EVOLUTION AND DEVELOPMENT OF A NOVEL TRAIT IN SEPSIDAE

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Dacotah Michael Melicher

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Biological Sciences

March 2016

Fargo, North Dakota

North Dakota State University
Graduate School

Title
EVOLUTION AND DEVELOPMENT OF A NOVEL TRAIT IN SEPSIDAE

By

Dacotah Michael Melicher

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Julia Bowsher
Chair

Dr. Steve Travers

Dr. Kendra Greenlee

Dr. Philip McClean

Approved:

3/23/16	Dr. Wendy Reed
Date	Department Chair

ABSTRACT

Evolutionary novelty, the appearance of new traits with no existing homology, is central to the adaptive radiation of new species. Novel traits inform our understanding of development and how developmental mechanisms can generate novelties. Sepsid flies (Diptera: Sepsidae) have a sexually dimorphic, jointed appendage used for courtship and mating. The appendage develops from the fourth abdominal histoblast nest rather than an imaginal disc. Histoblast nests in other species produce the adult epidermis and lack three-dimensional organization. The sepsid system is an opportunity to investigate the evolutionary history of a novel trait and the developmental mechanisms that pattern epidermal tissue into a complex structure.

The appendage has a complex history of gain, loss, and recovery over evolutionary time. Appendage morphology is highly variable between species and does not correlate to body size. I collected larval epidermal tissue from 16 species across Sepsidae and one outgroup to trace the evolutionary history of gain, secondary loss, and recovery. I characterized histoblast nests in all segments and sexes, determining the nest size, number, and size of cells. The appendage-producing nest is sexually dimorphic in species after primary gain. Loss of the appendage shows a return to ancestral state while regain shows an increase in nest size in both sexes. The loss of sex dimorphism may indicate that mechanisms involved in specification may be active in females while genes involved in patterning are not activated during pupation.

I assembled and annotated a reference transcriptome for the sepsid *Themira biloba* at using a custom bioinformatic pipeline that uses a merged assembly approach to maximize quality. This pipeline demonstrated an improvement over other methodologies using multiple published metrics for determining quality and completion. This pipeline also demonstrates how cloud computing architecture can complete bioinformatic tasks quickly and at low cost.

I used the *T. biloba* transcriptome to identify differentially expressed genes involved in appendage patterning during pupation. I sequenced the appendage producing fourth male larval segment and the third male and fourth female segments. Many of the differentially expressed transcripts are involved in cell signaling, epidermal growth, and transcripts involved morphological development in other species.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Julia Bowsher, for accepting me as her first doctoral graduate student. Over the term of my graduate career she has given me invaluable mentoring and advice while exhibiting great patience and optimism. I would like to thank my committee members, Dr. Kendra Greenlee, Dr. Steven Travers, and Dr. Phil McClean for their guidance and review of my work. I thank Dr. Wendy Reed for introducing me to my advisor and providing me with guidance entering the graduate program and for her continued advice and support. The NDSU Department of Biological Sciences as a whole has provided me with a comfortable and friendly research environment, financial support for research and conferences, a social community, and has been a second home to me during my undergraduate and graduate academic career. I thank members of the Bowsher lab for their continued support and friendship; Bodini Herath, Alex Torson, Garett Slater, and Bryan Helm.

Many collaborators assisted in the production of the research presented in this dissertation. I thank Dr. Ian Dworkin of Michigan State University for hosting myself and Alex Torson during a bioinformatic research exchange and assisting with the construction and review of our bioinformatic pipeline published in 2014. Thanks to Dr. Rudolf Meier of the National University of Singapore for hosting me during my data collection in 2012 and for his continued support with ongoing collaborative projects. Raphael Royette provided invaluable statistical help during data analysis. Pawel Borowicz of the NDSU Advanced Imaging and Microscopy core lab improved the sophistication of our imaging protocols and helped streamline our data collection. I thank George Yocum, Joe Reinhardt, and William Kemp from the Fargo, ND United States Department of Agriculture Agricultural Research Service (USDA ARS) for advice and guidance and internal review of my work as well as an introduction to collaborators.

Thanks to the Evo-Devo-Eco Network (EDEN) for providing travel funding for a research exchange to the Dworkin lab at Michigan State. The National Science Foundation East Asia and Pacific Summer Institute (EAPSI) fellowship allowed me to travel to the Meier lab in Singapore to collect materials and data presented in this dissertation as well as reinforce our collaborative relationship with another member of the sepsid research community. The NDSU Department of Biological Sciences provided startup funding which was used to fund the research presented here as well as a research assistantship and travel funds to attend conferences. The College of Science and Math also provided travel funds and facilitated the work presented here. The USDA ARS also provided financial support in the form of a research assistantship which was vital to the timely completion of my research.

I would like to thank Lauren Dennhardt who I met at the beginning of my graduate program and married. She has been my best friend and biggest supporter and convincing her to become my wife is arguably my greatest achievement during my graduate career. Thanks to my parents, Kathleen and Will Cooper, Steve and Cindy Melicher, my siblings, Mary Seelye, Jake, and Anna Melicher, for everything they have done for me over the years that prepared me for my academic and life achievements and being present during these milestones.

DEDICATION

This dissertation is dedicated to my wife Lauren, Kathleen and Will Cooper, Steve and Cindy Melicher, and several thousand sepsid flies, particularly *Themira biloba*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Arch.....*Archisepsis armata*

Asep .....*Allosepsis indica*

Dicra.....*Dicranosepsis sp.*

Malb .....*Meroplius albuquerque*

Mfas .....*Meroplius fasciculatus*

Marm.....*Microsepsid armillata*

Nnit .....*Nemopoda nitidula*

Oluc.....*Orygma luctuosum*

Pdik .....*Perochaeta dikowi*

Sful.....*Sepsid fulgens*

Slat .....*Sepsis latiforceps*

Spun .....*Sepsis punctum*

Tbil.....*Themira biloba*

Tfla.....*Themira flavicoxa*

Tluc .....*Themira lucida*

Tmin.....*Themira minor*

Tput.....*Themira putris*

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

CHAPTER ONE: INTRODUCTION

An introduction to evolutionary novelty

The concept of evolutionary novelty, the generation and diversification of new traits, is central to the generation and diversification of new species. An evolutionary novelty most generally is a trait possessed by an organism that which does not exist outside of a specific clade (Muller and Wagner 1991, Moczek 2008). Prior to the invention and refinement of gene sequencing technologies and for much of the history of taxonomy, organisms have been categorized broadly based on an evolutionary history of novel traits that result in very different morphologies. These novelties affect the symmetry and modularity of organism body plans, the function and organization of internal organs, the number, location, and morphological structure of appendages. The appearance of novel traits such as insect flight are often accompanied by rapid diversification and speciation (Jones and Teeling 2006). The adaptive radiation of a new trait is an important driver of speciation, while modifications of a trait by evolution through natural selection allow populations to adapt to changing ecosystems. Thus, novel morphologies often allow organisms to occupy ecological niches and exploit resources which were previously inaccessible (Dobzhansky 1963).

Understanding the evolution of novel traits allows us to gain insight into the development innovations that lie at the base of successful taxa and adaptive radiations. These innovations ultimately become widespread and derived traits in a large number of species. Traveling backward through evolutionary history, taxonomic diversification collapses into single species which represent individual instances of the evolution of novel traits. For this reason understanding novelty in existing systems is an opportunity to identify the mechanisms that result in diverse morphologies, physiologies, and life histories.

What defines a novelty? The definition of a novelty has been refined and become more specific as interest has shifted from a generalized historical interest in the origins of characteristics that are unique to one level of classification, such as feathers on birds, to the identification of genes and developmental mechanisms that produce a trait that is constrained to a family of organisms with no outside homologies (Moczek 2008). In 1859 at the time of publication of "*On the origin of species"*, there existed an intriguing gulf between minute adaptations which are the measureable product of evolution acting on a population and the vast diversity of life spread across the planet. As stated earlier all traits were novel at one point in evolutionary history and Darwin expresses concerns and intense curiosity over these novel moments which generated the first instance of a simple eye or feather which becomes refined, complex, and diverse across species with little else in common (Darwin 1859). Thus we can identify novelties at many different levels through evolutionary history and speciation, homology or lack thereof, a novel trait that allows organism's access to new ecological resources, and the genetic and molecular mechanisms directly responsible for producing novelties.

Novel traits can be traced back to an initial evolutionary occurrence in an ancestral species. The eye is a good example in that photosensitive opsin proteins allowed light detection which diversified into complex structures in many lineages (Darwin 1859). Other examples include the first occurrence of body coverings such as keratinized fur, feathers, and scales, multicellular, symmetrical and modular body plans, and the appearance of limbs (Wagner and Lynch 2010). The novelty itself is as specific as a single light-detecting protein or a co-opted network of genetic signaling that results in a localized multicellular structure that allows 'light detecting' organisms to become 'sighted' organisms with the ability to resolve patterns in the intensity of detected light. For this reason novelties are valuable for taxonomic purposes and to

identify the mechanisms and constraints of trait evolution. Novelties are not restricted to morphological traits. Novelties include new metabolic, physiological, and behavioral traits which may allow organisms to exploit new food resources, survive in an extreme environment, or form complex social structures. Morphological traits are the most visible and quantifiable and are the product of complex gene interactions and regulatory networks and environmental conditions. Novelty can be found in decreasing levels of relative genetic complexity in the diversity of snake venom toxins, antimicrobial peptides, and individual enzymes (Clark et al. 1994, Park et al. 1996, Goyal et al. 2005, Casewell et al. 2013). By shifting our focus from the evolutionary history, origin, and diversification of traits to recently evolved novel traits we can identify novelties which are in the process of diversification.

While a broad and historical view of novelty is useful for taxonomic purposes and identifying evolutionary forces that produce species and morphological diversity, a proximate investigation of genetic components helps us understand the mechanisms that produce novelties. Morphological evolution is the result of changes in the location, level, or duration of the expression of developmental genes. Existing genes may be repurposed or co-opted through evolutionary processes which modify the function of gene paralogues produced by duplication events (Abbasi 2010, Soshnikova et al. 2013). Processes such as alternative splicing also increase the number of unique developmental proteins and the diversity of their functions (Graveley 2001). This affects morphology using an existing genetic toolkit. Organisms which exhibit a segmented, or serially homologous body plan, are able to make changes in the number of segments, serial duplication of segments, the placement and identity of appendages, and appendage morphology by making regulatory changes to existing gene networks (Carroll 2005, Gompel et al. 2005, Prud'homme et al. 2006). These processes have also been shown to be a

source of evolutionary novelty in organisms such as butterflies which use a co-opted toolkit of appendage genes to produce novel eyespot colorations on their wings and horns of scarab beetles which use epidermal growth factors and apoptosis to form complex structures used for mating and combat (Nijhout 1991, Moczek 2005, 2006, Emlen et al. 2007).

Research into novel traits is challenging for several reasons. First, novel traits as described above are uncommon. Novelty occurs outside of model systems and many species are not amenable or adaptable to rearing in a laboratory. The nature of the novelty must be taken into consideration when attempting to assess the strength of a system and designing a research plan. Non-model organisms are often intractable systems for molecular techniques because genomic sequence data is limited or absent. For these practical and economic reasons, novelty in non-model systems are underrepresented in evolutionary and developmental research. Some of the obstacles that make investigating novelties and non-model systems so challenging have largely been solved in the form of improvements in sequencing technologies that allow increased accessibility to genomic and mRNA sequences and as a result many molecular tools for identifying, measuring, and modifying gene expression (Wang et al. 2009, Oshlack et al. 2010, Ekblom and Galindo 2011, Grabherr et al. 2011).

