1* expression in hemocyte cells, and their results clearly show reduced variability in cluster 9 inclusion. Virtually all of the signal obtained from hemocyte cells for cluster 9 was seen in variants 9.[6,9,13,30,and 31]. [Watson et al., 2005] also examined *Dscam1* expression in hemocyte cells, comparing it to that of neuronal cells. They propose that secreted forms of *Dscam1* are essential for a robust innate immune system in insects. These studies highlight how nature has applied one gene that produced extreme molecular diversity to multiple problems involving determing self from non-self [Hattori et al., 2008, Shi and Lee, 2012]. *Dscam1* use in these two very different biologically roles has been summarized previously [Hemani and Soller, 2012].

In 2005 the Zipersky lab published [Consortium, 2004] the first in a series of quality reports describing the function and diversity of *Dscam1* from a genetic approach. [Hattori et al., 2009, 2008]

Insert Hattori Series Summary

Finally, in another tour-de-force of genetic manipulation, once again the Zipersky lab advanced our understanding of *Dscam1*. [Miura et al., 2013] used a collection of *Dscam1* mutants allowing for visualization via GFP of specific cluster 4.X variants being used in real time. This was accomplished by creating frame shifting mutations when any varaint other than the one designer in the given mutant was included by the cell. This allowed Miura et al. [2013] to make the following conclusions: 1)

Need transition from Dscam to Rnl2!

1.4 Nucleic Acid Ligation

1.4.1 RNA Sequence investigation by ligation

In the late 1960's and early 1970's, the Lehman and Richardson labs characterized two workhorse-enzymes of modern molecular biology. Robert Lehman and colleagues, working at Standford Medical School, first described the activity of *polynucleotide-joining enzyme* from *Escherichia coli* (now known as *E.*

Coli DNA Ligase) [Olivera and Lehman, 1967]. Work on this enzyme paralleled that from the Richardson lab at Harvard Medical School, where they focused on *polynucleotide ligase* from *Escherichia coli* infected with T4 bacteriophage (now known as T4 DNA ligase) [Weiss and Richardson, 1967]. It became clear that while these two enzyme's shared a common mechanism—later elucidated by [Modrich et al., 1973]—they had important differences. First, T4 DNA ligase required ATP as a cofactor, which *E. Coli* DNA Ligase did not (though it was later discovered that DNA ligase required NAD as a cofactor). Second, only T4 DNA ligase could catalyze ligation of blunt-ended DNA [Tabor, 1987].

The general mechanism of ligation, shown in Figure 1.11, involves three steps: Step 1 (A) involves the ϵ-amino group from the active site lysine performs a nucleophilic attack on the α-phosphate of ATP in solution. B) The ligase is now charged with AMP and inorganic phosphate (PPi) is freed into solution. C) Step 2: Nucleophilic attack by the 5′ DNA phosphate on the 3′ side of the nick to the AMP:ligase phosphate. D) Adenylated DNA is now competent for DNA ligation. E) Step 3: the 3′ OH on the 5′ side of the nick performs a nucleophilic attack on the 5′ PO_4 across the DNA nick, liberating AMP into solution. F) Sealed nick resulting in: Ligase; AMP; and intact dsDNA.

In addition to elucidated the general mechanism of ligation, it was also discovered that T4 DNA ligase lacks a preference for terminal polynucleotide structures. The Khorana and Richardson labs both reported the activity of this enzyme on combinations of RNA and DNA duplexes [Fareed et al., 1971, Kleppe et al., 1970]. Both of these papers describe an activity on T4 DNA ligase, RNA-templated DNA to DNA ligation, that is of particular relevance to this thesis work. Unlike T4 DNA ligase, *E. Coli* DNA Ligase, will not join DNA strands on an RNA template [Bullard and Bowater, 2006]. Soon after demonstrating these activities in vitro, the Khorana lab reported detection of organism-generated DNA [Besmer et al., 1972], setting up an orthogonal field (respective to PCR) of nucleic acid sequence characterization [Conze et al., 2009].

An enzyme that can catalyze an RNA-templated DNA:DNA ligation is a very useful molecular biology tool for two main reasons. First, using RNA as a ligation guide means no modification is made to the template molecule. This contrasts cDNA analysis, where the RNA has been enzymatically converted by reverse transcription, potentially losing valuable RNA-coded information, such as modified bases. Second, synthesis of the DNA probes used in ligation is inherently easier and cheaper compared to synthesis of RNA probes. In addition to being cheaper, synthesis of DNA probes has become high-throughput since the adoption of microarrays as a standard gene expression measurement tool [Schena et al., 1995]. A pair of papers from the Landegren lab first reported the utility of RNA-templated DNA:DNA ligation for analysis of RNA transcripts

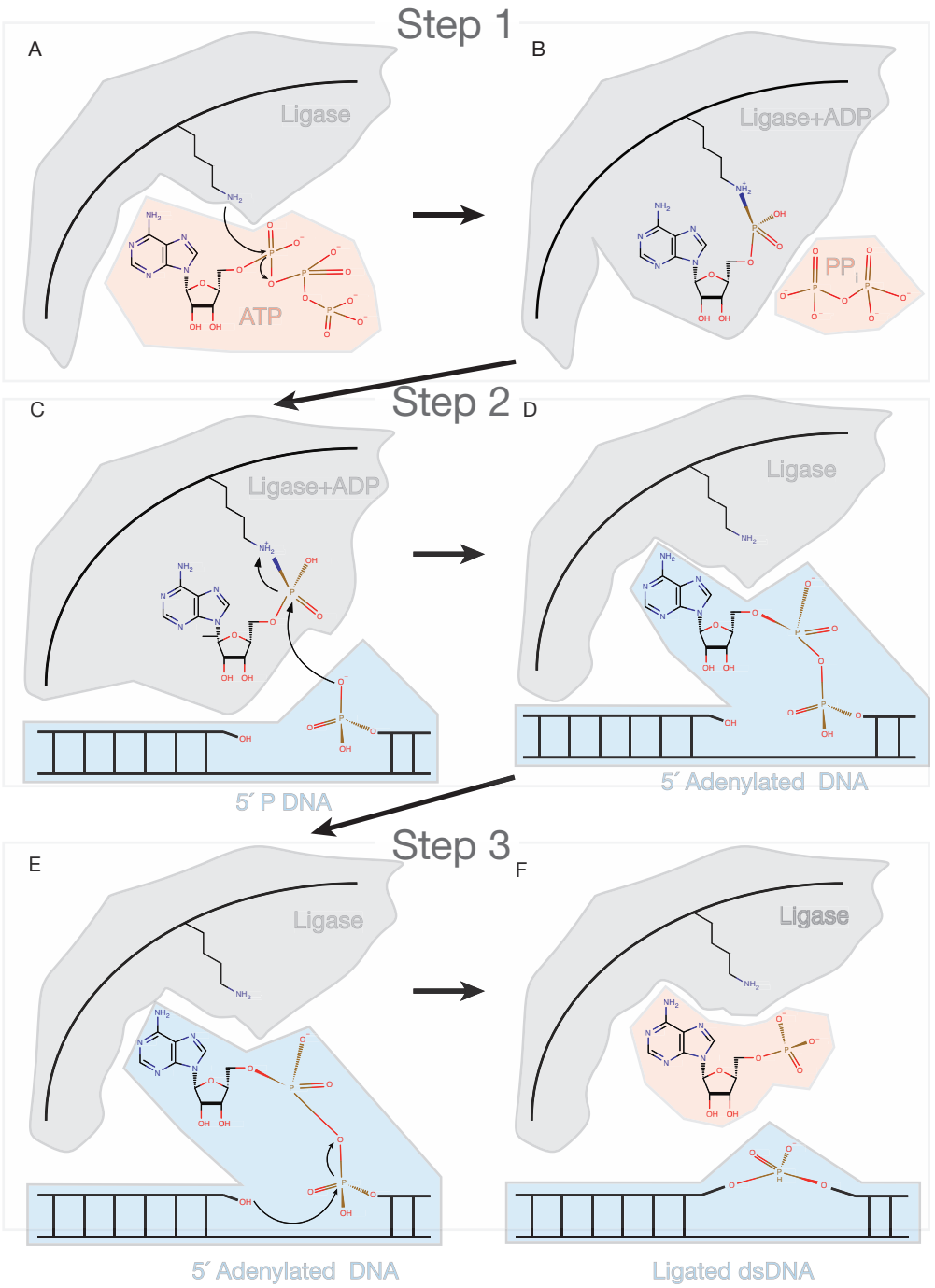


FIGURE 1.11: Mechanism of ATP-dependent ligation

Adapted from [Nandakumar et al., 2006] and specifically for that of T4 RNA ligase 2.

[Nilsson et al., 2001, 2000]. The Fu lab applied this approach in a multiplex experimental design in collaboration with Illumina [Li et al., 2012, Yeakley et al., 2002], while Mats Nilsson and Ulf Landegren developed a single molecule application [Conze et al., 2010]. It is important to note that *all* of these studies used T4 DNA ligase. Clearly, there is interest and utility in analyzing RNA in both high-throughput and multiplex experimental designs, using cheap DNA probes, and without cDNA conversion.

For more than 40 years after its first description, T4 DNA ligase was the only choice for RNA-templated DNA:DNA ligation. However, a recent publication from New England Biolabs (NEB) describes this activity by another well-studied ligase, Chlorella Virus PBCV-1 DNA ligase (herein Chlorella DNA ligase) [Lohman et al., 2013]. Chlorella DNA ligase is a long-studied enzyme and had been reported to *not* display RNA-templated DNA:DNA ligation activity [Ho et al., 1997, Sriskanda and Shuman, 1998]. However, at high enough concentrations and under special buffer conditions (specifically a critical concentration of ATP), Lohman et al have shown that Chlorella DNA ligase will join two DNA strands hybridized to an RNA template [Lohman et al., 2013]. They further demonstrated that it performs no worse in this activity than traditional T4 DNA ligase [Nilsson et al., 2001, Yeakley et al., 2002].

Building on the list of available enzymes that join hybrid polymer substrates Chapter 2 presents data supporting RNA-templated DNA:DNA ligation activity for another enzyme, T4 RNA Ligase 2.

1.4.2 T4 RNA Ligase 2 (Rnl2)

Proteins of the T4 and T7 bacteriophages have been a boon for molecular biology. Without enzymes like polynucleotide kinase [Richardson, 1965], T7 RNA polymerase [Summers and Siegel, 1970], and T4 DNA ligase [Weiss and Richardson, 1967], many essential manipulations of nucleic acids would have been impossible for decades. Obviously, these enzymes also have essential phage functions. T7 RNA polymerase is responsible for late stage replication of T7 phage transcripts, while T4 PNK works in concert with T4 DNA and RNA ligases to repair cleaved nucleic acids resulting from bacterial pathogens defense systems [Wang et al., 2002]. Specifically, T4 RNA ligase 1 (herein “Rnl1”, also known as *gene 63* maintains phage replication by repairing tRNAs cleaved by an anticodon nuclease produces from the *prr* locus [Amitsur et al., 1987].

Given the utility and importance of these enzymes, novel enzyme discovery is a fruitful area of research. The Shuman lab has a distinguished record of discovering and characterizing numerous such enzymes, including any involved in

nucleic acid synthesis, modification, and repair. Through a blast search looking for novel ligases with sequences related to *Trypanosoma brucei* RNA-editing ligases TbMP52 and TbMP48 [Ho and Shuman, 2002], they identified motifs in correct arrangement, spacing, and number indicative of an RNA ligase. The gene, identified as *gp24.1*, has quickly become an essential tool in the era of modern genomics.

Initial biochemical purification and characterization of *gp24.1* [Ho and Shuman, 2002] revealed that it indeed codes for an RNA ligase, which was renamed T4 RNA ligase 2 (herein "Rnl2"). Rnl2 is a 374 amino acid monomeric protein composed of 2 distinct domains initially purified as a 42-kDA His-tagged recombinant protein. The N-terminal domain (1–243) is responsible for steps (1) and (3) of the general ligation mechanisms (see Figure 1.11), while the C-terminal domain (244–329) is responsible for adenylation of the 5′ PO_4 on the 5′ residue at the 3′ side of the nick, as shown in step (2). Additionally, Rnl2 is routinely purified as a pre-adenylated and immediately poised for its first ligation. In contrast to the N-terminal domain, which is composed of motifs typical to main ligases, the C-terminal domain is significantly different from all other DNA ligases and has no structural homologue. While the biological function of Rnl1 is known, the biological function of Rnl2 remains a mystery, more than 12 years after its discovery [Chauleau and Shuman, 2013]. However, there is some speculation that the flurry of research into bacterial CRISPR phage defense may reveal a role for Rnl2 [Barrangou et al., 2007, Chauleau and Shuman, 2013].

Mutational analysis of Rnl2, and later a crystal structure of the enzyme, have identified key functional residues [Ho et al., 2004, Nandakumar et al., 2004, 2006, Yin et al., 2003]. The lysine residue at position 35 (K35) receives the AMP in Step 1. The K227 residue in the C-terminal domain is essential for both forward and reverse adenylation of the 5′ PO_4 at the nick [Viollet et al., 2011]. Mutation of H37 results in an 102 reduced ligation rate, and therefore indicates the essential nature of this residue. Finally, T39 has been shown to interact with the 2′ OH on the 3′ side of the nick, preferring a C3′ endo sugar pucker confirmation (see Figure 1.13). Rnl2 has a minimal footprint of 13nt, centered on the nick, and only requires magnesium for transfer of AMP to the 5′ phosphate. Work done in the Shuman lab [Nandakumar et al., 2006] observed that 2′ deoxyribose residues on the 5′ side of the nick (i.e. DNA) adopt an RNA-like sugar pucker, leading to the correct orientation of the 3′ OH relative to the AMP leaving group and resulting in ligation. This conformation is of particular importance to this results presented in Chapter 2.

While Rnl2 is extremely efficient at high concentration, displaying little or no

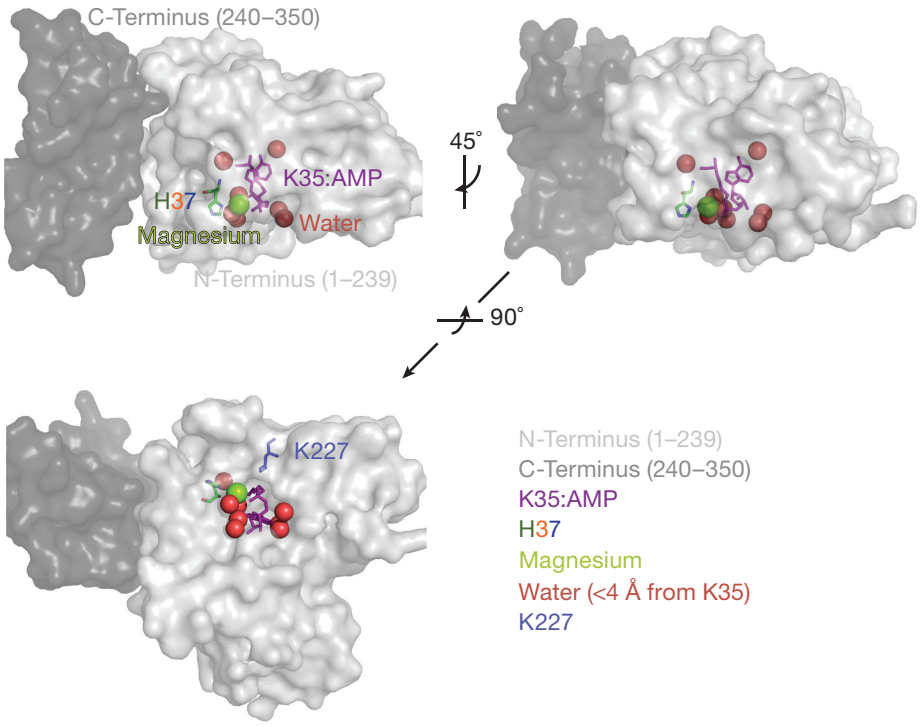


FIGURE 1.12: Structure and active site of pre-adenylated of Rnl2

Rnl2 as crystalized and described by [Nandakumar et al., 2006]. Structures from PDB:2HVQ were generated with PyMol. Top left) Rnl2 is composed of a C-terminal and N-terminal domain. Top Right) The active site of Rnl2 is highlighted. Bottom left) Active site of Rnl2 as shown from bottom. This face interacts with substrate. Residue numbering refers to that of the crystal structure.

reversible chemistry, a modified version of the enzyme containing only the N-terminal domain and a K227A point mutation ("Truncated mutant") has no adenyltransferase activity. In this case, adenyltransferase refers to the ligase transferring AMP from an adenylated substrate to itself; reverse chemistry of step 2 in Figure 1.11). This mutant has been used in specialized cloning applications [Ghildiyal et al., 2008, Hafner et al., 2008, Viollet et al., 2011] that take advantage of this activity. In these reactions, the use of pre-adenylated 3′ DNA adaptors allows for selective ligation among already phosphorylated species by limiting the enzyme-catalyzed transfer of AMP from the adaptor to other phosphorylated species. Use of this truncated mutant to create a hybrid RNA/DNA molecule has greatly improved many high-throughput sequencing work flows.

Ligation of hybrid substrates (eg. DNA-templated RNA:DNA vs DNA-templated DNA:DNA) have revealed general substrate preferences. DNA ligases appear

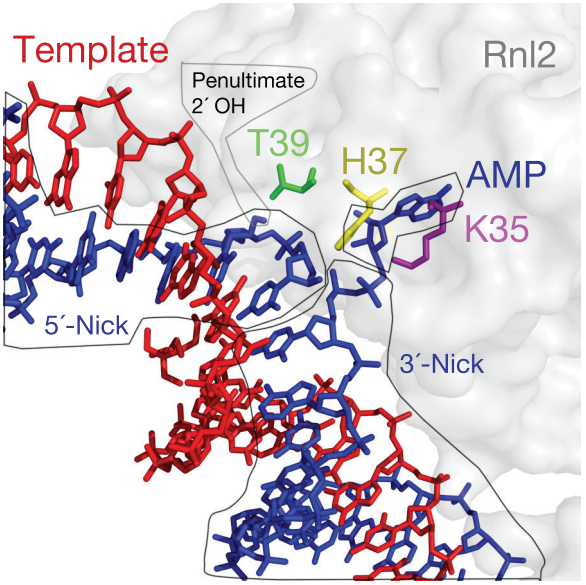


FIGURE 1.13: Structure and active site of pre-adenylated of Rnl2

Rnl2 complexed with nicked dsDNA as crystalized and described by [Nandakumar et al., 2006]. Structures from PDB:2HVR and generated with PyMol

to prefer the residue bearing the 5′ phosphate on the 3′ side of the nick to be 2′ deoxyribose, and have a relaxed requirement for the sugar on the 5′ side of the nick. RNA ligases have the reverse preference, demonstrating higher activities when the 5′ strand, 3′ OH residue also bears a 2′OH. Rnl2 has an additional preference for an RNA residue at the penultimate 3′ side of a nick [Ho and Shuman, 2002, Ho et al., 2004, Nandakumar et al., 2004, 2006]. The two base requirement for RNA at the 5′ side of the double stranded nick biases Rnl2 to join RNA:[RNA/DNA] strands. Independent labs have measured this preference and have reported that the RNA-templated DNA:DNA joining activity of Rnl2 is below assay limits of detection [Bullard and Bowater, 2006]. However, results discussed in this work clearly show that with enough enzyme, and sensitive downstream measurements, Rnl2 will catalyze RNA-templated DNA:DNA ligation (see Chapter 2. Previous reports of Rnl2 lacking this activity are likely due to a single turnover mechanism in this reaction, owning to the poor dissociation rate of nucleic acid-interacting enzymes.

Discuss Reuveni et al. [2014] on importance of enzyme unbinding to the speed of reactions.

Transition from Rnl2 to long RNAs/ piRNA precusors somehow.

1.5 Nucleic Acid Polymers

1.5.1 piRNAs original from very long PolII Transcripts

1.5.1.1 Brief history of piRNAs

piRNA section needs major work

Mammalian spermatogenesis is critical for the future of the species. Recently the importance a specific kind of small RNA—piRNAs—for proper spermatogenesis has become clear [Siomi et al., 2011]. Even after >12 years since their discovery in *Drosophila melanogaster*[Aravin et al., 2001], and >6 since their identification in rodents [Girard et al., 2006, Lau et al., 2006], the essential mechanisms of piRNA biogenesis to proper mammalian spermatogenesis remains largely unknown. These unknown mechanisms include: biogenesis from transcript to small RNA, physiological targets, and terminal function of sterility maintenance. What is known is that without a functioning piRNA pathway males are sterile. Studies in humans have also linked SNPs in the Argonaute proteins that bind piRNAs to decreased fertility [Gu et al., 2010].

piRNAs are so-called because they bind PIWI proteins, a sub-group of the Argonaute protein family, whose other members utilize small RNA as guides to target nucleic acids for many forms of post-transcriptional regulation [Siomi et al., 2011]. There are three PIWI proteins in mice, each displaying a distinct expression profile during development and an association with piRNAs of a specific length. The first PIWI protein expressed, even before a mouse is born, is MIWI2 [Carmell et al., 2007], followed quickly by the more consistent player, MILI (see Figure 1.15 [Aravin et al., 2006, Kuramochi-Miyagawa et al., 2004]. It is during the "fetal" stage of piRNA biogenesis in mice that MIWI2 and MILI undergo ping-pong amplification, similar to that observed in flies, in order to silence expression of transposons during germ line formation [Brennecke et al., 2007, Kuramochi-Miyagawa et al., 2008]. After birth, and during the "neonatal stage" only MILI is expressed, and piRNAs shift from mostly transposon-mapping to 3′ UTR mapping [Robine et al., 2009]. Once the "first wave" of spermatogenesis [Laiho et al., 2013, Oakberg and Oakberq, 1956] reaches meiosis, the pachytene piRNAs are expressed [Girard et al., 2006, Lau et al., 2006, Li et al., 2013a]. Pachytene piRNAs tend to map to intergenic "clusters" of unique genomic sequence. These clusters, from here after called "piRNA-generating genes" appear to produce a single, continuous, relatively long, and un-spliced Pol II transcript [Li et al., 2013a]. This is comparable to piRNA clusters in flies,

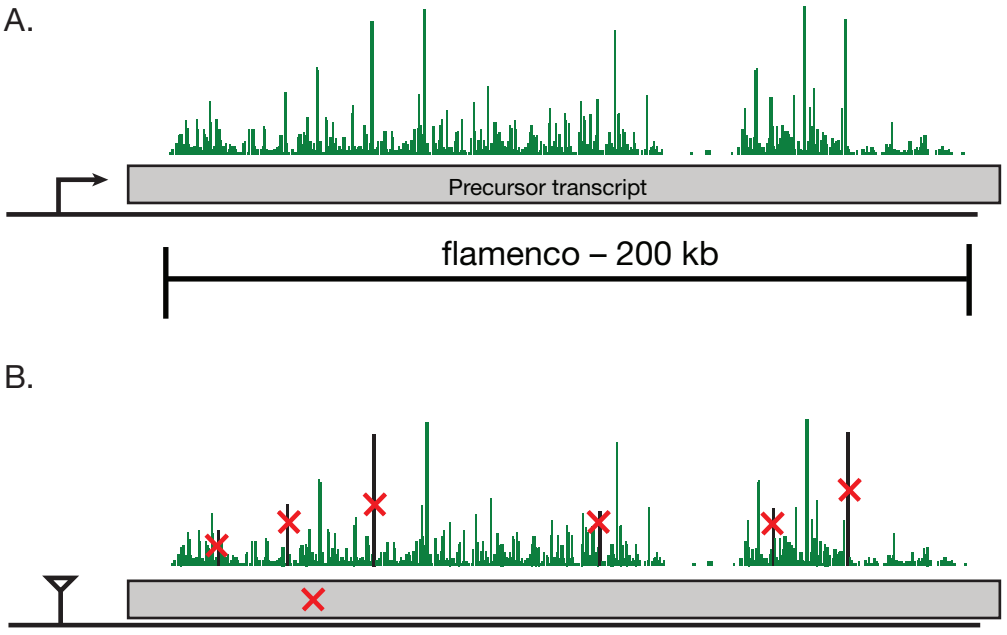


FIGURE 1.14: A the *Drosophila melanogaster* gene *flamenco* is a graveyard for transposon sequences [Pélisson et al., 1994]. Evidence for expression of a single-contiguous RNA transcript from *flamenco* (A) is provided by a P-element insertion into the suspected promoter region (B). [Brennecke et al., 2007] could not detect specific piRNAs (red X's) by northern blot in the P-element mutant.

such as flamenco, whose transcription can be abolished with a P-element insertion into a putative promoter, as measured by northern blot looking for piRNAs generated 168 kb downstream (see Figure 1.14 [Brennecke et al., 2007].

1.5.1.2 Fly and Mouse piRNAs have important differences

piRNAs are small RNAs with lengths 24–31 nt long, making them longer than other small RNAs (eg. miRNAs or siRNAs). They are believed to derive from single-stranded RNA precursors because also unlike other small RNAs, their biogenesis does not require double-stranded RNA-specific ribonuclease Dicer [Houwing et al., 2007, Vagin et al., 2006]. Yet, similar to other small RNAs, they bind a sub group of the Argonaute family of proteins, PIWI proteins, from which their name is derived (*PIWI Interacting RNAs*.

Mammalian piRNAs can be divided into three major classes (see Figure 1.15. *Fetal piRNAs* are present before birth. These piRNAs tend to be short, bind the PIWI protein MILI2 in mice, and have sequences found in transposable elements [Carmell et al., 2007]. The next class of piRNAs, historically but confusing grouped with the previous class, are called *Pre-pachytene piRNAs*.

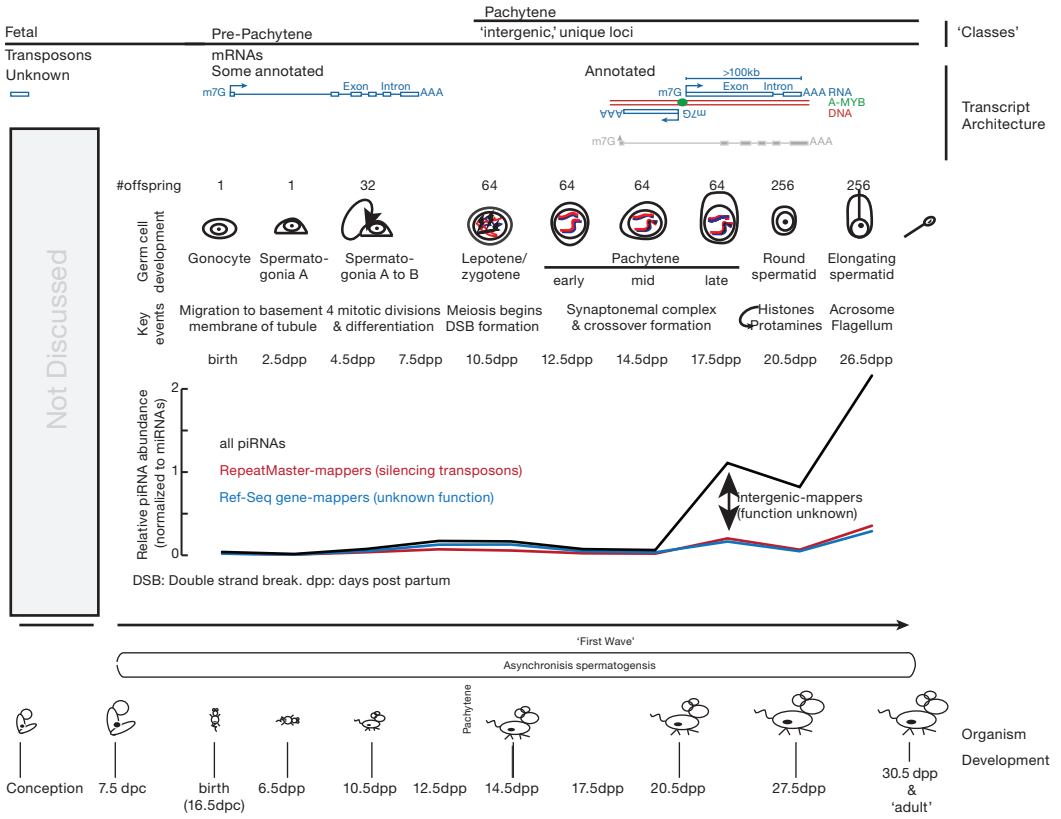


FIGURE 1.15: Write a nice caption here!

Pre-pachytene piRNAs are expressed just before birth, and continue to be expressed in functioning testes. These piRNAs tend to map to traditional, annotated, protein-coding genes. Finally, due to their unique sequence in the genome, the genetic origin of millions of piRNAs belonging to the third class, the *pachytene piRNAs*, was immediately known. Pachytene piRNAs are extremely abundant after the pachytene stage of meiosis I when chromosomes pair up, cross over, and exchange genetic material. The genomic origin of these piRNAs, while unique in terms of sequence, are often in "gene deserts"–unannotated and devoid of introns. This gene architecture makes the pachytene piRNA loci some of the most interesting RNA-producing regions of the mammalian genome.

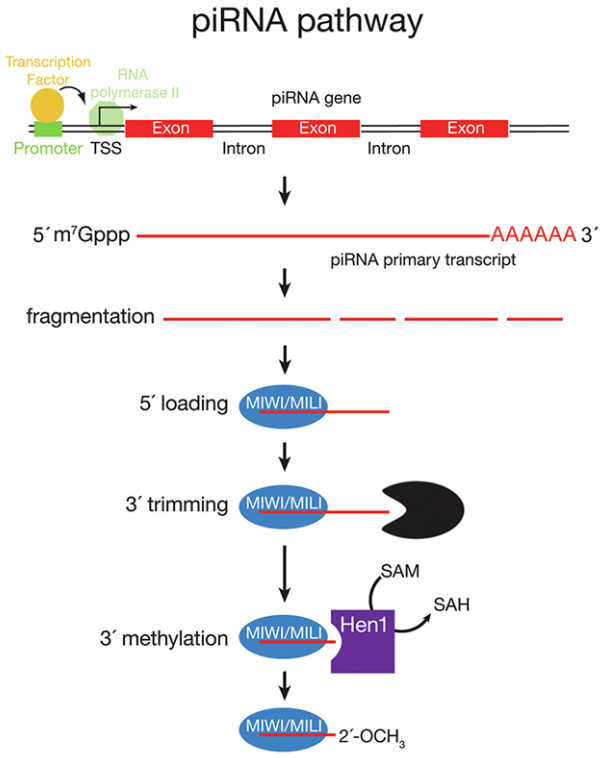


FIGURE 1.16: Figure taken from [Li et al., 2013b]: A model for piRNA biogenesis. Primary piRNA transcripts are transcribed by RNA polymerase II and contain 5′ caps, exons and introns and poly(A) tails. The transcription of pachytene piRNA genes is controlled by A-MYB; transcription factor(s) (TF) controlling pre-pachytene piRNA genes remain to be discovered. Current models of piRNA biogenesis propose that PLD6 determines the 5′ end of piRNA intermediates with lengths >30 nt. These intermediates are proposed to then be loaded into PIWI proteins. After PIWI binding, a nuclease is thought to trim the 3′ end of the piRNA to the length characteristic of the particular bound PIWI protein. Finally, further trimming is prevented by addition of a 2′ -O-methyl group to the 3′ end of the mature piRNA by the S-adenosylmethionine-dependent methyltransferase HEN1. Figure adapted from [Li et al., 2013c].

1.5.1.3 Known functions of mammalian piRNAs

Two studies [De Fazio et al., 2011, Reuter et al., 2011] used point mutations in the catalytic triad of MIWI2, MILI to remove slicer activity. The MIWI2 and MILI studies found that the mice were sterile, and did not accumulate transposon-mapping piRNAs. De Fazio et al. [2011] found that the mice were also sterile, and demonstrated increased LINE1 transcript accumulation. Reuter et al. [2011] states that much the biological activity of MIWI depends on its slicer activity.

Mouse piRNAs have also been implicated in gene imprinting [Watanabe et al., 2011b].

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

1.5.1.4 Integration of multiple HTS datatypes of piRNA analysis

the Resources paper!

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

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

piRNAs lend themselves to study using HTS

Require more datasets to work back to molecular precursors

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

Precision of 5′ 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 3′ end processing.

