UNIVERSITY OF CALIFORNIA
RIVERSIDE

A Genomic Exploration of Transposable Element and piRNA Occupancy, Abundance, and Functionality

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics, and Bioinformatics

by

Patrick A. Schreiner

June 2017

Dissertation Committee:
Dr. Peter Atkinson, Chairperson
Dr. Thomas Girke
Dr. Jason Stajich

The Dissertation of Patrick A. Schreiner is approved:

Committee Chairperson

University of California, Riverside

Acknowledgements

I acknowledge that the content in Chapter 2 has been published within the "The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species" article by Ppanicolaou *et. al* in Genome Biology on September 22, 2016. Chapter 3 has been published to the Biorxiv preprint server entitled "piClusterBusteR: Software For Automated Classification And Characterization Of piRNA Cluster Loci" with Dr. Peter Atkinson on May 1, 2017.

I am very grateful to my research advisor, Dr. Peter Atkinson, for the opportunity to pursue research in his laboratory. His knowledge, experience, and mentorship were critical in guiding the success of this research. The regular, open-minded scientific discussion and resources that Dr. Atkinson provided put me in an excellent position to succeed in graduate school and beyond.

I would also like to thank the members of my guidance, oral exam, and dissertation committees, Dr. Thomas Girke and Dr. Jason Stajich. Their knowledge and experience in the field, teaching ability, mentorship, and patience were invaluable in allowing my professional development as a scientist in the field of Bioinformatics. The contributions that Dr. Girke and Dr, Stajich made with regard to project development and execution were critical for the quality and applicability of this research. I cannot thank Dr. Atkinson, Dr. Girke, and Dr. Stajich enough for their support of my career goals and my job search.

I owe gratitude to my family, friends, and colleagues for their personal and professional support during the pursuit of my degree. My fiancé, Lily Maxham, was a stable, supportive, and understanding partner during a particularly unpredictable and demanding period of time. My parents, Janet and Edward Schreiner, as well as brothers, Timothy and Matthew Schreiner, were extraordinarily helpful in promoting physical, social, and mental well-being during graduate school. I can't express enough the appreciation that I have in knowing that I have that unwavering support of these individuals at all times.

I greatly appreciate my fellow lab members, Michael Han and Lee Doss, for their friendship and establishing the light and fun, yet productive environment that I enjoyed coming to day after day. Their knowledge of the lab and the field, as well as their availability for personal and professional discussion around the clock was extremely helpful.

I am also very lucky and appreciative to have additional staff and faculty that supported my career pursuit before, during, and after the pursuit of my Ph.D. at UCR. My undergraduate research advisor and continuing mentor, Dr. Catherine Putonti at Loyola University of Chicago, provided the opportunity to learn from undergraduate research and presentation. In doing so, she supported and facilitated my success before and during my Ph.D.

I am grateful for the opportunity to work with the Director of Graduate Student Professional Development, Dr. Margaret Gover, for the final two years at UCR.

Her support, experience, and commitment to my success played a major role in facilitating my degree progress and achieving my career goals directly after graduating with my Ph.D.

I owe thanks to Dr. Cheryl Hayashi for allowing independence and the opportunity to excel as a Teaching Assistant in the Biology department. Her personal and formal appreciation provided me the opportunity to succeed in the classroom and take on leadership roles with confidence.

ABSTRACT OF THE DISSERTATION

A Genomic Exploration of Transposable Element Occupancy, Abundance, and Functionality

by

Patrick A. Schreiner

Doctor of Philosophy, Graduate Program in Genetics, Genomics, and Bioinformatics
University of California, Riverside, June 2017
Dr. Peter W. Atkinson, Chairperson

The research herein aims to promote a better understanding of the occupancy, architecture, and diversity of transposable elements and piRNA clusters across species. Transposable elements are identified, characterized, and classified in an effort to pursue molecular strategies to control a species regarded among the most significant agricultural pests, *Ceratitis capitata*. I present a software tool available for general use, piClusterBusteR, that is capable of quickly and accurately annotating the contents of piRNA clusters in any species of interest. The conserved architecture, yet tissue-specific

nature of piRNA clusters is described in the ovaries and testes of 13 Metazoan species. A reproducible and statistically significant metric is also demonstrated regarding the relative utilization of piRNA amplification. This software, TruePaiR, generated benchmark values for comparison and context in notable tissues and species, as well as demonstrated the capability to identify subtle differences in piRNA biogenesis across control and experimental conditions. Finally, a bioinformatic workflow is generated to observe the potential for piRNA-mediated post-transcriptional regulation via poly(A) deadenylation of protein-coding genes. By further understanding the presence, architecture, and targets of transposable elements and piRNAs, this research contributes to knowledge of the organismal evolution and development, as well as the conservation and potential to engineer the piRNA pathway in a therapeutic context across species.

Table of Contents

Chapter 6: Summary and Conclusions

List of Tables

List of Figures

Chapter 1: Introduction

Section 1.1: Transposable Elements

Transposable elements (TEs) are mobile genetic elements that encode the ability to replicate themselves within a genome (McClintock 1956). TEs are referred to as "selfish" genetic elements since they encode to facilitate their transposition within a genome (Smit 1999; Doolittle & Sapienza 1980). TEs are critical drivers of evolution that have been identified ubiquitously in eukaryotic organisms from Protozaons to Fungi and Vertebrates (Feschotte & Pritham 2007). Their role in genome rearrangement results in a great deal genotypic and phenotypic variability observed within and between species (Whitelaw & Martin 2001; Barton & Keightley 2002). Tens of thousands of individual TEs have been identified and made in public databases such as RepBase, TEfam, and Dfam (Jurka et al. 2005; Terenius et al. 2008; Wheeler et al. 2013). Due to differential rates of transposition, TEs can have a range of genome occupation across eukaryotic species. For example, TEs compose over 70% of the Maize genome, almost 50% of the human and mosquito (*Aedes aegypti*) genome, almost 40% of the mouse genome, and over 15% of the fruity fly genome (Smit 1999; Feschotte & Pritham 2007; Arensburger et al. 2011).

Unregulated transposition can cause integration of the TE sequence into the genome and yield phenotypic distinctions within a particular organism (Levin & Moran 2011). Therefore, a regulatory mechanism – via sRNAs – is crucial in many species to suppress

genome perturbation by TEs and allow for proper development (Brennecke et al. 2007; Aravin et al. 2007; Malone & Hannon 2010; Malone et al. 2009).

Section 1.2: The Mechanisms of Transposition

TEs are a crucial driving force of evolution and variation within a population. TEs can facilitate mutagenesis, translocations, as well as gene fusions and duplications within a genome (Feschotte & Pritham 2007; Malone & Hannon 2010). TEs can vary greatly in their preference to integrate into genomic regions based upon gene density and chromatin state (Mills et al. 2011). Genome perturbation by TEs can yield advantageous or disadvantageous functional changes within the cell. Given the importance of the TE movement and the genetic cargo carried during a transposition event, the mechanisms that facilitate are important to consider.

TEs are most generally classified into RNA (Class I) or DNA (Class II) TEs, based on their mechanism of transposition (Wicker et al. 2007). Fully intact TEs encode the means by which they facilitate their own transposition. Although these classes of TEs vary greatly in their mechanism of transposition, both mechanisms ideally result in a duplicate copy of the original element in a new location within the genome. Both mechanisms of transposition also produce RNA, albeit performing unique functions, to facilitate their movement (Spradling & Rubin 1982; McClintock 1956). The RNA produced in the process of transposition is the target of silencing via piRNAs (Aravin et al. 2007).

RNA TEs, also referred to a Class I TEs, facilitate their own transposition via an RNA intermediate. In RNA TE transposition, the TE is transcribed into RNA via RNA polymerase II, converted to cDNA, and then integrated into another location within the genome (Sims et al. 2004). The TE copies itself via a RNA intermediate, without ever leaving its initial position within the genome. Therefore, this mechanism is often described as "copy-and-paste" mechanism of transposition (Feng et al. 1998; Cost et al. 2002; Schmidt 1999).

Class I TEs can be further categorized into long terminal repeat (LTR) or non-long terminal repeat (Non-LTR) elements (Malik et al. 1999). This difference in LTR and Non-LTR classification is with regard to difference in TE structure, mechanism of integration, and evolutionary origin of the encoded reverse transcriptase (Xiong & Eickbush 1990).

The structure of a fully intact LTR TE contains genes that are necessary for transposition: most notably the *gag*, *pol*, and *env* genes. LTR TEs have been characterized into four superfamilies (Jurka et al. 2005). LTR TEs can also encode for protease and integrase within the element. As the name suggests, LTR transposons have long terminal repeats at either end of the element (Cost et al. 2002). The mechanism of LTR transposition is referred to as replicative retrotransposition. The element is initially transcribed by RNA polymerase II (Sims et al. 2004). The *gag* gene encodes for a polyprotein that forms the retroviral core structure. The *pol* gene serves as the reverse transcriptase required for RNA TE transposition within the retroviral core. The *env* gene encodes an envelope protein that facilitates interaction with the target cell membrane (Finnegan 1989). Upon

generation of the cDNA from the template TE RNA, the cDNA is transported to the nucleus and integrated into target DNA (Wessler et al. 1995).

The other subclass of Class I TEs, Non-LTR TEs, has a quite different structure within the element. There are 33 superfamilies of Non-LTR that are currently recognized (Jurka et al. 2005). Non-LTR TEs do not contain long terminal repeats at either end of the element (Löwer et al. 1996). Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are the two major subgroups of Class I element classification (Smit 1996). Within the structure of a fully intact Non-LTR TE, a 5' UTR, open reading frame(s), 3' UTR, and poly(A) tail can generally be identified (Schmidt 1999). The Non-LTR mechanism of transposition is referred to as target-primed reverse transcription. Like LTR transposition, Non-LTR transposition begins with transcription by RNA polymerase II and is followed by generation of cDNA via the reverse transcriptase. The Non-LTR encoded protein is endonuclease then facilitates the integration of the cDNA into the target DNA (Cost et al. 2002; Feng et al. 1998).

Class II TEs are referred to as DNA transposons due to the necessity for a DNA donor in the transposition mechanism of this class of TEs (Yuan & Wessler 2011). There are 23 recognized superfamilies of DNA transposons (Jurka et al. 2005). The open reading frame(s) (ORF) of transposable element encodes for the transposase protein that facilitates movement at each end of the element. In facilitating the movement of a Class II TE, the donor element is excised from its site of origin, to be moved within a site of the target DNA. Therefore, Class II transposition is referred to as the "cut-and-paste" mechanism of TE mobility (Yuan & Wessler 2011).

Integration of TEs can be incomplete. Transposition that results in a TE that encodes for incomplete factors within the element, and therefore cannot facilitate its own transposition, is referred to as a non-autonomous TE. Terminal inverted repeats are generally, but not always found at each end of a complete or incomplete TE (Kapitonov & Jurka 2001). Miniature inverted-repeat transposable elements (MITEs) are minimal non-autonomous elements that are composed of nearly only TIRs (Wessler et al. 1995).

The presence and movement of TEs within a genome can affect transcriptional regulation and gene content. TE movement within a genome can affect cellular function by moving or altering promoters and enhancers, deriving novel genes, altering existing genes, promoting gene fusions, and affecting the chromatin availability of particular genomic regions. Also, after integration into the genome, TEs often accumulate mutations to render them incapable of transposition, and potentially to serve as an alternative function (Wessler et al. 1995).

Section 1.3: TE-derived Small RNAs

Small RNAs (sRNAs) are important non-protein coding RNAs that can regulate activity within the cell. sRNAs can be classified into one of three major categories in species within the Animalia kingdom: small interfering-RNA (siRNA), microRNA (miRNA), or PIWI-interacting RNA (piRNA). The three subclasses of sRNA in the Animalia kingdom differ in terms of their respective size, target specificity, and mechanisms of

biogenesis (Chung et al. 2008; Malone & Hannon 2010; Millar & Waterhouse 2005; Liu et al. 2008). I was particularly interested in studying the least thoroughly characterized among the subclasses of sRNA, the piRNA system.

Section 1.3.1: PIWI-interacting ribonucleic acids (piRNAs)

piRNAs are sRNAs that function via association with members of the PIWI protein family. piRNAs represent a subset of small RNAs that generally range between 24-33 nucleotides (nts) in length (Zhang et al. 2014). Their mechanism of biogenesis is unique from that of other sRNAs in that it is Dicer-independent (Brennecke et al. 2007; Aravin et al. 2007). Rather, piRNAs can be classified according to their primary or secondary mechanism of biogenesis. Those piRNAs generated from the primary mechanism of biogenesis are referred to as primary piRNAs. Due to the observation that primary piRNAs predominately originate from the antisense strand, it has been hypothesized that primary piRNAs originate from one long transcript. This long transcript is thought to originate from a particular region, or several particular heterochromatic genomic loci referred to as piRNA clusters (Brennecke et al. 2007).

Section 1.3.2: piRNA clusters

piRNA clusters have been defined as genomic regions that contain a high number of uniquely mapping piRNAs (Arensburger et al. 2011). These regions are generally found

in and around pericentromeric and telomeric heterochromatin in *Drosophila melanogaster* (Brennecke et al. 2007). piRNA clusters can widely vary in size, sequence, and distribution, depending on the organism of interest. These loci can be range from a few kilobases to hundreds of kilobases, consisting mainly of fragments of inactivated transposable elements (Brennecke et al. 2007; Arensburger et al. 2011; Schreiner & Atkinson 2017). However, known sequences of genic and viral origin, as well as large regions of unknown origin, have been noted in the piRNA clusters in Metazoan species. Although the genomic loci of piRNAs clusters is often well conserved within species, the primary sequence is much more variable between species (Schreiner & Atkinson 2017).

piRNA clusters can be classified as unidirectional, dual-strand or bidirectional, referring to the mechanism of transcription of this loci (Yamanaka et al. 2014). Unidirectional piRNA cluster transcription occurs in one direction, resulting in primary piRNAs that are generated in the same orientation with respect to their genomic origin. The most well-studied unidirectional piRNA cluster was among the first piRNA clusters to be identified in Drosophila melanogaster: the *flamenco* locus (Brennecke et al. 2007).

Dual-strand piRNA clusters have transcription occur at both terminal ends of the piRNA clusters, with the direction of transcription occurring towards the center of the piRNA cluster. Given the two opposite directions of transcription, two primary precursor transcripts are generated (Brennecke et al. 2007; Malone et al. 2009). The two precursor transcripts are processed by the same downstream factors as unidirectional piRNA clusters. The 42AB locus is a well-known dual-strand piRNA cluster in *Drosophila melanogaster* (Brennecke et al. 2007; Malone & Hannon 2010).

Bidirectional piRNA clusters are the least common amongst the mechanisms of piRNA cluster transcription (Aravin et al. 2008; Schreiner & Atkinson 2017). Bidirectional piRNA clusters have a central promoter which leads to transcription in each outward direction. It is interesting to note that transcripts from bidirectional piRNA clusters lack defined loci for the termination of transcription (Aravin et al. 2008; Yamanaka et al. 2014).

Section 1.3.3: piRNA Targets and Function

Given the repressive capability of piRNAs, identifying the targets of piRNAs is crucial in assessing their current and potential role within the cell. TEs are known to predominantly target TEs, but have the capability to repress protein-coding and viral RNA (Brennecke et al. 2007; Aravin et al. 2007; Arensburger et al. 2011; Schreiner & Atkinson 2017). Targets are likely dictated both by the piRNA clusters and TEs present in an individual organism's genome, as well as the environmental conditions encountered (Bregliano et al. 1980; Brennecke et al. 2008).

A well-studied mechanism of piRNA regulation is Argonaute-mediated RNA slicing. piRNAs facilitate target slicing by associating with PIWI proteins – Aubergine, Argonaute3, and Piwi – to guide protein-mediated cleavage via sequence complementarity to target transcripts (Brennecke et al. 2007). Primary piRNAs associate with Aubergine which guides the complex to its target mRNA molecule and results in Argonaute-mediated slicing of the target. The processed fragments of sense transcripts,

or secondary piRNAs, can then associate with Ago3, which then continues slicing via the same mechanism on other targets complementary to the associated secondary piRNA. In the continuous process of slicing, processing, and complementing to the next target, piRNA slicing provides a positive feedback loop of piRNA generation (Aravin et al. 2007).

Although the mechanism of piRNA-mediated silencing was initially thought to occur strictly via target slicing, the function of the nuclear PIWI protein, Piwi, is independent of its slicing activity in *Drosophila melanogaster*. Without its slicing activity, Piwi is still capable of inducing epigenetic modifications (Darricarrère et al. 2013). Piwi induces epigenetic modifications via association with other proteins. Piwi has been shown to recruit the proteins Heterochromatin Protein 1a (HP1a) and histone methyltransferase, Su(var)3-9. HP1a, guided by the Piwi-RISC, can bind to histones, where Su(var)3-9 can induce methylation at the H3K9 modification (Brower-Toland et al. 2007; Lu et al. 2013). Histone methylation at a particular site results in an even more probable binding site for further HP1a interaction (Huang et al. 2013). The association of HP1a is known to induce heterochromatin formation (James & Elgin 1986). As a result of increased heterochromatin at Piwi-RISC guided loci, the capability of association with RNA polymerase II is reduced (Sims et al. 2004). It has been well established that reduced RNA polymerase II association is correlated with lower levels of transcription (Sims et al. 2004).

piRNAs have also been implicated in an additional mechanism in the suppression of protein-coding mRNA in *Drosophila melanogaster*. Protein-coding gene regulation

mediated by piRNAs is occurs, much like the epigenetic mechanism of repression, via PIWI family proteins' association with other functional proteins. In piRNA-mediated deadenylation, as shown in the model of repression in the *Nanos* 3' UTR, requires the association of Aub and Ago3 in a common complex with Smaug, CAF1, and CCR4 (Rouget et al. 2010). Smaug is an RNA-binding protein (RBP) that facilitates the recruitment of other machinery necessary for deadenylation (Semotok et al. 2005). CAF1 and CCR4 form an exonuclease complex that is capable of modifying the 3' end of mRNA molecules (Chen et al. 2002; Temme et al. 2004). The mechanism of piRNA-mediated deadenylation occurs by piRNA association with a PIWI protein, either Aub or Ago3, leads to the recruitment of the Smaug, CAF1, and CCR4 complex to target molecules via sequence complementarity. The CCR4 exonuclease is then in a position to associate with the poly(A) tail of the transcript, leading to poly(A) tail shortening (Chen et al. 2002). Transcripts with a shortened poly(A) tails are recognized and lead to degradation (Salles et. al., 1999). It is important to note that although both miRNAs and piRNAs have been shown to affect deadenylation in Nanos in *Drosophila melanogaster*, it has also been demonstrated that the regulatory complex can bind to its target, to a lesser extent, in the absence of RNA (Pinder et. al., 2012). This observation is indicative of the sRNA's complementary, but not absolute, role in promoting transcript poly(A) deadenylation.

Section 1.4: Maternal Deposition of piRNAs

The functionality of piRNAs has been shown to play an essential role in fertility and in the maintenance of genome integrity. Initial observations leading to the concepts that demonstrate the importance in the presence of piRNAs was observed decades ago with the attempted cross of a female fly from a laboratory inbred line with a male fly taken from their natural environment. The offspring of this cross yielded hybrid dysgenesis: sterility due to high mutation rate, chromosomal rearrangement, or recombination (Bingham et al. 1982). On the other hand, when a wild female fly was crossed with a male fly from a laboratory inbred line, fertile offspring developed (Bregliano et al. 1980). More recent studies have concluded that this phenomenon is likely due to the naivety of the laboratory female to the wild male's active TEs (Brennecke et al. 2008). Without maternal deposition of piRNAs, the progeny was unable to inhibit the movement of active, genomic TEs. Therefore, it was shown that the piRNA mechanism of genome defense is inherited maternally in offspring, rather than paternally in *Drosophila melanogaster,* and that piRNA-mediated silencing of the active TEs present in the progeny is necessary for proper development (Bingham et al. 1982;

Brennecke et al. 2008). piRNAs are deposited along with other crucial sRNAs, mRNAs, and proteins to facilitate the development of the embryo (Malone et al. 2009; Bushati et al. 2008).

Section 1.5: Proteins in the piRNA Pathway

P-element induced wimpy testis (PIWI) proteins are essential for piRNA-mediated regulation. The PIWI proteins – Aubergine (Aub), Argonaute 3 (Ago3), and Piwi – are members of the Argonaute protein family that have demonstrated a non-redundant function in the piRNA pathway. The PIWI proteins can be differentiated by their subcellular localization, mechanism of repression, as well as their piRNA association (Table 1.1). Piwi is the nuclear PIWI protein that facilitates primary piRNA-mediated slicing and epigenetic regulation. Aub and Ago3 are the PIWI proteins that facilitate the amplification loop of secondary piRNA biogenesis. Aub associates with primary piRNAs, while Ago3 associates with secondary piRNAs to slice target RNA (Brennecke et al. 2007; Aravin et al. 2007).

Although, the number of PIWI proteins can vary between species. For example, there are two PIWI proteins in *Danio rerio*, three PIWI proteins in *Drosophila melanogaster* and *Mus musculus*, while there have been seven identified PIWI proteins in *Aedes aegypti* (Campbell et al. 2008; Nene et al. 2007). It is still under investigation whether the expansion of PIWI proteins in *Aedes aegypti* is due to redundant, compensatory, supplementary, or novel function.

piRNAs physically associate with the PIWI proteins, guide the proteins to the target molecule via sequence complementarity, before slicing the target molecule (Brennecke et al. 2007; Aravin et al. 2007). The PAZ domain of the PIWI protein aids in the physical association of the RISC with the RNA molecule, and the PIWI domain facilitates the

slicing of its target mRNA (Parker et al. 2005; Yan et al. 2003; Song et al. 2003; Cerutti et al. 2000). Argonaute proteins are known to slice between the tenth and eleventh nucleotide of bound RNA (Haley & Zamore 2004; Martinez & Tuschl 2004; Elbashir et al. 2001).

The nuclear PIWI protein, Piwi, can function without its slicing activity. Research has shown that mutation within the catalytic triad critical for Argonaute-mediated target slicing does not affect the stability, localization, nor function of Piwi (Darricarrère et al. 2013). Instead, the piRNA-guided Piwi recruits factors that can induce epigenetic modifications such as HP1a and Su(var)3-9 (Brower-Toland et al. 2007; Lin & Yin 2008). HP1a can facilitate the formation of heterochromatin and Su(var)3-9 can induce H3K9me2/3 histone modifications (Huang et al. 2013). Both of these epigenetic modifications can regulate transcriptional control of TEs.

	Piwi	**Aub**	**Ago3**
Localization	Nuclear	Cytoplasmic	Cytoplasmic
Regulation	Epigenetic	Post-Transcriptional	Post-Transcriptional
Strand of Associated piRNA Origin	Antisense	Antisense	Sense
Associated piRNA Bias	U at position 1	U at position 1	A at position 10

Table 1.1 | PIWI Protein Family Comparison. A comparison of PIWI proteins: Piwi, Aubergine (Aub), and Argonaute 3 (Ago3) in *Drosophila melanogaster*.

Several important proteins have been well-described regarding effects on the piRNA pathway. An endonuclease, Zucchini, slices the long, initial transcript that has originated from a piRNA cluster, into many fragments (Nishimasu et al. 2012). The 3' end of these

fragments then undergoes a 2'-O-methyl modification. This modification, modulated by HEN1, protects the 3' end of the piRNA from degradation or modification (Montgomery et al. 2012). Those piRNAs that have been sliced and whose 3' end has been modified are referred to as mature piRNAs (Aravin et al. 2007).

Although many of the contributors to piRNA biogenesis are still under investigation, several other proteins were identified as having association with the piRNA pathway. Control and knockdown piRNA populations have also provided insight into the function of these proteins (Malone et al. 2009). For example, the independent knockdowns of *krimper*, *spindle-E*, and *vasa* each led to aberrant localization of PIWI proteins, an increase in piRNA size, and a significant reduction in piRNA amplification. Krimper is a Tudor-domain containing protein, while spindle-E and vasa are putative RNA helicases (Malone et al. 2009). Further understanding of the factors involved in piRNA biogenesis, as well as a comprehensive knowledge of the proteins involved, is crucial to understanding the intricacies of the piRNA pathway.

Section 1.6: References

Aravin, A.A. et al., 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol. Cell*, 31(6), pp.785–799.

Aravin, A.A., Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (80-.).*, 318(5851), pp.761–764.

Arensburger, P. et al., 2011. The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics*, 12(1), p.606.

Barton, N.H. & Keightley, P.D., 2002. Understanding quantitative genetic variation. *Nature Reviews Genetics*, 3(1), pp.11–21.

Bingham, P.M., Kidwell, M.G. & Rubin, G.M., 1982. The molecular basis of PM hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell*, 29(3), pp.995–1004.

Bregliano, J. et al., 1980. Hybrid dysgenesis in Drosophila melanogaster. *Science (80-.).*, 207(4431), pp.606–611.

Brennecke, J. et al., 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science (New York, NY)*, 322(5906), p.1387.

Brennecke, J. et al., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6), pp.1089–103.

Brower-Toland, B. et al., 2007. Drosophila PIWI associates with chromatin and interacts directly with HP1a. *Genes \& development*, 21(18), pp.2300–2311.

Bushati, N. et al., 2008. Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila. *Current Biology*, 18(7), pp.501–506.

Campbell, C.L. et al., 2008. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics*, 9(1), p.425.

Cerutti, L., Mian, N. & Bateman, A., 2000. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends in biochemical sciences*, 25(10), pp.481–482.

Chen, J., Chiang, Y.-C. & Denis, C.L., 2002. CCR4, a 3′-5′ poly (A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. *The EMBO journal*, 21(6), pp.1414–1426.

Chung, W.-J. et al., 2008. Endogenous RNA Interference Provides a Somatic Defense against< i> Drosophila</i> Transposons. *Current Biology*, 18(11), pp.795–802.

Cost, G.J. et al., 2002. Human L1 element target-primed reverse transcription in vitro. *The EMBO Journal*, 21(21), pp.5899–5910.

Darricarrère, N. et al., 2013. Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. *Proceedings of the National Academy of Sciences*, 110(4), pp.1297–1302.

Doolittle, W.F. & Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757), pp.601–3.

Elbashir, S.M. et al., 2001. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *The EMBO journal*, 20(23), pp.6877–6888.

Feng, Q., Schumann, G. & Boeke, J.D., 1998. Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proceedings of the National Academy of Sciences*, 95(5), pp.2083–2088.

Feschotte, C. & Pritham, E.J., 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.*, 41, pp.331–368.

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. *Trends in genetics*, 5, pp.103–107.

Haley, B. & Zamore, P.D., 2004. Kinetic analysis of the RNAi enzyme complex. *Nature structural \& molecular biology*, 11(7), pp.599–606.

Huang, X.A. et al., 2013. A Major Epigenetic Programming Mechanism Guided by piRNAs. *Dev. Cell*.

James, T.C. & Elgin, S., 1986. Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. *Molecular and cellular biology*, 6(11), pp.3862–3872.

Jurka, J. et al., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4), pp.462–467.

Kapitonov, V.V. & Jurka, J., 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15), pp.8714–8719.

Levin, H.L. & Moran, J.V., 2011. Dynamic interactions between transposable elements and their hosts. *Nature reviews. Genetics*, 12(9), pp.615–27.

Lin, H. & Yin, H., 2008. A novel epigenetic mechanism in Drosophila somatic cells mediated by Piwi and piRNAs. In *Cold Spring Harbor symposia on quantitative biology*. p. sqb–2008.

Liu, X., Fortin, K. & Mourelatos, Z., 2008. MicroRNAs: biogenesis and molecular functions. *Brain Pathology*, 18(1), pp.113–121.

Lu, X. et al., 2013. Drosophila H1 regulates the genetic activity of heterochromatin by recruitment of Su (var) 3-9. *Science (80-.).*, 340(6128), pp.78–81.

Löwer, R., Löwer, J. & Kurth, R., 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences*, 93(11), pp.5177–5184.

Malik, H.S., Burke, W.D. & Eickbush, T.H., 1999. The age and evolution of non-LTR retrotransposable elements. *Molecular biology and evolution*, 16(6), pp.793–805.

Malone, C. & Hannon, G., 2010. Molecular evolution of piRNA and transposon control pathways in Drosophila. In *Cold Spring Harbor symposia on quantitative biology*. p. sqb–2009.

Malone, C.D. et al., 2009. Specialized piRNA Pathways Act in Germline and Somatic Tissues of the< i> Drosophila</i> Ovary. *Cell*, 137(3), pp.522–535.

Martinez, J. & Tuschl, T., 2004. RISC is a 5′ phosphomonoester-producing RNA endonuclease. *Genes \& development*, 18(9), pp.975–980.

McClintock, B., 1956. Controlling elements and the gene. In *Cold Spring Harbor symposia on quantitative biology*. pp. 197–216.

Millar, A.A. & Waterhouse, P.M., 2005. Plant and animal microRNAs: similarities and differences. *Functional \& integrative genomics*, 5(3), pp.129–135.

Mills, R.E. et al., 2011. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332), pp.59–65.

Montgomery, T.A. et al., 2012. PIWI associated siRNAs and piRNAs specifically require the Caenorhabditis elegans HEN1 ortholog henn-1. *PLoS genetics*, 8(4), p.e1002616.

Nene, V. et al., 2007. Genome sequence of Aedes aegypti, a major arbovirus vector. *Science (80-.).*, 316(5832), pp.1718–1723.

Nishimasu, H. et al., 2012. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature*, 491(7423), pp.284–7.

Parker, J.S., Roe, S.M. & Barford, D., 2005. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434(7033), pp.663–666.

Rouget, C. et al., 2010. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature*, 467(7319), pp.1128–1132.

Schmidt, T., 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant molecular biology*, 40(6), pp.903–910.

Schreiner, P.A. & Atkinson, P., 2017. piClusterBusteR: Software For Automated Classification And Characterization Of piRNA Cluster Loci. *bioRxiv*, p.133009.

Semotok, J.L. et al., 2005. Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early Drosophila embryo. *Current Biology*, 15(4), pp.284–294.

Sims, R.J., Mandal, S.S. & Reinberg, D., 2004. Recent highlights of RNA-polymerase-II-mediated transcription. *Curr. Opin. Cell Biol.*, 16(3), pp.263–271.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics \& development*, 9(6), pp.657–663.

Smit, A.F., 1996. The origin of interspersed repeats in the human genome. *Current opinion in genetics \& development*, 6(6), pp.743–748.

Song, J.-J. et al., 2003. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nature Structural \& Molecular Biology*, 10(12), pp.1026–1032.

Spradling, A.C. & Rubin, G.M., 1982. Transposition of cloned P elements into Drosophila germ line chromosomes. *Science (80-.).*, 218(4570), pp.341–347.

Temme, C. et al., 2004. A complex containing the CCR4 and CAF1 proteins is involved in mRNA deadenylation in Drosophila. *The EMBO journal*, 23(14), pp.2862–2871.

Terenius, O. et al., 2008. Molecular genetic manipulation of vector mosquitoes. *Cell host \& microbe*, 4(5), pp.417–423.

Wessler, S.R., Bureau, T.E. & White, S.E., 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Current opinion in genetics \& development*, 5(6), pp.814–821.

Wheeler, T.J. et al., 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.*, 41(D1), pp.D70–D82.

Whitelaw, E. & Martin, D.I., 2001. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nature genetics*, 27(4), pp.361–365.

Wicker, T. et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), pp.973–982.

Xiong, Y. & Eickbush, T.H., 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO journal*, 9(10), p.3353.

Yamanaka, S., Siomi, M.C. & Siomi, H., 2014. piRNA clusters and open chromatin structure. *Mobile DNA*, 5(1), p.22.

Yan, K.S. et al., 2003. Structure and conserved RNA binding of the PAZ domain. *Nature*, 426(6965), pp.469–474.

Yuan, Y.-W. & Wessler, S.R., 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U. S. A.*, 108(19), pp.7884–9.

Zhang, P. et al., 2014. piRBase: a web resource assisting piRNA functional study. *Database*, 2014, p.bau110.

Chapter 2: Repeat Element Characterization of the Mediterranean Fruit Fly, *Ceratitis capitata*

Section 2.1: Abstract

The Mediterranean fruit fly (medfly), *Ceratitis capitata*, is considered among the most significant economic and agricultural pests. The medfly can adapt to a wide variety of ecological niches and infest hundreds of species of fruits and vegetables. Using a high-quality assembly of the medfly scaffolds, I aimed to identify and characterize the locations and context of repeat loci within the known medfly genome. I further identify the overrepresentation of several TE families in the medfly genome. A phylogenetic analysis of the most overrepresented TE family, *Mariner/Tc1* elements, was performed to establish TE copy number and evolutionary divergence of the elements. This chapter focuses on TE genomic composition of the medfly to set the foundation for pursuit of molecular strategies for economic, agricultural, and population control of this pest, as well as to contribute to a better understanding of general TE activity.

Section 2.2: Introduction

The Mediterranean fruit fly is a pest that presents a major threat to agriculture across the globe (De Meyer et al. 2008; Thomas et al. 2001; Thomas et al. 2001; USDA 2017; USDA 2017; USDA 2017). *Ceratitis capitata*, is native to Africa and has been recorded in over 30 African countries and spread to southern Europe, Australia, Central and

Southern America, the Middle East, islands across the Atlantic, Pacific, and Indian Oceans. The first infestation of *C. capitata* in the United States was recorded in Florida in 1929 and was only first captured in California in 1975. California, Florida, Texas, and Hawaii are all at high risk for introductions due to amenable climate and abundant agriculture. Only California and Hawaii have confirmed Mediterranean fruit fly populations at this time (USDA NASS [United States Department of Agriculture 2012; USDA 2017; Thomas et al. 2001).

C. capitata populations are generally spread unintentionally via agricultural crop infestation with the fruit fly larvae. Crop damage occurs from larval feeding. The mother pierces the skin of ripening fruit and lays her eggs in the soft skin of the fruit. The eggs then hatch inside of the fruit and develop into their larval stage. As the larvae feed on the pulp of the fruit, the fruit decays, and eventually falls of the plant. Damage induced by this fruit fly renders effected crops inedible (APHIS 2017; USDA 2017; Carey 1984; Thomas et al. 2001).

C. capitata is amongst the most diverse in host range, infesting over 350 fruit and vegetable species such as apple, avocado, bell pepper, coffee, grape, grapefruit, lemon, lime, mango, orange, pomegranate, tomato, among many other known targets (APHIS 2017; USDA 2017; Liquido & Cunningham 1997; Thomas et al. 2001; Thomas et al. 2001; Carey 1984). The destruction of crop production significantly affects crop yield, which has incredible societal and economic repercussions. In California alone, the gross products of the crop loss had a gross value of over $16.5 billion in 2011. The significant

crop destruction is the rationale by which many consider *C. capitata* to be the most significant agricultural pest in the world (APHIS 2017).

