1-1-2014

Transposable Element Content in Non-Model Insect Genomes

Christine A. Lavoie

Transposable element content in non-model insect genomes

By

Christine A. Lavoie

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Agricultural Life Sciences
in the Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology

Mississippi State, Mississippi

May 2014

Transposable element content in non-model insect genomes

By

Christine A. Lavoie

Approved:

_____

David A. Ray
(Major Professor)

_____

Brian A. Counterman
(Co-Major Professor)

_____

Scott T. Willard
(Committee Member)

_____

Din-Pow Ma
(Graduate Coordinator)

_____

George M. Hopper
Dean
College of Agriculture and Life Sciences

Name: Christine A. Lavoie

Date of Degree: May 16, 2014

Institution: Mississippi State University

Major Field: Agricultural Life Sciences

Major Professors: David A. Ray and Brian A. Counterman

Title of Study: Transposable element content in non-model insect genomes

Pages in Study: 74

Candidate for Degree of Master of Science

While the study of transposable element evolution has been conducted in several model insect organisms such as *Anopheles gambiae*, *Drosophila melanogaster*, and *Bombyx mori*, little investigation has been conducted into the transposable element (TE) evolution within less commonly examined model and non-model taxa within Diptera. In this work we contributed two analyses to close this gap. First, TEs in the lepidopteran, *Heliconius melpomene*, were characterized, and it was determined that 25% of the genome is composed of TEs. Second, TEs in oestroid and muscid flies were characterized using survey sequencing rather than whole genomes. Comparative analyses were performed on *Haematobia irritans*, *Sarcophaga crassipalpis*, *Phormia regina*, and *Cochliomyia hominivorax*. TE proportions were 5.95%, 10.00%, 22.43%, and 30.67%, for *C. hominivorax*, *P. regina*, *S. crassipalpis* and *H. irritans*, respectively. These studies provide new insights into the diversity of TEs in Insecta and suggest that in general, TE diversity is high among insects.

DEDICATION

This thesis is dedicated to my parents, John and Helen Lavoie, and my family, Queenie, Brat, Scooby, Nikeeta, Simone, C.T., Bullet, Waddles, Jerry Lee, Mika, Teagan, Talan, and Tillie.

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. David Ray, for his patience and guidance through the past 3 ½ years. His insight and experience helped to refine and sharpen my scientific skills. Next, I would like to thank my co-major professor, Dr. Brian Counterman, for allowing me the opportunity to collaborate with him and for allowing me to become a member of his lab during my last two semesters at Mississippi State. I would also like to thank Dr. Scott Willard for his guidance during my transition into the Department of Biochemistry and Molecular Biology. Dr. Willard's support during my graduate studies was invaluable.

Next, I would like to thank Dr. Din-Pow Ma for his time and support during my graduate studies. Dr. Ma's attentiveness to detail ensured my studies met the requirements at MSU. I would like to thank Dr. Daniel Peterson for his knowledge and expertise which was shared during his graduate course in Genomes and Genomics. Additionally, Dr. Peterson provided support through his laboratory and other means during my graduate studies.

Finally, I would like to thank my current and past lab members including Heidi Pagan, Mike Vandewege, Neal Platt, Meganathan Ramakodi, and Ananya Sharma who have given me suggestions and support during my time as a graduate student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Transposable elements (TEs) are DNA sequences that can mobilize within a genome. TEs can be divided into two groups, Class I and Class II. Class I or retrotransposons utilize a "copy and paste" mechanism to insert themselves into a new location in the genome. The retrotransposon at the original site is transcribed into an RNA intermediate. An enzyme, reverse transcriptase, reverse transcribes the RNA intermediate into a complementary DNA and the complementary DNA is integrated into a new site in the genome. Retrotransposons can be classified as either long terminal repeats (LTRs) or non-LTRs. LTRs contain long terminal repeats on both ends of the element while non-LTRs do not contain any long terminal repeats. Non-autonomous retrotransposons do not encode reverse transcriptase and rely on an autonomous partner. Short interspersed elements or SINEs are an example of non-autonomous retrotransposons. SINEs can be derived from tRNAs, 5S rRNAs, or 7SL rRNAs. Class II elements consist of the DNA transposons which utilize a "copy and paste" mechanism to move around the genome. DNA transposons encode a transposase that is responsible for excising the TE from its original site and moving it to a new site. A few examples of Class II elements include Tc1/Mariner, hat, and piggyBac elements. Another category of DNA transposons includes the Helitrons which mobilize in the genome via a rolling circle mechanism (Figure 1.1).

TEs can affect the function and structure of genomes in several ways. One is through chromosomal rearrangements. For example, deletions, translocations, and inversions have been associated with TEs. In Drosophila, class I and class II elements have been linked to chromosomal rearrangements (Lim and Simmons 1994). In humans, LINEs and SINEs have been linked to at least 44% of inversions (Lee, Han et al. 2008). Another way in which TEs can affect the genome is through gene disruption. An insertion of a TE into a gene may be deleterious to the genome and result in a genetic disorder. For example, individuals diagnosed with Coffin-Lowry syndrome contained a L1 insertion in the *RPS6KA3* gene which resulted in exon 4 being skipped and a reading frame shift (Martinez-Garay, Ballesta et al. 2003).

While numerous studies have analyzed the importance of transposable elements throughout Insecta, much remains to be learned. The order Lepidoptera consists of butterflies and moths and there are approximately 200,000 species world-wide (Gilliot 2005). Within Lepidoptera, only three species' genomes have been fully sequenced, Bombyx mori, Danaus plexipus, and Heliconius melpomene. The TE landscape of B. mori has been examined and shown to harbor a wide range of TE diversity. Analyses of the TE landscape of D. plexipus are incomplete, so a TE-based comparison of D. plexipus to other lepidopterans is not yet available. In Chapter II, I examine the TE content and activity of Heliconius melpomene, the Postman butterfly. H. melpomene, is found throughout Central and South America (Brower 1996). The Heliconius Genome Consortium sequenced the genome of a male Heliconius melpomene melpomene from Panama with 454 sequencing and Illumina chemistries (2012). The TE content of H. melpomene was characterized and activity periods were estimated. Several novel

elements were identified and, while Class I elements exhibited a lack of recent activity, some Class II elements exhibited hallmarks of ongoing mobilization. Furthermore, I found evidence that longer elements are subject to removal from the genome, likely via ectopic recombination. The results suggest that selection is acting to retain a small genome and that lepidopterans in general will likely be a rich source of diverse TEs.

Another large insect order is Diptera, with approximately 150,000 species and approximately 180 families (Bertone 2009). Our understanding of transposable elements in Diptera is primarily due to the research of TEs in Drosophila and mosquitoes (Tu 1997; Tu 2001; Deninger 2002). Muscoidea and Oestroidea are superfamilies in the subsection Calyptratae (Yeates 2005). Little work has been done to characterize the TE landscape of either superfamily. The superfamily Oestroidea harbors six families that include Calliphoridae, Sarcophagidae, Oestridae, Mystacinobiidae, Rhinophoridae, and Tachinidae (Yeates 2005). In regards to the investigation of TEs in the superfamily Oestroidea, published findings include a hAT-like element and a P-like element that was identified in Lucilia cuprina (Perkins 1992; Coates 1996). A study involving Calliphora vicina identified a vast diversity of Class I and Class II elements in a 600 kb region of the genome (Negre 2013). In Chapter III, I examine TE content in two calliphorid flies, a sarcophagid fly, and a muscid fly using 454 survey sequencing. The analyses revealed that there is a substantial amount of TE diversity within the four species analyzed which may be a reflection of the evolutionary history and function of the TEs in each species.

Finally, Chapter IV discusses the implications that the results from the previous chapters have on the diversity of TEs. It also discusses how these results may be utilized in future studies.

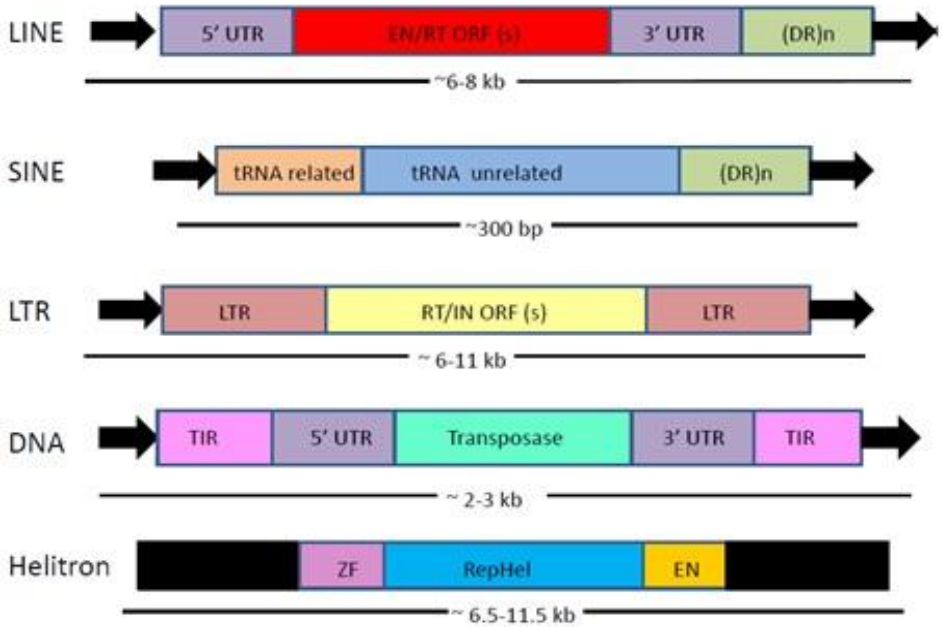

Figure 1.1 Schematic representation of Retrotransposons and DNA transposons. Figure modified from (Kapitonov and Jurka 2007; Ray, Platt et al. 2009)

Literature Cited

Bertone, M. a. W., B. (2009). True flies (Diptera). The Timetree of Life. S. B. a. K. Hedges, S. New York, Oxford University press.

Brower, A. (1996). "Parallel Race Formation and the Evolution of Mimicry in Heliconius Butterflies: A Phylogenetic Hypothesis from Mitochondrial DNA Sequences." Evolution 50(1): 195-221.

Coates, C. J. e. a. (1996). "The hermit transposable element of the Australian sheep blowfly, Lucilia cuprina, belongs to the hAT family of transposable elements." Genetica 97(1): 23-31.

Consortium, H. G. (2012). "A butterfly genome reveals promiscuous exchange of mimicry adaptations among species." Nature 487(7405): 94-98.

Deninger, P. L., and A.M. Roy-Engel (2002). Mobile Elements in Animal and Plant Genomes. Mobile DNA II. C. N.L. Craig, R., Gellert, M., Lambowitz, A.M. Washington D.C., ASM Press.

Gilliot, C. (2005). Entomology. The Netherlands, Springer.

Kapitonov, V. V. and J. Jurka (2007). "Helitrons on a roll: eukaryotic rolling-circle transposons." Trends Genet 23(10): 521-529.

Lee, J., K. Han, et al. (2008). "Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons." PLoS One 3(12): e4047.

Lim, J. K. and M. J. Simmons (1994). "Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster." Bioessays 16(4): 269-275.

Martinez-Garay, I., M. J. Ballesta, et al. (2003). "Intronic L1 insertion and F268S, novel mutations in RPS6KA3 (RSK2) causing Coffin-Lowry syndrome." Clinical Genetics 64(6): 491-496.

Negre, B., Simpson, P. (2013). "Diversity of transposable elements and repeats in a 600 kb region of the fly Calliphora vicina." Mobile DNA 4(13).

Perkins, H. D., and A.J. Howells (1992). "Genomic sequences with homology to the P element of Drosophila melanogaster occur in the blowfly Lucilia cuprina." Proc. Natl. Acad. Sci 89(22): 10753-10757.

Ray, D. A., R. N. Platt, et al. (2009). "Reading between the LINEs to see into the past." Trends Genet 25(11): 475-479.

Tu, Z. (1997). "Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, Aedes aegypti." Proceedings of the National Academy of Sciences of the United States of America 94(14): 7475-7480.

Tu, Z. (2001). "Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae." Proc. Natl. Acad. Sci 98(4): 1699-1704.

Yeates, D. K. a. W., B. (2005). The Evolutionary Biology of Flies, Columbia University Press.

CHAPTER II

TRANSPOSABLE ELEMENT EVOLUTION IN *HELICONIUS* SUGGEST GENOME DIVERSITY WITHIN LEPIDOPTERA

Abstract

In order to understand the contribution of transposable elements (TEs) to *Heliconius melpomene*, we queried the *H. melpomene* draft sequence to identify repetitive sequences. We determined that TEs comprise ~25% of the genome. The predominant class of TEs (~12% of the genome) was the non-long terminal repeat (non-LTR) retrotransposons, including a novel SINE family. However, this was only slightly higher than content derived from DNA transposons, which are diverse, with several families having mobilized in the recent past. Compared to the only other well-studied lepidopteran genome, *Bombyx mori*, *H. melpomene* exhibits a higher DNA transposon content and a distinct repertoire of retrotransposons. We also found that *H. melpomene* exhibits a high rate of TE turnover with few older elements accumulating in the genome, suggesting that TEs have an overall deleterious effect and/or that maintaining a small genome is advantageous for this taxon.

Introduction

Transposable elements (TEs) are segments of DNA that can mobilize in a genome. They impact the structure and function of the genomes they occupy. TEs can be divided into two classes. Class I TEs are the retrotransposons, which require an RNA intermediate and use a "copy and paste" mechanism to insert themselves into a new location in the genome. Retrotransposons are further divided into two groups, the long terminal repeat elements (LTRs) and non-LTR elements. LTR retrotransposons, such as members of the Gypsy and Copia superfamilies, are similar in structure to some retroviruses. Non-LTR retrotransposons lack LTR sequences and autonomous versions (Long INterspersed Elements or LINEs) usually harbor one or two open reading frames (ORFs) that are responsible for their mobilization. Examples include the LINE1, CR1, and RTE superfamilies and can be categorized into 28 monophyletic clades (Kapitonov 2009). Short INterspersed Elements (SINEs) are a group of nonautonomous non-LTR retrotransposons that are mobilized via the enzymatic machinery of LINEs (Dewannieux 2003).

Class II elements include the DNA transposons which use a "cut and paste" mechanism to mobilize in the genomes they occupy. Typically, DNA transposons require a transposase enzyme to recognize the terminal inverted repeats (TIRs) of the transposon and then excise and reinsert the element into another location in the genome (Biemont 2006). Examples of Class II elements include members of the TcMariner, hAT, and piggyBac superfamilies. There is a second group of Class II TEs known as the rolling circle transposable elements that includes the Helitrons (Kapitonov 2001).

The first lepidopteran to have its whole genome sequenced, the silkworm moth *Bombyx mori*, has accumulated a diverse array of retrotransposons and DNA transposons (Osanai-Futahashi 2008). For instance, a non-LTR retrotransposon, L1Bm, is abundant in the genome with copies of the 3‟ end numbering ~25,000. However, like many LINEs most copies are 5‟ truncated (Ichimura 1997). Multiple copies of a piggyBac-like DNA transposon that may harbor an intact transposase have also been found in *B. mori* and it appears to have been recently active along with other Class II elements (Daimon 2010).

Recently, the genome of *Heliconius melpomene* was released (Consortium 2012), providing new insights into lepidopteran genome evolution from a transposable element perspective. *H. melpomene* is a heliconiine butterfly that is widespread throughout Central America and South America (Brower 1996; Consortium 2012). The *H. melpomene* genome is the third lepidopteran and second butterfly genome to be sequenced. Unfortunately, the analysis of the second genome (and the first butterfly), the monarch, *Danaus plexipus*, was not comprehensive (Zhan 2011). Therefore, we confine our comparisons of the *H. melpomene* genome to *B. mori*.

Our analyses indicate that *H. melpomene* exhibits a high rate of TE turnover, with little accumulation of older elements, especially longer, autonomous elements, suggesting that TEs have an overall deleterious effect on the genome. Furthermore, the TE landscape of *H. melpomene* is distinct compared to the silkworm moth, consisting of substantially higher Class II content and a distinct set of retrotransposons. This suggests that lepidopterans in general will exhibit high levels of TE diversity as additional genomes are sequenced and characterized.

Methods

The genome sequence of a male *Heliconius melpomene melpomene* was recently described (Consortium 2012). Briefly, the specimen was acquired from Darien, Panama and the genome was sequenced using both 454 and Illumina platforms to generate a 38X draft genome. The sequenced male was inbred for five generations of sib mating. Repeat discovery was performed as summarized elsewhere (Consortium 2012) and described briefly here. Repetitive sequences in the *H. melpomene* draft sequence (Genbank accession number: CAEZ01000000) were identified *de novo* using RepeatModeler (Smit 2008-2010) To infer the consensus sequences for each repeat, we used the filtered RepeatModeler output to query the entire WGS draft using BLAST v2.2.23 (Altschul, Madden et al. 1997). Up to fifty of the top hits spanning at least 100 bases were extracted along with up to 1,000 bases of flanking sequence, and we aligned the extracted sequences with MUSCLE 4.0 (Edgar 2004) to generate 50% majority rule consensus sequences. Consensus sequences were considered „complete" when single copy sequence could be identified at the 5' and 3' ends in each component sequence. If this condition was not met, the process was repeated until single copy DNA sequence was identifiable at both ends. The resulting library was submitted to CENSOR (Kohany, Gentles et al. 2006), BLASTN and BLASTX to ascertain the identity of the consensus with regard to previously classified elements. The result was a custom library of elements, which served as our library for subsequent analyses. The library of TEs was passed through a locally implemented version of RepeatMasker (Smit 1996-2010) to estimate the TE content of the *H. melpomene* genome.