Sepsid novel abdominal appendages

Sepsid flies (Diptera: Sepsidae), also known as black scavenger flies, are an excellent candidate system for investigating the evolution of novel traits for several reasons. Sepsidae is a large and diverse family of approximately 250 known species with worldwide distribution (Pont and Meier 2002). Several clades within Sepsidae bear a novel modified sternite and brush-like appendage on their fourth abdominal segment (Figure 1.1.). The appendage is jointed and highly mobile. It is sexually dimorphic and used by male flies during courtship and mating

(Puniamoorthy et al. 2009). Appendage morphology and modification of the sternite vary greatly between species. The abdominal appendages have also evolved several times within Sepsidae with clades representing a primary gain, secondary loss, and tertiary regain or recovery of the appendage (Eberhard 2001b, Bowsher et al. 2013).

Figure 1.1. Abdomens from *Themira biloba* showing the modified sternite and abdominal appendages (left) with close-up on bristle sockets and joint (right).

While sepsid taxonomy and behavior are well-described in the literature (Pont 2002, Puniamoorthy et al. 2008, 2009, Iwasa and Thinh 2012), sepsid abdominal appendages have received little attention (Eberhard 2001b) and much is still unknown about the genetic mechanisms that specify and pattern them during development. In flies, "true" appendages, such as wings and legs, develop from imaginal discs which are identifiable clusters of cells set aside during larval development (Figure 1-3A). In contrast, sepsid abdominal appendages develop from histoblast nests which are small clusters of cells that lack three-dimensional organization (Figure 1.3.B) and form the epidermis in other segments during pupation (Bowsher and Nijhout 2007). Genes known to regulate limb development in *Drosophila* do not appear to be involved in appendage morphogenesis (Bowsher and Nijhout 2007, 2009). Histoblast cell numbers in the nests indicate that the size of the nest is not constant between segments, showing a marked increase in cell numbers in the fourth segment in males that have appendages (Bowsher et al.

2013). The process which patterns disorganized histoblast cells into the abdominal appendages is currently unknown.

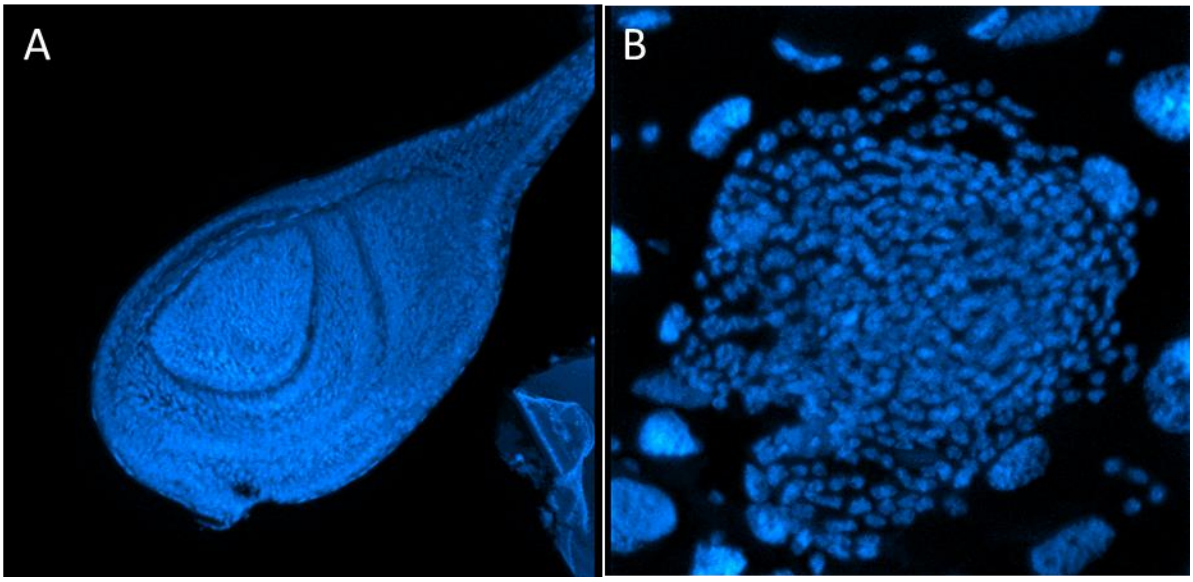

Figure 1.2. Themira biloba larval imaginal discs (A) produce adult appendages and are highly organized. Ventral histoblast nests (B) that produce the abdominal appendages lack the three-dimensional orgaization and complexity of imaginal discs.

Previous phylogenetic research in sepsid flies has shown that the evolutionary history of the sepsid abdominal appendage shows a complex pattern of primary gain, secondary loss, and regain of the appendage in different clades (Bowsher et al. 2013). The appendage develops from histoblast nests, which normally produce a single-cell sheet of adult epidermal tissue. The appendage and modified sternite develop from the 4th segment ventral histoblast nests during pupation. Characterizing the histoblast nests across Sepsidae in species that are representative of the pattern of gain, loss, and recovery may reveal developmental patterns relating to species, sex, organism size, cell number, and histoblast nest size. The sternite, joint, and bristles that form the appendage are highly diverse in morphology even between primary gain species, and species that secondarily lost and recovered the appendage are morphologically distinct (Figure 1.4) (Bowsher et al. 2013). Characterization of the histoblast nest that produces the appendage is essential to

understand the evolutionary history of the appendage itself, and it may identify whether distinct morphologies exist at earlier life stages which may indicate whether appendage patterning occurs in some form prior to pupation.

Investigation of the appendages in *T. biloba* using molecular tools has been limited by the lack of gene sequences available for gene expression analysis. Their distance from *Drosophila* also prevents us from using *Drosophila* sequences to generate molecular primers and probes. This critical limitation has prevented investigation of the appendages and the development of sepsid flies as an emerging model organism. Developing sequence resources for a sepsid species is essential to identifying the genes involved in specification of the histoblast nest and patterning of the developing appendage. Methods and best practices exist for developing *de novo* sequence resources in the form of reference genomes and transcriptomes for us in gene expression studies which can be implemented or modified for use in the sepsid system (Hornett and Wheat 2012, Vogel and Wheat 2012, Wheat and Vogel 2012). However, many genes of interest which specify and pattern the sepsid appendages are likely to be present at low expression levels so maximizing the retrieval of high quality, full length contiguous sequences (contigs) is critical for future gene expression studies, knockdowns and knockouts, and transgenic lines.

Objectives

My research goal is to identify the genetic mechanisms that specify and pattern the abdominal appendages in the sepsid fly *Themira biloba*. My first objective is to characterize the histoblast nests in species across Sepsidae that represent gain, loss, and recovery of the novel appendage and identify patterns in histoblast nest morphology between species and sexes. Second, I will generate a transcriptome from sequences obtained from multiple life stages of *T. biloba* to serve as a reference and improve my ability to use to molecular tools. I will then

sequence specific tissues that generate the appendage and perform a differential expression analysis to identify genes which are up or down-regulated in appendage producing tissues relative to tissues lacking the appendage.

Objective 1: Characterize histoblast nest morphology across Sepsidae

Adult sepsid abdominal appendages develop from the ventral larval histoblast nests during pupation. The histoblast nests in other segments and species produce adult epidermal tissue. It has been shown that the 4th segment male ventral histoblast nest in appendage-producing sepsid species is sexually dimorphic in some but not others, and varies significantly in total size, cell number, and cell size (Bowsher et al. 2013). The first objective will be to characterize histoblast nest size across sexes and species, sampling many species that describe the evolutionary history of gain, loss, and recovery of the appendage. I will also examine an out-group to identify the ancestral state of the histoblast nests. By collecting this information I will construct a model phylogeny that represents the evolutionary history of the appendage using these data as quantitative traits and compare it to an existing phylogeny constructed using gene sequence information.

Objective 2: Sequence mRNA from Themira biloba, assemble, and annotate a reference transcriptome

No reference transcriptome currently exists for any sepsid species, which limits the molecular techniques that are available to identify and interact with genes in this system. While antibodies to some *Drosophila* candidate genes exist (Gay et al. 1988, Rideout et al. 2007, Sanders and Arbeitman 2008), they may not bind with the same affinity in sepsid tissues. Without reference sequences it is not possible to use staining techniques such as *in situ* hybridization, real-time PCR, and mRNA-seq differential expression analysis. For these reasons

the second objective will be to develop a reference transcriptome using mRNA extracted from the sepsid fly *Themira biloba*, which will be used to perform downstream investigation of gene expression and allow interaction with genes of interest. The hypotheses listed under this objective describe the set protocols that have been designed using similar approaches described in the literature for the *de novo* sequencing, assembly, and annotation of a reference transcriptome.

Objective 3: Identify transcripts in specific tissues that are involved in appendage development

Very little is known about the genetic mechanisms that specify the fate of the histoblast cells to become the abdominal appendages, the genes that pattern the appendages during pupation, and those that regulate sexually dimorphic expression of these genes. Genes known to regulate limb development in *Drosophila* do not appear to be involved (Bowsher and Nijhout 2007, 2009). The proposed research described under this objective represents a broad attempt to characterize genes that may be involved in morphological appendage development during a critical stage in the histoblast nests just prior to pupation using an RNA-seq approach.

CHAPTER TWO: SEPSID HISTOBLAST NEST MORPHOLOGY

Abstract

The family Sepsidae shows a complex evolutionary history of gain, loss, and regain of a novel abdominal appendage. The appendage is sexually dimorphic, jointed, highly mobile, and used by males during courtship and mating. Sepsid abdominal appendages are highly diverse between species both morphologically and behaviorally. The appendage develops from histoblast nests rather than imaginal discs. I used fluorescent confocal microscopy to characterize the histoblast nest area, cell count, cell size, larval segment length of Sepsid flies from 17 species across 10 genera. Nest morphology in the ancestral species is not sexually dimorphic but there is a pattern that is different between segments. These species represent the evolutionary history of gains and losses of the appendage including one outgroup which retains the ancestral histoblast nest state. I found that histoblast nest morphology in species that gained, lost, and recovered the appendage matches the morphological patterns previously identified. However, a species that represents an independent loss has a different pattern of segment-specific nest morphology.

Introduction

Recovery of lost traits is a violation Dollo's Law. Dollo's Law states that complex biological structures cannot be recovered once lost and enough evolutionary time has passed, because genes that are not under selective pressure undergo rapid degradation and a non-reversible loss of coding sequences (Simpson 1955). Based on mutation rates, it is estimated that coding sequences in silenced genes that are not expressed and exposed to selective pressures would be rendered unrecoverable after >0.5 to 6 million years (Marshall et al. 1994). Modern sequencing methods also add complexity to our understanding of Dollo's law. A primitive

understanding of trait loss states that genes necessary at any stage of the specification or patterning of a trait are lost or silenced, the associated trait is also lost, and mutation causes degradation of coding sequences associated with the trait. Genomics and gene expression analysis have shown that many developmentally important genes are highly pleiotropic and are exposed to selection and maintained in their other functions. In one sense pleiotropic associations allow for the maintenance of genes but also constrain the evolution of function (Stern and Orgogozo 2009).