The TE content of the *C. capitata* genome was annotated in an effort to design molecular controls to contain pest (Torti et al. 2000). Both Class I and Class II elements have been observed in the medfly. Hundreds of Mariner TE family elements have been found in the medfly, as well as the *hAT*, *Tc1*, and *Gypsy* TE families to a lesser degree (Torti et al. 2000; Gomulski et al. 2004; Robertson & MacLeod 1993; Gomulski et al. 1997; Zhou & Haymer 1997).

Fluorescent in situ hybridization assays indicated a particularly large overrepresentation of *Mariner/Tc1* elements within the *C. capitata* genome (Torti et al. 2000). *Mariner/Tc1* elements have previously been classified into ten subclasses in the *Drosophila* genus: capitata, cecropia, drosophila, elegans, irritans (bytmar-like), irritans (himar-like), marmoratus, mauritiana, mellifera, and vertumnana (Wallau et al. 2014; Robertson & MacLeod 1993; Gomulski et al. 2004). However, a high quality assembly of the medfly genome and nucleotide-level resolution of TEs has not been observed in the medfly genome.

Section 2.3: Materials and Methods

In collaboration with the Baylor College of Medicine's i5k consortium, I set out to thoroughly annotate the genome of *C. capitata*. Thorough annotation of this genome will aid in developing genetic strategies that can control crop depletion due to Mediterranean

fruit fly populations. In its entirety, the consortium also contributed the genome assembly, gene annotations, phylogenomics, orthology, miRNA observation, and mechanics of several gene families in *C. capitata* (Papanicolaou et al. 2016).

This chapter is concerned with the repeat element classification within the previously established, *C. capitata* genomics scaffolds. Repeat elements contribute to the large variability within medfly populations, which likely contributes to the unique genetic plasticity of *Ceratitis capitata* (Malacrida et al. 1996). Therefore, the loci, origin, type, abundance, and distribution of repeat elements are of interest within the medfly genome.

A better understanding of the repeats within the medfly can provide information regarding the contents of the assembled genome, as well as the general activity of TEs in the medfly. This contribution can lead to the development of more effective molecular strategies to control the pest and the availability of foundational information that could lead to a fine investigation of TE mechanics in *C. capitata*.

Section 2.3.1: Annotation Pipeline

In order to thoroughly annotate repeat loci in *Ceratitis capitata*, I established a workflow to identify and characterize repeat families and individual elements within the medfly scaffolds (Figure 2.1). RepeatModeler was initially run to identify repeat elements and cluster similar repeats into profiles (Smit & Hubley 2010). RepeatModeler utilizes two previously established, complementary repeat finding programs, RECON and RepeatScout, to establish de novo identification of repeat element (Price et al. 2005; Bao

& Eddy 2002). The RECON algorithm relies upon single linkage cluster of local, pairwise and multiple sequence alignments between regions of the genome (Bao & Eddy 2002). RepeatScout also uses a method of sequential pairwise alignment, but in doing so, considers sequence on either side of an identified repeat element (Price et al. 2005). In doing so, RepeatScout is capable of clustering sequences whose repeats are of variable length within the genome. The ability to cluster sequences of varying lengths is particularly important in the context of TEs given imperfect integrations, MITEs, and single nucleotide variability. RepeatModeler then refines the consensus repeat profiles, obtained from RECON and RepeatScout, to be utilized downstream in the annotation pipeline (Smit & Hubley 2010).

Figure 2.1 | Repeat Annotation Workflow. General steps involved in identifying and characterizing the repeat elements in the genome of *Ceratitis capitata*.

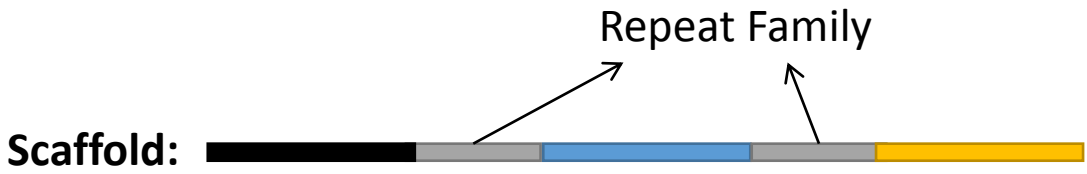


Figure 2.2 | Repeat Family Definition. A conceptual example of sequence content that would warrant repeat family identification. In this example, different colors represent different sequence clusters. The two gray clusters are converged into one representative repeat family.

Once the repeat elements have been identified, TE characterization of the defined repeat elements is attempted using an approach that is both structurally and homology-based. First, I performed a structural-based LTR search by running the LTR_STRUC program on the medfly scaffolds (McCarthy & McDonald 2003). A structurally-based approach is one that identifies motifs and features of the genome that appear to be consistent with other known TEs.

Features indicative of a LTR can be in terms of sequence length, order, and sequence specificity. These features have defined functional roles in LTR transposition. The identification of these functional domains is particularly useful in searching for intact and/or functional LTR elements. There are several features that the LTR_STRUC program takes into account as it scans the genome of interest. The LTR_STRUC program looks for LTR signatures, such as primer binding sites, poly-purine tracts, specific transposition-related genes, as well as dinucleotides indicative of the end of a LTR element (McCarthy & McDonald 2003). Sites with most, or all, of these features indicative of a LTR element are scored and characterized appropriately.

The primer-binding site (PBS) is a specific, roughly 18 nucleotide sequence that allows for binding, via base pair complementarity to the 3' end of a tRNA molecule (Hargittai et al. 2004). This binding of the tRNA at the PBS provides a primer site essential for the initiation of reverse transcription in the antisense orientation. The polypurine tract (PPT), on the other hand, is essential for reverse transcription to occur in the sense orientation (Wohrl & Moelling 1990). The reverse transcriptase cleaves an initial DNA/RNA hybrid

precisely at the 3' end of the PPT (Sarafianos et al. 2001). The mechanism by which this is accomplished is still under investigation. The remaining RNA then serves as a primer for reverse transcription in the sense orientation.

Several genes – the *gag, pol, and env* genes – are crucial for LTR transposition. The *gag* gene encodes for a polyprotein that forms the retroviral core structure. The *pol* gene serves as the reverse transcriptase required for RNA TE transposition within the retroviral core. The *env* gene encodes an envelope protein that facilitates interaction with the target cell membrane (Finnegan 1989). Confidence can be built on the TE characterization by considering the homology and orientation of these functionally critical, and therefore well-conserved retroviral genes (Figure 2.3).

Fully intact non-LTR TEs contain an untranslated region (UTR) at the 5' and 3' end of the element. Much like a typical gene, the 5'-UTR contains a promoter sequences and the 3'-UTR contains a termination sequence (Smit & Hubley 2010; Doucet et al. 2010). Within the translated region of the element, these non-LTR elements generally have two open reading frames (ORFs). In non-LTR TEs, depending on the nature of the element, the ORFs can be present or absent, or even responsible for various mechanics of transposition. For example, in Long Interspersed Elements (LINE) elements, the first ORF is encodes for a RNA binding protein, while the second ORF is responsible for the production of endonuclease and reverse transcriptase crucial for transposition (Doucet et al. 2010). In contrast, when considering Short Interspersed Elements (SINEs), reverse

transcriptase is rarely encoded in the element, although these elements can still encode for their own endonuclease to facilitate chromosomal breaks (Ohshima & Okada 2005).

Repeat family alignment to the ORFs of known TEs was used to attribute a higher degree of confidence when considering manual non-LTR TE characterization. Consistency in the orientation of the ORFs in an identified repeat family was also considered in defining non-LTR TE definition.

Similar to non-LTR elements, fully intact DNA TEs have an untranslated region on either end of the element, and ORF(s) within the translated region (Figure 2.3). The ORF(s) within the translated region of a DNA TE encode for the transposase protein that is responsible for the "cut and paste" mechanism of Class II transposition (Yuan & Wessler 2011). The degree of homology to the element and the localization of the homology (i.e. homology to the transposase or untranslated region) were prioritized in assessing DNA TE characterization.

A homology-based annotation of the RepeatModeler repeat families from the medfly scaffolds via tblastx searches using the CENSOR program against a database of known TEs within the RepBase database. (Altschul et al. 1990; Jurka et al. 1996; Smit et al. 1996; Smit et al. 1996).

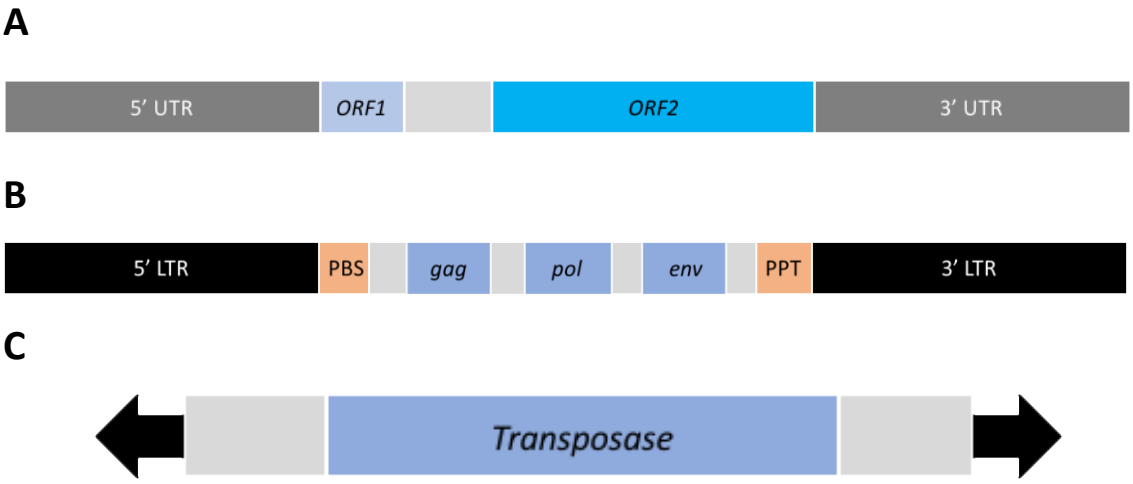


Figure 2.3 | Structural TE Characterization. Indicative structural features of (A) Non-LTR (B) LTR (C) DNA TEs.

The RepBase database is a widely utilized database in the scientific community which contains sequence, classification, and other basic information regarding previously established, eukaryotic repetitive DNA (Jurka et al. 2005). RepBase stores information regarding known simple repeats, autonomous and non-autonomous TEs. This database has been continuing to build its repository since it was established in 1992. The database contains over 40,000 consensus repeat family sequences in over 100 species of interest of animals, plants, and fungi (Jurka et al. 2005).

RepBase utilizes both automated and manual annotation in an attempt to maximize the advantages of each method, while minimizing the disadvantages (Jurka et al. 2005). Automated annotation allows for high throughput processing of given sequences, but is much more prone to error relative to manual annotation.

CENSOR is software associated with the RepBase database that masks TE sequences in order to make more meaningful transposable element comparisons (Jurka et al. 1996). Masking sequences entails hiding – or not considering – regions of a repeat sequence that could lead to an ambiguous result. Masking sequences before a homology-based sequence comparison removes simple and low complexity repeats before comparing sequences of interest.

In this homology-based annotation, only those putative TEs that had >80% of the query hit a known sequence with >30% amino acid identity were considered in this analysis (Neafsey et al. 2015). Manual annotation was then utilized to correct erroneous calls within these automated pipelines.

The RepeatMasker program then utilized the characterized repeat families – from the DNA, LTR, and Non-LTR TE discovery pipelines – to most effectively predict the origin of the repeat loci within the available *Ceratitis capitata* genome (Smit and Hubley 2010). RepeatMasker functions by effectively associating repeat families with the genome using a BLAST-based method called "cross_match." Cross_match uses a modified Smith-Waterman matrix, with significant speed improvements to allow for scaling, to associate local, unmasked regions of homology to known repeats (Gotoh 1982; Waterman et al. 1976).

Section 2.4: Results

A total of 26.7% of the available *Ceratitis capitata* genome was associated with repeat elements. Of the identified repeat loci, 73.7% of sequence was confidently associated with a known TE family. RNA TEs occupied 56.9% of known TE characterization and a total of 11.2% of the genome. Non-LTR TEs composed 75.9% of Class I TE definition, while the remaining 24.1% of known Class I TEs were identified as LTR TEs. DNA TEs comprised 43.1% of known TE characterization, and occupying 8.5% of the available genome. Simple and low complexity repeats, that could not be associated with known TEs, consisted of 26.2% of repeat element definition and occupied 7.0% of the genome (Table 2.1; Figure 2.4).

Class I					
Family	**Number of Elements**	**Sequence Occupied**	**Percent of Genome Occupied**	**Subclass**	**TOTAL**
I	2284	771565	0.16%	NonLTR	
Jockey	30693	7228917	1.49%	NonLTR	
Kiri	668	300422	0.06%	NonLTR	
L2	6324	1341355	0.28%	NonLTR	
Loner	2491	631521	0.13%	NonLTR	
R1	1072	400938	0.08%	NonLTR	
R4	191	57936	0.01%	NonLTR	
RTE	118743	27440177	5.66%	NonLTR	
Loa	2903	339440	0.07%	NonLTR	
CR1	15995	2475755	0.51%	NonLTR	**8.5%**
BEL	16865	4618600	0.95%	LTR	
Copia	1323	482145	0.10%	LTR	
Gypsy	22942	7669407	1.58%	LTR	
Pao	1256	425950	0.09%	LTR	**2.7%**
TOTAL:	**223750**	**54184128**	**11.2%**		
Class II					
Family	**Number of Elements**	**Sequence Occupied**	**Percent of Genome Occupied**		
CMC-Chapaev-3	842	125714	0.03%		

CMC-EnSpm	11077	942496	0.19%		
CMC-Transib	7603	2118330	0.44%		
hAT-Charlie	233	73205	0.02%		
hAT-hobo	258	59214	0.01%		
hAT-Tip100	127	55698	0.01%		
Helitron	641	228226	0.05%		
Kolobok-Hydra	6598	1490802	0.31%		
Maverick	185	107614	0.02%		
Merlin	2660	530238	0.11%		
MULE-MuDR	90	19496	0.00%		
P	439	84256	0.02%		
PIF-Harbinger	3683	945557	0.20%		
piggyback	350	132766	0.03%		
Polinton	446	264698	0.05%		
Sola	762	197779	0.04%		
TcMar-Tc1	141224	33911933	7.00%		
TOTAL:	**177218**	**41288022**	**8.5%**		
Other Repeats					
Family	**Number of Elements**	**Sequence Occupied**	**Percent of Genome Occupied**		
Low_complexity	77988	14798078	3.05%		

Simple_repeat	306302	19208738	3.96%		
TOTAL:	**384290**	**34006816**	**7.0%**		

Table 2.1 | Repeat Element Characterization in *Ceratitis capitata*. A description of the number of full or partial elements identified, the number of nucleotides characterized by the identified elements, and the percent of the available genome scaffold occupied by identified Non-LTR, LTR, and DNA TEs.

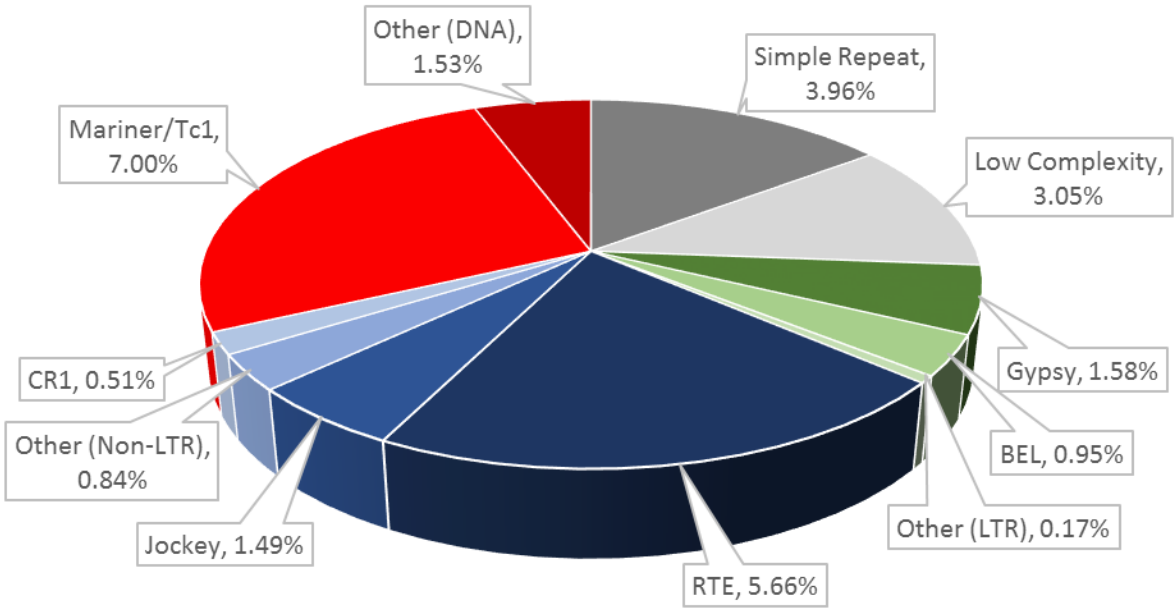


Figure 2.4 | Genomic Repeat Element Composition. Representation of TE Class and family occupancy with the 26.7% of the identified Ceratitis capitata scaffolds associated with repeat loci. The color scheme is as follows: Non-LTR (Class I): blue, LTR (Class I): green, DNA (Class II): red, and Simple/Low Complexity: gray.

In both Class I and Class II TEs, TE families were significantly overrepresented within the *Ceratitis capitata* scaffolds relative to expectation by random chance. On average, Class I TEs that were found within the *Ceratitis capitata* genome occupied 0.80%. RTE elements occupied 5.7% of the *Ceratitis capitata* genome, 50.6% of Class I TE characterization, and 70.6% of Non-LTR TE characterization. *Jockey* elements occupied 1.5% of the genome, 13.3% of Class I TE characterization, and 18.6% of Non-LTR TE characterization. *Gypsy* elements occupied 1.6% of the genome, 14.1% of Class I TE characterization, and 50.0% of LTR TE characterization (Table 2.2).

Class II TEs occupied 0.50% of the genome on average. *Mariner/Tc1* elements were the most prevalent Class II TEs by a factor of 100 in *Ceratitis capitata*. *Mariner/Tc1* elements occupied 7.0% of the genome and 82.4% of Class II TE characterization (Table 2.2).

	Genome Occupancy	TE Class	TE Class Occupancy	TE Subclass	TE Subclass Occupancy
Mariner/Tc1	**7.00%**	**Class II**	**82.14%**	**DNA**	**NA**
RTE	5.66%	Class I	50.63%	Non-LTR	70.57%
Jockey	1.49%	Class I	13.33%	Non-LTR	18.58%
Gypsy	1.58%	Class I	14.13%	LTR	50.00%

Table 2.2 | Overrepresented TE Families in *Ceratitis capitata*. A description of the top 4 overrepresented TE families, the percentage of the identified *Ceratitis capitata* scaffolds that are occupied by the TE family, the percentage of the respective TE class occupied by the TE family, and the percentage of TE subclass occupied for Non-LTR and LTR TE families.

Section 2.4.2: Evolutionary Relationship of *Mariner/Tc1* elements

Given the vast overrepresentation of *Mariner/Tc1* elements, I pursued a better understanding regarding a distinction amongst the individual *Mariner/Tc1* elements. In doing so, I hope to identify features of the particular active, or inactive, *Mariner/Tc1* elements that would contribute to a better understanding of the mechanisms of transposition in *Ceratitis capitata*.

The sequences of *Mariner/Tc1* elements – that were identified in our pipeline of TE characterization within the genomic scaffolds available for *Ceratitis capitata* – were extracted and utilized to infer an evolutionary relationship amongst the elements. A multiple sequence alignment of *Mariner/Tc1* elements was performed using MUSCLE. MUSCLE quickly determines the optimal multiple sequence alignment by calculating divergence profiles, using *k*-mer distance and progressive alignments, to reassess the highest scoring alignment available (Edgar 2004).

Next, the multiple sequence alignment was trimmed using trimAl (Capella-Gutiérrez et al. 2009). Alignment trimming is critical in phylogenetic analyses because the confidence associated with tree estimation is only as strong as the quality of sequence alignment. Poorly aligned regions within a sequence alignment will translate into a phylogenetic tree to which little confidence can be associated. On the other hand, relatively well-conserved regions of a protein tend to show little divergence between species. By considering well-conserved residues that appear to have diverged in a step-wise manner, more straight-forward and reproducible observation of divergence between species can be attained. trimAl uses scores, associated with the degree of similarity, gap, and identity at each residue, to remove residues in the alignment that would not contribute useful information to the distance calculations.

Using the trimmed and aligned residues of *Mariner/Tc1* elements, I then calculated a phylogenetic tree to represent the evolutionary divergence of the identified elements. A guide tree was produced using the neighbor-joining method of divergence clustering. The Neighbor-joining algorithm is a bottom-up approach to clustering (Saitou & Nei

1987). The neighbor-joining algorithm begins by identifying the two *Mariner/Tc1* sequences that are most similar. The algorithm then considers Euclidean distance, which indicates the number of changes that would have to be made to make the sequences identical. A General Time Reversible (GTR) model was used to quantify divergence considering the time-reversible, independent, and finite divergent states (Tavaré 1986; Huelsenbeck et al. 2001). An edge length is created corresponding to the number of steps that was required and a common node is created representing the profile from which the two most similar sequences diverged. The algorithm then continues to find the two next most similar sequences, while replacing the two previously converged sequence with their common node (Figure 2.5). This clustering method is appropriate for our analysis of the evolutionary divergence of *Mariner/Tc1* elements given that I was particularly interested differentiating the leaves of the tree: the Mariner element sequences.

The guide tree was used as a reference to optimize the parameters of the model. The guide tree was optimized for edge length (i.e. maximum parsimony), base frequencies, and variance distribution, as well as using a tree optimization algorithm, nearest neighbor interchange (NNI), to return the best tree (Li et al. 1996). NNI exchanges the

connectivity of sequences deriving from a particular node with those diverging from a different node within the tree (Figure 2.6). The likelihood of each tree is then recalculated to maintain the most probable representation of divergence.

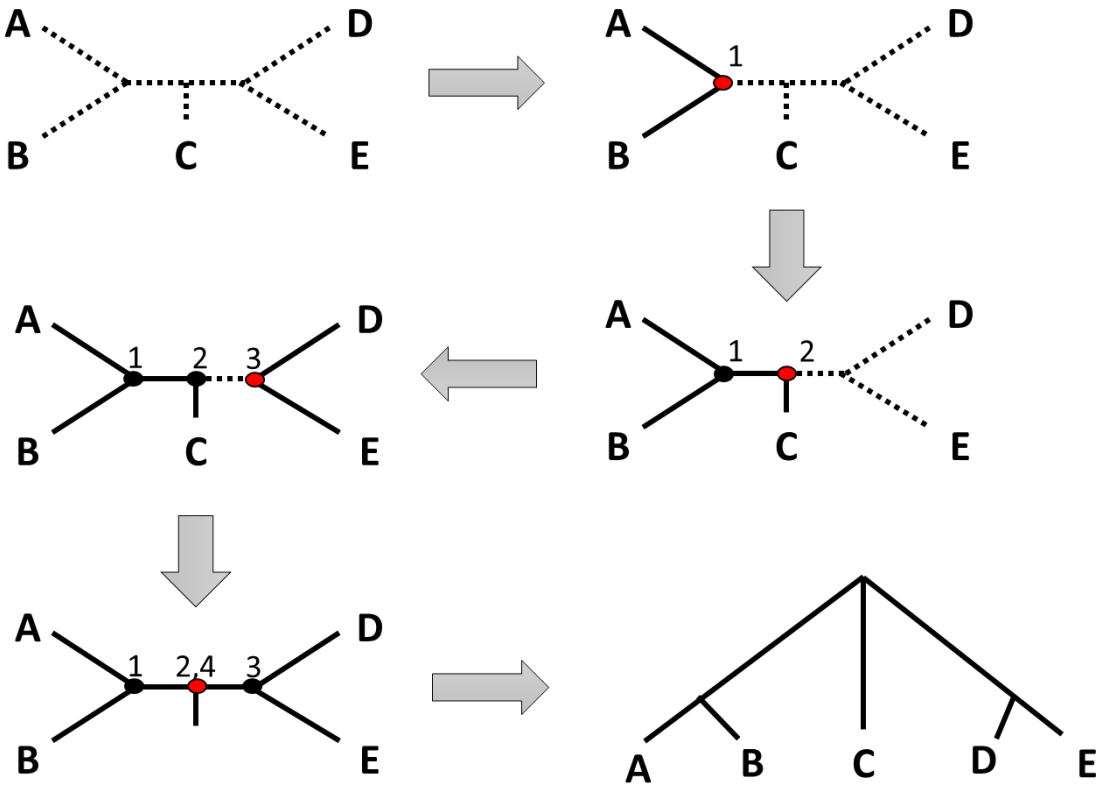


Figure 2.5 | Neighbor-joining Algorithm. A-D represent sequences from a unique *Mariner/Tc1* repeat family. A red circle represents a new profile created for the two most similar sequences in each step.

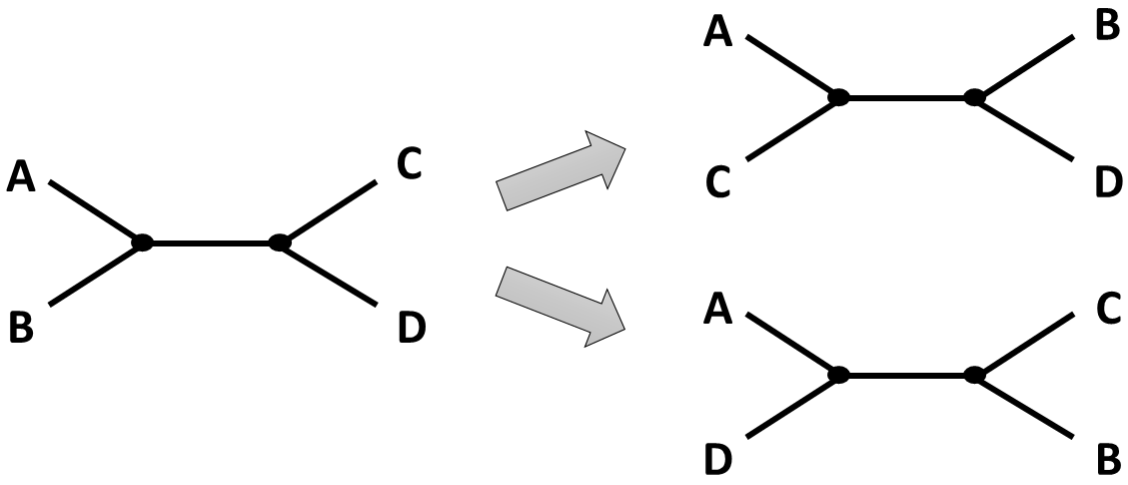


Figure 2.6 | Nearest Neighbor Interchange Algorithm. The tree on the left is the current tree. The trees on the right are transformations of the current tree. The best tree is maintained and this process will continue until all nodes have been optimized.

Assessment of node reproducibility within the tree was performed using 1000 tree calculations with bootstrap resampling. In the context of this divergence calculation, bootstrapping resampling refers to the reconstruction of biologically insignificant sequences, using random resampling of divergent residues within the multiple sequence alignment of *Mariner* nucleotides with replacement, to assess the statistical reproducibility of each node with the tree (Hall & Martin 1988; Schliep 2011; Hall & Martin 1988). Percent reproducibility of each node is indicated to the left of the respective node. Critical nodes in differentiating Mariner subclass identification demonstrated a relatively high degree of reproducibility (Figure 2.7).

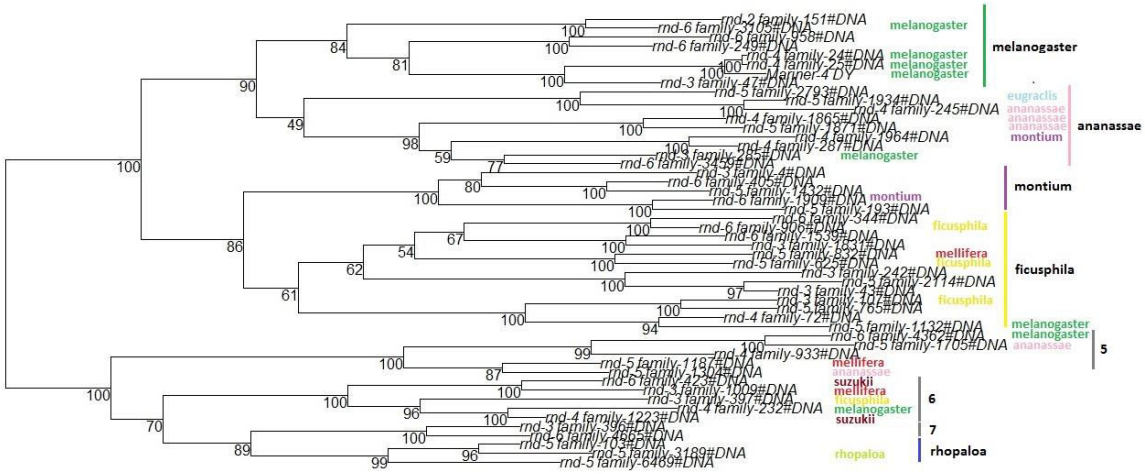


Figure 2.7 | Phylogenetic Tree of *Mariner/Tc1* elements in *Ceratitis capitata*. A guide tree was created using the neighbor-joining (bottom-up) clustering method with a GTR substitution model. The parameters of the model were optimized to maintain the most probable tree. The percentage of reproducibility is represented to the left of each node for 1000 bootstrap replicates.

Section 2.4.3: Subclass Characterization

Using the phylogenetic tree of evolutionary divergence of *Mariner/Tc1* elements, I aimed to identify the most similar subclasses amongst the identified elements. K-means clustering was utilized, considering a value of k between 2 and 15, to estimate the number of cluster among the identified *Mariner/Tc1* elements (Kanungo et al. 2002). Bayesian inference, as well as the bootstrap reproducibility of downstream nodes, was considered in identifying the eight subclasses of *Mariner/Tc1* elements in *Ceratitis capitata.*

Five of the ten previously established *Mariner/Tc1* subclasses were also represented to be represented in *Ceratitis capitata* (Robertson & MacLeod 1993; Gomulski et al. 2004). I identified eight repeat families in the melanogaster *Mariner/Tc1* subclass, nine repeat families in the ananassae subclass, six repeat families in the montium subclass, thirteen in ficusphila subclass, six repeat families in the unidentified subclass 5, six repeat families in the unidentified subclass 6, two repeat families in the unidentified subclass 7, and three repeat families in the rhopaloa *Mariner/Tc1* subclass.

I identified three subclasses (subclasses 5-7) that were unable to be resolved into the previously established TE subgroups in *Ceratitis capitata*. The TEs within subclasses 5-7 have either diverged beyond recognition from their ancestral elements, or they represent a novel subclass of previously unknown origin.

Section 2.5: Future Directions

Using the repeat element loci and information regarding origin presented in this chapter, I suggest that an individual TE investigation be performed. By performing an investigation of individual TEs, one can begin to ascertain the degree of active elements and MITEs within the *Ceratitis capitata* genome. Potentially active elements can be inferred based upon the maintenance of fully intact ORFs of the crucial elements necessary for transposition of a particular element (Figure 2.3).

An informative result could also derive from a more in-depth observation of MITEs within the *Ceratitis capitata* genome. A revised workflow, with a focus on the identification of MITEs could provide a more accurate account of repeat TE copy number, and therefore degree of transposition. Although the workflow described in this research is capable of detecting MITEs, software exists that was designed specifically for the detection of MITEs, even when a small amount of sequence remains between the inverted repeats of the TE. A de novo detection of MITEs could be performed by running previously established programs, such as MITE Digger and MITE Hunter, to better understand the content and quantity of MITEs in *Ceratitis capitata* (Yang 2013; Han & Wessler 2010).

A further exploration of the identified *Mariner/Tc1* elements could contribute to a better understanding of TE activity in *Ceratitis capitata*. I have identified 8 subclasses of *Mariner/Tc1* elements in *Ceratitis capitata*: five of which are similar to previously identified *Mariner/Tc1* subclasses and three subclasses that appear to be distinct (Figure

2.7). I propose that an evolutionary analysis be performed to further understand the nature and origin of the subclasses 5-7, which did not demonstrate significant similarity relative to the ten previously established *Mariner/Tc1* subclasses.

Using similar methodology from the workflow that was utilized to identify these repeat elements, subclasses 5-7 could be compared to profiles of the *Mariner/Tc1* subclasses in *Ceratitis capitata* to identify to which *Mariner/Tc1* subclass it is most closely affiliated. When subclasses 5-7 are associated with their most similar, known *Mariner/Tc1* subclass, one can begin to observe the regions of particular divergence. One can infer loss or gain of functional attributes of *Mariner/Tc1* subclasses by identifying the regions of divergence, the degree of divergence by region, and copy number within the genome. This analysis would provide additional information regarding whether these *Mariner/Tc1* subclasses are indeed distinct, novel subclasses or if they represent a subset of elements within a previously described subclass, as seen within the irritans *Mariner/Tc1* subclass. *Mariner/Tc1* ORFs and terminal ends of the elements would be of particular interest due to their capability to encode for transposase and facilitate the movement of the elements, respectively.

Further, the mechanics of *Mariner/Tc1* element movement can be examined on a fine scale. Given that *Mariner/Tc1* subclasses have been identified in *Ceratitis capitata* in this work, the number of elements associated with each subclass can be observed. Subclasses of interest would be those with a particularly high or low copy number, since their evolutionary distinction from other *Mariner/Tc1* subclasses could lead to the increased or decreased transposition activity. It is important to note that in this analysis,

the date of the *Mariner/Tc1* derivation with the *Ceratitis capitata* genome would have to be considered to produce a meaningful result. Information regarding the timeline of *Mariner/Tc1*subclass derivation in the *Ceratitis capitata* genome could then be considered to eliminate the possibility of the identification of a over- or underrepresentation of a particular *Mariner/Tc1*subclass strictly due to a the time it has been available for, and therefore the increased probability of, a transposition event. A large number of transposition events strictly due to the amount of time the *Mariner/Tc1* subclass has been available for transposition in the genome would likely provide no insight into the TE mechanics of that subclass and would therefore not be of interest in this context.

Section 2.6: References

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

APHIS, U., 2017. Plant Pests and Diseases Program: Insects - Fruit Flies.