Identification of SINE subfamilies

We identified 14,196 intact insertions of *Metulj* between 240 - 294 bases in length (+/- 10% of the general consensus) and passed them to COSEG (Price, Eskin et al. 2004; Smit and Hubley 2008-2013) for subfamily identification. COSEG examines multiple instances of TE insertions and identifies significant co-segregating (2-3 bp) sites in an effort to determine subfamily structure. A perl script provided by R. Hubley was used to refine the consensus sequence for each subfamily and is available upon request. We created a custom RepeatMasker library consisting of the suggested *Metulj* subfamily consensus sequences and extracted the top 150 hits for each from the genome. We aligned the extracted sequences with their respective subfamily consensus sequence to confirm the presence of each in the genome.

Identification of Intact ORFs

We submitted the consensus sequence of each TE to NCBI ORF finder to identify potential open reading frames (ORFs). We classified any elements with identifiable ORFs spanning 1000 bp or more as potentially full length. ORF sequences were translated and BLASTP was used to confirm identity. ORFs of BEL-1_HMM, BEL-2_HMM, Copia-1_HMM, Gypsy-10_HMM, Gypsy-1_HMM, Gypsy-2_HMM, Gypsy-3_HMM, Gypsy-4_HMM, Gypsy-5_HMM, Gypsy-6_HMM, Gypsy-7_HMM, Gypsy-8_HMM, Gypsy-9_HMM as well as RTE-1_HMe, R4-1_Hme were identified by other parties and were obtained from RepBase.

We estimated the number of intact ORFs for each family of autonomous elements by passing the ORF sequences through a local version of TBLASTN, after which, up to 50 of the top hits based on bit score were extracted with 1000 bp of buffer and aligned.

Extracted sequences were trimmed so they began and ended at the same position as the ORF query sequence. We defined an intact ORF as one that is greater than or equal to 90% of the expected amino acid length, contains a single, terminal stop codon, and begins with a methionine start codon.

Age analyses and relative insertion periods:

We used the TinT online server (http://www.compgen.uni-muenster.de) as a method to determine periods of relative TE activity and succession patterns (Churakov 2010). Due to low copy numbers, analysis of LTR elements could not be performed. Furthermore, DNA transposons utilize a cut-and-paste mechanism of transposition that makes a nested insertion analysis of this type less informative. Thus, we analyzed only non-LTR retrotransposons.

We also estimated activity periods based on genetic distances between individual insertions and the consensus of each subfamily as described previously (Ray 2008; Pagan 2010). Briefly, we created a modified TE library consisting of the full consensus of all *Metulj* subfamilies and non-autonomous DNA transposons, the full ORFs of all DNA transposons and 500 bp from the 3‟ end of non-LTR ORFs. This library was then used to query the genome using RepeatMasker. We estimated Kimura2-parameter (Kimura 1980) distances (including CpG sites) between each insertion and its respective consensus (Pagan 2010). A neutral mutation rate is not available for *H. melpomene.* We applied an estimated mutation rate of 0.01909 substitutions per site/per million years which was taken from Papilioninae, a subfamily of the butterfly family Papilionidae (Simonsen 2010).

The nearly vertical succession of non-LTR retrotransposons seen in the TinT plot (Figure A.1) suggests a rapid turnover of longer elements. One mechanism through which elements can be removed from a genome is non-homologous recombination leading to large deletions. By taking each RepeatMasker hit from each TE subfamily and mapping its location along the consensus element, we were able to examine decay patterns among selected elements.

Evolutionary relationships among autonomous Non-LTR retrotransposons

From Genbank and Repbase, we collected non-LTR retrotransposon protein sequences from diverse known clades (Jurka 2005; Benson 2010). We aligned these sequences with the consensus sequences retrieved from the *H. melpomene* genome using Clustal W in BioEdit (Hall 1999). The most conserved region (about 300 amino acids) from the reverse transcriptase domain was identified and used in the phylogenetic analysis. Newly identified families missing this region were excluded. We inferred a maximum-likelihood tree with 1,000 bootstrap replicates using MEGA5 (Tamura 2011).

Horizontal Transfer

We investigated the taxonomic distribution of all *H. melpomene* TEs by querying the full WGS database at NCBI with BLAST. We considered an element to be a likely candidate for HT if a BLASTN search indicated that the consensus shared >95% sequence identity over at least 80% of its length. Any hits matching these criteria were examined by extracting the highest scoring hits, alignment to the query sequence and manual examination.

Results

TEs comprise ~25% of the *H. melpomene* genome (Table 2.1). The majority are non-LTR retrotransposons (12.07% of the genome), and among these, Short INterspersed Elements (SINEs) make up the greatest proportion (8.22%). The second most common group in *H. melpomene* are the DNA transposons, comprising 10.05% of the genome and dominated by Helitrons (~5.37% of the genome). LTR elements were also found, but occupy a much smaller proportion of the genome (0.45%).

Identification of *Metulj* and its subfamilies

One novel element from the genome was a SINE family we have dubbed Metulj (meh-TOOL), Slovenian for butterfly. The Metulj general consensus is ~267 bases in length with minor length differences depending on subfamily. The 5' region of Metulj contains the typical RNA polymerase III promoters separated by 30 bp (Figure 2.1). We identified a secondary structure reminiscent of a tRNA using the methods described in (Okada 2004), suggesting that the family, like many SINEs, is tRNA derived and consists of two regions, a tRNA head and a non-tRNA tail. Results from COSEG (Price, Eskin et al. 2004; Smit and Hubley 2008-2013) suggest that Metulj comprises eight major subfamilies (Figure 2.2). However, subfamilies 3 and 4, appear to be composite TEs, instances where Metulj elements inserted into other active elements which then continued to mobilize. For example, Metulj subfamily 3 is embedded within a non-autonomous Mariner element, nMar-16_Hm (7,770 copies), while an unidentified repetitive sequence (21,461 copies) includes both Metulj subfamily 4 and a Helitron-like element (data not shown). Because these two predicted subfamilies were likely distributed throughout in the genome by mechanisms other than retrotransposition, they were not included in

analyses of SINE dynamics. Metulj subfamily 3 likely expanded as a consequence of nMar-16_Hm mobilization. Given that the identity of the repetitive element into which Metulj subfamily 4 has embedded is unknown, we cannot speculate on its expansion mechanism.

Age analyses and relative insertion rates

Divergence estimates indicate that the majority of *Metulj* activity occurred in the distant past (Table 2.2), *Metulj-2_Hm* appears to be the youngest, with an average divergence from the consensus of ~5%. The topology of the *Metulj* tree generated as part of a COSEG (Price, Eskin et al. 2004) analysis supports the divergence analyses (Figure 2.2). For example, *Metulj-2_Hm* is a near-terminal node and exhibits the lowest level of divergence, while *Metulj-0_Hm* and *7*, which are estimated to be older are found nearer the root. Analyses of nested insertions via TinT (Churakov 2010) also supports this arrangement with *Metulj-0* and *7*, both of which exhibit high divergence levels, harboring proportionally more nested insertions than other subfamilies (Figure A.1). There does not appear to be any recent SINE activity in the *H. melpomene* genome. This could be due to inactivation and subsequent removal (see below) of the autonomous LINE partner for *Metulj*. Indeed, we are unable to identify the likely autonomous partner for this SINE family, because most older LINE families are present only as incomplete „fossils" in the genome.

Autonomous non-LTR elements exhibit a similar lack of recent activity with mean periods of activity ranging from ~2.7 mya to over 21 mya. A general lack of retrotransposition competence is suggested when examining numbers of potentially intact ORFs. We were unable to identify intact ORFs for most autonomous retrotransposon

families and, of the families with identifiable, intact ORFs, the numbers were generally small. The largest number of intact ORFs was for RTE-3_Hm, with six (Table 2.3). The lack of success in identifying intact ORFs could be attributed to problems with the assembly. Most breaks in an assembly are associated with highly similar TE insertions. However, we were able to identify multiple instances of relatively long and highly similar sequences (see the discussion of Tc3-1_Hm below), suggesting instead that intact non-LTR ORFs, if present, would not evade detection.

DNA transposons exhibit a much different pattern of succession with multiple lineages exhibiting relatively recent activity (i.e. mean activity periods estimated within the last 2 my; Table A.1). Only three autonomous DNA transposon families were identified in the genome but one stands out. Tc3-1_Hm exhibits an average divergence of 0.002% among 113 full length insertions. A total of 43 intact ORFs are present, suggesting that this family is a recent and active addition to the TE repertoire of *H. melpomene*. However, no intact transposase ORFs other than Tc3-1_Hm were evident. A second standout is the Helitron superfamily, which also appears to have undergone a relatively recent amplification and is the most prevalent Class II element, occupying ~5% of the genome. Several other element families also appear to be young and active. These include multiple nonautonomous families of the piggyBac, Mariner, hAT and Helitron superfamilies and the two autonomous piggyBac elements. For the purposes of this study, MITEs (miniature inverted repeat transposable elements) were considered a subset of non-autonomous DNA transposons.

Evidence of TE removal

As part of their mobilization non-LTR retrotransposons are reverse transcribed from their 3' end. Large non-LTR retrotransposons are often truncated at the 5‟ end and this is thought to be a consequence of either premature dissociation of reverse transcriptase or the activity of cellular RNases (Ustyugova 2005). However, the presence of a 5‟ region without the corresponding 3‟ region is not likely to be result of either process. Thus, LINE fragments that lack their 3‟ ends or consist solely of internal sections are considered evidence of genomic deletions as described previously (Novick 2009; Blass 2012). We found that many *H. melpomene* LINE families exhibited patterns consistent with large deletions acting to remove them from the genome (Figure 2.3 and Figure A.2). As expected given their insertion mechanism, we observe an abundance of 3‟ fragments for LINE families. However, unlike what is observed in mammals (Blass 2012), which exhibit a low rate of DNA loss, we see a large number of 5‟ fragments and orphaned internal LINE fragments. This suggests ectopic recombination acting to remove these elements from the genome at a high rate.

Evolutionary relationships among autonomous Non-LTR retrotransposons

A maximum-likelihood tree of autonomous non-LTR retrotransposons (Figure 2.4) reveals that the *H. melpomene* genome harbors 56 families from 10 diverse clades (L2, CR1, Vingi, Daphne, R1, I, Jockey, Proto2, RTE and R4). Although most clades (7/10) have relatively low diversity (three or fewer representatives within the clade), the remaining clades are represented by many families. The L2 and RTE clades are each represented by 13 families, while the Jockey and CR1 clades each contain seven. Zenon is sometimes considered a member of the CR1 clade, thereby raising the count to ten for

that family. Although most of the non-LTR consensus sequences that were generated cluster with their appropriate clade, three CR1 families (CR1-6_Hm, CR1-8_Hm, CR1-1_Hm) fail to do so bootstrap support (greater than 65). Despite the fact that RepeatMasker identifies these elements as CR1, these families form a monophyletic group sister to Daphne elements and may represent a novel clade.

Horizontal Transfer

We considered an element to be a likely candidate for horizontal transfer (HT) if a BLASTN search indicated that the consensus shared >95% sequence identity over at least 80% of its length. BLAST results from querying NCBI"s WGS database suggest three candidate elements for horizontal transfer between *H. melpomene* and other animals (Figure 2.5). The first involves a non-autonomous hAT-like element, *nhAT-10_Hm* with hits to scaffolds in *Rhodnius prolixus* (best hit = 97% identity over 83% of the query, E-value = 0), *Mengenilla moldrzyki* (96% identity over 83% of the query, E-value = 0), and *Schmidtea mediterranea* (95% identity over 83% of the query, E-value = 0). *R. prolixus* and *M. moldrzyki* are insects from the orders Hemiptera and Strepsiptera, respectively. The fact that similar hits were not observed in more closely related taxa such as *B. mori* or *D. plexipus* is evidence that these elements were likely transferred to the genome by mechanisms other than vertical transmission.

The other two candidates were piggyBac-1_Hm and piggyBac-2_Hm with hits matching our criteria in *Manduca sexta* (pibbyBac-1_Hm, 99% identity over 100% of the query, E-value = 0), *Bombyx mori* (and piggyBac-2_Hm, 99% identity over 100% of the query, E-value = 0), and *D. plexipus* (and piggyBac-2_Hm, 98% identity over 85% of the

query, E-value = 0). In the case of *D. plexipus* , the reduced coverage is due to the fact that the insertion terminates with the scaffold (AGBW01001888).

Discussion

TE content in Heliconius compared to Bombyx

The genome *of Heliconius melpomene* is the third lepidopteran genome to be fully sequenced. Unfortunately, the authors of the monarch genome manuscript did not complete a comprehensive analysis of the TE landscape (Zhan 2011), and our comparisons were therefore limited to *B. mori*.

TEs make up 35% of the *B. mori* genome, with the largest fraction (26.6%) being non-LTR retrotransposons (Osanai-Futahashi 2008). Of the non-LTR content, around half is derived from SINEs, 48%. A smaller fraction, ~25%, of the *H. melpomene* genome is composed of TEs. 12.5% consists of non-LTR retrotransposons, and 8% of the genome is occupied by SINEs (68% of the non-LTR content). Thirty-two non-LTR families belonging to 12 clades (Jockey, RTE, CR1, CRE, R1, R2, R4, I, Vingi, Daphne, Proto2 and L2) were identified and classified from *B. mori*. This is two more than were identified in *H. melpomene*. However, despite harboring two fewer clades than B. mori, the *H. melpomene* genome contains more families in total and this can be attributed to higher within-family diversity in some clades. For instance, 13 families of L2 and 10 families of CR1 were identified in *H. melpomene*, while only one and two are present in *B. mori*, respectively. Of the available lepidopteran genomes (including the monarch butterfly), Metulj is restricted to Heliconius.

In *H. melpomene*, LTRs make up only ~0.45% of the genome. This is within the same range as what was described for *B. mori* by Osanai-Futahashi et al. in 2008

(Osanai-Futahashi 2008), 1.7%, but substantially different from a second estimate of LTR content in B. mori by Jin-Shan et al. (Jin-Shan 2005), 11.8%. Given that Osanai-Futahashi examined a more complete assembly of the silkworm genome, we suspect that their estimate is closer to reality. Both genomes harbor Gypsy and Copia elements. *B. mori* however has two additional families which include Pao and Micropia (Osanai-Futahashi 2008). That being said, ~2.4% of the genome consists of candidate TEs that remain unidentified by our analyses and could belong in the LTR category.

While the retrotransposon content of *B. mori* and *H. melpomene* are similar, with regard to Class II elements, the DNA transposons, the two species are strikingly different in both content and quantity. Only ~3% of the *B. mori* genome consists of Class II elements (Osanai-Futahashi 2008) while ~10% of the *H. melpomene* genome is derived from DNA transposons. Indeed, the butterfly genome has been the subject of considerable DNA transposon activity within the recent past. This includes massive amplification by the Helitron superfamily and very recent, if not ongoing activity, from one member of the Tc-Mariner family. At least 43 intact members of the Tc3-1_Hm autonomous element are present in the genome draft and they are 99.4% identical, indicating that these elements are likely active.

Turnover of non-LTR element families in Heliconius

The lack of intact, older LINE elements in the genome suggests that they have a high fitness cost and that they may be preferentially removed. Mechanisms to accomplish removal include ectopic recombination between similar elements and removal of individual insertions via selection. Indeed, increased rates of ectopic recombination have been suggested as a mechanism for the differences in TE

accumulation in both mammals and insects (Eickbush 2002). Our results suggest that this mechanism is in play in the *H. melpomene* genome. Figure 2.3 indicates that deletions of large portions of LINE elements occur at relatively high frequency.

That being said, we note that other elements families have accumulated to relatively high numbers. In particular, this is true of *Metulj* and many of the Helitron elements. However, those elements with high copy numbers are typically under 500 nt in length. Previous authors have noted that shorter elements are likely less prone to recombination than their longer cousins (Cooper 1998; Song 2007), allowing them to remain in the genome.

Hierarchical insertion patterns (TinT) indicate short periods of activity for the longer, autonomous elements, which exhibit a clear pattern of succession (Figure A.1). If one ignores the wide distributions of *Metulj*, the only SINE, each non-LTR family occupies a relatively narrow temporal space indicating that they experience brief periods of activity before ceasing mobilization. This is similar to what has been observed in some other taxa, including the lizard *Anolis carolinensis*, but is distinct from mammals, which have a single lineage of LINE-1 that has accumulated high copy numbers (Furano 2004). The same analysis was performed for *B. mori*, with similar results (Figure A.1). Like many insects, the *H. melpomene* genome is relatively small, ~269 Mb. These results suggest that, while TE activity occurs and novel elements can invade the genome with some success, strong selection is working against the accumulation of large TEs and that homologous recombination acts to rapidly disable elements and keep the genome compact.

Evidence of horizontal transfer

We found evidence of horizontal transfer of three DNA transposons between *H. melpomene* and other taxa. Multiple elements matching nhAT10_Hm were identified in three taxa, the triatomine bug, *Rhodnius prolixus*, a strepsipteran insect, *Mengenilla moldrzyki*, and the planarian, *Schmidtea mediterranea*. In each case, the entire nhAT10_Hm is present as part of a larger element. For example, when compared to the planarian autonomous element, hAT-11_SM, nhAT10_Hm_has the hallmarks of an internal deletion variant. The first 70 bases are essentially identical between both TEs, as are the last 420 (Figure 2.5). The same regions overlap with as yet unnamed repeats in *R. prolixus* and *M moldrzyki*. The top hit for *R. prolixus* can be found on contig ACPB02011601.1, nt 29253-30319, and the top hits for *M. moldrzyki* can be found on contigs AGDA01050831.1, nt 10068-10485 and AGDA01007612.1, nt 6860-6920, respectively. In these two taxa, the overlaps are with elements that are likely nonautonomous. This suggests that a hAT-22_SM-like element has been invading multiple genomes and produced similar nonautonomous variants in each. Indeed, we subsequently used BLASTN to query the genome drafts of *H. melpomene*, *M. moldrzyki* and *R. prolixus* using the consensus sequence of hAT-11_SM and, while no full-length elements were obvious, we identified high scoring (E-value = 0) hits from various portions of the consensus in each. Interestingly, both *S. mediterranea* and *R. prolixus* have been implicated in horizontal transfer previously (GARCIA-FERNANDEZ, BAYASCAS-RAMIREZ ET AL. 1995; GILBERT, SCHAACK ET AL. 2010; NOVICK, SMITH ET AL. 2010).