1.5.2 From short reads to full-length transcripts

Assembly of full length transcripts is difficult for at least 3 reasons. 1) The transcriptome is expressed across many 5 orders of magnitude, with an RNA-Seq library containg many reads from a few highly-expressed genes, and much fewer reads from many lowly-expressed genes [Blencowe et al., 2009]; 2) RNA-Seq libraries are often not created from a completely pure source of mRNA and can contain reads from other RNA classes (e.g. tRNAs) or intronic reads from pre-mRNAs; and 3) Reads are often much shorter than a typical mRNA, making it difficult to assign which read goes to which isoform of a given gene (see the "connectivity problem" discussed in section 1.3.3. With these challenges

in mind, what is the current state of transcript reconsuction (herein transcript assembly).

Computational transcriptome assembly of short reads is currently performed in one of two modes: genome-guided and genome-independent [Garber et al., 2011]. The difference between these two approaches is use of a high-quality genome during the reconsruction process. Popular assembly programs such as Cufflinks [Trapnell et al., 2010] and Scripture [**?**], use genome-aligned short reads as the bases for

Constraints imposed by the huge dynamic range of mRNA expression is the source of the biggest issue with current transcript assembly approaches: they frequently generate short transcript fragments, or contigs, due to poor coverage on long and lowly-expressed transcripts. Merging these contigs into a continuous annotation is a major goal. Improvements will surely come from greater sequencing depth, longer reads, and mRNA enrichment schemes. Longer-term barriers include repetive sequences, transcript secondary structure [Wan et al., 2014], and mRNA processing including hydrolysis and RT processivity [Sharon et al., 2013].

Use Blower2014 to continue/refine discussion

Chapter 2

SeqZip - Development and Applications

2.1 SeqZip Overview

Development of the SeqZip methodology began with an attempt to circumvent an obvious shortcoming in second generation HTS—the short nature of the reads. Until second generation HTS (i.e. reads <100nt on either the Illumina or SOLiD platforms), most sequencing was done using cloned fragments, stored in a bacteria, and analyzed using dideoxy Sanger Sequencing (see 1.2.1. Indeed, this is how most ESTs where analyzed. An extremely powerful feature of these ESTs is that as they represented the sequence of a single clone, from a single original molecule of RNA, the connectivity between sequences that were far apart (>1,000 nt) in the original sequence was maintained. It is this very feature, the continuity of sequence, that allowed whole genome shotgun sequenced to be used, and ESTs to be assembled, into complete genomes, despite sometimes lengthly, and highly repetitive, sretches of DNA (see 1.2.1). Once research transitioned over to heavy use of the second generation HTS, all of that connectivity was lost, and all the inherent information with it.

Second generation HTS can be supplemented with other technologies. This has been demonstrated perhaps most successfully with long-read assisted genome assembly [Koren et al., 2012]. Why not supplemental the disconnected nature of short reads with another technilogy? To that end, Phillip Zamore proposed an RNA-templated DNA to DNA ligation approach as drawn in Figure 2.1 (see also US Patent application 12/906,678). Using this approach, 2 or more distant sequences of RNA are investigated using short DNA oligonucleotides that

hybridize to target sequences, forcing the intervening sequences to loop out. Incorporation of the hybridized DNA via ligation with those of DNAs adjacently hybridized generates a positive read out of sequence presence.

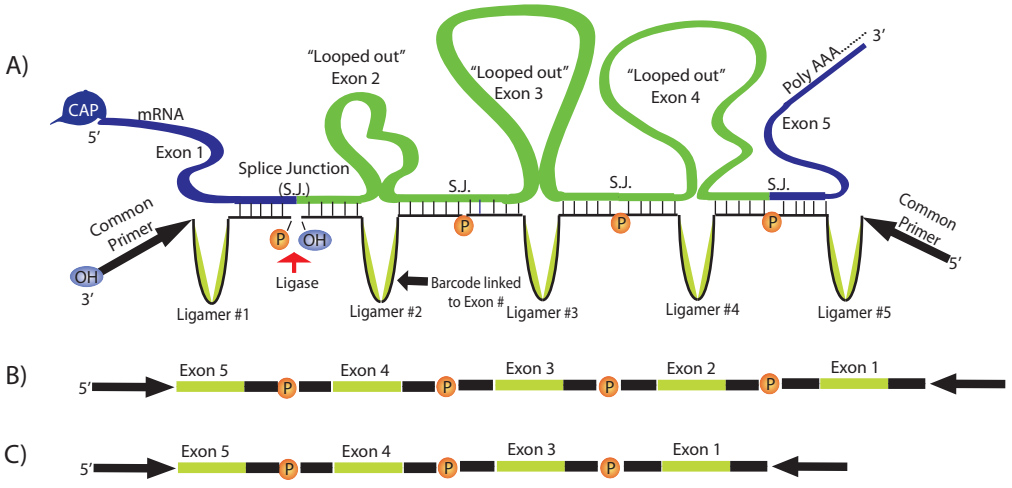


FIGURE 2.1: Original SeqZip Diagram
This is the original concept diagram of the SeqZip methodology. (A) Specific DNA oligos target an mRNA and loop out the RNA sequence. Ligases is added to join the DNA oligos together; (B) & (C) Two different possibilities of ligation products templated from the RNA in (A), where Exon 2 is an cassette exon.

Along with Patent 12/906,678, Chapter 3 presents much of the early and important developmental work demonstrating reduction to practice of this method (termed “SeqZip"), and its application to investgating connectivity of sequence content in biologically-interesting genes *Fn1* and *Dscam1*.

Presented in this Chapter are experiences demonstrating SeqZip application to the following questions and issues:

- Section 2.2: Investgation of 10 genes, simultaneously ("Multiplex") for connected AS decisions
- Section 2.3: Using SeqZip to investigate the interigy of RNA molecules
- Section 2.4: Using SeqZip to demonstrate the presence of long, continuous piRNA precusors

The three sections contrast with Chapter 3 in some important ways. First, Section 2.2 demonstrates that the SeqZip method can not only be used to investigate one, extremely complex Alternatively splice gene, such as *Dscam1* in a c comprehensive manner, but can also be applied to looking at multiple genes at once. Seciton 2.3 expoits an important subtle feature of the method—that the

RNA must be intact in order to produce a ligation product. This can be used to report on a fraction of RNA that is intact, and deduce meaningful information such as the amount of RNA virus that is intact (see 2.3.2), or the existance of as-yet unobserved mega transcripts, like mammalian piRNA precursors (see 2.4 and [Li et al., 2013a,c]).

2.2 Multiplex Gene Study

Is the coordination discussed in section 1.3.4 a general phenomenon? One of the major goals of developing the SeqZip methodology was investigating potential coordination genome-wide. By genome-wide, what we really mean is to analyze many (or all) of the RNA transcripts in a tissue for evidence of coordinated splicing decisions. When development of the method reached the point that it could be applied in a multiplex study, I did not posses the bioinformatic skills necessary to 1) design ligamers in an automated and high-throughput fashion and 2) identify target transcripts, exons, and sequences to investigate for potential connectivity. Both of these points are discussed later(see C**??**.

In order to make some progress on applying the technique to multiple genes at once, I used data presented by Fagnani et al. [2007]. This paper identified genes displaying tissue-specific splicing patterns, focusing on those with CNS-specific patterns. Once section focused on "Coordination between AS events belonging to the same genes," and seemed to be the exact type of data we were interested in applying the SeqZip method too. Five hundred of the 3,044 genes investigated by their microarrays contained 2–5 alternative exons. Fagnani et al. [2007] contained an additional data file listing all pair-wise combinations of alternative exons in the same gene (with that gene having significant expression in >20 different tissues), along with the standard and partial spearman correlations.

It is important to note that the genes above also contain alternative first exons, a prominant type of AS (see figure 1.7). Indeed, from microarrays studies, it has been estimated that approximately 16%–23% of all AS events involve alternative first and last exons [Bingham et al., 2008]. It is known that, through alternative use of first and last exons, cells can fine-tune a transcript's untranslated region (UTR) and control many aspects of mRNA regulation including nuclear export, localization, expression, and stability [Hughes, 2006]. In support of the importance of alternative UTRs in tuning of gene expression, a landmark RNA-Seq study demonstrated a high occurrence of alternative first and last exon splicing, with alternative tandem $3'$ UTR usage being the most highly tissue-dependent form of AS observed [Wang et al., 2008]. The current model

of spatial proximity between 5′ and 3′ UTRs is suggestive of their possible interdependence. In our multiplez analysis, we included genes potentially displaying interdependence between first and last exons. Discovery of interdependence would lead to many questions into how specific combinations of UTRs can influence mRNA processing downstream of AS.

Using the Fagnani et al. [2007] data, I filtered exon pairs to those with a distance >350 nt in the final pre-mRNA. I also visualized their transcript achetecture, and EST evidence using NCBIs AceView tool [Thierry-Mieg and Thierry-Mieg, 2006]. For example, the exons with strong correlation of expression in *Chl1* are in the beginning (second exon) and end (fourth from last exon, accession BC060216) with plenty of supporting evidence for these exons being expressed and skipped. After combing through [Fagnani et al., 2007] data for a group of 11 genes. The genes examined are presented in Table 2.1.

TABLE 2.1: caption

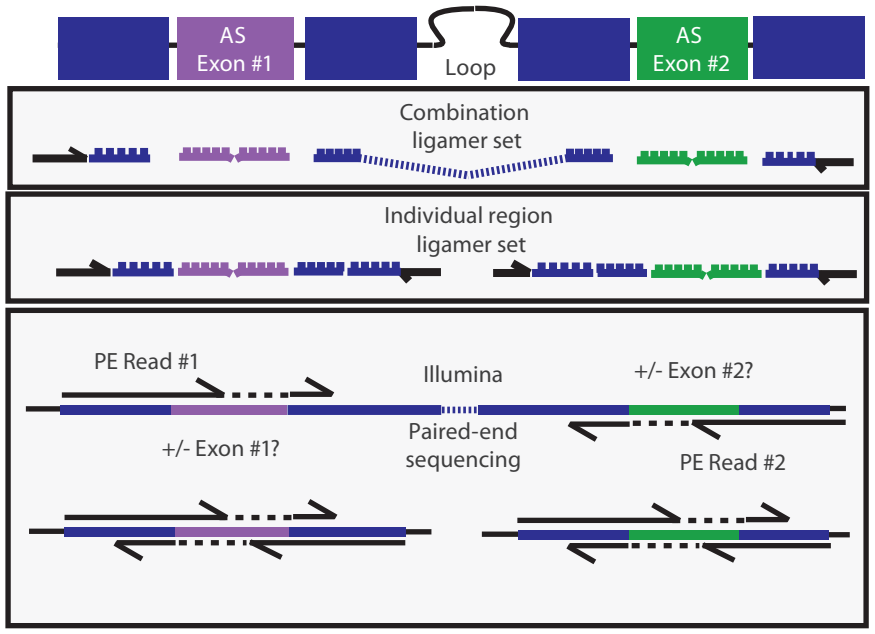
Gene name	nt mRNA between	possible isoforms	Exon 1	Exon 2
Chl1	4665	18	2	24
Mdm1	1846	4	EDA	IIICS
PTPRF-Y	1633	4	2	13
Cacna1c	1403	4	15	21/22
PTPRF-X	936	4	9/10	21
FN1	813	8	13/14	21/22
Apbb1	802	260	1/2b	2/3e
Agrn	736	8	33/34c	33/34a
Exoc7	513	4	7	13
Prom1	512	4	7	9
Lphn2	396	32	19	24/25a

I hand-designed ligamers to observe potentially coordinated splicing decisions. These oligos were then ordered from IDT in a 96-well plate format, pooled according to gene, and used to develop a multiplex approach to applying SeqZip, as well as investigate coordination between these exons, in these genes, using mouse total RNA from brains.

After attempts at performing the SeqZip assay on all 10 genes in one ligation failed, I reverted back to per-gene ligation reactions in order to trouble shoot and optimize the assay. Once I had obtained ligaiton products from per-gene ligation reactions for both the individual and combination ligamers pools, I pooled all the ligation products and amplified them via PCR. Amplified products were sent for PE100 sequencing on the Illumina GEIIx platform.

After considerable delay and optimization from the Umass Sequencing Core (likely due to little sequence diversity in the library), the analyzed data demonstrated little alternative splicing in the genes examined. Put another way- most of the transcripts observed via SeqZip were uniform in exon inclusion, and

A.



B.

		#1	
	(f)	(+) 0.50	(-) 0.50
#2	(+) 0.5	0.25	0.25
	(-) 0.5	0.25	0.25

		% Abundance	
#1	#2	Expected	Observed
+	+	25	50
+	-	25	0
-	+	25	0
-	-	25	50

FIGURE 2.2: 10 Gene set schematic
caption next

showed little variability for cassette exon inclusion. These results were disappointing, and forced us to rethink a better way to apply the SeqZip methodology to multiple genes, or complex alternative splicing (i.e. *Dscam1*; see 3. For a discussion of an "ideal" multiplex study, see section 5.2.2.

2.3 Determining RNA integrity using SeqZip

An exciting use of SeqZip would be rapid quantification of the integrity of RNA molecules. Integrity here is defined as the faction of molecules that are a continuous and unbroken nucleic acid polymers, from the original site of transcript to

3′ -processed end. Quantification of integrity has many uses including: quanlity control of RNA before downstream analysis such as RT or sequencing, and 2) (discussed below) implications of infectivity for viruses that package and RNA genome in the virion.

2.3.1 Demonstration of Concept

In order to demonstrate the feasibility of the SeqZip assay toward performing these type of analysis, I in vitro transcribed a 9,800 nt long RNA that I digested using $ZnCl_2$ at two different concentrations and times (see Figure 2.3. The RNA was probed using 3 ligamers, 2 to the very edges of the RNA, and one that looped out the intervening 8,000 nt. Theoretically, the amount of ligamer product observed would be directly tied to the abundance of the full length product.

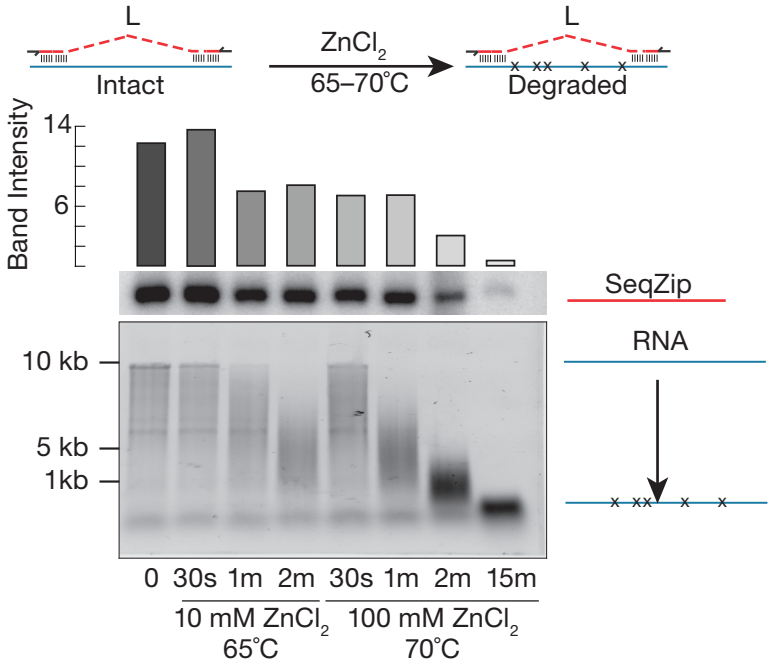


FIGURE 2.3: Ligation product tied to RNA integrity
Top) Schematic demonstrating the experimental design. A middle ligamer (L), that hybridizes to the edges of a 8,000 nt section of RNA should only ligate to flanking ligamers when the template RNA is intact, and not when hydrolyzed using $ZnCl_2$. Middle) Intensity of PCR products ampilified using end-labeled primers such that the intensities of all bands can be quantitatively compared (i.e. semi-quantitative PCR). Bottom) A denature agarose gel stained with EtBr showing the intactness of the template RNA used in position-matched ligation reactions in the middle panel. Time and concentrations of $ZnCl_2$ are shown.

Figure 2.3 produced promising results in that the aparent intesity of the bands shown in (middle) was tied to the amount of intact RNA seen in (bottom). However, the lanes where the RNA was degraded for 2m with 10 mM $ZnCl_2$ compared to 30s with 100 mM $ZnCl_2$ were not in good agreement, with clearly less intact RNA in the 2m lane, but just as much ligaiton product. We hypothesized that this was due to inherent secondary strucutre in the template we used (a section of the HIV genome, discussed below in section 2.3.2).

To see how well the SeqZip assay could distinguish a pool of containing RNA fragments from full-length templates, two different pools of ligamers were used on the same template as that used in Figure 2.3. At what concentratio of template does the "long" ligamer generate ligation products from fragments of the template message? Using pools of ligamers targeting fragments and a complete message, the SeqZip assay was performed using a 1:1 ratio of RNA fragments. The results of the experiment are shown in Figure 2.4.

Results shown that the SeqZip assay accurately reports on the presence of fragments, and not full length transcripts at concentrations less than 1 nM template. This is in good agreement with results presented in Chapter 3.

These results are encouraging, but bear repeating in order to address the issus of potential secondary stucture and repetitive regions inherent to the template RNA used. The should be repeated with a template RNA of mRNA original, instead of a highly-structured and repetive template such as the HIV genome used in the experiments described above.

2.3.2 Investigating HIV viral genome integrity using SeqZip

In late 2010, early 2011, a Graduate student in the Gottlinger lab, Anna Kristina Serquiña obervered that a cell line expressing ATPase-defective forms of the SF1 helicase UPF1 [**?**] did not infect a reporter cell line to the same degree as control. A previous Mass-spec study had reported MOV10, another SF1 family helicase [**?**] was packed in extracellular viral particles. She hypothezied that the decrease in infectivity was due to a problem with RT of the genetic material injected into the target cells. The results of this study were recently published [**?**].

Anna was interested in using the SeqZip methodology to quanitfy the amount of intact HIV virus in virus-producing cells and extracellular virions. The first step in applying SeqZip to HIV was to design ligamers.

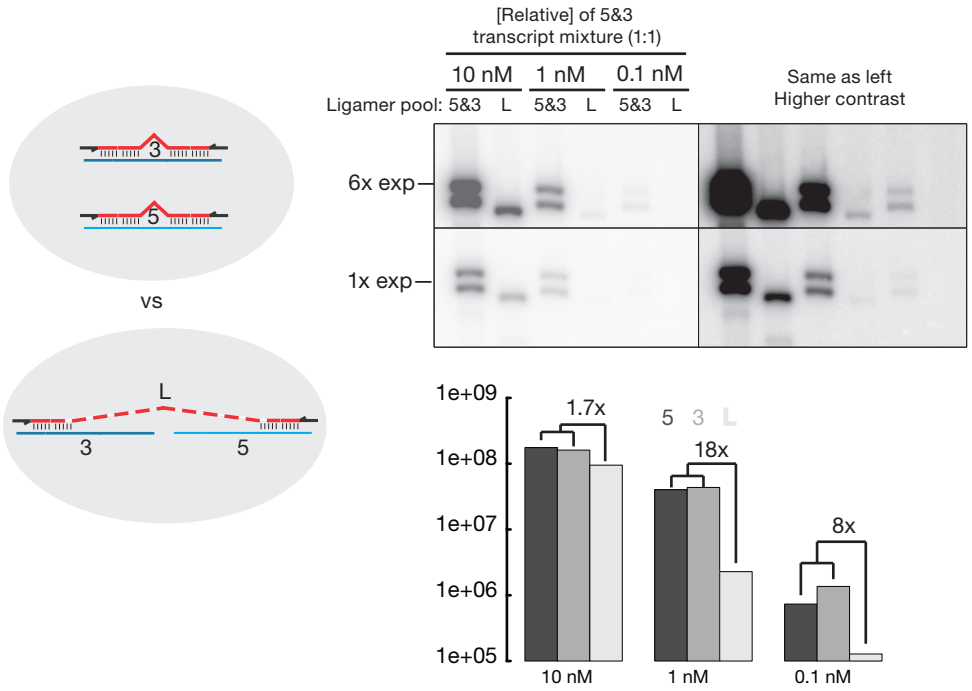


FIGURE 2.4: Trans Transcript investigation

Left) Schematic of experimental design: Three pools of ligamers were used. Two of them (labeled "3" and "5" hybridize to the 5′ and 3′ sections of a 9,800 nt template RNA. The last, labeled "L" connects these two regions via a long longer with target 5′ and 3′ regions of complementarity.

Right Top) Combinations of the ligamer pools were used with different concentrations of template RNA in the SeqZip assay. Ligation products were amplified with end-labeled PCR primers and amplified using radioactive PCR. Shown are low (left column) and high (right column) verions of two different exposure times (1x on bottom and 6x on top).

Right Bottom) Quantification of the bands shown in the gel above, grouped by input template RNA concentration. The fold difference in band intensity between the lowest signal "5" or "3" ligamer pool and the "L" pool is indicated. Y-Axis is the raw band intensity.

2.3.3 Design of HIV ligamers

Research into the integrity of the HIV RNA genome using SeqZip began with designing a set of ligamers against two different clones. The first clone, targeting transcripts from the M19921 plasmid (so called "M" clone), and transcripts from the K03455 clone contain nearly identical sequences with respect to the genome itself, and differ mostly in plasmid originating sequences. We targeted a difference in sequence for one site of ligation (Fig3-11A). Three different pools of ligamers were created initially: a Five(5) ligamer pool, with three ligamers design to test for the presence of sequence in the first 1,140 nt of the HIV genome,

importantly the first site of ligation in the 5 region pool should contain a mismatch in the K clone sequence; a three(3) pool, testing the last 1,210 nt of the genome, and a Long (L) ligamer pool, also containing three ligamers, but the middle ligamer of which would span the 5 and 3 regions, looping out 8,633 nt of sequence in the middle of the HIV genome. In vitro transcripts were created using both the K and M clone plasmids. These transcripts were added to a background of total MEF RNA, and the SeqZIp assay was performed. Ligation products were successfully amplified from all ligamer pools when using the M clone transcript and all three ligamer pools. Also the abundance of these ligation products, as measured by endpoint PCR, seemed to be spike-concentration dependent. Notably, Ligation products were not obtained from the K clone using either the 5 or L ligamer pools, likely due to the mismatch between the transcript and the ligamers at the site of ligation. Also of note was the appearance of ligation products from purified endogenous virons of the M clone from all three ligamer pools, and the absence of products from virions purified from plasmids containing a defective protein, Gag, essential for viral packaging.

The results shown in Figure 2.11 clearly show that the SeqZip assay, and these three pools of ligamers can be used to profile HIV in vitro transcripts, and RNA from puried virions. Important features of the figure are 1) Ligation products are *not* observed for ligation reactions using the K clone template RNA and the Five(5) pool of ligamers, verifying the specificty of the ligamers to the different base containined in the M clone; and 2) That the amount of product from reactions using the L pool of ligamers required more cycles (22 vs 12) in order to be visualized, as would be expected given the physical constrain of hybridizing to two sequences seperated by >8,000 nt.

Again, while these results were encouraging, access to purified material and a general push to publish Anna's UPF1 story lead the Gottlinger lab to substatiate the viral genome intregrity claims effecting infectivity using a traditional northern blot [**?**]. However, this work clearly warrents additional optimization and application

2.4 Continuity of piRNA precursor transcripts

Write more SeqZip experiment work

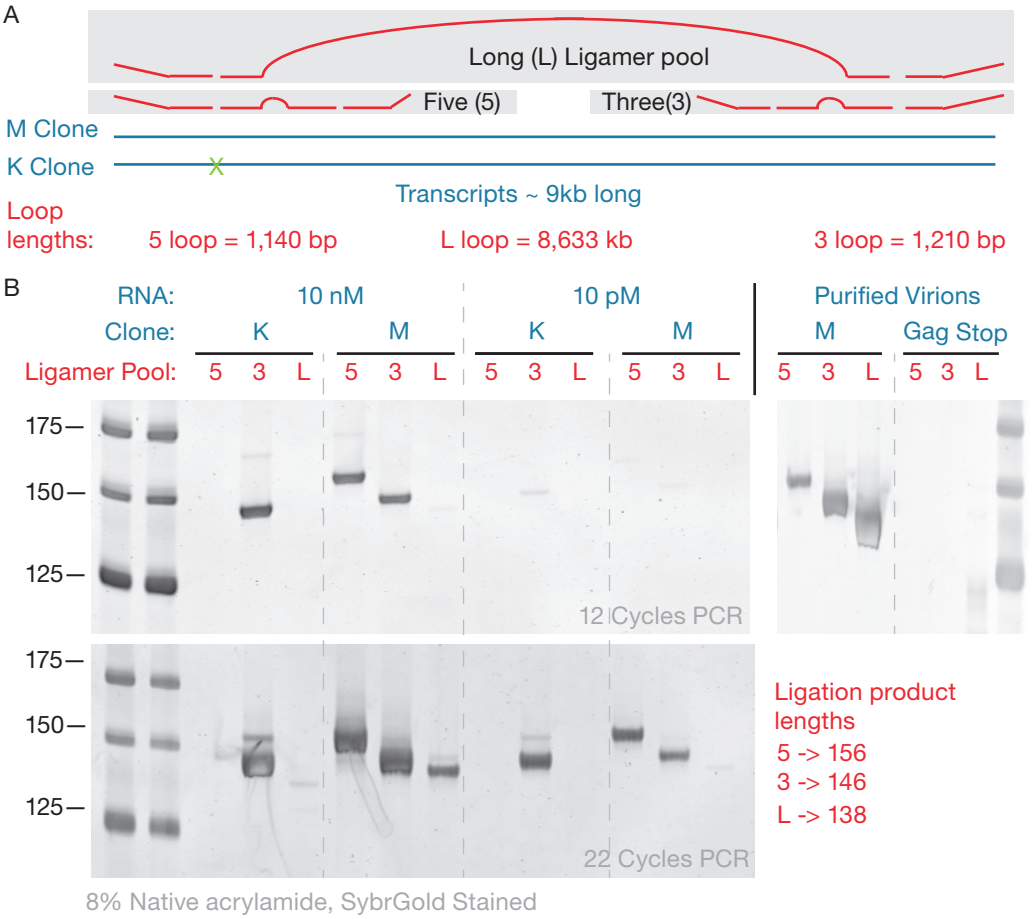


FIGURE 2.5: SeqZip can examine HIV transcript integrity
A) Schematic demonstrating the experimental design. Three different pools are used to probe for connectivity on the 5′ (Five(5)) and 3′ (Three(3)) ennds. Additionally, a Long (L) ligamer is used to check for connectivity between the two ends. We used two different clones of the HIV genome, described in the text and denoted as “M” and “K”. Important here is that the “K” contains difference base at a ligation site of the 5 ligamer pool. B) A series of end-point PCR gels showing amplified ligation products templated with in vitro transcribed RNA at 10 nM or 10 pM of either the K or M clones, or from purified virions of (M clone origin). Show are two different end points of PCR, 12 cycles (top) or 22 cycles (bottom). Also shown is a legend of expected ligation products lengths

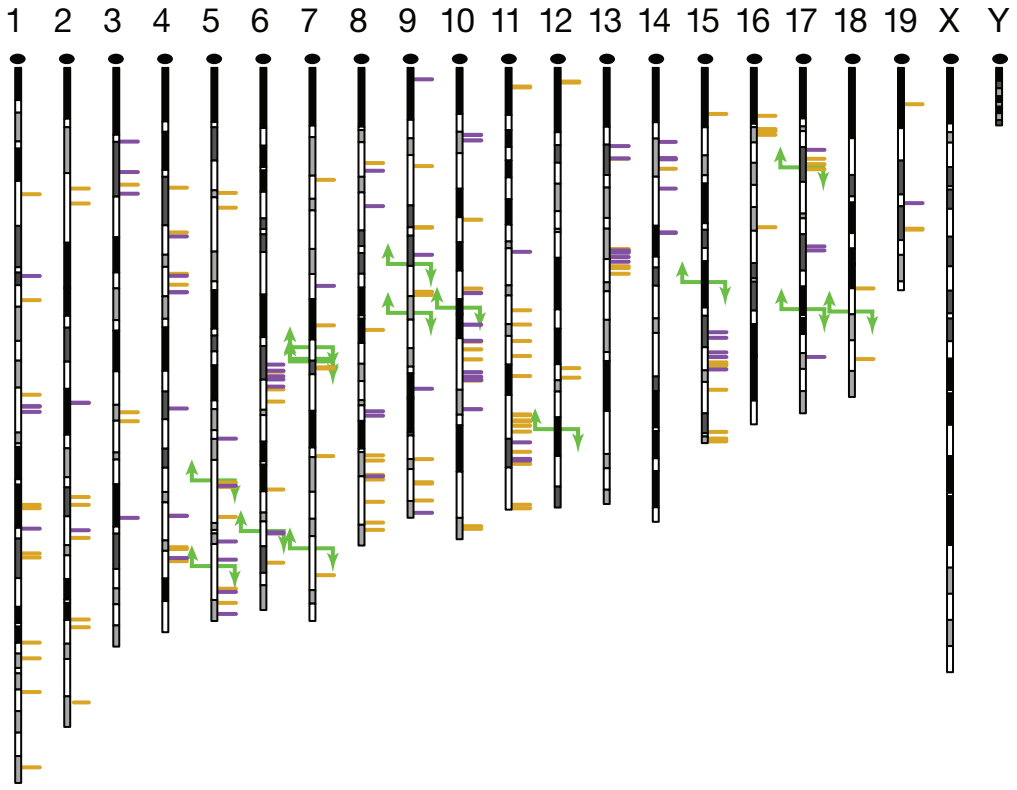


FIGURE 2.6: piRNA precursor locations
figure Caption

TABLE 2.2: caption

Cluster Name	Matched Cluster	Unique-mapping piRNAs @ wt.14dpp	Fraction of pachytene piRNAs	Cumulative pachytene piRNAs
17-qA3.3-26735.1	17-qA3.3-27363	3,021,022	17.2	17.2
17-qA3.3-27363.1	17-qA3.3-26735	1,742,695	9.9	27.2
9-qC-31469.1	9-qC-10667	1,006,333	5.7	32.9
9-qC-10667.1	9-qC-31469	272,385	1.6	34.5
7-qD2-24830.1	7-qD2-11976	652,564	3.7	38.2
7-qD2-11976.1	7-qD2-24830	280,312	1.6	39.8
6-qF3-28913.1	6-qF3-8009	564,930	3.2	43.0
6-qF3-8009.1	6-qF3-28913	180,210	1.0	44.0
2-qE1-35981.1	NA	1121042	6.4	50.4