Bao, Z. & Eddy, S.R., 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research*, 12(8), pp.1269–1276.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15), pp.1972–3.

Carey, J., 1984. Host-specific demographic studies of the Mediterranean fruit fly Ceratitis capitata. *Ecological Entomology*, 9(3), pp.261–270.

Doucet, A.J. et al., 2010. Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet*, 6(10), p.e1001150.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), pp.1792–1797.

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. *Trends in genetics*, 5, pp.103–107.

Gomulski, L. et al., 1997. Ccmar1, a full-length mariner element from the Mediterranean fruit fly, Ceratitis capitata. *Insect molecular biology*, 6(3), pp.241–253.

Gomulski, L.M. et al., 2004. Medfly transposable elements: diversity, evolution, genomic impact and possible applications. *Insect biochemistry and molecular biology*, 34(2), pp.139–148.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. *Journal of molecular biology*, 162(3), pp.705–708.

Hall, P. & Martin, M.A., 1988. On bootstrap resampling and iteration. *Biometrika*, 75(4), pp.661–671.

Han, Y. & Wessler, S.R., 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research*, 38(22), pp.e199–e199.

Hargittai, M.R. et al., 2004. Mechanistic insights into the kinetics of HIV-1 nucleocapsid protein-facilitated tRNA annealing to the primer binding site. *Journal of molecular biology*, 337(4), pp.951–968.

Huelsenbeck, J.P. et al., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550), pp.2310–2314.

Jurka, J. et al., 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Computers \& chemistry*, 20(1), pp.119–121.

Jurka, J. et al., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110(1-4), pp.462–467.

Kanungo, T. et al., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7), pp.881–892.

Li, M., Tromp, J. & Zhang, L., 1996. On the nearest neighbour interchange distance between evolutionary trees. *Journal of theoretical biology*, 182(4), pp.463–7.

Liquido, N.J.P.G.B. & Cunningham, R.T., 1997. Medhost: an encyclopedic bibliography of the host plants of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann) (electronic database/program). *USDA ARS-144*.

Malacrida, A.R. et al., 1996. Allozyme divergence and phylogenetic relationships among species of tephritid flies. *Heredity*, 76(6), pp.592–602.

McCarthy, E.M. & McDonald, J.F., 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, 19(3), pp.362–367.

De Meyer, M. et al., 2008. Ecological niches and potential geographical distributions of Mediterranean fruit fly (Ceratitis capitata) and Natal fruit fly (Ceratitis rosa). *Journal of Biogeography*, 35(2), pp.270–281.

Neafsey, D.E. et al., 2015. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, 347(6217), p.1258522.

Ohshima, K. & Okada, N., 2005. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenetic and genome research*, 110(1-4), pp.475–490.

Papanicolaou, A. et al., 2016. The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biology*, 17(1), p.192.

Price, A.L., Jones, N.C. & Pevzner, P.A., 2005. De novo identification of repeat families in large genomes. *Bioinformatics*, 21(suppl 1), pp.i351–i358.

Robertson, H. & MacLeod, E., 1993. Five major subfamilies of mariner transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. *Insect molecular biology*, 2(3), pp.125–139.

Saitou, N. & Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), pp.406–25.

Sarafianos, S.G. et al., 2001. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA: DNA. *The EMBO journal*, 20(6), pp.1449–1461.

Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)*, 27(4), pp.592–3.

Smit, A. & Hubley, R., 2010. RepeatModeler Open-1.0. *Repeat Masker Website*.

Smit, A.F., Hubley, R. & Green, P., 1996. RepeatMasker Open-3.0.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17, pp.57–86.

Thomas, M. et al., 2001. Mediterranean Fruit Fly, Ceratitis capitata (Wiedemann)(Insecta: Diptera: Tephritidae). *Fla Depart Agr Cons Serv, DPI. Entomol Cir*.

Torti, C. et al., 2000. Evolution of different subfamilies of mariner elements within the medfly genome inferred from abundance and chromosomal distribution. *Chromosoma*, 108(8), pp.523–532.

USDA, 2017. Hungry Pests: The Threat - Mediterranean Fruit Fly.

USDA NASS [United States Department of Agriculture, N.A.S.S., 2012. California County Agricultural Commissioners' Reports, 2011. United States Department of Agriculture. National Agricultural Statistics Service. California Field Office, Sacramento. .

Wallau, G.L. et al., 2014. Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. *BMC genomics*, 15(1), p.727.

Waterman, M.S., Smith, T.F. & Beyer, W.A., 1976. Some biological sequence metrics. *Advances in Mathematics*, 20(3), pp.367–387.

Wohrl, B.M. & Moelling, K., 1990. Interaction of HIV-1 ribonuclease H with polypurine tract containing RNA-DNA hybrids. *Biochemistry*, 29(44), pp.10141–10147.

Yang, G., 2013. MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC bioinformatics*, 14(1), p.186.

Yuan, Y.-W. & Wessler, S.R., 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19), pp.7884–9.

Zhou, Q. & Haymer, D.S., 1997. Molecular structure of yoyo, a gypsy-like retrotransposon from the Mediterranean fruit fly, Ceratitis capitata. *Genetica*, 101(3), pp.167–178.

Chapter 3: piClusterBusteR - Software for Automated Classification and Characterization of piRNA Cluster Loci

Section 3.1: Abstract

Background

Piwi-interacting RNAs (piRNAs) are sRNAs that have a distinct biogenesis and molecular function from siRNAs and miRNAs. The piRNA pathway is well-conserved and shown to play an important role in the regulatory capacity of germline cells in Metazoans. Significant subsets of piRNAs are generated from discrete genomic loci referred to as piRNA clusters. Given that the contents of piRNA clusters dictate the target specificity of primary piRNAs, and therefore the generation of secondary piRNAs, they are of great significance when considering transcriptional and post-transcriptional regulation on a genomic scale. A quantitative comparison of top piRNA cluster composition can provide further insight into piRNA cluster biogenesis and function.

Results

I have developed software for general use, piClusterBusteR, which performs nested annotation of piRNA cluster contents to ensure high-quality characterization, provides a quantitative representation of piRNA cluster composition by feature, and makes available annotated and unannotated piRNA cluster sequences that can be utilized for downstream analysis. The data necessary to run piClusterBusteR and the skills necessary to execute this software on any species of interest are not overly burdensome for biological researchers.

piClusterBusteR has been utilized to compare the composition of top piRNA generating loci amongst 13 Metazoan species. Characterization and quantification of cluster composition allows for comparison within piRNA clusters of the same species and between piRNA clusters of different species.

Conclusions

I have developed a tool that accurately, automatically, and efficiently describes the contents of piRNA clusters in any biological system that utilizes the piRNA pathway. The results from piClusterBusteR have provided an in-depth description and comparison of the architecture of top piRNA clusters within and between 13 species, as well as a description of annotated and unannotated sequences from top piRNA cluster loci in these Metazoans.

piClusterBusteR is available for download on GitHub:
https://github.com/pschreiner/piClusterBuster

Section 3.2: Introduction

P-element induced wimpy testis (PIWI) proteins and the utilization of the PIWI-interacting RNA (piRNA) pathway has been conserved in a diverse range of Metazoans, including sponges, roundworms, fruit flies, and humans (Grimson et al. 2008). The importance of the role of piRNAs in fertility was demonstrated in Metazoans by the observation of crosses after exposure of *Drosophila melanogaster* to the *P* transposable

element (TE) (Kidwell 1983). When female flies containing the *P* element were mated with males that were lacking it, the progeny were fertile. However, when males containing the *P* element were mated with females lacking it, hybrid dysgenesis occurred, leading to sterility of the progeny (Kidwell et al. 1977). It was later discovered that exposure to the *P* element prompted maternal deposition of piRNAs to effectively silence the *P* element and allow for fertile progeny (Brennecke et al. 2008). Perturbations to the piRNA pathway have also demonstrated gametogenic defects in *M. musculus* and *D. rerio* (Kuramochi-Miyagawa et al. 2004; Houwing et al. 2007). Although, piRNAs are notably absent in plant and fungal species (Grimson et al. 2008).

piRNAs are a subset of sRNAs between 24-31 nucleotides in length, although the range of the piRNA size distribution varies across species (Aravin et al. 2007; Arensburger et al. 2011). piRNAs are generated via a primary or secondary mechanism of biogenesis (Aravin et al. 2007).

Primary piRNAs derive from discrete genomic loci that are referred to as piRNA clusters. These loci can vastly range in size from under one thousand nucleotides to over one hundred thousand nucleotides in length. Transcription of piRNA clusters can occur in several distinct manners depending on the nature of the piRNA clusters (Brennecke et al. 2007).

piRNA clusters are characterized as unidirectional, bidirectional, or dual-stranded based on the direction transcription at the locus (Brennecke et al. 2007; Malone et al. 2009). The transcripts generated from piRNA clusters serve as precursor molecules for piRNAs,

undergoing dicer-independent slicing and modification at their 3' end (Aravin et al. 2007; Saito et al. 2007). The processing of primary piRNAs has been shown to demonstrate a bias of U at position 1 of the piRNAs (Brennecke et al. 2007). When post-transcriptional processing of piRNAs is complete, the molecules are referred to as mature piRNAs.

Mature piRNAs then associate with an Argonaute family, PIWI protein to form a RNA-induced silencing complex (RISC). The RISC complex has the capability to facilitate both transcriptional and post-transcriptional regulation. RISC-mediated transcriptional regulation occurs via piRNA association and guiding of a PIWI protein which facilitates epigenetic modification in *Drosophila* (Yin & Lin 2007). RISC-mediated post-transcriptional regulation occurs via piRNA association with PIWI or AGO3, which leads to piRNA-directed cleavage of mRNAs in *Drosophila* (Aravin et al. 2007). piRNAs have also been implicated in post-transcriptional silencing of mRNAs via poly(A) deadenylation (Rouget et al. 2010; Barckmann et al. 2015).

The number of PIWI proteins can differ in Metazoan species. While three PIWI proteins have been identified in *D. melanogaster, H. sapiens,* and *M. musculus,* as few as two PIWI proteins have been identified in *D. rerio* and as many as seven PIWI proteins have been identified in *Ae. aegypti* (Aravin et al. 2007; Keam et al. 2014; Kuramochi-Miyagawa et al. 2001; Houwing et al. 2007; Vodovar et al. 2012). It has not yet been determined whether the variation in the number of PIWI proteins between these species is a result of redundant, compensatory, or additional functionality.

Secondary piRNAs are generated by the slicing mechanism of RISC regulation, resulting in what is referred to as the amplification loop, or ping-pong pathway (Brennecke et al. 2007). The amplification loop functions by primary piRNA targeting of mRNA via sequence complementarity, followed by PIWI-mediated slicing of the target mRNA. The remaining fragment of the mRNA can be processed into a secondary piRNA. A secondary, mature piRNA can then associate with AGO3, and slice other mRNA targets via sequence complementarity in *Drosophila*. The overlap of complementarity between the piRNAs and their mRNA targets is generally ten base pairs in the opposite orientation, leaving an A10 bias in secondary piRNAs (Brennecke et al. 2007). piRNAs have been known to target TEs, genic mRNAs, viral mRNAs, and even rRNA molecules (Brennecke et al. 2007; Aravin et al. 2007; Yin & Lin 2007; Aravin et al. 2008; Garc'\ia-López et al. 2014).

Given that the contents of piRNA clusters dictate target specificity, finding the origin of these sequences is of great importance in understanding the biogenesis and function of piRNA clusters. I have developed software, piClusterBusteR, to be capable of automatically, consistently, and efficiently detecting top piRNA cluster loci and thoroughly describing the contents of those loci on a large scale. The capability that piClusterBusteR has to quantify piRNA cluster composition and describe annotated, as well as unannotated sequences in diverse Metazoan species allows for meaningful comparisons that can aid in facilitating a better understanding of top piRNA cluster

biogenesis and function across species. Exploring piRNA cluster composition on a large scale can provide insight into conserved piRNA cluster architecture that dictates piRNA cluster biogenesis and function.

Section 3.3: Materials and Methods

piClusterBusteR is a series of integrated R and bash scripts that interact along with other standalone bioinformatics programs to perform piRNA cluster characterization and annotation. The tool supports a variety of user input data, customization of the analyses and computational resources to be used in executing the program. piClusterBusteR is intended to be executed in a Unix environment and has a series of required software dependencies (Table 2).

Flag	Default	Description
Data Input		
-fa <file>		Indicates a FASTA input file containing piRNA cluster sequences of interest
-fq <file>		Indicates a FASTQ input file containing quality trimmed sRNAs
-bed <file>		Indicates a BED input file containing the location of the piRNA clusters of interest
-gid	"Genome"	Name of the piClusterBusteR Run
Databases (provide in FASTA format)		
-x <file>		Reference Genome
-gndb <file>		Organism-specific Gene Set
-tedb <file>		Transposable Element (TE) Set
Additional Analysis		
-n	5	Indicates the number of piRNA clusters to be analyzed
-ncbidb <file>	None	NCBI Nucleotide Database
--verbose	FALSE	Retain intermediate results
--go	FALSE	Perform a Gene Ontology enrichment analysis on sequence of genic origin
--all-srna	FALSE	Observe all sRNA, not just piRNAs
Performance Enhancement		
--qsub	FALSE	Submit jobs via Torque/Maui Resource Allocation
--srun	FALSE	Submit jobs via Slurm Resource Allocation

Flag	Default	Description
-p	1	Number of processors to utilize

Table 3.1 | Program Parameters. A list of the options, corresponding runtime flags, and default values available for use in piClusterBusteR. A flag is an indicator to specify the type of input information to the application. A blank in the "Default" column constitutes a required parameter.

Database	Website	Reference
NCBI (nt)	http://www.ncbi.nlm.nih.gov/	(Altschul et al. 1990)
RepBase	http://www.girinst.org/repbase/	(Jurka et al. 2005)

Standalone Programs	Website	Reference
BLAST+ (v2.2.30+)	http://blast.ncbi.nlm.nih.gov/	(Sayers et al. 2011)
CENSOR	http://www.girinst.org/censor/	(Jurka et al. 1996)
proTRAC	http://www.smallrnagroup.uni-mainz.de/	(Rosenkranz & Zischler 2012)
RepeatMasker	http://www.repeatmasker.org/	(Smit et al. 1996)

R Packages	Reference
Biostrings	(Pages et al. 2009)
doMC	(Analytics 2014)
GenomicRanges	(Lawrence et al. 2013)
gProfileR	(Reimand et al. 2007)
Plyr	(Wickham 2009)
Qcc	(Scrucca 2004)
Seqinr	(Charif & Lobry 2007)
systemPipeR	(Girke 2014)

Table 3.2 | List of Software and Databases Utilized in piClusterBusteR
Depending on the user-specified analyses to be performed, piClusterBusteR may require these **(A)** standalone software and **(B)** R libraries.

piClusterBusteR only requires four input parameters from the user on the command line: (1) input data, (2) a reference genome, (3) a species-specific gene set, and (4) a set of known TEs. Additional options are available to increase the efficiency of the software and to customize the program output.

piClusterBusteR allows for data input in the form of sRNA reads, piRNA cluster sequences, or piRNA cluster chromosomal loci. When sRNA reads are provided as the data input, piClusterBusteR must perform additional steps in order to assign piRNA cluster loci. First, all of the reads are filtered in order to analyze only those that are 24 nucleotides in length or greater. The piRNAs from the filtered FASTQ file are then mapped to the user-provided reference genome using proTRAC's sRNA mapping tool. The piClusterBusteR-generated map file is then utilized to define the top piRNA cluster loci using proTRAC (Rosenkranz & Zischler 2012).

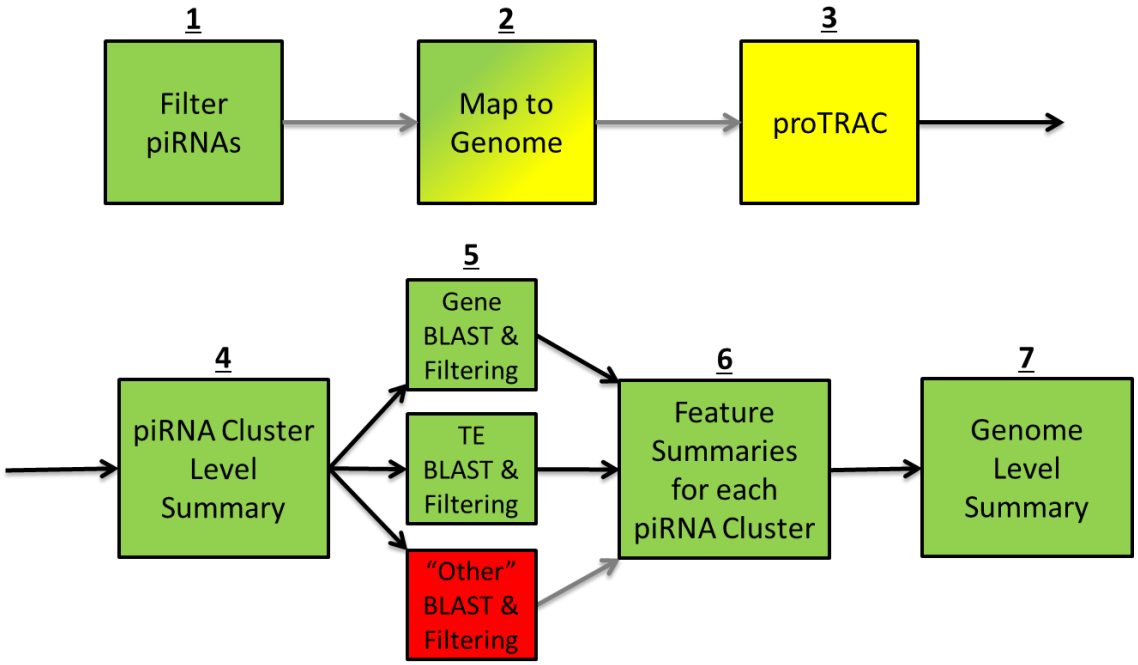


Figure 3.1 | Algorithm Overview. A workflow of the steps taken by piClusterBusteR to annotate and characterize piRNA clusters. The step number is indicated above the boxes that describe the analysis in each step. The relative time requirement of each step is indicated by green, yellow, and red from fastest to slowest. The gray arrows indicate that the previous step may be skipped if it is unnecessary.

proTRAC is an standalone tool designed for the definition of piRNA clusters (Rosenkranz & Zischler 2012). proTRAC considers features of small RNA sequence reads features that are indicative of piRNAs such as read length and a U1 or A10 bias. proTRAC uses a density-based approach to identify genomic regions that have piRNA accumulation, as defined by a significant deviation from a hypothetical uniform distribution, which then defines the degree of confidence of the piRNA cluster call. proTRAC has demonstrated efficacy in piRNA cluster definition relative to previously established methods of piRNA cluster detection (Rosenkranz & Zischler 2012). The

proTRAC output is then processed to identify the top piRNA cluster loci, as defined by the number of normalized reads per piRNA cluster, and converted to a BED file of piRNA cluster loci. The BED file is then utilized to analyze the contents of the individual top piRNA cluster loci.

In the individual piRNA cluster level analysis, piClusterBusteR performs a detailed characterization and quantification regarding the contents of each individual piRNA cluster. The user has the option to analyze piRNA clusters sequentially (default), or in parallel for each piRNA cluster of interest.

piClusterBusteR first extracts the sequence using the chromosomal coordinates and reference genome that was provided by the user. piClusterBusteR then attempts to identify the origin of the sequences within the piRNA clusters of interest.

In order to best infer the origin of the sequences within a given piRNA cluster, piClusterBusteR utilizes what I will refer to as nested annotation using RepeatMasker, CENSOR, and BLAST (Smit et al. 1996; Jurka et al. 1996; Altschul et al. 1990). Nested annotation allows for sequential and non-redundant definition of known sequences with the piRNA cluster sequences under observation. RepeatMasker is run initially on the piRNA cluster of interest using the TE database and organism-specific gene set provided by the user (Smit et al. 1996). TE and organism-specific data sets were extracted from RepBase and NCBI non-redundant nucleotide databases, respectively (Jurka et al. 2005; Sayers et al. 2011). Any of the unannotated sequence remaining in the piRNA cluster of interest is extracted and subjected to TE and genic analysis via CENSOR (Jurka et al.

1996). Finally, the remainder of the unannotated piRNA cluster sequence is subjected to a blastn search, with a word size of 7 and maximum E-value of 1e-3, against the NCBI nucleotide database (Jurka et al. 1996; Altschul et al. 1990). Any of the hits returned in the BLAST of sequences within the NCBI non-redundant (nt) database are classified as "Other," in comparison to sequence originating from known TEs or genes (Figure 3.2). Regions of the piRNA cluster loci that have not been defined with a known sequence origin are then extracted and printed reported. In doing so, piRNA cluster sequence of unknown origin can be easily accessed for downstream analysis.

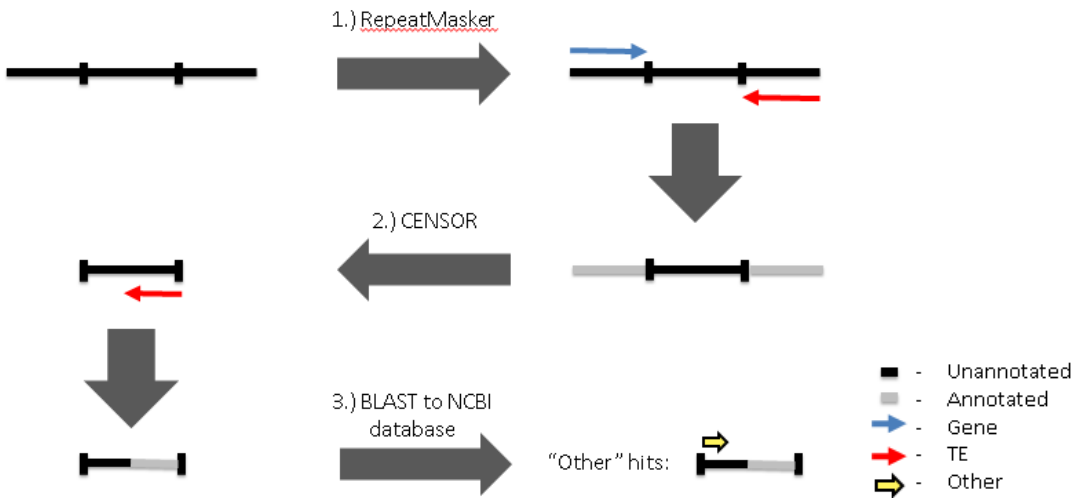


Figure 3.2 | Nested Annotation. Workflow regarding the characterization of unannotated loci are using RepeatMasker, CENSOR, and BLAST (Smit et al. 1996; Jurka et al. 1996; Altschul et al. 1990). Sequence characterization in the former steps excludes sequences from being passed to the latter.

To ensure that there is no redundancy in the sequence characterization, feature filtering is performed to only retain the best available annotation for a given piRNA cluster

sequence. The best available annotation is defined as the hit with the longest available alignment length and highest similarity percentage to a known feature.

Non-redundant TE, genic, and "other" annotation is then summarized and plotted. A directory containing all of the intermediate annotation files and summary files is output in an individual directory to represent the analysis of each individual piRNA cluster. The results of the piRNA cluster level analyses of each piRNA cluster are stored so that they can be used in the genome level analysis.

In the genome level analysis, annotation is graphically compared between individual piRNA clusters. The piRNA clusters are compared in terms of their length, contents, degree of strand specificity, and percent genome occupancy. Top piRNA cluster loci can then be compared between piRNA clusters within the same species and between species on a genomic level (Figure 3.1).

The main directory represents the outcome of the genome-level analysis. Four output files are generated in the genome-level analysis: (1) a BED file containing the piRNA cluster coordinates, (2) an aggregate file describing the total occupancy of piRNA clusters relative to the size of the organism's genome, as well as the final data necessary to make the genome summary plots in a (3) graphical and (4) text format (Quinlan & Hall 2010). The genome-level graphical output contains a comparison of piRNA cluster size, piRNAs associated with each piRNA cluster, feature composition, and strandedness of feature calls, followed by the average feature content composition across all piRNA cluster loci analyzed (Figure 3.3).

The genome level analysis also provides an individual directory for each piRNA cluster of interest in the order specified within the BED file of piRNA cluster loci. Within each piRNA cluster directory resides intermediate data files and summary files that are necessary to produce the piRNA cluster-level graphical output. The intermediate data files that were used in the data collection are available in the respective program output format defaults for each utilized tool (Table 1). The unfiltered BLAST output for each piRNA cluster, however, can often be large in size and is therefore removed by default. The piRNA cluster-level summary is also available in a text and graphical output.

The piRNA cluster-level graphical output contains a representation of the number of each feature that was characterized within the piRNA cluster, the nucleotide occupancy of each feature called, the nucleotide occupancy of all feature calls in both orientations, a representation of the prominent TE superfamilies within the piRNA cluster, the prominent specific TEs called within the piRNA cluster, and optionally, the most significant GO terms associated with genic hits within the piRNA cluster, a GO enrichment analysis of genic hits within the piRNA cluster, and stranded sRNA coverage plot with annotated features (Figure 3.4).

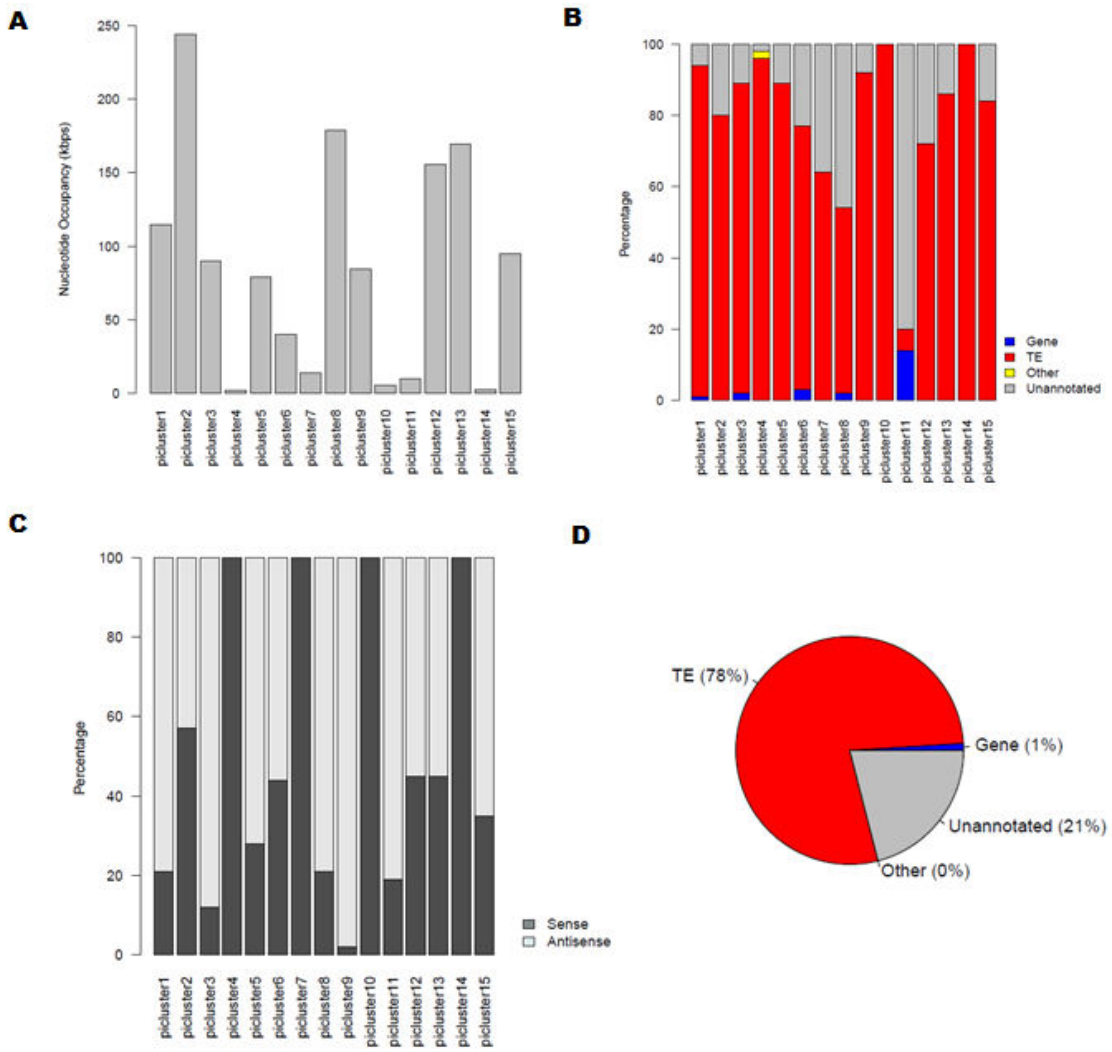
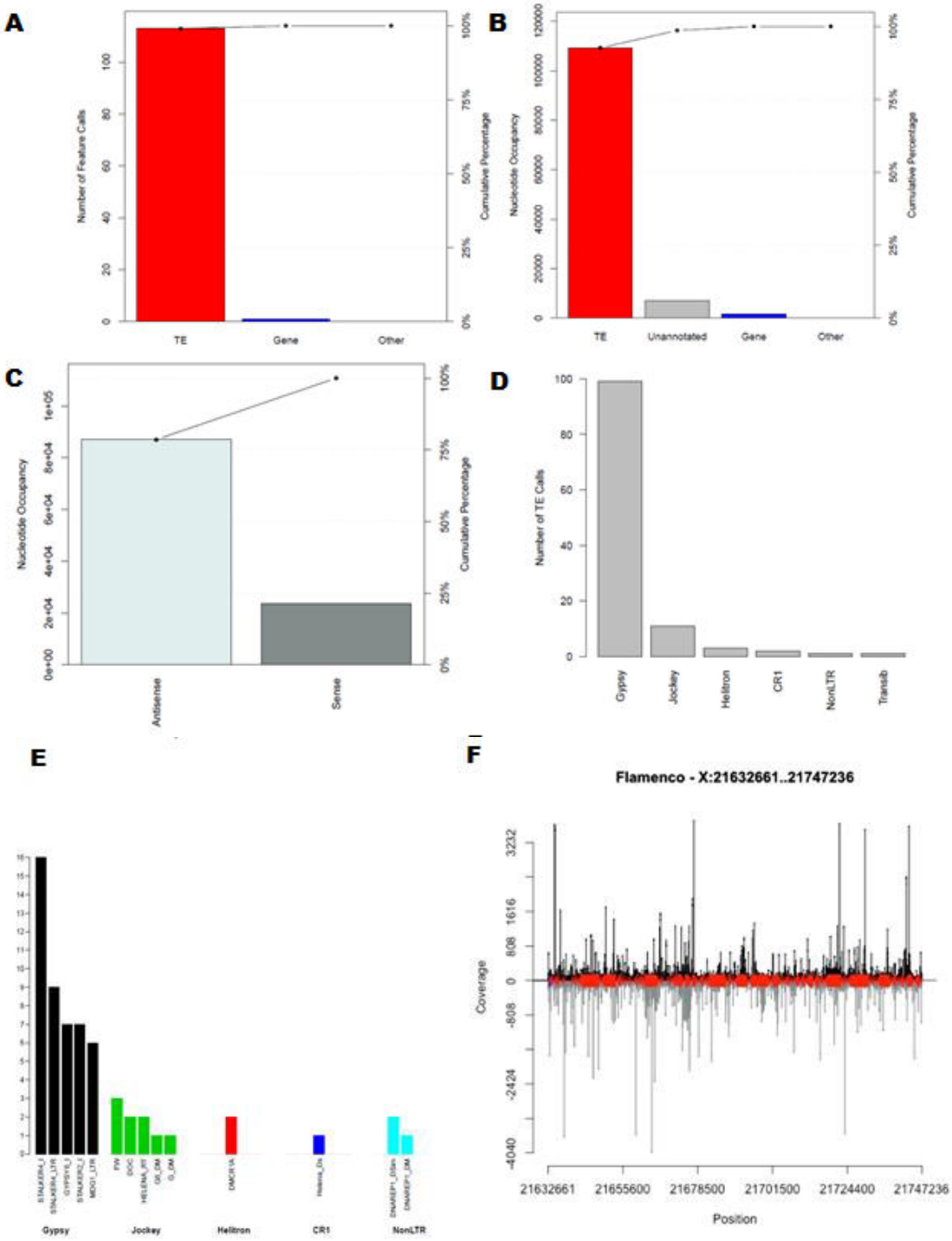


Figure 3.3 | Genome-Level Analysis of Top 15 piRNA Cluster Contents in *Drosophila melanogaster.* **(A)** Comparison of piRNA clusters by the total number of nucleotides occupied **(B)** Relative nucleotide occupancy occupied by each feature **(C)** Stranded nucleotide occupancy of feature calls. Unannotated sequences are not considered in this representation **(D)** Average nucleotide occupancy occupied by each feature across the top 15 *D. melanogaster* piRNA clusters previously identified (Brennecke et al. 2007). The *flamenco* and *42AB* loci are represented as piRNA clusters 1 and 2, respectively (Aravin et al. 2007; Brennecke et al. 2007).


A
Number of Feature Calls
Cumulative Percentage
TE
Gene
Other
B
Nucleotide Occupancy
Cumulative Percentage
TE
Unannotated
Gene
Other
C
Nucleotide Occupancy
Cumulative Percentage
Antisense
Sense
D
Number of TE Calls
Gypsy
Jockey
Helitron
CR1
NonLTR
Transib
E
Gypsy
Jockey
Helitron
CR1
NonLTR
F
Flamenco - X:21632661..21747236
Coverage
Position
21632661
21655600
21678500
21701500
21724400
21747236

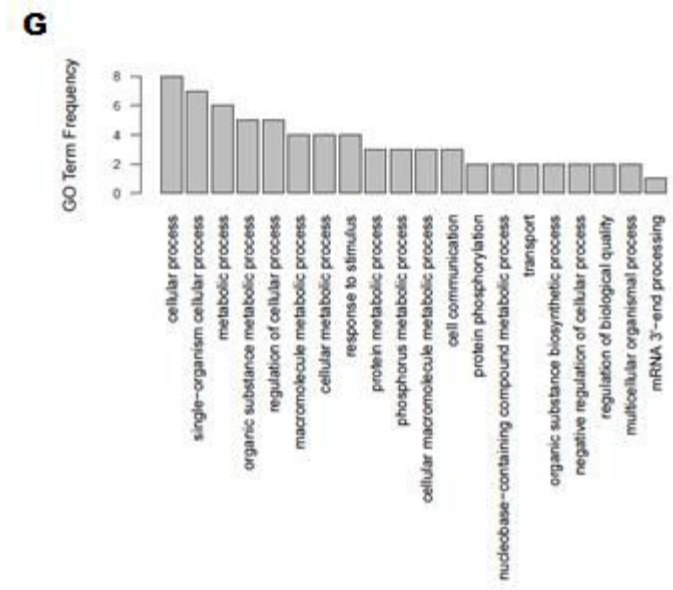


Figure 3.4 | piRNA Cluster-Level Analysis of the *Flamenco* Locus of *Drosophila melanogaster*. **(A)** Number of known TE, Gene, or "Other" feature calls **(B)** Nucleotide occupancy of the feature calls **(C)** Orientation of feature calls **(D)** Number TE calls within the piRNA cluster for the most represented TE superfamilies **(E)** Number of individual TEs called in the top 5 represented superfamilies. Additional functionality can optionally be specified by the user to prompt production of a **(F)** sRNA coverage plot with feature content and orientation in 0-2hr eggs libraries **(G)** GO term frequency plot regarding all gene hits within the piRNA cluster.