The other candidates are the autonomous piggyBac elements, piggyBac-1_Hm and piggyBac-2_Hm. A single instance of *piggyBac-1_Hm* was identified in the

Manduca sexta genome draft (scaffold AIXA01012877) with 99% identity over its entire length. Two full length copies of *piggyBac-2_Hm* in the Dazao strain of *B. mori* (scaffolds AADK01008943 and AADK01013248) with the same values. The final hit, to the monarch butterfly genome, is incomplete due to the termination of the scaffold ~350 bp prior to the end of the consensus. Both moths would have diverged from the lineage leading to butterflies ~145 mya (Tu 1997) while the monarch is thought to have diverged from *Heliconius* ~89.79 mya (Wahlberg, Leneveu et al. 2009) and, given the high rate of change observed in lepidopteran genomes, it is unlikely that they would have been conserved over such an extended period. This suggests to us that horizontal transfer explains their presence in each. However, as additional genomes are characterized this interpretation could change.

Conclusions

In conclusion, by conducting the first full TE analysis of a butterfly we have demonstrated that TEs, specifically SINEs and Helitrons, make up a large portion of the *H. melpomene* genome. We identified a novel SINE family which is found only in *Heliconius* and demonstrated that the genome of *H. melpomene* has experienced recent DNA transposon activity, most notably a Tc3 element. We have also shown that older, intact LINE elements are not found within the genome and that their activity period in the genome is short due to their rapid removal. Further studies of other lepidopteran genomes will be beneficial to our understanding of TEs in lepidopterans.

Table 2.1 Summary of TE content in *Heliconius melpomene*

Class	Family	% Genome
DNA Transposons		**10.05%**
	Helitron	5.37%
	Mariner	2.13%
	Tc3	1.49%
	PiggyBac	0.32%
	hobo/Activator/Tam	0.38%
	Other/Unidentified	0.36%
LTR elements		**0.45%**
	Gypsy	0.21%
	Copia	0.00%
	Unknown	0.24%
Non-LTR elements		**12.07%**
SINE	Metulj	8.22%
LINEs		3.85%
	Daphne	0.45%
	RTE	0.89%
	Jockey	0.34%
	L2	0.41%
	Zenon	0.32%
	Other/Unidentified	1.44%
Unclassified		**2.37%**
Total		**24.94%**

Table 2.2 Divergence values and estimated activity periods for *Metulj* subfamilies

Metulj subfamily	Mean Distance	Standard Deviation	Range	Time (mya)
Metulj-0_Hm	0.20747	0.06289	0.14458-0.27036	7.6-14.2
Metulj-1_Hm	0.17328	0.07409	0.09919-0.24737	5.2-13.0
Metulj-2_Hm	0.1597	0.09649	0.06321-0.25619	3.3-13.4
Metulj-5_Hm	0.20597	0.06798	0.13799-0.27395	7.2-14.4
Metulj-6_Hm	0.20272	0.07116	0.13156-0.27388	6.9-14.3
Metulj-7_Hm	0.24241	0.06665	0.17576-0.30906	9.2-16.2

Table 2.3 Counts of intact open reading frames for each element class.

Class	Element Name	Coordinates of ORF	# Intact
DNA Transposon	**Tc3-1_Hm**	**120 - 1208**	**43**
NonLTR	Jockey-1_Hm	2980 - 4896	3
	Jockey-3_Hm	2051 - 4969	2
	L2-1_Hm	534 - 2924	1
	L2-7_Hm	95 - 1828	1
	L2-9_Hm	55 - 1530	3
	L2-13_Hm	543 - 2975	1
	L2-14_Hm	1468 - 4407	1
	L2-15_Hm	505 - 1986	1
	Proto2-3_Hm	111 - 1280	1
	R1-2_Hm	1411 - 4557	2
	R4-2_Hm	119 - 4207	1
	RTE-1_Hme	616 - 3636	1
	RTE-3_Hm	**264 - 3233**	**6**
	RTE-5_Hm	1334 - 3874	2
	RTE-9_Hm	723 - 1724	2
	RTE-10_Hm	323 - 1639	1
	RTE-15_Hm	69 - 1130	1
	RTE-20_Hm	181 - 3144	3
	TRAS1_R1_Hm	1299 - 3611	2
	Zenon-1_Hm	172 - 3333	1
	Zenon-2_Hm	590 - 3517	2
LTR	Gypsy-1_HMM-I	13 - 4542	1
	Gypsy-2_HMM-I	1071 - 5060	1
	Gypsy-3_HMM-I	2741 - 4402	1
	Gypsy-5_HMM-I	52 - 1716	1
	Gypsy-5_HMM-I	2601 - 4148	1
	Gypsy-6_HMM-I	84 - 2540	1
	Gypsy-6_HMM-I	**3008 - 4198**	**5**
	Gypsy-7_HMM-I	49 - 1272	1
	Gypsy-7_HMM-I	1694 - 3433	1
	Gypsy-8_HMM-I	1260 - 3167	1
	Gypsy-10_HMM-I	1525 - 3489	1

Counts of intact open reading frames for full length consensus sequences of each element class. Counts were determined as describe in the text. Bolded elements indicate the highest count in each category.

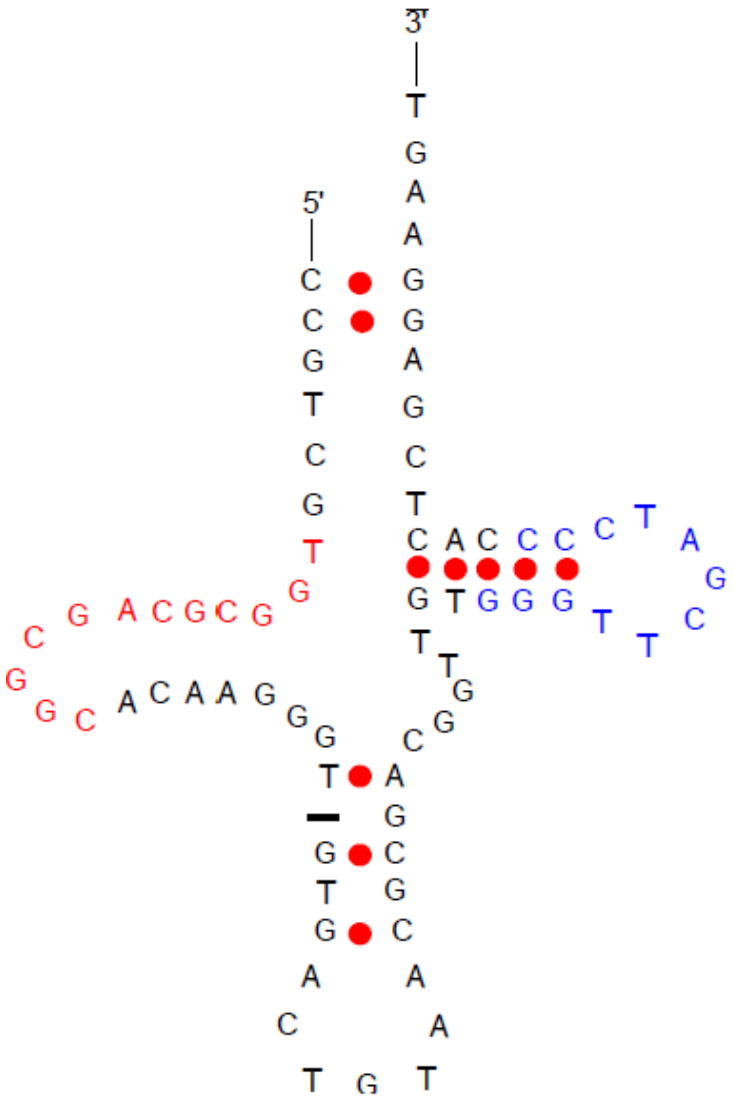

Figure 2.1 Pr

The first 73 bases of the H. melpomene SINE, Metulj, illustrating the predicted seconda structure of the presumed tRNA-derived region. The colored nucleotides identify the putative A (red) and B (blue) promoter regions.

ry

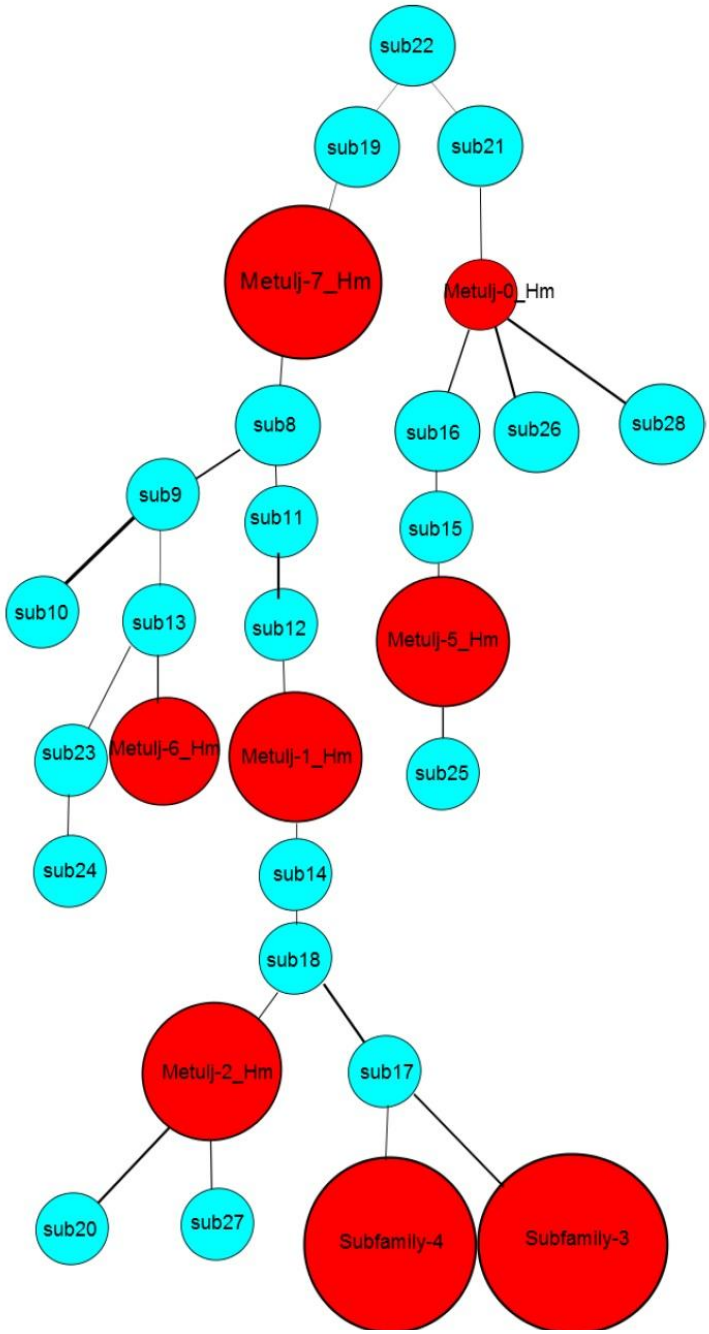

Figure 2.2 Results of the COSEG analysis.

Red circles are proposed subfamilies.

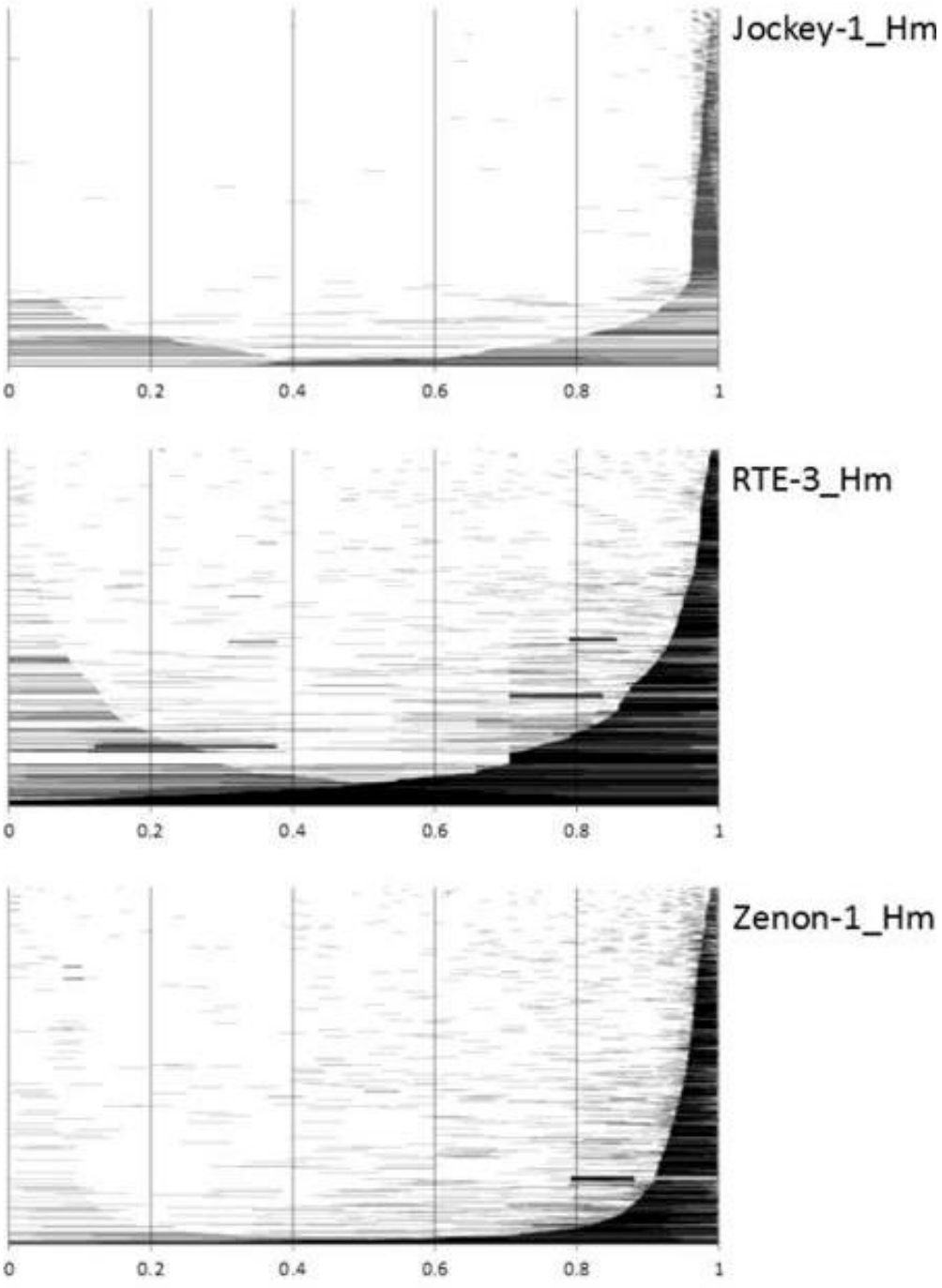

Figure 2.3 Length distribution of three *H. melpomene* LINE insertions.

Insertions are ordered from bottom to top by length (longest insertions at the bottom). Numbers along the x-axis are normalized to reflect length proportions relative to the total length of the family consensus.

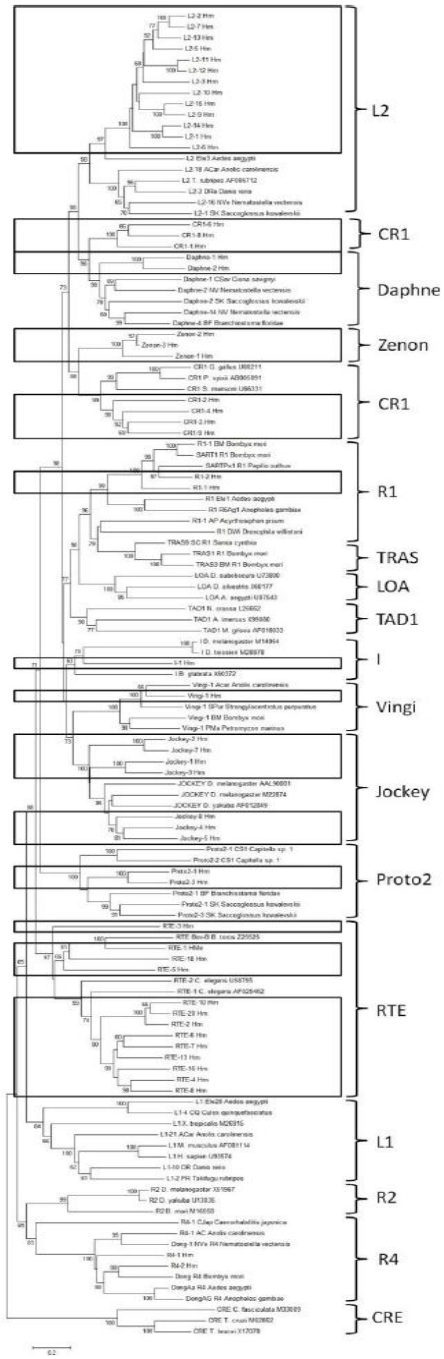

Figure 2.4 Phylogenetic relationships of autonomous non-LTR elements. Relatively weak bootstrap values (< 65) were not included.

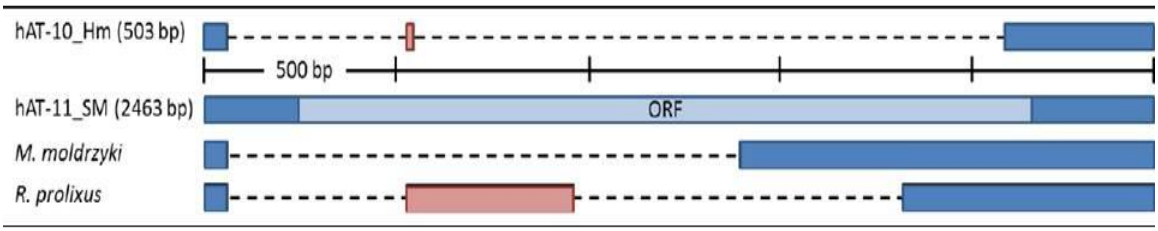

Figure 2.5 Relationships among hits highly similar to hAT-10_Hm in other taxa.