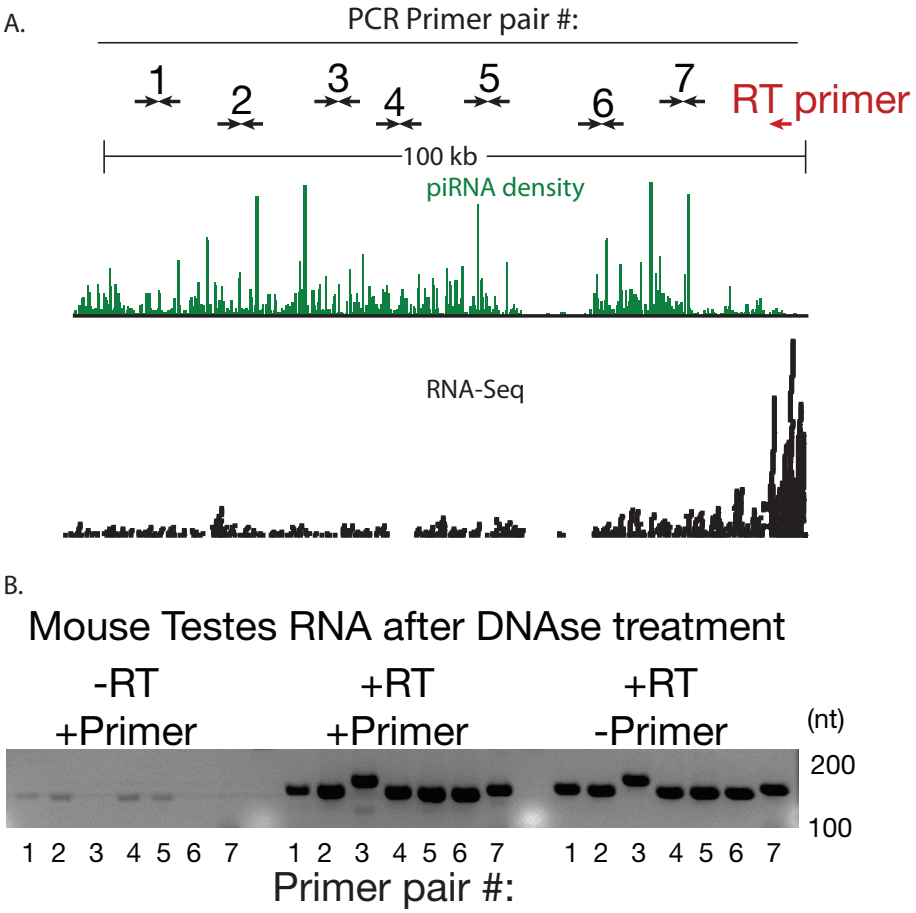


FIGURE 2.7: RT Doesn't Work for piRNA precursors
figure Caption

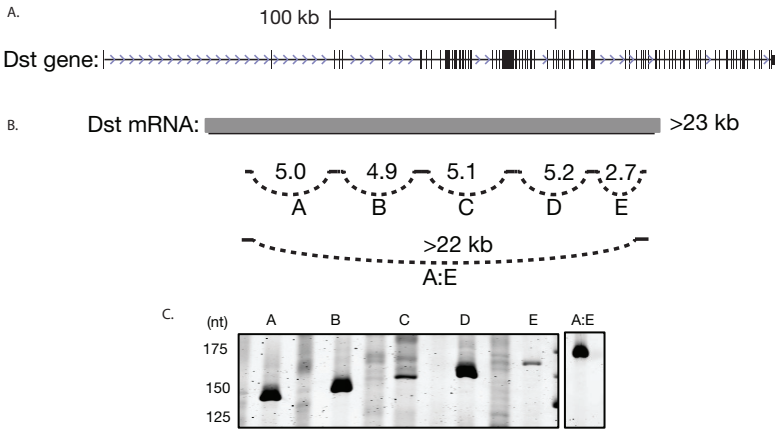


FIGURE 2.8: Dst1 by SeqZip
figure Caption

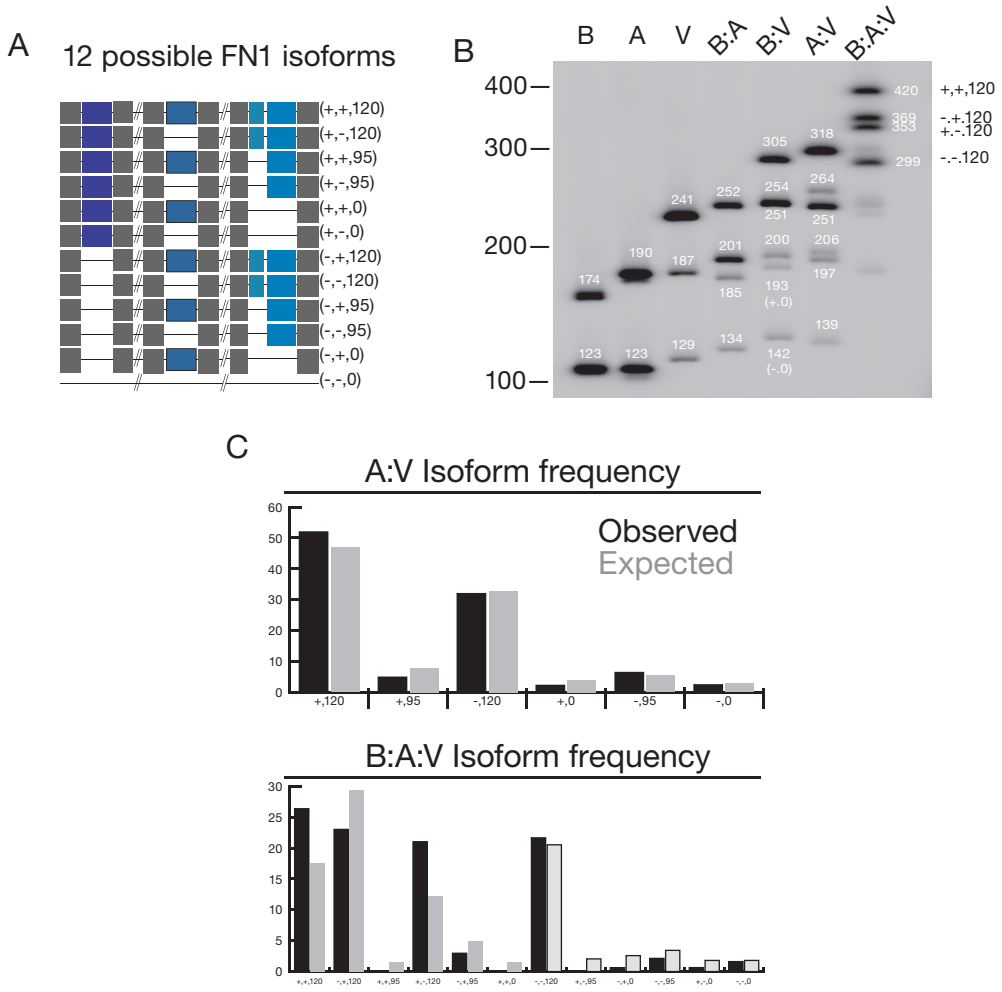


FIGURE 2.9: Capturing AS at three sites and >2kb of mRNA
figure Caption

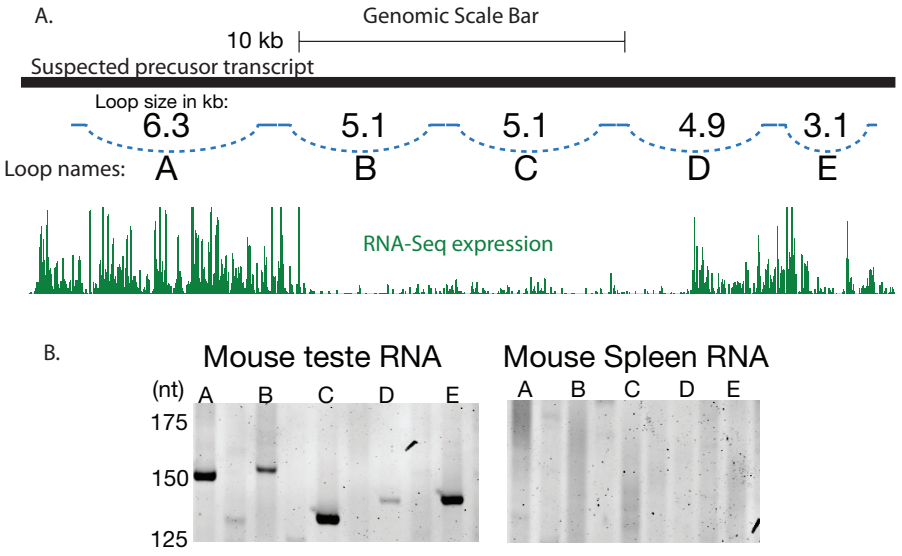


FIGURE 2.10: Testes Specific RNA precursor expression
figure Caption

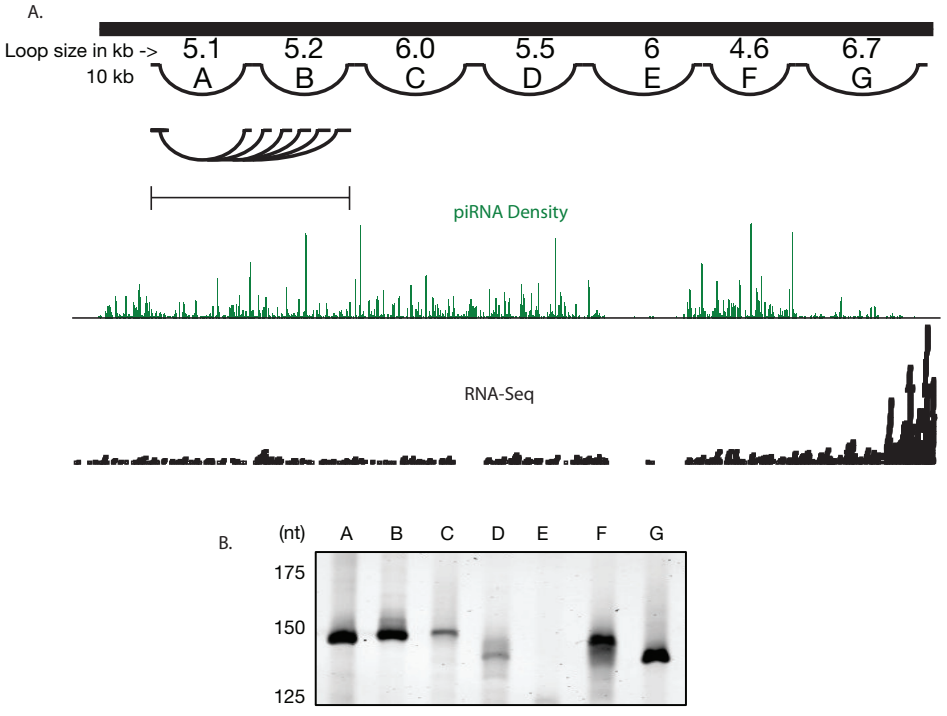


FIGURE 2.11: piRNA precursor analysis via SeqZip
figure Caption

Chapter 3

SeqZip Publication

3.1 Abstract

3.2 Introduction

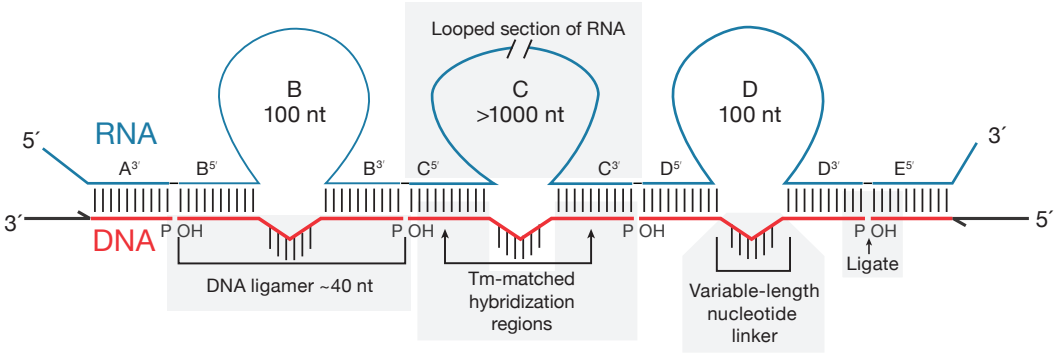


FIGURE 3.1: SeqZip Diagram]
Insert Figure Text

3.3 Results

3.3.1 Subsection 2

]

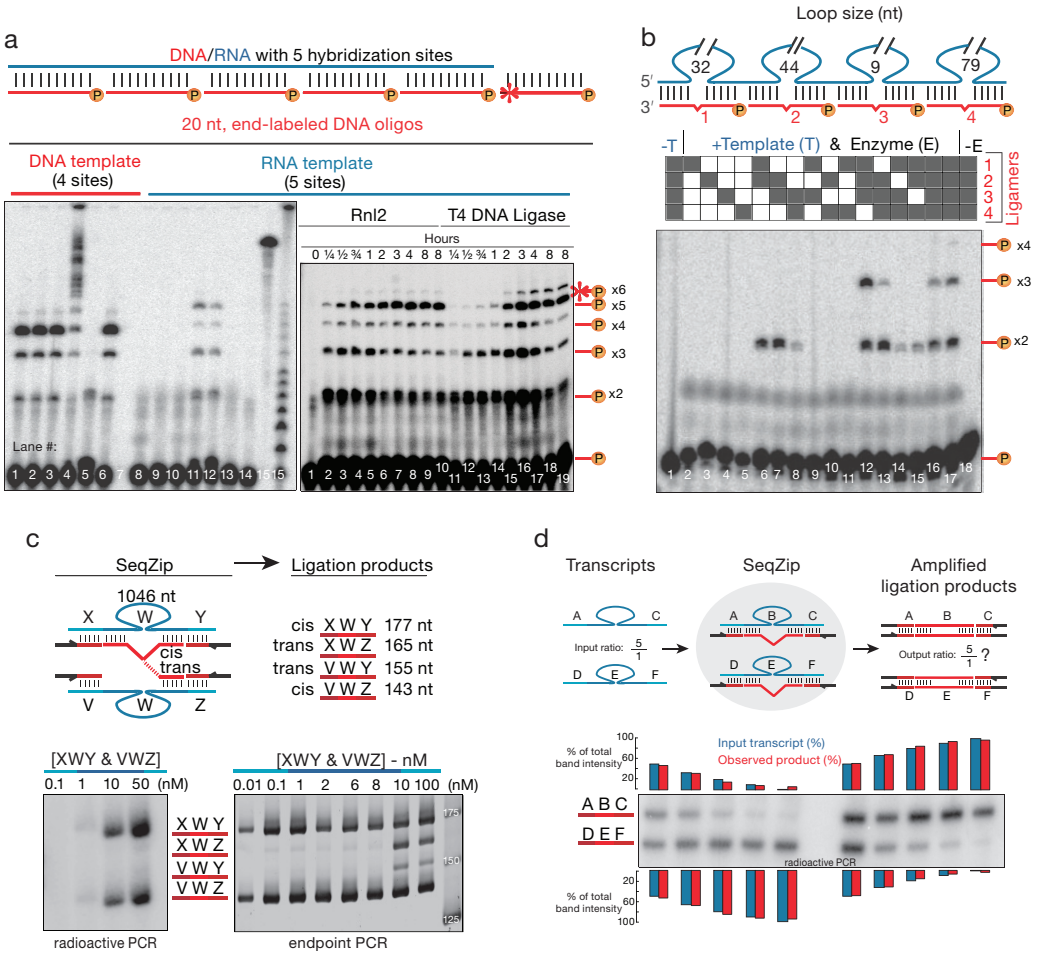


FIGURE 3.2: Roy et al Figure 2]
Insert Figure Text

3.4 Discussion

3.5 Methods

3.6 Supplemental Text

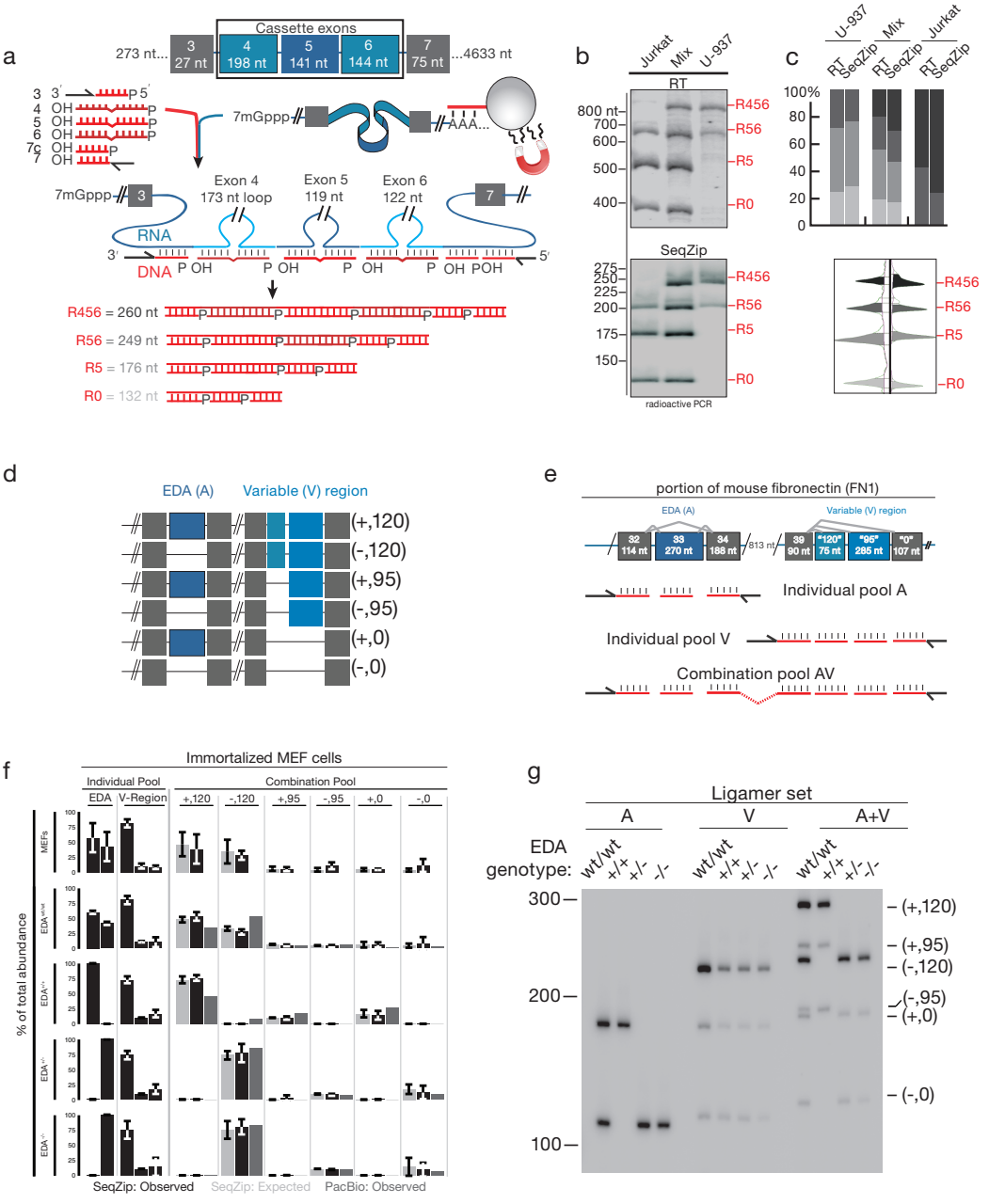


FIGURE 3.3: Roy et al Figure 3l]
Insert Figure Text

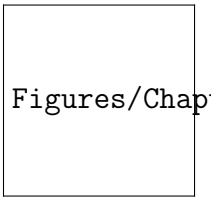


FIGURE 3.4: Rnl2 Panel]
Insert Figure Text

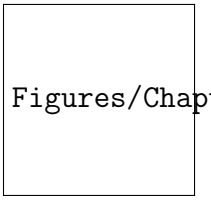


FIGURE 3.5: Roy et al Figure 5]
Insert Figure Text

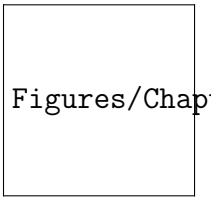


FIGURE 3.6: Roy et al Figure 6]
Insert Figure Text

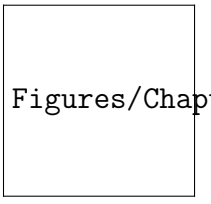


FIGURE 3.7: Roy et al Figure 7 Insert Figure Text

Chapter 4

An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes

How do I include the supplemental tables?!

4.1 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 [Cenik and Zamore, 2011, Farazi et al., 2008, Kim et al., 2009, Kumar and Carmichael, 1998, Thomson and Lin, 2009, **?**]. 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 [Houwing et al., 2007, Vagin et al., 2006].

In most bilateral animals, germline piRNAs protect the genome from transposon activation, but also have other functions [Aravin and Hannon, 2008, Aravin et al., 2007a, 2001, Ashe et al., 2012, Brennecke et al., 2007, Carmell et al., 2007, Hartig et al., 2007, Kuramochi-Miyagawa et al., 2008, Lee et al., 2012,

Shirayama et al., 2012, Vagin et al., 2004]. 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 [Aravin et al., 2001, Deng and Lin, 2002, 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., 2012]. 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, Brennecke et al., 2007, Girard et al., 2006, Grivna et al., 2006, Lau et al., 2006, 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, Brennecke and Malone, 2008, Vourekas et al., 2012] and protein-coding genes that also produce piRNAs [Aravin et al., 2007b, 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 [Bolcun-Filas et al., 2011, Trauth et al., 1994]. 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.

4.2 RESULTS

4.2.1 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 4.1A). 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].

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

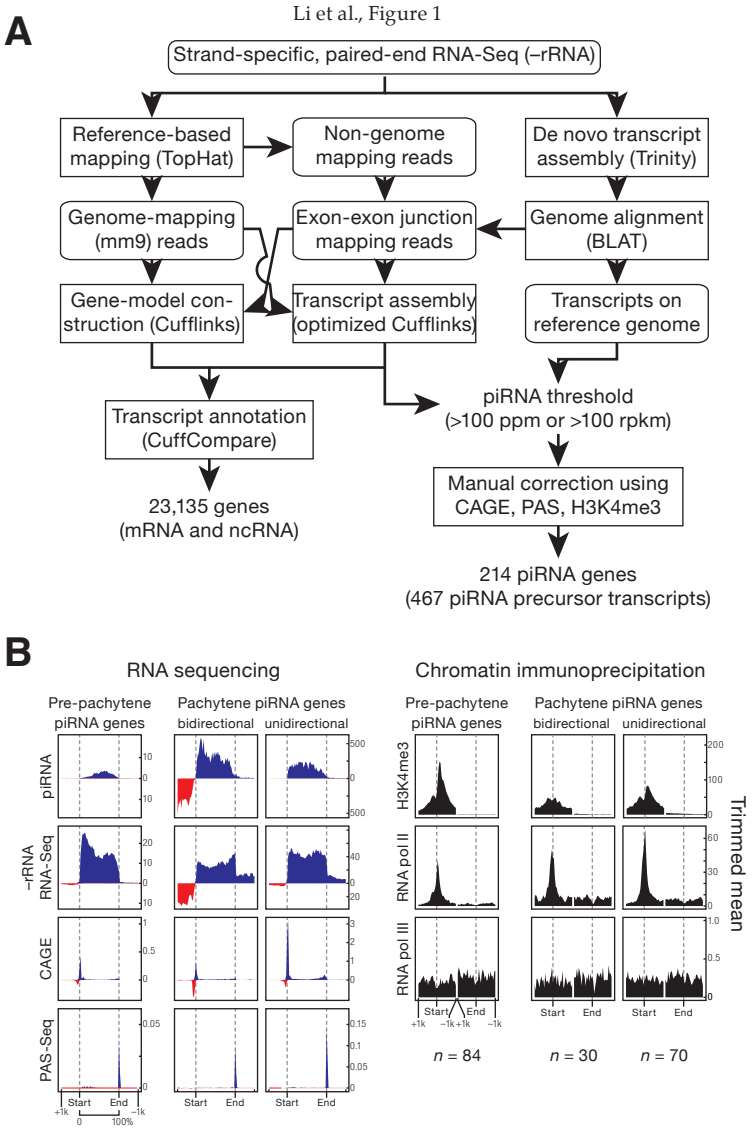


FIGURE 4.1: piRNA Precursors are RNA Pol II Transcripts

(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 4.2.

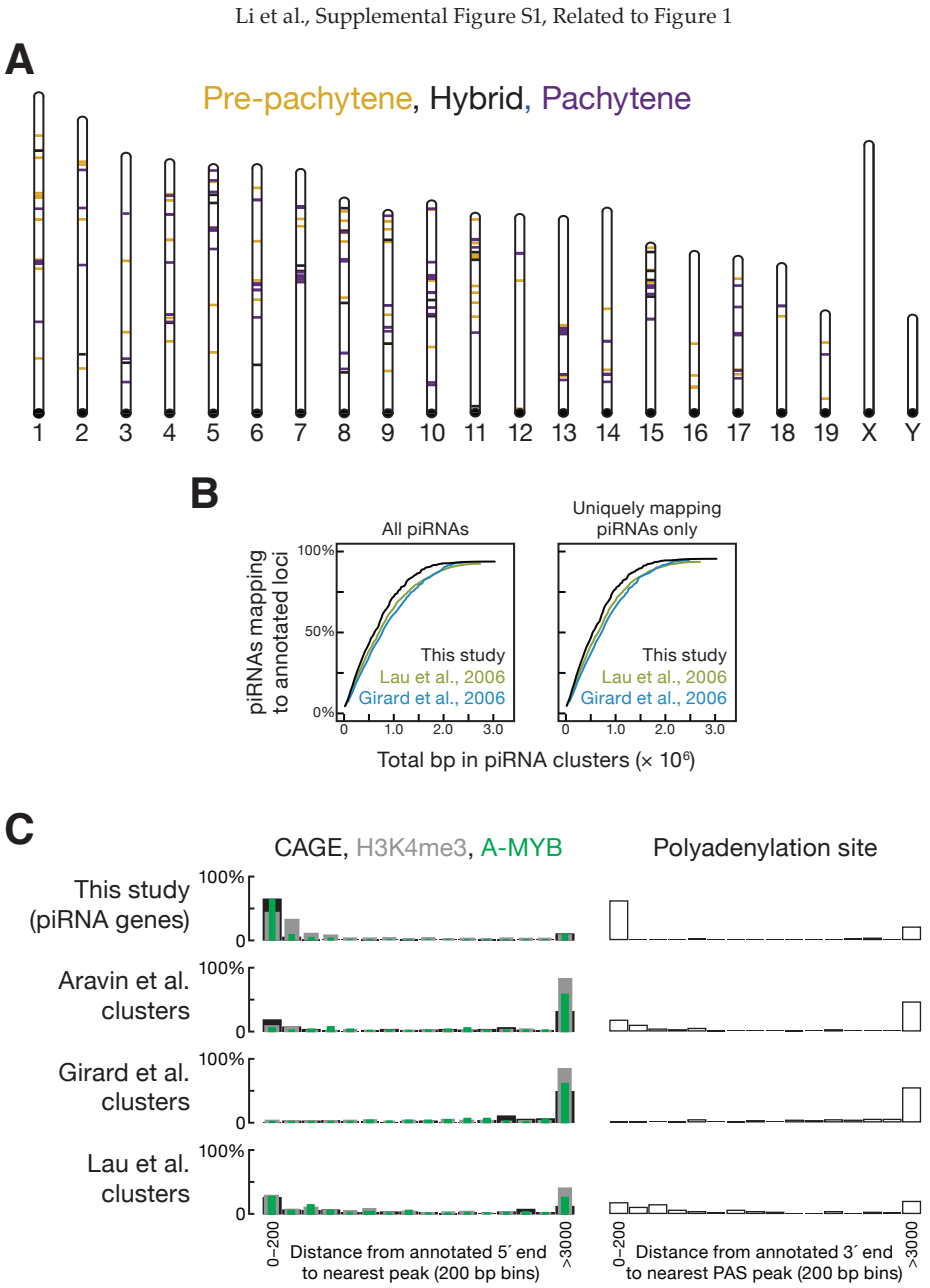


FIGURE 4.2: (A) Positions of the 214 major piRNA-producing genes on the 19 autosomes of mice. We detected no loci on the X or Y chromosomes. (B) Cumulative distributions for all piRNAs and for uniquely mapping piRNAs comparing the piRNA loci defined by our methods and by previous approaches [Girard et 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.

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 4.1B). 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].

4.2.2 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 4.1. 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 4.1B; [Kutter et al., 2011]. We conclude that piRNA transcripts are conventional RNA pol II transcripts bearing 5′ caps and 3′ poly(A) tails.

4.2.3 A Transcript-based Set of piRNA Loci

Our transcriptome assembly yielded 467 piRNA-producing transcripts that define 214 genomic loci (Figure 4.2A 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 [Aravin et al., 2007a, Girard et al., 2006, Lau et al., 2006]. 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 4.2B and 4.2C; [Girard et al., 2006, Lau et al., 2006]. Our piRNA-producing loci include 41 piRNA loci that escaped previous detection [Aravin et al., 2007a, Girard et al., 2006, Lau et al., 2006], 37 of which contain introns. The 41 loci account for 2% of piRNAs at 10.5 dpp and 0.36% in the adult testis.

4.2.4 Three Classes of piRNAs During Post-Natal Spermatogenesis

Mice produce three PIWI proteins: MIWI2 (PIWIL4), which binds piRNAs in perinatal testis [Aravin and Hannon, 2008, Carmell et al., 2007]; MILI (PIWIL2), which binds piRNAs at least until the round spermatid stage of spermatogenesis [Aravin et al., 2006, 2007a, Kuramochi-Miyagawa et al., 2004]; and MIWI (PIWIL1), 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 [Aravin et al., 2006, Aravin and Hannon, 2008, Montgomery et al., 1998, Robine et al., 2009], to MIWI-bound piRNAs, which have a 30 nt modal length (Figure 4.4A; [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 4.3A and 4.4A 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 4.3B). 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.

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

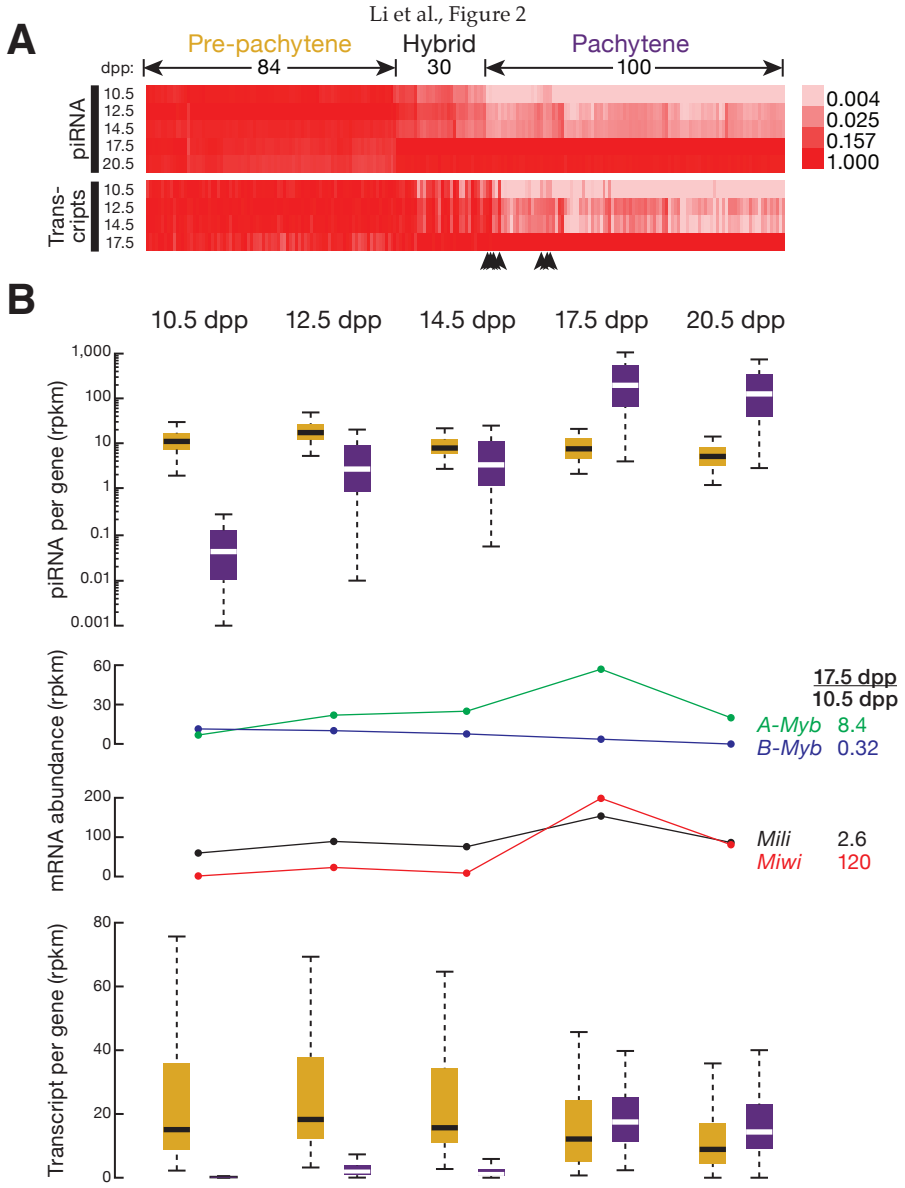


FIGURE 4.3: (A) Normalized piRNA density (rpkm) for each piRNA-producing gene is shown as a heatmap across the developmental stages. Hierarchical clustering divided the genes into three classes. Arrowheads mark seven pachytene piRNA genes that were not classified as pachytene according to the change 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 4.4 and Table S2.