Section 3.4: Results

Benchmarking Software Performance

piClusterBusteR was timed for the analysis of the top 5 piRNA clusters identified in *Drosophila melanogaster* ovarian samples. When running sequentially on a single Intel(R) Xeon(R) CPU E5-2683 v4 at 2.10GHz, piClusterBusteR took approximately 3 hours to complete.

When utilizing the multithreading and multitasking capability of piClusterBusteR to analyze the same 5 piRNA clusters, using 5 nodes and 6 cores per node using the same processor speed, the timing of the piClusterBusteR run took approximately 20 minutes to run. One compute node was designated per piRNA cluster and six threads were utilized on each node. This run represents the enhanced capability of piClusterBusteR if additional resources, such as a computing cluster and queue submission system, are available to the user. Output from these independent analyses was identical.

piClusterBusteR results were observed on the previous established contents of the *flamenco* locus in *Drosophila melanogaster* which were extracted from the FlyBase database (Brennecke et al. 2007; Attrill et al. 2016).

Previous exploration with regard to the contents of the *flamenco* locus used RepeatMasker to characterize sequence content (Malone et al. 2009). The results of this method were extracted from the UCSC Table Browser retrieval tool (Karolchik et al. 2004). The contents and strand specificity of feature calls within the *flamenco* locus identified by piClusterBusteR were consistent with the previous observation (Figure 3.5).

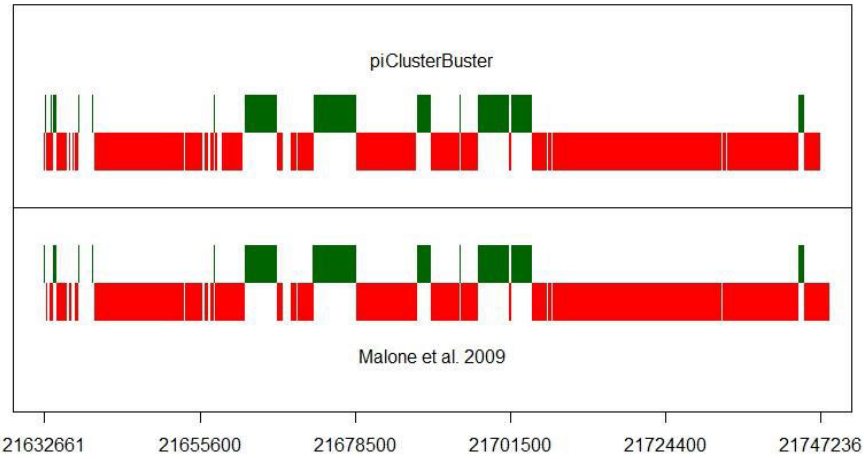
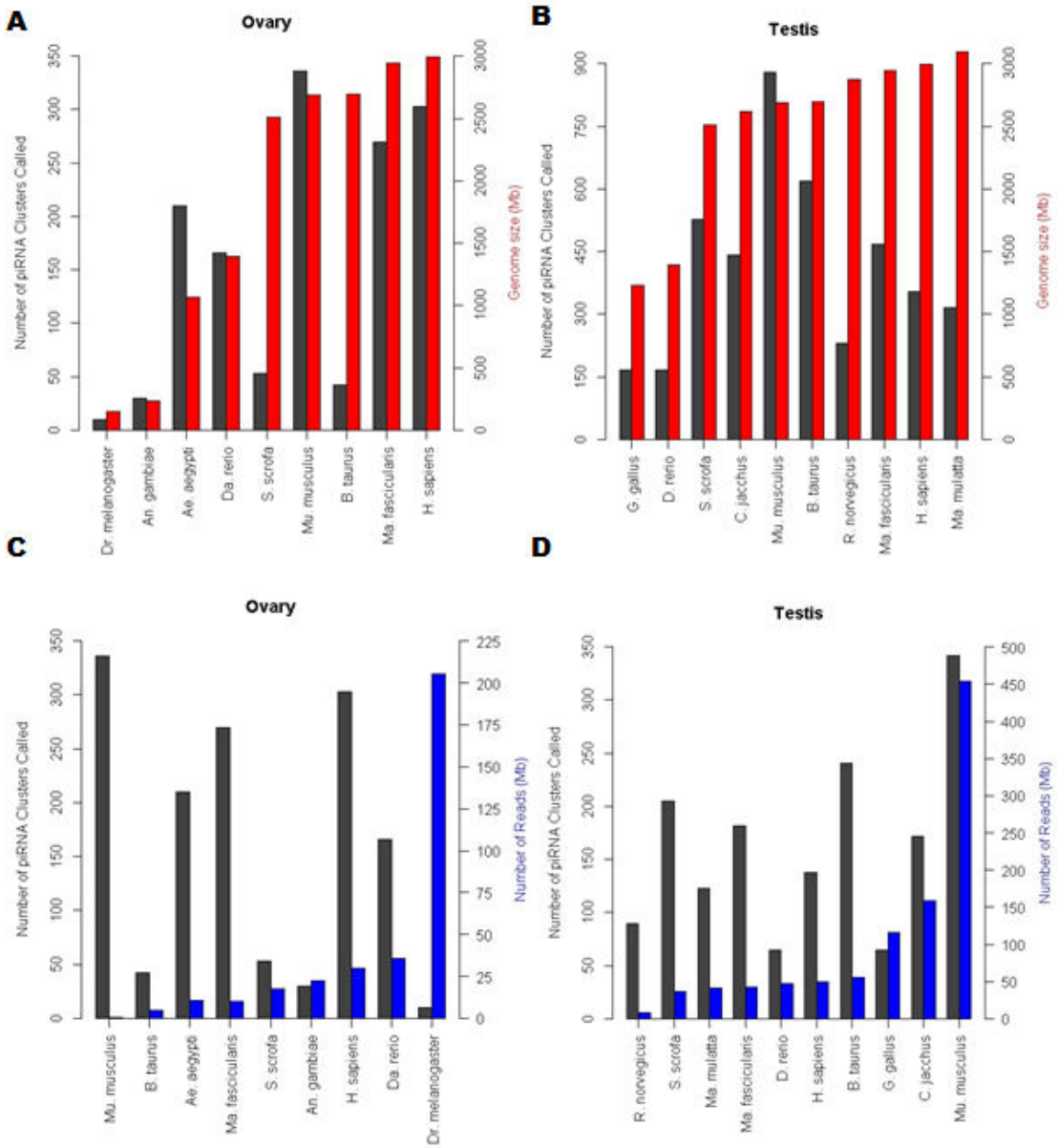


Figure 3.5 | Comparison of *Flamenco* TE Annotation. A depiction of the agreement of piClusterBusteR characterization of the *flamenco* locus in comparison to previous reported characterization of TE contents in this locus in *Drosophila melanogaster* (Malone et al. 2009) (Figure 3.1S). Green boxes represent a sense orientation of TE calls and the red boxes represent an antisense orientation.

piRNA Cluster Definition is Unaffected by Genome Size and Read Coverage

The 13 Metazoan species analyzed were selected based on data availability. A density-based approach of piRNA definition was implemented via use of the previous established software, proTRAC (Rosenkranz & Zischler 2012). A Pearson correlation test demonstrated that piRNA cluster definition by proTRAC appears to be irrespective of genome size of the organism and the number of piRNAs available for analysis at the 1% confidence level (Figure 3.6A-D, Figure 3.5S).


A
Ovary
Number of piRNA Clusters Called
Genome size (Mb)
Dr. melanogaster
An. gambiae
Ae. aegypti
Da. rerio
S. scrofa
Mu. musculus
B. taurus
Ma. fascicularis
H. sapiens
B
Testis
Number of piRNA Clusters Called
Genome size (Mb)
G. gallus
D. rerio
S. scrofa
C. jacchus
Mu. musculus
B. taurus
R. norvegicus
Ma. fascicularis
H. sapiens
Ma. mulatta
C
Ovary
Number of piRNA Clusters Called
Number of Reads (Mb)
Mu. musculus
B. taurus
Ae. aegypti
Ma. fascicularis
S. scrofa
An. gambiae
H. sapiens
Da. rerio
Dr. melanogaster
D
Testis
Number of piRNA Clusters Called
Number of Reads (Mb)
R. norvegicus
S. scrofa
Ma. mulatta
Ma. fascicularis
D. rerio
H. sapiens
B. taurus
G. gallus
C. jacchus
Mu. musculus

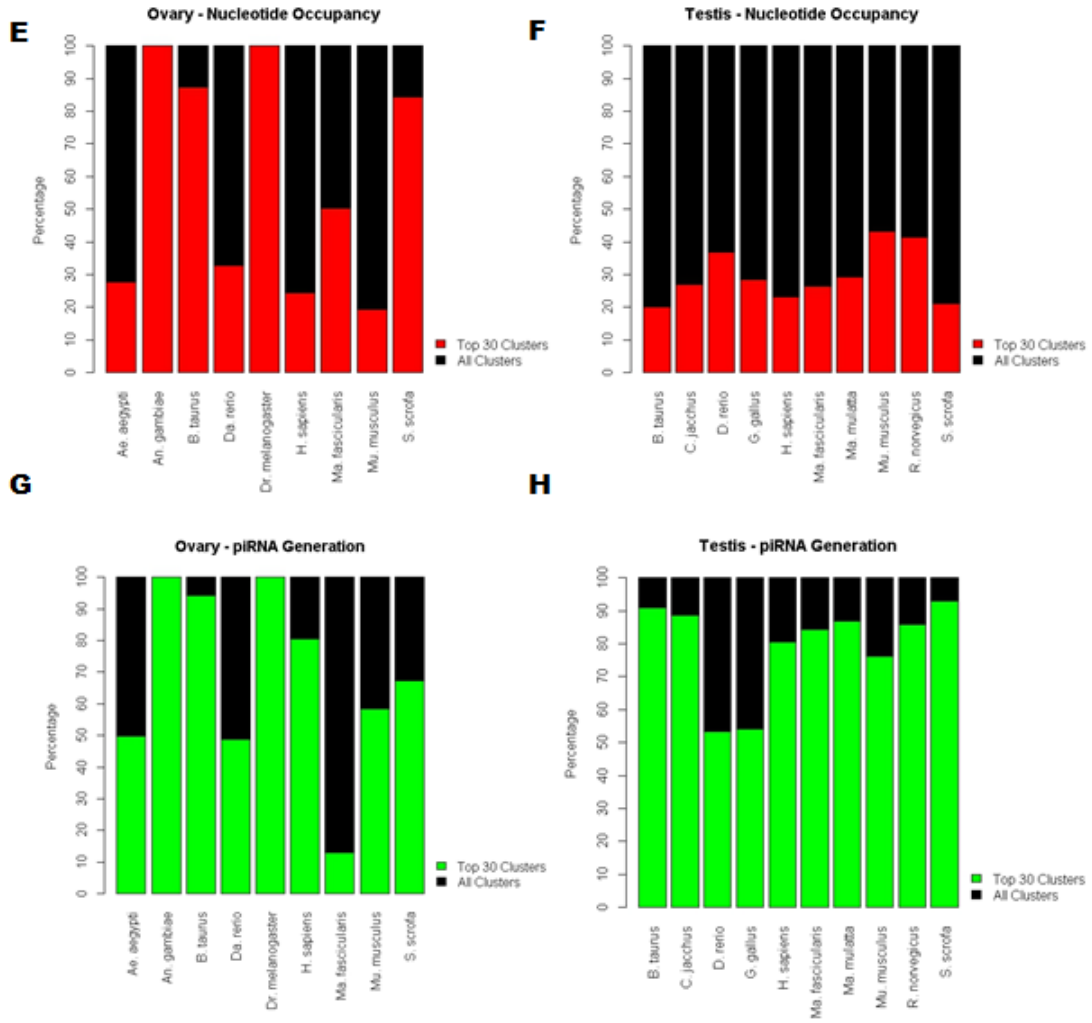


Figure 3.6 | Representation of piRNA Cluster Loci. Number of piRNA cluster calls relative to the genome size of the organism in **(A)** ovary and **(B)** testis. Number of piRNA cluster calls relative to the number of reads available in **(C)** ovary and **(D)** testis. The percent composition of the top 30 piRNA clusters (red/blue) relative to the full repertoire of piRNA clusters (black) called by proTRAC with regard to the percent of nucleotides occupied by piRNA clusters in **(E)** ovary and **(F)** testis. The percentage of piRNA generated from the top piRNA cluster loci in **(G)** ovary and **(H)** testis.

Given that the breadth of piRNA clusters is difficult to define in a given organism, due to the concern of false positives, I have used only the top 30 major contributing piRNA cluster loci in the between species comparisons of piRNA cluster composition. piRNA cluster definition required a length of at least five kilobases, at least 75% of the piRNAs deriving from a putative piRNA cluster with a U-1 or A-10, at least 50% of the piRNAs deriving from a putative piRNA cluster with a U-1 and A-10, and the top 1% of piRNA sequences cannot comprise more than 90% of the piRNAs that were used to define a particular piRNA cluster.

In ovarian samples, the nucleotide occupancy of the top 30 piRNA clusters ranged from 19.3% to all of the piRNA clusters defined in a tissue with an average of 58.5% and median of 50.2% in these species (Figure 3.6E). The percent piRNA generation of the top 30 piRNA cluster loci ranged from 13.0% to all of the piRNAs generated in a tissue with an average of 68.0% and median of 67.3% relative to total piRNA generation (Figure 3.6G).

In testes samples, the nucleotide occupancy ranged from 20.1% to 43.1% relative to all of the piRNA clusters defined with an average of 29.7% and median of 27.8% in these species (Figure 3.6F). The percent piRNA generation of the top 30 piRNA cluster loci ranged from 53.3% to 93% with an average of 79.3% and median of 85.0% relative to total piRNA generation (Figure 3.6H).

Therefore, I consider the top 30 piRNA clusters to be representative of large scale architecture of genomic piRNA clusters based on the large proportion of the nucleotide occupancy and piRNA generation that is correlated with these loci.

Top piRNA Cluster Architecture is Conserved in Metazoans on a Large Scale

The analysis of piRNA cluster architecture focused on the number of piRNA clusters, piRNA cluster size, the known features within the piRNA cluster, and the orientation of the known feature.

Certain features of piRNA cluster architecture were conserved better than others. In all of the Metazoan species observed in this analysis, the majority of piRNA cluster sequence was unable to be attributed to any known origin. Unannotated sequence ranged between 18% and 70% of piRNA cluster composition. TEs were the major known contributor to piRNA cluster loci. TEs occupied up to 78% of ovarian piRNA cluster loci and 62% of testis piRNA clusters, with an average piRNA cluster occupancy of 40% to 32% in ovarian and testes libraries, respectively. Sequences of known genic origin ranged from 1 to 11%, with an average of 3% and 3.5% piRNA cluster occupancy in ovaries and testes, respectively. Non-genic, non-TE sequences within the NCBI database were the least significant contributor to piRNA cluster loci in these species, ranging from 0 to 9% of piRNA cluster composition, with an average piRNA cluster occupancy between 4 to 3.5% in ovarian and testes libraries (Figure 3.7).

The strand specificity of feature calls within top piRNA cluster loci was also summarized. Features were predominantly characterized on the sense strand of piRNA clusters. The nucleotide occupancy of sense features accounted for between 38.0% to 61.0% of feature calls with a piRNA cluster with an average of 50.0% and 52.8% in ovarian and testes samples, respectively, in these species.

Tissue-Specificity of piRNA Cluster Loci

piRNA cluster definition can vary between sRNA libraries that derived from the same tissue. The number of defined piRNA clusters differed from 3 to 398 calls between two samples of the same tissue with a mean difference of 75 and median difference of 45 piRNA cluster calls. Although, at least 52.4%, and up to 92.2% of the lesser piRNA cluster definitions were also represented in the larger sample of piRNA cluster calls. piRNA cluster definition demonstrated an average of 67.8% overlap in same tissue samples (Figure 3.8).

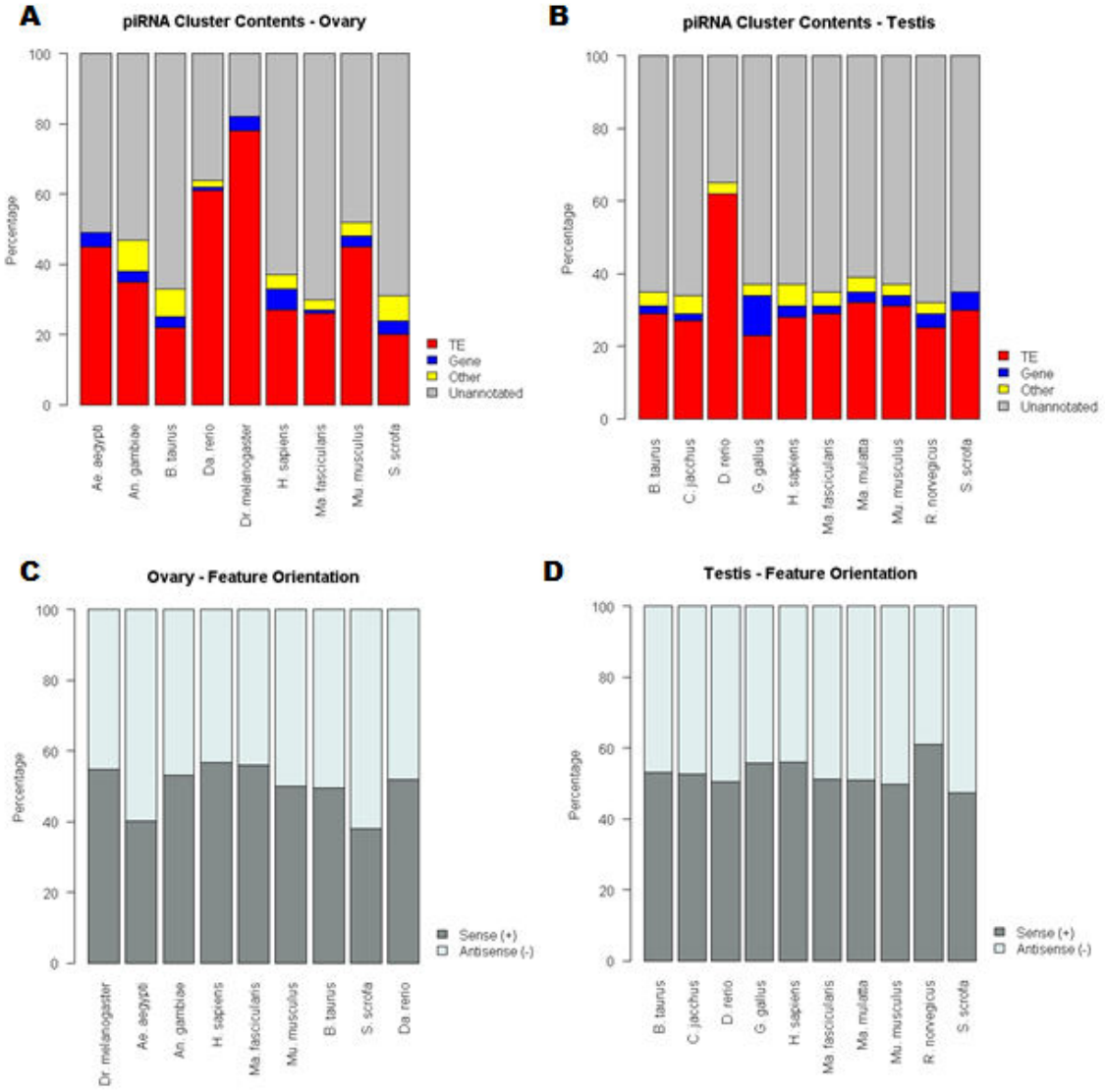
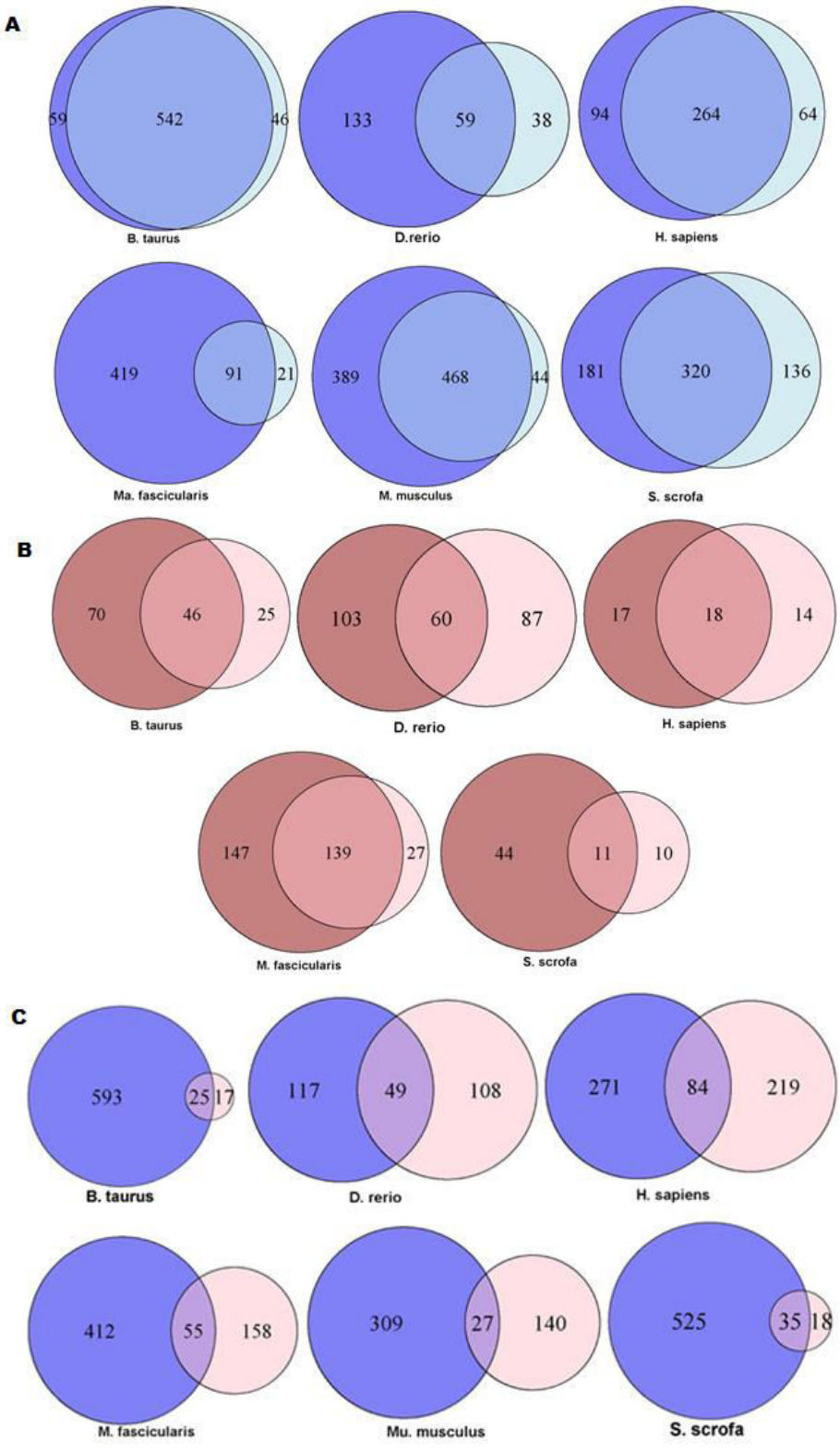


Figure 3.7 | Comparison of Top piRNA Cluster Composition. piRNA cluster content comparison between species in **(A)** ovary and **(B)** testis. Orientation of the feature calls in **(C)** ovary and **(D)** testis samples. These data represent an analysis of the top 30 piRNA cluster loci in each species. Available ovarian and testes datasets from the piRNA cluster database and the Short Read Archive were used to run piClusterBusteR (Rosenkranz 2016; Kodama et al. 2012). Only ten piRNA clusters were called in the *Dr. melanogaster* ovarian library.


A
59
542
46
B. taurus
133
59
38
D.rerio
94
264
64
H. sapiens
419
91
21
Ma. fascicularis
389
468
44
M. musculus
181
320
136
S. scrofa
B
70
46
25
B. taurus
103
60
87
D. rerio
17
18
14
H. sapiens
147
139
27
M. fascicularis
44
11
10
S. scrofa
C
593
25
17
B. taurus
117
49
108
D. rerio
271
84
219
H. sapiens
412
55
158
M. fascicularis
309
27
140
Mu. musculus
525
35
18
S. scrofa

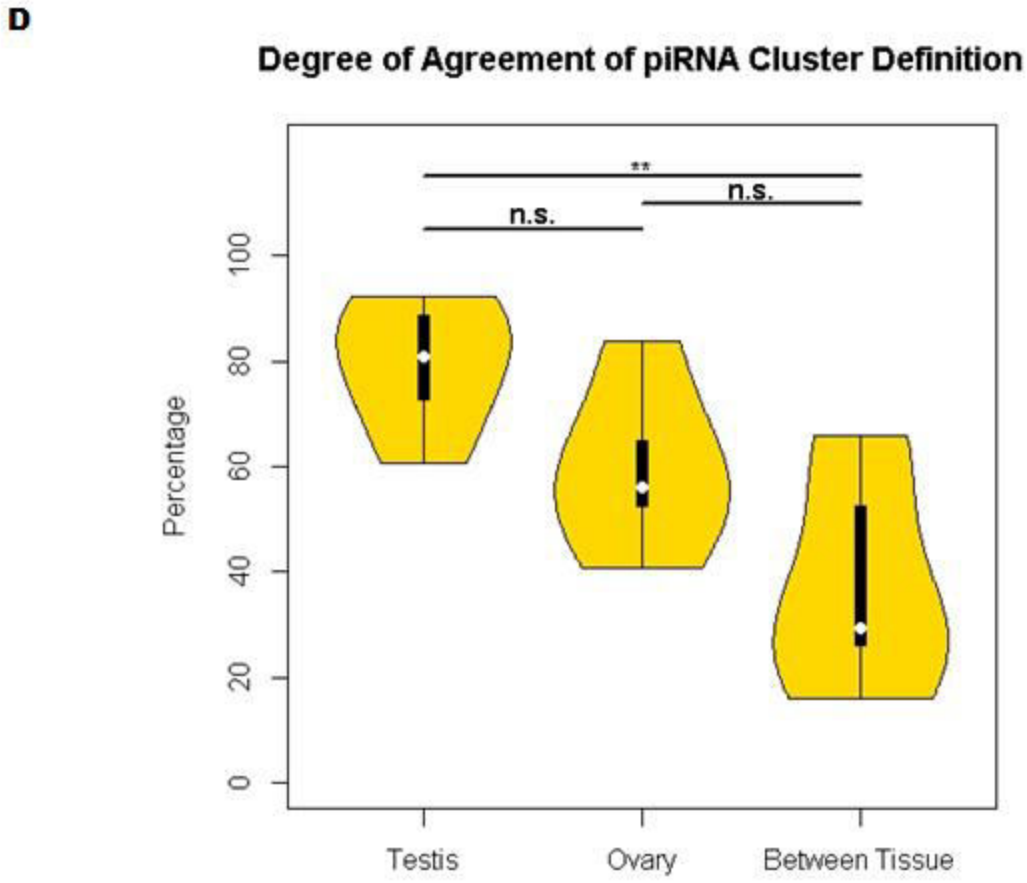


Figure 3.8 | Tissue-specificity of piRNA Cluster Definition. Venn Diagrams representing the degree of overlap of piRNA cluster definitions between two independent libraries of the same tissue. **(A)** Blue circles represent testes samples and **(B)** red circles represent ovarian samples. Only one ovarian sRNA library was analyzed in *M. musculus* and *D. melanogaster* testis defined no piRNA clusters. **(C)** Venn Diagrams representing the degree of overlap of piRNA cluster definitions between ovary and testis piRNAs. **(D)** Violin plot representing the agreement of piRNA cluster calls within same species testes, within same species ovaries, and a comparison between testes and ovaries piRNA cluster definition. (Figure 3.12S)

I observed a significantly lesser degree of agreement piRNA cluster calls relative to same sample testis libraries. Same sample ovary libraries also showed a small increase in piRNA cluster agreement relative to samples between tissues (Figure 3.8D). The number of piRNA cluster definition differed from 9 to 576 calls between two samples of the same tissue with a mean difference of 255 and median difference of 211 piRNA cluster calls. The lesser sample of piRNA cluster definitions ranged in agreement from 16.2% to 66.0% with an average of 34.6% agreement and a median of 29.5% between samples (Figure 3.8C).

Section 3.5: Discussion

The options for piClusterBusteR performance enhancement allow for utilization of multitasking and multithreading. Use of multithreading allows for the execution of piClusterBusteR processes by multiple nodes simultaneously. Utilization of the multitasking capability of piClusterBusteR prompts independent, parallel submission for each piRNA cluster of interest to independent compute nodes. Multithreading and multitasking piClusterBusteR runs allows the user the capability to significantly increase the number of piRNA clusters under observation without significantly increasing the timing of the piClusterBusteR run. piClusterBusteR supports both Torque/Maui or Slurm resource management software.

This comparison of piRNA cluster architecture focuses on the major genomic loci contributing to piRNA populations. By only considering only the top piRNA cluster loci

in this analysis, I can be relatively confident in piRNA cluster definition relative to other piRNA-generating loci. The top 30 piRNA clusters also were a large representation of the total nucleotides occupied by piRNA clusters in these genomes, as well as disproportionally large contributors to total piRNA populations in these species (Figure 3.6). Taken together, the contents of the piRNA clusters in the analysis serve as the best representation of piRNA cluster architecture in these species.

Since it is difficult to determine whether RepeatMasker and CENSOR will annotate a piRNA cluster of interest more thoroughly, with higher confidence, I implemented nested annotation. A nested annotation approach allows for both of the programs that performed well in annotating sequences that are dense with repeats, RepeatMasker and CENSOR, the opportunity to characterize the sequence of interest, while only maintaining the best annotation in the description of the contents of piRNA cluster sequence (Figure 3.2) (Smit et al. 1996; Jurka et al. 1996). This method allows for consistent and accurate characterization amongst diverse piRNA clusters on a large scale.

I also noted that the degree of sense or antisense orientation of feature calls within individual piRNA clusters correlated with the direction of transcription in known piRNA clusters, *flamenco* and 42AB (Brennecke et al. 2007; Malone et al. 2009). Therefore, the orientation of feature calls within a piRNA cluster may be informative in the prediction of the nature of piRNA cluster transcription.

Components of piRNA architecture were strikingly similar across species. With regard to known piRNA cluster features, TEs consistently composed the majority by nucleotide

occupancy and a relatively low percentage of known genic and "other" calls. The majority of informative, "other" hits within the NCBI nucleotide database were associated with mRNAs that were not available in the organism-specific gene set. Other informative non-genic, non-TE sequence appeared to be of viral and rRNA origin. The orientation of feature calls within piRNA clusters were also highly conserved in these species. Taken together, these data suggest highly conserved nature, yet dynamic capacity within piRNA cluster architecture with regard to known features in Metazoans (Figure 3.6).

I observed that a significant portion of the piRNA cluster sequence was unable to be characterized in the species observed in this study (Figure 3.7). This observation prompts an interesting question regarding the derivation of piRNA cluster sequence whose origin is currently undetectable and its purpose within the piRNA clusters. This sequence is of particular biological interest given that these sequences occupy significant regions of piRNA clusters and may further inform scientific knowledge of piRNA cluster biogenesis and function.

Sets of piRNA clusters were differentially represented between different, and within the same, independent tissue samples. It is worth noting that differential representation of piRNA generating loci between same, independent tissue samples may be due to the previous observation that sRNA libraries represent only a subset of the complete sRNA populations within the cell, even when deep sequencing is performed (Yamtich et al. 2015). However, the variability in piRNA cluster overlap was far greater between tissues

than within tissues when comparing piRNA cluster definitions between libraries. Therefore, preliminary observation of these data supports a model in which different regions of the genome appear to be responsible for generating the majority of piRNAs in ovaries and testes samples in Metazoans and it may be advantageous for an organism to have a diverse, dynamic set of piRNA cluster activity in a unique cellular environment.

Section 3.6: References

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

Analytics, R., 2014. doMC: Foreach parallel adaptor for the multicore package. *R package version*, 1(3).

Aravin, A.A. et al., 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular cell*, 31(6), pp.785–799.

Aravin, A.A., Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *science*, 318(5851), pp.761–764.

Arensburger, P. et al., 2011. The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC genomics*, 12(1), p.606.

Attrill, H. et al., 2016. FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic acids research*, 44(D1), pp.D786–D792.

Barckmann, B. et al., 2015. Aubergine iCLIP reveals piRNA-dependent decay of mRNAs involved in germ cell development in the early embryo. *Cell reports*, 12(7), pp.1205–1216.

Brennecke, J. et al., 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science (New York, NY)*, 322(5906), p.1387.

Brennecke, J. et al., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6), pp.1089–103.

Charif, D. & Lobry, J.R., 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution*. Springer, pp. 207–232.

Garc'\ia-López, J. et al., 2014. Global characterization and target identification of piRNAs and endo-siRNAs in mouse gametes and zygotes. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(6), pp.463–475.

Girke, T., 2014. systemPipeR: NGS workflow and report generation environment. *UC Riverside. https://github. com/tgirke/systemPipeR*.

Grimson, A. et al., 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217), pp.1193–1197.

Houwing, S. et al., 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, 129(1), pp.69–82.

Jurka, J. et al., 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Computers \& chemistry*, 20(1), pp.119–121.

Jurka, J. et al., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110(1-4), pp.462–467.

Karolchik, D. et al., 2004. The UCSC Table Browser data retrieval tool. *Nucleic acids research*, 32(suppl 1), pp.D493–D496.

Keam, S.P. et al., 2014. The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. *Nucleic acids research*, 42(14), pp.8984–8995.

Kidwell, M.G., 1983. Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 80(6), pp.1655–1659.