Comparisons are to hAT-11_SM, the consensus sequence of a known autonomous DNA transposon from the planarian, *Schmidtea mediterranea*, and contigs from *M. moldrzyki* and *R. prolixus*. Blue boxes exhibit high similarity within the corresponding regions. Red boxes found for *H. melpomene* (nine bases) and *R. prolixus* (410 bases) indicate regions with no similarity to any corresponding sequence in the other taxa. Contig IDs and sequence similarity values are available from the text.

Literature Cited

Altschul, S., T. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucl. Acids Res. **25**(17): 3389-3402.

Benson, D. A. (2010). "Genbank." Nucleic Acids Research **38**: D46-D51.

Biemont, C., Vieira, C. (2006). "Junk DNA as an evolutionary force." Nature **443**(5): 521-524.

Blass, E., Bell, M., Boissinot, S. (2012). "Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback." Genome Biology and Evolution **4**(5): 687-702.

Brower, A. V. Z. (1996). "Parallel race formation and the evolution of mimicry in *Heliconius* butterflies: A phylogenetic hypothesis from mitochondrial DNA sequences." Evolution **50**(1): 195-221.

Churakov, G., Grundmann, N., Kuritzin, A., Brosius, J., Makalowski, W., Schmitz J (2010). "A novel web-based TinT application and the chronology of the primate Alu retroposon activity." BMC Evolutionary Biology **10**: 376.

Consortium, H. G. (2012). "Butterfly genome reveals promiscuous exchange of mimicry adaptations among species." Nature **487**(7405): 94-98.

Consortium, H. G. (2012). "A butterfly genome reveals promiscuous exchange of mimicry adaptations among species." Nature **487**(7405): 94-98.

Cooper, J., Watanabe, Y., Nurse, P. (1998). "Fission yeast Taz1 protein is required for meiotic telomere clustering and recombination." Nature **392**: 828-831.

Daimon, T., Mitsuhiro, M., Katsuma, S., Abe, H., Mita, K., Shimada, T. (2010). "Recent transposition of yabusame, a novel piggyBac-like transposable element in the genome of the silkworm, *Bombyx mori*." Genome **53**: 585-593.

Dewannieux, M., Esnault,C., Heidmann, T. (2003). "LINE-mediated retrotransposition of marked Alu Sequences." Nature Genetics **35**: 41-48.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucl. Acids Res. **32**(5): 1792-1797.

Eickbush, T., Furano, A. (2002). "Fruit flies and humans respond differently to retrotransposons." Genomes and Evolution **12**: 669-674.

Furano, A., Duvernell, D., Boissinot, S. (2004). "L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish." Trends in Genetics **20**(1): 9-14.

Garcia-Fernandez, J., J. R. Bayascas-Ramirez, et al. (1995). "High copy number of highly similar mariner-like transposons in planarian (Platyhelminthe): evidence for a trans-phyla horizontal transfer." Molecular Biology and Evolution **12**(3): 421-431.

Gilbert, C., S. Schaack, et al. (2010). "A role for host-parasite interactions in the horizontal transfer of transposons across phyla." Nature **464**(7293): 1347-1350.

Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symposium Series **41**: 95-98.

Ichimura, S., Mita, K., Sugaya, K. (1997). "A Major Non-LTR Retrotransposon of *Bombyx mori*, L1Bm." Journal of Molecular Evolution **45**(3): 253-264.

Jin-Shan, X., Qing-You,X., Jun, L., Guo-Qing, P., Ze-Yang, Z. (2005). "Survey of long terminal repeat retrotransposons of domesticated silkworm (Bombyx mori)." Insect Biochemistry and Molecular Biology **35**: 921-929.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and Genome Research **110**: 462-467.

Kapitonov, V., Jurka, J. (2001). "Rolling-circle transposons in eukaryotes." Proceedings of the National Academy of Sciences of the United States of America **98**(15): 8714-8719.

Kapitonov, V., Tempel, S., Jurka, J. (2009). "Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences." Gene **448**(2): 207-213.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences " Journal of Molecular Evolution **16**: 111-120.

Kohany, O., A. Gentles, et al. (2006). "Annotation, submission and screening of repetitive elements in Repbase: RebaseSubmitter and Censor." BMC Bioinformatics **25**(7).

Novick, P., Basta H., Floumanhaft, M., McClure, M., Boissinot, S. (2009). "The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis carolinensis* shows more similarity to fish than mammals." Molecular Biology and Evolution **26**(8): 1811-1822.

Novick, P., J. Smith, et al. (2010). "Independent and parallel lateral transfer of DNA transposons in tetrapod genomes." Gene **449**(1-2): 85-94.

Okada, N., Shedlock, A., Nikaido, M. (2004). Retroposon Mapping in Molecular Systematics. Mobile Genetic Elements: Protocols and Genomic Applications. W. Miller, Pierre, C. NJ, Humana Press. **260:** 189-226.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K., Fujiwara, H. (2008). "Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori." Insect Biochemistry and Molecular Biology **38**: 1046-1057.

Pagan, H., Smith, J., Hubley, R., Ray, D. (2010). "PiggyBac-ing on a primate genome: Novel elements, recent activity and horizontal transfer." Genome Biology and Evolution **2**: 293-303.

Price, A. L., E. Eskin, et al. (2004). "Whole-genome analysis of Alu repeat elements reveals complex evolutionary history." Genome Research **14**(11): 2245-2252.

Ray, D., Feschotte, C., Pagan, H., Smith, J., Pritham, E., Arensburger, P., Atkinson, P., Craig, N. (2008). "Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*." Genome Research **18**: 717-728.

Simonsen, T., Zakharov,E., Djernaes, M., Cotton, A., Vane-Wright, R., Sperling, F. (2010). "Phylogenetics and divergence times of Papilioninae (Lepidoptera) with special reference to the enigmatic genera Teinopalpus and Meandrusa." Cladistics **26**: 1-25.

Smit, A., Hubley, R. (2008-2010). "RepeatModeler Open-1.0." from http://www.repeatmasker.org/RMDownload.html.

Smit, A., Hubley, R., Green, P. (1996-2010). "RepeatMasker Open-3.0." from http://www.repeatmasker.org.

Smit, A. F. A. and R. Hubley (2008-2013). "COSEG 0.2.1." from http://www.repeatmasker.org.

Song, M., Boissinot, S. (2007). "Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination." Gene **390**: 206-213.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S. (2011). "MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods." Molecular Biology and Evolution **28**: 2731-2739.

Tu, Z. (1997). "Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, Aedes aegypti." Proceedings of the National Academy of Sciences of the United States of America **94**(14): 7475-7480.

Ustyugova, S. V., Amosova, A.L., Lebedev, Y.B., Sverdlov, E.D. (2005). "Cell line fingerprinting using retroelement insertion polymorphism." Biotechniques **38**: 561-565.

Wahlberg, N., J. Leneveu, et al. (2009). "Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary." Proc Biol Sci **276**(1677): 4295-4302.

Zhan, S., Merlin, C., Boore, J., Reppert, S. (2011). "The Monarch butterfly genome yields insights into long-distance migration." Cell **147**: 1-15.

CHAPTER III

IDENTIFICATION OF TRANSPOSABLE ELEMENTS IN MUSCID AND OESTROID FLIES

Abstract

Transposable elements (TEs) mobile DNA fragments found in genomes, are divided into two classes, class I and class II. Class I elements use a "copy and paste" mechanism while class II elements use a "cut and paste" mechanism. Outside Drosophila and other model species, very little has been done to characterize the TE landscapes of insects, especially when considering the huge diversity represented within Insecta. Lesser investigated clades include the superfamily Oestroidea and the family Muscidae. Members of both clades include a large number of agricultural pests and/or forensic indicator species. In this study, the genomes *Haematobia irritans* (horn fly), *Sarcophaga crassipalpis* (flesh fly), *Phormia regina* (black blow fly), and *Cochliomyia hominivorax* (the New World screw-worm fly) were investigated using 454 sequencing. The data were analyzed to determine the TE landscapes of these taxa and compared to well-characterized model insects and to one another. The TE proportions were 5.95%, 10.00%, 22.43%, and 30.67%, for *C. hominivorax, P. regina*, *S. crassipalpis* and *H. irritans*, respectively. DNA transposons were the predominant class in each taxon, except for *S. crassipalpis* whose predominant class of TEs was non-LTRs. These results indicate that TE content among these taxa and within insects in general is highly variable.

Introduction

Transposable elements (TEs) encompass two classes of DNA sequences that can move or copy themselves from one site to another in a genome. Class I TEs, or the retrotransposons, use a "copy and paste" mechanism to insert themselves into a new location. LTR (Long Terminal Repeat) retrotransposons are structurally similar to retroviruses, with direct repeats at each end and coding sequences such as GAG, POL and ENV. Non-LTR retrotransposons lack terminal repeats and instead harbor a poly-A tail or some other repetitive motif. Non-LTR retrotransposons can be divided into autonomous and non-autonomous elements. Autonomous non-LTR retrotransposons (often referred to collectively as LINEs) consist of one or two open reading frames (ORFs) that encode proteins involved in mobilization. Non-autonomous non-LTR retrotransposons include Short INterspersed Elements (SINEs) which require the enzymatic machinery of LINEs to mobilize (Dewannieux 2003).

DNA transposons belong to Class II and utilize a "cut and paste" mechanism to mobilize. Class II elements include the hAT, piggyBac, and TcMariner superfamilies, which require an enzyme known as transposase for their activity (Biemont 2006). Class II also encompasses the rolling-circle transposons such as Helitrons and Mavericks (Munoz-Lopez and Garcia-Perez 2010).

Regardless of class, TEs have substantial impacts on genome structure and function. As they mobilize, TEs can impact genomes through insertion mutagenesis and transduction (Ivics and Izsvak 2004). However, even after mobilization, they can influence genome structure by mediating chromosome rearrangements (Biemont 2006)

and non-homologous recombination (Shalev and Levy 1997). They can also serve as a source of novel coding sequences and regulatory sites (Rebollo, Romanish et al. 2012).

TEs are found in nearly all eukaryotic organisms, and TE landscapes can vary widely. For example, TEs in humans make up as much as 70% of the genome (Lander 2001; de Koning, Gu et al. 2011) while in *Fugu rubripes*, TEs comprise only ~2.7% (Aparicio, Chapman et al. 2002). This is true among even closely related taxa. In *Drosophila melanogaster*, 18% of the genome is made up of TEs compared to 5% in *Drosophila simulans* (Capy and Gibert 2004), species separated by less than 3 million years (Hedges, Dudley et al. 2006).

The TEs of some insects, *Drosophila, Bombyx,* and mosquitos have been examined from a transposable element perspective (Kaminker, Bergman et al. 2002; Boulesteix and Biemont 2005; Osanai-Futahashi 2008) but most recently sequenced insect genomes have yet to receive a detailed analysis of their TE content. For example, outside of *Calliphora vicina* (the bluebottle fly) which exhibits a diverse array of non-LTR retrotransposons and DNA transposons (Negre and Simpson 2013) very little work has been completed to characterize TE landscapes in oestroid or muscid flies, two very diverse groups. These two clades include large numbers of taxa. For example, Muscidae encompasses ~4000 species (Resh and Cardé 2003). The superfamily Oestroidea can be divided into two families, Calliphoridae, with ~1000 species, and Sarcophagidae, with ~2000 species (Rognes 1991). In addition to being a major repository of insect diversity, many of these flies are important economically, medically and forensically. Thus, knowledge of their genome structure could provide insight into their evolution and potential management strategies.

In this study, the genomes of four less commonly studied insects from these two clades, *Cochliomyia hominivorax* (the New World screw-worm fly), *Phormia regina* (the black blow fly), *Sarcophaga crassipalpis* (the flesh fly), and *Haematobia irritans* (the horn fly), were surveyed using 454 sequencing. The former three are calliphorid flies while the last is a muscid. Each represents a taxon with some importance agriculturally and/or forensically. For example, *C. hominivorax, P. regina,* and *S. crassipalpis* can enter wounds of animals and feed on the healthy, underlying flesh, which if left untreated can lead to death (Capinera 2008). *H. irritans* primarily feed on the blood of cattle and can cause a decrease in a cow"s weight which can lead to a reduction in milk production (Floate 2002). *S. crassipalpis* is a model for endocrinology (Verleyen, Huybrechts et al. 2004), cold resistance (Lee, Chen et al. 1987) and diapause (Denlinger 2002). Finally, these taxa can also be considered forensic indicator species in that they are known to colonize human remains and can be used to infer the postmortem interval (PMI) (Wells 2001).

Methods

Samples and 454 Sequencing

The genomes were surveyed using methods similar to those described previously (Pagan, Smith et al. 2010). Briefly, adult specimens of *P. regina* and *S. crassipalpis* were collected from colonies maintained at West Virginia University in the laboratory of JDW. Founders for the *P. regina* and *S. crassipalpis* colonies were obtained from the wild in Pullman, WA and Morgantown, WV, respectively, in 2008. DNA was isolated as described in Singh et al. (2011).

Adult specimens of *H. irritans* were collected in 2003 from the Pressler Cattle Ranch in Kerrville, TX and a single individual fly was used for this study. *C. hominivorax* adults were obtained from a production facility in Panama. DNA was isolated by grinding the head of individual samples in liquid nitrogen followed by standard phenol-chloroform purification and ethanol precipitation.

All four samples were sequenced commercially at Georgia Genomics Facility on a single microtiter plate using Roche"s standard FLX chemistry. Sample preparation, including indexing of each DNA library and post-sequencing deconvolution of the data followed Roche protocols (October 2008). All raw data were parsed locally using 454 Replicate Filter (http://microbiomes.msu.edu/replicates/) to remove emulsion PCR artifacts as described in Pagan et al (2012) .

Repeat Discovery

In order to determine the repeat content of each genome, we followed the protocols of Macas (2002) and Novak (2010) which examines reads and clusters them into groups exhibiting sequence similarity. These pipelines assemble clustered reads into contigs. However, visual inspection suggested that some contigs may have been misassembled. We therefore, reassembled clusters of reads using SeqMan, a part of the DNAStar package, using the following parameters: match size = 50, minimum match = 80, and minimum sequence length = 30. The number of clusters obtained and analyzed from 454 sequencing differed in each taxon (Table 3.1). Clusters represented by large numbers of sequences are likely to be repetitive and could be TEs. To isolate these sequences, clusters with greater than 100 reads were assembled into contigs in all taxa

except C. hominivorax. Because C. hominivorax reads were underrepresented, clusters with 50 reads or greater were assembled.

For each taxon, consensus sequences were generated from the reassembled contigs. Tandem Repeat Finder was used to filter contigs consisting of satellite sequences. The remaining contigs were passed through Censor, blastn, and blastx to identify known transposable elements. The result was a custom library of TEs which consisted of these contigs and the existing repertoire of TEs from insects as found in RepBase (Jurka 2005). We then filtered the survey sequence data to remove any reads under 100 nt in length and the library was used in conjunction with a locally implemented version of RepeatMasker (A.F.A. Smit and R. Hubley; see http://www.repeatmasker.org/) to estimate the TE content of the survey sequence data and, by extension, each taxon.

Identification of SINE subfamilies:

Two novel SINE elements were identified in the data, one each in *S. crassipalpis* and *H. irritans*. The *S. crassipalpis* SINE has been dubbed *Wingman*. We used COSEG (Smit 1996-2010) to predict subfamilies for *Wingman*. COSEG examines multiple instances of TE insertions and identifies co-segregating (2-3 bp) sites in an effort to determine subfamily structure. Approximately 48,000 instances of the SINE element were identified in the *S. crassipalpis* sequence data using RepeatMasker. The consensus is ~324 bp in length. We identified 5,449 full- or near full-length insertions spanning 292-356 bases (+/- 10% of the general consensus) and passed them to COSEG for subfamily identification. A custom perl script provided by R. Hubley was used to refine the consensus sequence for each subfamily. A custom RepeatMasker library consisting of the suggested *Wingman* subfamilies consensus sequences was created and the top 150

hits extracted from the genome. The extracted sequences were aligned with their respective subfamily consensus sequence to confirm the presence of each subfamily in the survey data.

We also used COSEG to predict possible subfamilies for the novel SINE element in *H. irritans*, which we named *Bloodsucker*. The consensus is ~349 bp in length. We identified 23 intact insertions spanning 314-384 bases (+/-10% of the general consensus). Only 214 instances of the SINE element were identified in *H. irritans* and no subfamily structure was indicated by COSEG.

Age analyses and activity periods:

We estimated activity periods by utilizing genetic distances to measure the difference between individual insertions and the consensus of each element (Ray 2008; Pagan 2010). Older elements are expected to have accumulated independent mutations, resulting in a higher genetic distance values, than recently active elements, which would have less time for mutations to accumulate. Given that we were analyzing survey sequence data with an average read length of ~360 bp, we were limited to comparing only full-length elements with consensus sequences below that size and fragments of any longer elements. We therefore created a modified TE library which included Wingman and Bloodsucker, DNA transposons, and autonomous non-LTR families (LINEs). We divided DNA transposons into two categories: 1) short, full-length DNA transposons (<360 bp), and 2) fragments chosen from coding regions of longer, full length elements. For LINE elements, the last 300 bp of the 3' end of the LINE elements was utilized. The library was used to query the 454 survey data with RepeatMasker.

We estimated distances between the insertions and their respective consensus by using a modified version of the calcDivergencefromAlign script that is part of the RepeatMasker package to calculate the Kimura-2 parameter (Kimura 1980) distance value (including CpG sites) (Pagan 2010). No neutral mutation rate is available for any of the sampled taxa. Instead a neutral substitution rate for *Drosophila*, 0.016 substitutions per site/Myr was used(Beeman 1996).