Exceptions to this law have been identified in many distantly related taxa, (Gould 1970, Domes et al. 2007). The advent of sequencing technologies and the application of multiple genetic markers to phylogenetics has identified phylogenetic anomalies that do not reflect morphological parsimony and appear to be exceptions to Dollo's law, such as the coiling of shells in snails and the recovery of sexual reproduction in several parthenogenetic species, the recovery of wings, digits, tooth morphology, entire stages of life history, and the abdominal appendages in the sepsid *Perochaeta dikowi* (Collin and Miglietta 2008, Bowsher et al. 2013).

Sepsid flies possess an abdominal appendage with a complex evolutionary history of gains, losses, and recovery (Eberhard 2001b). The recovery of this complex trait appears to violate Dollo's law. The appendage is a complex structure consisting of a modified sternite, a moveable joint, and bristles. The abdominal appendage is sexually dimorphic, occuring only in adult males, and is used for complex courtship behaviors. The abdominal appendage is an evolutionarily novel structure and has no known homology in Diptera (Eberhard 2001a, 2001b). The abdominal appendage also has a novel developmental mechanism in that it develops from histoblast nests, clumps of cells without three-dimensional organization that produce sheets of adult epithelium. Until recently, histoblast nests were not known to produce appendages

(Bowsher and Nijhout 2007). Histoblast nests in non-sepsid Dipterans show no segment-specific or sexually dimorphic patterns in cell number or nest size (Bowsher and Nijhout 2007). Males in some species of sepsids have enlarged histoblast nests with cell number peaking in the appendage-producing segment (Figures 2.1 - 2.2). This pattern disappears with the loss of the appendage. Interestingly, recovery of the appendage results in an enlarged histoblast nest in both sexes, yet the appendage still only occurs in adult males, not females (Bowsher et al. 2013).

Sepsid flies appear to have rapidly speciated after acquiring the appendage and some species are morphologically indistinct. Previous research has shown that sepsid appendages have evolved multiple times and may have multiple developmental mechanisms (Eberhard 2001b, Bowsher et al. 2013). It is my objective to identify patterns in histoblast nest morphology constrained by evolutionary history and developmental processes. A greater sampling of Sepsidae may identify currently unknown histoblast nest morphologies and will identify the extent of variation that exists between species. Further investigation and comparison between closely related species will increase knowledge about these mechanisms and determine if differences in histoblast nest morphology between species represent distinct character states which describe nest morphology within a clade after gain or loss or if transitional morphologies exist between clades. If the evolutionary history gains and losses of the abdominal appendage in sepsid flies were indeed distinct character states, it would add evidence to the hypothesis that appendage recover is a result of recovery of a modified histoblast nest specification pathway. It is also necessary to measure segment length to correct for morphological variation between sexes and to control for body size when comparing histoblast nest morphology between species. I have identified 17 species which represent the evolutionary history of gains and losses of the appendage including closely related species and clades that have not been investigated in

previous research including an outgroup which is expected to retain the ancestral state of the appendage. Measurements of histoblast nest area, cell number, cell nuclei size, and segment length were taken. These efforts identified an ancestral pattern of histoblast nest morphology that is not sexually dimorphic, but is not consistent between segments. It also appears that independent losses of the appendage result in a loss in sexual dimorphism but the segment-specific pattern is different between the two loss groups.

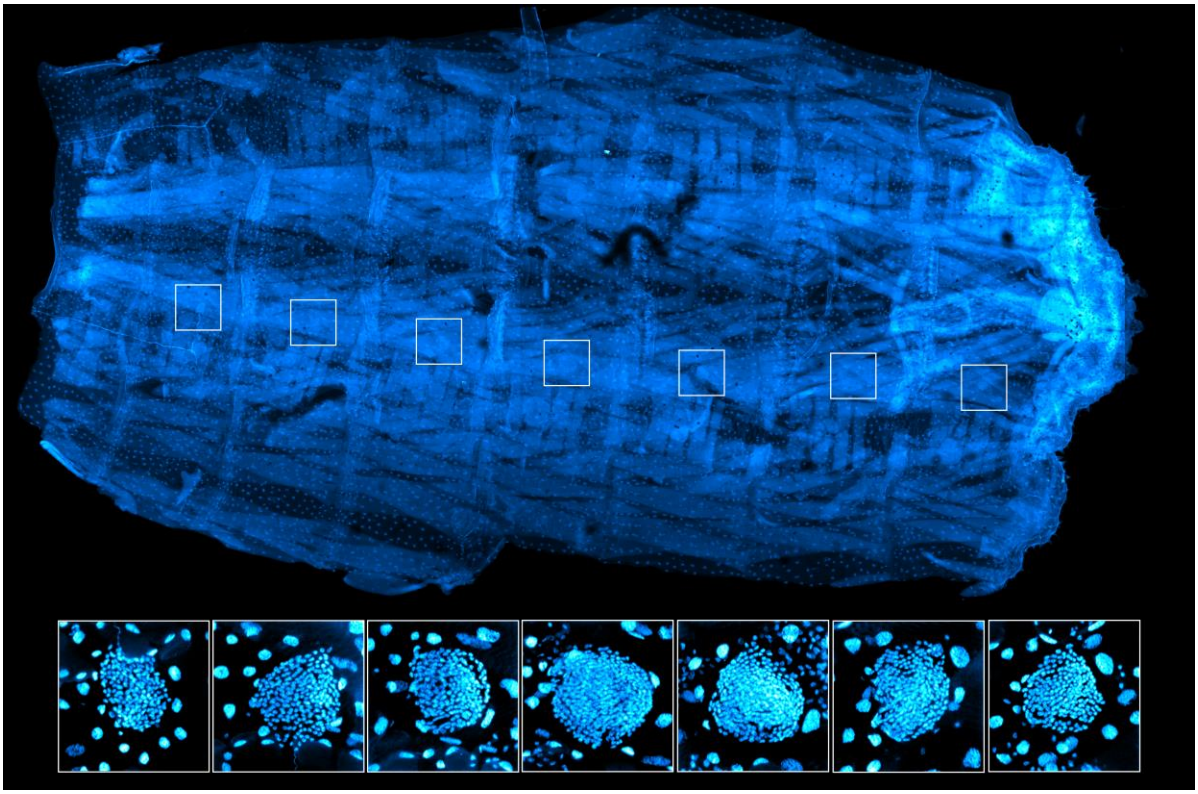

Figure 2.2. *Themira biloba* larval epidermis with ventral histoblast nests in abdominal segments enlarged showing increased nest size, cell count, and cell density peaking in the 4th abdominal segment.

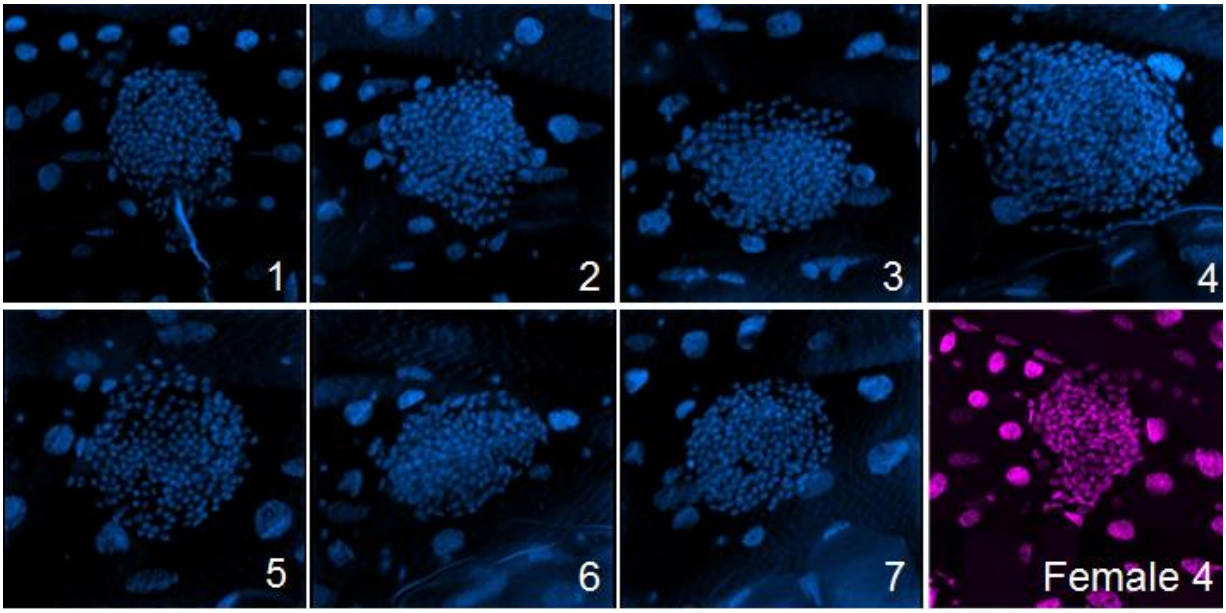

Figure 2.3. *Themira biloba* ventral histoblast nests from abdominal segments 1-7. Histoblast nest size, cell number, and cell density peak in the 4th male segment. The 4th female segment is shown for comparison.

Table 2.1. Sampled species

Species	Abbreviation	Males	Females
Allsepsis indica	Asep	14	13
Archisespsis armata	Arch	15	11
Dicranosepsis	Dicra	12	13
Meroplius albequerqui	Malb	17	17
Meroplius fasciculatus	Mfas	5	6
Microsepsis armillata	Marm	12	15
Nemopoda nitidula	Nnit	8	6
Orygma lucuosum	Oluc	8	7
Perochaeta dikowi	Pdik	13	8
Sepsis fulgens	Sful	10	11
Sepsis latiforceps	Slat	17	18
Sepsis punctum	Spun	9	9
Themira biloba	Tbil	6	6
Themira flavicoxa	Tfla	14	11
Themira lucida	Tluc	12	6
Themira minor	Tmin	8	11
Themira putris	Tput	8	7

Results

The objective of this study was to determine if histoblast nest morphology between sexes and segments was consistent across the evolutionary history of gains and losses of the appendage. Histoblast nest size, cell count, and cell density of the sampled species had three distinct patterns which represent distinct character states described by a previous study (Bowsher et al. 2013). The purpose of this analysis was to determine if these morphological patterns are distinct, track the evolutionary history of gains and losses, or if there are other patterns or transitional patterns present. All of the species that represent the primary gain of the appendage have a distinct, sexually dimorphic increase in histoblast nest size and cell count, which peak in the 4th abdominal segment. Mean histoblast nest size varies between species (Table 2.2, 2.3) but segment-specific and sexually dimorphic patterns are maintained. A likelihood ratio test using chi-square values identified the significance of sex and abdominal segment identity on histoblast cell number (Table 2.4). Species with a significant segment effect have variation in cell number between segments that is consistent between individuals. A significant sex effect indicates a sexually dimorphic pattern. Species such as in *T. biloba* and *N. nitidula* with both significant sex and segment effects indicate a pattern of histoblast nest morphology that is both sexually dimorphic and segment-specific. A linear mixed-effects model revealed the effects of sex and abdominal segment identity on histoblast cell number (Tables 2.5-2.8). *T. biloba* had the largest nests of the species sampled, while *N. nitidula* had largest sexual dimorphism in cell count (Figure 2.5, Table 2.6). The outgroup *O. luctuosum*, and the species from the clades that lost the appendage also lost the enlarged fourth-segment histoblast nest and sexual dimorphism of the nests which indicates that the species have returned to the ancestral state. There appears to be a segment-specific pattern within these species in that histoblast nest cell counts vary between

segments however the pattern is consistent between sexes and not between species (Figure 2.4, 2.6, Table 2.4, 2.7). The species that regained the appendage has enlarged histoblast nests in both sexes peaking in the 4th segment (Figure 2.7, Table 2.8). Charts of the linear mixed-effects model estimated mean for each sex and segment for all sampled species can be found in Appendix B. The patterns of sexual dimorphism and histoblast nest morphology are distinct morphological patterns that track the gains and losses of the appendage in the sepsid evolutionary history as described by Bowsher et. al.