Kidwell, M.G., Kidwell, J.F. & Sved, J.A., 1977. Hybrid dysgenesis in Drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, 86(4), pp.813–833.

Kodama, Y., Shumway, M. & Leinonen, R., 2012. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research*, 40(D1), pp.D54–D56.

Kuramochi-Miyagawa, S. et al., 2004. Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development*, 131(4), pp.839–849.

Kuramochi-Miyagawa, S. et al., 2001. Two mouse piwi-related genes: miwi and mili. *Mechanisms of development*, 108(1), pp.121–133.

Lawrence, M. et al., 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8), p.e1003118.

Malone, C.D. et al., 2009. Specialized piRNA Pathways Act in Germline and Somatic Tissues of the< i> Drosophila</i> Ovary. *Cell*, 137(3), pp.522–535.

Pages, H. et al., 2009. String objects representing biological sequences, and matching algorithms. *R package version*, 2(2).

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

Reimand, J. et al., 2007. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl 2), pp.W193–W200.

Rosenkranz, D., 2016. piRNA cluster database: a web resource for piRNA producing loci. *Nucleic acids research*, 44(D1), pp.D223–D230.

Rosenkranz, D. & Zischler, H., 2012. proTRAC-a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC bioinformatics*, 13, p.5.

Rouget, C. et al., 2010. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature*, 467(7319), pp.1128–1132.

Saito, K. et al., 2007. Pimet, the Drosophila homolog of HEN1, mediates 2′-O-methylation of Piwi-interacting RNAs at their 3′ ends. *Genes \& development*, 21(13), pp.1603–1608.

Sayers, E.W. et al., 2011. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1), pp.D38–D51.

Scrucca, L., 2004. qcc: an R package for quality control charting and statistical process control. *dim (pistonrings)*, 1(200), p.3.

Smit, A.F., Hubley, R. & Green, P., 1996. RepeatMasker Open-3.0.

Vodovar, N. et al., 2012. Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PloS one*, 7(1), p.e30861.

Wickham, H., 2009. plyr: Tools for splitting, applying and combining data. *R package version 0.1*, 9, p.651.

Yamtich, J. et al., 2015. piRNA-like small RNAs mark extended 3'UTRs present in germ and somatic cells. *BMC genomics*, 16(1), p.462.

Yin, H. & Lin, H., 2007. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature*, 450(7167), pp.304–308.

Section 3.7: Supplementary Material

	Malone *et al.*, 2009	piClusterBuster
# of + TE calls	27	34
Nucleotide Occupancy of + TE calls	22,311 (20.3%)	22,606 (20.6%)
# of - TE calls	74	95
Nucleotide Occupancy of - TE calls	87,583 (79.7%)	86,867 (79.3%)

Table 3.1S: Improved Annotation with piClusterBusteR

Species	Datasets
Dr. melanogaster	♀ - SRR618933, SRR797070, SRR797071, SRR797172, SRR797179, SRR797182, SRR797187, SRR797193, SRR797195, SRR797196, SRR797197, SRR797198, SRR797199, SRR797200, SRR797203 ♂ - GSM280085
Ae. aegypti	RNAlib17
An. gambiae	SRR1927173
H. sapiens	♀ - SRR1755247,SRR1755248, SRR1755251, SRR1755252, SRR1755255, SRR1755256 ♂ - ERR328151, SRR835324, SRR835325, SRR950451
Ma. mulatta	♂ - SRR116839, SRR553581, SRR606728
Ma. fascicularis	♀ - SRR1755241, SRR1755242 ♂ - SRR1755243, SRR1755244
C. jacchus	♂ - SRR606715, SRR1041905, SRR1041906, SRR1041907
Mu. musculus	♀ - SRR014234 ♂ - SRR014231, SRR014232, SRR014233, SRR014235, SRR772028, SRR772029, SRR772030, SRR772031, SRR772032, SRR772033, SRR772050, SRR772051, SRR772052, SRR772053, SRR772054, SRR772055
R. norvegicus	♂ - SRR035663
B. taurus	♀ - SRR1755229, SRR1755230 ♂ - SRR1755231, SRR1755232
S. scrofa	♀ - SRR066809, SRR1274763 ♂ - SRR066810, SRR1274764, SRR1654828
G. gallus	♂ - SRR772069
Da. rerio	♀ - SRR578904, SRR578905, SRR578906, SRR578913, SRR578914, SRR578915 ♂ - SRR578922, SRR578923

Table 3.2S: Description of Datasets

Species	Number of piRNA Clusters	Genome Size (Mb)	Total Read Count
Dr. melanogaster	10	148	205,554,806
Ae. aegypti	210	1064	10,655,147
An. gambiae	30	236	22,370,627
H. sapiens	303	2996	29,968,314
Ma. fascicularis	213	2946	10,117,534
Mu. musculus	336	2689	269,297
B. taurus	42	2697	4,528,165
S. scrofa	48	2508	17,661,429
Da. rerio	166	1391	35,936,571

Table 3.3S: Ovary Genome Size and Read Count

Species	Number of piRNA Clusters	Genome Size (Mb)	Total Read Count
H. sapiens	355	2996	50,251,433
Ma. mulatta	316	3097	41,304,681
Ma. fascicularis	467	2946	42,234,678
C. jacchus	441	2621	158,431,380
Mu. musculus	879	2689	454,610,320
R. norvegicus	231	2870	8,407,181
B. taurus	618	2697	55,482,891
G. gallus	166	1230	116,234,474
D. rerio	166	1391	47,798,654
S. scrofa	527	2508	37,221,910

Table 3.4S: Testis Genome Size and Read Count

Pearson Correlation Test Relative to piRNA Cluster Calls	Tissue	T value	Pearson Correlation Value	P value
Genome Size	Ovary	1.8787	0.5789624	0.1024
	Testis	-1.3173	-0.4457003	0.2292
Number of Reads	Ovary	1.7516	0.5520267	0.1233
	Testis	2.936	0.7428724	0.02184

Table 3.5S: Correlations of piRNA Cluster Definition

Species	Number of piRNA Clusters Called	Average piRNA Cluster Size (kb)	Total Nucleotides Occupied by piRNA Clusters	Genome Occupancy	% TE	% Gene	% Other	% Unannotated
Dr. melanogaster	10	11.4	113,593	0.07%	78	4	0	18
Ae. aegypti	210	10	2,109,482	0.15%	45	4	0	51
An. gambiae	30	8.9	268,084	0.10%	35	3	9	53
H. sapiens	303	10	3,023,129	0.09%	27	6	4	63
Ma. fascicularis	213	21.8	4,635,040	0.15%	26	1	3	70
Mu. musculus	336	7.67	2,576,831	0.09%	45	3	4	48
B. taurus	42	15.5	652,933	0.02%	22	3	8	67
S. scrofa	48	17.8	856,408	0.03%	20	4	7	69
Da. rerio	166	10.6	1,761,421	0.13%	61	1	2	36

Table 3.6S: piRNA Cluster Contents - Ovary

Species	Number of piRNA Clusters Called	Average piRNA Cluster Size (kb)	Total Nucleotides Occupied by piRNA Clusters	Genome Occupancy	% TE	% Gene	% Other	% Unannotated
H. sapiens	355	12.9	4578738	0.14	28	3	6	63
Ma. mulatta	300	11.2	3360121	0.12	32	3	4	61
Ma. fascicularis	213	12	5602830	0.15	29	2	4	65
C. jacchus	441	13.5	5959699	0.22	27	2	5	66
Mu. musculus	167	21.8	3642073	0.13	31	3	3	63
R. norvegicus	231	18	4146688	0.14	25	4	3	68
B. taurus	618	12	7434999	0.28	29	2	4	65
G. gallus	166	10.7	1772287	0.18	23	11	3	63
D. rerio	157	10.8	1691347	0.12	62	0	3	35
S. scrofa	527	43.2	6147214	0.24	30	5	0	65

Table 3.7S: piRNA Cluster Contents - Testis

Species	Nucleotide Occupancy of Top 30 piRNA clusters	Nucleotide Occupancy of All piRNA Cluster Calls	Percent Composition of Top 30 piRNA Clusters
Dr. melanogaster	113,593	113,593	100%
Ae. aegypti	585,679	2,109,482	27.8%
An. gambiae	268,084	268,084	100%
H. sapiens	741,925	3,023,129	24.5%
Ma. fascicularis	2,326,841	4,635,040	50.2%
Mu. musculus	496,848	2,576,831	19.3%
B. taurus	569,934	652,933	87.3%
S. scrofa	722,238	856,408	84.3%
Da. rerio	575,932	1,761,421	32.7%

Table 3.8S: Nucleotide Occupancy of the Top 30 piRNA Clusters - Ovary

Species	Nucleotide Occupancy of Top 30 piRNA clusters	Nucleotide Occupancy of All piRNA Cluster Calls	Percent Composition of Top 30 piRNA Clusters
H. sapiens	1,054,062	4,578,738	23.0%
Ma. mulatta	979,655	3,360,121	29.2%
Ma. fascicularis	1,472,922	5,602,830	26.3%
C. jacchus	1,609,193	5,959,699	27.0%
Mu. musculus	1,570,168	3,642,073	43.1%
R. norvegicus	1,719,232	4,146,688	41.5%
B. taurus	1,496,074	7,434,999	20.1%
G. gallus	505,336	1,772,287	28.5%
D. rerio	622,435	1,691,347	36.8%
S. scrofa	1,297,435	6,147,214	21.1%

Table 3.9S: Nucleotide Occupancy of the Top 30 piRNA Clusters - Testis

Species	piRNA Generation of Top 30 piRNA clusters	piRNA Generation of All piRNA Cluster Calls	Percent piRNA Generation of Top 30 piRNA Clusters
Dr. melanogaster	27,422	27,422	100%
Ae. aegypti	80,000	161,016	49.7%
An. gambiae	124,402	124,402	100%
H. sapiens	97,707	121,570	80.4%
Ma. fascicularis	3,280	25,222	13.0%
Mu. musculus	95,879	164,121	58.4%
B. taurus	3,630	3,854	94.2%
S. scrofa	792	1,176	67.3%
Da. rerio	8,5872	175,856	48.8%

Table 3.10S: piRNA Generation of the Top 30 piRNA Clusters - Ovary

Species	piRNA Generation of Top 30 piRNA clusters	piRNA Generation of All piRNA Cluster Calls	Percent piRNA Generation of Top 30 piRNA Clusters
H. sapiens	308061	383126	80.4%
Ma. mulatta	414155	477018	86.8%
Ma. fascicularis	411479	488199	84.3%
C. jacchus	559813	631692	88.6%
Mu. musculus	363370	476957	76.2%
R. norvegicus	751565	876417	85.8%
B. taurus	820381	903538	90.8%
G. gallus	195819	362278	54.1%
D. rerio	95038	178252	53.3%
S. scrofa	578459	621844	93.0%

Table 3.11S: piRNA Generation of the Top 30 piRNA Clusters - Testis

Wilcoxon Rank Sum Test with Continuity Correction

Testis vs Ovary

W = 25, p-value = 0.08225

Testis vs Between Tissue

W = 1, p-value = 0.002165

Ovary vs Between Tissue

W = 7, p-value = 0.08874

Table 3.12S: Degree of Agreement of piRNA Cluster Definition

Chapter 4: TruePaiR - Software for the Accurate Identification of Complementary piRNA Read Pairs in High-Throughput Sequencing Data

Section 4.1: Abstract

piRNAs and their biogenesis pathways are well-conserved in Metazoans (Grimson et al. 2008). piRNAs have been implicated in transcriptional, post-transcriptional, and translational regulation (Grivna et al. 2006; Lin & Yin 2008; Brennecke et al. 2008; Brennecke et al. 2007; Aravin et al. 2007). I analyzed the signatures of a critical process in the primary and secondary mechanism of piRNA biogenesis, referred to as the amplification loop.

The presence of U-1 and A-10 bias within piRNA populations is an indicator, but not an absolute measure of piRNA amplification. By further considering imperfect and perfect sequence complementarity within the first ten base pairs of piRNAs, the active site promoting secondary piRNA biogenesis, I developed practical and statistically powerful metrics to observe relative piRNA amplification. TruePaiR is a fast and effective general software tool to assess the relative utilization of piRNA amplification in high throughput sRNA sequencing data.

The results of TruePaiR runs in seven species and five tissues serve as a benchmark for meaningful context of piRNA amplification. The TruePaiR metrics provide foundational data regarding the in terms of species specificity, tissue specificity, as well as the relative participation based upon origin-based piRNA subsets regarding piRNA amplification. The low degree of variability of same sample TruePaiR runs allows for metric reliability,

reproducibility, as well as the ability to detect subtle differences in piRNA amplification within and between species and tissues. Given that TruePaiR serves as an effective and consistent metric of piRNA amplification across species, it can represent a new, meaningful standard in the degree of piRNA amplification in a specific organism and tissue that is or is not expected to undergo piRNA amplification.

Section 4.2: Introduction

piRNAs are the largest, in both size and number, distinct subclass of sRNAs (Zhang et al. 2014). Yet, piRNA biogenesis, targeting, and function are less well-understood relative to other sRNA pathways: siRNAs and miRNAs. piRNAs are quite distinct from the siRNA and miRNA pathways (Grimson et al. 2008; Aravin et al. 2007; Brennecke et al. 2007; Murchison & Hannon 2004).

piRNAs are noticeably distinct in that they are longer in sequence length, from 24-33 nts, relative to siRNAs and miRNAs (Aravin et al. 2007; Brennecke et al. 2007; Zhang et al. 2014). piRNAs are not known to form a hairpin secondary structure, and therefore have a Dicer-independent biogenesis (Aravin et al. 2007; Brennecke et al. 2007; Grimson et al. 2008). piRNA contain a 3'-O-methyl modification, modulated by HEN1, to protect from modification at the 3' end of piRNAs (Horwich et al. 2007; Saito et al. 2007; Yang et al. 2006).

piRNAs are generated via a primary and secondary mechanism of biogenesis. Primary piRNAs derive from discrete genomic loci, referred to as piRNA clusters. piRNA

clusters range from five to several hundred kbps in length and generally persist in heterochromatin (Arensburger et al. 2011; Brennecke et al. 2007). TE remnants are the major known component of piRNA clusters, which also can contain sequences of genic, viral, and unknown origin (Aravin et al. 2007; Schreiner & Atkinson 2017). Hundreds of millions of unique piRNA sequences have been identified, since piRNA sequences are not well-conserved amongst Metazoans (Zhang et al. 2014). piRNA cluster loci, however, are well-conserved by species (Schreiner & Atkinson 2017; Malone & Hannon 2010; Zanni et al. 2013; Malone & Hannon 2009; Grimson et al. 2008). piRNAs have been implicated in transcriptional, post-transcriptional, and translational regulation within the cell (Grivna et al. 2006; Lin & Yin 2008; Brennecke et al. 2008; Brennecke et al. 2007; Aravin et al. 2007).

Primary piRNA biogenesis is initiated via the transcription of a single, long piRNA precursor transcript (Brennecke et al. 2007). The Zucchini endonuclease slices the primary piRNA precursor molecule, generally resulting with a U at the first position of mature piRNAs (Nishimasu et al. 2012). Mature piRNAs then associate with the PIWI protein, Aub, to form a RNA-induced silencing complex (RISC) (Schwarz et al. 2004; Brennecke et al. 2007). The RISC is then guided to secondary piRNAs via complementarity of the associated primary piRNA.

Secondary piRNAs are generated as a result of the slicing mechanism of the RISC (Aravin et al. 2007). Argonaute, and therefore PIWI, proteins slice between the tenth and eleventh base pair of target molecules (Tolia & Joshua-Tor 2007). Initially, the Ping-Pong model of piRNA amplification suggested that given that adenine complements the

uracil at the first position of the primary piRNA, and the secondary piRNA complements in the reverse orientation, the tenth position of secondary piRNAs generally have an A at position ten (Holbrook et al. 1991; Brennecke et al. 2007). An alternative model challenged this hypothesis, suggesting rather that the A-10 bias arises as a result of intrinsic preference of the target molecules of Aubergine (Wang et al. 2014). The 3' end of the piRNAs trail on the opposite ends of the complex, and therefore, do not necessarily compliment (Zamore 2010; Aravin et al. 2007).

The degree of piRNA amplification is an important metric for assessing the activity of the piRNA biogenesis pathways. A metric exists to assess the degree of piRNA amplification in high throughput sRNA sequencing data considering the extent of the U-1 and A-10 bias. A Z-score test statistic can be calculated a to quantitate the significance of the observed bias within piRNA populations using the "pingpong" function to quantitate U-1 and A-10 bias overrepresentation within the NGS Toolbox of the piRNA cluster database (Zhang et al. 2011; Rosenkranz & Zischler 2012). Although, a method has not been developed to consider the sequence complementarity of piRNAs within a sRNA dataset of interest.

In order to correctly specifically identify sRNA pairs that have the potential to complement, the sequences of the piRNAs must be considered for compatibility. TruePaiR uses sequence complementarity to detect read pairs that are likely to facilitate Ping-Pong amplification in sRNA high throughput sequencing data.

Section 4.3: Materials and Methods

Workflow

sRNA reads are the required input for TruePaiR in FASTA or FASTQ format. piRNAs are distinguished from other sRNAs using a length threshold greater than 23 nucleotides. Under the current model of Ping-Pong amplification loop, piRNA base pairs 11 and beyond don't facilitate complementarity. Therefore, piRNA reads are trimmed to include only the first ten base pairs of the piRNAs. piRNA reads are then binned by those exhibiting only a U at the first position, those exhibiting only an A at the tenth position, and those exhibiting both a U at the first position and an A at the tenth position. Reads without a piRNA signature are not considered in assigning sRNA read pairs.

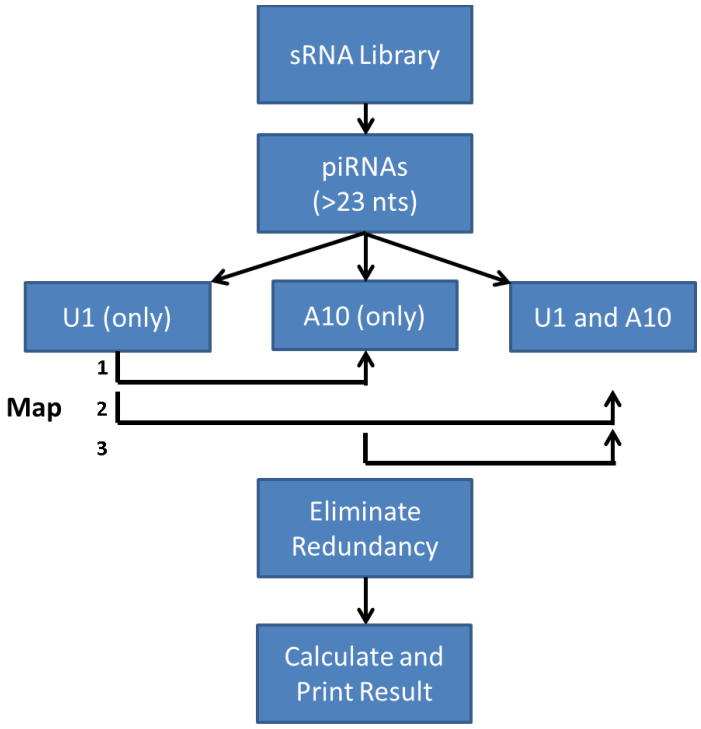


Figure 4.1 | TruePaiR Workflow. A depiction of the steps that are utilized in the assessment of relative piRNA amplification using TruePaiR.

Mapping is then performed using Bowtie2 on the three partitions to predict complementary piRNA reads: (1) U-1 piRNAs as the subject and A-10 piRNAs as the reference set (2) U-1 piRNAs as the subject and piRNAs with both U-1 and A-10 as the reference set, and (3) A-10 piRNAs as the subject and piRNAs with both U-1 and A-10 as the reference set (Figure 1).

A verbose option is available to maintain intermediate files for downstream analysis. The intermediate files can give insight into the only U-1, only A-10, and both U-1 and A-10 subsets, as well as the mapping metadata associated with the TruePaiR run.

The three resulting SAM files, from each mapping run, are appended into a single file. Redundancy is removed within the file to be certain that each read is associated or not associated with a pair a maximum of one time. That is, each piRNA has a binary state in the TruePaiR algorithm: zero if the piRNA has no piRNA complement and one if the piRNA has at least one piRNA complement.

TruePaiR reports metrics regarding the number of piRNAs and the percentage of piRNAs with a U at the first position, an A at the tenth position, piRNAs that have a *possible* piRNA complement (0-2 mismatches), and piRNAs that have a *perfect* piRNA complement.

Section 4.4: Results

Software Performance

TruePaiR is written and executed using R software. When the number of piRNA reads was under ten million, TruePaiR consistently completed under 10 minutes. However, the timing of TruePaiR runs can vary based upon library size and degree of complementarity of the piRNAs.

Benchmarking Degree of piRNA Amplification

In order to appropriately interpret the TruePaiR output, I established benchmark values based on five tissues within seven species that have known or implicated activity in piRNA amplification (Brennecke et al. 2007; Aravin et al. 2007). Benchmarking using model organisms, and tissues with known piRNA pathway activity, serves to assess the ability of TruePaiR to make *a posteriori* assessments regarding the utilization of the Ping-Pong pathway using piRNA reads.

Variability of TruePaiR metrics was relatively low amongst same tissue samples within the same species. On average in the species and tissues observed in this analysis, the standard deviation relative to the observed values for U-1 presence in piRNAs was 9.5%, A-10 presence was 7.8%, possible pairs was 11.3%, and perfect pairs was 24.7% (Figure 4.2). The numbers of piRNAs in each library varied from 269,297 to 77,761,751 with an average of 9,498,527 and median of 4,674,392 piRNAs per library.

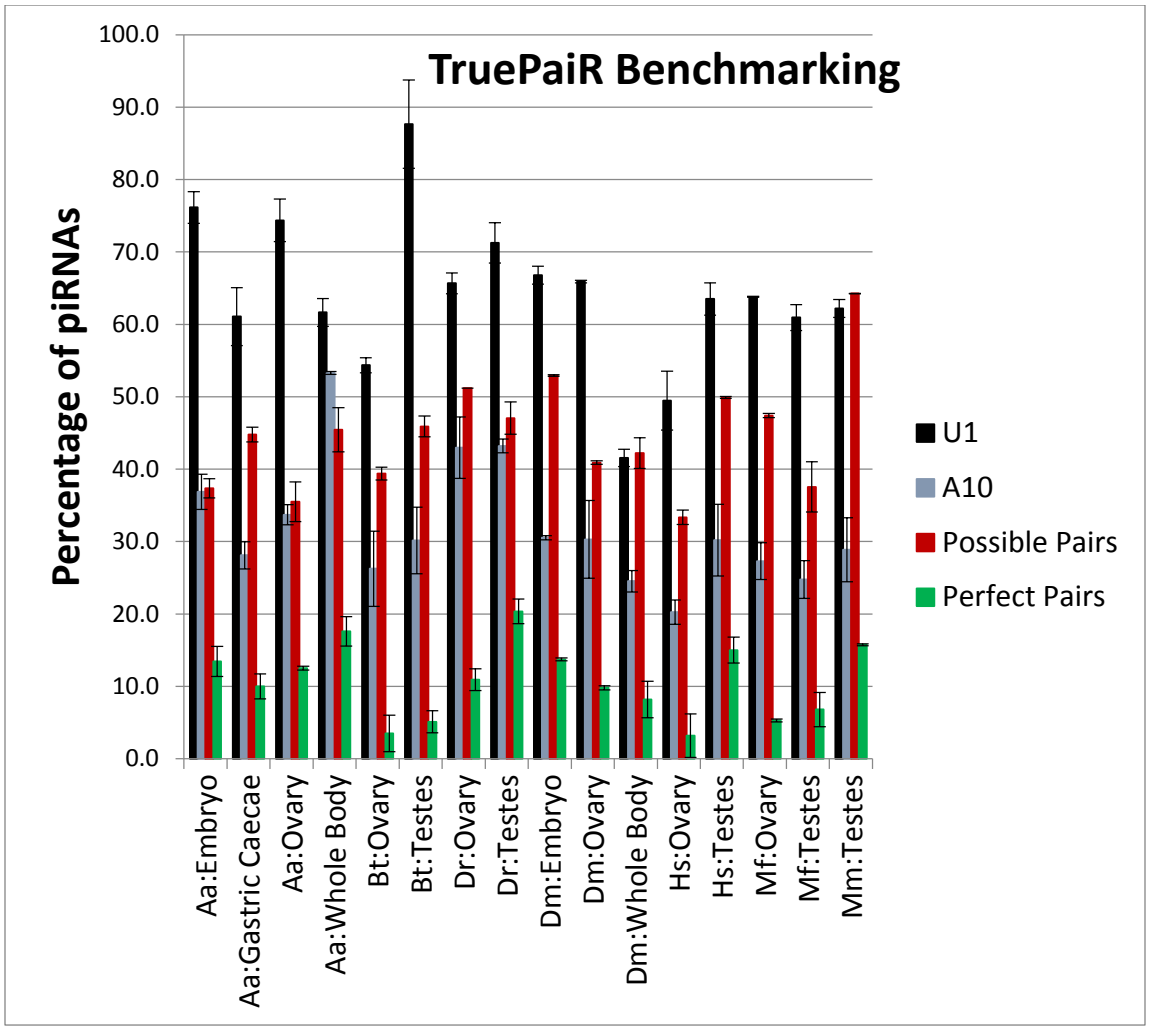


Figure 4.2 | TruePaiR Benchmarking. Representation of the TruePaiR metrics oberseved in seven species and five tissues with replicate libraries. TruePaiR metrics include the percentage of piRNAs with a U at the first position (black), an A at the tenth position (blue), at least one possible piRNA complement (0-2 mismatches - red), and at least one perfect piRNA complement (green).

Species-specific piRNA Amplification

TruePaiR metrics within the same tissue highlighted fundamental differences in piRNA populations and amplification among Metazoans. In ovarian tissue, the degree of U-1 bias within the piRNAs did not significantly differ between *D. melanogaster, D. rerio*, and *M. fascicularis*. *A. aegypti* had a significantly higher, while *B. Taurus* and *H. sapiens* had a significantly lower degree of U-1 bias in the piRNAs. The degree of A10 bias differed significantly between *D. rerio*, *A. aegypti*, *D. melanogaster*, *M. fascicularis*, *B. taurus*, and *H. sapiens* from greatest to least in A-10 representation amongst the piRNAs.

piRNA amplification, defined with indefinite sequence complementarity, was not significantly different between *A. aegypti*, *B. taurus*, *D. melanogaster*, and *H. sapiens*. The potential for imperfect piRNA complementarity was significantly higher in *D. rerio* and *M. fascicularis*. piRNA amplification with perfect sequence complementarity was significantly lower in each species relative to imperfect pairing. *A. aegypti*, *D. rerio*, and *D. melanogaster* had significantly more potential for perfect piRNA complements relative to *B. taurus* and *H. sapiens* (Figure 4.3A).

In testes tissue, the degree of U-1 bias did not significantly differ between *H. sapiens*, *M. fascicularis*, and *M. musculus*. piRNAs from *B. taurus* testes were significantly higher in the degree of U-1 bias relative to the other species observed. The U-1 bias varied greatly (59.3%) *M. fascicularis* testes piRNAs. The degree of A-10 bias did not significantly differ between *B. taurus*, *Homo sapiens*, *M. fascicularis*, and *M. musculus* testes samples.

Possible piRNA amplification was not significantly different in *B. taurus*, *D. rerio*, and *H. sapiens*. *M. fasciularis* had significantly less, while *M. musculus* had significantly more potential for possible piRNA pairs. Perfect complementarity in facilitating piRNA amplification varied significantly in testes samples across species. *D. rerio*, *M. musculus*, *H. sapiens*, *M. fascicularis*, and *B. taurus* demonstrated the greatest to least potential for perfect piRNA pairs (Figure 4.3B).

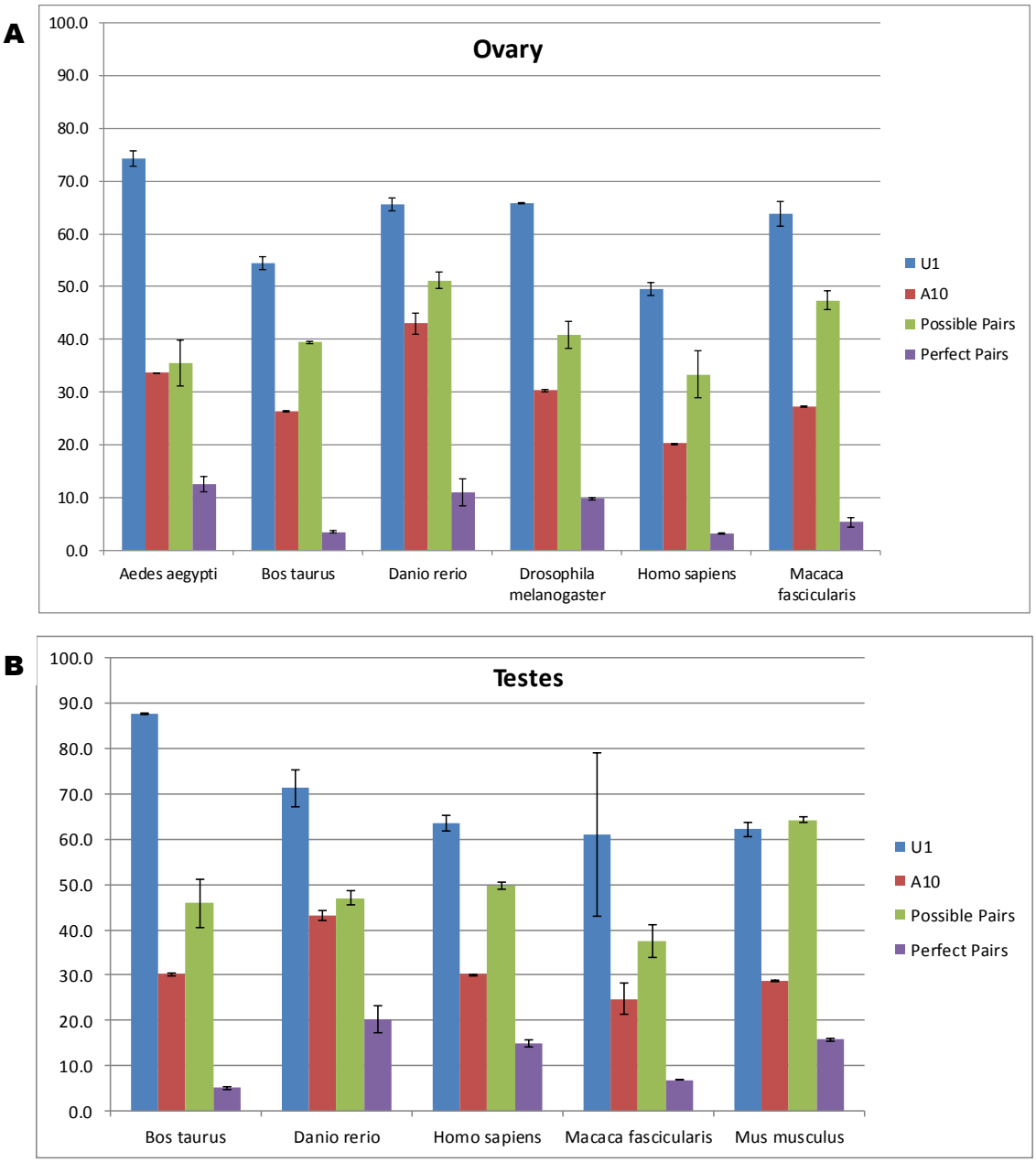


Figure 4.3 | Species-specific Degree of piRNA Amplification. Histogram of the TruePaiR results for **(A)** ovary and **(B)** testes samples across species.

Tissue-specific piRNA Amplification

The observation of TruePaiR metrics regarding relative piRNA amplification within different tissues of the same species allows for an objective assessment of piRNA amplification between tissues.

In *A. aegypti*, consistent with previous research in germline and somatic piRNA amplification, ovary and embryo tissue demonstrated a significantly higher degree of U-1 bias in the piRNAs relative to gastric caecae and whole body (Brennecke et al. 2007; Aravin et al. 2007). The degree of A-10 piRNA bias differed significantly between whole body, embryo, ovary, and gastric caecae from greatest to least. The number of perfect and imperfect possible pairs did not significantly differ bwteen the tissue (Figure 4.4A).

In *B. taurus*, testes demonstrated a significantly higher degree of U-1 bias, A-10 bias, possibility of imperfect pairs, and possibility of perfect pairs relative to ovarian tissue (Figure 4.4B).

In *D. rerio*, ovarian tissue exhibited a significantly greater degree of U-1 bias relative to testes. No significant difference was observed between the degree of A-10 bias between tissues. Ovarian tissue demonstrated a higher potential of imperfect piRNA pairs, but less of a potential for perfect pairs relative to testes (Figure 4.4C).

In *H. sapiens*, testes tissue exhibited a significantly higher degree of U-1 bias, A-10 bias, possibility of imperfect pairs, and possibility of perfect pairs relative to ovarian samples (Figure 4.4D).

In *M. fascicularis*, no significant difference was observed in the degree of U-1 nor A-10 bias. Ovarian tissue demonstrated a significantly greater degree of possible piRNA complements, but a significantly lower degree of perfect piRNA complements (Figure 4.4E).

In *M. musculus*, replicate libraries of piRNAs from ovarian tissue were not available to assess deviation between samples. However, piRNAs from ovarian tissue demonstrated a higher degree U-1 bias relative to testes samples. The difference in the degree of A-10 piRNA bias and imperfect pairing was consistent between tissues. Although, testes piRNAs exhibited a higher proportion of perfect piRNA pairs relative to ovarian tissue (Figure 4.4F).

piRNA Origin and Relative Amplification

Further, observing subsets of the piRNAs in a particular library by their sequence of origin can provide insight into the nature of the piRNAs that are facilitating piRNA amplification. All piRNAs represent metrics gathered from sRNA reads greater than 23 nucleotides in length. TE-derived piRNAs were determined by piRNA homology to TEs available in the RepBase database (Jurka et al. 2005). Gene-derived piRNAs were determined by transcript reference datasets respective to the species under observation. Virus sequences were extracted from the NCBI database (Sayers et al. 2011).