Horizontal Transfer

To investigate the taxonomic distribution of the elements we identified from each genome, we queried the full WGS database at NCBI. If the results indicated that the consensus sequence shared at least 95% sequence identity over at least 80% of its length then the element was considered a candidate for horizontal transfer. Any hits matching the criteria above were extracted, aligned to the query sequence, and examined by eye.

Results

Summary of 454 Sequencing

In all ~300 million bp of useable data was generated for the four taxa. Genome coverage was calculated by dividing the total base pairs by the estimated genome size (Picard, Johnston et al. 2012). Since the genome size for *S. crassipalpis* is unknown, *Sarcophaga bullata*, with the assumption of similar genome size, was used. (www.genomesize.com) The percentage of genome coverage for *C. hominivorax*, *S. crassipalpis*, *P. regina*, and *H. irritans* was 9.7%, 11.4%, 20.8%, and 5.1%, respectively. The read lengths ranged from 29 bp to 1174 bp, and the average read length was ~360 bp. (Table 3.1).

Repeat Discovery

Representatives from all major TE orders are present in most taxa (Table 2). Total TE composition in each genome ranged from 5.95% to 30.67%. The highest non-LTR percentage of the genome (13.82%) was found in *S. crassipalpis*. The estimated LTR content in each taxon was relatively low with proportions ranging from 0.39% in *C. hominivorax* to 2.78% in *S. crassipalpis*. DNA transposons were most prevalent in *H. irritans* (9.75%).

Both *H. irritans* and *S. crassipalpis* harbor elements exhibiting the hallmarks of SINEs. For example, the 5' region of both *Wingman* and *Bloodsucker* contain the RNA polymerase III promoters boxes, A and B, which are separated by approximately 30 bp (Miller and Capy 2004). Both presumptive SINEs can be folded into secondary structures that would indicate that they are tRNA derived (Figure 3.1).

Age Analyses

Three taxa exhibited very little recent activity. In *C. hominivorax,* there appears to be an overall paucity of transposable element activity (Table 3.4). Only one DNA transposon family, DNA/zator, was identified in the genome, and it is a relatively old element (mean activity periods are greater than 2 my). The only autonomous non-LTR element to be identified, an R1, does not exhibit any recent activity. In *P. regina*, the only TE to exhibit recent activity (~1.38 mya) was a Mariner element. *S. crassipalpis* does not exhibit any recent activity of DNA transposons or autonomous non-LTR retrotransposons. There is, however, some evidence for relatively recent SINE activity in *S. crassipalpis* via *Wingman*. We were unable to identify an active autonomous LINE

partner of *Wingman* due to the short read lengths obtained from the 454 survey sequencing.

While the majority of *Wingman* activity occurred in the relatively distant past, analyses of genetic distances suggest *Wingman1* and *Wingman2* activity as recent as 0.04 and 0.48 mya, respectively (Table 3.3). For the most part the topology of the COSEG tree and the estimated activity periods are similar to one another (Figure 3.2). For example, according to the genetic distances *Wingman1* and *Wingman2* are the most recently active elements, and they are located at the termini on the COSEG tree which is where the most recently active elements should be found. *Wingman7* is the oldest subfamily according to the genetic distances, and COSEG has positioned *Wingman7* at the base of the tree which indicates that it is the oldest.

In H. irritans, three mariner elements were relatively young which suggests that they may be recently active while another three mariner elements were older (Table 3.4). The autonomous non-LTR retrotransposons do not appear to exhibit any recent activity in the genome. With regard to the non-autonomous non-LTR retrotransposons, we estimated the activity period of the H. irritans SINE element, Bloodsucker (7,732 copies), and found a mean activity period of 4.76 mya. This suggests a lack of activity in the recent past (Table 3.4).

Horizontal Transfer

We identified two potential horizontal transfer events. First, a non-autonomous Mariner element from *S. crassipalpis* (Mariner5_SC) is found as a single copy in *Bombyx mori* (accession number BAAB01003695.1) with 95% identity over its entire length (query coverage = 94%, E value = 0.0). The total length of the query was 482 bp.

The second candidate for horizontal transfer is a mariner element (Mariner4_HI) between *H. irritans* and *Anopheles gambiae* (accession number ABKQ02017564.1) with an identity of 97% over 1300 bp (query coverage 100%, E value 0.0).

Discussion

Comparison of TE content

Our data compared with other studies suggests that the genome proportion attributable to TEs varies greatly among dipteran genomes. Approximately 16% of the genome of *Anopheles gambiae* is attributable to TEs (Holt, Subramanian et al. 2002), 28% in *Culex quinquefaciatus* (Arensburger, Megy et al. 2010), 47% in *Aedes aegypti* (Nene, Wortman et al. 2007), 3% in *Drosophila grimshawii* (Clark, Eisen et al. 2007), 14% in *Drosphila virilis* (Clark, Eisen et al. 2007), and 25% in *Drosphila ananassae* (Clark, Eisen et al. 2007). Very few studies have analyzed the TE landscapes of non-model insect organisms. However, recently a portion of the genome of a blowfly, *Calliphora vicina*, was analyzed from a TE perspective. In the 600 kb region analyzed, TEs make up 24% (Negre and Simpson 2013). The four taxa that we surveyed offer additional insights into the diversity of TE landscapes of dipteran genomes that are not model organisms.

Among the four taxa that we analyzed, *H. irritans* exhibited the greatest proportion of TEs (30.67%) and *C. hominivorax* (5.95%) the smallest (Table 3.2). In the 600 kb region analyzed in *C. vicina*, DNA transposons comprised the largest fraction of TEs (12.87%) and the most common DNA transposons were Mariner and Helitron elements (Negre and Simpson 2013). When the fraction of DNA transposons of *C. vicina* is compared to the fraction of DNA transposons in the three oestroid flies (*C.*

hominivorax, P. regina, and *S. crassipalpis*), there are dramatic differences. For example, only 2.48% of the sequenced region of *C. hominivorax* arises from DNA transposons, DNA transposons make up ~5% of the survey sequences of *P. regina*, and *S. crassipalpis* harbors only slightly higher DNA transposon content (5.83%). On the other hand, the fraction of DNA transposons in the survey sequences of the muscid fly, *H. irritans*, is more similar to *C. vicina*, ~9.75% DNA transposons.

Of the non-LTR elements, only 2.95% make up the sequenced region of *C. vicina* (Negre and Simpson 2013). This is similar to the low fraction of non-LTRs found in the survey sequences of *C. hominivorax* (0.62%) and *P. regina* (1.08%). However the third oestroid fly, *S. crassipalpis*, is dramatically different with regard to non-LTR content. 13.8% of our survey sequences were identifiable as non-LTR elements with SINEs making up the greatest proportion at 8.86% of the survey sequences. The fraction of non-LTR content from the survey sequences of *H. irritans* (7.39%). Interestingly, a general pattern has emerged among several insect species that the activity periods of autonomous non-LTR elements have occurred in the distant past similar to what is observed in *C. hominivorax, P. regina, S. crassipalpis*, and *H. irritans*. For example, the LINE elements in *C. vicina* were found to be older (Negre and Simpson 2013). A similar pattern was observed in *Heliconius melpomene* in which most autonomous non-LTR elements exhibited a lack of recent activity. This was explained by increased rates of ectopic recombination acting to remove the elements and the same mechanism may be at play here (Lavoie 2013).

With regard to LTR elements, 3.54% make up the 600 kb region sequenced in *C. vicina* (Negre and Simpson 2013). This is similar to what was found in the *S.*

crassipalpis (2.78%) and *H. irritans* (2.12%) sequence data, but differs from the rather low estimates from *C. hominivorax* (0.39%) and *P. regina* (0.62%). However, we must point out that *C. hominvorax* and *P. regina* exhibit relatively large fractions of unidentified, potentially TE-derived content (2.46% and 2.99%, respectively). Some of these unknown elements may be as yet LTR retrotransposons which could change the fraction of LTRs found in *C. hominivorax* and *P. regina* as future analyes proceed.

Diversity of TEs

Our study demonstrates how the accumulation of TEs in insects can be dramatically different among insect taxa. For example, 18% of the genome of *D. melanogaster* is made up of TEs (Biemont and Cizeron 1999), the genome of *B. mori* is composed of 35% TEs (Osanai-Futahashi 2008), and the genome of *Heliconius melpomene* consists of 24.94% of TEs (Lavoie 2013). With regard to the TEs identified in *B. mori* and *H. melpomene*, the types of TEs that make up these proportions in each of the genomes are different. For example, in *B. mori* 93% of the non-LTR elements are SINEs compared to 68% in *H. melpomene*. Also, DNA transposons make up only 3% of the *B. mori* genome compared to 10% of the *H. melpomene* genome (Lavoie 2013) and nearly 13% of the *C. vicina* genome. However, our analysis of *H. irritans* (~31% of our survey sequences) suggests that DNA transposon levels can rise substantially higher.

The low estimation of TEs in *C. hominivorax* (5.95%) may be a representation of the actual percentage of TEs or it may indicate that because of the low coverage that the TE estimate may have been affected which resulted in a TE library that was not comprehensive. However, when comparing our data to coverage of the *H. irritans* genome and the others, we note that genome coverage estimates are all comparable

(9.7%, 11.4%, 20.8%, 5.1% in *C. hominivorax, S. crassipalpis, P. regina*, and *H. irritans*, respectively) Given the small genome size of *C. hominivorax* compared to *H. irritans*, it appears entirely likely that a paucity of TEs may be the case. Indeed, if genome coverage is used to estimate our ability to recover TEs using survey sequence data, we have no reason to suspect that our efforts with *C. hominivorax* were biased in such a way as to disallow TE discovery.

Two elements were identified as candidates for horizontal transfer. The first candidate is a mariner element that was found in both *S. crassipalpis* and *B. mori*. The second candidate is another mariner element that was found in both *H. irritans* and *A. gambiae.* Considering that Calliphoridae and *B. mori* diverged approximately 360.9 my, Muscidae and *A. gambiae* diverged 274.9 mya (Hedges, Dudley et al. 2006), and that DNA transposons are more likely to be involved in horizontal transfer events (Loreto, Carareto et al. 2008), it appears that these elements have undergone horizontal transfer.

Future Studies

The investigation of the TE landscape of these four fly taxa may have several impacts such as utilizing TEs as genetic vectors in controlling pests and contributing to the phylogeny of Diptera. TEs, specifically the DNA transposons, can be used as genetic vectors in order to genetically modify insect pests. TEs have been used in previous studies to modify pests. For example, modification using a piggyBac element has been investigated in *Lucilia cuprina* (Heinrich 2002). Thus, the identification of additional DNA transposons in these fly taxa may allow for the development of additional tools derived from transposable elements.

The retrotransposons identified could make a different contribution. The retrotransposons, specifically SINE elements, have been shown to be particularly useful at resolving phylogenetic uncertainties (Shedlock, Takahashi et al. 2004; Ray, Xing et al. 2006). Retrotransposons might also be utilized in a forensic context to identify a particular fly species found on a corpse in order to determine the post mortem interval (PMI).

In conclusion, we have analyzed the TE landscape of four non-model fly taxa, and we have estimated the TE content for each taxa. We also identified two novel SINE elements unique to *S. crassipalpis* and *H. irritans* which could be utilzed in further phylogenetic studies. We have also identified mariner elements in each family which could potentially be utilized as tools for genetic manipulation. Further studies of other dipteran genomes will be beneficial in understanding the evolution of TEs in the order Diptera.

Table 3.1 Summary of 454 Sequencing

Taxon	Reads	Mean Read Length	Total Bases	Estimated Genome Size (Mb)	% of Genome Coverage	Clusters (> 100 reads)	# Reads in Largest Cluster
C. hominivorax	112037	381	42,631,023	441.5	9.7%	14	676
S. crassipalpis	237214	357	84,614,591	743.3	11.4%	33	16654
P. regina	313869	350	109,851,114	529.3	20.8%	41	6926
H. irritans	168190	364	61,196,231	1,197.4	5.1%	49	8833
Totals	**831310**		**298,292,959**				

Table 3.2 Summary of TE content in each taxon

Class	Family	*C. hominivorax*	*P. regina*	*S. crassipalpis*	*H. irritans*
DNA	Zator	0.10%	0.22%		
	Helitron	1.54%	2.56%	0.01%	2.22%
	Mariner	0.79%	2.14%	5.67%	7.32%
	hat	0.03%			
	Other	0.03%	0.11%	0.15%	0.21%
Total DNA transposons		**2.48%**	**5.04%**	**5.83%**	**9.75%**
Non-LTR	I	0.02%	0.26%	0.40%	
	R1	0.30%		1.05%	
	Vingi	0.13%			0.63%
	Jockey	0.11%	0.30%	0.17%	
	Loa		0.14%	0.16%	4.22%
	CR1		0.07%	0.40%	0.14%
	SINE			8.86%	1.99%
	L2			0.14%	
	RTE			1.69%	
	BILBO			0.83%	
	Other	0.07%	0.31%	0.10%	0.41%
Total Non-LTR		**0.62%**	**1.08%**	**13.82%**	**7.39%**
LTR	Gypsy	0.16%	0.62%	2.33%	1.35%
	Copia	0.03%	0.08%	0.07%	0.16%
	Pao	0.18%	0.13%	0.33%	0.54%
	Other	0.01%	0.05%	0.05%	0.08%
Total LTR		**0.39%**	**0.89%**	**2.78%**	**2.12%**
Unknown	Unknown	2.46%	2.99%	0.85%	11.39%
Total TEs		**5.95%**	**10.00%**	**23.28%**	**30.67%**

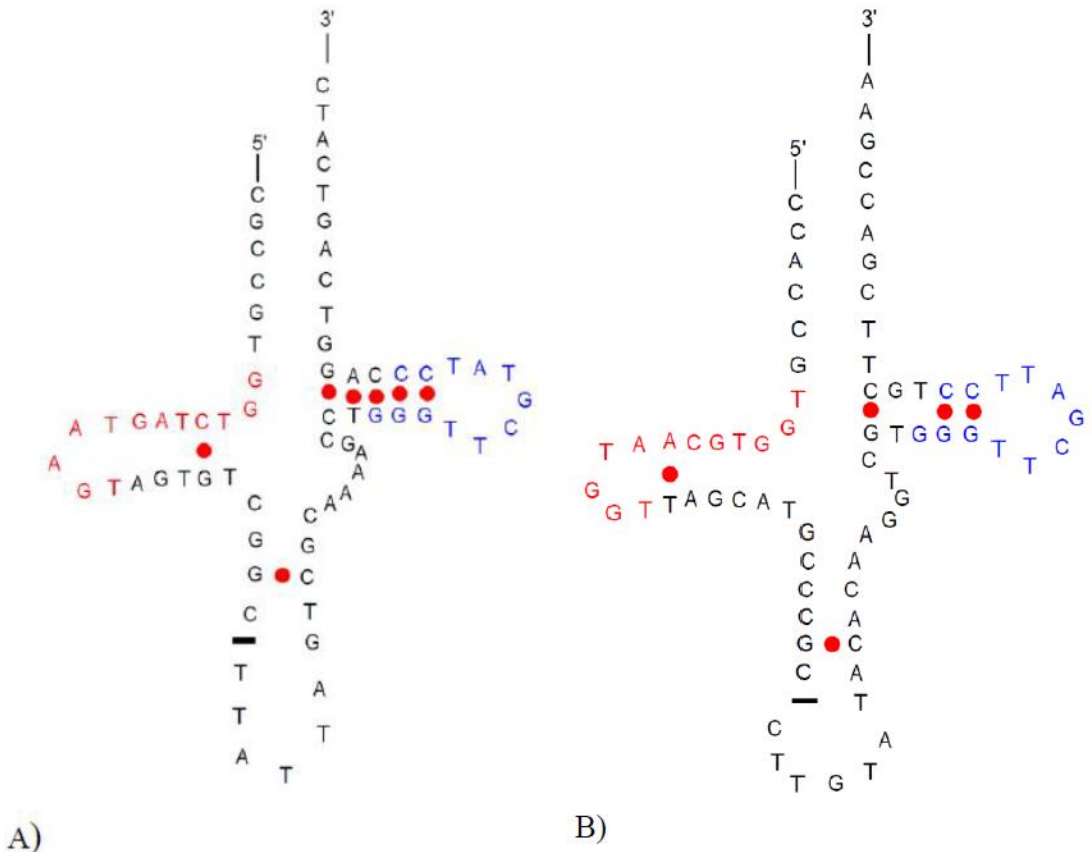

Figure 3.1 The predicted tRNA secondary structure of (a) *S. crassipalpis* and (b) *H. irritans* SINE elements.