Mean segment lengths for each species were measured to correct for body size and phylogenetic distance (Figure 2.8). Histoblast cell number and segment length do not appear to be correlated within species. Using an existing sepsid phylogeny created using multiple genetic markers (Figure 2.1) I created a phylogenetic tree of the study species and mapped trait values for the number of histoblast nest cells corrected for organism size effects using the segment lengths with the formula Trait value = (log(MaleCellCount)/(log(MaleSegmentLength)) - (log(FemaleCellCount)/(log(FemaleSegmentLength)) and mapped these values on to the phylogeny using the R package Phytools (Figure 2.9). The resulting phylogeny is color coded with values that represent the degree of sexual dimorphism in each species with cooler colors representing increased dimorphism in the histoblast nests. The primary gain species all have a high degree of histoblast nest dimorphism with *N. nitidula* showing the maximum value. The *Meroplius* and *Microsepsis* primary-gain species are not significantly sexually dimorphic in histoblast nest cell number. These three species have relatively small appendages with a sternite that is not dramatically modified compared to the *Themira* and *Nemopoda* species. This may indicate that sternite modification is an important factor in histoblast nest size. Secondary loss species all lose the sexual dimorphism (represented by the warm colors). The species that

recovered the lost appendage, *P. dikowi*, does not show sexual dimorphism of the histoblast nest. This species does have an enlarged 4th segment histoblast nest, but lack of sexual dimorphism appears to be a result of the females expressing the male phenotype.

Table 2.2. Male mean histoblast nest cell counts per segment

Species	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7
Arch	20.6	23.1	23.9	24	24.3	25.2	25.2
Asep	46.6	53.2	56.3	50.3	49.2	41.8	48.7
Dicra	42.2	47.4	47.1	39.8	34.8	36.3	36.7
Malb	35.1	42.3	43.8	41.6	45.6	42	41.5
Marm	26.4	39.3	33.6	31.2	28.7	28.3	26.7
Mfas	79.1	90	89.8	103.7	101.7	99	73.8
Nnit	78.3	83.7	82.3	83	77.5	76.2	57.5
Oluc	127.6	135.8	141.5	159.4	131.9	164.4	177.4
Pdik	66.9	77	87.6	94.8	84.6	80.1	78.1
Slat	27.8	33.6	29.8	29	32.5	29.2	33
Spun	65.1	78.7	69.8	68	69.8	67.3	60.3
Sful	28.6	34.2	34.5	37.4	30.6	28.4	28.6
Tbil	125.5	165.3	160.8	167.8	155.3	155	159.2
Tfla	46.9	54.5	60.6	56.4	54.7	55.5	47.8
Tluc	68	64.4	63.6	69.4	55.5	56.5	55.9
Tmin	67.5	90	88.3	80.5	72.3	72.9	61.4
Tput	109.6	101.4	107.7	95.3	112.9	90	88

Table 2.3. Female mean histoblast nest cell counts per segment

Species	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7
Arch	23.3	24.5	23.1	24.9	25	24.3	25
Asep	43	52.3	55.3	55.8	49.9	40.5	48.3
Dicra	47.3	52.1	51.1	48.4	40.8	36.9	45.4
Malb	37.8	43.1	42.7	55	45.1	44.6	38
Marm	31.3	43.4	40.8	37.8	32.3	30.4	30.4
Mfas	79.3	94.3	124.7	166.1	126.5	110.3	70.5
Nnit	98	97.1	132.8	293	166.8	119.9	80.9
Oluc	130.3	153.1	164.6	162.5	151.4	160.7	185.4
Pdik	65.3	83.5	104.1	101.7	84.2	80.9	75.5
Slat	30.3	34	34.3	37.1	41	38.6	33.9
Spun	61.4	73.4	66.3	68.8	64.3	63.2	62
Sful	22.6	24.9	22.7	28.5	26.9	22.6	23.1
Tbil	164.2	174.8	216.8	371.2	273.8	222.5	211.7
Tfla	58	64.1	83.9	126.8	83.9	69.5	57.1
Tluc	68.3	87.8	79.4	133.6	80	61.1	58.7
Tmin	66	96.9	127.2	211.4	107.4	80.2	72.4
Tput	94.3	103.1	114.9	205.5	132.6	114.5	107.9

Table 2.4. Likelihood ratio test of sex * segment effects

Species	Appendage history		LRT	P	
Orygma luctuosum	Ancestral	Sex	0.411	0.5214	
		Segment	34.462	5.48E-06	***
Meroplius albuquerque	Gain	Sex	0.139	0.7093	
		Segment	31.289	2.23E-05	***
Meroplius fasciculatus	Gain	Sex	3.833	0.2917	
		Segment	41.616	5.92E-04	***
Microsepsis armata	Gain	Sex	6.632	0.01002	*
		Segment	70.543	3.16E-13	***
Nemopoda nitidula	Gain	Sex	14.532	0.0001378	***
		Segment	67.856	1.13E-12	***
Themira biloba	Gain	Sex	30.188	3.92E-08	***
		Segment	52.502	1.48E-09	***
Themira flavicoxa	Gain	Sex	12.798	0.0003471	***
		Segment	112.557	2.20E-16	***
Themira lucida	Gain	Sex	3.394	0.06542	.
		Segment	38.729	8.09E-07	***
Themira minor	Gain	Sex	21.94	2.81E-06	***
		Segment	127.65	2.20E-16	***
Themira putris	Gain	Sex	1.451	0.2284	
		Segment	34.657	5.02E-06	***
Archisepsis armata	Loss	Sex	0.646	0.4216	
		Segment	7.6648	0.2637	
Allosepsis indica	Loss	Sex	0.042	0.837	
		Segment	62.153	1.64E-11	***
Dicranosepsis sp.	Loss	Sex	5.283	0.02153	*
		Segment	93.235	2.00E-16	***
Sepsis fulgens	Loss	Sex	6.624	0.01006	*
		Segment	37.204	1.61E-06	***
Sepsis latiforceps	Loss	Sex	4.5408	0.033097	*
		Segment	19.2743	0.003725	**
Sepsis punctum	Loss	Sex	0.4842	0.486514	
		Segment	22.2383	1.10E-03	**
Perokita dikowi	Regain	Sex	0.47	0.4931	
		Segment	59.238	6.43E-11	***

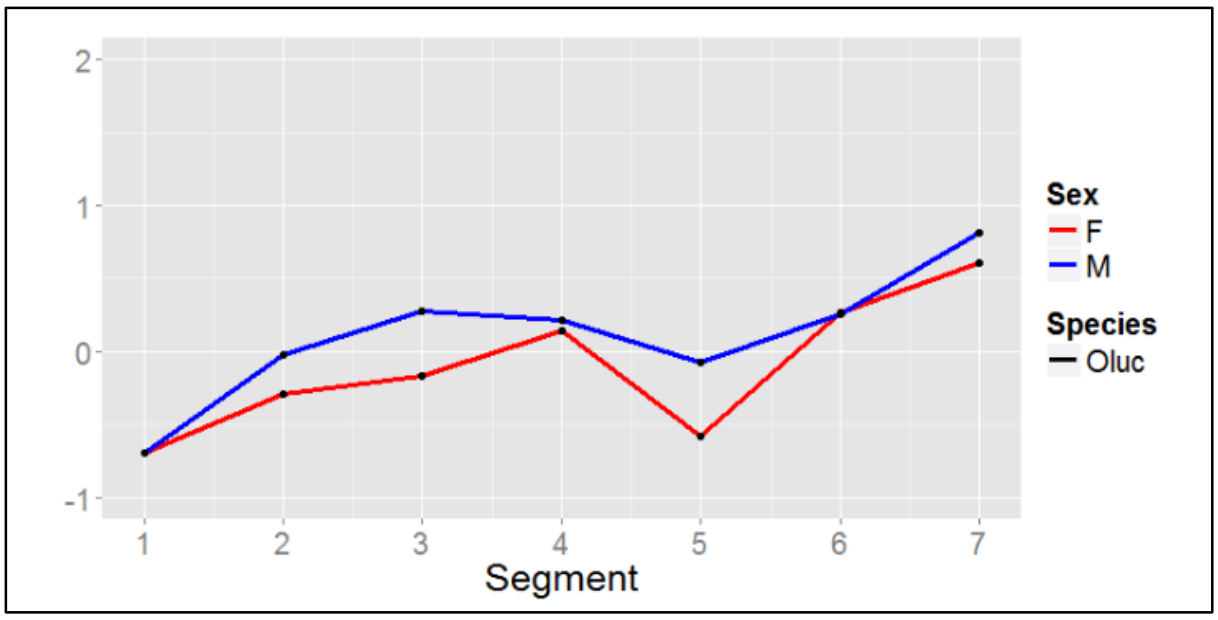

Figure 2.4. The ancestral state of the ventral histoblast nest in *O. luctuosum* shows no sexual dimorphism or increase in cell count or nest density in the 4th segment.

Table 2.5. Linear mixed effects model of sex * segment effects for *Orygma luctuosum*

Term	Estimate	SE	t-value	P	
Intercept	-0.69292	0.372636	-1.86	0.07942	.
SegmentSeg2	0.404726	0.358894	1.128	0.26362	
SegmentSeg3	0.525953	0.36589	1.437	0.15489	
SegmentSeg4	0.832755	0.363264	2.292	0.02494	*
SegmentSeg5	0.11203	0.381429	0.294	0.77026	
SegmentSeg6	0.963457	0.403585	2.387	0.02374	*
SegmentSeg7	1.303281	0.429114	3.037	0.00693	**
SexM	0.003823	0.519898	0.007	0.99421	
SexM:SegmentSeg2	0.259429	0.490591	0.529	0.59874	
SexM:SegmentSeg3	0.438815	0.500822	0.876	0.3838	
SexM:SegmentSeg4	0.076465	0.507309	0.151	0.88063	
SexM:SegmentSeg5	0.506379	0.531721	0.952	0.34566	
SexM:SegmentSeg6	-0.01822	0.569531	-0.032	0.97469	
SexM:SegmentSeg7	0.203899	0.595984	0.342	0.73597	

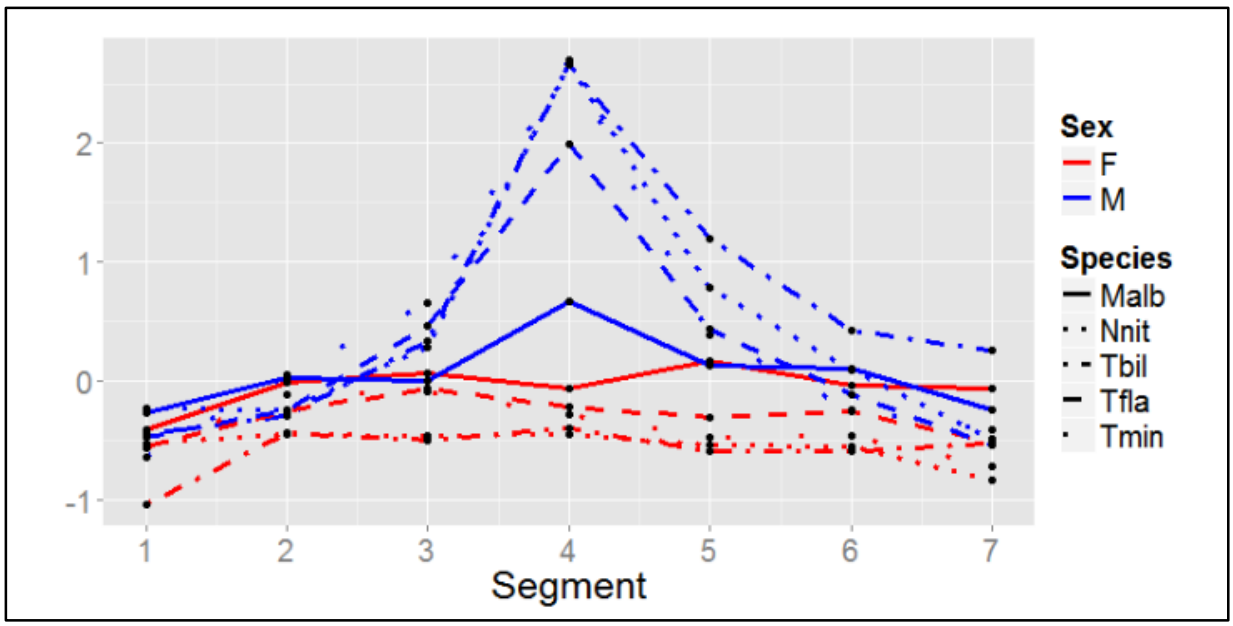

Figure 2.5. Primary gain of the appendage results in a strong sexual dimorphism and increase in cell count and nest density peaking in the 4th abdominal segment which produces the novel appendage.