Relative amplification based upon piRNA origin varied greatly between species and tissues. However, TE- and gene-derived piRNAs were consistently a more prevalent

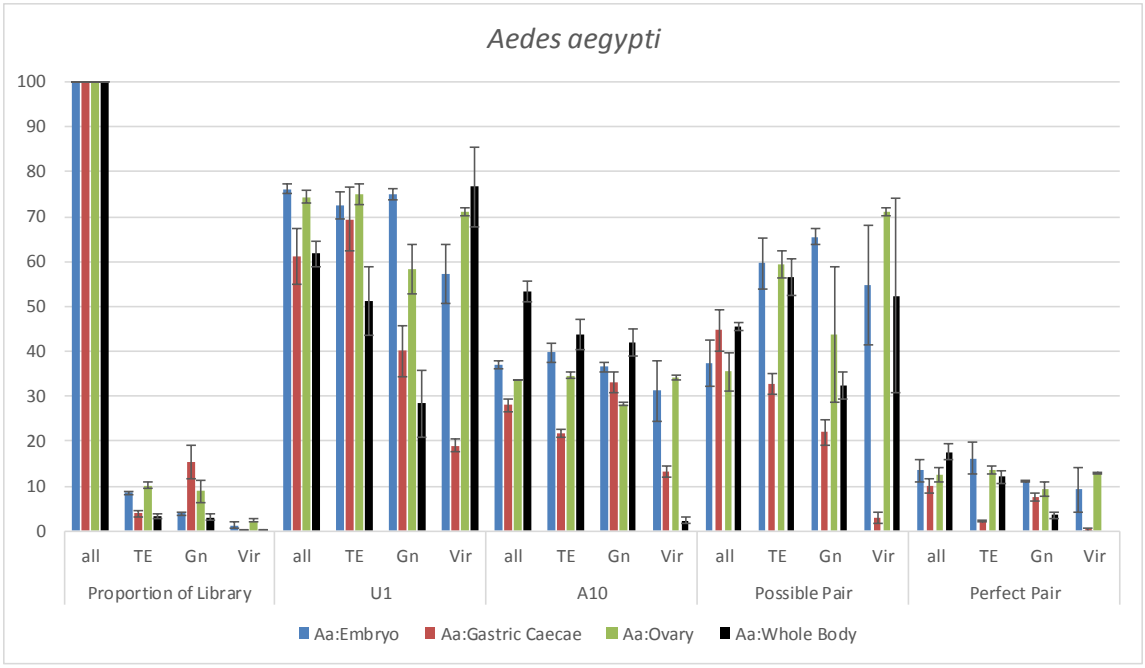
subset of the total piRNA population and had a higher capability of participation in piRNA amplification in the species and tissues observed relative to viral-derived piRNAs (Figure 4.4).

Application in piRNA Pathway Knockdowns (KDs)

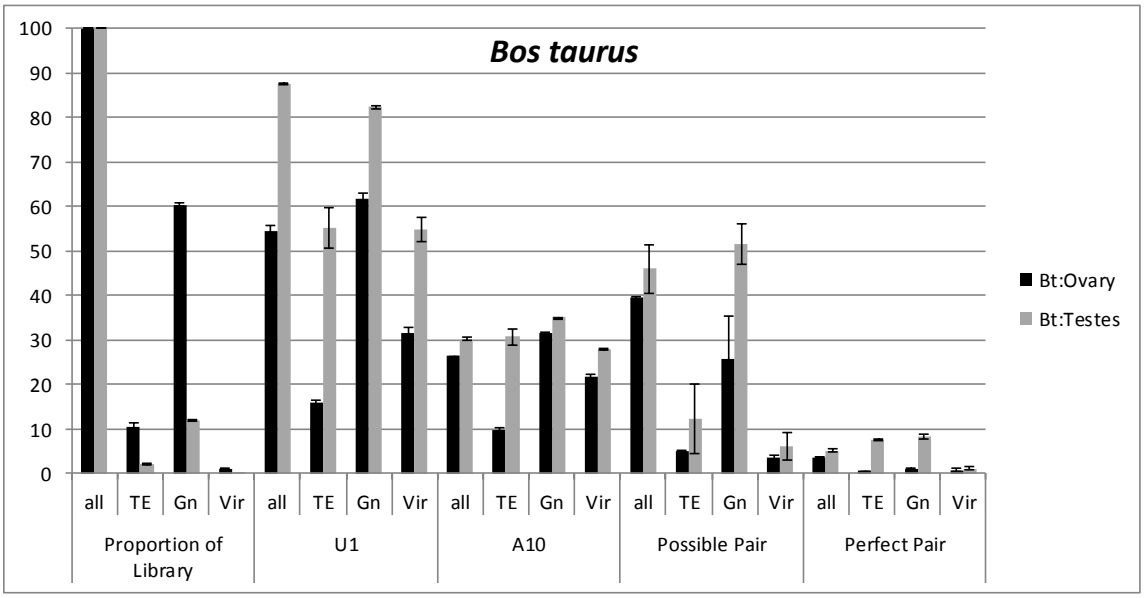
Data available within the Short Read Archive and Gene Expression Omnibus is available for heterozygous and knockdown conditions of several proteins that are critical in promoting piRNA amplification (Leinonen et al. 2010; Edgar et al. 2002). Most notably, I observed knockdown effect on the piRNA populations of Piwi, Aubergine, Zucchini, and Argonaute 3.

Although replicate libraries were not available to establish statistical significance, a distinction can be noted between heterozygous and knockdown libraries in *Drosophila melanogaster* ovaries (Malone et al. 2009). In Piwi KD, the proportion of piRNA populations with a U-1 bias was higher by 45.4%, A-10 bias was lower by 18.4%, possible pairs was higher by 44.2%, and perfect pairs was higher by 3.0% in control libraries relative to KD. In Aubergine KD, the proportion of piRNA populations with a U-1 bias increased by 41.3%, A-10 bias decreased by 23.0%, possible piRNA pairs increased by 11.9%, and perfect piRNA pairs increased by 6.6%. In Zucchini KD, the proportion of piRNA populations with a U-1 bias increased by 6.5%, A-10 bias increased by 1.9%, possible piRNA pairs increased by 3.4%, and perfect piRNA pairs increased by 5.1% (Figure 4.5).

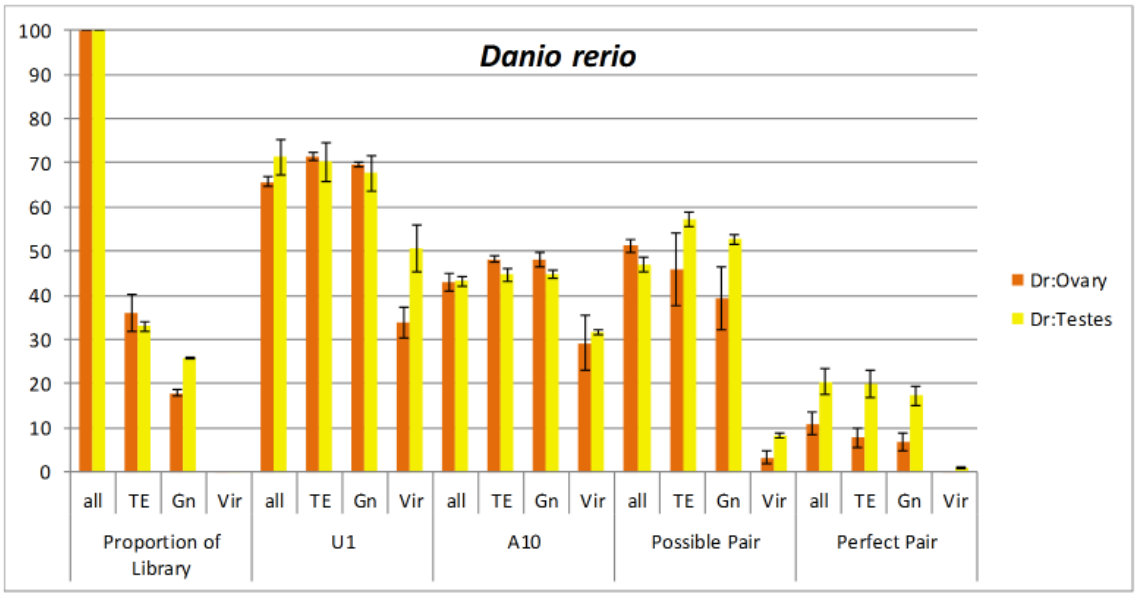
A


Aedes aegypti
100
90
80
70
60
50
40
30
20
10
0
all TE Gn Vir all TE Gn Vir all TE Gn Vir all TE Gn Vir all TE Gn Vir
Proportion of Library
U1
A10
Possible Pair
Perfect Pair
Aa:Embryo
Aa:Gastric Caecae
Aa:Ovary
Aa:Whole Body

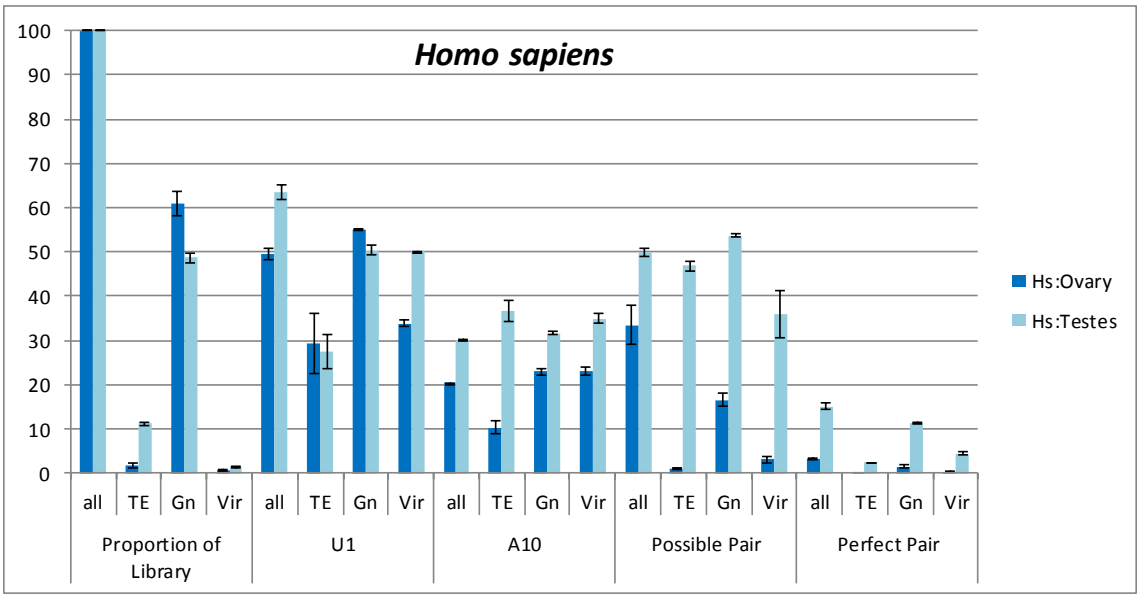
B


Bos taurus
100
90
80
70
60
50
40
30
20
10
0
all TE Gn Vir all TE Gn Vir all TE Gn Vir all TE Gn Vir all TE Gn Vir
Proportion of Library
U1
A10
Possible Pair
Perfect Pair
Bt:Ovary
Bt:Testes

C


Danio rerio
100
90
80
70
60
50
40
30
20
10
0
all
TE
Gn
Vir
Proportion of Library
U1
A10
Possible Pair
Perfect Pair
Dr:Ovary
Dr:Testes

D


Homo sapiens
100
90
80
70
60
50
40
30
20
10
0
all
TE
Gn
Vir
Proportion of Library
U1
A10
Possible Pair
Perfect Pair
Hs:Ovary
Hs:Testes

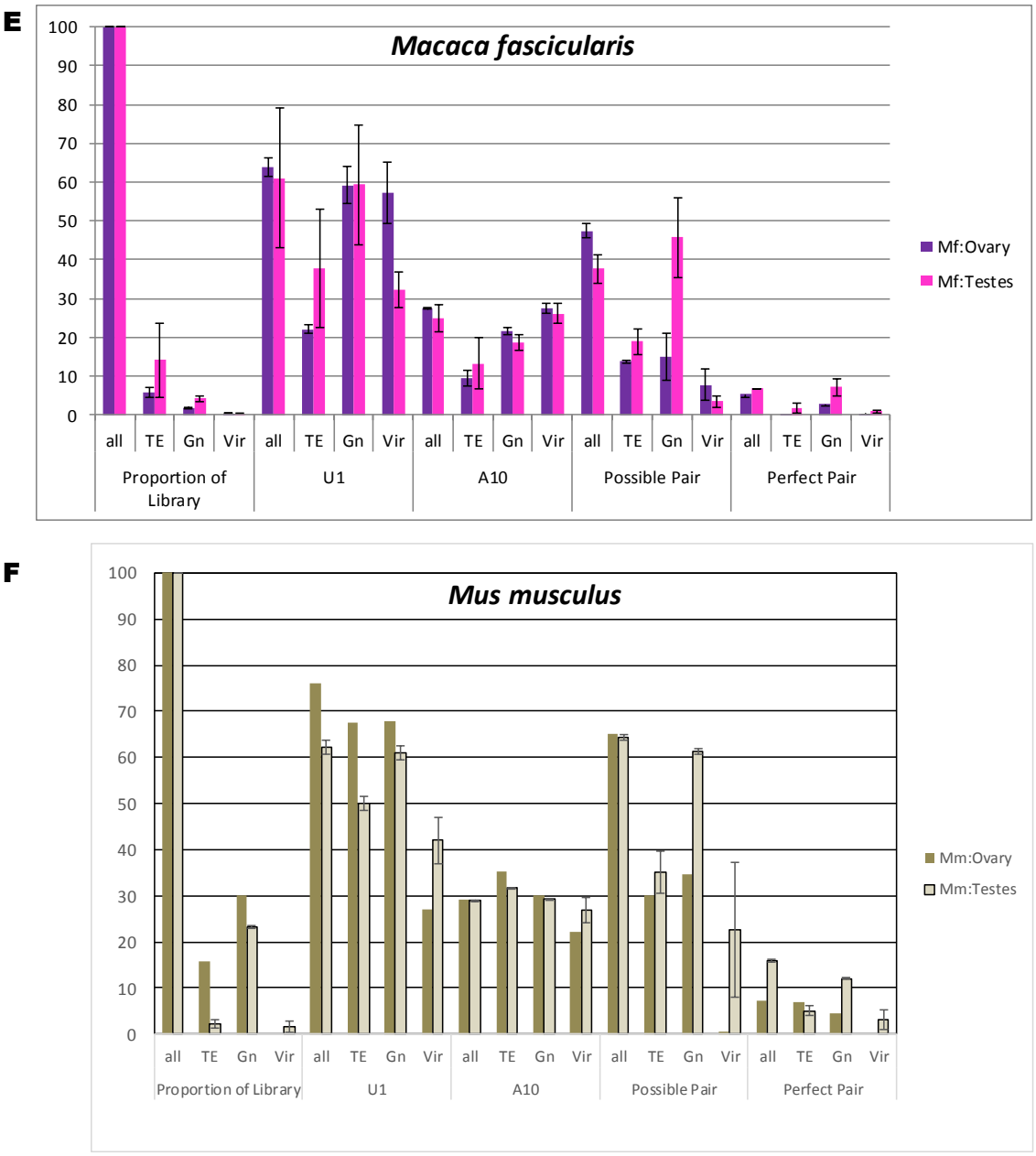


Figure 4.4 | Tissue-Specific Degree of piRNA Amplification. Histogram representing the differential proportions of piRNAs with U-1, A-10, a possible piRNA pair, and a perfect piRNA pair within ovary and testes tissues of **(A)** *Aedes aegypti* **(B)** *Bos taurus* **(C)** *Danio rerio* **(D)** *Homo sapiens* **(E)** *Macaca fascicularis* and **(F)** *Mus musculus*.

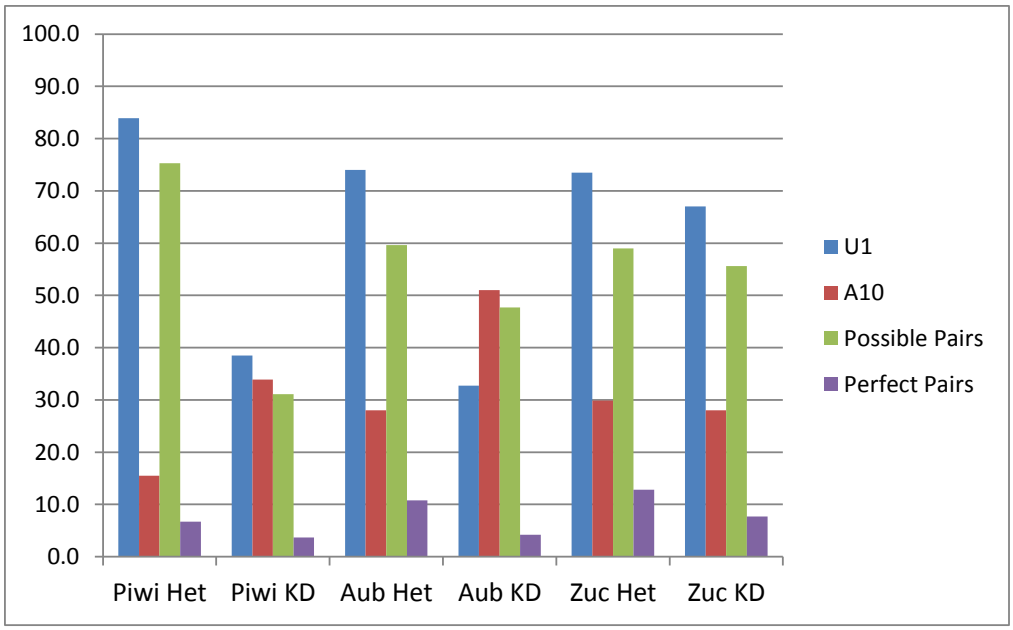


Figure 4.5 | Relative piRNA Amplification in Available Knockdown Libraries. Observation of TruePaiR metrics in publicly available Piwi, Aubergine (Aub), and Zucchini (Zuc) heterozygous and knockdown ovary in *Drosophila melanogaster* (Malone et al. 2009).

Further, piRNAs from male and female fourth instar larvae in *Aedes aegypti* in the presence and absence of AGO3 (Han and Atkinson, unpublished). The model of piRNA biogenesis suggests that AGO3 is a critical protein involved in the promotion of piRNA amplification (Brennecke et al. 2007). Upon successful AGO3 KD in males, a reproducible and statistically significant difference is observed in the degree of piRNA amplification relative to uninduced male fourth instar larvae. However, due to an explanation that is still under investigation, AGO3 transcript was not suppressed in female fourth instar larvae upon induction (Han and Atkinson, unpublished). TruePaiR detected no significant difference of piRNA amplification in female fourth instar larvae (Figure 4.6).

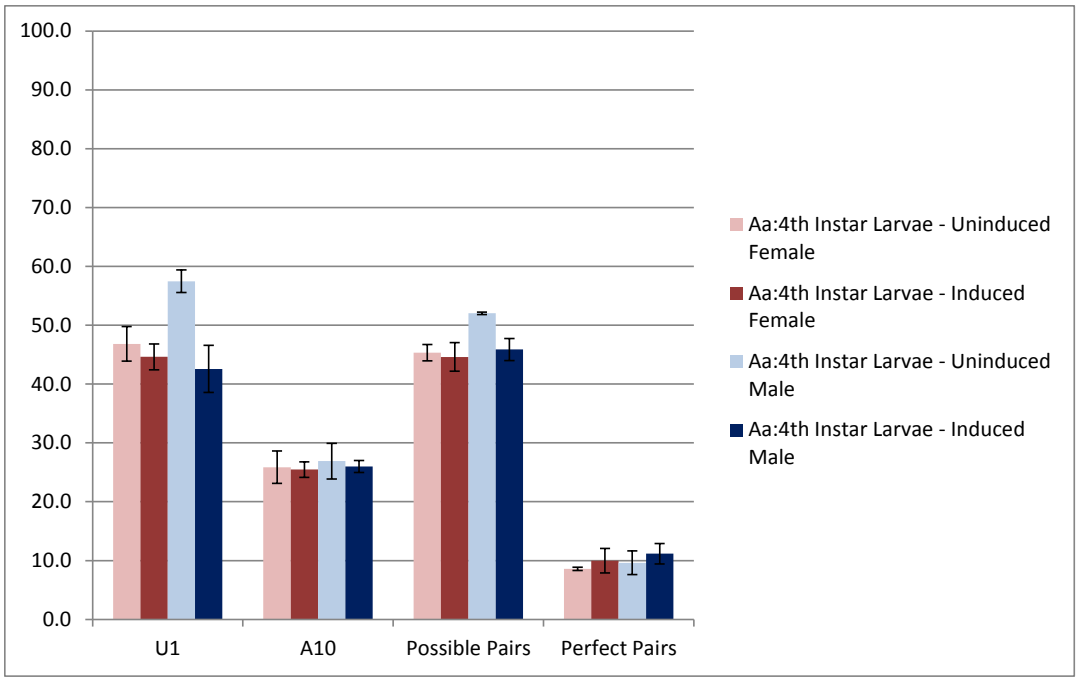


Figure 4.6 | Relative piRNA Amplification in Ago3 Knockdown. TruePaiR runs in *Aedes aegypti* fourth instar larvae males and females in the presence (uninduced) and absence (induced) of Ago3.

Section 4.5: Discussion

The presence of U-1 and A-10 bias within piRNA populations is an indicator, but not an absolute measure of piRNA amplification. By further considering imperfect and perfect sequence complementarity within the first ten base pairs of piRNAs, the active site promoting secondary piRNA biogenesis, I developed practical and statistically powerful metrics to observe relative piRNA amplification (Brennecke et al. 2007).

TruePaiR is a fast and effective tool to determine the relative utilization of the piRNA amplification using high-throughput sRNA sequencing data. TruePaiR is an effective and robust gauge of the piRNA amplification using a single sRNA subpopulation.

The TruePaiR results were accurate in detecting relatively high piRNA amplification in tissue, such as ovary and testes, that has been well-studied to participate in the piRNA pathway. Further, TruePaiR was able to detect piRNA amplification in *Aedes aegypti* gastric caecae and fourth instar larvae, which are tissues that have not been previously known to utilize this pathway (Han and Atkinson, unpublished). TruePaiR was also capable of detecting low levels of piRNA amplification in species, tissues, and piRNA subsets in which the secondary pathway of piRNA biogenesis has little or no activity.

Established benchmark values, in model species and tissues known to undergo piRNA amplification, allow for the observation of meaningful context of the TruePaiR metrics for species or tissue in which the degree of piRNA amplification is not well-understood (Figure 4.2-4.3). The results presented herein provide foundational data regarding piRNA amplification in terms of species specificity, tissue specificity, as well as the relative participation based upon piRNA origin. General trends in the proportion of U-1 piRNAs, A-10 piRNAs, and number of piRNA complements are consistent with conserved model of piRNA biogenesis via the amplification loop (Brennecke et al. 2007). The TruePaiR benchmark values characterize the difference in relative piRNA amplification, which can lead to downstream experimentation to identify species- or tissue-specific factors that affect piRNA biogenesis.

Sample variation was minor in independent sRNA samples of the same tissue. The detected differences in the TruePaiR metrics between species and tissues may be due to species-specific factors that facilitate or inhibit piRNA amplification, the number of active piRNA clusters, the number of generated piRNAs, or the sequence content of

generated piRNAs (Figure 4.2-4.3). Even considering the innate variability between organisms and library preparations, the results of TruePaiR were very consistent in assessing piRNA amplification in particular species and tissues. Consistency across same sample TruePaiR runs allows for a reliable and reproducible assessment of relative piRNA amplification.

TruePaiR demonstrated capability to detect differences in relative piRNA amplification between conditions. Differences were observed between hetereozygous and knockdown sRNA libraries of Piwi, Aub, and Zucchini, in a similar magnitude as previously described, while providing specific metrics regarding the effects of each particular knockdown (Malone et al. 2009) (Figure 4.5). Further, the TruePaiR metrics of possible piRNA pairs was capable of distinguishing, with both reproducibly and statistical significance, minor differences in piRNA amplification in triplicate sRNA libraries of *Aedes aegypti* fourth instar larvae upon Ago3 control and knockdown (Han and Atkinson, unpublished) (Figure 4.6).

The TruePaiR results showed consistently low levels of perfect piRNA pairs within the first ten base pairs, even in tissue that are known to have the highest levels of piRNA amplification. The relative proportion of possible piRNA pairs, allowing up to two mismatches in the first ten base pairs, increased significantly in germline tissue known to be involved in piRNA amplification. (Figure 4.4). These results support a piRNA amplification model of imperfect complementarity in the first ten base pairs of piRNA complements.

Given that TruePaiR serves as an effective and consistent metric of piRNA amplification across species, it can represent a new, meaningful standard in the degree of piRNA amplification in a specific organism and tissue that is or is not expected to undergo piRNA amplification.

Section 4.6: References

Aravin, A.A., Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (80-.).*, 318(5851), pp.761–764.

Arensburger, P. et al., 2011. The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics*, 12(1), p.606.

Brennecke, J. et al., 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science (New York, NY)*, 322(5906), p.1387.

Brennecke, J. et al., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6), pp.1089–103.

Edgar, R., Domrachev, M. & Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1), pp.207–210.

Grimson, A. et al., 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217), pp.1193–1197.

Grivna, S.T., Pyhtila, B. & Lin, H., 2006. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proceedings of the National Academy of Sciences*, 103(36), pp.13415–13420.

Holbrook, S.R., Cheong, C. & others, 1991. Crystal Structure of an RNA Double Helix Incorporating a Track of Non-Watson-Crick Base Pairs. *Nature*, 353(6344), p.579.

Horwich, M.D. et al., 2007. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Current Biology*, 17(14), pp.1265–1272.

Jurka, J. et al., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4), pp.462–467.

Leinonen, R., Sugawara, H. & Shumway, M., 2010. The sequence read archive. *Nucleic Acids Res.*, p.gkq1019.

Lin, H. & Yin, H., 2008. A novel epigenetic mechanism in Drosophila somatic cells mediated by Piwi and piRNAs. In *Cold Spring Harbor symposia on quantitative biology*. p. sqb–2008.

Malone, C. & Hannon, G., 2010. Molecular evolution of piRNA and transposon control pathways in Drosophila. In *Cold Spring Harbor symposia on quantitative biology*. p. sqb–2009.

Malone, C.D. et al., 2009. Specialized piRNA Pathways Act in Germline and Somatic Tissues of the< i> Drosophila</i> Ovary. *Cell*, 137(3), pp.522–535.

Malone, C.D. & Hannon, G.J., 2009. Small RNAs as guardians of the genome. *Cell*, 136(4), pp.656–668.

Murchison, E.P. & Hannon, G.J., 2004. miRNAs on the move: miRNA biogenesis and the RNAi machinery. *Curr. Opin. Cell Biol.*, 16(3), pp.223–229.

Nishimasu, H. et al., 2012. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature*, 491(7423), pp.284–7.

Rosenkranz, D. & Zischler, H., 2012. proTRAC-a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics*, 13, p.5.

Saito, K. et al., 2007. Pimet, the Drosophila homolog of HEN1, mediates 2′-O-methylation of Piwi-interacting RNAs at their 3′ ends. *Genes \& development*, 21(13), pp.1603–1608.

Sayers, E.W. et al., 2011. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 39(suppl 1), pp.D38–D51.

Schreiner, P.A. & Atkinson, P., 2017. piClusterBusteR: Software For Automated Classification And Characterization Of piRNA Cluster Loci. *bioRxiv*, p.133009.

Schwarz, D.S., Tomari, Y. & Zamore, P.D., 2004. The RNA-induced silencing complex is a Mg 2+-dependent endonuclease. *Current Biology*, 14(9), pp.787–791.

Tolia, N.H. & Joshua-Tor, L., 2007. Slicer and the argonautes. *Nature chemical biology*, 3(1), pp.36–43.

Wang, W. et al., 2014. The initial uridine of primary piRNAs does not create the tenth adenine that is the hallmark of secondary piRNAs. *Mol. Cell*, 56(5), pp.708–716.

Yang, Z. et al., 2006. HEN1 recognizes 21-24 nt small RNA duplexes and deposits a methyl group onto the 2′ OH of the 3′ terminal nucleotide. *Nucleic Acids Res.*, 34(2), pp.667–675.

Zamore, P.D., 2010. Somatic piRNA biogenesis. *The EMBO journal*, 29(19), pp.3219–21.

Zanni, V. et al., 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences*, 110(49), pp.19842–19847.

Zhang, P. et al., 2014. piRBase: a web resource assisting piRNA functional study. *Database*, 2014, p.bau110.

Zhang, Z. et al., 2011. Heterotypic piRNA Ping-Pong requires qin, a protein with both E3 ligase and Tudor domains. *Mol. Cell*, 44(4), pp.572–84.

Section 4.7: Supplemental Material

Species	Tissue	Library ID	Subset	Number of Reads	Perc_U1	Perc_A10	Perc_Pos	Perc_Perf
Drosophila melanogaster	Whole Body - Male	GSM286602	pis	980097	44.3	21.3	43.5	3.7
Drosophila melanogaster	Whole Body - Male	GSM399107	pis	4299126	43.2	20.3	44.1	5.4
Drosophila melanogaster	Whole Body - Female	GSM286603	pis	397534	39.3	20.5	32.6	3.2
Drosophila melanogaster	Whole Body - Female	GSM399106	pis	3887244	37.1	21.1	41.8	10
Drosophila melanogaster	Whole Body	RNAlib14	pis	10622157	46.3	38	42.9	14.5
Drosophila melanogaster	Whole Body	GSM811191	pis	1239278	39.1	25.9	48.4	12.3
Drosophila melanogaster	Ovary	GSM280082	pis	987689	66	30.7	44.5	10.1
Drosophila melanogaster	Ovary	GSM379050	pis	1841437	65.8	29.9	37.3	9.5
Drosophila melanogaster	Testes	GSM399106	pis	522848	68.7	13.9	9	1.3
Drosophila melanogaster	Embryo	GSM2186328	pis	965390	61.9	30.3	64.4	17.8
Drosophila melanogaster	Embryo	GSM2186329	pis	9834482	67.8	30.8	47.5	12.4
Drosophila melanogaster	Embryo	GSM2186330	pis	8628247	70.6	30.5	46.9	11
Drosophila erecta	Ovary	GSM379301	pis	3090905	61.4	22.2	52.5	9.8
Drosophila yakuba	Ovary	GSM1528802	pis	10547877	42.1	19.8	40	8.7
Aedes aegypti	Orlando - Whole Body (RNAlib1)		pis	8889695	61.5	45.6	47	17.7
Aedes aegypti	Orlando - Whole Body (RNAlib2)		pis	9998589	60.9	57.1	42.5	17.6
Aedes aegypti	Orlando - Whole Body (RNAlib4)		pis	8181467	72.1	50.8	45.8	12.8
Aedes aegypti	Orlando - Whole Body (RNAlib6)		pis	9351484	58	54.5	47	20.6
Aedes aegypti	Orlando - Whole Body (RNAlib10)		pis	7875769	56.2	57.4	43.8	15
Aedes aegypti	Orlando - Whole Body (RNAlib11)		pis	10333029	61.2	54.3	46.6	21.9
Aedes aegypti	Orlando - Embryos (RNAlib16)		pis	31738783	74.7	38.1	44.7	17

Aedes aegypti	Liverpool - Embryos (RNAlib21)		pis	51687336	77.6	35.6	30	9.9
Aedes aegypti	Liverpool - Ovaries (RNAlib17)		pis	26936731	72.3	33.7	41.5	14.6
Aedes aegypti	Orlando - Ovaries (RNAlib18)		pis	77761751	76.4	33.7	29.5	10.4
Aedes aegypti	Orlando - Gastric Caecae 1		pis	39069830	63.9	31.2	34.3	8.7
Aedes aegypti	Orlando - Gastric Caecae 2		pis	7676105	71.6	27.6	51.7	7.8
Aedes aegypti	Orlando - Gastric Caecae 3		pis	6682523	47.7	25.5	48.3	13.5
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 1		pis	1129798	54.9	32.3	52.5	7.2
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 2		pis	3698077	55.6	20.3	51.7	7.4
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 3		pis	8863156	61.9	28.1	51.9	14.3
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 1		pis	1148271	44.7	24.1	45.2	7.2
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 2		pis	5556466	49.3	25.8	49.9	13.4
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 3		pis	2017646	33.7	28.1	42.5	12.9
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 1		pis	664541	53	31.6	42.2	9.2
Aedes aegypti	Orlando - Uninduced -		pis	1294657	41.3	25.4	47.4	8.2

	Female - 4th instar Larvae 2							
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 3		pis	5049658	46.2	20.6	46.4	8.4
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 1		pis	814562	39.6	28.1	39.1	5.4
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 2		pis	6817664	46.6	22.8	46.4	11
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 3		pis	6682523	47.7	25.5	48.3	13.5
Homo sapiens	Brain	GSM2257348	pis	479170	11.4	5.5	1.2	0
Homo sapiens	Ovary	GSM1584521	pis	1112698	51.2	20.3	27.1	3
Homo sapiens	Ovary	GSM1584522	pis	1448615	47.7	20.2	39.6	3.4
Homo sapiens	Testes	SRR835324	pis	15759157	61.2	30.1	51	16
Homo sapiens	Testes	SRR835325	pis	16564549	65.8	30.3	48.8	14
Mus musculus	Ovary	SRR014234	pis	269297	76.2	29.3	65.2	7.2
Mus musculus	Testes	GSM802671	pis	2630008	64.4	29	65.2	15.3
Mus musculus	Testes	GSM802674	pis	2699611	60	28.7	63.3	16.2
Macaca fascicularis	Ovary	GSM1584515	pis	3730190	60.4	27.1	44.8	4
Macaca fascicularis	Ovary	GSM1584516	pis	7866404	67.2	27.5	50	6.6
Macaca fascicularis	Testes	GSM1584519	pis	20045731	86.5	29.7	42.6	6.8
Macaca fascicularis	Testes	GSM1584520	pis	2562594	35.4	19.8	32.5	6.8
Bos taurus	Ovary	GSM1584503	pis	3386440	52.6	26.4	39.8	3.8
Bos taurus	Ovary	GSM1584504	pis	2091771	56.1	26.1	39	3.2
Bos taurus	Testes	GSM1584507	pis	27396744	87.4	29.8	38.3	5.5
Bos taurus	Testes	GSM1584508	pis	2679619	87.9	30.5	53.5	4.7
Danio rerio	Ovary	SRR578904	pis	403044	68.4	46.7	54.6	6.4