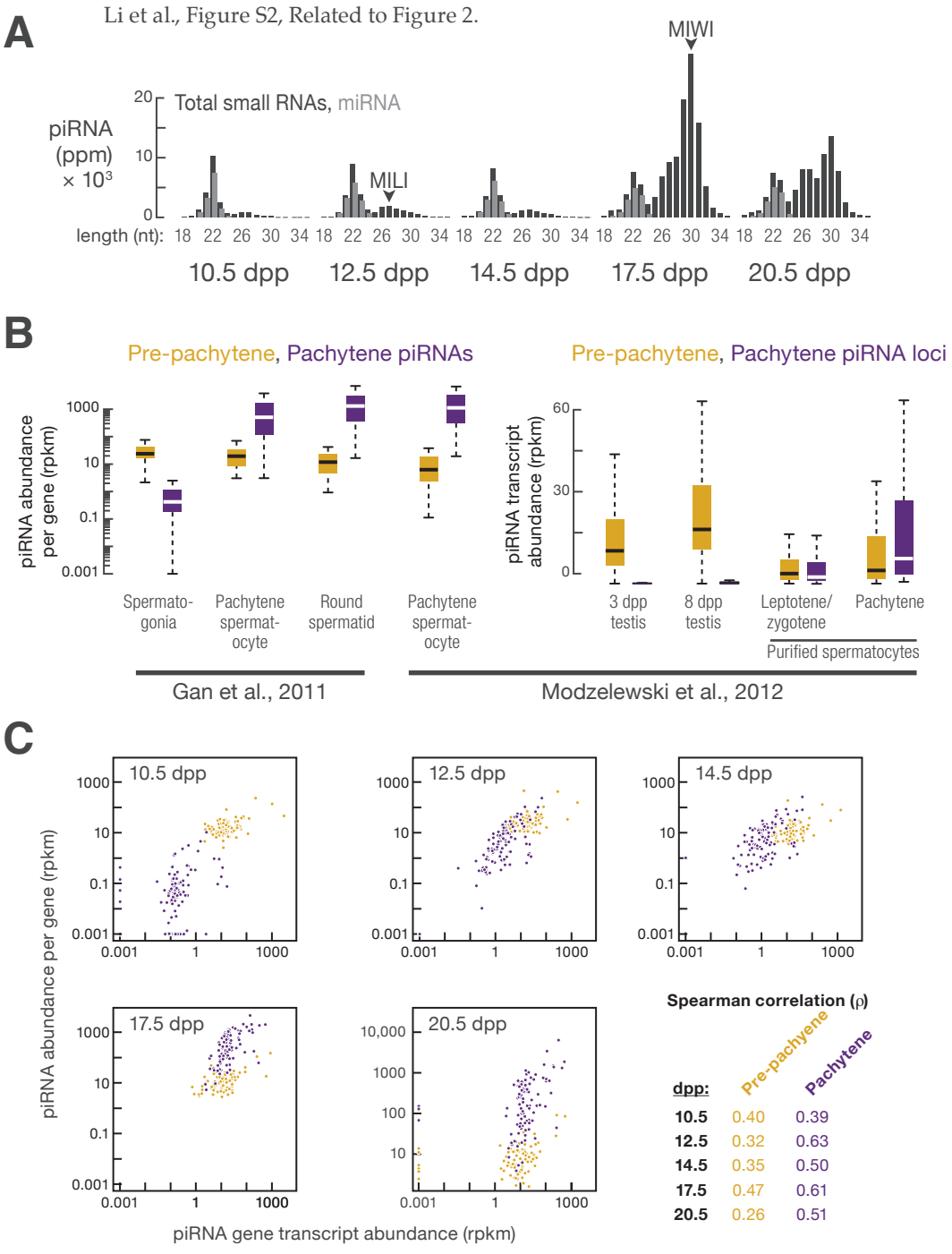


FIGURE 4.4: (A) As shown previously by others using lower temporal resolution, the modal length of piRNAs increases as spermatogenesis proceeds to more advanced stages. (B) Total piRNA rpkm abundance and piRNA transcript abundance per locus by class, from purified spermatogonia, spermatocytes, round 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 Figures, gold indicates pre-pachytene and purple indicates pachytene piRNA loci.

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 4.4B; [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; 4.4B). 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 4.4B); [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 4.4B).

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 ($0.32 \geq \rho \leq 0.40$; Figure 4.4C). 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 4.3B). 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 ($0.39 \geq \rho \leq 0.63$; $pvalue \leq 7.3x10-5$; Figure 4.4C). 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 **??** and 4.6 provide examples of pachytene and pre-pachytene piRNA genes defined by our data.

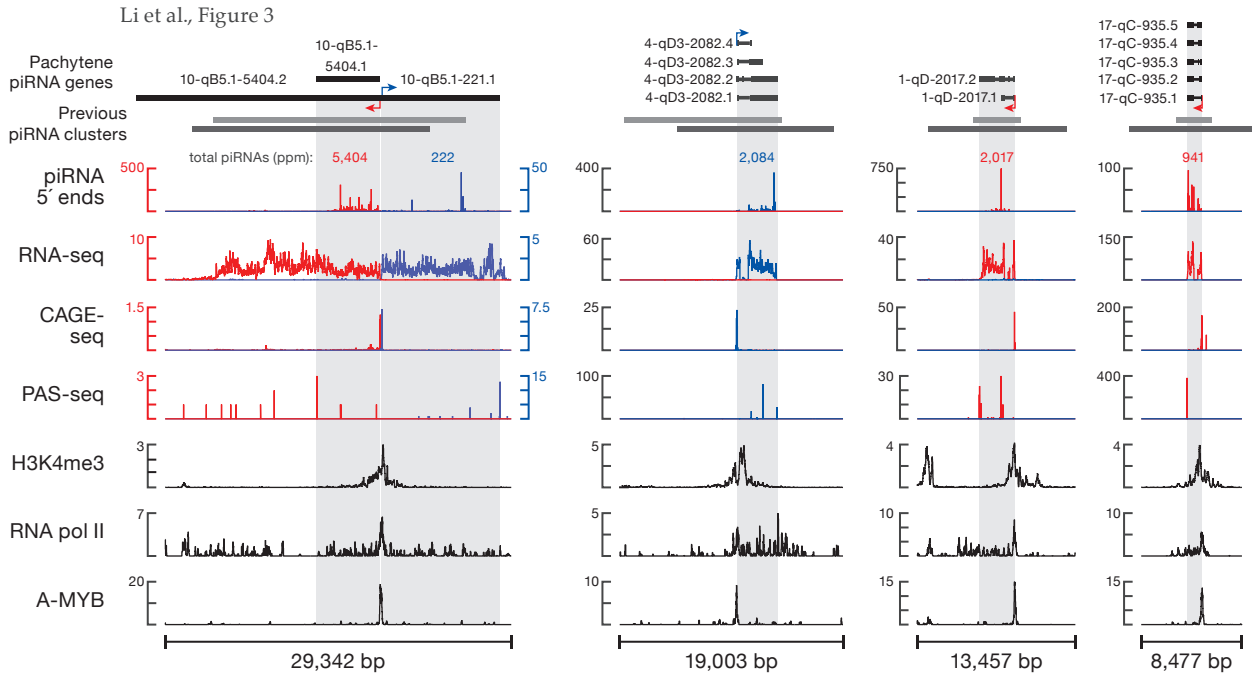


FIGURE 4.5: Previous cluster boundaries are from Lau et al. [2006] in gray and Girard et al. [2006] in dark gray).

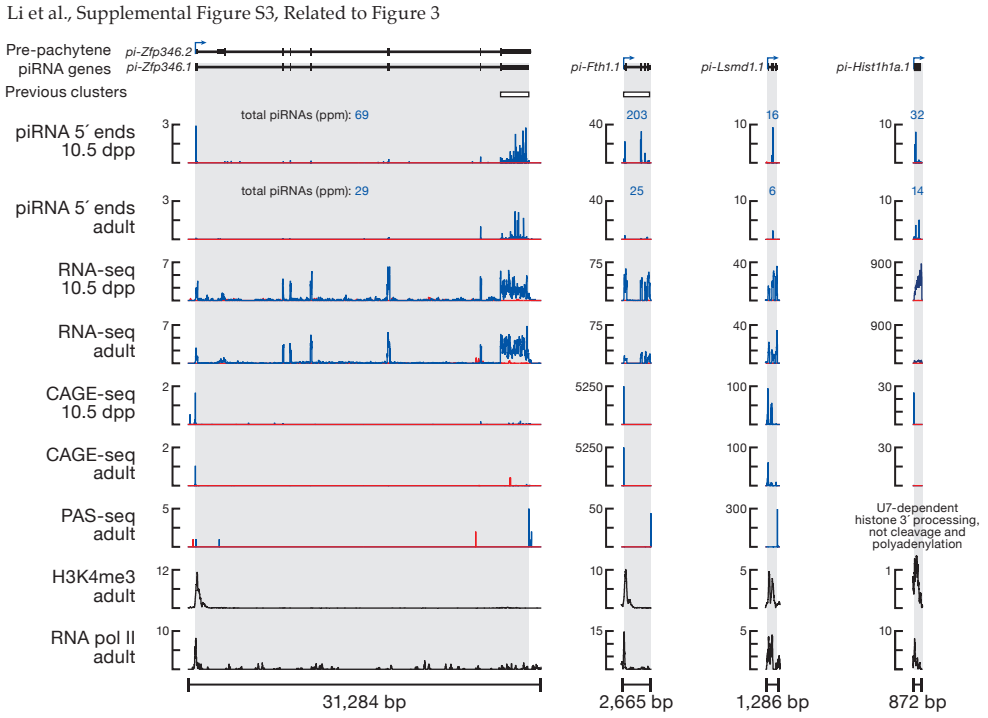


FIGURE 4.6: Previous cluster boundaries are from Lau et al. [2006] in gray and Girard et al. [2006] in dark gray).

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]).

4.2.5 *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 x 10^{12}; Figure 4.7A). This motif matches the binding site of the Myb family of transcription factors (Figure 4.7A; [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 x 10^{-28}; Figure 4.7B).

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 [Latham et al., 1996, Mettus et al., 1994, Oh and Reddy, 1999, Trauth et al., 1994]. 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

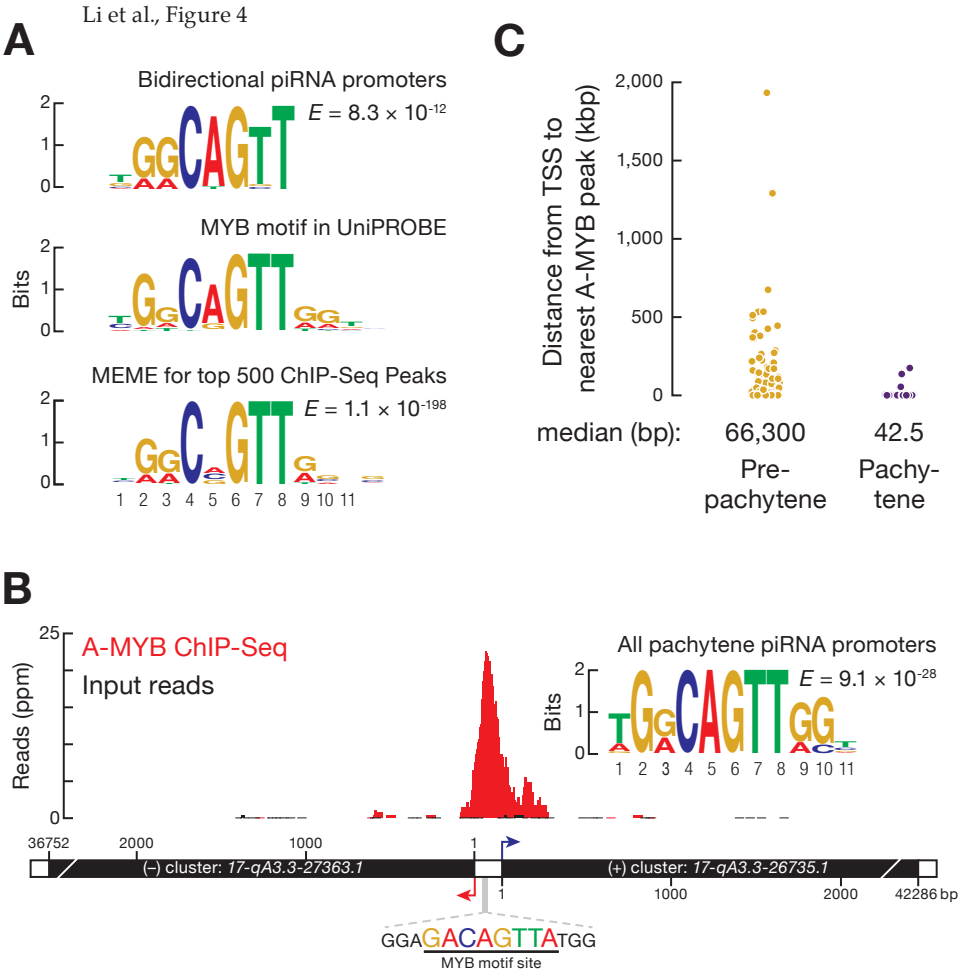


FIGURE 4.7: (A) Top: MEME identified a sequence motif in the bidirectional promoters of the 15 pairs of divergently transcribed pachytene piRNA genes. E value computed by MEME measures the statistical significance of the motif. Middle: Myb motif from the mouse UniPROBE database. Bottom: MEME-reported 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 4.8.

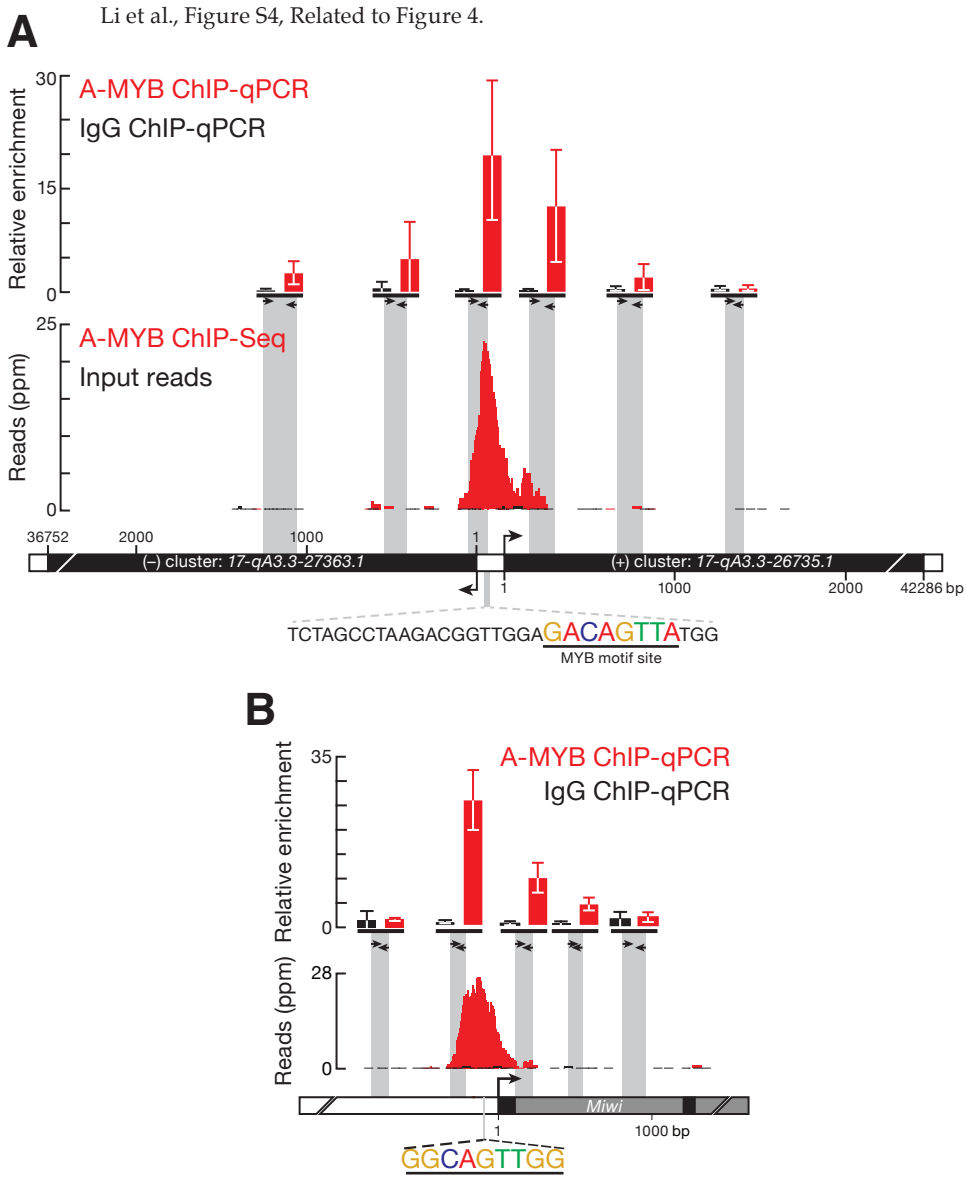


FIGURE 4.8: (A) A-MYB binds to the common promoter of divergently transcribed pachytene piRNA loci *17-qA3.3-27363.1* and *17-qA3.3-26735.1*. The abundance of DNA fragments at the amplified region relative to a control region (mean ± standard deviation; n = 3) was measured by qPCR (top). The A-MYB ChIP-seq (red) and input (black) data for this pair of genes is presented as in Figure 4.7B. (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.

4.3B; [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 4.3B) 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, *Mybl1*^{repro9}, 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 [Bolcun-Filas et al., 2011, Toscani et al., 1997]. 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 ($\rho < 10^4$) for the MYB binding motif (Figure 4.7A). Figure 4.7B 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 4.8A).

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 4.7C). 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 (*Mybl1*^{repro9}) mice and their heterozygous littermates (Figure 4.9). 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 x 10^{-13}) and 248-fold at 17.5 dpp (q

= 3.9 x 10^{-23}). The abundance of pre-pachytene piRNA transcripts was not significantly changed (q $\geq$ 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.

4.2.6 *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 ($\rho < 2.2X10^{-16}$; Figure 4.9. By 17.5 dpp, median pachytene piRNA abundance was >9,000 times lower in the *A-Myb* mutant than the heterozygotes (P < 2.2 x 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.

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 4.10). 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 4.12). We conclude that the loss of pachytene piRNA

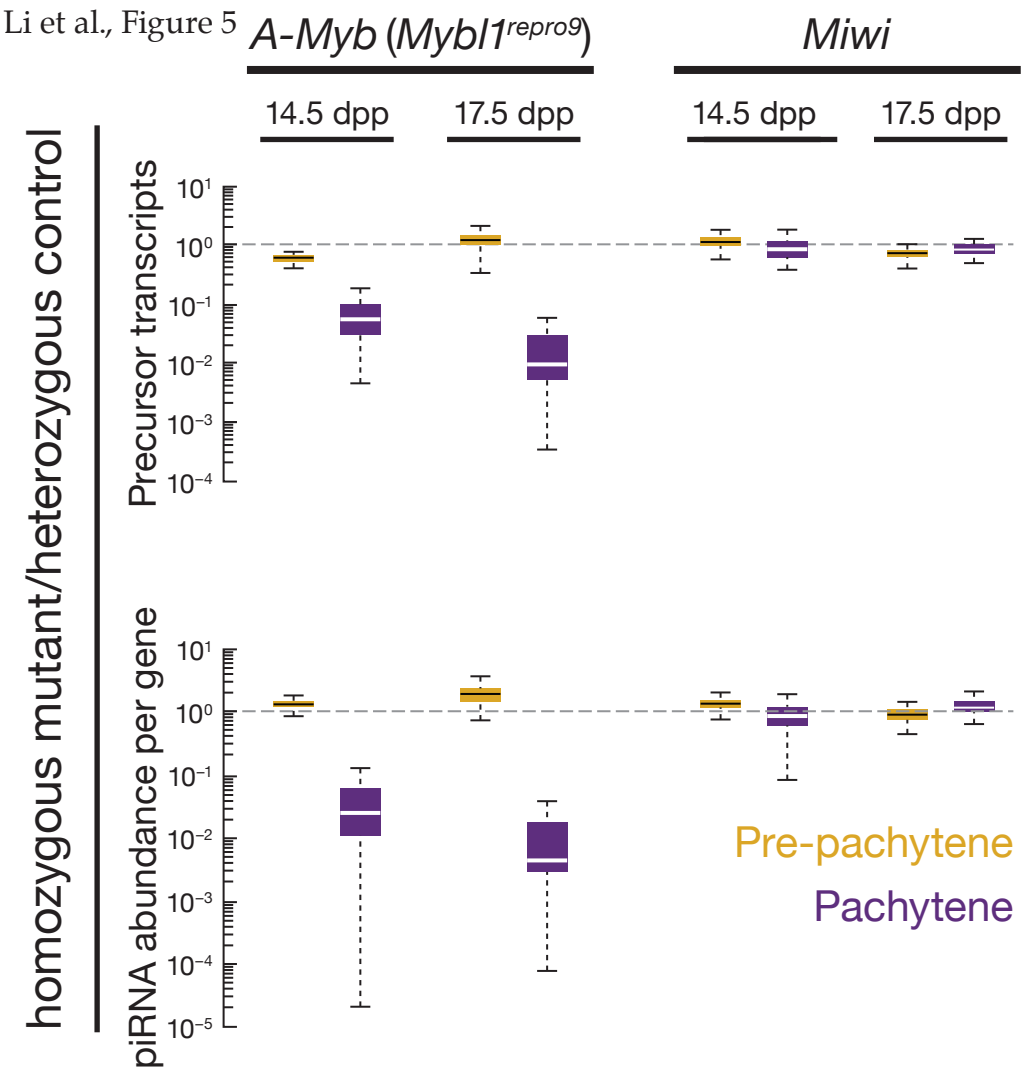


FIGURE 4.9: The change in transcript or piRNA abundance per gene in *A-Myb* (n = 3) and *Miwi* (n = 1) mutants compared to heterozygotes in testes isolated at 14.5 and 17.5 dpp. See also Figure 4.10.

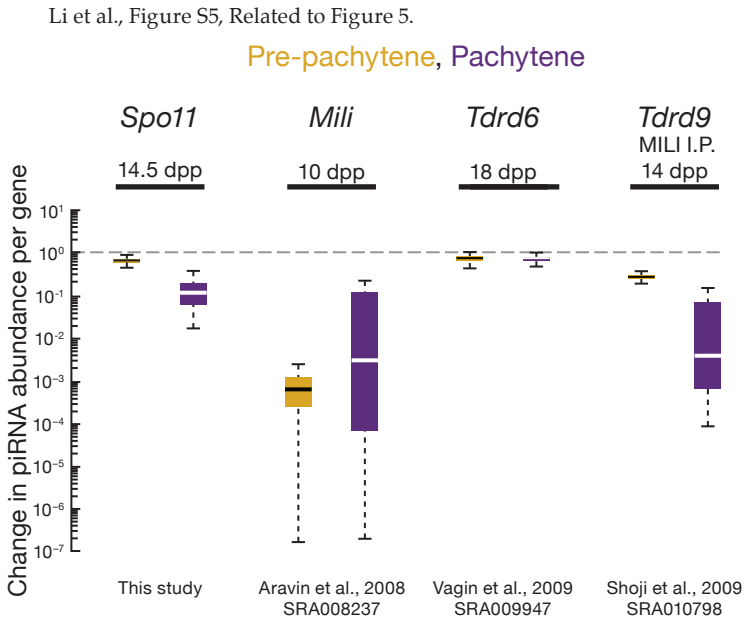


FIGURE 4.10: Change in piRNA abundance per locus (rpkm) for *Spo11* (14.5 dpp), *Miwi* (*Piwil2*; 10.5 dpp), *Tdrd6* (18 dpp), and *Tdrd9* (14 dpp) mutants compared to heterozygous controls.

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.

4.2.7 *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 4.9; 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 4.13A and 4.14 and Table S3). Furthermore, our ChIP data confirm that A-MYB binds the *Miwi* promoter in vivo (Figures 4.13B, 4.8B, and 4.14). Like pachytene piRNAs, *Miwi* transcripts first appear at 12.5 dpp (Figure 4.3B), 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].

A previous study reported that piRNAs fail to accumulate to wild-type levels in *Miwi* mutant testes [Grivna et al., 2006]. 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 x 10-6) in *Miwi* mutant testes (Figure 4.9). 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 4.9). 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 4.13B 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 x 10-5) and *Tdrd5* (7.5-fold decrease; q = 1.0 x 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]

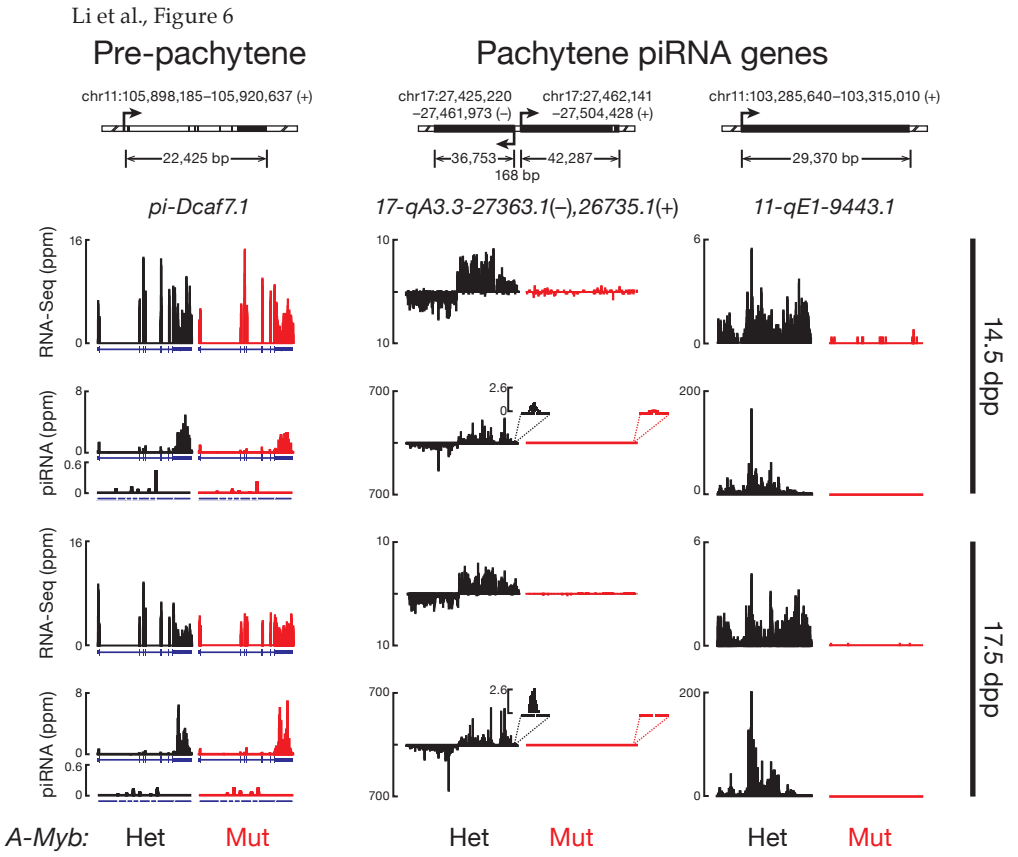


FIGURE 4.11: Transcript and piRNA abundance in heterozygous (Het) and homozygous *A-Myb* (Mut) point-mutant testes is shown for four illustrative examples at 14.5 and 17.5 dpp. Also shown is the abundance of piRNA sequencing reads that map to the exon-exon junctions. Gene *11-qE1-9443* does not have an intron. Exons, blue boxes; splice junctions, gaps; the last exon is compressed and not to scale. See also Figure 4.12.

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 4.6; [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

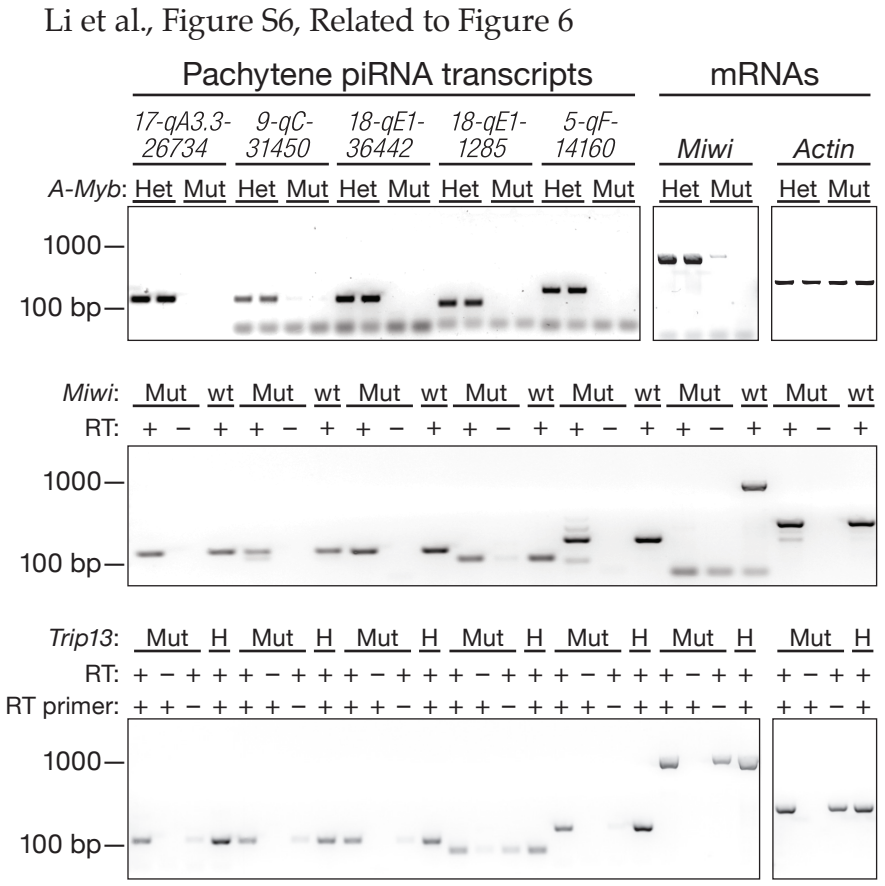


FIGURE 4.12: Transcripts were detected in total RNA from adult testes by RT-PCR (using random primers) for five pachytene piRNA loci as well as *Miwi* and *Actin*. Mut, mutant; Het or H, heterozygote; wt, wild type.

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 [Haase et al., 2010, Houwing et al., 2007, Huang et al., 2011, Ipsaro et al., 2012, Nishimasu et al., 2012, Pane et al., 2007, Watanabe et al., 2011a]. 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 and Hosokawa, 2006]. In Tdrd1 mutant testes, spermatogenesis fails, with no spermatocytes progressing past the round spermatid stage [Chuma and Hosokawa, 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.

4.2.8 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 4.13B and 4.14C). 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 4.13C). 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 [Osella et al., 2011, Shen-Orr et al., 2002].

A-MYB also bound to the *A-Myb* promoter (Figure 4.13B), and *A-Myb* transcripts decreased 4.2-fold in testes from an *A-Myb* point mutant (*Mybl1*^{repro9}; Figure 4.13B). 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.