Table 2.6. Linear mixed effects model of sex * segment effects for *Themira biloba*

Term	Estimate	SE	t-value	P	
Intercept	-1.0397	0.1719	-6.05	2.49E-07	***
SegmentSeg2	0.5996	0.2357	2.543	0.013565	*
SegmentSeg3	0.5319	0.2359	2.255	0.027757	*
SegmentSeg4	0.6372	0.2361	2.699	0.008968	**
SegmentSeg5	0.4491	0.2363	1.9	0.062135	.
SegmentSeg6	0.4441	0.2367	1.876	0.065523	.
SegmentSeg7	0.5068	0.2371	2.137	0.036916	*
SexM	0.582	0.2431	2.395	0.020788	*
SexM:SegmentSeg2	-0.439	0.3334	-1.317	0.192883	
SexM:SegmentSeg3	0.2609	0.3336	0.782	0.437145	
SexM:SegmentSeg4	2.4787	0.3338	7.425	4.22E-10	***
SexM:SegmentSeg5	1.2017	0.3342	3.595	0.000648	***
SexM:SegmentSeg6	0.434	0.3347	1.297	0.199747	
SexM:SegmentSeg7	0.2082	0.3354	0.621	0.537138	

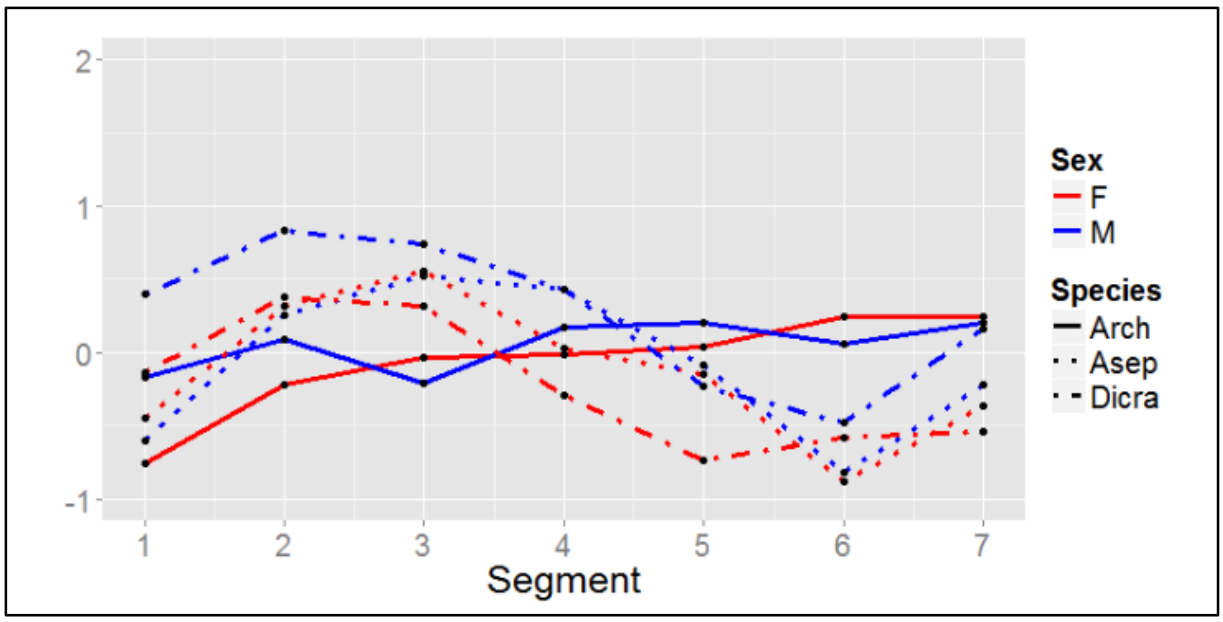

Figure 2.6. Secondary loss of the abdominal appendage results in a loss of sexual dimorphism and segment specific difference in histoblast nest size and cell number.

Table 2.7. Linear mixed effects model of sex * segment effects for *Archisepsis armata*

Term	Estimate	SE	t-value	P	
Intercept	-0.7545	0.3012	-2.505	0.0134	*
SegmentSeg2	0.5391	0.4255	1.267	0.2072	
SegmentSeg3	0.7188	0.4257	1.689	0.0934	.
SegmentSeg4	0.7388	0.426	1.734	0.0849	.
SegmentSeg5	0.7987	0.4264	1.873	0.063	.
SegmentSeg6	0.9983	0.4269	2.339	0.0208	*
SegmentSeg7	0.9983	0.4275	2.335	0.0212	*
SexM	0.5924	0.3966	1.494	0.1375	
SexM:SegmentSeg2	-0.2902	0.5602	-0.518	0.6053	
SexM:SegmentSeg3	-0.7627	0.5604	-1.361	0.1756	
SexM:SegmentSeg4	-0.402	0.5608	-0.717	0.4746	
SexM:SegmentSeg5	-0.4326	0.5613	-0.771	0.4421	
SexM:SegmentSeg6	-0.7787	0.562	-1.386	0.1681	
SexM:SegmentSeg7	-0.6323	0.5628	-1.123	0.2635	

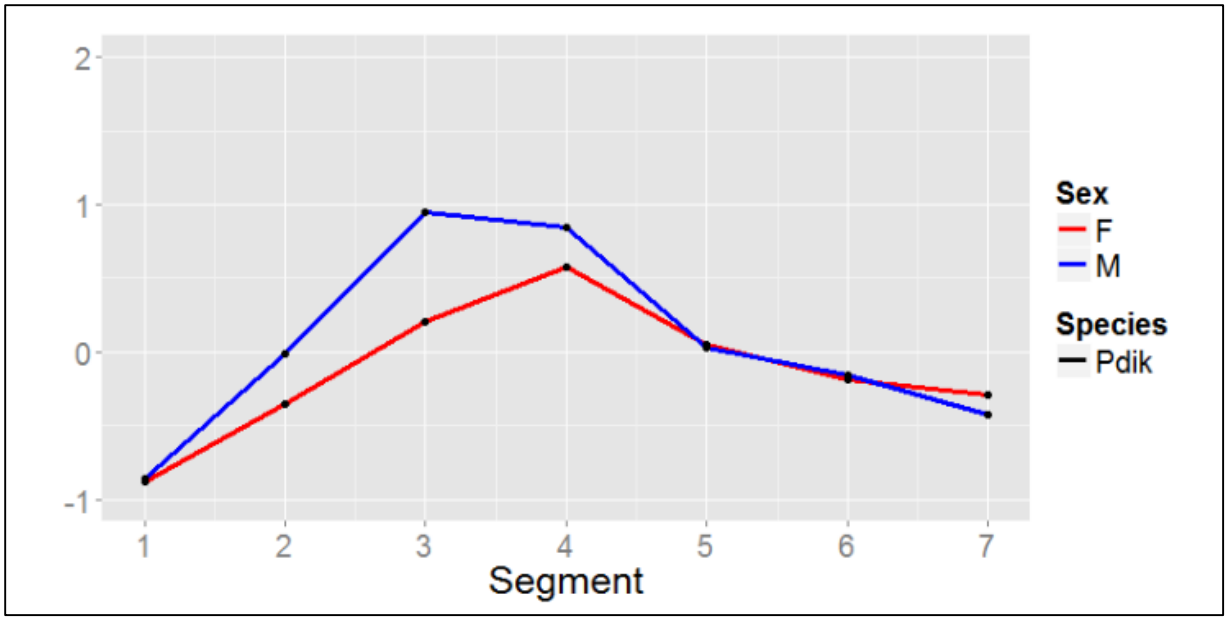

Figure 2.7. Tertiary recovery of the appendage results in recovery of increased histoblast nest size and cell density but not sexual dimorphism as female *P. dikowi* now show a male histoblast phenotype.

Table 2.8. Linear mixed effects model of sex * segment effects for *Perokita dikowi*

Term	Estimate	SE	t-value	P	
Intercept	-0.87654	0.300568	-2.916	0.00573	
SegmentSeg2	0.526654	0.347971	1.514	0.13314	
SegmentSeg3	1.079316	0.348047	3.101	0.00247	**
SegmentSeg4	1.449925	0.348174	4.164	6.40E-05	***
SegmentSeg5	0.92327	0.348351	2.65	0.00928	**
SegmentSeg6	0.689202	0.34858	1.977	0.05067	.
SegmentSeg7	0.585171	0.348858	1.677	0.09654	.
SexM	0.020806	0.405522	0.051	0.9593	
SexM:SegmentSeg2	0.315292	0.462719	0.681	0.49711	
SexM:SegmentSeg3	0.726351	0.470511	1.544	0.12562	
SexM:SegmentSeg4	0.249452	0.467262	0.534	0.59455	
SexM:SegmentSeg5	-0.04131	0.463181	-0.089	0.9291	
SexM:SegmentSeg6	0.006757	0.466304	0.014	0.98847	
SexM:SegmentSeg7	-0.15935	0.463797	-0.344	0.73186	