Danio rerio	Ovary	SRR578905	pis	3776787	63.8	38.4	49.2	10
Danio rerio	Ovary	SRR578906	pis	9446136	64.8	43.8	49.8	16.4
Danio rerio	Testes	SRR578922	pis	24199051	77	41.8	44.7	16.1
Danio rerio	Testes	SRR578923	pis	23599603	65.5	44.6	49.4	24.6
Drosophila melanogaster	Whole Body - Male	GSM286602	tedev	105824	37.6	21.2	19.6	0.4
Drosophila melanogaster	Whole Body - Male	GSM399107	tedev	1034742	33.5	23.5	24	1.1
Drosophila melanogaster	Whole Body - Female	GSM286603	tedev	31373	29.2	22.3	8	0.4
Drosophila melanogaster	Whole Body - Female	GSM399106	tedev	595920	30.9	26	28.2	3.9
Drosophila melanogaster	Whole Body	RNAlib14	tedev	15148	17.8	40.3	20.5	0.3
Drosophila melanogaster	Whole Body	GSM811191	tedev	322940	39.2	25.8	40	7.8
Drosophila melanogaster	Ovary	GSM280082	tedev	498534	58.3	30.3	41.9	12.8
Drosophila melanogaster	Ovary	GSM379050	tedev	876258	52.1	32.8	40.4	8.1
Drosophila melanogaster	Testes	GSM399106	tedev	108679	35.7	21.5	2.8	1.6
Drosophila melanogaster	Embryo	GSM2186328	tedev	135545	71.7	31.5	44.9	11
Drosophila melanogaster	Embryo	GSM2186329	tedev	3762481	74.4	32.3	50.7	13.1
Drosophila melanogaster	Embryo	GSM2186330	tedev	3356786	75.9	29.9	49.8	10.4
Drosophila erecta	Ovary	GSM379301	tedev	673148	60.6	24	29.2	7.9
Drosophila yakuba	Ovary	GSM1528802	tedev	7281906	39.8	17	40.6	3.6
Aedes aegypti	Orlando - Whole Body (RNAlib1)		tedev	491092	60.5	37.6	58.2	11.1
Aedes aegypti	Orlando - Whole Body (RNAlib2)		tedev	207946	44.9	53.2	66.3	14.8
Aedes aegypti	Orlando - Whole Body (RNAlib4)		tedev	383740	77.2	37.2	61.3	9
Aedes aegypti	Orlando - Whole Body (RNAlib6)		tedev	264787	41.1	49.8	44.7	15
Aedes aegypti	Orlando - Whole Body (RNAlib10)		tedev	154176	35.3	40.5	48	8.6
Aedes aegypti	Orlando - Whole Body (RNAlib11)		tedev	320011	47.9	44.5	60.5	13.6
Aedes aegypti	Orlando - Embryos (RNAlib16)		tedev	2491299	68.2	42.7	67.5	21.1

Aedes aegypti	Liverpool - Embryos (RNAlib21)		tedev	4679034	76.6	36.8	51.7	11.1
Aedes aegypti	Liverpool - Ovaries (RNAlib17)		tedev	2979590	71.7	35.5	63.4	14.9
Aedes aegypti	Orlando - Ovaries (RNAlib18)		tedev	7070813	78.2	33.9	54.9	12.2
Aedes aegypti	Orlando - Gastric Caecae 1		tedev	1039281	80.7	19.7	37.5	2.3
Aedes aegypti	Orlando - Gastric Caecae 2		tedev	384251	73.7	23.3	32.4	2
Aedes aegypti	Orlando - Gastric Caecae 3		tedev	261073	53.8	22.2	28.3	2.3
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 1		tedev	39394	16.6	31.2	19.4	0.8
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 2		tedev	265839	17	23.5	20.6	0.9
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 3		tedev	461605	65.9	21.6	37.5	2.9
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 1		tedev	49139	50.9	24	18.1	1.2
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 2		tedev	257188	53.2	23.7	33.1	2.5
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 3		tedev	132292	31.9	28.9	23.2	0.6
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 1		tedev	22230	81.8	22.2	14.1	0.7
Aedes aegypti	Orlando - Uninduced -		tedev	18410	59.6	21.5	11.3	0.5

	Female - 4th instar Larvae 2							
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 3		tedev	335095	44.3	18.9	21.3	1.1
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 1		tedev	69392	36.6	26.7	17.1	0.4
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 2		tedev	343381	46	20.7	26.4	2.3
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 3		tedev	261073	53.8	22.2	28.3	2.3
Homo sapiens	Brain	GSM2257348	tedev	348450	5.2	1.9	0	0
Homo sapiens	Ovary	GSM1584521	tedev	7909	39.1	8.4	0.6	0
Homo sapiens	Ovary	GSM1584522	tedev	34985	19.4	12.2	1.1	0.1
Homo sapiens	Testes	SRR835324	tedev	1821331	21.9	40.2	45.3	2.3
Homo sapiens	Testes	SRR835325	tedev	1776535	33.1	33.2	48.3	2.3
Mus musculus	Ovary	SRR014234	tedev	42418	67.5	35.3	30	6.9
Mus musculus	Testes	GSM802671	tedev	26606	52.2	31.5	28.5	3.5
Mus musculus	Testes	GSM802674	tedev	92527	47.8	31.8	41.7	6.6
Macaca fascicularis	Ovary	GSM1584515	tedev	280167	20.5	6.6	14.1	0
Macaca fascicularis	Ovary	GSM1584516	tedev	320716	23.7	12.1	13.2	0.3
Macaca fascicularis	Testes	GSM1584519	tedev	107906	59.2	22.6	23.3	3.5
Macaca fascicularis	Testes	GSM1584520	tedev	709653	16.2	3.8	14.2	0
Bos taurus	Ovary	GSM1584503	tedev	386315	14.8	9.2	5.1	0.5
Bos taurus	Ovary	GSM1584504	tedev	195333	16.6	10.3	4.5	0.3
Bos taurus	Testes	GSM1584507	tedev	578764	48.8	28.2	23.3	7.9
Bos taurus	Testes	GSM1584508	tedev	50969	61.7	33.2	1	7.2
Danio rerio	Ovary	SRR578904	tedev	181079	69.7	47.6	27.4	3.7

Danio rerio	Ovary	SRR578905	tedev	1064417	73.2	47.5	52.1	7.2
Danio rerio	Ovary	SRR578906	tedev	3259572	71.9	49.6	58.1	12.5
Danio rerio	Testes	SRR578922	tedev	8347496	76.5	42.5	54.8	15.4
Danio rerio	Testes	SRR578923	tedev	7439386	64.1	46.6	59.5	24.3
Drosophila melanogaster	Whole Body - Male	GSM286602	gndev	21505	22.3	40.2	30	0.5
Drosophila melanogaster	Whole Body - Male	GSM399107	gndev	145277	24.1	27.8	29.2	2.3
Drosophila melanogaster	Whole Body - Female	GSM286603	gndev	14142	35.7	24.7	13	1.7
Drosophila melanogaster	Whole Body - Female	GSM399106	gndev	152055	30.6	39	29	3.1
Drosophila melanogaster	Whole Body	RNAlib14	gndev	24743	34.8	39.7	24.7	2.1
Drosophila melanogaster	Whole Body	GSM811191	gndev	62323	46.3	28.1	32.6	4.1
Drosophila melanogaster	Ovary	GSM280082	gndev	49650	72.2	27.3	21.9	4.8
Drosophila melanogaster	Ovary	GSM379050	gndev	92060	82.3	22.6	17.4	2.5
Drosophila melanogaster	Testes	GSM399106	gndev	247	74.1	25.9	1.6	0.8
Drosophila melanogaster	Embryo	GSM2186328	gndev	55523	67	33.3	38.9	8.1
Drosophila melanogaster	Embryo	GSM2186329	gndev	524898	78.2	37.4	48.6	9.5
Drosophila melanogaster	Embryo	GSM2186330	gndev	447831	83.3	33.6	49	6.8
Drosophila erecta	Ovary	GSM379301	gndev	2615590	72.6	21.7	56.8	8
Drosophila yakuba	Ovary	GSM1528802	gndev	925038	33.3	16.7	26.5	1.7
Aedes aegypti	Orlando - Whole Body (RNAlib1)		gndev	495876	25.1	35.7	23.9	3.3
Aedes aegypti	Orlando - Whole Body (RNAlib2)		gndev	143896	36.4	52.2	36.8	3.2
Aedes aegypti	Orlando - Whole Body (RNAlib4)		gndev	181333	53.8	44.6	37.7	5.5
Aedes aegypti	Orlando - Whole Body (RNAlib6)		gndev	276354	23.1	40.9	32	2.8
Aedes aegypti	Orlando - Whole Body (RNAlib10)		gndev	229922	11.8	36.1	26.4	1.6
Aedes aegypti	Orlando - Whole Body (RNAlib11)		gndev	361364	20	41.9	37.1	5.3
Aedes aegypti	Orlando - Embryos (RNAlib16)		gndev	1388001	76.7	38	68	11.1

Aedes aegypti	Liverpool - Embryos (RNAlib21)		gndev	1789794	73.2	34.9	62.9	11
Aedes aegypti	Liverpool - Ovaries (RNAlib17)		gndev	1421660	66.2	29	65.2	11.4
Aedes aegypti	Orlando - Ovaries (RNAlib18)		gndev	9677114	50.3	27.6	22.4	6.9
Aedes aegypti	Orlando - Gastric Caecae 1		gndev	6592238	41.2	28.8	16.5	6.7
Aedes aegypti	Orlando - Gastric Caecae 2		gndev	550524	50.9	37.8	27.7	9.4
Aedes aegypti	Orlando - Gastric Caecae 3		gndev	1490320	28.1	32.4	21.7	6.5
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 1		gndev	139166	46.7	49.3	50.2	0.8
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 2		gndev	235249	42.3	20.3	41.4	1.9
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 3		gndev	966669	31.4	43.3	38.8	3.4
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 1		gndev	61791	30	27.3	22.7	1.1
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 2		gndev	443474	29.4	34.9	45.3	3.8
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 3		gndev	265102	30.6	33.6	43.6	3.2
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 1		gndev	50388	49.6	48.3	50.5	0.5
Aedes aegypti	Orlando - Uninduced -		gndev	89513	47.9	38.6	62.1	0.6

	Female - 4th instar Larvae 2							
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 3		gndev	504336	31.4	22	29	3.4
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 1		gndev	30392	41.9	42.1	25.4	0.4
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 2		gndev	474553	36.5	32.4	53.6	18.1
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 3		gndev	745160	28.1	32.4	43.7	13
Homo sapiens	Brain	GSM2257348	gndev	370319	3.3	1.9	1.6	0
Homo sapiens	Ovary	GSM1584521	gndev	721503	54.8	22	14.5	0.9
Homo sapiens	Ovary	GSM1584522	gndev	826360	55.1	23.8	18.5	2
Homo sapiens	Testes	SRR835324	gndev	7901441	48.7	32.1	54.1	11.6
Homo sapiens	Testes	SRR835325	gndev	7821414	51.8	31	53.1	11.2
Mus musculus	Ovary	SRR014234	gndev	80684	67.8	30.1	34.6	4.4
Mus musculus	Testes	GSM802671	gndev	597067	62.7	29.3	62	11.8
Mus musculus	Testes	GSM802674	gndev	637145	58.4	29.2	60.3	12.4
Macaca fascicularis	Ovary	GSM1584515	gndev	58120	52.4	20.1	6.3	1.9
Macaca fascicularis	Ovary	GSM1584516	gndev	162382	66	22.8	23.7	3.5
Macaca fascicularis	Testes	GSM1584519	gndev	993003	81.3	31.4	60.2	10.2
Macaca fascicularis	Testes	GSM1584520	gndev	82322	37.5	25.7	31.1	4
Bos taurus	Ovary	GSM1584503	gndev	2027280	59.9	31.8	39.2	1.1
Bos taurus	Ovary	GSM1584504	gndev	1273483	63.6	31	12.1	0.9
Bos taurus	Testes	GSM1584507	gndev	3201088	82.9	34.8	57.9	8.9
Bos taurus	Testes	GSM1584508	gndev	321802	81.6	35.2	44.9	7.6
Danio rerio	Ovary	SRR578904	gndev	75535	68.7	48.1	27.1	3.1

Danio rerio	Ovary	SRR578905	gndev	620256	70.6	45.1	35.9	6.3
Danio rerio	Ovary	SRR578906	gndev	1746361	69.1	51.4	54.7	11.3
Danio rerio	Testes	SRR578922	gndev	6173634	73.4	43.6	51.1	13.8
Danio rerio	Testes	SRR578923	gndev	6148568	61.9	45.9	54.2	20.5
Drosophila melanogaster	Whole Body - Male	GSM286602	virdev	53406	39.1	27.5	36.6	8.9
Drosophila melanogaster	Whole Body - Male	GSM399107	virdev	256573	44.5	19.9	24.7	0.9
Drosophila melanogaster	Whole Body - Female	GSM286603	virdev	21745	40	14.9	1.2	0.1
Drosophila melanogaster	Whole Body - Female	GSM399106	virdev	204821	25.3	28.9	24.6	0.8
Drosophila melanogaster	Whole Body	RNAlib14	virdev	2514	32.4	8.1	25.4	0
Drosophila melanogaster	Whole Body	GSM811191	virdev	273825	25.9	21.8	30.8	3.7
Drosophila melanogaster	Ovary	GSM280082	virdev	33603	38.1	20.2	12.4	1.1
Drosophila melanogaster	Ovary	GSM379050	virdev	40121	26.8	15.4	0.6	0.1
Drosophila melanogaster	Testes	GSM399106	virdev	18525	52.4	14.8	0.4	0.2
Drosophila melanogaster	Embryo	GSM2186328	virdev	19732	43.3	31.5	30.6	5.9
Drosophila melanogaster	Embryo	GSM2186329	virdev	154253	47.7	29.9	6.5	2.7
Drosophila melanogaster	Embryo	GSM2186330	virdev	229220	53.6	27.9	38.2	3.1
Drosophila erecta	Ovary	GSM379301	virdev	11574	30.8	9.5	0.8	0.1
Drosophila yakuba	Ovary	GSM1528802	virdev	2930	31.5	18.9	0.7	0.4
Aedes aegypti	Orlando - Whole Body (RNAlib1)		virdev	58583	88.6	0.4	85.7	0
Aedes aegypti	Orlando - Whole Body (RNAlib2)		virdev	38495	95.3	1.2	91.7	0
Aedes aegypti	Orlando - Whole Body (RNAlib4)		virdev	8266	89.9	3	87.6	0
Aedes aegypti	Orlando - Whole Body (RNAlib6)		virdev	12634	62	4	0.1	0
Aedes aegypti	Orlando - Whole Body (RNAlib10)		virdev	11626	72.3	2.1	0	0
Aedes aegypti	Orlando - Whole Body (RNAlib11)		virdev	7351	51.3	3.4	48.6	0
Aedes aegypti	Orlando - Embryos (RNAlib16)		virdev	716079	66.5	40.8	73.4	16.3

Aedes aegypti	Liverpool - Embryos (RNAlib21)		virdev	150946	47.8	21.7	35.9	2
Aedes aegypti	Liverpool - Ovaries (RNAlib17)		virdev	724212	69.7	34.8	72.5	12.9
Aedes aegypti	Orlando - Ovaries (RNAlib18)		virdev	1684373	72.3	33.5	69.9	12.8
Aedes aegypti	Orlando - Gastric Caecae 1		virdev	83932	15.7	13.6	5.7	1.3
Aedes aegypti	Orlando - Gastric Caecae 2		virdev	7047	19.6	15.5	2.9	0
Aedes aegypti	Orlando - Gastric Caecae 3		virdev	4324	21.6	10.4	0.8	0
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 1		virdev	959	17.1	20.3	0.7	0
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 2		virdev	8638	13.3	11.6	3.6	1.6
Aedes aegypti	Orlando - Uninduced - Male - 4th instar Larvae 3		virdev	7017	24.1	18.3	3.8	2.3
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 1		virdev	3180	27.9	10.9	0.3	0
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 2		virdev	3957	24.7	9.4	0.3	0.1
Aedes aegypti	Orlando - Induced - Male - 4th instar Larvae 3		virdev	1324	25.2	26.1	0	0
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 1		virdev	546	17.6	17.9	0	0
Aedes aegypti	Orlando - Uninduced -		virdev	208	19.7	15.4	0.5	0

	Female - 4th instar Larvae 2							
Aedes aegypti	Orlando - Uninduced - Female - 4th instar Larvae 3		virdev	9730	14.6	13.7	0.9	0
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 1		virdev	165	24.2	21.8	0	0
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 2		virdev	11794	18.4	5.9	0.4	0
Aedes aegypti	Orlando - Induced - Female - 4th instar Larvae 3		virdev	4324	21.6	10.4	0.8	0
Homo sapiens	Brain	GSM2257348	virdev	12	0	0	0	0
Homo sapiens	Ovary	GSM1584521	virdev	9457	32.6	21.5	4.1	0.3
Homo sapiens	Ovary	GSM1584522	virdev	8416	34.9	24.5	2.1	0.4
Homo sapiens	Testes	SRR835324	virdev	237987	50.2	33.2	43.2	4.9
Homo sapiens	Testes	SRR835325	virdev	158453	50	36.6	28.5	3.7
Mus musculus	Ovary	SRR014234	virdev	1079	27	22.2	0.4	0
Mus musculus	Testes	GSM802671	virdev	84304	49	30.7	43.2	6.2
Mus musculus	Testes	GSM802674	virdev	2699611	60	28.7	63.3	16.2
Macaca fascicularis	Ovary	GSM1584515	virdev	12483	46.1	25.6	1.9	0.1
Macaca fascicularis	Ovary	GSM1584516	virdev	27687	68.5	29.2	13.5	0.4
Macaca fascicularis	Testes	GSM1584519	virdev	16613	38.8	29.6	5.4	1.3
Macaca fascicularis	Testes	GSM1584520	virdev	8923	25.8	22.4	1.3	0.2
Bos taurus	Ovary	GSM1584503	virdev	35331	29.8	20.9	4.4	1.1
Bos taurus	Ovary	GSM1584504	virdev	17586	33.2	22.2	2.7	0.3
Bos taurus	Testes	GSM1584507	virdev	21406	50.9	28	10.2	1.7
Bos taurus	Testes	GSM1584508	virdev	3179	58.5	27.6	1.8	0.4
Danio rerio	Ovary	SRR578904	virdev	473	40.8	43.3	0	0

Danio rerio	Ovary	SRR578905	virdev	7473	26.9	19.6	4	0
Danio rerio	Ovary	SRR578906	virdev	18375	34	24.6	6	0.1
Danio rerio	Testes	SRR578922	virdev	16274	58	32.3	8.8	0.7
Danio rerio	Testes	SRR578923	virdev	14657	43.2	30.7	7.5	1.1

Table 4.1S: TruePaiR Benchmark Values for piRNA Amplification

Chapter 5: A Genomic Exploration of piRNA-mediated Deadenylation of Protein-coding Genes in *Drosophila melanogaster*

Section 5.1: Abstract

The Piwi-interacting sRNA (piRNA) subclass is the least characterized of the small non-coding RNAs. piRNAs have been shown to function in suppressing transposable elements (TEs) via Argonaute-mediated transcript slicing and epigenetic modification. Further, the pi-RNA induced silencing complex (RISC) has been shown to have the ability to regulate protein-coding genes as well. A complex involving PIWI proteins – Aub and Ago3 – along with Smaug, and the CCR4 deadenylase has been shown to associate with target transcripts in a sequence-specific manner within the 3' UTR to promote poly(A) shortening, resulting in downstream transcript degradation. Although piRNA-mediated deadenylation has been demonstrated in a protein-coding gene in *Drosophila melanogaster*, Nanos, this mechanism of regulation has yet to be thoroughly explored on a genome-wide scale. Our goal is to further characterize protein-coding gene targets of piRNA-mediated regulation via deadenylation by identifying several independent, genomic factors indicative of this interaction.

Section 5.2: Introduction

Beyond the piRNA and PIWI capability to repress target RNA by slicing, the PIWI RISC complex has also been implicated in repression indirectly via the modification of the

poly(A) tails of target mRNAs (Rouget et al. 2010). The physical association of piRNAs and PIWI proteins is well established (Aravin et al. 2007). Further, PIWI proteins – AUB and AGO3 – have been shown to be in a common complex with other proteins that directly facilitate poly(A) deadenylation: Smaug, CAF1, and CCR4 (Rouget et al. 2010; Smibert et al. 1999; Temme et al. 2004; Temme et al. 2004).

Smaug is a RNA-binding protein capable of physically associating with target mRNA molecules (Smibert et al. 1999). Smaug generally associates with 5-mer regions – generally CNGGN but most often CTGGC – in the target mRNA, referred to as Smaug Recognition Elements (SREs) (Chen et al. 2014). The association of Smaug to its target mRNA positions the CAF1 and CCR4 proteins to execute the poly(A) deadenylation (Temme et al. 2004).

CCR4 is the major 3' exonuclease subunit whose presence dominates in wild-type conditions (Tucker et al. 2001). CAF1 is a secondary 3' exonuclease subunit within the deadenylase complex (Tucker et al. 2001). Both CCR4 and CAF1 have the capability of association with the 3' poly(A) tail of a target mRNA and cleaving adenine residues from the 3' end . (Temme et al. 2004; Tucker et al. 2001; Tucker et al. 2001; Temme et al. 2004). The absence of a poly(A) tail results in the degradation of mRNAs (Tucker et al. 2001; Temme et al. 2004).

piRNAs have been shown to play a critical role in guiding the PIWI-Smaug-CCR4 complex to its targets. When eliminating the *412* and *roo* TE-derived piRNAs complementary the *Nanos* 3' UTR in *Drosophila melanogaster* early embryonic

development, poly(A) deadenylation ceased (Rouget et al. 2010). Consequentially, the *Nanos* mRNA molecules were not degraded, as shown in wild-type embryos (Rouget et al. 2010).

Although piRNA association within the Smaug-PIWI-CCR4 protein complex is necessary to promote deadenylation in *Nanos* in *Drosophila melanogaster*, the analysis did not return a comprehensive set of genes under this regulatory mechanism (Rouget et al. 2010; Dahanukar et al. 1999). I set to explore the possibility of other protein-coding genes that may also be affected by a similar mechanism of regulation by utilizing several independent factors related to the structure of the *Nanos* 3' UTR. The factors that I assess include transcript depletion in early embryos, an accumulation of TE-derived piRNAs that exhibit a region of complementarity to a specific 3' UTR, mapping of a potentially functional TE-derived piRNA to a defined piRNA cluster, as well as the 3' UTR of the gene of interest's structural consistency with the *Nanos* 3' UTR.

Given that Smaug – in cooperation with piRNAs, AUB and AGO3 – has been implicated in recruiting the CCR4 deadenylase to transcripts to promote poly(A) tail shortening, Smaug's association with mRNA molecules can provide additional information into genic targets that are likely to be regulated by piRNA-mediated deadenylation (Dahanukar et al. 1999). Factors regarding a physical Smaug association, as well as regulation of transcripts may also be useful in the assessment of genes that undergo piRNA-mediated deadenylation. Using additional factors to those described in the Smaug-Independent analysis, I further assess potential targets of piRNA-mediated

deadenylation by the presence of a Smaug recognition element (SRE), direct transcript interaction with Smaug, and transcript derepression in the absence of a functional Smaug protein (Smibert et al. 1999).

Section 5.3: Methods and Results

Smaug-Independent Pipeline for Gene Target Prediction of piRNA-mediated Deadenylation

The only previous study regarding piRNA-mediated deadenylation was investigated primarily in one gene, *Nanos*, in *Drosophila melanogaster*. I was also interested in the assessment of other potential genes that may be under a similar mechanism of piRNA-mediated deadenylation.

Nearly 30 million piRNAs, from seven publicly available sRNA libraries from the previous investigation of Nanos piRNA-mediated deadenylation in early *Drosophila melanogaster* embryos, 0-2 hours old, were used to assess the piRNA landscape (Leinonen et al. 2010; Rouget et al. 2010).

miRNAs generally have an indicative 21-23 nucleotide length profile, depending on the species of interest (Elbashir et al. 2001). piRNAs, however, typically have a more broad length spectrum (Aravin et al. 2007). Primary piRNAs tend to exhibit an uracil (U) at position 1, and secondary piRNAs tend to exhibit an adenine (A) at position 10. These biases exist due to the mechanism of piRNA biogenesis via the primary pathway and

amplification loop, respectively. These biases, though, are not absolutely exclusive of piRNAs. That is, many piRNAs exist that do not show these biases, while other sRNA, which are not piRNAs, may exhibit the biases (Aravin et al. 2007). Therefore, the only sRNA filter that can be used to predict piRNAs is by the use of a sRNA length cutoff value. piRNAs tend to be identified as longer, generally at 24-32 nts, in comparison to the other types of sRNA found in *Drosophila melanogaster* (Aravin et al. 2007). Therefore, in this research, any sRNA that has a length greater than 24 nts is considered a putative piRNA.

As a proof of concept that I can detect functional piRNAs within genomic 3' UTRs, I first aimed to confirm the presence of those previously identified functional piRNA with the 3' UTR of *Nanos* within our genomic screen. I used the BLAST algorithm to assess piRNAs that have a region of complementarity to genomic 3' UTRs (Altschul et al. 1990). I ran a NCBI BLAST with a word size of 14 and an E-value of 100, as described in the discovery of the functional piRNAs complementary to the Nanos 3' UTR (Altschul et al. 1990). The piRNA motif mediating target specificity has been shown to exist anywhere within the piRNA sequence, rather than strictly the 2-8 nucleotide seed region that is generally observed in miRNAs (Rouget et al. 2010; Lewis et al. 2003). Therefore, any piRNA motif that is complementary to a 3' UTR, as returned in the BLAST search, was considered as a potential guide of Aub or Ago3-RISC transcript association.

Further, the functional piRNAs observed in the Nanos 3' UTR were associated with the TEs from which they had derived. I acknowledge that all piRNAs do not appear to be derived from TEs, but for the purposes of this analysis, I only considered those piRNAs

that can be mapped to a known TE in *Drosophila melanogaster*. TE sequences were extracted from the RepBase database (Jurka et al. 2005). I then performed a Bowtie2 alignment, with default parameters, of all piRNA sequences to the sequences of all TEs in *Drosophila melanogaster* (Langmead & Salzberg 2012). I defined those piRNAs that uniquely mapped to a *Drosophila melanogaster* TE sequence as a TE-derived piRNA.

Those piRNAs that were both TE-derived and contain a significant region of complementarity to a 3' UTR were used in downstream analysis. This analysis returned a gene list of about 4,000 genes with over 500,000 TE-derived piRNAs with regions of complementarity to a 3' UTR.

Next, I subset the TE-derived piRNAs to specifically analyze the *Nanos* 3' UTR for the presence of previously established and functional TE-derived piRNAs. For the functional piRNAs deriving from the *412* LTR retrotransposon of the *Gypsy* family, I identified 50 piRNAs that contain the functional motif (Figure 5.1A). I also showed about 600 other piRNAs that are TE-derived, have a region of complementarity to the *Nanos* 3' UTR, and map to a defined piRNA cluster in *Drosophila melanogaster* (Figure 5.1B). However, many of these TE-derived piRNAs are not downstream of the Smaug recognition element (SRE), and therefore are unlikely to function in promoting deadenylation (Smibert et al. 1999). The function of these additional piRNAs is unknown. Since I made parameter calls as described in the previous research, I attributed this difference to variation deriving from software improvements within the local alignments and read mapping tools utilized, the improved annotation of *D. melanogaster* TEs, and innate stochasticity amongst biological piRNA populations. With that said, since these parameters were able

to generate a significant hit of the known functional piRNA motif, I deemed these parameters sufficient to detect potential piRNA complementarity and can be utilized within the workflow. Once the previously established functional piRNAs were detected, I was then confident in an extrapolation of the workflow to detect additional genes that may be acted upon by a similar regulatory mechanism (Figure 5.2).

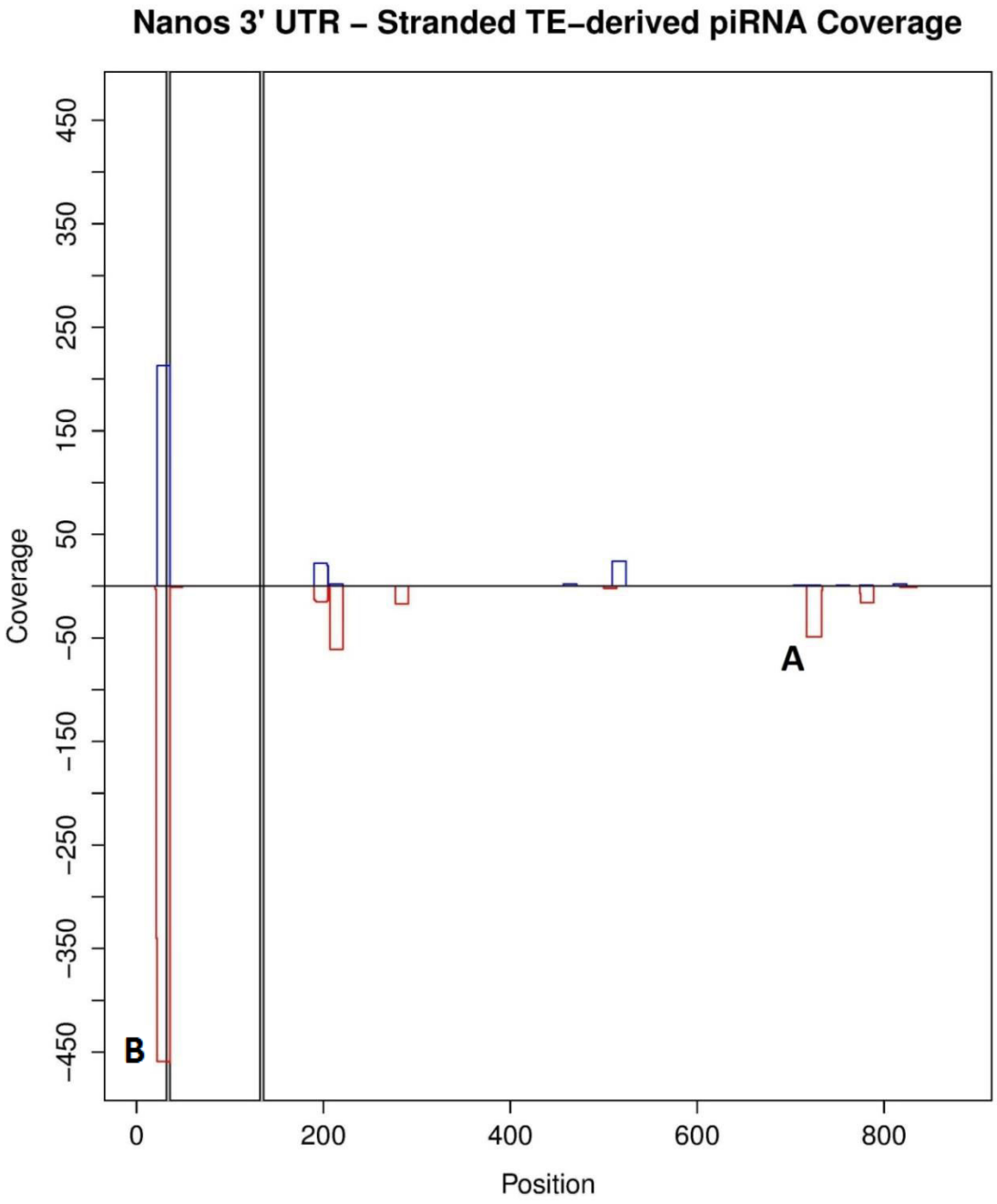


Figure 5.1 | piRNA Landscape in the *Nanos* 3' UTR. (**A**) Validation of the known, functional 412 transposable element-derived piRNAs to the *Nanos* 3' UTR. The gray regions are the defined SREs within the TCE of the *Nanos* 3' UTR. Sense (blue) and antisense (red) 412-derived piRNAs are represented as having a region of complementarity to the *Nanos* 3' UTR. (**B**) DMRT1B transposable element-derived piRNAs that map to a defined piRNA cluster, but are not downstream of the SREs.

Since not every TE-derived piRNA with complementarity to a 3' UTR has an easily ascertained function, I focused gene target filtering based on other independent factors indicative piRNA-mediated deadenylation. Further, as there is no published data regarding the sequences to which the Aub or Ago3 RISC complexes associate, it is most advantageous to initially filter out those genes whose 3' UTRs appear to be incapable of regulation via piRNA-mediated deadenylation. Factors indicative of an incapability to undergo piRNA-mediated regulation via deadenylation include no apparent degradation of transcript in early embryos, a lack of TE-derived piRNA accumulation in a defined 3' UTR, and a lack of complementary TE-derived piRNAs that can be mapped to a defined piRNA cluster in *Drosophila melanogaster*.

Previous investigation of *Nanos* has shown that Smaug is present and able to degrade transcript between 0-2 hr and 2-4 hr embyos in *Drosophila melanogaster* (Rouget et al. 2010; Pinder & Smibert 2013; Dahanukar et al. 1999). I assessed the degree of degradation between 0-2 hr embryos and 2-4 hr embryos based upon RNA-seq expression profile data from the modENCODE group, as seen on FlyBase (Washington et al. 2011; dos Santos et al. 2015). As the Nanos expression profile has an RPKM-normalized expression profile of 243 transcripts in 0-2 hr embryo, and 22 in 2-4 hr embryos, I explored other genes with similar, significant transcript during this timeframe (Figure 5.4) (Celniker et al. 2009). With that said, I acknowledge that a larger hairpin within the secondary structure of the 3' UTR, as well as other, potentially unknown factors, may influence the rate of degradation as well. Therefore, I do consider the

possibility of genes being degraded by this mechanism to various extents (Laver et al. 2013).

In the established model of piRNA-mediated deadenylation, TE-derived piRNAs complementarity to the Nanos 3' UTR are shown to guide the piRNA-mediated deadenylation complex to target transcripts. Therefore, I explore other genes of interest by the observation of a substantial number of TE-derived piRNAs that have a region of complementary to their specific 3' UTRs. In this analysis, I only considered genes that had more than 300 TE-derived piRNAs complementary to their 3' UTR.

Finally, I utilized the default settings of Bowtie2 to map the TE-derived piRNAs, with regions of complementarity to a specific 3' UTR, to sequences that have been defined as piRNA clusters in *D. melanogaster* (Langmead & Salzberg 2012). I extracted genomic loci from the piRNABank that define regions of the genome that have been considered piRNA clusters (Sai Lakshmi & Agrawal 2008). The mapping of TE-derived piRNAs to a specific, defined piRNA cluster provides an indication for the origin of the piRNAs.

Taken together, these independent factors returned a subset of 54 genes that appear to be regulated via piRNA-mediated deadenylation. Probabilities have been calculated to represent the chance that a gene satisfies each independent factor (Table 5.1). Further investigation of the genes that satisfied these filters is required to prioritize the functional

assay of potential piRNA-mediated deadenylation. Additional factors such as expression profiles, piRNA sequence, and putative function can be useful in prioritizing the assay of these genes (Figure 5.2).