The colored nucleotides identify the putative A (red) and B (blue) promoter regions

Table 3.3 Estimated activity periods for *Wingman* subfamilies

Subfamily	Mean Distance	Standard Deviation	Range	Time (mya)	Copy Number
Wingman0	0.093499	0.064202	0.029298-0.157701	1.8-9.8	12,268
Wingman1	0.067675	0.067011	0.000664-0.134686	0.04-8.4	10,581
Wingman2	0.082686	0.074977	0.007708-0.157663	0.48-9.8	3,731
Wingman3	0.168888	0.076649	0.092239-0.245536	5.7-15.3	692
Wingman4	0.117533	0.084565	0.032968-0.202098	2.1-12.6	1,506
Wingman5	0.128902	0.064681	0.064221-0.193582	4.0-12.1	5,572
Wingman6	0.093045	0.057969512	0.035076-0.151015	2.1-9.4	5,194
Wingman7	0.182275	0.064109	0.118166-0.246384	7.3-15.3	13,349

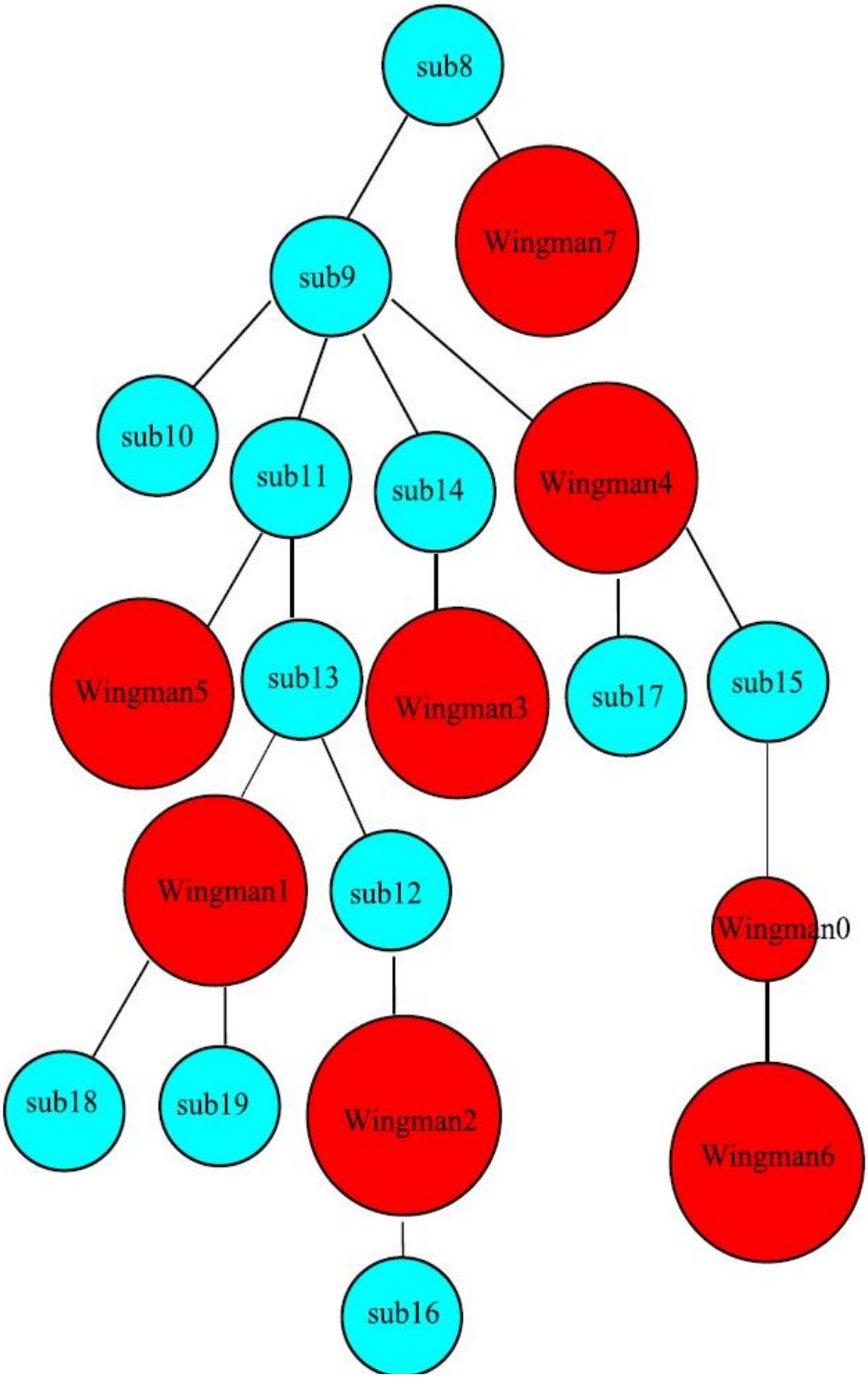

Figure 3.2 Results of the COSEG analysis.

Red circles are proposed subfamilies.

Table 3.4 Ages of non-LTR elements and DNA transposons elements in all four taxa surveyed.

Element Name:	Cluster Number	Mean Distance	Standard Deviation	Range	Mean activity period (mya)	Time Range (mya)
Mariner1_HI	CL08	0.012504	0.046699	0-0.059203	0.78	0-3.70
Mariner2_HI	CL06	0.018148	0.013723	0.004425-0.031871	1.13	0.28-1.99
Mariner1_PR	CL04	0.022098	0.028718	0-0.050817	1.38	0-3.18
Mariner3_HI	CL23	0.030824	0.042545	0 - 0.073369	1.93	0 - 4.59
Mariner2_PR	CL13	0.040599	0.076176	0-0.116775	2.54	0-7.30
R1_1_SC	CL22	0.04162	0.05886	0 - 0.10048	2.6	0 - 6.28
R1_1_CH	CL13	0.04347	0.06677	0-0.11024	2.72	0 - 6.89
Mariner4_HI	CL05	0.049421	0.081113	0-0.130535	3.09	0-8.16
Mariner1_SC	CL17	0.066518	0.027232	0.039286-0.093749	4.16	2.46-5.86
Loa1_HI	CL03	0.069561	0.047254	0.022307 - 0.116814	4.35	1.39 - 7.30
Mariner2_SC	CL08	0.0705	0.03192	0.03858-0.10242	4.41	2.41-6.40
CR1_1_SC	CL13	0.071515	0.084605	0 - 0.15612	4.47	0 - 9.76
Bloodsucker_HI	CL09	0.0763	0.077404	0-0.153704	4.76	0-9.6
Mariner3_SC	CL24	0.076058	0.036981	0.039077-0.11304	4.75	2.44-7.06
RTE1_SC	CL05	0.076123	0.057122	0.019002 - 0.133245	4.76	1.19 - 8.33
Mariner3_PR	CL12	0.076569	0.081454	0-0.158023	4.79	0-9.88
R1_2_SC	CL27	0.104593	0.061506	0.043087 - 0.166099	6.54	2.69 - 10.38
Vingi1_HI	CL19	0.105997	0.105234	0.000763 - 0.211232	6.62	0.048 - 13.20
I_1_SC	CL21	0.107192	0.181057	0 - 0.288249	6.7	0 - 18.02
Mariner5_HI	CL24	0.111807	0.177843	0-0.28965	6.99	0-18.10
CR1_2_SC	CL23	0.113193	0.11337	0 - 0.226563	7.07	0 - 14.16
Zator1_CH	CL10	0.12541	0.05997	0.065439-0.185388	7.84	4.09 -11.59
Mariner4_SC	CL18	0.12858	0.152578	0-0.281158	8.04	0-17.57
I_1_PR	CL35	0.128753	0.134034	0 - 0.262787	8.05	0 - 16.42
Mariner4_PR	CL05	0.130208	0.096043	0.034165-0.226251	8.14	2.14-14.14
Mariner5_SC	CL12	0.131178	0.113934	0.017244-0.245111	8.2	1.08-15.31
BILBO1_SC	CL20	0.137507	0.12429	0.013217 - 0.261797	8.59	0.83 - 16.36
R1_1_PR	CL23	0.137927	0.127141	0.010787 - 0.265068	8.62	0.67 - 16.57
R1_3_SC	CL19	0.194387	0.016274	0.178112 - 0.210661	12.15	11.13 -13.17
I_2_SC	CL16	0.198908	0.147222	0.051686 - 0.346131	12.43	3.23 - 21.63
BILBO2_SC	CL28	0.202934	0.119072	0.083861 - 0.322006	12.68	5.24 - 20.13
Mariner6_HI	CL12	0.228269	0.101312	0.126957-0.32958	14.27	7.93-20.60

Literature Cited

Aparicio, S., J. Chapman, et al. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes." Science **297**(5585): 1301-1310.

Arensburger, P., K. Megy, et al. (2010). "Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics." Science **330**(6000): 86-88.

Beeman, R. W. e. a. (1996). "Woot, an active gypsy-class retrotransposon in the flour beetle, Tribolium castaneum, is associated with a recent mutation." Genetics **143**(1): 417-426.

Biemont, C. and G. Cizeron (1999). "Distribution of transposable elements in Drosophila species." Genetica **105**(1): 43-62.

Biemont, C., Vieira, C. (2006). "Junk DNA as an evolutionary force." Nature **443**(5): 521-524.

Boulesteix, M. and C. Biemont (2005). "Transposable elements in mosquitoes." Cytogenet Genome Res **110**(1-4): 500-509.

Capinera, J. (2008). Encyclopedia of Entomology. Heidelberg, Springer.

Capy, P. and P. Gibert (2004). "Drosophila melanogaster, Drosophila simulans: so similar yet so different." Genetica **120**(1-3): 5-16.

Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-218.

de Koning, A. P., W. Gu, et al. (2011). "Repetitive elements may comprise over two-thirds of the human genome." PLoS Genet **7**(12): e1002384.

Dewannieux, M., Esnault,C., Heidmann, T. (2003). "LINE-mediated retrotransposition of marked Alu Sequences." Nature Genetics **35**: 41-48.

Floate, L. (2002). Biological control programmes in Canada. New York, NY, CABI Publishing.

Hedges, S. B., J. Dudley, et al. (2006). "TimeTree: a public knowledge-base of divergence times among organisms." Bioinformatics **22**(23): 2971-2972.

Heinrich, J. C., et al. (2002). "Germ-line transformation of the Australian sheep blowfly Lucilia cuprina." Insect Molecular Biology **11**(1): 1-10.

Holt, R. A., G. M. Subramanian, et al. (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." Science **298**(5591): 129-149.

Ivics, Z. and Z. Izsvak (2004). "Transposable elements for transgenesis and insertional mutagenesis in vertebrates: a contemporary review of experimental strategies." Methods in Molecular Biology **260**: 255-276.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and Genome Research **110**: 462-467.

Kaminker, J. S., C. M. Bergman, et al. (2002). "The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective." Genome Biol **3**(12): RESEARCH0084.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences " Journal of Molecular Evolution **16**: 111-120.

Lander, E. S., Linton, L.M., Birren, B. Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K. Doyle, M. FitzHugh, W., et al. (2001). "Initial Sequencing and analysis of the human genome." Nature **409**: 860-921.

Lavoie, C., Platt, R.N., Novick, P.A., Counterman, B.A., Ray. D.A. (2013). "Transposable element evolution in *Heliconius* suggests genome diversity within Lepidoptera." Mobile DNA.

Lee, R. E., Jr., C. P. Chen, et al. (1987). "A rapid cold-hardening process in insects." Science **238**(4832): 1415-1417.

Loreto, E. L., C. M. Carareto, et al. (2008). "Revisiting horizontal transfer of transposable elements in Drosophila." Heredity (Edinb) **100**(6): 545-554.

Miller, W. J. and P. Capy (2004). Mobile genetic elements : protocols and genomic applications. Totowa, N.J., Humana Press.

Munoz-Lopez, M. and J. L. Garcia-Perez (2010). "DNA transposons: nature and applications in genomics." Curr Genomics **11**(2): 115-128.

Negre, B. and P. Simpson (2013). "Diversity of transposable elements and repeats in a 600 kb region of the fly Calliphora vicina." Mob DNA **4**(1): 13.

Nene, V., J. R. Wortman, et al. (2007). "Genome sequence of Aedes aegypti, a major arbovirus vector." Science **316**(5832): 1718-1723.

Novak, P., P. Neumann, and J. Macas (2010). "Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data" BMC Bioinformatics **11**: 378.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K., Fujiwara, H. (2008). "Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori." Insect Biochemistry and Molecular Biology **38**: 1046-1057.

Pagan, H., Smith, J., Hubley, R., Ray, D. (2010). "PiggyBac-ing on a primate genome: Novel elements, recent activity and horizontal transfer." Genome Biology and Evolution **2**: 293-303.

Pagan, H. J., J. Macas, et al. (2012). "Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats." Genome Biol Evol **4**(4): 575-585.

Pagan, H. J., J. D. Smith, et al. (2010). "PiggyBac-ing on a primate genome: novel elements, recent activity and horizontal transfer." Genome Biol Evol **2**: 293-303.

Picard, C. J., J. S. Johnston, et al. (2012). "Genome sizes of forensically relevant Diptera." Journal of Medical Entomology **49**(1): 192-197.

Ray, D., Feschotte, C., Pagan, H., Smith, J., Pritham, E., Arensburger, P., Atkinson, P., Craig, N. (2008). "Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*." Genome Research **18**: 717-728.

Ray, D. A., J. Xing, et al. (2006). "SINEs of a nearly perfect character." Syst Biol **55**(6): 928-935.

Rebollo, R., M. T. Romanish, et al. (2012). "Transposable elements: an abundant and natural source of regulatory sequences for host genes." Annual Review of Genetics **46**: 21-42.

Resh, V. H. and R. T. Cardé (2003). Encyclopedia of insects. Amsterdam ; Boston, Academic Press.

Rognes, K. (1991). Blowflies (Diptera, Calliphoridae) of Fennoscandia and Denmark. Leiden ; New York, E.J. Brill/Scandinavian Science Press.

Shalev, G. and A. A. Levy (1997). "The maize transposable element Ac induces recombination between the donor site and an homologous ectopic sequence." Genetics **146**(3): 1143-1151.

Shedlock, A. M., K. Takahashi, et al. (2004). "SINEs of speciation: tracking lineages with retroposons." Trends Ecol Evol **19**(10): 545-553.

Singh, B., H. Kurahashi, et al. (2011). "Molecular phylogeny of the blowfly genus Chrysomya." Med Vet Entomol 25(2): 126-134.

Smit, A., Hubley, R., Green, P. (1996-2010). "RepeatMasker Open-3.0." from http://www.repeatmasker.org.

Verleyen, P., J. Huybrechts, et al. (2004). "Neuropeptidomics of the grey flesh fly, Neobellieria bullata." Biochemical and Biophysical Research Communications 316(3): 763-770.

Wells, J. D. a. L. R. L. (2001). Forensic Entomology: Utility of Arthropods in Legal Investigations, CRC Press.

CHAPTER IV

CONCLUSIONS

For this thesis, I utilized 454 pyrosequencing technology and whole genome analyses to characterize the TE landscapes of non-model organisms in insects. The majority of comprehensive TE studies performed to date in insects have utilized model organisms such as *Bombyx mori* (Osanai-Futahashi 2008) and *Drosophila melanogaster* (Kaminker, Bergman et al. 2002) and mosquitoes (Boulesteix and Biemont 2005). The TE content among the insect species selected for this study varies by the class of TEs and fractional representation. The data presented in these two studies suggest TE diversity within Insecta is extensive.

In chapter 2, I conducted the first comprehensive analysis of TEs in a butterfly using the whole genome draft of *H. melpomene*. The study showed that the genome has accumulated a diverse array of DNA transposons and retrotransposons. While the DNA transposons were recently active, LINE elements exhibited a short activity period before being purged from the genome. This suggests that the genome might have defense mechanisms that influence the diversity of TEs. Studies have shown that organisms with a high TE diversity have more compact genomes (Volff, Bouneau et al. 2003; Furano 2004), and it is theorized that organisms with more compact genomes usually have increased rates of ectopic recombination. The work presented in this study provides the basis for additional studies. For example, further investigation of the active DNA

transposons, specifically the Tc3 element, may help to better understand the genome structure and evolution of *H. melpomene*. Investigating active TEs is important due to the ability of the active TE to produce genetic diversity in populations (Bennett, Coleman et al. 2004) and serve as raw material for evolutionary change (McDonald 2000).

We also characterized the TE landscapes of three oestroid flies and a muscid fly using 454 pyrosequencing. In this case, I did not have the advantage of whole genome sequences. Instead, the study relied on the assumption that survey sequencing (between 5.1% and 20.8% coverage of the genomes) would provide basic information on the TE complements of each genome. Any analyses suggested that the TE landscapes varied greatly. *C. hominivorax* exhibited the lowest total of TEs at only 5.95%. The low estimation could be a result of the genome acting to keep itself compact, or it could be a result of low coverage; *C. hominivorax* obtained the lowest amount of coverage compared to the other three genomes. Thus, the low estimate of TEs may be a result of not being able to acquire a comprehensive TE library from a limited data set. However, arguments are made that this is not the case.

My efforts to identify TEs in the other genomes were more successful. For example, two unique SINE elements were identified in *S. crassipalpis* and *H. irritans*. Both SINE elements appear to be tRNA derived. SINE subfamilies were identified in *S. crassipalpis* while no subfamilies were found in *H. irritan.* In each taxon, representatives of most major DNA transposon and autonomous non-LTR retrotransposon families were identified.

The TE content obtained from the four fly taxa may be utilized in future studies. For example, since all four taxa are considered agricultural pests, the DNA transposons

identified in each taxon may be investigated further to determine which transposon(s) may be utilized as genetic vectors to help modify its respective taxon. The retrotransposons, specifically the SINE elements, may be utilized for a different purpose. While SINE elements have been characterized in different types of insects, they have not been extensively investigated. The SINE elements that were identified in S. crassipalpis and H. irritans could be utilized in phylogenetic analyses in order to gain more information on the evolutionary relationships of these taxa (Nei and Kumar 2000; Shedlock, Takahashi et al. 2004).

The data presented in this thesis has utilized non-model organisms to further enhance our understanding of TE diversity in insects. This has laid the foundation for future studies of non-model insect genomes. Insects make up most of the species that inhabit Earth (Hoy 2013) and studying their evolutionary history allows for a greater understanding of the evolution of life.

Literature Cited

Bennett, E. A., L. E. Coleman, et al. (2004). "Natural genetic variation caused by transposable elements in humans." Genetics **168**(2): 933-951.

Boulesteix, M. and C. Biemont (2005). "Transposable elements in mosquitoes." Cytogenet Genome Res **110**(1-4): 500-509.

Furano, A., Duvernell, D., Boissinot, S. (2004). "L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish." Trends in Genetics **20**(1): 9-14.

Hoy, M. A. (2013). Insect molecular genetics : an introduction to principles and applications.

Kaminker, J. S., C. M. Bergman, et al. (2002). "The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective." Genome Biol **3**(12): RESEARCH0084.

McDonald, J. F. (2000). Transposable elements and genome evolution. Dordrecht ; London, Kluwer Academic.

Nei, M. and S. Kumar (2000). Molecular evolution and phylogenetics. Oxford ; New York, Oxford University Press.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K., Fujiwara, H. (2008). "Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori." Insect Biochemistry and Molecular Biology **38**: 1046-1057.