4.2.9 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

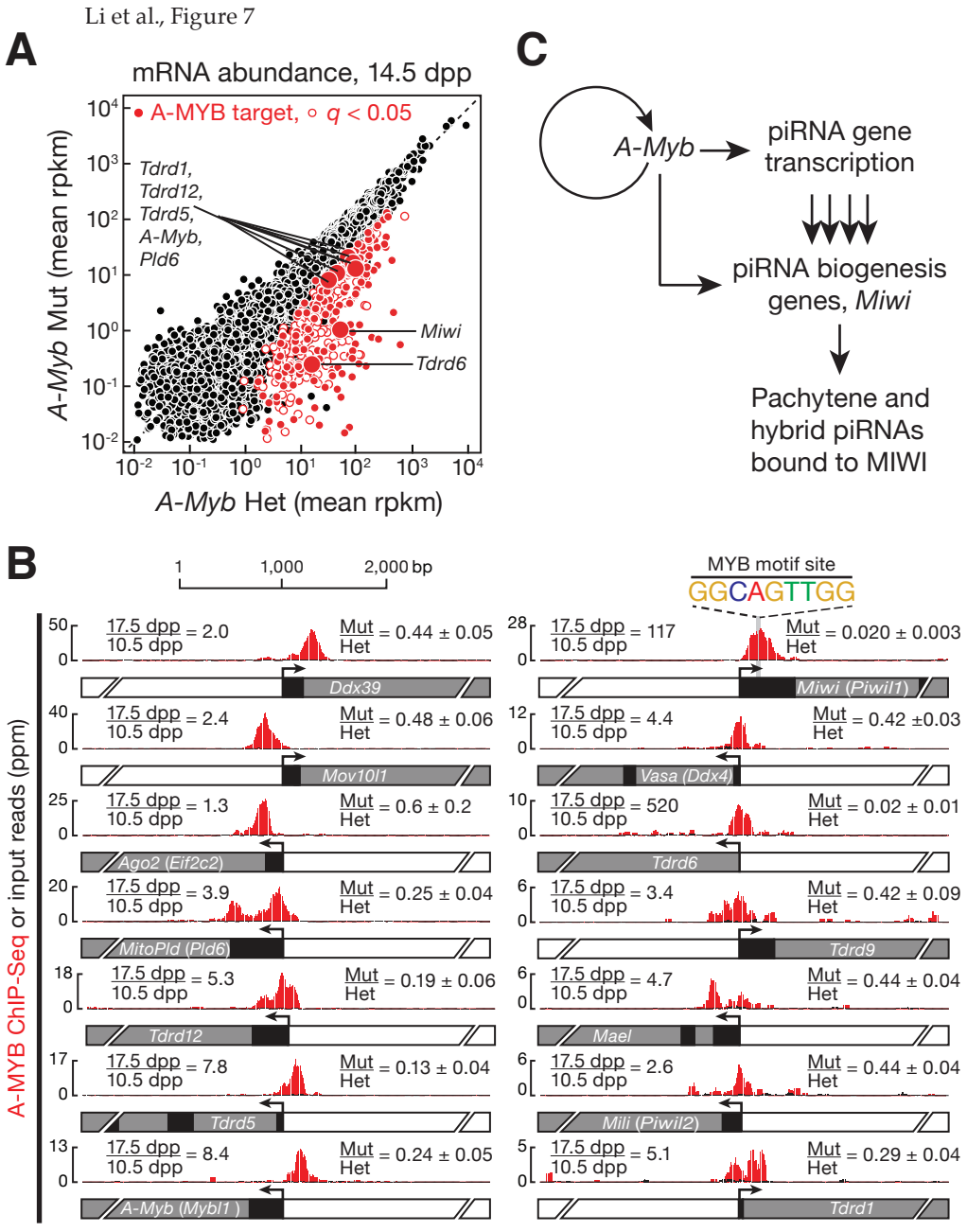
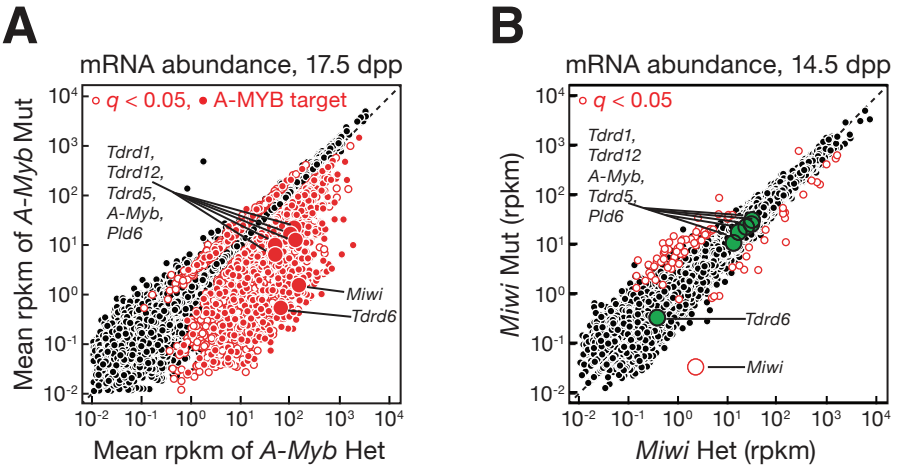


FIGURE 4.13: (A) mRNA abundance in *A-Myb* mutant versus heterozygous testes. The 407 genes with a significant (q < 0.05) change in steady-state mRNA levels are shown as red circles. The 203 with A-MYB peaks within 500 bp of their transcription start site are filled. (B) A-MYB ChIP-seq signal at the 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 4.14 and Table S3.

Li et al., Figure S7, Related to Figure 7



C

	***A-Myb* Het/Mut**		***Miwi* Het/Mut**	
	14.5 dpp	**17.5 dpp**	**14.5 dpp**	**17.5 dpp**
Ddx39	2.3 ± 0.2	2.9 ± 0.1	1.1	0.82
Vasa (Ddx4)	2.4 ± 0.2	3 ± 1	1.4	1.0
Ago2 (EiF2c2)	1.7 ± 0.7	1.0 ± 0.4	1.0	0.59
Mael	2.3 ± 0.2	3.5 ± 0.5	1.2	0.84
Mov10l1	2.1 ± 0.3	1.6 ± 0.2	1.1	0.85
A-Myb (Mybl1)	4.1 ± 0.7	5.0 ± 0.2	1.2	0.99
Miwi (Piwil1)	50 ± 8	100 ± 100	66	41
Mili (Piwil2)	2.3 ± 0.2	2.3 ± 0.6	0.85	0.85
MitoPld (Pld6)	3.9 ± 0.6	8.0 ± 2.1	1.2	1.4
Tdrd1	3.4 ± 0.4	5 ± 1	0.99	0.90
Tdrd12	5 ± 2	6 ± 1	1.0	0.72
Tdrd5	8 ± 1	10 ± 3	0.93	0.84
Tdrd6	64 ± 30	100 ± 100	1.1	0.84
Tdrd9	2.4 ± 0.5	2.8 ± 0.5	1.1	0.79

FIGURE 4.14: 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).

both 23 and 25 nt (Figure 4.15A). 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 4.15B), and we detected a significant Ping-Pong amplification signature (Z score = 31; Figure 4.15C). 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.

Using 24–30 nt piRNAs from oxidized libraries, we identified 327 rooster piRNA clusters (Figure 4.16). 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 x 10-201; Figure 4.15D) similar to that found in the mouse (Figure 4.7A). 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 [Deng et al., 1996, Weston, 1992].

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.

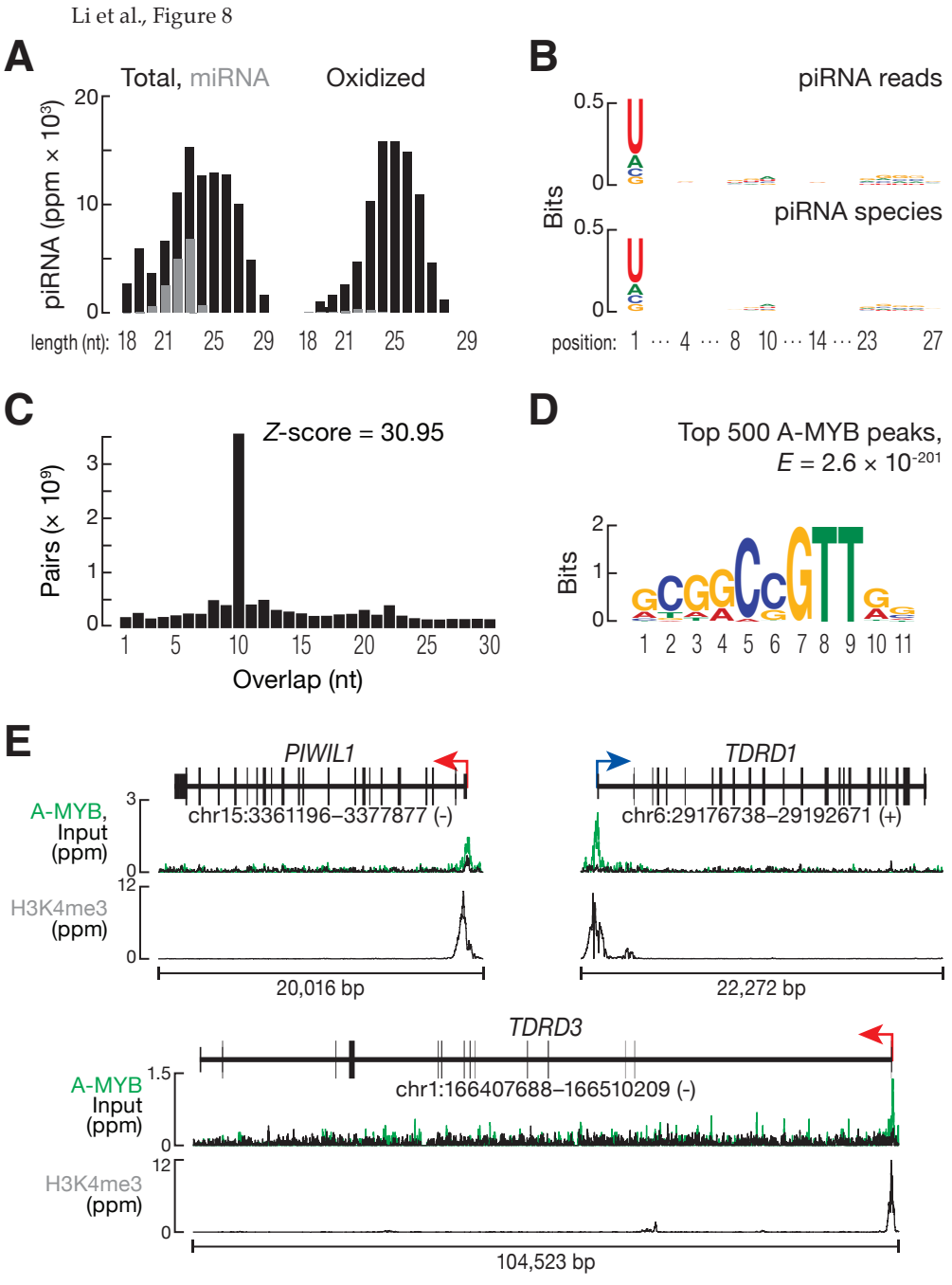


FIGURE 4.15: (A) Length distributions of total rooster testis small RNAs (black) and miRNAs (gray).(B) Sequence logo showing the nucleotide composition of piRNA reads and species.(C) The 5′ -5′ overlap between piRNAs from opposite strands was analyzed to determine if rooster piRNAs display Ping-Pong 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.

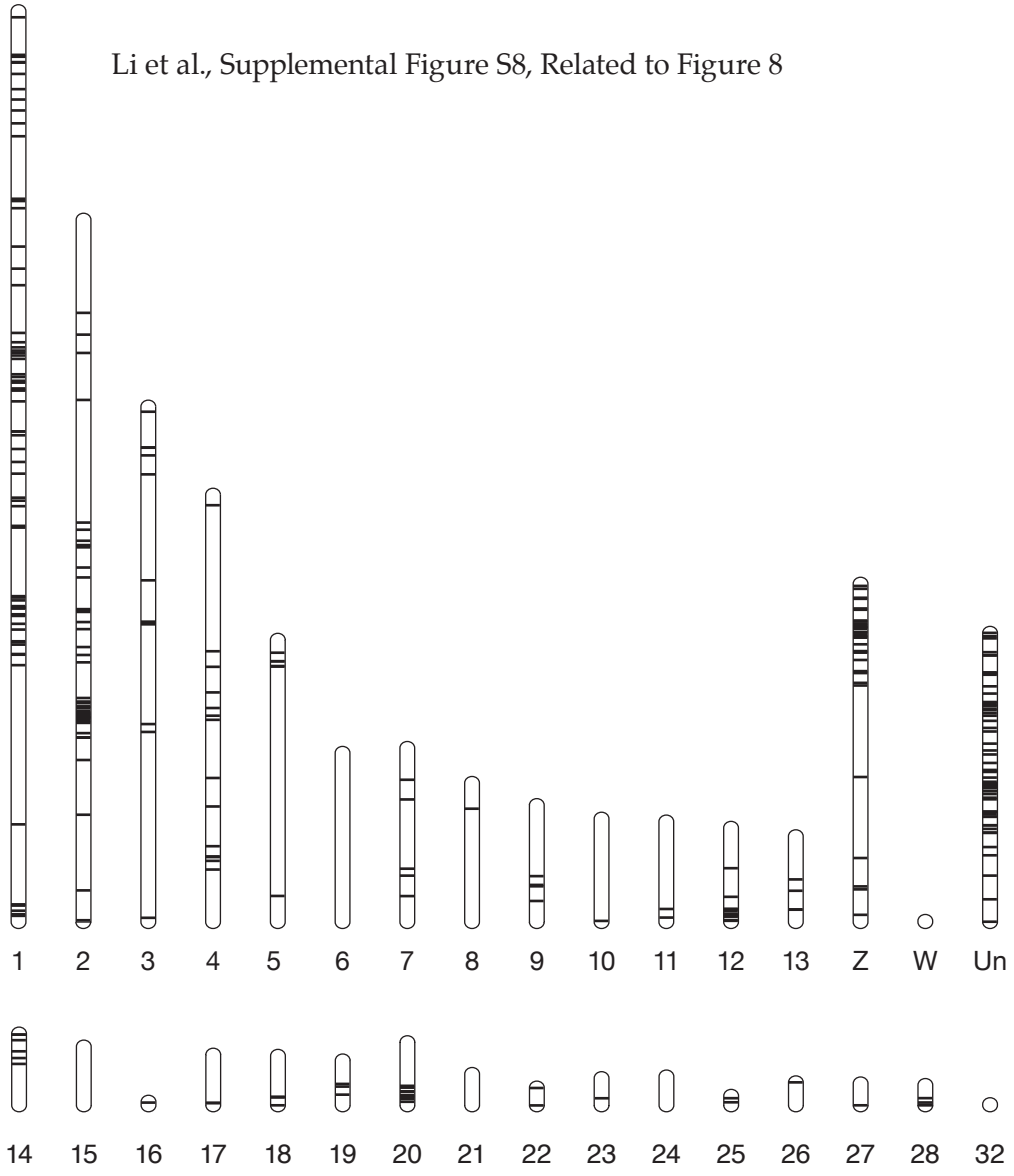


FIGURE 4.16: Black horizontal lines denote the locations on the Gallus gallus (galGal3) chromosomes of the piRNA clusters identified by small RNA sequencing. The figure shows 324 clusters; clusters on E64 (cluster 370) and E22C19W28_E50C23 (clusters 109 and 563) are not shown.

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 4.15E). TDRD1 and TDRD3 also showed A-MYB peaks (Figure 4.15E). 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 4.2A), perhaps because mouse sex chromosomes are silenced during the pachytene stage [Li et al., 2009a]. 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 4.16). Of the 39 Z chromosome clusters, 18 had an A-MYB peak at their promoter.

4.3 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 4.3A). 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, *Mybl1*^{repro9}, 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.

4.4 EXPERIMENTAL PROCEDURES

Mice

Mybl1^{repro9}, *Spo11*^{tm1Sky}, and *Piwil1*^{tm1Hf} 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, Bolcun-Filas et al., 2011, Deng and Lin, 2002].

Sequencing

Small [Ghildiyal et al., 2008, Seitz et al., 2008] and long RNA-seq [Zhang et al., 2012b] and analysis [Li et al., 2009b] 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.

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. C57BL/6J (Jackson Labs, Bar Harbor, ME, USA; stock number 664); *Mybl1*^{repro9} in a mixed 129X1/SvJ x C57BL/6J background; Spo11tm1Sky in a C57BL/6J background; and Piwil^{1tm1Hf} in a mixed 129X1/SvJ x C57BL/6J background ("*Miwi*") mice were genotyped as described [Baudat et al., 2000, Bolcun-Filas et al., 2011, Deng and Lin, 2002]. Rooster testes from White Leghorn of the Cornell Special C strain, about 15 months old, were used for small RNA analysis; and testes from the Brown Leghorn strain, about one year old, were used for ChIP analysis.

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 2S rRNA depletion omitted. Sequencing was performed using either a Genome Analyzer GAII (36 or 76 nt reads) or HiSeq 2000 (50 nt) instrument (Illumina, San Diego, CA, USA). We analyzed published small RNA libraries from purified mouse spermatogonia (SRR069809), spermatocytes (SRR069810, GSE39652), or spermatids (SRR069811; [Gan et al., 2011, Modzelewski et al., 2012]; from *Mili* mutant or heterozygous testes at 10 dpp (SRX003089 and SRX003088; [Aravin and Hannon, 2008]; from Tdrd6 mutant or heterozygous testes at 18 dpp (SRX012165 and SRX012166; [Vagin et al., 2009]; and MILI IP samples from Tdrd9 mutant or heterozygous testes at 14 dpp (SRX015795, SRX015796, SRX015797, and SRX015798; [Shoji et al., 2009].

Strand-specific RNA-seq libraries [Zhang et al., 2012a] using Ribo-Zero Gold (Epicentre Biotechnologies, Madison, WI, USA) were sequenced using the paired-end protocol on a HiSeq 2000.

Small RNA Analysis

Small RNA sequence analysis was as described [Li et al., 2009b] using mouse genome release mm9 and chicken genome release galGal3. Noncoding RNA annotations comprised data from ncRNAscan, the known tRNAs from UCSC, and 18S, 28S and 5.8S rRNAs. miRNA hairpin and mature miRNA annotation was from miRBase Release 19. Mouse and chicken transposons were annotated using Repeat Masker from UCSC. Reads that did not map to the 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. Table S1 reports the statistics for high-throughput sequencing. For oxidized (i.e., piRNA-enriched) samples, uniquely mapping small RNAs >23 nt were mapped to each assembled piRNA precursor transcript and reported as reads per kilobase pair per million reads mapped to the genome (rpkm) using a pseudo count of 0.001.

Small RNA Analysis

RNA-seq reads were aligned to the genome (NCBI 37/mm9) using TopHat 2.0.4 [Trapnell et al., 2009]. Reads were mapped uniquely using the '-g 1' switch. We assembled the mouse testes transcriptome (see below). For genes with multiple isoforms, the transcript with the highest average rpkm value among the three replicates of adult testes was selected for further analysis. Fragments with both reads mapped within a transcript, or to piRNA precursor transcripts, were counted using BEDTools [Quinlan and Hall, 2010]. The sum of the reads aligning to the top quartile of expressed transcripts per library was used to calibrate the samples. The number of reads per transcript was normalized by length, divided by the library-specific calibration factor, and reported as rpkm with a pseudo count of 0.001. Table S1 presents the statistics for the RNA-seq data. Sequences mapping to five genes (Table S1) that overlapped with or were embedded within a piRNA gene were excluded when calculating piRNA precursor transcript abundance.

PAS-seq Library Construction and Analysis

PAS-seq libraries (Table S1) were prepared essentially as described [Shepard et al., 2011] and sequenced using a HiSeq 2000 (100 nt read length). We first removed adaptors and performed quality control using Flexbar 2.2 (http://sourceforge.net/projects/theflexibleadap) with the parameters "-at 3 -ao 10 –min-readlength 30 –max-uncalled 70 –phred-pre-trim 10." For reads beginning with GGG including (NGG, NNG or GNG) and ending with three or more adenosines, we removed the first three nucleotides and mapped the remaining sequence with and without the tailing adenosines to the mouse genome using TopHat 2.0.4. We retained only those reads that could be mapped to the genome without the trailing adenosine residues. Genome-mapping reads containing trailing adenosines were regarded as potentially originating from internal priming and thus discarded. The 3′ end of the mapped, retained read was reported as the site of cleavage and polyadenylation.

CAGE Library Construction and Analysis

CAGE (cap analysis of gene expression; Table S1) was as described

[Yang et al., 2011] and sequenced using a HiSeq 2000 (100 nt reads). After removing adaptor sequences and checking read quality using Flexbar 2.2 with the parameters of "-at 3 -ao 10 –min-readlength 20 –max-uncalled 70 –phred-pre-trim 10", we retained only reads beginning with NG or GG (the last two nucleotides on the 5′ adaptor). We then removed the first two nucleotides and mapped the sequences to the mouse genome using TopHat 2.0.4. All unique 5′ ends of the mapped positions were considered as CAGE-tag starting sites and grouped into tag clusters using a distance-based method in which the maximal distance between two neighboring tags was required to be <20 bp. The peak position of a tag cluster was then reported as the transcription start site.

Transcriptome Assembly and Annotation

De novo transcriptome assembly from three biological replicates of strand-specific RNA-seq data from adult testes was performed using Trinity (r2012-06-08) with default parameters [Grabherr et al., 2011]. The assembled RNA sequences were aligned to the mouse genome (mm9) with BLAT [Kent, 2002], and the alignments with more than 95% of sequence length mapped and fewer than 1% mismatches retained.

We extracted novel junctions from Trinity (i.e., reads with [0-9]+M[0-9]+N[0-9]+M pattern in the CIGAR string of SAM output), and re-mapped all RNA-seq reads to these junctions, rescuing 1,402,444 reads in three replicates. Rescued reads were combined with TopHat alignments (supplied with "–max-multihits 100" to assembly through repetitive regions) and used as input for reference-based assembly.

We used Cufflinks v2.0.2 [Trapnell et al., 2010] with parameters of "-u -j 0.2 –min-frags-per-transfrag 40" to assemble transcripts. To join small transcript fragments caused by insufficient read coverage or embedded repetitive elements, two different gap-joining distance cutoffs were used for the assembly of genes ("–overlap-radius 100") and piRNA loci ("–overlap-radius 250"). We used Cuffcompare v2.0.2 [Trapnell et al., 2010] to annotate the 49,840 Cufflinks-assembled transcripts using parameters optimized for genic conditions ("–overlap-radius 100").

piRNA Precursor Transcript Annotation

We combined transcripts from the two Cufflink assemblies with those from the Trinity assembly, producing 136,069 unique transcripts. Those transcripts with 100 ppm or 100 rpkm unique mapping piRNAs at any time point (10.5, 12.5, 14.5, 17.5, 20.5 dpp and adult oxidized small RNA from testis) were selected for manual annotation.

To refine the termini of the piRNA-producing transcripts, we supplemented the RNA-seq data with high-throughput sequencing of 5′ ends of RNAs bearing (5′)ppp(5′) cap structures (CAGE) and of the 3′ ends of transcripts flanking the poly(A) tail (PAS-seq). To provide independent confirmation of the 5′ ends of each piRNA-producing transcript, we used chromatin immunoprecipitation (ChIP-seq) of RNA polymerase II (pol II) and histone H3 bearing trimethylated lysine-4 (H3K4me3). Refinement of transcriptional starts required both a CAGE and a H3K4me3 peak to support the 5′ end of the transcript. When no H3K4me3 peak corroborated alternative transcription start sites proposed by the CAGE data, the alternative transcripts were merged with the fully substantiated transcript.

piRNA Gene Nomenclature

When piRNA-producing genes overlap an annotated protein coding gene, we refer to them using the name of the overlapping gene preceded by 'pi-'; when they do not, their names refer to their genomic location followed by a number indicating the piRNA abundance in ppm at 6 weeks postpartum. The last digit of a piRNA gene name specifies the rank order of expression among isoforms, determined by the highest abundance of transcripts (rpkm) observed for that gene among the six developmental stages of testis.

Grouping piRNA Precursor Transcripts

For the most abundant transcript in each locus, the abundance (rpkm) of piRNAs at each stage was expressed as a fraction of the maximum abundance reached during the developmental time course. These data were then analyzed by hierarchical clustering according to Euclidean distance and complete linkage using Cluster 3.0. Clustering results were visualized using Java Tree View 1.1.3.

Analysis of Differential Gene Expression

We determined differential gene expression using DESeq R [Anders and Huber, 2010]. For each annotated mRNA, reads from each library were aligned to the most abundant assembled transcript. Transcripts with $q < 0.05$ were considered to be differentially expressed. Table S3 lists the genes that were differentially expressed in *A-Myb* at 14.5 dpp. Three biologically independent replicates were used for *A-Myb*homozygotes and heterozygotes at 14.5 and at 17.5 dpp.

Motif Discovery

For divergently transcribed piRNA gene pairs, the promoter region was defined as the region between the transcription start sites defined by CAGE peaks. Sequence motifs in these putative promoter regions were detected

ab initio using MEME [Bailey et al., 2009, Bailey and Elkan, 1994] in TCM mod (any number of repetitions per sequence) and compared to existing JASPAR and TRANSFAC libraries via TOMTOM [Gupta et al., 2007]. FIMO was used to detect motif sites within the putative promoters (default $p < 10^{-4}$; [Grant et al., 2011].

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 further crushed using 20 strokes with a 'B' pestle in a Dounce homogenizer (Kimble-Chase, Vineland, NJ, USA). Chromatin was sheared by sonication and immunoprecipitated using anti-A-MYB (HPA008791; Sigma, St. Louis, MO, USA) or anti-H3K4me3 (ab8580; Abcam, Cambridge, MA, USA) antibody; immunoglobulin G (IgG; Sigma, item 2729) served as a control. ChIP-quantitative PCR (qPCR) was performed using the CFX96 Real-Time PCR Detection System with SsoFast EvaGreen Supermix (Bio-Rad, Hercules, CA, USA). Data were analyzed using DART-PCR [Peirson, 2003]. Relative ChIP enrichment values were normalized to *MyoD1*, a gene not expressed in testes. Table S1 lists ChIP-qPCR primers. ChIP-seq libraries for anti-A-MYB and control input DNA were prepared following the Illumina ChIP-seq protocol and sequenced on a HiSeq 2000 (50 nt reads).

ChIP-seq Analysis

ChIP-seq reads were aligned to the genome using Bowtie version 0.12.7 [Langmead et al., 2009]. Reads were mapped uniquely using the '-M 1 –best –strata' switches and one mismatch was allowed (-v 1). ChIP peaks were identified using MACS version 1.4.1 [Zhang et al., 2008] using default arguments, input as control, and a cutoff p-value = 10^{-25} was used. BEDTools was used to assign peaks to the nearest 5′ end of genes. Table S1 reports sequencing statistics for ChIP-seq.

RT-PCR

Total RNA was treated with Turbo DNase (Ambion, Austin, TX, USA), and then reverse transcribed using SuperScript III (Invitrogen, Eugene, OR, USA) with random primers (Promega, Madison, WI, USA). The resulting cDNA was analyzed by conventional PCR. Table S1 lists the primers used in Figure 4.12.

Ping-Pong Analysis

Ping-Pong amplification was analyzed by the 5′–5′ overlap between piRNA pairs from opposite genomic strands [Li et al., 2009b]. Overlap scores for each overlapping pair were the product of the number of reads of each

of the piRNAs from opposite strands. The overall score for each overlap extend (1–30) was the sum of all such products for all chromosomes. Heterogeneity at the 3′ ends of small RNAs was neglected. Z-score for 10 bp overlap was calculated using the scores of overlaps from 1–9 and 11–30 as background.

Rooster piRNA Cluster Detection

We developed a dynamic programming algorithm to identify the genomic regions with the highest piRNA density. We used oxidized small RNA reads (>23 nt) to detect clusters. We used the conservative assumption that piRNA clusters compose at most 2% of the chicken genome. We first split the genome into 1 kbp non-overlapping windows and computed piRNA abundance for each window. The mean of the top 2% of windows was used as the penalty score for the dynamic programming algorithm. The algorithm computes the cumulative piRNA abundance score as a function of the window index along each chromosome. The score at a window is the sum of the score in the previous window and the piRNA abundance in the current window, minus the penalty score; if the resulting score was negative it was reset to 0. The maximal score points to the largest piRNA cluster. We extracted the largest piRNA cluster, recomputed the scores at the corresponding windows, and searched for the next cluster. The process continued until the scores for all windows were zero. The boundaries of each cluster were further refined by including those base pairs for which piRNA abundance exceeded the mean piRNA abundance of the top 2% windows. We considered only those clusters with abundance >10 ppm for uniquely mapping piRNAs. In Figure 4.15E, gene models were corrected using data from our unpublished adult rooster testis RNA-seq data.

Chapter 5

Discussion

5.1 The Future of Dynamic long RNAs

Deep sequencing of transcriptomes has revolutionized biology. Previously, transcript discovery was a cumbersome task. Transcript identification and characterization involved significant labor, cost, and materials. In the mid-90's, microarray technology [Schena et al., 1995] gave us a tantalizing glimpse into how genes were expressed, but were limited to probed, and therefore known, sequences. Yet, the green and red landscapes of a microarray analysis hinted at incredible complexity —a complexity that would have to wait for technology to catch up.

Like many transformative technologies, RNA-seq was made possible by incremental improvements to numerous supportive technologies such as: 1) digital optics; 2) microscopy; 3) slide chemistry and on-slide PCR; and 4) nucleic-acid alignment. A HiSeq 2500 relies on all of these technologies (and others) to produce the 100M+ sequences that allow scientists to peer every day into the transcriptional output of a genome.

In the past 5 years, biologists have started to think way beyond mRNAs and small RNAs. The former captured out interest for 30+ years [Furuichi, 1975, Wei et al., 1975], and the later has been on a run-away train since capturing out attention in 1998 [Fire et al., 1998]. HTS has added long RNAs to these classes of gene products. However, many biologically-trained and minded Scientists find themselves overwhelmed by the complete different methods and approaches to tackling the "big data" created by modern genome-wide experiments. Experimental training does not currently provide students with the required skills in statistics, computer programing, and experimental design that are needed to

work with genome-wide data (see section 5.5.1. The richness of this data often leaves many unasked (and unanswered) test-able hypothesis just sitting in public repositories [Plocik and Graveley, 2013].

At this point, it is important to remember that in this document *long RNAs* may also refer to products containing characteristics of traditional mRNAs, that is a 5′ m7G Cap, ligated exons, and a Poly(A) tail. However, many of these long mRNAs are extremely dynamic. So much so that until HTS and RNA-Seq, comprehensive investigation of their complexity was not possible.

5.1.1 Pervasive transcription

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 cells in culture simply perpetuating their existence.

The ENCODE papers from late 2012 suggested that 95% of the genome is functional (REF). Djebali et al. [2012] focused on issues of transcription in these cell lines, as discussed in section 1.3.5. The ENCODE studies were performed an human cancerous cell lines from different sources, and they still saw tremendous transcriptional diversity. Will better HTS tools and resolution, we should fully expect ever more diversity from the transcriptome once we can accurate catalogue and assemble transcripts from tissues, cells, and single single cells *over time*.

Insert ENCODE References and Graur References

5.1.2 Tissue and cell specificity

Your feelings on Specificity of long RNA expression

5.1.3 Function of long RNAs

We *know* that some long RNAs are functional. What are these?

5.1.4 Chromatin regulation

We also know that some long RNAs regulation Chromatin structure. What are your feelings as to the importance of this fact?

5.1.5 PTGS

Long RNAs ability to do post transcriptional gene regulation, including piRNAs, and Xist, etc....

5.1.6 Accurate and complete transcript annotation

A deck of cards has only 52 cards, but can delt into $2,598,960$ different 5 cards hands, as when playing Poker. There are $1,098,240$ different single-pair combinations, with a probability of obtaining one being almost 50%. Compare this to "Royal Flush", for which there are only 4 options, and a probability of $649,739:1 or 1.54 * 10^{-6}$!. It is these numbers that makes possible to play Poker for hours on end. Long ago, biologic evolved to arrange genes into unique and rare combinations, especially in eukaryotic organisms. Indeed the process of splicing is closely correlated with organism complexity. The process of AS is even more closely tied to organism complexity (see figure 1.4).

Accurate determination and assembly of each card (exon) that comprises a hand (transcript) is a major known unknown of research into long RNA. The current state of the field is described in section 1.5.2. This field is in its infancy. Each gain in resolution or sensitivity reveals more complexity. Yet transcript assembly algorithms only provide predictions and probabilities for the existence of real molecules. Until RNA can be directly sequenced, in their entirety, from single cells, researchers will always be making compromises for transcript annotation and quantification. Once technology advances to the point where a transcriptome can be as accurately and quickly determined as a genome, extremely exiting research into the more subtle and nuanced complexity (e.g. what makes one twin molecularly different from another?) of biology can be unlocked.

5.1.7 How are they important?

Conservation of these things is not obvious - if they are not conserved - are they important? Maybe talk about how MALAT1 is highly expressed, but seems to be dispensable.

5.1.8 What regulates their tissue-specific expression?

Do they important some of the special sauce that makes tissues different from one other, more so then the mRNAs changes which can be extreme, but not terribly so....

What determines AS splicing decisions? It is not connectivity in splicing, and it seems to be SR and hnRNP proteins. Splicing is tissue specific. So is it the tissue-specific expression of SR and hnRNP proteins the is the main determinant of AS outcomes? How does chromatin state and organization play into AS decisions?

5.1.9 Technological Improvements

How does one perform deep and broad analysis of mRNAs using second-generation HTS give the tremendous log-range over which they are expressed?

This is the area of knowledge keeps many motivated to perform basic research every day. What secrets does the transcriptome have in store that we haven't even *thought* about? Only through pushing the boundaries of the last two sections can we begin to think beyond the edge of map and formulate testable hypothesis. Here I propose a few outlandish ideas for Unknown Unknowns.

5.2 Future SeqZip development and use

5.2.1 Assay Modifications

5.2.1.1 Rnl2 T29A mutation

One of the most potentially exciting improvements in the SeqZip assay comes from a mutant discussed in Nandakumar et al. [2006]. In this, and past studies, the T39A mutation alleviates structural constrains on having a penultimate $2'$ OH on the $3'$ side of the nick.