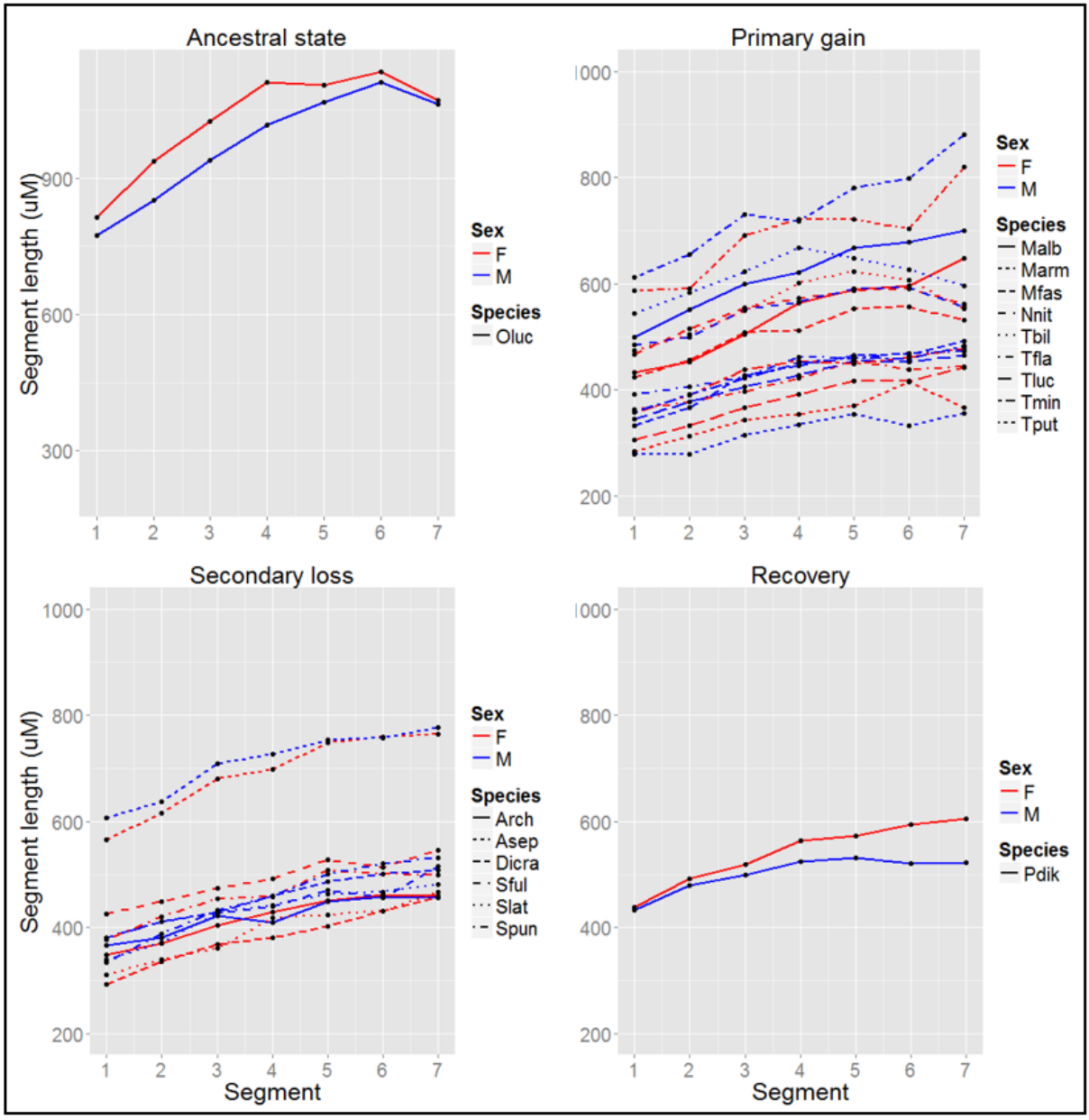

Figure 2.8. Mean length of abdominal segments 1-7 (uM) in each species.

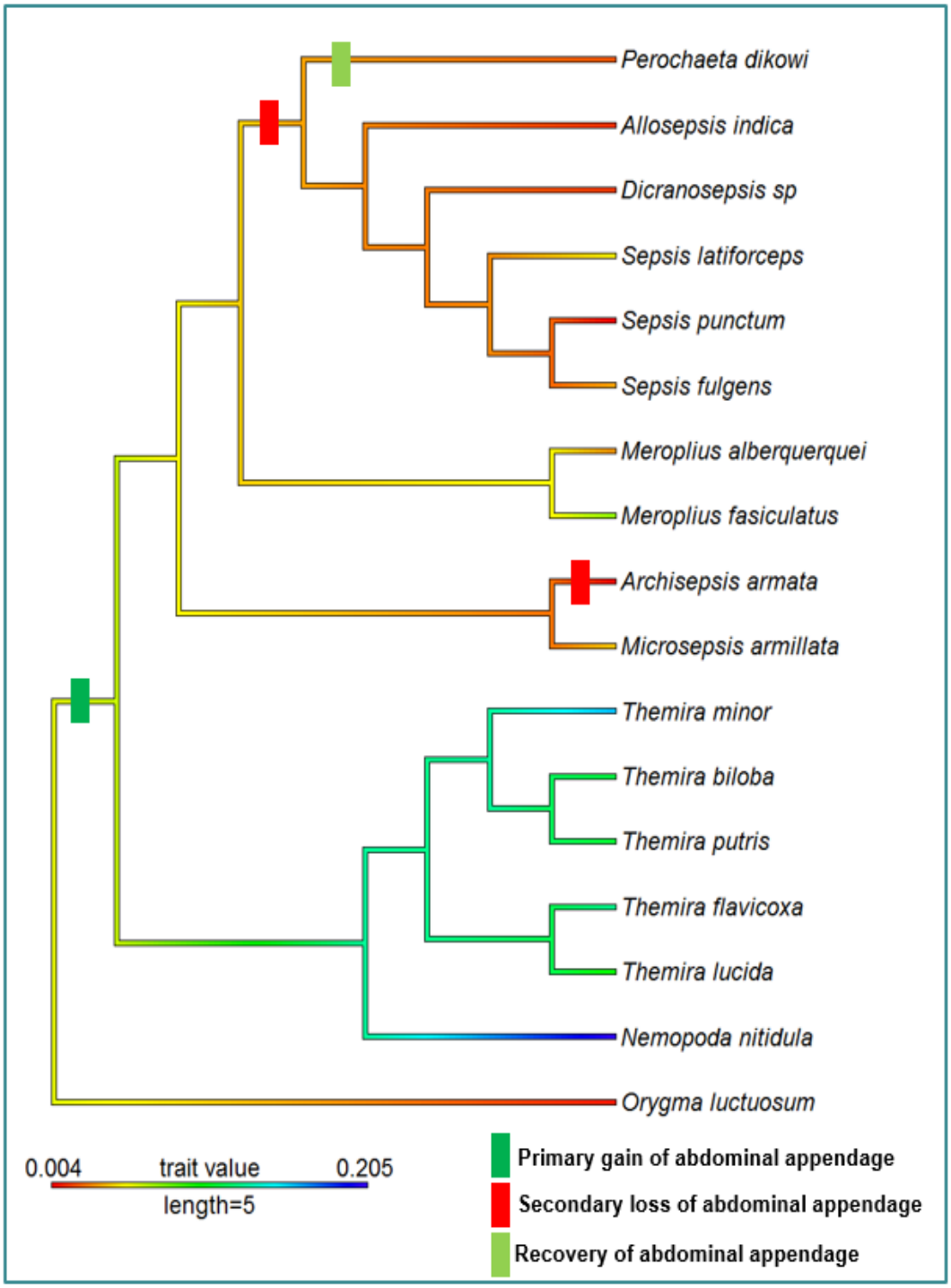

Figure 2.9. Phylogeny of sampled species with cooler colors indicating higher sexual dimorphism of histoblast nest cell counts corrected for organism size. Trait value = (log(MaleCellCount)/(log(MaleSegmentLength)) - (log(FemaleCellCount)/(log(FemaleSegmentLength))

Discussion

The primary objective of this study was to identify patterns in histoblast nest morphology between segments, sexes, and species that are constrained by the evolutionary history of the appendage. The pattern of gains and losses has previously been described in four species (Bowsher et al. 2013). These species represent histoblast nest morphology in three clades that identify distinct patterns in histoblast nest cell number. I identified and sampled species closely related to those previously investigated to determine if these patterns represent distinct character states or if patterns of histoblast nest morphology are highly variable across Sepsidae. I also sampled species from an additional independent loss and five additional genera to determine if other patterns of histoblast nest morphology exist in other clades. Including the larval segment length also allowed me to control for the effect of organism size. The ancestral state of the histoblast nest was also unknown and by including *O. luctuosum* in this data set I was able to confirm that species that have lost the appendage do return to the ancestral state. The increased size of the histoblast nest in male sepsids after primary gain of the appendage is consistent between species, although the increase in size and the degree of sexual dimorphism varies between species. Loss of the appendage results in the loss of sexual dimorphism, although some species have significant differences between segments, which is consistent between individuals and sexes but varies among species. The segment-specific pattern of histoblast nest morphology in the *Sepsis* clade is similar in these species. This group also contains *P. dikowi* which recovered the appendage. The recovery of the appendage results in increased histoblast nest size, but there is no sexual dimorphism because females have a pattern of enlarged nests that is not significantly different from males. This indicates that the ancestor of *P. dikowi* had a monomorphic nest morphology.

The loss and recovery of sepsid appendages is consistent with the hypothesis that lost traits "flicker" or appear and disappear in different clades during adaptive radiation which may cause misinterpretation of the ancestral relationships between sepsid species and inaccurate tree construction using morphological parsimony (Marshall et al. 1994). The phylogeny of Sepsidae is constructed using multiple genomic markers, but it is not known if the timescale of appendage gain and loss exceeds the timescale allowed for the recoverability of coding sequences under Marshall's model of Dollo's Law. The genetic pathway which produces the appendage may be under ongoing selection to remain intact, consistent with the morphological similarity of highly complex appendage sternite, bristle, and joint structure, as well as the associated musculature and behaviors between species which have primarily gained the appendage and *P. dikowi* which has recovered the appendage (Bowsher et al. 2013). If sepsid flies underwent rapid speciation due to sexual selection after the initial gain of the appendage it is possible that this "flickering" effect described by Marshall during radiation has produced a diversity of gains and losses which appear today as distinct and distant islands that were previously connected more recently by extinct species. The recovery of the lost sepsid abdominal appendage may also be an artifact of adaptive radiation which produced rapid gains and losses and a diversity of closely-related species that are now extinct. It is possible that the amount of evolutionary time that has passed between them is brief enough that the mechanisms that specify the appendage have not degraded.

Dollo's law describes many atavistic recoveries of long absent traits as misinterpretation or the product of co-option of genes. Several of the genes that pattern the sepsid abdominal appendage have broad developmental functions and are expressed in other tissues (Bowsher and Nijhout 2010). Although the appendage is lost, pleiotropy maintains the selection pressure and functionality of the rest of the appendage pathway. In *P. dikowi* the abdominal appendage is

morphologically consistent with other species. The loss of sexual dimorphism and the increased size of the female larval histoblast nest indicate that the loss of appendages in the Sepsis clade is the result of a change in specification. The investigation of gene expression during embryogenesis and a comparison of gene expression between species representing the gain, loss, and recovery of the appendage would provide insight in to how the specification pathway of the histoblast nest has been restructured through co-option.

Methods

Tissue collection

The epidermis was collected from dissected 3rd instar sepsid larva from live colonies maintained at North Dakota State University and at the National University of Singapore (NUS) by a collaborator Rudolf Meier. Prior to tissue collection, eggs were collected on the preferred dung substrate of each species to increase the population and ensure consistent age and nutrition at the time of collection. Flies raised at NDSU received a soy infant formula supplement solidified in agar under 1 cm of cow dung in a 15 cm petri dish. Dung was collected from a local organic cattle herd. Flies raised at NUS received the dung substrate preferred by that particular genus or species. Duck dung was collected from the local poultry industry, and horse or cow dung was collected from the Singapore Zoo.