Shedlock, A. M., K. Takahashi, et al. (2004). "SINEs of speciation: tracking lineages with retroposons." Trends Ecol Evol **19**(10): 545-553.

Volff, J. N., L. Bouneau, et al. (2003). "Diversity of retrotransposable elements in compact pufferfish genomes." Trends Genet **19**(12): 674-678.

APPENDIX A

CHAPTER II SUPPLEMENTARY FIGURES AND TABLES

Table A.1 The estimated ages of DNA transposons

Element Name:	Intact ORFs	Mean Distance	Standard Deviation	Range		Mean Activity Period (mya)	Time Range (mya)		Copy #
npiggyBac-5_Hm		0.0098	0.0228	0.0000	0.0325	0.51	0.00	1.70	28
Tc3-1_Hm	43	0.0131	0.0614	0.0000	0.0745	0.69	0.00	3.90	449
nMar-19_Hm		0.0167	0.0338	0.0000	0.0505	0.88	0.00	2.64	34
nMar-3_Hm		0.0198	0.0114	0.0084	0.0311	1.03	0.44	1.63	169
Helitron-like-3_Hm		0.0209	0.0365	0.0000	0.0574	1.10	0.00	3.01	2093
nhAT-10_Hm		0.0245	0.0186	0.0059	0.0431	1.28	0.31	2.26	73
nPIF-2_Hm_PIF		0.0268	0.0273	0.0000	0.0542	1.41	0.00	2.84	27
DNA-like-1_Hm		0.0270	0.0249	0.0021	0.0520	1.42	0.11	2.72	381
piggyBac-2_Hm		0.0297	0.0724	0.0000	0.1022	1.56	0.00	5.35	21
nPIF-6_Hm_PIF		0.0298	0.0486	0.0000	0.0784	1.56	0.00	4.11	180
npiggyBac-1_Hm		0.0298	0.0309	0.0000	0.0608	1.56	0.00	3.18	357
Helitron-like-6a_Hm		0.0322	0.0279	0.0043	0.0601	1.69	0.23	3.15	1621
npiggyBac-4_Hm		0.0350	0.0887	0.0000	0.1237	1.83	0.00	6.48	20
piggyBac-1_Hm		0.0350	0.0945	0.0000	0.1294	1.83	0.00	6.78	82
nMar-13_Hm		0.0371	0.0193	0.0178	0.0565	1.95	0.93	2.96	38
nhAT-3_Hm		0.0384	0.0248	0.0136	0.0632	2.01	0.71	3.31	31
nMar-17_Hm		0.0386	0.0436	0.0000	0.0823	2.02	0.00	4.31	94
nMar-2_Hm		0.0396	0.0934	0.0000	0.1330	2.08	0.00	6.97	29
nMar-11_Hm		0.0409	0.0502	0.0000	0.0911	2.14	0.00	4.77	52
nMar-21_Hm		0.0444	0.1049	0.0000	0.1493	2.32	0.00	7.82	139
Helitron-like-4a_Hm		0.0496	0.0626	0.0000	0.1122	2.60	0.00	5.88	713
nMar-10_Hm		0.0515	0.0306	0.0210	0.0821	2.70	1.10	4.30	56
Helitron-like-11_Hm		0.0519	0.0736	0.0000	0.1255	2.72	0.00	6.57	946
DNA-like-10_Hm		0.0529	0.0746	0.0000	0.1275	2.77	0.00	6.68	79
Helitron-like-16_Hm		0.0545	0.0682	0.0000	0.1227	2.86	0.00	6.43	90
nPIF-1_Hm_PIF		0.0555	0.0283	0.0272	0.0839	2.91	1.43	4.39	241
Helitron-like-6b_Hm		0.0561	0.0810	0.0000	0.1371	2.94	0.00	7.18	1184
nMar-15_Hm		0.0562	0.0204	0.0358	0.0766	2.95	1.88	4.01	23
DNA-like-8_Hm		0.0571	0.0620	0.0000	0.1191	2.99	0.00	6.24	209
nhAT-6_Hm		0.0599	0.0483	0.0116	0.1083	3.14	0.61	5.67	219
nhAT-4_Hm		0.0635	0.0418	0.0217	0.1053	3.33	1.14	5.52	144
nhAT-8_Hm		0.0641	0.0416	0.0224	0.1057	3.36	1.18	5.54	270
Helitron-like-4b_Hm		0.0648	0.0733	0.0000	0.1381	3.40	0.00	7.23	1241
npiggyBac-2_Hm		0.0665	0.0984	0.0000	0.1649	3.48	0.00	8.64	39
npiggyBac-3_Hm		0.0666	0.0899	0.0000	0.1565	3.49	0.00	8.20	59
nhAT-7_Hm		0.0666	0.0481	0.0185	0.1147	3.49	0.97	6.01	165
nTc3-1_Hm		0.0670	0.0536	0.0135	0.1206	3.51	0.70	6.32	1237
Helitron-like-12_Hm		0.0673	0.0632	0.0041	0.1305	3.53	0.22	6.84	2243
Helitron-like-7_Hm		0.0678	0.0648	0.0031	0.1326	3.55	0.16	6.95	2822
nhAT-5_Hm		0.0683	0.0502	0.0182	0.1185	3.58	0.95	6.21	120

Table A. 1 Continued

Element Name:	Intact ORFs	Mean Distance	Standard Deviation	Range		Mean Activity Period (mya)	Time Range (mya)		Copy #
npiggyBac-5_Hm		0.0098	0.0228	0.0000	0.0325	0.51	0.00	1.70	28
Tc3-1_Hm	43	0.0131	0.0614	0.0000	0.0745	0.69	0.00	3.90	449
nMar-19_Hm		0.0167	0.0338	0.0000	0.0505	0.88	0.00	2.64	34
nMar-3_Hm		0.0198	0.0114	0.0084	0.0311	1.03	0.44	1.63	169
Helitron-like-3_Hm		0.0209	0.0365	0.0000	0.0574	1.10	0.00	3.01	2093
nhAT-10_Hm		0.0245	0.0186	0.0059	0.0431	1.28	0.31	2.26	73
nPIF-2_Hm_PIF		0.0268	0.0273	0.0000	0.0542	1.41	0.00	2.84	27
DNA-like-1_Hm		0.0270	0.0249	0.0021	0.0520	1.42	0.11	2.72	381
piggyBac-2_Hm		0.0297	0.0724	0.0000	0.1022	1.56	0.00	5.35	21
nPIF-6_Hm_PIF		0.0298	0.0486	0.0000	0.0784	1.56	0.00	4.11	180
npiggyBac-1_Hm		0.0298	0.0309	0.0000	0.0608	1.56	0.00	3.18	357
Helitron-like-6a_Hm		0.0322	0.0279	0.0043	0.0601	1.69	0.23	3.15	1621
npiggyBac-4_Hm		0.0350	0.0887	0.0000	0.1237	1.83	0.00	6.48	20
piggyBac-1_Hm		0.0350	0.0945	0.0000	0.1294	1.83	0.00	6.78	82
nMar-13_Hm		0.0371	0.0193	0.0178	0.0565	1.95	0.93	2.96	38
nhAT-3_Hm		0.0384	0.0248	0.0136	0.0632	2.01	0.71	3.31	31
nMar-17_Hm		0.0386	0.0436	0.0000	0.0823	2.02	0.00	4.31	94
nMar-2_Hm		0.0396	0.0934	0.0000	0.1330	2.08	0.00	6.97	29
nMar-11_Hm		0.0409	0.0502	0.0000	0.0911	2.14	0.00	4.77	52
nMar-21_Hm		0.0444	0.1049	0.0000	0.1493	2.32	0.00	7.82	139
Helitron-like-4a_Hm		0.0496	0.0626	0.0000	0.1122	2.60	0.00	5.88	713
nMar-10_Hm		0.0515	0.0306	0.0210	0.0821	2.70	1.10	4.30	56
Helitron-like-11_Hm		0.0519	0.0736	0.0000	0.1255	2.72	0.00	6.57	946
DNA-like-10_Hm		0.0529	0.0746	0.0000	0.1275	2.77	0.00	6.68	79
Helitron-like-16_Hm		0.0545	0.0682	0.0000	0.1227	2.86	0.00	6.43	90
nPIF-1_Hm_PIF		0.0555	0.0283	0.0272	0.0839	2.91	1.43	4.39	241
Helitron-like-6b_Hm		0.0561	0.0810	0.0000	0.1371	2.94	0.00	7.18	1184
nMar-15_Hm		0.0562	0.0204	0.0358	0.0766	2.95	1.88	4.01	23
DNA-like-8_Hm		0.0571	0.0620	0.0000	0.1191	2.99	0.00	6.24	209
nhAT-6_Hm		0.0599	0.0483	0.0116	0.1083	3.14	0.61	5.67	219
nhAT-4_Hm		0.0635	0.0418	0.0217	0.1053	3.33	1.14	5.52	144
nhAT-8_Hm		0.0641	0.0416	0.0224	0.1057	3.36	1.18	5.54	270
Helitron-like-4b_Hm		0.0648	0.0733	0.0000	0.1381	3.40	0.00	7.23	1241
npiggyBac-2_Hm		0.0665	0.0984	0.0000	0.1649	3.48	0.00	8.64	39
npiggyBac-3_Hm		0.0666	0.0899	0.0000	0.1565	3.49	0.00	8.20	59
nhAT-7_Hm		0.0666	0.0481	0.0185	0.1147	3.49	0.97	6.01	165
nTc3-1_Hm		0.0670	0.0536	0.0135	0.1206	3.51	0.70	6.32	1237
Helitron-like-12_Hm		0.0673	0.0632	0.0041	0.1305	3.53	0.22	6.84	2243
Helitron-like-7_Hm		0.0678	0.0648	0.0031	0.1326	3.55	0.16	6.95	2822
nhAT-5_Hm		0.0683	0.0502	0.0182	0.1185	3.58	0.95	6.21	120

Table A.2 The estimated ages of Non-LTRs

Element Name:	Intact ORFs	Mean Distance	Standard Deviation	Range		Mean Activity Period (mya)	Time Range (mya)		Copy #
Jockey-3_Hm	2	0.0521	0.0948	0.0000	0.1469	2.73	0.0000	7.6973	247
Daphne-1_Hm		0.0569	0.1293	0.0000	0.1862	2.98	0.0000	9.7544	544
R4-1_HMe		0.0618	0.0864	0.0000	0.1482	3.24	0.0000	7.7639	1171
Vingi-1_Hm		0.0727	0.0976	0.0000	0.1703	3.81	0.0000	8.9192	251
Daphne-2_Hm		0.0833	0.0803	0.0030	0.1637	4.36	0.1560	8.5730	3326
CR1-8_Hm		0.0834	0.1495	0.0000	0.2330	4.37	0.0000	12.2042	101
RTE-9_Hm	2	0.0907	0.1053	0.0000	0.1960	4.75	0.0000	10.2689	130
RTE-3_Hm	6	0.0951	0.0836	0.0116	0.1787	4.98	0.6060	9.3593	2893
L2-6_Hm		0.0999	0.1062	0.0000	0.2061	5.23	0.0000	10.7953	269
CR1-4.Hm		0.1092	0.1259	0.0000	0.2352	5.72	0.0000	12.3181	54
RTE-15_Hm	1	0.1109	0.0874	0.0235	0.1983	5.81	1.2332	10.3864	291
CR1-9_Hm		0.1154	0.1490	0.0000	0.2644	6.05	0.0000	13.8508	125
CR1-2.Hm		0.1190	0.0995	0.0195	0.2185	6.23	1.0192	11.4450	90
R1-2_Hm	2	0.1319	0.1341	0.0000	0.2660	6.91	0.0000	13.9337	430
Jockey-1_Hm	3	0.1335	0.1237	0.0098	0.2572	6.99	0.5132	13.4734	1030
Zenon-1_Hm	1	0.1347	0.0949	0.0398	0.2296	7.06	2.0858	12.0269	2980
RTE-18_Hm		0.1410	0.1298	0.0112	0.2708	7.39	0.5850	14.1867	58
Jockey-4_Hm		0.1446	0.1688	0.0000	0.3135	7.58	0.0000	16.4221	62
L2-2_Hm		0.1459	0.1388	0.0071	0.2846	7.64	0.3737	14.9107	149
Jockey-7_Hm		0.1549	0.1351	0.0198	0.2900	8.12	1.0396	15.1929	70
L2-1_Hm	1	0.1606	0.1361	0.0245	0.2967	8.41	1.2831	15.5431	287
I-1_Hm		0.1666	0.1102	0.0564	0.2767	8.73	2.9539	14.4968	2039
RTE-4_Hm		0.1720	0.1469	0.0251	0.3189	9.01	1.3132	16.7037	65
RTE-10_Hm	1	0.1752	0.1424	0.0328	0.3175	9.18	1.7167	16.6336	325
RTE-2_Hm		0.1768	0.1301	0.0467	0.3070	9.26	2.4480	16.0799	1038

Table A.2 Continued

Element Name:	Intact ORFs	Mean Distance	Standard Deviation	Range		Mean Activity Period (mya)	Time Range (mya)		Copy #
RTE-5_Hm	2	0.1806	0.1632	0.0173	0.3438	9.46	0.9082	18.0093	65
CR1-3.Hm		0.1870	0.1758	0.0112	0.3628	9.80	0.5881	19.0051	108
R1-1_Hm		0.1917	0.1292	0.0625	0.3208	10.04	3.2755	16.8063	288
Zenon-3_Hm		0.1999	0.1341	0.0658	0.3340	10.47	3.4467	17.4975	246
RTE-7_Hm		0.2036	0.1604	0.0433	0.3640	10.67	2.2664	19.0674	114
RTE-1_HMe	1	0.2142	0.1770	0.0372	0.3912	11.22	1.9477	20.4903	146
L2-14_Hm	1	0.2201	0.1348	0.0853	0.3549	11.53	4.4679	18.5884	386
Jockey-2_Hm		0.2215	0.1187	0.1028	0.3402	11.60	5.3859	17.8218	186
L2-15_Hm	1	0.2263	0.1420	0.0843	0.3682	11.85	4.4157	19.2897	87
Zenon-2_Hm	2	0.2269	0.1428	0.0840	0.3697	11.88	4.4018	19.3664	327
RTE-20_Hm	1	0.2336	0.1110	0.1226	0.3445	12.23	6.4222	18.0463	383
L2-12_Hm		0.2433	0.1893	0.0541	0.4326	12.75	2.8326	22.6614	263
Jockey-8_Hm		0.2563	0.1889	0.0675	0.4452	13.43	3.5334	23.3213	43
L2-9_Hm	3	0.2655	0.1479	0.1176	0.4134	13.91	6.1603	21.6552	160
L2-13_Hm	1	0.2669	0.1115	0.1553	0.3784	13.98	8.1375	19.8214	297
R4-2_Hm	1	0.2699	0.1552	0.1146	0.4251	14.14	6.0051	22.2676	86
RTE-13_Hm		0.2761	0.1349	0.1412	0.4109	14.46	7.3972	21.5262	63
Proto2-1_Hm		0.2773	0.1747	0.1026	0.4520	14.53	5.3740	23.6777	59
RTE-6_Hm		0.2906	0.1356	0.1550	0.4261	15.22	8.1197	22.3214	82
L2-11_Hm		0.2911	0.1530	0.1381	0.4441	15.25	7.2342	23.2656	164
L2-7_Hm	1	0.2998	0.1673	0.1325	0.4671	15.70	6.9412	24.4687	245
CR1-6_Hm		0.3023	0.1792	0.1231	0.4815	15.84	6.4492	25.2250	246
RTE-8_Hm		0.3029	0.1415	0.1615	0.4444	15.87	8.4583	23.2780	111
L2-10_Hm		0.3065	0.1183	0.1882	0.4248	16.05	9.8571	22.2519	433
Proto2-3_Hm	1	0.3095	0.1943	0.1152	0.5038	16.21	6.0365	26.3910	37
RTE-16_Hm		0.3150	0.1519	0.1632	0.4669	16.50	8.5472	24.4586	121
Jockey-5_Hm		0.3200	0.1511	0.1689	0.4711	16.76	8.8463	24.6796	140
L2-5_Hm		0.3300	0.1090	0.2209	0.4390	17.28	11.5729	22.9967	182
CR1-1.Hm		0.3331	0.1511	0.1819	0.4842	17.45	9.5304	25.3639	107
L2-3_Hm		0.4132	0.0919	0.3213	0.5050	21.64	16.8309	26.4555	187

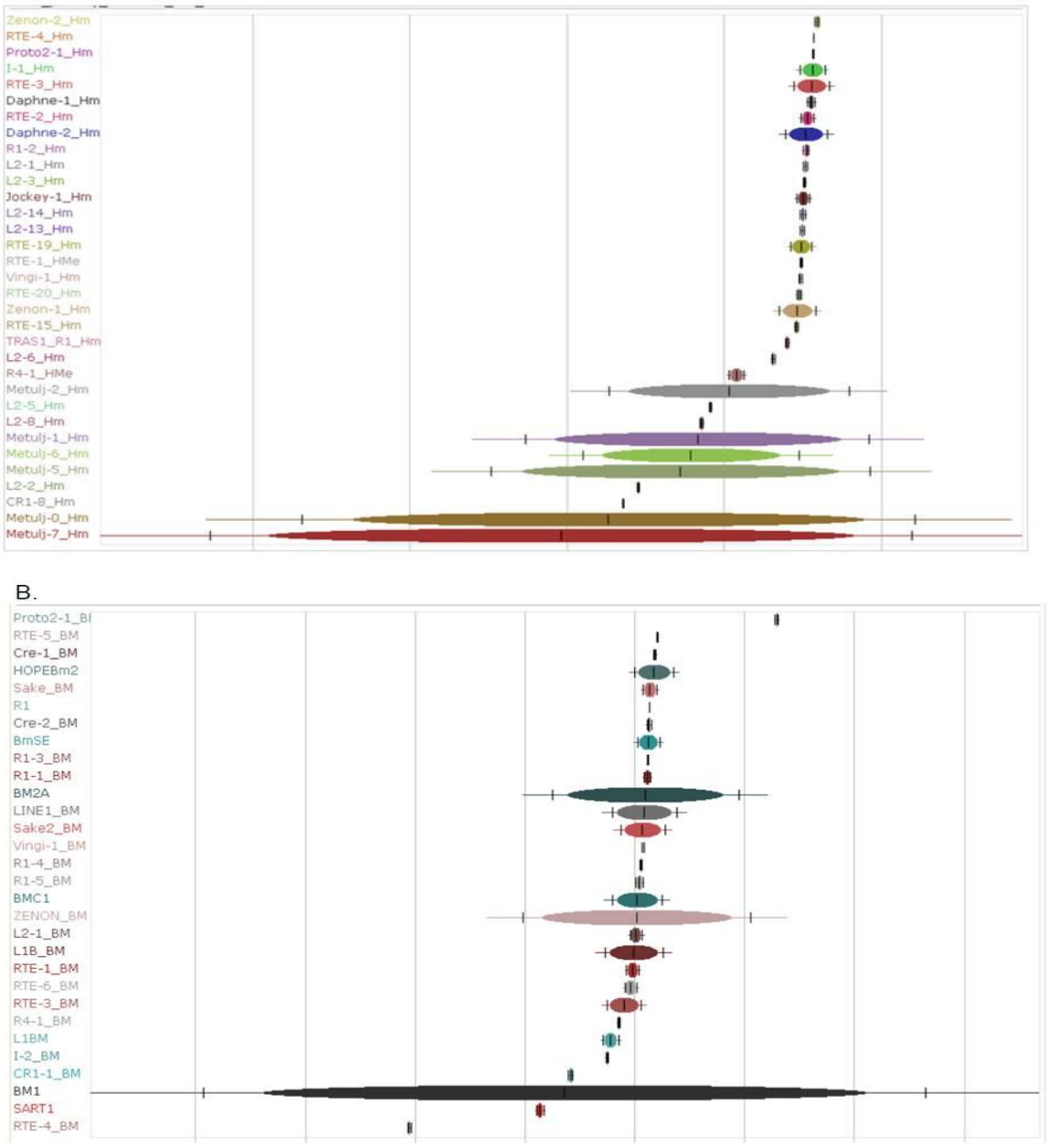

Figure A.1 Results of the TinT analysis for *H.melpomene* and *B. mori* non-LTR elements.