Use of T39A mutation to alleviate penultimate $2'$ OH requirement of T4 Rnl2 (See Nandakumar...Lima, Cell 2006)

Make a note into the future directions that you would like to explore LNA's at the 3 extprime OH position of all ligation results, leading to increased ligation efficiency

however both this and the use of penultimate 2 extprime OH (Ribosome) suger in your ligamers would lead to added costs, and the latter maybe better served with a T39A mutation. Giggity

5.2.1.2 Thermostable Ligases

Use of thermostable ligase, allowing for multiple rounds of ligation. Need a good reference, DO NOT USE Ref 27 from Conze et al 2009! Also Elevated ligation temperatures, minimizing blut-ended NTL events

5.2.1.3 Repurposing the SOLiD Platform

5.2.1.4 Quantifying ligation and PCR events

Digital PCR of the PCR products ala [Shiroguchi et al., 2012].

5.2.1.5 Reducing required input RNA and ligamer concentration

SeqZip on single-cell RNA samples.

5.2.2 An ideal SeqZip experiment to query coordinated splicing

If I could go back in time 4 years and still possess the knowledge and abilities that I do now, I would have approached a genome-wide study of coordination in splicing using SeqZip much differently. First, I would have focused on alternative first exon (or promoter, or TSS) and potential coordination with downstream cassette exons. I would have mined newly-generated RNA-Seq data [Pan et al., 2008, Wang et al., 2008] for alternative first exons and cassette exons of sufficient expression. Then, I would have used my automated ligamer design software (see Appendix C), to create a database of the required ligamers. As this would require at least 3 ligamers per event, with very little duplicated use of ligamers, the number of ligamers would make standard synthesis, even using 384-well plates, impossible. Therefore, I would have pursued printing the ligamers on a custom microarray, and cleaved them into solution, similar to products offered by

If I could have done the experiment designed above, I feel the full potential and utility of the SeqZip method could have been realized and generated new and valuable knowledge for the field of gene expression.

5.2.3 SeqZip and single-molecule FISH

The cellular location of precursor piRNA transcript processing is not known. The most accepted hypothesis is that precursor transcripts are processed into mature piRNAs with machinery tethered to chromatoid bodies (REF) or another structure similar to Drosophila Nuage (REF) near the mitochondrial cement (REF). Knowledge of *where* mature piRNAs are generated would provide clues into larger mechanist details of their biogenesis.

Need Chromatoid References

I can think of two broadly different ways in which we could pinpoint the physical location of mature piRNA generation. The first is to chemically label, in some manner, primary piRNA transcripts. The label would need to 1) not interfere with processing and 2) be durable to later methods used to analyze the presence or absence of the modification. A second way to identify the location of mature piRNA processing would be to actually observe, through in vivo FISH, the various products and by-products of biogenesis. In this particular approach, SeqZip maybe useful.

Another idea here would be to use the MCP x MBS mouse as published by the Singer lab [Park et al., 2014]. In this paper they insert multiple MS2 binding loops into the 3 extprime -UTR of the $beta$-actin gene, and cross it with another mouse that has MCP (MS2 Bacteriophage capsid protein) fused to GFP. Using this system they can visualize extitendogenous mRNAs in cultured MEF cells, and brain sections. How could we apply this to imaging of piRNA precursor transcripts? If we inserted an MS2 loop into the tail end of transcripts, what would we see? It is certainly worth it!

5.3 In the haystack: piRNA precursor transcripts surrounded by RNA

5.3.1 In vivo chemical labeling of precursor piRNA transcripts

SeqZip accurately reports on the presence of multiple, distant sequences contained in the same RNA molecule (see Chapters 2 and 3). Yet I was not able to observe ligation products templated off some of the most highly-expressed

precursor transcripts (see section 2.4. We hypothesize that the most likely explanation for being unable to observe ligation products for these transcripts, but successfully observing those for a very long, but lowly expressed gene extitDst, suggests that precursor transcripts are rapidly processed into mature piRNAs. What if this is not the case if we could hybridize ligamers to precursor transcripts in vivo? Put another way extemdash What if the steady-state amount of precursor transcript is very low due to rapid processing, but when visualized in real time, the transcripts are of sufficient abundance for FISH? Ligamers could be engineered to contain different combinations of fluorescent dyes, and the proximity and intensity of the signal could be used to infer precursor transcripts. Importantly, this approach could distinguish long, continuous precursor transcripts from processing intermediates and mature piRNAs.

5.3.2 In vivo imaging of precursor piRNA transcripts

The most important question for mammalian extitpachytene piRNAs is extitWhat are they doing?. We know that they are essential for the health of the species, as discussed in section 4.1, and piRNA-pathway mutants are sterile. What could these small RNAs, with complementarity to nothing but themselves, be doing? Building on the last section discussing chemical modification precursor transcripts, if the modification was again durable enough for downstream analysis, perhaps the modification would remain in the mature piRNAs that are incorporated into mature sperm. These modifications could be tracked as the sperm move through the semineferous tubues and into the $SPERMANATOMY$. One could even track the piRNAs as they fuse with the oocyte to create an embryo. If this modification was labile, piRNA interacting proteins or nucleic acids could be captured or marked as well, providing additional clues as to the function of piRNAs in sperm and early embryogenesis.

5.3.3 What are they doing?

5.3.4 How are they generated?

What determines that a seemingly mRNA-like piRNA precursor is processed into piRNAs, and not translated like crazy by Ribosomes? Everything meets the ribosome (REF), so why is it that precursors are given a different lot in life?

5.3.5 Why should we care?

5.4 Lingering Questions for *Dscam1*

What controls the stochastic and probabilistic splicing of *Dscam1*? Why is it different between hemocytes and neurons? Why would hemocytes need less apparent diversity, given the range of antigen they could potentially encounter?

5.5 Final thoughts

5.5.1 Biologists need Computation Biological Skills

Just 10 years ago, Graduate students and PhDs in the fields of Molecular Biology or Biochemistry need not venture far from data analysis within Excel or perhaps a statistical program with an advanced graphical interface (examples include Prism or Graphpad). Software knowledge that stops at these tools and the rest of the Microsoft Office suite of tools is no longer enough to generate big strides in Biomedical research.

Working with tens of even hundreds of lines of data within a spreadsheet is manageable and computers from 20 years ago had more then enough computing power to process the data. Yet, this type of data is longer then endpoint of most cutting edge projects. Many students and post docs often find that they are unable to analyze the data generated from months or years of tireless bench work. Faced with learning what is effectively a collection of new languages and awash in a sea of acronyms (LINUX, BASH, GNU, PERL, R) they reach out for help from a "Bioinformatics person." Perhaps the relationship and interaction with this personal is productive, leading to a collaboration and exciting new knowledge. Sometimes it isn't, and the bench scientist shifts into one of three modes: 1) Wait; 2) Find another bioinformatic-minded collaborator; or 3) collect more data.

In my experience, the most often chosen mode is "wait." This is also the most damaging, as it delays the progress of one's work, and the advancement of science in general. Personally, I did not want to fall into this mode, and once the multiplex study described in section 2.2 reached a point where I had millions of sequencing reads, but I could not find anyone to help me analyze the data, that I decided to educate myself on the basic principles of Linux, the command line, and analysis of HTS data.

A Biologically-train individual who posses the knowledge of analysis of HTS datasets is an extremely powerful and empowering situation. This was recently commucated in Plocik and Graveley [2013]:

> *Such exercises will empower students to explore and assess the quantitative data published in the manuscripts that they read, which can no longer be assessed at a glance like the qualitative gel-based results on which molecular biology was founded. Ultimately, it will be equally important to know how to write code as it is to pipette. - [Plocik and Graveley, 2013]*

The fact is that no one will care about a project as much as the Graduate student or PostDoc who is the main project driver. Learning and training of computational skills bent on analyzing large datasets should be central to the education in Biomedical sciences in the future.

5.5.2 Science versus Engineering: Two thoughts in one school

> *"There is a general attitude among the scientific community that science is superior to engineering." - [Macilwain, 2010]*
>
> *"Science is about what is; engineering is about what can be. Engineers are dedicated to solving problems and creating new, useful, and efficient things." - Niel Armstrong*

A common schism between technically-oriented individuals is whether or not they identify themselves as an engineer or a scientist. The first quote, from an article published in Nature, communicates a clear bias in academic circles of the importance of the extitwhy over the extithow. In essence, how one priorities these questions may categories individuals as a scientists (why is important) or an engineer (how is more important). The second quote, from the first man to walk on the Moon, Neil Armstrong, highlights what motivates a self described "engineer" and "geek." How does a single graduate system, training PhDs for careers in life science, educate individuals who fall into these two fundamentally different belief systems?

In short extemdash not well. When searching for a lab to call home, I told professors that I wanted to work on a technology development project. A typical response was, "That's not what we do here." As someone who is first interested in the "how" over the "why," this began a brief period when I thought I had made the wrong choice in leaving industry to go back to graduate school. What was

the basis for this aversion to technology development? The same article in Nature states that this feeling toward engineering may be attributed:

> *...partly to a "linear" model of innovation, which holds that scientific discovery leads to technology, which in turn leads to human betterment. This model is as firmly entrenched in policy-makers' minds as it is intellectually discredited. As any engineer will tell you, innovations, such as aviation and the steam engine, commonly precede scientific understanding of how things work.*

If policy-makers value basic discovery over technological application, perhaps this explains why many of my professors tried to steer me away from a technology development project.

In spite of policy-makers and my professors holding the viewpoint that discovery precedes technology, some of the most notable breakthrough scientific discoveries, including many made by Nobel Laureates, demonstrate a clear integration of both the scientific method and technological application. For example, the 2007 award in Physiology and Medicine was given for "discoveries of principles for introducing specific gene modifications in mice by the use of embryonic stem cells." By combining these principle discoveries, an indispensable technique in modern genetics was created – gene targeting. The feeling of which is more important, the principle discoveries or the application thereof, is likely what separates a scientist from an engineer.

The importance of technology to the advancement of science in general is not limited to anecdotes resulting in a Nobel prize. A quick scan of the most highly-cited papers in the journal PNAS reveals that the top 13, indeed extitall 13, are about a novel methodology or technique. Sequencing of DNA, microarray analysis, tetracycline-ineducable promoters, recombinant adenovirus, and site-specific mutagenesis are just a handful of the tools on this list. This effect can be seen in computation biology, with transformative techniques, such as BLAT [Altschul et al., 1990] and Bowtie [Langmead et al., 2009] attaining citations well beyond a typical paper in their journal of publication and far more than most primary research-centered articles.

How does someone who is motivated by the engineering of science best contribute in an academic setting? Luckily, I did not have to question my decision to return to school for long. I found a pair of labs where I could learn how to practice traditional hypothesis-driven research while also developing new tools necessary to do so. My project is a perfect fit for me and has been terrific fun to work on. In the past five years, the technique that I developed, SeqZip, allows for the more efficient study of mRNA isoforms produced from alternative splicing

of pre-mRNA. This is a very engineering-type accomplishment. However, the technique uses short DNA oligonucleotides and a novel activity (that I discovered) of an known RNA ligase to shorten and simplify isoform sequence information. Use of Rnl2 to perform RNA-templated DNA:DNA ligation is completely new knowledge, and falls squarely within a scientific purview. The technique simplifies many of the experimental issues that researchers struggle with when studying the often complex products of alternative splicing.

5.5.3 Dealing with the data deluge

How do we work with all this data? Mention LabKey, GenomeBridge, Define the scope of the problem, etc...

I would love to have a database of complete transcripts of a cell - think about how you were panning the data for the MolCel paper, and you could see ChIP marks, Pol II occupancy, RNA-Seq, and piRNA data. It would be great to be able to do that more often, and we greater precision

Appendix A

Appendix - Misc Information

A.1 Buffers

TABLE A.1: SeqZip Hybridization and Ligation Buffer

Component	Concentration
Tris-HCl	50 mM
MgCl2	2 mM
DTT	1 mM
ATP	400 μM
pH	7.5 @ 25° C

A.2 Equations

A.2.1 Determining [RNA] from ^{32}P-α-UTP used during vitro transcription

$$\mu\text{M} = \left(\frac{\text{pmol}}{\mu\text{L}}\right) = \left(\frac{\text{cpm after purification} \times \text{dilution factor}}{\text{cpm before purification} \times \text{dilution factor}}\right) \times \left(\frac{\text{mol UTP in original reaction}}{\text{Reaction Volume}}\right) \times \left(\frac{1}{\text{Number UTPs in transcript}}\right) \times 10^{-12}$$

A.2.2 Determining [RNA] based on A_{260}

$$\text{[RNA in M]} = \left(\frac{\text{A}_{260} \times \text{Dilution Factor}}{10,313 < \text{note 1} > \times \text{nucleotides in message}} \right)$$

note 1: This value represents an average RNA extinction (ϵ) coefficient value

A.2.3 Normalize oxidized small RNA libraries size to time-matched unoxidized library

NB: this equation assumes calibration against a specific time-point , in this case data obtained from 6 week-old testes.

$$\text{unox } \tau \text{ norm}_1 = \left(\frac{\left(\frac{\sum \text{miRNA reads } \tau}{\sum \text{miRNA reads 6wk}} \right) \times \text{ depth 6wk}}{1,000,000} \right)$$

$$\text{ox } \tau \text{ norm}_1 = \text{unox } \tau \text{ norm}_1 \times \left(\frac{\sum \text{oxidized shared } \geq 23 \text{ nt reads}}{\sum \text{unoxidized shared } \geq 23 \text{ nt reads}} \right)$$

A.3 PCR Programs

Ligamer Hybridization ROY-H2 | Ligamer Hybridization
Steps 1–9 are 10 minute incubations at the following temperatures:
69;66;63;58;54;52;50;48;46° C
Step 10 is a 45° C incubation for 1 hour
Steps 11–14 are 10 minute incubations at the following templates:
43;41;39;37° C
Final incubation is at 37° C for ∞

SeqZip ligation program ROY-37-4 | T4 Rnl2 RNA-template DNA:DNA ligation
1. 37° C for 18 hours
2. 10° C for ∞

Appendix B

Appendix B: piRNA precurors are spliced

B.1 Evidence for spliced piRNA precusor transcripts

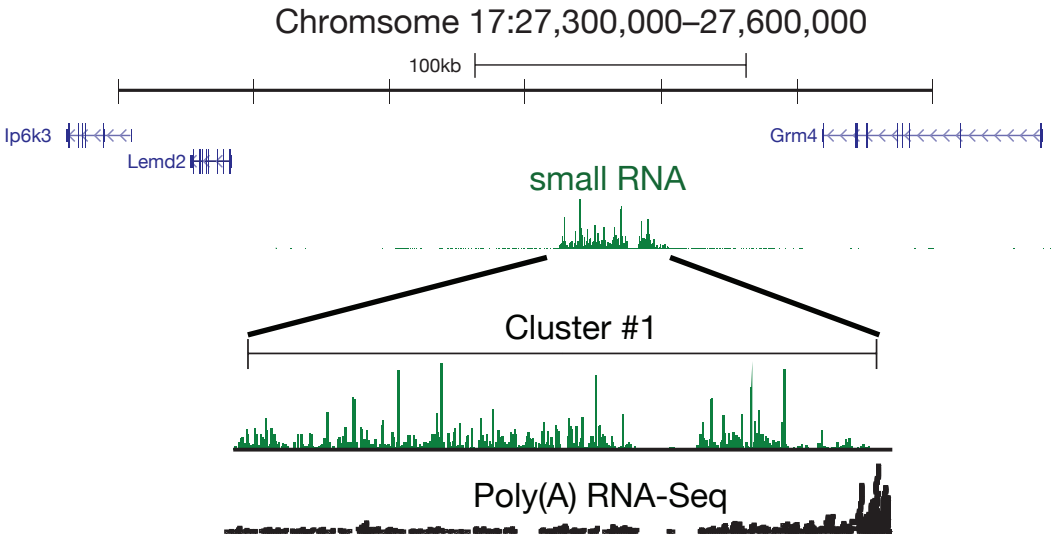


FIGURE B.1: RNA-Seq evidence for piRNA precursor splicing]
Insert Figure Text

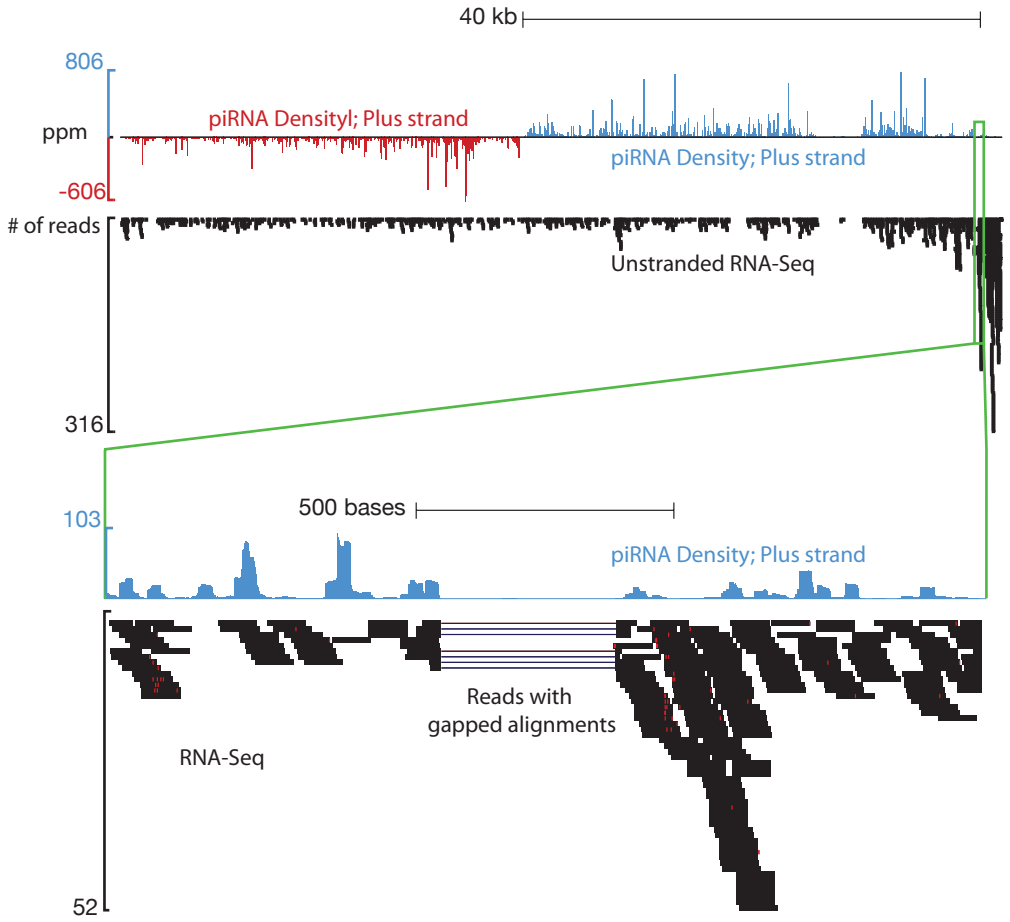


FIGURE B.2: Lack of piRNAs within precusor introns
Insert Figure Text

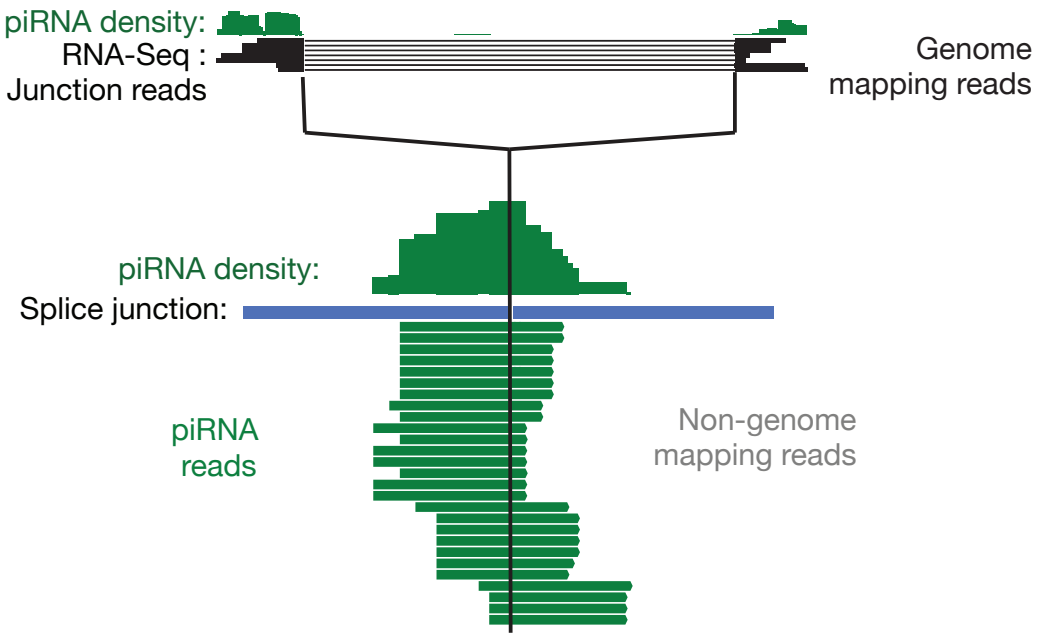


FIGURE B.3: piRNAs map to Splice Junctions of precursor transcripts
Insert Figure Text

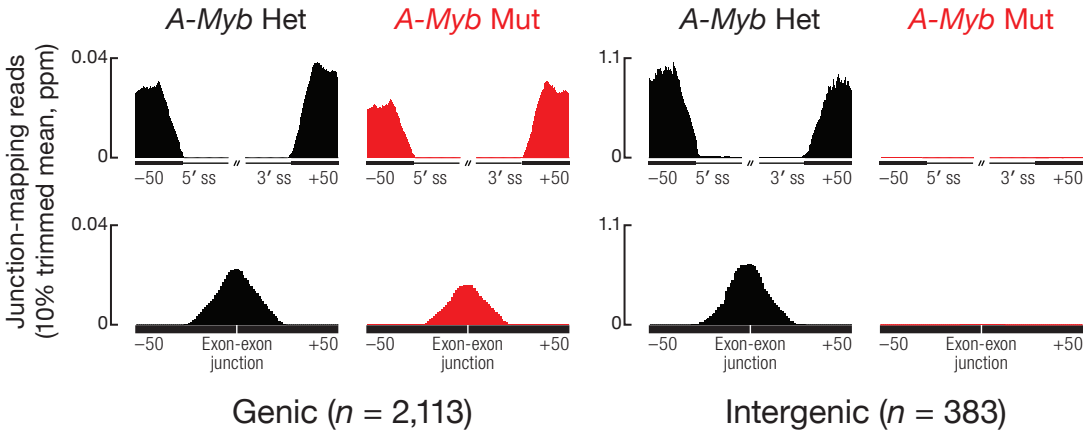


FIGURE B.4: *A-Myb* Mutants produce no splice-junction mapping piRNAs
Insert Figure Text

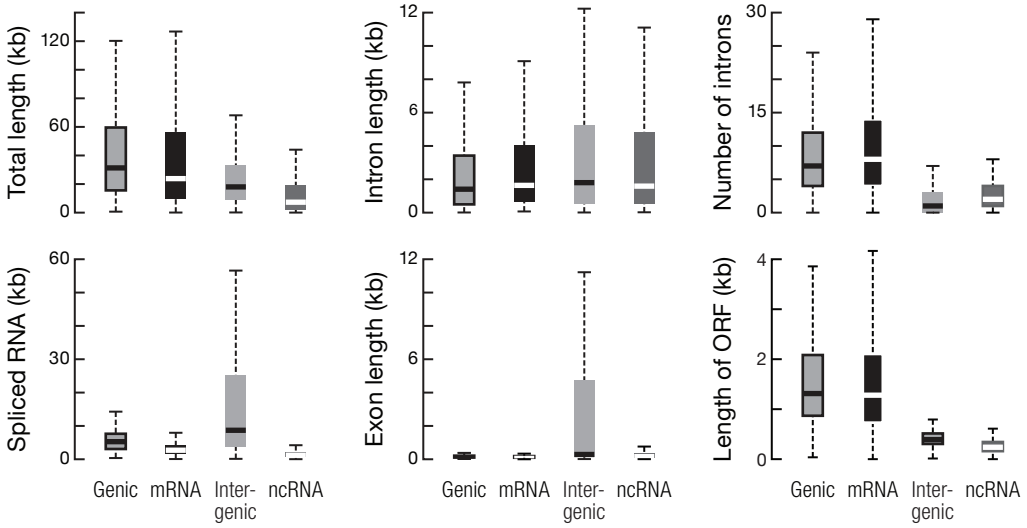


FIGURE B.5: General features of piRNA transcripts
Insert Figure Text

Appendix C

Automated Ligamer Assembly

C.1 Installation

Major Steps:

- Create an input csv file with required information
- Run this information sequentially through the scripts
- Use the results to order oligos from IDT

Required Tools:

- Perl
- BioPerl
- Ensembl Perl APIs
- String::Random Perl Package

Items to future improvements

- Use Ensembl Database to initilize queries
- Make the use of BioPerl more flexible
- Make more web-friedly

Helpful hints on installing BioPerl and Emsembl Perl APIs:

```
## Install BioPerl, use git
    cpan App::cpanminus # First prep cpan
    cpanm DBI ## Install necessary DBI perl module
```

```
        mkdir ~/src; cd ~/src
        git clone git://github.com/bioperl/bioperl-live.git
        cp ~/.bash_profile ~/.bash_profile.bak
        echo -e 'PERL5LIB=$HOME/src/bioperl-live:$PERL5LIB' >>
~/.bash_profile
        source ~/.bash_profile
    # Install ensembl perl apis
        mkdir ~/src; cd ~/src
        wget ftp://ftp.ensembl.org/pub/ensembl-api.tar.gz
        tar xvfz ensembl-api.tar.gz
    # Add locations to perlfile libs to $PATH
        echo -e '
        PERL5LIB=${PERL5LIB}:${HOME}/src/ensembl/modules
        PERL5LIB=${PERL5LIB}:${HOME}/src/ensembl-compara/modules
        PERL5LIB=${PERL5LIB}:${HOME}/src/ensembl-variation/modules

PERL5LIB=${PERL5LIB}:${HOME}/src/ensembl-functgenomics/modules
        export PERL5LIB' >> ~/.bash_profile
```

C.2 Example Input Format

Here is an example input file to create ligamers investigating the *Gria3* gene in Rats:

```
# Comment lines are ignored
#Gene name
 GRIA3
# PCR primers used - Solexa PE adaptor sequences
# Five prime
PCR-Primer-5'-ATCTGAGCGGGCTGGCAAGGC
#Three Prime
PCR-Primer-3'-GCCTCCCTCGCGCCATCAGA
```

ExonId	LigID	Name	Strand	Code	TargetPrime	bedLoc	SetID	ConstID
<Gria3_201/202-Shared-I14	10	rn4	minus	TC	5	X:127903250-127903350	NANNNNNN	201_2_intron
<Gria33_201/202-Shared-I14	9	rn4	minus	T	3	X:127903210-127903249	NNNNNNNN	201_2_intron
<Gria33_201/202-E15	8	rn4	minus	TC	5	X:127914822-127915069	NNNNNNN	201,202
<Gria33_202-I14:15	7	rn4	minus	I	N	X:127912345-127914821	TACACAT	202
<Gria33_202-E14	6	rn4	minus	I	N	X:127912230-127912344	ACCCCAG	202
<Gria33_201-I14:15	5	rn4	minus	I	N	X:127897499-127914821	CGCGCAC	201
<Gria33_201-E14	4	rn4	minus	I	N	X:127897384-127897498	GTCTCAA	201
<Gria33_202-I13:14	3	rn4	minus	I	N	X:127896828-127912229	ACCGATT	202
<Gria33_201-I13:14	2	rn4	minus	I	N	X:127896828-127897383	CGCTATG	201
<Gria33_201/202-E13	1	rn4	minus	T	3	X:127896580-127896827	NNNNNNN	201,202