Third instar wandering-phase larvae were identified by their pale yellow coloration, which is easily identifiable after purging dung from their digestive tract, signaling a commitment to pupation. Sex of the larvae was determined by the presence or absence of testes, which appear as two large, clear ovoid masses between the 4th and 6th abdominal segments. Males and females were placed in separate collection dishes. Larvae were sacrificed by immersion into water heated to 55°C. This method causes muscle relaxation, which aids in dissection, and a consistent

appearance of the larval epidermis. The larvae were dissected in an isotonic phosphate-buffered saline (PBS) solution (See Appendix A). Lateral cuts at the location of the anterior and posterior tracheal cross-branches remove the head and open the larval posterior while leaving abdominal segments 1-8 intact. A longitudinal incision was then made along the dorsal line between the two main tracheal trunks. The tracheae, organ systems, and fat body were removed leaving a flat fillet of epidermal tissue.

Tissue fixation and staining

Fixation and staining protocol was been adapted from Madhaven and Madhaven (see Appendix A). After dissection tissues were rinsed in dissection buffer to remove remaining fat cells and other detritus and then transferred to a micro-centrifuge tube containing Kahle's fixative (12% formalin, 32% absolute EtOH, 2% glacial acetic acid, 60% ddH_2O) for 18-24 hours at room temperature. Tissue was dehydrated in increasing concentrations of ethanol and may be stored in 100% ethanol until for up to several days at room temperature. Prior to staining tissue was rehydrated in decreasing concentrations of ethanol and finally in ddH_2O.

Staining took place in a cell culture dish. First 2mL of 6N HCl was added for hydrolysis with 10 minute incubation. HCl was removed and the tissue and well was rinsed several times to remove all traces of the acid. 2mL of Schiff's reagent (see Appendix A) was added and the cell culture dish was placed in a light-proof box for 90 minutes at room temperature. After the 90 minute incubation, the Schiff's reagent was removed and the tissue was rinsed with distilled water. Exposure to light causes the nuclear stain to develop. After the desired level of stain was achieved the tissue was rinsed with distilled water several times to prevent further development. Every 30 minutes the distilled water was changed until it no longer colored pink after 30 minutes. The tissue was stored in 100% ethanol in micro-centrifuge tubes prior to mounting.

Tissue was transferred to a glass dish containing 50% ethanol and Histoclear or HemoDe equilibrating agent for 5 minutes after which it was replaced by 100% equilibrating agent. The equilibrating agent prevents the mounting medium from hardening quickly. Excess equilibrating agent was removed by immersing each larval epidermis in mounting medium (PerMount) immediately prior to mounting. The flattened tissue was placed inside a drop of mounting medium and held in place by a weighted or clamped coverslip for 2-5 days while the mounting medium hardens. With thicker tissue, it was often necessary to add mounting medium around the outside of the coverslip to fill in and prevent desiccation or the formation of bubbles.

Histoblast imaging and characterization

Histoblast image acquisition was performed with an inverted Zeiss LSM microscope and Zeiss ZEN digital imaging software to generate z-stacked images of the histoblast nests. The Schiff's-stained histoblast nests fluoresce distinctly with little or no background fluorescence. Pixel dwell time was set to 1.11 seconds and image averaging was set to 2. These settings remained consistent across all samples and produced consistent images, although the z-stack slice depth and window-size reflected the size of the nest between species. This information was preserved and accounted for during image data processing as part of the image file.

Histoblast nest nuclei were counted using the FIJI Image-J (Abràmoff et al. 2004) package using a custom macro. The LSM z-stack images were flattened to produce a single TIFF image file which preserves scaling information. De-speckling, Gaussian, and median filters were applied to remove noise and the contrast against background was enhanced to resolve individual cell nuclei. An automatic local threshold was applied to convert the image to binary white nuclei on black background. Then a watershed tool was used to separate nuclei, which may overlap slightly as a product of collapsing the z-stack image. The histoblast area of interest

was then selected and an automated cell counter was used to collect the number of cell nuclei and the total, average, and individual area of the nuclei (Figure 2.10). The total area measurement is a sum of individual nuclei and does not represent the total size of each nest. FIJI was also used to measure total nest area by connecting the outermost nuclei at a point that includes all individual areas and measuring the internal space. Segment length information was recorded using a calibrated measurement tool in AxioVision per segment per sample per species.

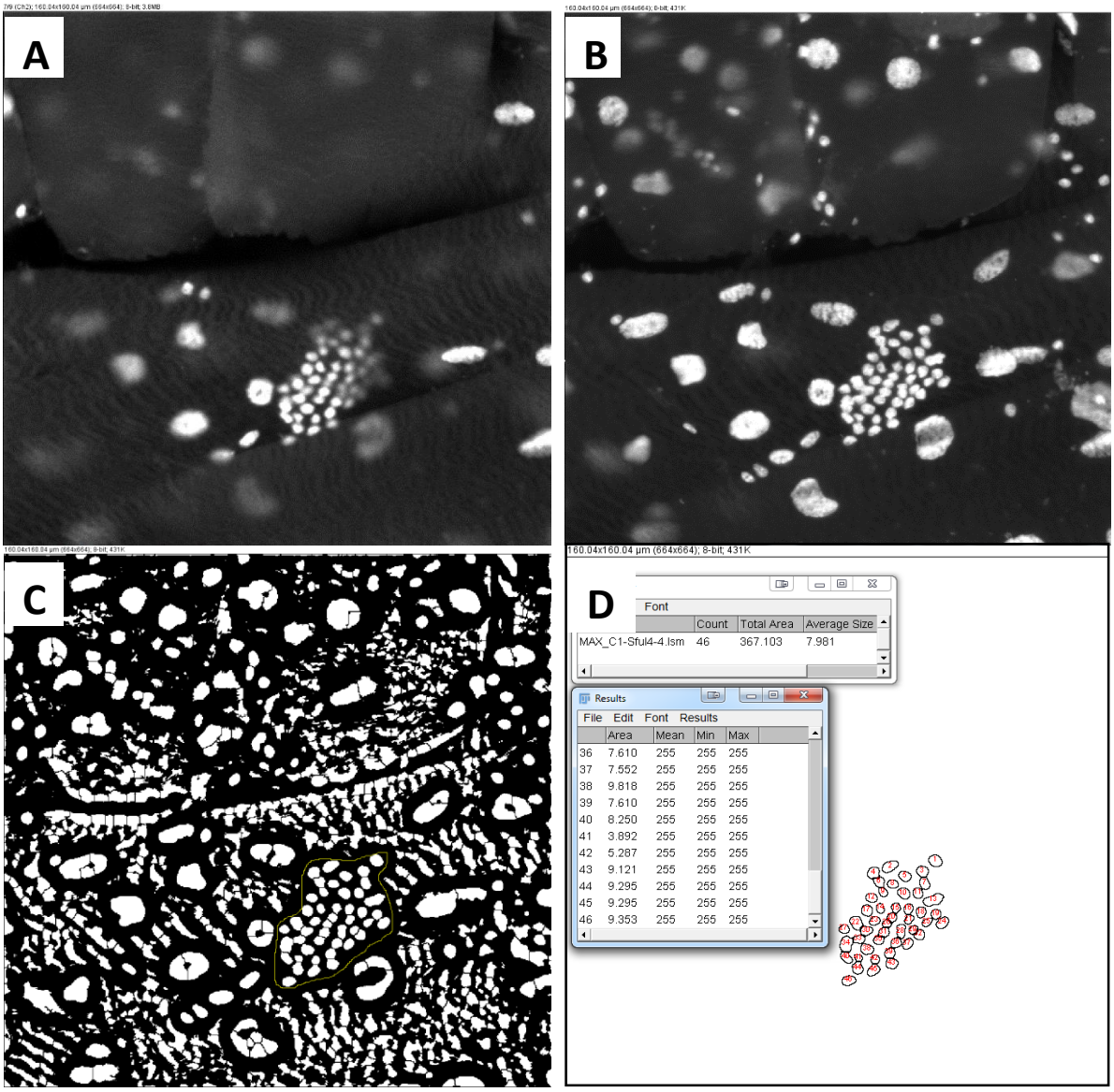

Figure 2.10. Image processing and cell counting protocol using FIJI. Raw LSM files (A) are collapsed to a representative single image which is contrast and gamma corrected, passed through filters to remove noise and resolve nuclei (B). A local threshold is applied to convert the image to binary (C). FIJI is used to count and measure cell nuclei (D).

Statistical analysis

Linear mixed-effect model (LMM) analysis was performed using the R package Lme4 (Bates et al. 2015 p. 4). Sex was used as a categorical fixed effect and segment identity was used as a continuous fixed effect. Sampled individuals were used as random effects. Cell number was scaled to standard deviation units. The LMM also generated an estimated mean for each segment, which was plotted for each sex with 95% confidence intervals. A likelihood ratio test (LRT) was performed for each species using chi-squared values with the R package lmerTest. This was also used to determine significant differences between sexes and segments. See Appendix B for sample statistical analysis using R.

A phylogeny of sampled species was manually constructed using Newick phylogenetic tree format by assigning branch points using an existing phylogenetic tree that was constructed using multiple genetic markers. The R package Phytools was used to map trait values on to the phylogeny of sampled species (Revell 2012). Phytools takes trait values for each species and assigns color-coded values that represent similarity between species and clades. Branch lengths are unknown between these species because the time of divergence has not been identified. The effect of organism size on trait value are accounted for before mapping using the equation Trait value = (log(MaleCellCount)/(log(MaleSegmentLength)) - (log(FemaleCellCount)/(log(FemaleSegmentLength)).

CHAPTER THREE: *DE NOVO* SEPSID TRANSCRIPTOME ASSEMBLY[1]

Abstract

The Sepsidae family of flies is a model for investigating how sexual selection shapes courtship and sexual dimorphism in a comparative framework. However, like many non-model systems, there are few molecular resources available. Large-scale sequencing and assembly have not been performed in any sepsid, and the lack of a closely related genome makes investigation of gene expression challenging. Our goal was to develop an automated pipeline for *de novo* transcriptome assembly, and to use that pipeline to assemble and analyze the transcriptome of the sepsid *Themira biloba*.

Our bioinformatics pipeline uses cloud computing services to assemble and analyze the transcriptome with off-site data management, processing, and backup. It uses a multiple k-mer length approach combined with a second meta-assembly to extend transcripts and recover more bases of transcript sequences than standard single k-mer assembly. We used 454 sequencing to generate 1.48 million reads from cDNA generated from embryo, larva, and pupae of *T. biloba* and assembled a transcriptome consisting of 24,495 contigs. Annotation identified 16,705 transcripts, including those involved in embryogenesis and limb patterning. We assembled transcriptomes from an additional three non-model organisms to demonstrate that our pipeline assembled a higher-quality transcriptome than single k-mer approaches across multiple species.

[1] This chapter was co-authored by Dacotah Melicher, Alex S. Torson, Ian Dworkin, Julia H Bowsher and published in BMC Genomics in 12 March 2014 under the title "A pipeline for the de novo assembly of the Themira biloba (Sepsidae: Diptera) transcriptome using a multiple k-mer length approach". Dacotah Melicher had primary responsibility for all of the methods described in this chapter including, collection and processing tissue and construction of the bioinformatic pipeline. Dacotah Melicher drafted and revised all versions of this chapter. Julia H Bowsher proofread the text written by Dacotah Melicher.

The pipeline we have developed for assembly and analysis increases contig length, recovers unique transcripts, and assembles more base pairs than other methods through the use of a meta-assembly. The *T. biloba* transcriptome is a critical resource for performing large-scale RNA-Seq investigations of gene expression patterns, and is the first transcriptome sequenced in this Dipteran family.