Independent Factors	Number of hits	Total Hits Possible	Probability of Occurrence
Transcript Depletion in Early Embryos	286	17,262	1.66%
Presence of >300 TE-derived piRNAs complementarity to a specific 3' UTR	15,888	30,277	52.48%
TE-derived piRNAs complementary to genic 3' UTR also maps to defined piRNA cluster in *Drosophila melanogaster*	1,855	5,548	33.44%
Probability of all 3 independent factors			**2.91E-3**

Table 5.1 | Smaug-Independent Filters. Probability that a gene or transcript satisfies each of the Smaug-Independent filter criteria considered individually, and combined.

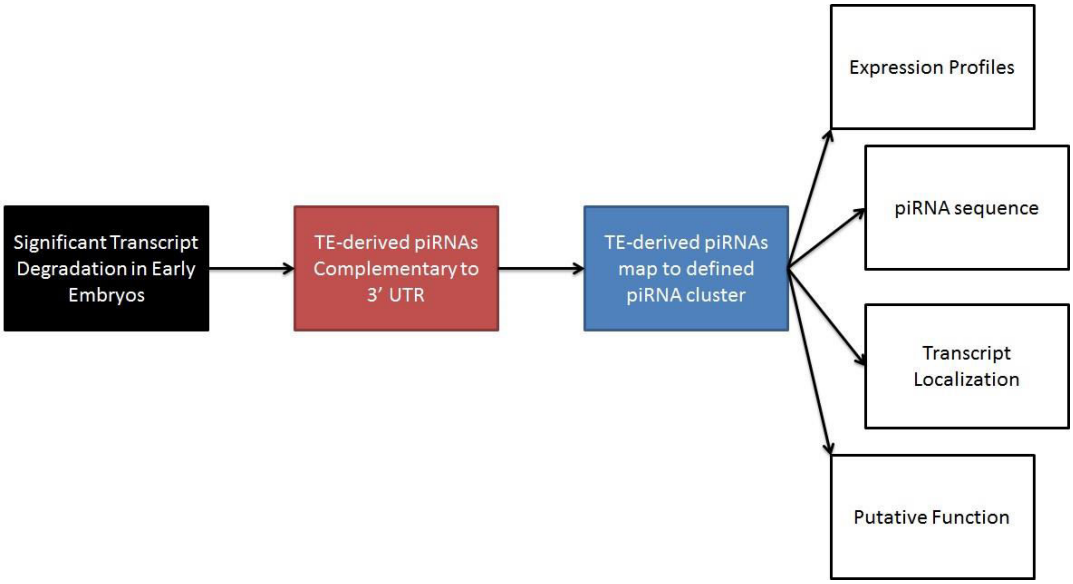


Figure 5.2 | Smaug-Independent Workflow. The series of filters regarding the

prediction of piRNA functionality in facilitating deadenylation of protein-coding transcripts.

Smaug-Dependent Pipeline for Gene Target Prediction of piRNA-mediated Deadenylation

The RNA-binding protein, Smaug, has been shown to immunoprecipitate (IP) with other proteins that are supposed to be in the same complex: Aub, Ago3, CAF1, and CCR4 (Rouget et al. 2010). Smaug, like most RBPs, has many targets to which it associates and represses (Chen et al. 2014; Dahanukar et al. 1999). Smaug's ability to bind transcripts and guide protein complexes to target transcripts potentially renders it an interesting asset to this analysis. Although, I acknowledge that the guiding of this deadenylase complex to target transcripts may be a result of strictly Smaug binding, strictly piRNA binding, or a cooperative effort amongst the two. Previous data suggests that piRNAs are likely playing a supplementary role in the recruitment of the deadenylase complex (Rouget et al. 2010). Therefore, building upon our previous Smaug-Independent workflow, I further assess genes of interest based on factors indicative of Smaug interaction and repression.

Therefore, it is likely informative to consider RNA targets of Smaug as further confidence in the prediction of piRNA-mediated recruitment of the deadenylase complex. The recruitment of the deadenylase complex to target mRNA has been shown to induce poly(A) tail shortening and downstream degradation in *Nanos* in *Drosophila melanogaster*.

In this analysis, I further assess potential targets of piRNA-mediated deadenylation by Smaug-dependent factors such as the presence of a SRE, overrepresentation of Smaug binding to target mRNA in immunoprecipitations (IPs), and transcript recovery in the absence of functional Smaug.

The first step in the workflow is to identify 3' UTRs that contain SREs. A SREs is a 5-mer region within the 3' UTR, generally towards the 5' end of the 3' UTR, which is indicative of Smaug association. In the *Nanos* 3'UTR, there are two SREs, each of the sequence, CTGGC. Both of the SREs are contained in the *Nanos* translational control elements, which is found in the 5' region of the 3' UTR (Dahanukar et al. 1999; Forrest et al. 2004). Only one of these SREs appears to be necessary for the recruitment of the piRNA-mediated deadenylase complex (Crucs et al. 2000). The TE-derived piRNAs, with a region of complementarity to the 3' UTR, exist downstream of the SREs in the *Nanos* model. Given that RNA has the ability to fold into proper conformation, I further consider any targets that contain an accumulation of TE-derived piRNAs at any region downstream of the SRE.

It is important to note that through a comparative analysis of SRE conservation within the 3' UTRS of other *Nanos* homologs, as well as evidence from recent literature, suggests that SREs may have a more general motif, CNGGN (Clark et al. 2007; Chen et al. 2014). Therefore, I also assess the presence of a SRE by incorporating the detection of this more general SRE motif.

Next, I investigate previous research that has shown that Smaug binding to 3' UTRs of transcripts facilitates target degradation (Dahanukar et al. 1999). There exists data regarding the transcript targets of Smaug binding and their expression levels in early embryos in the absence of a functional Smaug protein. Smaug IPs, followed by RNA elutions, were performed to assess a transcriptome-wide scale overrepresentation of Smaug association relative to control IPs (Chen et al. 2014). Also, independently, a transcriptome-wide assay in which the RNA-binding domain of Smaug was made non-functional, and transcript level within the cells was assessed. With these data, I can further assess our genes of interest on whether or not the transcripts of interest were shown to significantly associate with Smaug, as well as whether or not the gene of interest's transcripts recovered in the absence of functional Smaug (Chen et al. 2014). Presumably, in the absence of functional Smaug, there is an absence of this mechanism of deadenylation, and recovery of transcript values.

I have calculated the probability that a particular gene would satisfy each particular, Smaug-dependent criterion. Assuming that each of these factors occurs independently of one another, I also return the probability that a gene would satisfy all of the six Smaug-Independent and Smaug-Dependent criteria strictly by chance (Table 5.2).

When applying the six independent filters to all transcripts in Drosophila melanogaster, only two genes were returned: *Nanos* and *CG5010*. As piRNA-mediated deadenylation has already been described in *Nanos*, I focused on further exploration of *CG5010*. The function of *CG5010* has not been previously identified.

Independent Factors	Number of hits	Total Hits Possible	Probability of Occurrence
Presence of SRE in 3' UTR	2,976 (CNGGN: 10,943)	30,277	9.83% (36.14% with more general motif: CNGGN)
Overrepresented Smaug Association	312	17,262	1.81%
Transcript Recovery in the Absence of Functional Smaug	1,814	17,262	10.51%
Probability of all 6 independent factors			**5.45E-7**

Table 5.2 | Smaug-Dependent Filters. Probability that a gene or transcript satisfies each of the Smaug-dependent filter criteria considered individually. The final probability calculation takes into account both Smaug-independent and Smaug-dependent filters.

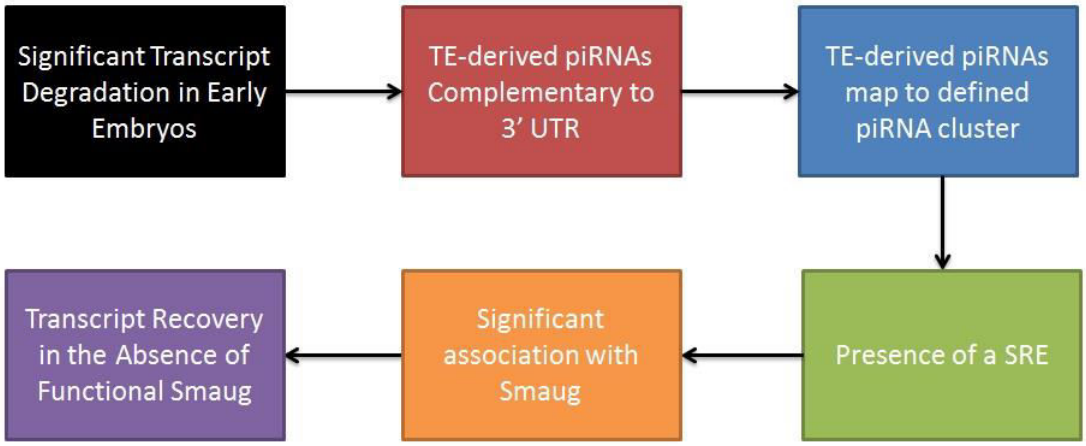


Figure 5.3 | Smaug-Dependent Workflow. The series of filters regarding the Smaug-dependent prediction of piRNA functionality in facilitating deadenylation of protein-coding transcripts.

I performed further investigation regarding whether or not the *CG5010* expression profile is consistent with the *Nanos* profile. Although, *Nanos* transcript is more abundant in *Drosophila melanogaster* embryos 0-2 hrs relative to *CG5010*, both *Nanos* and *CG5010* transcript abundance falls nearly 90% in 2-4 hr embryos (Fig. 5.4). Since initial transcript abundance presumably has little to do with the recruitment of the deadenylase complex, yet transcript abundance falls on a similar scale, I concluded that the *CG5010* expression profile is consistent with the Nanos profile in terms of having the capability of being regulated via a similar mechanism.

CG5010 TE-derived piRNAs complemented the 3' UTR downstream of the SRE and mapped to a non-LTR, Jockey family retrotransposon named FW (Fig. 5.5). It is worth noting that the previously established functional piRNAs in the *Nanos* 3' UTR derived from a Gypsy family transposon, 412. The motif regarding the TE-derived piRNA region of complementarity to the *CG5010* 3' UTR of those that mapped to a defined piRNA cluster is CAAAACGAAAACGTA. Although the motif doesn't occur in the same position in all of the piRNAs, most of the piRNAs exhibit this motif from the 11th to the 25th nucleotides. With that said, about 30% of the piRNAs exhibited the motif start in the first 4 nucleotides (Table 5.3).

Therefore, I present the possibility that piRNAs may promote the deadenylation of *CG5010*, among other potential protein-coding candidates, in early *Drosophila melanogaster* embryos. Neither the post-transcriptional regulation of *CG5010* transcripts, nor the effect of *CG5010* derepression in early embryos has been previously explored, but may lead to a better understanding as to the breadth of the piRNA system.

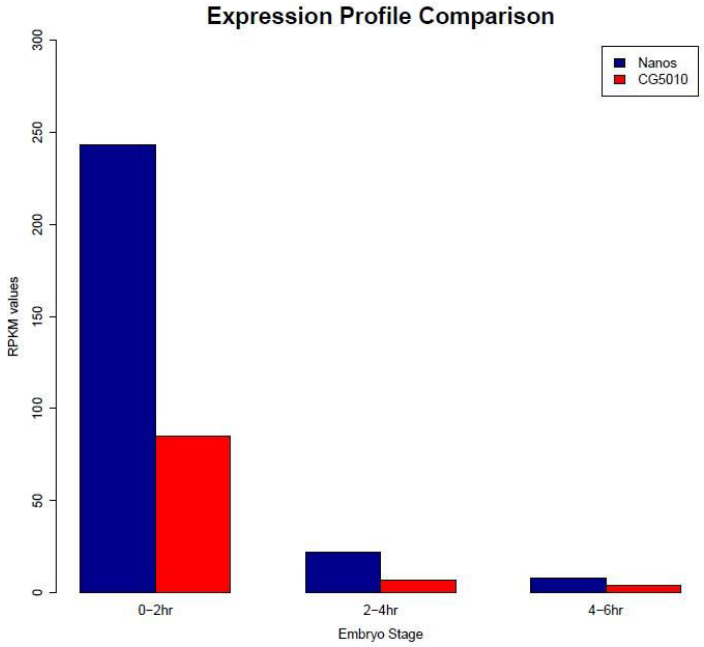


Figure 5.4 | Expression Profile Comparison. Bar plot comparing the RPKM values of transcript expression of Nanos and CG5010 in early embryos.

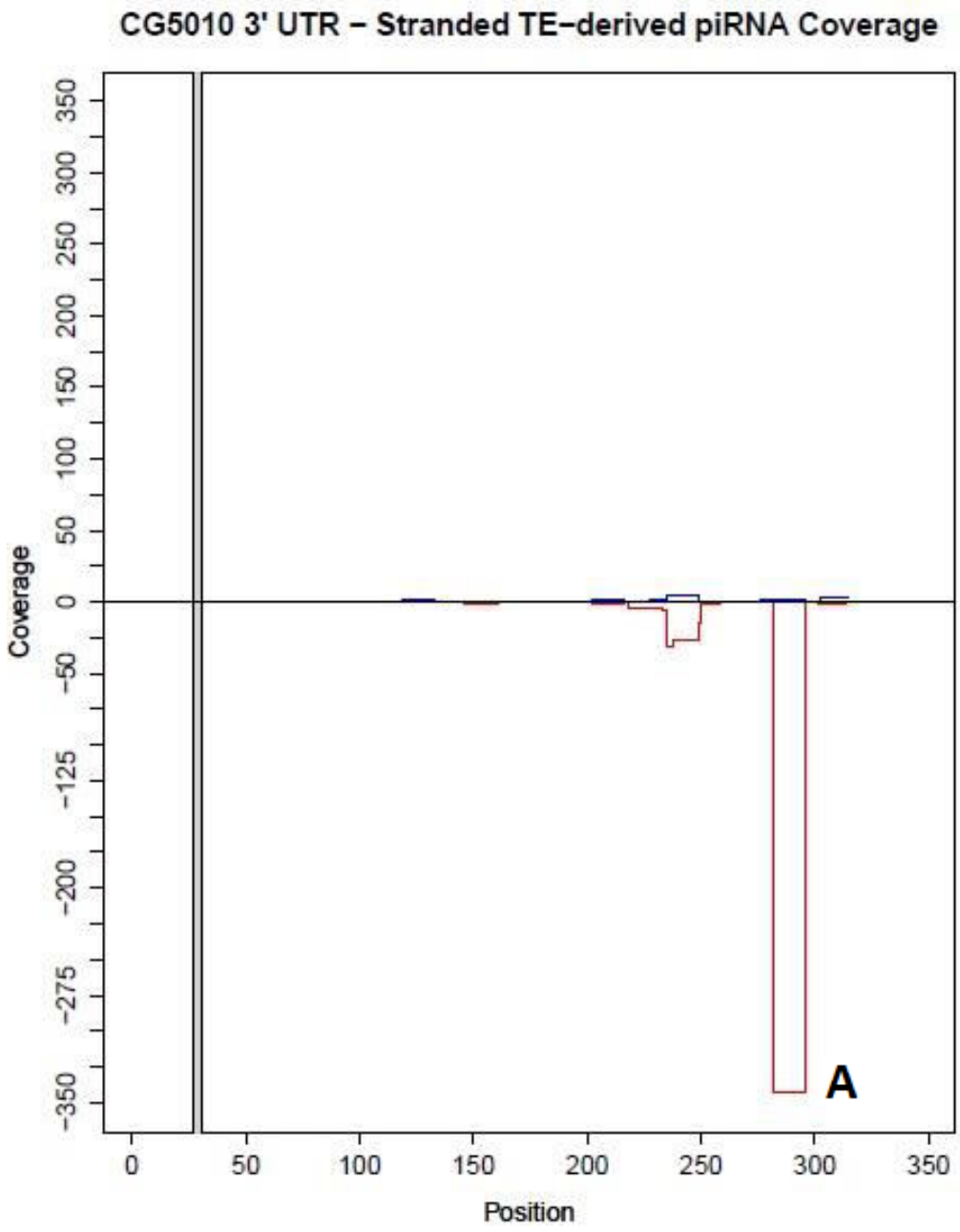


Figure 5.5 | piRNA Landscape in the *CG5010* 3' UTR. The gray regions are the defined SREs within the *CG5010* 3' UTR. Sense (blue) and antisense (red) 412-derived piRNAs are represented as having a region of complementarity to the *CG5010* 3' UTR. (**A**) The FW transposable element-derived piRNAs that also mapped to a defined piRNA cluster.

piRNA Position of Motif Start Site	Number of piRNAs
1-4	119
5-9	43
10-14	227
15-18	2

Table 5.3 | piRNA Motif Position. Table representing the number of piRNAs that had the complementary motif to the 3' UTR start at the beginning (nts 1-4), early middle(nts 5-9), late middle (nts 10-14) and end (nts 15-18) of the piRNA.

Section 5.4: Future Directions

Upon the establishment of genes of interest, like *CG5010*, potentially under piRNA-mediated deadenylation, experimental validation is required to establish the significance of the regulation.

First, it is necessary to demonstrate that deadenylation of *CG5010* transcripts is taking place in early, wild-type embryos. In order to assess whether or not deadenylation is occurring in CG5010 transcripts, I will observe the length of the poly(A) tails of *CG5010* transcripts in a time series in early embryos via poly(A) tests (PAT). PATs are Polymerase Chain Reactions (PCRs) that indicate the length of poly(A) regions in transcripts of interest. Initially, an adaptor is ligated onto the 3' end of RNAs within the sample. The adaptor sequence can then act as a priming site for reverse transcriptase to convert the molecules from RNA to cDNA. Then, using the cDNA, primers can be chosen to amplify the region between a unique sequence within the transcript of interest to the adapter site beyond the poly(A) tails of the transcripts of interest. Finally, size

selection of the cDNA amplified product on an agarose gel electrophoresis would provide the length of the poly(A) tail (Sallés et al. 1999). PAT assays in *Nanos* showed poly(A) tails between 100-250 nucleotides in length. It is reasonable to predict that the length of the poly(A) tails of *CG5010* transcripts would be similar. If *CG5010* is undergoing deadenylation, then the poly(A) tails should shorten in the wild-type time series.

Upon the establishment of deadenylation in early, wild-type embryos, I can proceed to investigate whether piRNAs are playing a role in facilitating the regulation. I would explore the necessity of the piRNA complementarity to the target 3' UTR by injecting anti-piRNAs into a separate time series in early *Drosophila melanogaster* embryos. Anti-piRNAs are single-stranded oligonucleotides that have a complementary sequence to the piRNAs supposedly guiding the deadenylation complex. In theory, the anti-piRNA should bind to the complementary, potentially functional piRNA to render the guiding element non-functional. Therefore, if the piRNAs of interest are playing a role in guiding the deadenylation complex to *CG5010* transcripts, the efficiency of this mechanism should be compromised. A change in phenotype could be assessed in this experiment by observing the poly(A) tail length in the subsequent stages of embryos: 2-6 hrs relative to 0-2 hrs embryos. If piRNAs are promoting recruitment of the CCR4 deadenylase complex to shorten poly(A) tails of target transcripts, then upon injection of anti-piRNAs, poly(A) tail shortening should become less efficient or cease altogether. I would perform a control injection using only injection buffer to demonstrate that any observed phenotypic aberration is not resulting from the innate stress of an injection in the embryos. Further, I would also inject a piRNAs that do not contain the potentially

functional motif in order to demonstrate that the piRNA-mediated deadenylation is occurring in a sequence-specific manner. The degree of phenotypic difference between the wild-type embryos, the previous Nanos anti-piRNA assays, as well as our assays in *CG5010* will provide further information as to the extent of regulation resulting from piRNAs in this mechanism.

Section 5.5: Conclusions

Taken together, the Smaug-independent criteria aim to predict other protein-coding genes undergoing piRNA-mediated deadenylation. Although piRNA-mediated deadenylation has already been established in a protein-coding gene, *Nanos*, the deadenylation mechanism of piRNA-mediated regulation has yet to be explored on a transcriptome-wide scale. Based on genic 3' UTR structure and piRNA complementarity, I have generated novel genes that have the potential of undergoing piRNA-mediated deadenylation. I suggest the continued exploration of target protein-coding transcripts that may be under regulation via piRNA-mediated deadenylation using several independent factors indicative of this mechanism of regulation in the Nanos 3' UTR.

Further, as well as to facilitate poly(A) deadenylation and transcript degradation in *Nanos*, Smaug has also been shown to be involved in a common complex with PIWI proteins. Immunoprecipitations suggest that Smaug, Aub and Ago3 form a complex that can associate with the CCR4 exonuclease to promote poly(A) deadenylation on its target

transcripts (Rouget et al. 2010). Therefore, the investigation of the genic targets of Smaug-mediated deadenylation can aid in the identification of transcripts potentially undergoing piRNA-mediated deadenylation. I have identified, and seek to further investigate the protein-coding gene, *CG5010*, as a potential target of piRNA-mediated deadenylation. *CG5010* is the one protein-coding gene, with the exception of *Nanos*, that satisfied all six of the filters specified in Smaug-Independent and Smaug-Dependent workflows regarding target identification of piRNA-mediated deadenylation in *Drosophila melanogaster*.

The concept of PAT assays can be utilized to assess whether the other potential genic targets of piRNA-mediated deadenylation also undergo poly(A) tail degradation in early embryo, and are also reliant on piRNAs to facilitate the poly(A) degradation, as observed in *Nanos*.

Further demonstration of piRNA purpose, targets, and functionality would contribute to a better understanding of this relatively poorly understood class of sRNA. A better understanding of the piRNA class of sRNA can provide further explanation in piRNA function, transposable element regulation, and even the regulation of protein-coding genes. Also, as the functional mechanisms of the sRNA system are relatively well conserved among the Metazoans, these findings within *Drosophila melanogaster* embryos can likely be extrapolated to other species of interest.

Section 5.6: References

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

Aravin, A.A., Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *science*, 318(5851), pp.761–764.

Celniker, S.E. et al., 2009. Unlocking the secrets of the genome. *Nature*, 459(7249), pp.927–930.

Chen, L. et al., 2014. Global regulation of mRNA translation and stability in the early Drosophila embryo by the Smaug RNA-binding protein. *Genome Biol*, 15(1), p.R4.

Clark, A.G. et al., 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167), pp.203–218.

Crucs, S., Chatterjee, S. & Gavis, E.R., 2000. Overlapping but Distinct RNA Elements Control Repression and Activation of< i> nanos</i> Translation. *Molecular cell*, 5(3), pp.457–467.

Dahanukar, A., Walker, J.A. & Wharton, R.P., 1999. Smaug, a Novel RNA-Binding Protein that Operates a Translational Switch in< i> Drosophila</i>. *Molecular cell*, 4(2), pp.209–218.

Elbashir, S.M. et al., 2001. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *The EMBO journal*, 20(23), pp.6877–6888.

Forrest, K.M. et al., 2004. Temporal complexity within a translational control element in the nanos mRNA. *Development*, 131(23), pp.5849–5857.

Jurka, J. et al., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110(1-4), pp.462–467.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–359.

Laver, J.D. et al., 2013. Genome-wide analysis of Staufen-associated mRNAs identifies secondary structures that confer target specificity. *Nucleic acids research*, p.gkt702.

Leinonen, R., Sugawara, H. & Shumway, M., 2010. The sequence read archive. *Nucleic acids research*, p.gkq1019.

Lewis, B.P. et al., 2003. Prediction of mammalian microRNA targets. *Cell*, 115(7), pp.787–98.

Pinder, B.D. & Smibert, C.A., 2013. microRNA-independent recruitment of Argonaute 1 to nanos mRNA through the Smaug RNA-binding protein. *EMBO reports*, 14(1), pp.80–86.

Rouget, C. et al., 2010. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature*, 467(7319), pp.1128–1132.

Sai Lakshmi, S. & Agrawal, S., 2008. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research*, 36(Database issue), pp.D173–7.

Sallés, F.J., Richards, W.G. & Strickland, S., 1999. Assaying the polyadenylation state of mRNAs. *Methods*, 17(1), pp.38–45.

Dos Santos, G. et al., 2015. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic acids research*, 43(D1), pp.D690–D697.

Smibert, C.A. et al., 1999. Smaug, a novel and conserved protein, contributes to repression of nanos mRNA translation in vitro. *Rna*, 5(12), pp.1535–1547.

Temme, C. et al., 2004. A complex containing the CCR4 and CAF1 proteins is involved in mRNA deadenylation in Drosophila. *The EMBO journal*, 23(14), pp.2862–71.

Tucker, M. et al., 2001. The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in Saccharomyces cerevisiae. *Cell*, 104(3), pp.377–386.

Washington, N.L. et al., 2011. The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details. *Database*, 2011, p.bar023.

Chapter 6: Summary and Conclusions

Section 6.1: Summary

The research presented herein the thesis investigates the occupancy, abundance, distribution and functionality of repeat elements in Metazoans. Repeat loci were identified and characterized on a genome-scale in *Ceratitis capitata*. An overrepresentation of particular TE families was then identified. Given that *Mariner/Tc1* TEs were the most prevalent and demonstrated the highest degree of overrepresentation, I further classified the elements, from an evolutionary perspective, with the goal of inferring *Mariner/Tc1* element origin and better understanding TE mechanics in the Mediterranean fruit fly.

The remaining research presented herein focuses on the repressive sRNA of TEs: piRNAs. I developed software that serves to be informative to a better understanding of the piRNA system. The software provides a general tool for the analysis of primary and secondary piRNAs, as well as the piRNA cluster loci from which piRNAs originate, in any species of interest. I also explored mRNA target prediction of a less well-understood mechanism of piRNA-mediated repression.

Section 6.2: Repeat Element Identification, Characterization, and Evolutionary Origin in the Mediterranean Fruit Fly, *Ceratitis capitata*

In collaboration with the i5k Consortium, I contributed to characterizing the genome contents of the agriculturally and economically important species, the Mediterranean Fruit Fly. The repeat elements of the genome were identified, characterized, and associated with chromosomal loci using well-established algorithms.

Hyperactive TE families were identified to best understand the TE mechanics, as well as the functional chromosomal loci and origin within *Ceratitis capitata*. The evolutionary origin regarding the individual TEs within the most abundant and overrepresented TE superfamily, *Mariner/Tc1*, was observed. Previously established and novel *Mariner/Tc1* subclasses were observed in the medfly. The results presented herein lay the foundation for continued research in identifying a molecular rationale for the differential activity of *Mariner/Tc1* TE subclasses.

Section 6.3: piClusterBusteR: A Program for Automated piRNA Cluster Characterization

piRNA clusters are the regions of the chromosome that serve as the precursors for primary piRNA generation (Aravin et al. 2007). Therefore, the sequences within piRNA clusters dictate the sequences of the piRNAs and the downstream targets of the piRNAs. piRNA clusters have been annotated on a small scale, but despite the critical nature of these sequences, had not been analyzed on a large-scale.

Initially, I mined datasets within the Short Read Archive and the Gene Expression Omnibus (Sayers et al. 2011; Edgar et al. 2002). The extracted data featured over one hundred libraries, billions of reads, and data in 13 Metazoan species.

The large data that was extracted was utilized as an input for software that I developed, piClusterBusteR. piClusterBusteR is a tool made available for the general analysis of piRNA clusters in any species of interest. The software requires information to produce a meaningful result such as sRNA library or defined piRNA cluster loci, a reference genome, and an organism-specific gene set. piClusterBusteR has the capability to utilize high throughput and cluster computing infrastructure – via multitasking and multithreading – to automatically, accurately, and efficiently define piRNA cluster loci, and characterize fragments of sequences within piRNA clusters on a large scale.

I established that piRNA cluster definition, as defined in the proTRAC algorithm, is not significantly influenced by the number of reads in a library. piRNA cluster definition was also not significantly influenced by the size of the genome of interest (Rosenkranz & Zischler 2012). Despite the potential definition of hundreds and thousands of piRNA cluster loci using this algorithm, I showed that the top 30 piRNA clusters are generally the major contributors to piRNA populations in ovary and testis tissue in Metazoans.

My analysis of top piRNA cluster architecture demonstrated significantly similarity in the consistency of piRNA cluster loci within and across species, despite little to no syntenic or piRNA sequence conservation between species (Grimson et al. 2008; Chirn et al. 2015). TEs were the main known constituent in ovary and testis of the 13 Metazoans

investigated in this analysis, with a large degree of sequence that is not of a known origin. Further, contrary to small-scale piRNA cluster characterization, I observed the features of a piRNA clusters to be in similar proportions of sense and antisense orientation, rather than predominantly antisense (Malone & Hannon 2009). The unanticipated observation of sense oriented features within piRNA clusters contributes to a better understanding of the mechanisms by which the contents of piRNA clusters are formed, as well as the mechanisms by which piRNAs are generated.

Finally, I observed the consistency of piRNA cluster definition within and between the species for which biological replicates were available. The consistency of piRNA cluster definition was statistically significant between two same tissue libraries within the same species and comparison across germline tissues within the same species. This finding suggests that when observing piRNA clusters on a large scale, contrary to previous suggestion of master piRNA cluster loci, distinct sets of piRNA clusters appear to be producing the majority of the piRNAs (Brennecke et al. 2007). It is worth noting that I did also observe a small subset of master piRNA cluster loci in the 13 Metazoans observed: piRNA cluster loci that were producing the majority of piRNAs in both germline tissues.

Section 6.4: Bioinformatics Method Improvement in piRNA Biology

While examining TEs and piRNAs, and in pursuit of experimentation to contribute to scientific knowledge of these systems, I developed computational method improvements for general use in the field.

I improved the capability for the detection of relative utilization of the secondary piRNA pathway, also referred to as the Ping-Pong pathway or amplification loop, using software that I developed, referred to as TruePaiR (Brennecke et al. 2007). TruePaiR utilizes sRNA sequencing data to find read pairs within the sRNA reads that have the capability of complementing one another in the first ten base pairs via the Ping-Pong pathway. The degree of complementarity in the first ten base pairs of piRNAs dictates the relative degree of amplification of the secondary pathway of piRNA biogenesis.

TruePaiR was run in whole body, embryonic, ovary, testis, and brain tissues in nine species. The results demonstrate the efficacy and reproducibility of the detection of piRNA amplification using TruePaiR.

The TruePaiR results serve as a benchmark of piRNA amplification in the tissues of model organisms that were observed. piRNA amplification was also observed separately for piRNAs of TE, genic, and viral origin. In tissues in which the piRNA pathway was substantially utilized, piRNAs of TE and genic origin varied in their relative utilization of piRNA amplification. However, piRNAs of TE and genic origin ubiquitously participated in piRNA amplification to a substantially higher degree relative to piRNAs of viral origin.

Section 6.5: piRNA-Mediated Deadenylation

Genomic piRNA-mediated regulation was observed considering an alternative mechanism of target suppression. piRNAs have been implicated in the capability to associate with an exonuclease complex and facilitate poly(A) deadenylation of *Nanos* in *Drosophila melanogaster*. I predicted other potential protein-coding mRNA targets of piRNA-mediated deadenylation.

I established a workflow to consider all other 3' UTRs in *Drosophila melanogaster* and identify 3' UTRs that have a similar profile to *Nanos*. 3' UTRs were filtered by considering mRNA depletion in early embryos in *D. melanogaster*, piRNAs with complementarity to the 3' UTR of the target gene, and confirmation that the piRNAs with complementarity to the 3' UTR derived from a known piRNA cluster. Using the Smaug-Independent criterion, 55 genic targets have been identified as undergoing potential piRNA-mediated deadenylation.

After the advent of functional genomic Smaug assays, further consideration was made with regard to data related to another guiding factor of the PIWI-Smaug-CCR4 exonuclease complex, the Smaug protein, in assessing potential protein-coding targets of piRNA-mediated deadenylation (Rouget et al. 2010). 3' UTRs were further filtered by the identification of a SRE, physical association of genic mRNA with Smaug, and recovery of transcript abundance in the absence of Smaug. Targets were prioritized by expression profiles, piRNA complementarity downstream of the SRE, presence of

expected piRNA sequence bias, transcript localization, and putative gene function associated with development.

PAT assays have been described, as continuing work, to serve as wet-lab validation of the genic targets. PAT assays involve the ligation of a known sequence adaptor 3' of the mRNA poly(A) tail. Primers can then be designed, using the 3' translated region and the known adaptor sequence, to assess the mRNA poly(A) tail length (Sallés et al. 1999).

Section 6.6: Significance and Future Direction

TEs and piRNAs have been ubiquitously found amongst Metazoans (Grimson et al. 2008). The foundational research described herein concerns piRNA biogenesis, targeting, and function contributes to the conceptual understand of the piRNA pathway, which has the potential to lead to robust and specific target suppression. TEs and piRNAs have various known implications in disease including, but not limited to fundamental embryonic development, Alzheimer's, ALS, Cancer, and viral immunity (Cheng et al. 2011; Li et al. 2012; Boudreau & Davidson 2006; Ding & Lu 2011; Vodovar et al. 2012). A better understanding of TEs and piRNAs is crucial in continuing to understand disease development and progression, genome evolution, and the exhaustive capability of RNAi.

Section 6.7: References

Aravin, A.A., Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *science*, 318(5851), pp.761–764.

Boudreau, R.L. & Davidson, B.L., 2006. RNAi therapy for neurodegenerative diseases. *Current topics in developmental biology*, 75, pp.73–92.

Brennecke, J. et al., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6), pp.1089–103.

Cheng, J. et al., 2011. piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. *Clinica chimica acta*, 412(17), pp.1621–1625.

Chirn, G. et al., 2015. Conserved piRNA expression from a distinct set of piRNA cluster loci in eutherian mammals. *PLoS Genet*, 11(11), p.e1005652.

Ding, S.-W. & Lu, R., 2011. Virus-derived siRNAs and piRNAs in immunity and pathogenesis. *Current opinion in virology*, 1(6), pp.533–544.

Edgar, R., Domrachev, M. & Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1), pp.207–210.

Grimson, A. et al., 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217), pp.1193–1197.

Li, W. et al., 2012. Transposable elements in TDP-43-mediated neurodegenerative disorders. *PloS one*, 7(9), p.e44099.

Malone, C.D. & Hannon, G.J., 2009. Small RNAs as guardians of the genome. *Cell*, 136(4), pp.656–68.

Rosenkranz, D. & Zischler, H., 2012. proTRAC-a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC bioinformatics*, 13, p.5.

Rouget, C. et al., 2010. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature*, 467(7319), pp.1128–1132.

Sallés, F.J., Richards, W.G. & Strickland, S., 1999. Assaying the polyadenylation state of mRNAs. *Methods*, 17(1), pp.38–45.

Sayers, E.W. et al., 2011. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1), pp.D38–D51.

Vodovar, N. et al., 2012. Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PloS one*, 7(1), p.e30861.