C.3 Ligamer Assmebly Source Code

```
#! /usr/bin/perl

#Pre requisites
# These are working on 02/19/13
use lib "/home/royc/perl5/lib/perl5/"; # BioPerl location
#use lib "/home/royc/lib/ensembl.perl.zpi/ensembl/modules"; #ensembl
    packages

=head1 Ligamer Assembler

        This script will automatically create ligamers.

=head2 Contact information

        Script made by Christian Roy, Umass Medical School
        christian.roy@umassmed.edu

=cut

use strict; # To help wtih variable control
use warnings; # To help me catch mistakes

use Bio::EnsEMBL::Registry; # To load remote EnsEMBL Registry
use Bio::EnsEMBL::Slice; # To retreave sequences from EnsEMBL registry
use Bio::DB::Fasta; # BioPerl tool to retreave sequnce from local FastA
    file
use Bio::SeqFeature::Primer; # BioPerl Tool for Tm normalization
use Cwd; # To retreave current working directory information

my $dir = getcwd; # Assign current working directory to scalar
my $timestamp = localtime(); # Grab the time at script start

## Variables
my      (
        $file_input, # Name of specified input file
        $output_file, # Name of file to print results too
        $species, # The species to grab from Ensembl
        $strand, # The strand to grab for ligamer sequences
        $working_sequence, # The slice sequence variable
        $line_counter, #Keep track of stepping through input file
        @arguments, # Keep track of input arguments
```

```
        $fa_reference, # Fill if using a local FASTA Reference file
        $chr, # Obvious
        $coordinates, # Interim variable for splitting UCSC
        $start, # obvious
        $end, # Obvious
        $gene, # target gene name
        $lig_location, # Ligamer prime variable
        $target_prime, # Broad variable to define ligamer type - see man
        $UCSCcoordinates, # Obvious
        $pcrsequence, # fill with appropriate PCR sequence for terminal
    oligos
        $barcode, # Fill will barcode for sequence between regions of
    comp.
        $note_line, # Fill with notes for a ligamer query
        $three_prime_PCR_sequence, # Fill with three prime PCR sequence
        $five_prime_PCR_sequence, # Fill with five prime PCR sequence
        $lig_joiner_code, # Internal varialbe for assembling ligamers
    see man
        $set, # Move set assembly information input to output file
        );

#Variables with Defaults
my $verbose=0; # Verbose loading of ensembl databases
my $db_version=62; # Default database version for ensembl database
    loading
my $temp="58"; # Defalt temp for Tm normalization
my $salt="0.05"; # Default salt concetration for Tm calculation in M
my $lig_conc="0.00000025"; # Defeult ligamer conc for Tm calc in M
my $man_print=0; # for printing manual information
my $help_print=0; # For printing help informatio to HTML file
my $ligamer_name=0; # Internal variable for sequental numbering of
    ligamers
my $remote=0; # set to 1 for ensembl database loading
my $control_length=20; # Default length for control variables in nt
my $plname=$0; # assign $plname scalar to script name (for help printing)

#Print Usage information if nothing is entered at commandline
if (@ARGV==0) {system "pod2text $0 | less"; die}

=head2 Usage


        -hp = Print HTML POD data for scriptname
        -mp = Print and view Manual POD data for scriptname
```

```
        -i [File] = File Input
        -o [File] = File output
        -v [#] = Verbose for Ensembl loading
        -d [#] = data_base version for Ensembl loading
        -t [#] = Temp in degrees celcius
        -salt [#] = Salt concentration for Tm in mM
        -lig_conc [#] = Ligamer concentration for Tm in nM
        -c [#} = Minimum length for Control ligamers (default=20)

=cut
## Finish message if run with no arguments

#Parse the command line
while(@ARGV>0)
{
  @arguments = @ARGV;  #Store the command line for printing later

  my $next_arg=shift(@ARGV);

  if ($next_arg eq "-hp") { # Do you want to print HTML POD Data?
    $help_print=1;
    }
  if ($next_arg eq "-mp") { # Do you want to print a manual?
    $man_print=1;
    }
  if ($next_arg eq "-i") { # What is the name of the input file?
    $file_input = shift @ARGV;
   }
  if ($next_arg eq "-f") { #n Name of the fasta file your sequences are
    in?
    $fa_reference = shift @ARGV ;
    }
  if ($next_arg eq "-r") { # Do you want to fetch sequences from ensembl?
    $remote = 1
    }
  if ($next_arg eq "-o") { # Name of output file
    $output_file = shift(@ARGV);
    }
  if ($next_arg eq "-v") { # Do you want to see the ensembl load data?
    $verbose = shift @ARGV;
    }
  if ($next_arg eq "-d") { # What version of ensembl do you want to use?#
    $db_version = shift @ARGV;
    }
```

```
  if ($next_arg eq "-t") { # What temperature in degrees C do you want
    to norm ?
    $temp = shift @ARGV;
    }
  if ($next_arg eq "-salt") { # Salt concentration for Tm calculations?
    $salt = shift @ARGV ;
    $salt = $salt / 1000; # from micro Molar to Molar
    }
  if ($next_arg eq "-lig_conc") { # Concentration for Tm calculations?
    $lig_conc=shift@ARGV;
    $lig_conc = $lig_conc / 1000000000; # nM to M
    }
  if ($next_arg eq "-c") { # What length would you like (min) for
    control ligs?
    $control_length = shift @ARGV ;
    $control_length = $control_length-1
    }
} ## Finish Parsing the command line

###################### POD HELP SUBROUTINE
    CALLS#################################
my $scriptname=$0;
podhelp( $scriptname, $help_print, $man_print, $dir);
###################### POD HTML Subroutine
    CALLS#################################

#################### open the ensembl
    registry#################################
my $db;
if ($fa_reference) {
  $db = Bio::DB::Fasta->new($fa_reference);
  }

if ($remote==1) {
  $db = ensembl_database($verbose, $db_version)
  }
##############################################################################
#open the output file
open (OUT, '>'.$output_file) || die "The output file could not be
    created.\n";
## Print the headers
print OUT
# Start with general assembler information
">Source Program\t",$dir,$0,"\n".
```

```
">Date Run \t$timestamp \n".
">Arguments entered \t", "@arguments"," \n".
">Input filename \t$file_input\n".
">Output filename \t$output_file\n".
">Control Seq Length \t$control_length plus 1\n".
">Normalization temperature \t$temp\n".
##### now all on 1 line print the ligmaer-specific information
">Gene\t". #1
"Ligamer_Number\t". #2
"Species\t". #3
"Strand\t". #4
"Ligamer Joiner Code\t". #5
"Target Prime\t". #6
"UCSC coordinates\t". #7
"PCR Used\t". #8
"Barcode Used\t". #9
"Total Query span\t". #10
"Five Prime Sequence\t". #11
"5 Prime Length\t". #12
"Five Prime Tm\t".
"3 Prime Sequence\t".
"3 Prime Length\t".
"3 Prime Tm\t".
#"Ligamer Identifier\t".
"Ligamer Sequence\t".
"Ligamer Length\t".
"Notes\t".
"Set\t".
"\n";
################################################################################
## open the input file
open (INPUT, $file_input) || die "The file $file_input couldn't be
    opened.\n";
################################################################################

##################read and analyze each line of the input file
    ##################
while (my $line=<INPUT>) { ## starting brace to read through csv

  if ($line=~/^#/){next} #skips comments
  if ($line=~/^>/){print OUT $line; next} #skips and trans.these lines
  if ($line=~/^~/){$line=~s/~//;chomp $line; $gene=$line;} # find gene
    identifier
  $gene=~s/[\s]+//g;
```

```
if ($line=~/^\@/){chomp $line;$note_line=$line;next} #store notes

chomp $line;

if ($line=~/^PCR-Primer-5'-/g) { #Find the 5 adaptor
  $five_prime_PCR_sequence=$line;
  $five_prime_PCR_sequence=~s/PCR-Primer-5'-//;
  $five_prime_PCR_sequence=~s/[\s]+//g;
  print OUT ">5_pcr\t".$five_prime_PCR_sequence."\n";
  next
  }

if ($line=~/^PCR-Primer-3'-/) {## find the 3 adaptor
  $three_prime_PCR_sequence=$line;
  $three_prime_PCR_sequence=~s/PCR-Primer-3'-//;
  $three_prime_PCR_sequence=~s/[\s]+//g;
  print OUT ">3_pcr\t".$three_prime_PCR_sequence."\n";
  next
  }

if ($line=~/^</) {#ligamer query lines start with a '<'
  unless ($note_line) {$note_line=" ";}

  my $lig_joiner_code;
  my $slice_sequence;

  $line_counter++;

   (
   $gene,
   $species,
   $strand,
   $lig_location,
   $target_prime,
   $UCSCcoordinates,
   $barcode,
   $set
   ) = parse_the_line($line);

  $lig_location = uc $lig_location;

  # Parse the ligamer query line
  ($chr, $start, $end)= parse_coordinates($UCSCcoordinates);
```

```
    my $target_seq_length = ($end-$start);

############ Get genomic slice from ensembl registry ###############
    if ($remote==1) {
      $slice_sequence =
        get_genomic_sequence
        (
        $chr,
        $start,
        $end,
        $species,
        $db,
        )
      }
#########################################################################

############ Get the genomic slice from local Fasta #################
    if ($fa_reference) {
      #$chr="chr".$chr;
      my $obj = $db -> get_Seq_by_id($chr);
      $slice_sequence = $obj -> subseq ($start => $end);
      }
#########################################################################

#### get the correct orientation
    my ($working_sequence) =
      revcom_slice_based_on_strand
        (
         $strand,
         $slice_sequence
         );

  # Get the T5 end
    my ($T5_seq, $T5_tm, $T5_seq_length)=
      obtain_T5_tm_sequence
          (
          $working_sequence,
          $temp,
          $lig_location,
          $control_length,
          $salt,$lig_conc
          );
  # Get the T3 end
```

```
  my ($T3_seq, $T3_tm, $T3_seq_length) =
    obtain_T3_tm_sequence
        (
        $working_sequence,
        $temp,
        $lig_location,
        $control_length,
        $salt,
        $lig_conc
        );

#start to build your working HASH
  my %common =
    (
    working_sequence          => $working_sequence,
    temp                      => $temp,
    UCSCcoordinates           => $UCSCcoordinates,
    UCSC_chr                  => $chr,
    UCSC_start                => $start,
    UCSC_end                  => $end,
    gene                      => $gene,
    ligamer_name              => $ligamer_name,
    species                   => $species,
    strand                    => $strand,
    target_prime              => $target_prime,
    five_prime_PCR_sequence   => $five_prime_PCR_sequence,
    three_prime_PCR_sequence  => $three_prime_PCR_sequence,
    barcode                   => $barcode,
    target_seq_length         => $target_seq_length,
    seed                      => $control_length,
    T3_seq                    => $T3_seq,
    T3_tm                     => $T3_tm,
    T3_seq_length             => $T3_seq_length,
    T5_seq                    => $T5_seq,
    T5_tm                     => $T5_tm,
    T5_seq_length             => $T5_seq_length,
    notes                     => $note_line,
    set                       =>$set,
    );

  if ($lig_location eq "T" && $target_prime eq "5") { #Terminal 5
 targeted
    #Advance the ligamer number
    $ligamer_name++;
```

```
    $common{ligamer_name} = $ligamer_name;
    # Add to the hash table
    $lig_joiner_code = "T-5";
    $common {lig_joiner_code} = $lig_joiner_code;
    my %lig_results = Terminal_5(%common);
    my %final = ligamer_piece_joiner(%lig_results);
    my %bed_output = %final;
    output (%final);
    };

  if ($lig_location eq "TC" &&  $target_prime eq "5") {# Grab the
 internal
    $ligamer_name++;
    $common{ligamer_name} = $ligamer_name;
    $lig_joiner_code = "T-C-5-I";
    $common {lig_joiner_code} = $lig_joiner_code;
    my %lig_results = Terminal_5(%common);

    my %final_internal =
       ligamer_piece_joiner
           (%lig_results);

    my $working_sequence = $lig_results{working_sequence};
    my $T5_seq_length = $lig_results{T5_seq_length};

    # Now grab the sequence inside of the control
    $working_sequence = $common{working_sequence};
    my $T5_ctrl_length = $common{T5_seq_length};
    $common{T5_ctrl_length} = $T5_ctrl_length;
    $working_sequence = substr ($working_sequence,$T5_ctrl_length);
    $lig_location = "IC";
    ($T5_seq, $T5_tm, $T5_seq_length) =
    obtain_T5_tm_sequence
        (
        $working_sequence,
        $temp,
        $lig_location,
        $salt,
        $lig_conc
        );

    $common{working_sequence} = $working_sequence;
    $common{T5_seq} = $T5_seq;
    $common{T5_tm} = $T5_tm;
```

```
  $common{T5_seq_length} = $T5_seq_length;
  $ligamer_name++;
  $common{ligamer_name}= $ligamer_name;
  $lig_joiner_code="T-C-5-T";
  $common {lig_joiner_code}= $lig_joiner_code;
  %lig_results = Terminal_5 (%common);
  my %final = ligamer_piece_joiner(%lig_results);
  output (%final_internal);
  output (%final);
  }

if ($lig_location eq "T" && $target_prime eq "3") {
  $ligamer_name++;
  $common{ligamer_name} = $ligamer_name;
  $lig_joiner_code = "T-3";
  $common {lig_joiner_code} = $lig_joiner_code;
  my %lig_results = Terminal_3 (%common);
  my %final = ligamer_piece_joiner (%lig_results);
  my %bed_output = %final;
  output (%final);
  };

if ($lig_location eq "TC" &&  $target_prime eq "3") {
  #Grab the control
  $ligamer_name++;
  $common{ligamer_name} = $ligamer_name;
  $lig_joiner_code = "T-C-3-I";
  $common {lig_joiner_code} = $lig_joiner_code;
  my %lig_results = Terminal_3              (%common);
  my %final_internal = ligamer_piece_joiner (%lig_results);
  # Grab the sequence internal of the control
  $working_sequence = $common{working_sequence};
  my $T3_ctrl_length = $common{T3_seq_length};
  $common{T3_ctrl_length} = $T3_ctrl_length;
  $T3_seq_length = $common{T3_seq_length};
  $working_sequence = substr ($working_sequence,0, $T3_ctrl_length);
  $lig_location = "IC";
  ($T3_seq, $T3_tm, $T3_seq_length) =
  obtain_T3_tm_sequence
    (
    $working_sequence,
    $temp,
    $lig_location
     );
```

```
  $common{working_sequence} = $working_sequence;
  $common{T3_seq} = $T3_seq;
  $common{T3_tm} = $T3_tm;
  $common{T3_seq_length} = $T3_seq_length;
  $common{bed_start} = $start;
  $common{bed_end} = $end;
  $ligamer_name++;
  $common{ligamer_name} = $ligamer_name;
  $lig_joiner_code = "T-C-3-T";
  $common {lig_joiner_code} = $lig_joiner_code;
  %lig_results = Terminal_3 (%common);
  my %final = ligamer_piece_joiner (%lig_results);
  output (%final);
  output (%final_internal);
  };

if ($lig_location eq "I" && $target_seq_length>60) {
  $ligamer_name++;
  $common{ligamer_name} = $ligamer_name;

    if ($lig_location eq "I" && $target_prime eq "C") {
    $lig_joiner_code = "I-L-C";
    $common {lig_joiner_code} = $lig_joiner_code;
    my %lig_results = (%common);
    my %final = ligamer_piece_joiner(%lig_results);
    $final{pcrsequence} = "";
    my %bed_output = %final;
    output (%final);
    }

    if ($lig_location eq "I" && $target_prime eq "N") {
    $lig_joiner_code = "I-L";
    $common {lig_joiner_code} = $lig_joiner_code;
    my %lig_results = (%common);
    my %final = ligamer_piece_joiner(%lig_results);
    $final{pcrsequence} = "";
    my %bed_output = %final;
    output (%final);
    #my %bed_final = prep_bed (%bed_output);
    }
  }

if ($lig_location eq "I" && $target_seq_length<=60) {
```

```
        $ligamer_name++;
        $common{ligamer_name} = $ligamer_name;
        $lig_joiner_code = "I-S";
        $common {lig_joiner_code} = $lig_joiner_code;
        my %lig_results = obtain_short_interal_tm (%common);
        my %final = ligamer_piece_joiner (%lig_results);
        $final{pcrsequence} = "";
        my %bed_output = %final;
        output (%final);
        #my %bed_final = prep_bed (%bed_output);
        };

     }## matching brace for ligamer data lines

   else {next};
}## Matching brace for csv file input test
####### END LIGAMERS ASSEMBLY PORTION

close OUT;
close INPUT;
print "Program Finished.\n";
exit;

####### END MAJOR WORK OF PROGRAM !!
################################################################################
#### Begin Subroutine section of program.
sub Terminal_5 {
my %results = (@_);

my $T3_seq = "";
my $T3_tm = "";
my $T3_seq_length = "";


$results {T3_seq} = $T3_seq;
$results {T3_tm} = $T3_tm;
$results {T3_seq_length} = $T3_seq_length;

return %results;
}
################################################################################
##
################################################################################
sub Terminal_3 {
```

```
my %results = (@_);

my $T5_seq="";
my $T5_tm="";
my $T5_seq_length="";

$results {T5_seq} = $T5_seq;
$results {T5_tm} = $T5_tm;
$results {T5_seq_length} = $T5_seq_length;

return %results;
}
################################################################################

################ BEGIN Subroutine for short slices
   ##############################
sub obtain_short_interal_tm {

my %results = (@_);
my $working_sequence = $results{working_sequence};

my $working_sequence_tm_obj=
Bio::SeqFeature::Primer -> new(-seq=>$working_sequence);
my $T5_tm = $working_sequence_tm_obj->
        Tm
                (
                -salt => $salt,
                -oligo => $lig_conc
                );

$T5_tm=substr($T5_tm,0,5);
my $T5_seq = $working_sequence;

$results{working_sequence} = $working_sequence;
$results{T5_seq} = $T5_seq;
$results{T5_tm} = $T5_tm;

return (%results);
}
################################################################################

################################################################################
sub output { my %results=(@_);
```

```
print OUT $results{gene                 },"\t"; # 0
print OUT $results{ligamer_name         },"\t"; # 1
print OUT $results{species              },"\t"; # 2
print OUT $results{strand               },"\t"; # 3
print OUT $results{lig_joiner_code      },"\t"; # 4
print OUT $results{target_prime         },"\t"; # 5
print OUT $results{UCSCcoordinates      },"\t"; # 6
print OUT $results{pcrsequence          },"\t"; # 7
print OUT $results{barcode              },"\t"; # 8
print OUT $results{target_seq_length    },"\t"; # 9
print OUT $results{T5_seq               },"\t"; # 10
print OUT $results{T5_seq_length        },"\t"; # 11
print OUT $results{T5_tm                },"\t"; # 12
print OUT $results{T3_seq               },"\t"; # 13
print OUT $results{T3_seq_length        },"\t"; # 14
print OUT $results{T3_tm                },"\t"; # 15
print OUT $results{ligamer              },"\t"; # 16
print OUT $results{warning              };      #
print OUT $results{ligamer_length       },"\t"; # 17
print OUT $results{notes                },"\t"; # 18
print OUT $results{set                  },"\t"; # 19


#Commented on 022013
#if ( defined $results{T5_ctrl_length} ) {
#  print OUT $results{  T5_ctrl_length  },"\t";
#  }

#if ( defined $results{T3_ctrl_length} ) {
#  print OUT $results{  T3_ctrl_length  },"\t";
#  }

print OUT "\n";

}
################################################################################

######BEGIN Subroutine to parse csv file into variables
    ########################
sub parse_the_line {

my $line = shift(@_);

my      (
```

```
        $gene,
        $ligamer_name,
        $species,
        $strand,
        $lig_location,
        $target_prime,
        $UCSCcoordinates,
        $barcode,
        $set
        )
        = split /\t/ , $line ;

$gene=~s/^<//;
print "Gene - $gene\n";
print "Ligmamer name - $ligamer_name\n";
print "species - $species\n";
print "strand - $strand\n";
print "lig_location - $lig_location\n";
print "target_prime - $target_prime\n";
print "UCSC - $UCSCcoordinates\n";
print "barcode - [$barcode]\n";
print "Set - [$set]\n";

if ($barcode=~/ /){$barcode=~s/ //}  ## GO HERE!

$gene=~s/<//;

return
  (
  $gene,
  $species,
  $strand,
  $lig_location,
  $target_prime,
  $UCSCcoordinates,
  $barcode,
  $set
  );

}

######END Subroutine to parse csv file into variables ############
```

```
######BEGIN Subroutine to parse genomic coordnates into variables
    ############
sub parse_coordinates {

my $input=shift(@_);

my ($chr,$coordinates)  =split /\:/,$input;

my ($start,$end)        =split /\-/,$coordinates;
#$chr=~s/chr//;  # I have comment out this to behave with local fasta
    files!

$start=~s/\,//g;
$end=~s/\,//g;

return ($chr, $start, $end);

}
######END Subroutine to parse genomic coordnates into variables
    #################

######Subroutine to make revcom depending on strand
    annoation####################
sub revcom_slice_based_on_strand {

my ($strand, $slice_sequence) = @_;

#if the strand is positive - make the reverse compliment
$strand=lc($strand);

if ($strand eq 'plus') {
  $working_sequence = reverse($slice_sequence);
  $working_sequence =~ tr/ACGTacgt/TGCAtgca/;
  }
# if the strand is minus - do nothing
if ($strand eq 'minus') {
  $working_sequence = $slice_sequence;
  }

return $working_sequence
}

###### END SUBROUTINE revcom_slice_based_on_strand ####################
```

```
################ BEGIN Subroutine to obtained only 5' end of working
    sequence##
sub obtain_T5_tm_sequence {

my $working_sequence = shift (@_);
my $temp = shift (@_);
my $lig_location = shift (@_);
my $control_length=shift (@_);
my $salt=shift (@_);
my $lig_conc=shift (@_);
my $working_sequence_length = length ($working_sequence);
my $T5_seq_length;

if ($lig_location eq "TC") {$T5_seq_length=$control_length};
if ($lig_location eq "T") {$T5_seq_length=19};
if ($lig_location eq "I") {$T5_seq_length=19};

my $T5_tm=0;
my $T5_seq;
my $T5_seq_out;

while($T5_tm < $temp)
  {
  $T5_seq_length++;
  print ".";
  $T5_seq=substr $working_sequence,0, $T5_seq_length;
  my $T5_seq_primer=
  Bio::SeqFeature::Primer ->
    new
    (
    -seq=>$T5_seq
    );
$T5_tm = $T5_seq_primer ->
    Tm
    (
    -salt=>$salt,
    -oligo=>$lig_conc
    );
  $T5_tm=substr($T5_tm,0,5);
  if ($T5_seq_length eq $working_sequence_length) {last;}
  if ($T5_tm>=$temp)
    {
    $T5_seq_out = $T5_seq;
```

```
    print "\n";
    last
    }

  if ($T5_seq_length eq 33)
    {
    $T5_seq_out=$T5_seq;
    print "\n";
    print STDERR "Warning: ".
    "Assembly at line $. T5 side cut".
    " off due to low Tm \n";
    last
    }
  elsif ($T5_tm<=$temp){next}
  }
return ($T5_seq_out, $T5_tm, $T5_seq_length);
}

################ BEGIN Subroutine to obtained only 3' end of working
    sequence##
sub obtain_T3_tm_sequence {

my $working_sequence = shift (@_);
my $temp = shift (@_);
my $lig_location = shift (@_);
my $control_length=shift (@_);
my $salt=shift (@_);
my $lig_conc=shift (@_);
my $working_sequence_length = length ($working_sequence);
my $T3_seq_length;

if ($lig_location eq "TC") {$T3_seq_length=(-$control_length)};
if ($lig_location eq "T") {$T3_seq_length=(-19)};
if ($lig_location eq "I") {$T3_seq_length=(-19)};

my $T3_tm=0;
my $T3_seq;
my $T3_seq_out;

while ($T3_tm < $temp )
  {
  print ".";
  $T3_seq_length--;
  $T3_seq=
```

```
  substr $working_sequence, $T3_seq_length;
  my $T3_seq_primer=
  Bio::SeqFeature::Primer ->
    new
    (
   -seq=>$T3_seq
    );
  $T3_tm = $T3_seq_primer ->
    Tm
    (
    -salt=>$salt,
    -oligo=>$lig_conc
     );
  $T3_tm=substr($T3_tm,0,5);
  if ($T3_seq_length eq (-$working_sequence_length)) {last;}
  if ($T3_seq_length<(-80)){die}
  if ($T3_tm>=$temp)
    {
    $T3_seq_out=$T3_seq;
    print "\n";
    last
    }

  if ($T3_seq_length eq (-33))
    {
    $T3_seq_out=$T3_seq;
    print "\n";
    print STDERR "Warning: ".
    "Assembly at line $. T3 side cut".
    " off due to low Tm \n";
    last
    }
  if ($T3_tm<$temp){next}
  }
return ($T3_seq_out, $T3_tm, $T3_seq_length);
}

################ END Subroutine to obtained only 3' end of working
    sequence####

################ BEGIN Subroutine to joined pieces of ligamer   #########
sub ligamer_piece_joiner{

my %results = @_;
```

```
my $lig_joiner_code     = $results{lig_joiner_code};
my $T5_seq              = $results{T5_seq};
my $barcode             = $results{barcode};
my $T3_seq              = $results{T3_seq};
my $pcrsequence;
my $short_sequence=$T5_seq;
my $ligamer;
my $ligamer_length;
my $warning=" ";
my $Phos_mod_code="\/5Phos\/";

if ($lig_joiner_code eq "T-5")
  {
  $pcrsequence=$results{three_prime_PCR_sequence};
  $ligamer = join ("",$Phos_mod_code, $T5_seq, $barcode,$pcrsequence);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($lig_joiner_code eq "T-C-5-I")
  {
  $pcrsequence=$results{three_prime_PCR_sequence};
  $ligamer = join ("",$Phos_mod_code,$short_sequence);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($lig_joiner_code eq "T-C-5-T")
  {
  $pcrsequence=$results{three_prime_PCR_sequence};
  $ligamer = join ("",$Phos_mod_code, $T5_seq, $barcode, $pcrsequence);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($lig_joiner_code eq "T-3")
  {
  $pcrsequence=$results{five_prime_PCR_sequence};
  $ligamer = join ("",$pcrsequence,$barcode,$T3_seq);
  $ligamer_length = length $ligamer;
  }

if ($lig_joiner_code eq "T-C-3-I")
```

```
  {
  $pcrsequence=$results{five_prime_PCR_sequence};
  $ligamer = join ("",$Phos_mod_code,$T3_seq);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($lig_joiner_code eq "T-C-3-T")
  {
  $pcrsequence=$results{five_prime_PCR_sequence};
  $ligamer = join ("",$pcrsequence,$barcode,$T3_seq);
  $ligamer_length = length $ligamer;
  }

if ($lig_joiner_code eq "I-S")
  {
  $ligamer = join ("",$Phos_mod_code,$short_sequence);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($lig_joiner_code eq "I-L")
  {
  $ligamer = join ("",$Phos_mod_code,$T5_seq,$barcode,$T3_seq);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($lig_joiner_code eq "I-L-C")
  {
  $ligamer = join ("",$Phos_mod_code,$T5_seq,$barcode,$T3_seq);
  $ligamer_length = length $ligamer;
  $ligamer_length = $ligamer_length-7;
  }

if ($ligamer_length > 60)
  {
  print STDERR
  "Warning! The ligamer from input file data line $.".
  " has a length greater than 60!\n";
  };


$results{pcrsequence}           = $pcrsequence;
```

```
$results{ligamer}               = $ligamer;
$results{ligamer_length}        = $ligamer_length;
$results{warning}               = $warning;

return %results;

}
################################################################################

################################################################################
## Load the latest Ensembl Registry
sub ensembl_database{

my $verbose=shift@_;
my $db_version=shift@_;

my $registry = 'Bio::EnsEMBL::Registry';
print "Beginning to login to Ensembl database version $db_version.\n";
$registry->load_registry_from_db
  (
  -host => 'ensembldb.ensembl.org',
  -user => 'anonymous',
  -db_version => $db_version,
  -verbose => $verbose,
  );

print "Done loading ensembl database.\n";
return $registry;
}
################################################################################

################ BEGIN Subroutine to obtained get genomic sequence slice
   #######
sub get_genomic_sequence {

my ($chr, $start, $end, $species, $db ) = @_;

my $slice_adaptor = $db->get_adaptor( $species, 'Core', 'Slice');

$chr=~s/^chr//;

my $slice = $slice_adaptor->
  fetch_by_region
    (
```

```
    'chromosome',
    $chr,
    $start,
    $end,
    );

my $slice_sequence = ($slice->seq);

return $slice_sequence;
}
################ END Subroutine to obtained get genomic sequence slice
   ######
#############################################################################
################ BEGIN Subroutine to PROVIDE POD HELP DATA ######
sub podhelp {

my $scriptname=         shift@_;
my $help_print=         shift@_;
my $man_print=          shift@_;
my $perlname=$scriptname;
my $htmlname=$scriptname;
my $manname=$scriptname;

if ($help_print eq 1)
  {
  $htmlname =~ s/\.pl/\.html/;
  system "pod2html $perlname --title=$perlname --outfile=$htmlname";
  print "\n\t$htmlname printed in cwd.\n\n";
  exit
  }

if ($man_print eq 1)
  {
  $manname =~ s/\.pl/\.man/;
  system "pod2man $perlname $manname";
  print "\n\t$manname printed in $dir.\n\n";
  system "man -l $manname|less";
  exit
  }
}
################ END Subroutine to PROVIDE POD HELP DATA ######
```

Bibliography

Altschul, S., Gish, W., and Miller, W. (1990). Basic local alignment search tool. *Journal of molecular . . . .*

Amitsur, M., Levitz, R., and Kaufmann, G. (1987). Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA. *The EMBO journal*, 6(8):2499–503.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Anstey, M. L., Rogers, S. M., Ott, S. R., Burrows, M., and Simpson, S. J. (2009). Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. *Science (New York, N.Y.)*, 323(5914):627–30.

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(7099):203–207.

Aravin, A. A. and Hannon, G. J. (2008). Small RNA silencing pathways in germ and stem cells. *Cold Spring Harbor symposia on quantitative biology*, 73:283–90.

Aravin, A. A., Hannon, G. J., and Brennecke, J. (2007a). The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race. *Science*, 318(5851):761–764.

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. *Current Biology*, 11(13):1017–1027.

Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007b). Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science (New York, N.Y.)*, 316(5825):744–7.

Ashe, A., Sapetschnig, A., Weick, E.-M., Mitchell, J., Bagijn, M. P., Cording, A. C., Doebley, A.-L., Goldstein, L. D., Lehrbach, N. J., Le Pen, J., Pintacuda,

G., Sakaguchi, A., Sarkies, P., Ahmed, S., and Miska, E. a. (2012). piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell*, 150(1):88–99.

Bailey, T. L., Boden, M., Buske, F. a., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue):W202–8.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings International Conference on Intelligent Systems for Molecular Biology ISMB International Conference on Intelligent Systems for Molecular Biology*, 2:28–36.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. (2010). Deciphering the splicing code. *Nature*, 465(7294):53–59.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)*, 338(6114):1587–93.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. a., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, 315(5819):1709–12.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. a., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5.

Baudat, F., Manova, K., Yuen, J. P., Jasin, M., and Keeney, S. (2000). Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Molecular cell*, 6(5):989–98.

Bell, G. I., Pictet, R. L., Rutter, W. J., Cordell, B., Tischer, E., and Goodman, H. M. (1980). Sequence of the human insulin gene. *Nature*, 284(5751):26–32.

Benson, D. a., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). GenBank. *Nucleic acids research*, 39(Database issue):D32–7.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. a., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T.,

Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. a., Benoit, V. a., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. a., Brown, R. C., Brown, A. a., Buermann, D. H., Bundu, A. a., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. a., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. a., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. a., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. a., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. a., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. a., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9.

Benton, M. J. and Donoghue, P. C. J. (2007). Paleontological evidence to date the tree of life. *Molecular biology and evolution*, 24(1):26–53.

Bernstein, S. I., Mogami, K., Donady, J. J., and Emerson, C. P. (1983). Drosophila muscle myosin heavy chain encoded by a single gene in a cluster of muscle mutations. *Nature*, 302(5907):393–397.

Besmer, P., Jr., R. C. M., Caruthers, M. H., Kumar, A., Minamoto, K., van de Sande, J., Sidarova, N., and Khorana, H. (1972). Studies on polynucleotides.

CXVII. Hybridization of polydeoxynucleotides with tyrosine transfer RNA sequences to the r-strand of phi80psu + 3 DNA. *Journal of Molecular Biology*, 72:503–522.

Bingham, J. L., Carrigan, P. E., Miller, L. J., and Srinivasan, S. (2008). Extent and diversity of human alternative splicing established by complementary database annotation and microarray analysis. *Omics: A Journal of Integrative Biology*, 12(1):83–92.

Blencowe, B. J., Ahmad, S., and Lee, L. J. (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & development*, 23(12):1379–86.

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(15):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(6):1089–1103.

Brennecke, J. and Malone, C. (2008). An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. *Science (New York, ...*, 322(November):1387–1392.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–4.

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., Davis, C. a., Frise, E., Hammonds, A. S., Olson, S., Shenker, S., Sturgill, D., Samsonova, A. a., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J., Bickel, P. J., Carninci, P., Cherbas, P., Gingeras, T. R., Hoskins, R. a., Kaufman, T. C., Lai, E. C., Oliver, B., Perrimon, N., Graveley, B. R., and Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*, pages 1–7.

Bullard, D. and Bowater, R. (2006). Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4. *Biochemical Journal*, 398(Pt 1):135–144.

Burnette, J. M., Miyamoto-Sato, E., Schaub, M. a., Conklin, J., and Lopez, a. J. (2005). Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics*, 170(2):661–74.

Burnside, J., Ouyang, M., Anderson, A., Bernberg, E., Lu, C., Meyers, B. C., Green, P. J., Markis, M., Isaacs, G., Huang, E., and Morgan, R. W. (2008). Deep sequencing of chicken microRNAs. *BMC genomics*, 9:185.

Calarco, J. A., Saltzman, A. L., Ip, J. Y., and Blencowe, B. J. (2007a). Technologies for the global discovery and analysis of alternative splicing. In *Advances in Experimental Medicine and Biology*, volume 623, pages 64–84.

Calarco, J. A., Xing, Y., Cáceres, M., Calarco, J. P., Xiao, X., Pan, Q., Lee, C., Preuss, T. M., and Blencowe, B. J. (2007b). Global analysis of alternative splicing differences between humans and chimpanzees. *Genes & Development*, 21(22):2963–75.

Carmell, M. A., Girard, A., van de Kant, H. J., Bourc'his, D., Bestor, T. H., de Rooij, D. G., and Hannon, G. J. (2007). MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. *Developmental Cell*, 12(4):503–514.

Celotto, a. M. and Graveley, B. R. (2001). Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics*, 159(2):599–608.

Cenik, E. S. and Zamore, P. D. (2011). Argonaute proteins. *Current biology : CB*, 21(12):R446–9.

Chang, H., Lim, J., Ha, M., and Kim, V. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3 End Modifications. *Molecular Cell*, pages 1–9.

Chauleau, M. and Shuman, S. (2013). Kinetic mechanism of nick sealing by T4 RNA ligase 2 and effects of 3'-OH base mispairs and damaged base lesions. *RNA (New York, N.Y.)*, 19(12):1840–1847.

Chen, C., Jin, J., James, D. A., Adams-cioaba, M. A., Gyoon, J., Guo, Y., Tenaglia, E., Xu, C., Gish, G., Min, J., and Pawson, T. (2009). Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. *Proceedings of the National Academy of Sciences*, 106(48):20336–20341.

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(6):1106–17.

Chuma, S. and Hosokawa, M. (2006). Tdrd1/Mtr-1 ,a male tudor-related gene, is essential for male germ-cell differentiationand nuage/germinal grandule formation in mice. *Proceedings of the . . .*, 103(43):1–6.

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 50(2):162–8.

Conze, T., Goransson, J., Razzaghian, H. R., Ericsson, O., Oberg, D., Akusjarvi, G., Landegren, U., and Nilsson, M. (2010). Single molecule analysis of combinatorial splicing. *Nucl. Acids Res.*, page gkq581.

Conze, T., Shetye, A., Tanaka, Y., Gu, J., Larsson, C., Göransson, J., Tavoosidana, G., Söderberg, O., Nilsson, M., and Landegren, U. (2009). Analysis of genes, transcripts, and proteins via DNA ligation. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 2:215–39.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322:1845–1848.

Cramer, P., Pesce, C. G., Baralle, F. E., and Kornblihtt, A. R. (1997). Functional association between promoter structure and transcript alternative splicing. *Proceedings of the National Academy of Sciences*, 94(21):11456–11460.

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*, advance on.

Deng, Q. Q.-l., Ishii, S., and Sarai, A. (1996). Binding site analysis of c-Myb : screening of potential binding sites by using the mutation matrix derived from systematic binding affinity measurements. *Nucleic Acids Research*, 24(4):766–774.