TinT analysis for *H.melpomene* (top) and *B. mori* (bottom). TinT uses patterns of nested insertion to predict relative activity periods among TEs. In the graph, periods of probable activity are depicted by an oval (period of maximum activity), vertical lines (95% of the probable activity period), and horizontal lines (99% of the probabl activity period).

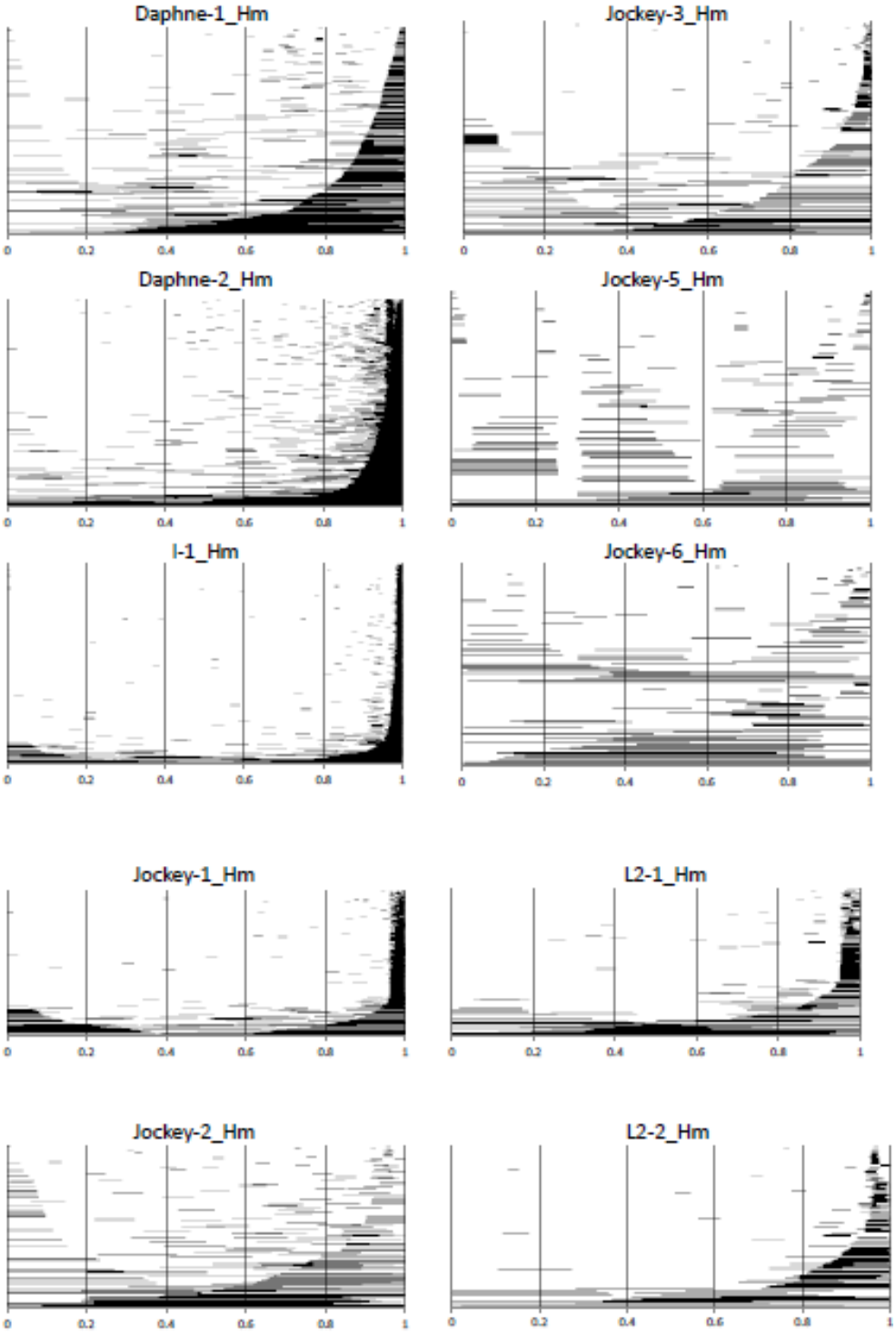

Figure A.2 Length distributions of *H. melpomene* LINE insertions.

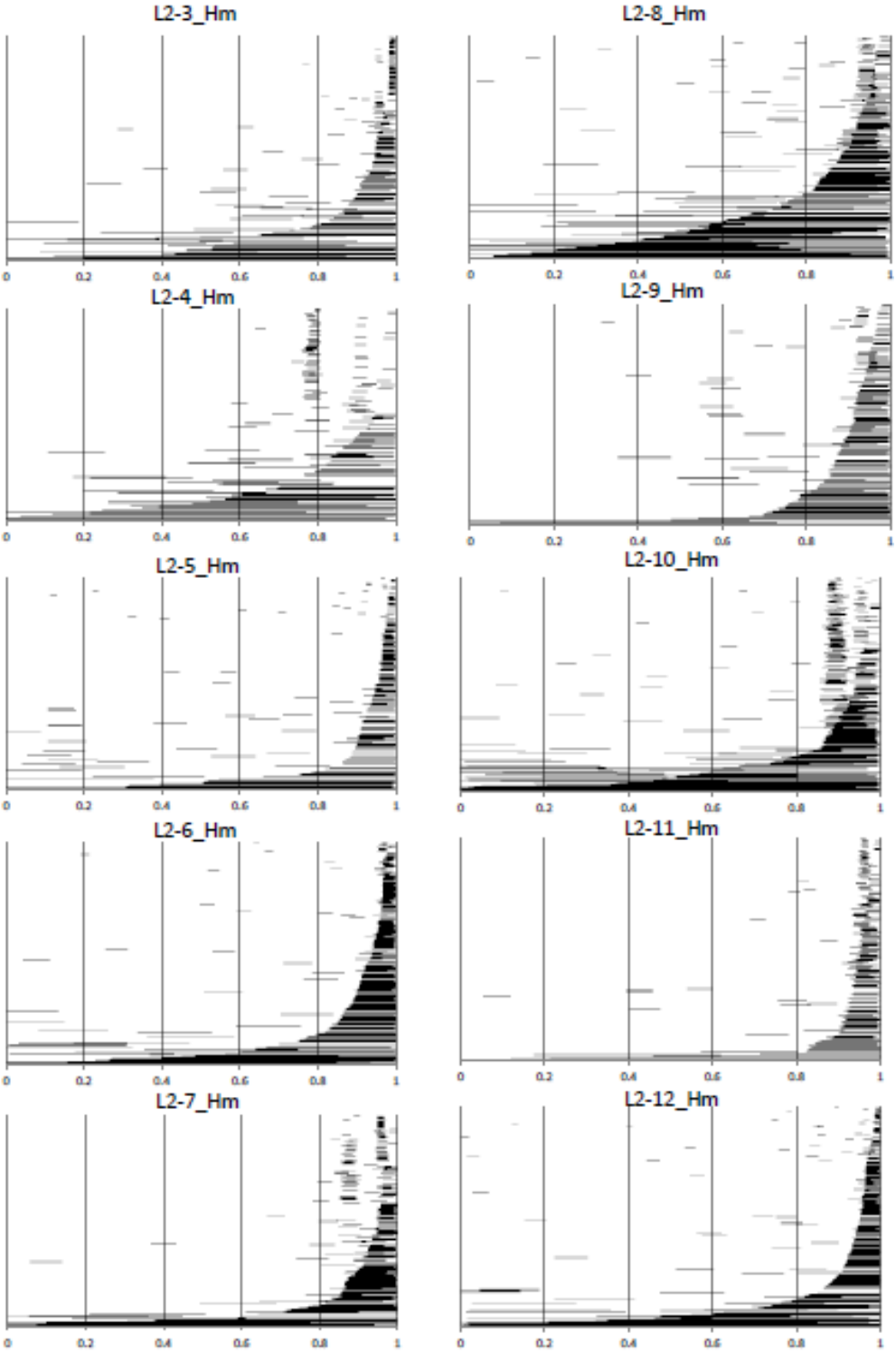

Figure A.2 Continued

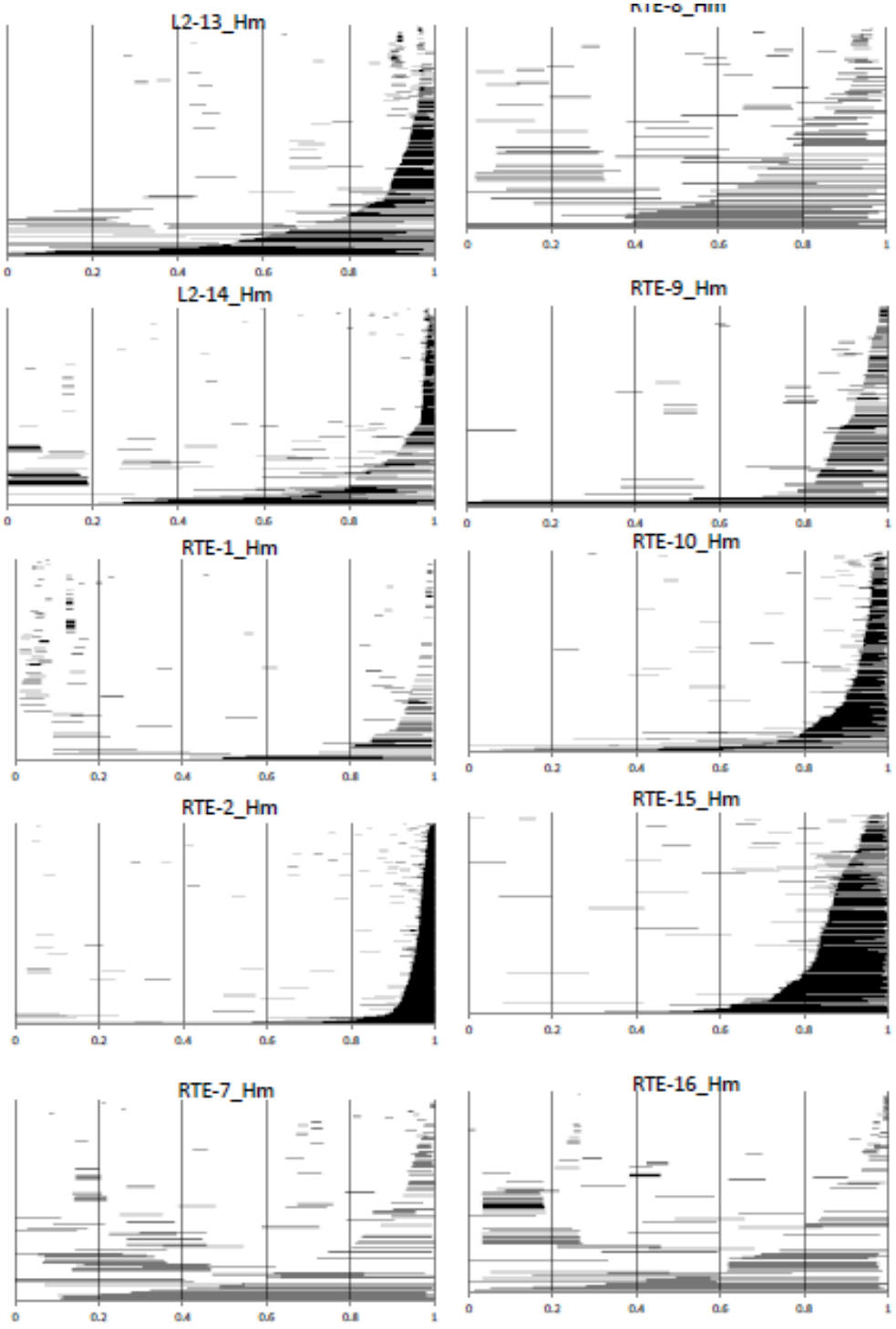

Figure A.2 Continued

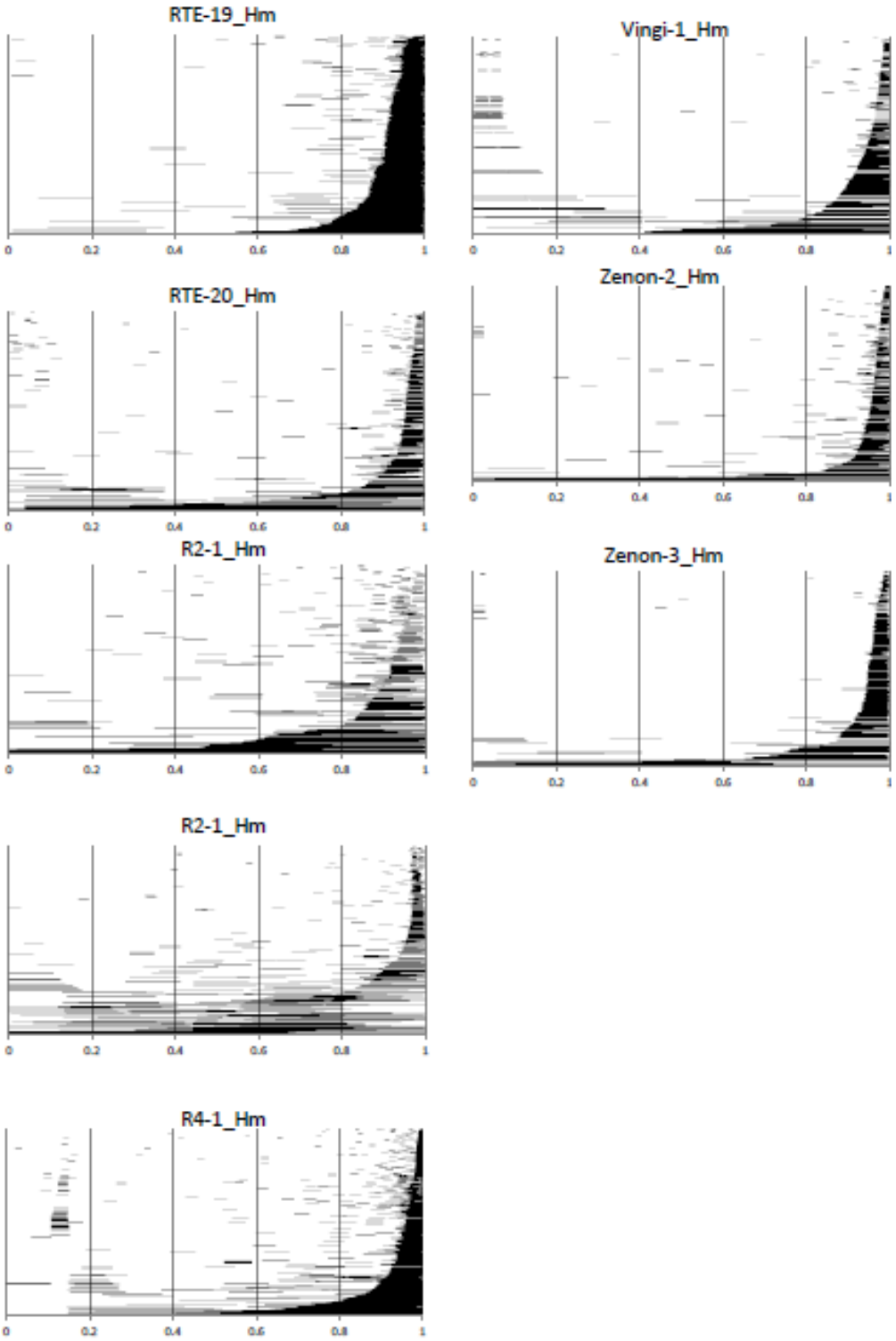

Figure A.2 Continued