Deng, W. and Lin, H. (2002). miwi, a Murine Homolog of piwi, Encodes a Cytoplasmic Protein Essential for Spermatogenesis. *Developmental cell*, 2(6):819–830.

Djebali, S., Davis, C. a., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. a., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–8.

Djebali, S., Kapranov, P., Foissac, S., Lagarde, J., Reymond, A., Ucla, C., Wyss, C., Drenkow, J., Dumais, E., Murray, R. R., Lin, C., Szeto, D., Denoeud, F., Calvo, M., Frankish, A., Harrow, J., Makrythanasis, P., Vidal, M., Salehi-Ashtiani, K., Antonarakis, S. E., Gingeras, T. R., and Guigo, R. (2008). Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat Meth*, 5(7):629–635.

Emerick, M., Parmigiani, G., and Agnew, W. (2007). Multivariate Analysis and Visualization of Splicing Correlations in Single-Gene Transcriptomes. *BMC Bioinformatics*, 8(1):16.

Fagnani, M., Barash, Y., Ip, J. Y., Misquitta, C., Pan, Q., Saltzman, A. L., Shai, O., Lee, L., Rozenhek, A., Mohammad, N., Willaime-Morawek, S., Babak, T., Zhang, W., Hughes, T. R., van der Kooy, D., Frey, B. J., and Blencowe, B. J. (2007). Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biology*, 8(6):R108.

Farazi, T. a., Juranek, S. a., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development (Cambridge, England)*, 135(7):1201–14.

Fareed, G. C., Wilt, E. M., and Richardson, C. C. (1971). Enzymatic Breakage and Joining of Deoxyribonucleic Acid. *Journal of Biological Chemistry*, 246:925–932.

Fededa, J. P., Petrillo, E., Gelfand, M. S., Neverov, A. D., Kadener, S., Nogués, G., Pelisch, F., Baralle, F. E., Muro, A. F., and Kornblihtt, A. R. (2005). A Polar Mechanism Coordinates Different Regions of Alternative Splicing within a Single Gene. *Molecular Cell*, 19(3):393–404.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391:806–811.

Fodor, A. a. and Aldrich, R. W. (2009). Convergent evolution of alternative splices at domain boundaries of the BK channel. *Annual review of physiology*, 71:19–36.

Furuichi, Y. (1975). 5'-terminal m7G(5')ppp(5')Gmp In Vivo: Identification in Reovirus Genome RNA. *Proceedings of the . . .*, 72(2):742–745.

Gan, H., Lin, X., Zhang, Z., Zhang, W., Liao, S., Wang, L., and Han, C. (2011). piRNA profiling during specific stages of mouse spermatogenesis. *RNA*.

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477.

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E. L. W., Zapp, M. L., Weng, Z., and Zamore, P. D. (2008). Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. *Science*, 320(5879):1077–1081.

Ghosh, S. and Jacobson, A. (2010). RNA decay modulates gene expression and controls its fidelity. *WIREs RNA*, 1:351–361.

Gilbert, W. (1978). Why genes in pieces? *Nature*, 271(5645):501.

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(7099):199–202.

Glauser, D. A., Johnson, B. E., Aldrich, R. W., and Goodman, M. B. (2011). Intragenic alternative splicing coordination is essential for Caenorhabditis elegans slo-1 gene function. *Proceedings of the National Academy of Sciences of the United States of America*, 108(51):20790–5.

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. *Nature Biotechnology*, 29(7):644–652.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, 27(7):1017–8.

Graveley, B. R. (2000). Sorting out the complexity of SR protein functions. *RNA (New York, N.Y.)*, 6(9):1197–211.

Graveley, B. R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell*, 123(1):65–73.

Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. *Genes & development*, 20(13):1709–14.

Gu, A., Ji, G., Shi, X., Long, Y., Xia, Y., Song, L., Wang, S., and Wang, X. (2010). Genetic Variants in Piwi-Interacting RNA Pathway Genes Confer Susceptibility to Spermatogenic Failure in a Chinese Population. *Human Reproduction*, 25(12):2955–2961.

Gu, W., Lee, H.-C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D., and Mello, C. C. (2012). CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell*, 151(7):1488–500.

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.

Gupta, S., Stamatoyannopoulos, J. a., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, 8(2):R24.

Haase, A. D., Fenoglio, S., Muerdter, F., Guzzardo, P. M., Czech, B., Pappin, D. J., Chen, C., Gordon, A., and Hannon, G. J. (2010). Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. *Genes &*

development, 24(22):2499–504.

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C., and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods (San Diego, Calif.)*, 44(1):3–12.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr., M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141.

Handler, D., Olivieri, D., Novatchkova, M., Gruber, F. S., Meixner, K., Mechtler, K., Stark, A., Sachidanandam, R., and Brennecke, J. (2011). A systematic analysis of Drosophila TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *The EMBO journal*, 30(19):3977–93.

Hartig, J. V., Tomari, Y., and Förstemann, K. (2007). piRNAs–the ancient hunters of genome invaders. *Genes & development*, 21(14):1707–13.

Hattori, D., Chen, Y., Matthews, B. J., Salwinski, L., Sabatti, C., Grueber, W. B., and Zipursky, S. L. (2009). Robust discrimination between self and non-self neurites requires thousands of Dscam1 isoforms. *Nature*, 461(7264):644–648.

Hattori, D., Millard, S. S., Wojtowicz, W. M., and Zipursky, S. L. (2008). Dscam-Mediated Cell Recognition Regulates Neural Circuit Formation. *Annual Review of Cell and Developmental Biology*, 24(1):597–620.

Hemani, Y. and Soller, M. (2012). Mechanisms of Drosophila Dscam mutually exclusive splicing regulation. *Biochemical Society transactions*, 40(4):804–9.

Ho, C. K., Etten, J. L. V., and Shuman, S. (1997). Characterization of an ATP-dependent DNA ligase encoded by Chlorella virus PBCV-1. *Journal of virology*, 71(3).

Ho, C. K. and Shuman, S. (2002). Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains. *Proceedings of the National Academy of Sciences*, 99(20):12709–12714.

Ho, C. K., Wang, L. K., Lima, C. D., and Shuman, S. (2004). Structure and mechanism of RNA ligase. *Structure (London, England: 1993)*, 12(2):327–339.

Hoeffer, C. A., Sanyal, S., and Ramaswami, M. (2003). Acute induction of conserved synaptic signaling pathways in Drosophila melanogaster. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23:6362–6372.

Horvath, G. C., Kistler, M. K., and Kistler, W. S. (2009). RFX2 is a candidate downstream amplifier of A-MYB regulation in mouse spermatogenesis. *BMC*

Developmental Biology, 9(1):63.

Hosokawa, M., Shoji, M., Kitamura, K., Tanaka, T., Noce, T., Chuma, S., and Nakatsuji, N. (2007). Tudor-related proteins TDRD1/MTR-1, TDRD6 and TDRD7/TRAP: domain composition, intracellular localization, and function in male germ cells in mice. *Developmental biology*, 301(1):38–52.

House, A. E. and Lynch, K. W. (2008). Regulation of alternative splicing: more than just the ABCs. *The Journal of Biological Chemistry*, 283(3):1217–21.

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(1):69–82.

Huang, H., Gao, Q., Peng, X., Choi, S.-Y., Sarma, K., Ren, H., Morris, A. J., and Frohman, M. a. (2011). piRNA-associated germline nuage formation and spermatogenesis require MitoPLD profusogenic mitochondrial-surface lipid signaling. *Developmental cell*, 20(3):376–87.

Hughes, T. A. (2006). Regulation of gene expression by alternative untranslated regions. *Trends in Genetics: TIG*, 22(3):119–22.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324:218–223.

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802.

Ipsaro, J. J., Haase, A. D., Knott, S. R., Joshua-Tor, L., and Hannon, G. J. (2012). The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature*, 491(7423):279–83.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *The EMBO Journal*, 22(4):905–12.

Johnson, B. E., Glauser, D. A., Dan-Glauser, E. S., Halling, D. B., Aldrich, R. W., and Goodman, M. B. (2011). Alternatively Spliced Domains Interact to Regulate BK Potassium Channel Gating. *Proceedings of the National Academy of Sciences*, 108(51):20784–20789.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., and Shoemaker, D. D. (2003). Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science*, 302(5653):2141–2144.

Kawaoka, S., Hara, K., Shoji, K., Kobayashi, M., Shimada, T., Sugano, S., Tomari, Y., Suzuki, Y., and Katsuma, S. (2012). The comprehensive

epigenome map of piRNA clusters. pages 1–10.

Kennard, E. H. (1927). Zur Quantenmechanik einfacher Bewegungstypen. *Zeitschrift for Physik*, 44(4-5):326–352.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology*, 10(2):126–39.

Kleppe, K., Van de Sande, J. H., and Khorana, H. G. (1970). Polynucleotide ligase-catalyzed joining of deoxyribo-oligonucleotides on ribopolynucleotide templates and of ribo-oligonucleotides on deoxyribopolynucleotide templates. *Proceedings of the National Academy of Sciences of the United States of America*, 67(1):68–73.

Kojima, K., Kuramochi-Miyagawa, S., Chuma, S., Tanaka, T., Nakatsuji, N., Kimura, T., and Nakano, T. (2009). Associations between PIWI proteins and TDRD1/MTR-1 are critical for integrated subcellular localization in murine male germ cells. *Genes to cells : devoted to molecular & cellular mechanisms*, 14(10):1155–65.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. a., McCombie, W. R., Jarvis, E. D., and Adam M Phillippy (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700.

Kreahling, J. M. and Graveley, B. R. (2005). The iStem, a long-range RNA secondary structure element required for efficient exon inclusion in the Drosophila Dscam pre-mRNA. *Molecular and Cellular Biology*, 25(23):10251–60.

Kumar, M. and Carmichael, G. G. (1998). Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiology and molecular biology reviews : MMBR*, 62(4):1415–34.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T. W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., Lin, H., Matsuda, Y., and Nakano, T. (2004). Mili, a Mammalian Member of Piwi Family Gene, Is Essential for Spermatogenesis. *Development*, 131(4):839–849.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T. W., Hata, K., Li, E., Matsuda, Y., Kimura, T., Okabe, M., Sakaki, Y., Sasaki, H., and Nakano, T. (2008). DNA Methylation of Retrotransposon Genes Is Regulated by Piwi Family Members MILI and MIWI2 in Murine Fetal Testes. *Genes & Development*, 22(7):908–917.

Kutter, C., Brown, G. D., Gonçalves, 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. *Nature genetics*, 43(10):948–55.

Ladd, A. N. and Cooper, T. A. (2002). Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology*, 3(11):reviews0008.

Laiho, A., Kotaja, N., Gyenesei, A., and Sironen, A. (2013). Transcriptome profiling of the murine testis during the first wave of spermatogenesis. *PloS one*, 8(4):e61558.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

Latham, K. E., Litvin, J., Orth, J. M., Patel, B., Mettus, R., and Reddy, E. P. (1996). Temporal patterns of A-myb and B-myb gene expression during testis development. *Oncogene*, 13:1161–1168.

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 (New York, N.Y.)*, 313(5785):363–367.

Lee, H.-C., Gu, W., Shirayama, M., Youngman, E., Conte, D., and Mello, C. C. (2012). C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell*, 150(1):78–87.

Lenasi, T., Peterlin, B. M., and Dovc, P. (2006). Distal regulation of alternative splicing by splicing enhancer in equine β-casein intron 1. *RNA*, 12(3):498 –507.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079.

Li, H., Qiu, J., and Fu, X.-D. (2012). RASL-seq for massively parallel and quantitative analysis of gene expression. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 4(April):Unit 4.13.1–9.

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., and Cheung, V. G. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. *Science (New York, N.Y.)*, 333(6038):53–8.

Li, X. C., Barringer, B. C., and Barbash, D. a. (2009b). The pachytene checkpoint and its relationship to evolutionary patterns of polyploidization and hybrid sterility. *Heredity*, 102(1):24–30.

Li, X. C. and Schimenti, J. C. (2007). Mouse pachytene checkpoint 2 (trip13) is required for completing meiotic recombination but not synapsis. *PLoS genetics*, 3(8):e130.

Li, X. Z., Roy, C. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W., Xu, J., Moore, M. J., Schimenti, J. C., Weng, Z., and Zamore, P. D. (2013a). An Ancient Transcription Factor Initiates the Burst of piRNA Production during

Early Meiosis in Mouse Testes. *Molecular Cell*, 50(1):1–15.

Li, X. Z., Roy, C. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W., Xu, J., Moore, M. J., Schimenti, J. C., Weng, Z., and Zamore, P. D. (2013b). An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. *Molecular Cell*, pages 1–15.

Li, X. Z., Roy, C. K., Moore, M. J., and Zamore, P. D. (2013c). Defining piRNA primary transcripts. *Cell cycle (Georgetown, Tex.)*, 12(11):1657–8.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–36.

Lohman, G. J. S., Zhang, Y., Zhelkovsky, A. M., Cantor, E. J., Evans, T. C., and Jr, T. C. E. (2013). Efficient DNA ligation in DNA – RNA hybrid helices by Chlorella virus DNA ligase. *Nucleic acids research*, 42(36):1–14.

Long, J. C. and Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *The Biochemical journal*, 417(1):15–27.

Lynch, K. W. (2004). Consequences of regulated pre-mRNA splicing in the immune system. *Nat Rev Immunol*, 4(12):931–940.

Macilwain, C. (2010). Scientists vs engineers: this time it's financial. *Nature*, 467(7318):885.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980 –11985.

Marinov, G. K., Williams, B. a., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., and Wold, B. J. (2013). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome research*, pages 496–510.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.-F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. *Advances in experimental medicine and biology*, 623:123–147.

Maxam, A. M. and Gilbert, W. (1992). A new method for sequencing DNA. 1977. *Biotechnology (Reading, Mass.)*, 24(2):99–103.

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*,

338(6114):1593–1599.

Mettus, R. V., Litvin, J., Wali, A., Toscani, A., Latham, K., Hatton, K., and Reddy, E. P. (1994). Murine A-myb: evidence for differential splicing and tissue-specific expression. *Oncogene*, 9(10):3077–86.

Miura, S. K., Martins, A., Zhang, K. X., Graveley, B. R., and Zipursky, S. L. (2013). Probabilistic splicing of Dscam1 establishes identity at the level of single neurons. *Cell*, 155(5):1166–77.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–9.

Modrich, P., Anraku, Y., and Lehman, I. R. (1973). Deoxyribonucleic Acid Ligase ISOLATION AND PHYSICAL CHARACTERIZATION OF THE HOMOGENEOUS ENZYME FROM ESCHERICHIA COLI. *Journal of Biological Chemistry*, 248(21):7495–7501.

Modzelewski, A. J., Holmes, R. J., Hilz, S., Grimson, A., and Cohen, P. E. (2012). AGO4 regulates entry into meiosis and influences silencing of sex chromosomes in the male mouse germline. *Developmental cell*, 23(2):251–64.

Montgomery, M. K., Xu, S., and Fire, a. (1998). RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26):15502–7.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621–628.

Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, 24(1):22–30.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–9.

Nakano, M., Komatsu, J., Matsuura, S.-i., Takashima, K., Katsura, S., and Mizuno, A. (2003). Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology*, 102(2):117–124.

Namekawa, S. H. and Lee, J. T. (2009). XY and ZW: is meiotic sex chromosome inactivation the rule in evolution? *PLoS genetics*, 5(5):e1000493.

Nandakumar, J., Ho, C. K., Lima, C. D., and Shuman, S. (2004). RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2. *The Journal of biological chemistry*, 279(30):31337–47.

Nandakumar, J., Shuman, S., and Lima, C. D. (2006). RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward. *Cell*, 127(1):71–84.

NEBEL, B. R., AMAROSE, A. P., and HACKET, E. M. (1961). Calendar of gametogenic development in the prepuberal male mouse. *Science (New York, N.Y.)*, 134(3482):832–3.

Neves, G., Zucker, J., Daly, M., and Chess, A. (2004). Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nature genetics*, 36(3):240–6.

Newburger, D. E. and Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research*, 37(Database issue):D77–82.

Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463.

Nilsson, M., Antson, D.-O., Barbany, G., and Landegren, U. (2001). RNA-templated DNA ligation for transcript analysis. *Nucleic Acids Research*, 29(2):578–581.

Nilsson, M., Barbany, G., Antson, D. O., Gertow, K., and Landegren, U. (2000). Enhanced detection and distinction of RNA by enzymatic probe ligation. *Nature biotechnology*, 18(7):791–3.

Nishimasu, H., Ishizu, H., Saito, K., Fukuhara, S., Kamatani, M. K., Bonnefond, L., Matsumoto, N., Nishizawa, T., Nakanaga, K., Aoki, J., Ishitani, R., Siomi, H., Siomi, M. C., and Nureki, O. (2012). Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature*, 491(7423):284–7.

Oakberg, E. and Oakberq, F. (1956). A Description of Spermiogenesis in the mouse and its use in analysis of the cycle of the seminiferous epithelium and germ cell renewal. *The American journal of anatomy*, 99(3):391–419.

Oh, I. H. and Reddy, E. P. (1999). The myb gene family in cell growth, differentiation and apoptosis. *Oncogene*, 18(19):3017–33.

Olivera, B. and Lehman, I. (1967). Linkage of polynucleotides through phosphodiester bonds by an enzyme from Escherichia coli. *Proceedings of the National Academy of . . .* , pages 1426–1433.

Olson, S., Blanchette, M., Park, J., Savva, Y., Yeo, G. W., Yeakley, J. M., Rio, D. C., and Graveley, B. R. (2007). A regulator of Dscam mutually exclusive splicing fidelity. *Nat Struct Mol Biol*, 14(12):1134–1140.

Osella, M., Bosia, C., Corá, D., and Caselle, M. (2011). The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS computational biology*, 7(3):e1001101.

Osheim, Y. N., Miller, O. L., and Beyer, a. L. (1985). RNP particles at splice junction sequences on Drosophila chorion transcripts. *Cell*, 43(1):143–51.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–5.

Pane, A., Wehr, K., and Schüpbach, T. (2007). zucchini and squash encode

two putative nucleases required for rasiRNA production in the Drosophila germline. *Developmental cell*, 12(6):851–62.

Park, H. Y., Lim, H., Yoon, Y. J., Follenzi, A., Nwokafor, C., Meng, X., Singer, R. H., and Yoon, H. (2014). Visualization of dynamics of single endogenous mRNA labeled in live mouse. *Science (New York, N.Y.)*, 343(6169):422–4.

Park, J. W. and Graveley, B. R. (2007). Complex alternative splicing. *Advances in Experimental Medicine and Biology*, 623:50–63.

Parra, M. K., Gallagher, T. L., Amacher, S. L., Mohandas, N., and Conboy, J. G. (2012). Deep intron elements mediate nested splicing events at consecutive AG dinucleotides to regulate alternative 3' splice site choice in vertebrate 4.1 genes. *Molecular and cellular biology*, 32(11):2044–53.

Parra, M. K., Tan, J. S., Mohandas, N., and Conboy, J. G. (2008). Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene. *EMBO J*, 27(1):122–131.

Peirson, S. N. (2003). Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. *Nucleic Acids Research*, 31(14):73e–73.

Pélisson, a., Song, S. U., Prud'homme, N., Smith, P. a., Bucheton, a., and Corces, V. G. (1994). Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *The EMBO journal*, 13(18):4401–11.

Peng, T., Xue, C., Bi, J., Li, T., Wang, X., Zhang, X., and Li, Y. (2008). Functional importance of different patterns of correlation between adjacent cassette exons in human and mouse. *BMC Genomics*, 9(1):191.

Plocik, A. M. A. and Graveley, B. R. (2013). New insights from existing sequence data: generating breakthroughs without a pipette. *Molecular cell*, 49(4):605–17.

Purandare, S. R., Tenhumberg, B., and Brisson, J. a. (2014). Comparison of the wing polyphenic response of pea aphids (Acyrthosiphon pisum) to crowding and predator cues. *Ecological Entomology*, 39(2):263–266.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842.

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*, advance on.

Reuter, M., Chuma, S., Tanaka, T., Franz, T., Stark, A., and Pillai, R. S. (2009). Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. *Nature structural & molecular biology*, 16(6):639–46.

Reuveni, S., Urbakh, M., and Klafter, J. (2014). Role of substrate unbinding in Michaelis-Menten enzymatic reactions. *Proceedings of the National Academy of Sciences of the United States of America*, 111(12).

Richardson, C. (1965). Phosphorylation of Nucleic Acid by and Enzyme from T4 Bacteriophate-infected Escherichia Coli. *Proceedings of the National Academy of . . .* , 1372(1961):158–165.

Riley, K. J. and Steitz, J. A. (2013). Minireview The " Observer Effect " in Genome-wide Surveys of Protein-RNA Interactions Minireview. pages 2011–2014.

Ro, S., Park, C., Song, R., Nguyen, D., Jin, J., Sanders, K. M., McCarrey, J. R., and Yan, W. (2007). Cloning and Expression Profiling of Testis-Expressed piRNA-Like RNAs. *RNA*, 13(10):1693–1702.

Roach, J. J. C., Boysen, C., Wang, K., Hood, L., and Wang, I. K. A. I. (1995). Pairwise End Sequencing: A unified approach to genomic mapping and sequencing. *Genomics*, 353:345–353.

Robine, N., Lau, N. C., Balla, S., Jin, Z., Okamura, K., Kuramochi-Miyagawa, S., Blower, M. D., and Lai, E. C. (2009). A Broadly Conserved Pathway Generates 3' UTR-Directed Primary piRNAs. *Current Biology*, 19(24):2066–2076.

Romanienko, P. J. and Camerini-Otero, R. D. (2000). The mouse Spo11 gene is required for meiotic chromosome synapsis. *Molecular cell*, 6(5):975–87.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281:363, 365.

Saito, K., Inagaki, S., Mituyama, T., Kawamura, Y., Ono, Y., Sakota, E., Kotani, H., Asai, K., Siomi, H., and Siomi, M. C. (2009). A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. *Nature*, 461(7268):1296–9.

Salehi-Ashtiani, K., Yang, X., Derti, A., Tian, W., Hao, T., Lin, C., Makowski, K., Shen, L., Murray, R. R., Szeto, D., Tusneem, N., Smith, D. R., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2008). Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. *Nature Methods*, 5(7):597–600.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–8.

Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., and Smith, M. (1978). The nucleotide sequence of bacteriophage phiX174. *Journal of molecular biology*, 125(2):225–46.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative

monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–70.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., and Zipursky, S. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *Cell*, 101(6):671–684.

Schoenmakers, S., Wassenaar, E., Hoogerbrugge, J. W., Laven, J. S. E., Grootegoed, J. A., and Baarends, W. M. (2009). Female meiotic sex chromosome inactivation in chicken. *PLoS genetics*, 5(5):e1000466.

Schwarzbauer, J. E., Tamkun, J. W., Lemischka, I. R., and Hynes, R. O. (1983). Three different fibronectin mRNAs arise by alternative splicing within the coding region. *Cell*, 35(2 Pt 1):421–31.

Seitz, H., Ghildiyal, M., and Zamore, P. D. (2008). Argonaute loading improves the 5' precision of both MicroRNAs and their miRNA* strands in flies. *Current biology : CB*, 18(2):147–51.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. T., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., and Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, pages 1–5.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9):618–630.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, 31(11):1009–1014.

Sharp, P. A. (2014). Nobel Lectures in Physiology or Medicine 1991-1995.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–8.

Shendure, J. and Aiden, E. L. (2012). The expanding scope of DNA sequencing. *Nature Biotechnology*, 30(11):1084–1094.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

Shendure, J., Mitra, R. D., Varma, C., and Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews. Genetics*, 5(5):335–44.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741):1728–1732.

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(4):761 –772.

Shi, L. and Lee, T. (2012). Molecular diversity of Dscam and self-recognition. *Advances in experimental medicine and biology*, 739:262–75.

Shingleton, A., Sisk, G., and Stern, D. (2003). Diapause in the pea aphid (Acyrthosiphon pisum) is a slowing but not a cessation of development. *BMC developmental biology*, 12:1–12.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, 100:15776–15781.

Shirayama, M., Seth, M., Lee, H.-C., Gu, W., Ishidate, T., Conte, D., and Mello, C. C. (2012). piRNAs initiate an epigenetic memory of nonself RNA in the C. elegans germline. *Cell*, 150(1):65–77.

Shiroguchi, K., Jia, T. Z., Sims, P. a., and Xie, X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1347–52.

Shoji, M., Tanaka, T., Hosokawa, M., Reuter, M., Stark, A., Kato, Y., Kondoh, G., Okawa, K., Chujo, T., Suzuki, T., Hata, K., Martin, S. L., Noce, T., Kuramochi-Miyagawa, S., Nakano, T., Sasaki, H., Pillai, R. S., Nakatsuji, N., and Chuma, S. (2009). The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germline. *Developmental cell*, 17(6):775–87.

Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol*, 12(4):246–258.

Smith, J. M., Bowles, J., Wilson, M., Teasdale, R. D., and Koopman, P. (2004). Expression of the tudor-related gene Tdrd5 during development of the male germline in mice. *Gene expression patterns : GEP*, 4(6):701–5.

Southern, E. M. (2001). DNA microarrays. History and overview. *Methods in Molecular Biology (Clifton, N.J.)*, 170:1–15.

Sriskanda, V. and Shuman, S. (1998). Specificity and fidelity of strand joining by Chlorella virus DNA ligase. *Nucleic acids research*, 26(15):3536–41.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610.

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., and Bartel, D. P. (2014).

Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956–960.

Summers, W. and Siegel, R. (1970). Transcription of Late Phage RNA by T7 RNA Polymerase. *Nature*, 228:1160–1162.

Tabor, S. (1987). DNA Ligases. *Current Protocols in Molecular Biology*, pages 3.14.1–3.14.4.

Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome research*, 22(5):947–56.

Tazi, J., Bakkour, N., and Stamm, S. (2009). Alternative splicing and disease. *Biochimica et biophysica acta*, 1792(1):14–26.

Thierry-Mieg, D. and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome biology*, 7 Suppl 1(Suppl 1):S12.1–14.

Thomson, T. and Lin, H. (2009). The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annual review of cell and developmental biology*, 25:355–76.

Toscani, A., Mettus, R. V., Coupland, R., Simpkins, H., Litvin, J., Orth, J., Hatton, K. S., and Reddy, E. P. (1997). Arrest of spermatogenesis and defective breast development in mice lacking A-myb. *Nature*, 386(6626):713–717.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105 –1111.

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. *Nature Protocols*, 7(3):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 Biotech*, 28(5):511–515.

Trauth, K., Mutschler, B., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., and Klempnauer, K.-h. (1994). Mouse A-myb encodes a trans-activator and is expressed in mitotically active cells of the developing central nervous system, adult testis and B lymphocytes. *EMBO*, 13(24):5994–6005.

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: A method for

identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–386.

Vagin, V. V., Klenov, M. S., Stolyarenko, A. D., and Kotelnikov, R. N. (2004). The RNA Interference Proteins and Vasa Locus are Involved in the Silencing of Retrotransposons in the Female Germline of Drosophila melanogaster. *RNA Biology*, (June):54–58.

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(5785):320 –324.

Vagin, V. V., Wohlschlegel, J., Qu, J., Jonsson, Z., Huang, X., Chuma, S., Girard, A., Sachidanandam, R., Hannon, G. J., and Aravin, A. a. (2009). Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes & development*, 23(15):1749–62.

Vasileva, A., Tiedau, D., Firooznia, A., Müller-Reichert, T., and Jessberger, R. (2009). Tdrd6 is required for spermiogenesis, chromatoid body architecture, and regulation of miRNA expression. *Current biology : CB*, 19(8):630–9.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science (New York, N.Y.)*, 270(5235):484–7.

Venter, J. (2007). *A life decoded: my genome, my life*. Penguin (Non-Classics).

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. a., Holt, R. a., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, a. G., Nadeau, J., McKusick, V. a., Zinder, N., Levine, a. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, a. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. a., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, a. K., Narayan, V. a., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K.,

Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51.

Viollet, S., Fuchs, R. T., Munafo, D. B., Zhuang, F., and Robb, G. B. (2011). T4 RNA Ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotechnology*, 11(1):72.

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. *Nature Structural & Molecular Biology*, 19(8):773–781.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. a., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E., and Chang, H. Y. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–9.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6.

Wang, L. K., Lima, C. D., and Shuman, S. (2002). Structure and mechanism of T4 polynucleotide kinase: an RNA repair enzyme. *The EMBO journal*,

21(14):3873–80.

Watanabe, T., Chuma, S., Yamamoto, Y., Kuramochi-Miyagawa, S., Totoki, Y., Toyoda, A., Hoki, Y., Fujiyama, A., Shibata, T., Sado, T., Noce, T., Nakano, T., Nakatsuji, N., Lin, H., and Sasaki, H. (2011a). MITOPLD Is a Mitochondrial Protein Essential for Nuage Formation and piRNA Biogenesis in the Mouse Germline. *Developmental Cell*, 20(3):364–375.

Watanabe, T., Tomizawa, S.-i., 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. (2011b). Role for piRNAs and Noncoding RNA in de Novo DNA Methylation of the Imprinted Mouse Rasgrf1 Locus. *Science*, 332(6031):848 –852.

Watson, F. L., Püttmann-Holgado, R., Thomas, F., Lamar, D. L., Hughes, M., Kondo, M., Rebel, V. I., and Schmucker, D. (2005). Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science (New York, N.Y.)*, 309(5742):1874–8.

Watson, J. D., Gann, A., and Witkowski, J. (2012). *The Annotated and Illustrated Double Helix*. Simon & Schuster.

Watson, J. D. J. and Crick, F. (1953). Molecular structure of nucleic acids. *Nature*, 4356(171):737–738.

Wei, C. M., Gershowitz, a., and Moss, B. (1975). Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. *Cell*, 4(4):379–86.

Weiss, B. and Richardson, C. (1967). Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. *Proceedings of the National Academy of . . . .*

Weston, K. (1992). Extension of the DNA binding consensus of the chicken c-Myb and v-Myb proteins. *Nucleic acids research*, 20(12).

White, E. S. and Muro, A. F. (2011). Fibronectin splice variants: understanding their multiple roles in health and disease using engineered mouse models. *IUBMB life*, 63(7):538–46.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8):748–752.

Yabuta, Y., Ohta, H., Abe, T., Kurimoto, K., Chuma, S., and Saitou, M. (2011). TDRD5 is required for retrotransposon silencing, chromatoid body assembly, and spermiogenesis in mice. *The Journal of cell biology*, 192(5):781–95.

Yang, Y., Zhan, L., Zhang, W., Sun, F., Wang, W., Tian, N., Bi, J., Wang, H., Shi, D., Jiang, Y., Zhang, Y., and Jin, Y. (2011). RNA secondary structure in mutually exclusive splicing. *Nat Struct Mol Biol*, 18(2):159–168.

Yeakley, J. M., Fan, J.-B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M. S., and Fu, X.-D. (2002). Profiling alternative splicing on fiber-optic arrays. *Nature Biotechnology*, 20(4):353–358.

Yin, S., Ho, C. K., and Shuman, S. (2003). Structure-function analysis of T4 RNA ligase 2. *The Journal of biological chemistry*, 278(20):17601–8.

Zhan, X.-L., Clemens, J. C., Neves, G., Hattori, D., Flanagan, J. J., Hummel, T., Vasconcelos, M. L., Chess, A., and Zipursky, S. L. (2004). Analysis of Dscam Diversity in Regulating Axon Guidance in Drosophila Mushroom Bodies. *Neuron*, 43(5):673–686.

Zhang, F., Wang, J., Xu, J., Zhang, Z., Koppetsch, B. S., Schultz, N., Vreven, T., Meignin, C., Davis, I., Zamore, P. D., Weng, Z., and Theurkauf, W. E. (2012a). UAP56 Couples piRNA Clusters to the Perinuclear Transposon Silencing Machinery. *Cell*, 151(4):871–884.

Zhang, Y., Liu, T., Meyer, C. a., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137.

Zhang, Z., Theurkauf, W. E., Weng, Z., and Zamore, P. D. (2012b). Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly (A) selection. *Silence*, 3(1):9.

Zhu, J., Shendure, J., Mitra, R. D., and Church, G. M. (2003). Single molecule profiling of alternative pre-mRNA splicing. *Science (New York, N.Y.)*, 301(5634):836–8.

Zikherman, J. and Weiss, A. (2008). Alternative Splicing of CD45: The Tip of the Iceberg. *Immunity*, 29(6):839–841.