Introduction

The Sepsidae family of flies consists of over 200 species with a global distribution (Pont 2002). Sepsids are a model system for the investigation of sexual selection and how it affects courtship and sexual dimorphism(Bowsher et al. 2013). Sepsids have complex courtship behaviors that include elements of male display, female choice, and sexual conflict(Martin and Hosken 2003, Baena and Eberhard 2007, Ingram et al. 2008, Puniamoorthy et al. 2009). Specialized male traits have evolved alongside these complex courtship behaviors. Sexual selection has resulted in the evolution of modified forelimbs, body size, and abdominal appendage-like structures, which are articulated and have long bristles attached to their distal ends(Eberhard 2001a, 2001b, 2005, 2012, Blanckenhorn et al. 2004, Bowsher and Nijhout 2007, 2009, Puniamoorthy et al. 2008, 2012). Next-generation sequencing in combination with gene expression analysis has the potential to answer multiple questions including: how new morphologies evolve, whether shared developmental mechanisms underlie traits that have evolved multiple times, what the genetic basis of sexual dimorphism is and how to resolve the phylogenetic relationships within Sepsidae. Despite the potential of sepsids as a model to test a wide variety of evolutionary hypotheses, almost no molecular resources exist in this family, nor are any genomes or EST databases available.

Most Dipteran families have few genomic resources compared to drosophilids and mosquitoes. Sepsids shared a common ancestor with *Drosophila* and houseflies between 74 and 98 MYA, and are not closely related to any taxon with significant genomic resources(Wiegmann et al. 2003, 2011). A detailed investigation of the *even-skipped* locus revealed that approximately twice as many nucleotide substitutions exist between coding regions of *D. melanogaster* and sepsid species as exists between *D. melanogaster* and the most distantly related *Drosophila* species(Hare et al. 2008). The Sepsidae are a sister taxon to the Tephritoidea or true "fruit flies," which contains four species with genomic and transcriptomic resources (Schwarz et al. 2009, Zheng et al. 2012, Hsu et al. 2012, Nirmala et al. 2013), but these are not as well annotated as *Drosophila* and the level of sequence similarity with sepsids is unknown. A sepsid transcriptome would not only facilitate gene expression studies across the Sepsidae, but would also enhance comparative bioinformatics within Diptera.

For non-model organisms, the challenge of gene discovery no longer resides in a dearth of sequence data, but from the computational challenges of large and complex datasets(Sboner et al. 2011). This challenge is particularly true for *de novo* assembly, which is more computationally intensive than syntenic assembly via mapping to a reference genome. Another hurdle to *de novo* assembly is recovering rare transcripts from a datasets with heterogeneous sequence coverage. Assemblies that combine multiple k-mer lengths generally recover a greater number of unique transcripts during *de novo* assembly than single k-mer approaches(Surget-Groba and Montoya-Burgos 2010, Gruenheit et al. 2012), but with additional potential for mis-assembly. Although both cloud computing and multiple k-mer approaches are widely available, they have not been employed as broadly as reference-based pipelines because some programing knowledge is required.

Our objectives were two-fold: 1) to construct a general purpose *de novo* transcriptome assembly pipeline that compares the output of multiple programs and automatically analyzes this data for downstream applications, and 2) to use that pipeline to assemble the transcriptome of the sepsid *T. biloba*. Our pipeline uses Velvet-Oases and Trinity for the initial assembly and constructs a meta-assembly with CAP3 followed by analysis with various downstream programs, including BLAST and Blast2GO("Oases: a transcriptome assembler for very short reads" n.d., "Velvet: a sequence assembler for very short reads" n.d., Huang and Madan 1999, Conesa et al. 2005). The pipeline functions on a low-cost cloud computing network, and can be operated from a standard desktop computer. In addition to assembling the *de novo* transcriptome of the sepsid fly *T. biloba*, we used this pipeline to re-assemble previously published transcriptomes that used both 454 and Illumina sequencing platforms. Compared to the standard single k-mer assembly, our pipeline assembles longer contigs and more base pairs in all four species. By comparing annotated transcripts from different assemblies of the *Themira biloba* transcriptome, we demonstrate that our pipeline recovers a greater number of transcripts than standard approaches by pooling unique transcripts from multiple assemblies.

Results

General overview of computational pipeline

This pipeline was designed to automate a large number of intermediate bioinformatic activities such as trimming and filtering reads, converting sequence files through various formats, performing a large number of sequential assemblies using different assemblers and parameters, and formatting the output for downstream use (Figure 3-1). This pipeline was also designed to circumvent what have traditionally been significant limitations for small research groups, a lack of computing facilities and programing knowledge. This pipeline, while

functional on a local network, is designed to make use of virtual cloud computing units, which provide scalable resources with direct interaction. Our pipeline produces intermediate products that are compatible with graphical user interface (GUI) based platforms such as The iPlant Collaborative and Galaxy, so that researchers can use these interfaces for downstream applications if desired(Giardine et al. 2005, Goecks et al. 2010, Blankenberg et al. 2010, Goff et al. 2011).

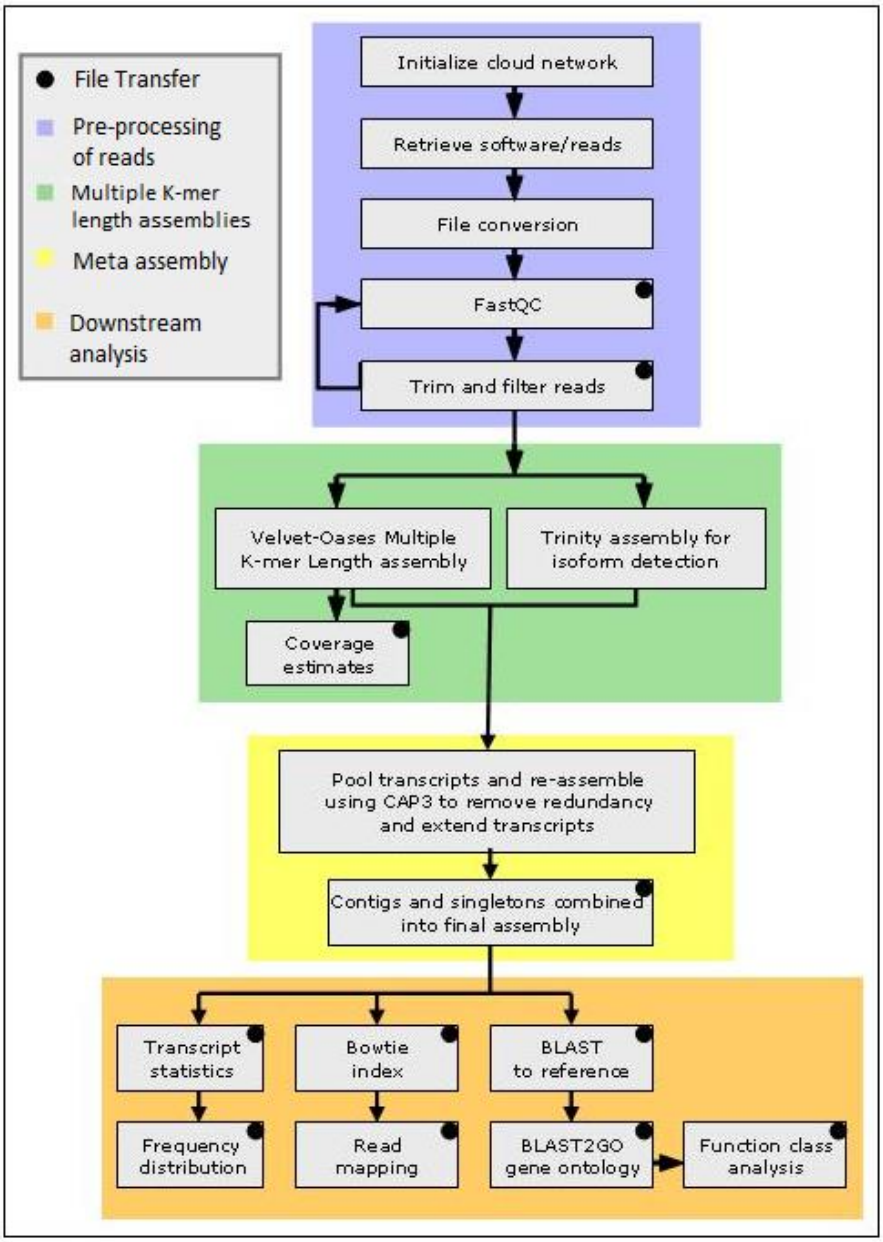

Figure 3.1. Flowchart of the bioinformatic pipeline. The pipeline performs multiple operations from sequence editing to annotation. First, a cloud network is initialized and algorithms are retrieved and installed. The sequence reads are parsed and filtered for quality and removal of adaptor sequences (blue). Next, assemblies are generated using various k-mer lengths and algorithms to create a diversity of transcript fragments (green). Then, the transcripts from all assemblies are pooled and re-assembled to remove redundant contigs and extend sequences based on overlap (yellow). The resulting multiple k-mer length meta-assembly is then analyzed and formatted for various downstream applications. Reads are mapped back to contigs, genes are annotated, and gene ontology is applied using BLAST and Blast2GO (orange). The pipeline generates an analysis of the assembly and the quantity and distribution of sequences. The resulting data is packaged in an archive for transfer and the cloud network is disbanded.

We used this pipeline to perform the *de novo* assembly of the *T. biloba* transcriptome, the first transcriptome assembly for any species for the family Sepsidae. We also used the pipeline to re-assemble archived RNA-seq reads from other studies to assess the performance of the multiple k-mer length assembly process compared to a single k-mer assembly. Archived sequence from an arthropod (the milkweed bug, *Oncopeltus fasciatus*: [SRR:057573]), a plant (*Silene vulgaris*: [SRR:245489]), and a mammal (the ground squirrel *Ictidomys tridecemlineatus*: [SRR:352220]) were selected to test the performance of the pipeline across taxa and genome sizes. Each of these data sets consists of 454 sequence reads of approximately 3.2-4x coverage, the same coverage as our *T. biloba* data set. The *O. fasciatus* and *S. vulgaris* sequence reads were generated for *de novo* assembly of the entire transcriptome of the organism while the *I. tridecemlineatus* sequences were generated for differential expression analysis (Ewen-Campen et al. 2011, Hampton et al. 2011, Sloan et al. 2012).

Cloud computing network and data management

All of the data presented here were generated using Amazon Web Services Elastic Cloud Compute (AWS EC2) using a Debian Linux operating system (version 6.0.3). Software, sequence reads, reference assemblies, and other files are stored persistently on AWS Elastic Block Storage (EBS) volumes for the purpose of off-site backup, reduced network traffic, and storage. Data produced by the pipeline may be parsed and manipulated further through AWS or downloaded locally as needed. As presented here, the pipeline runs software in series. However, it is simple to create many duplicate systems through AWS, which may then run the processes in parallel.

Cloud computing instances were initialized using memory-optimized architecture to memory requirements the high memory requirements of Velvet-Oases assembly of 454 sequence

reads. An instance with 64 gigabytes (GB) of available memory was used to during initial analysis of assembly performance at different k-mer lengths. This was sufficient to produce assemblies with a k-mer length up to 31bp after which available memory became a limiting factor which coincided with a reduction in assembly quality. At the time of this writing high-memory instance types with up to 244GB of available memory are available for larger data sets. Instances were initialized using a publically available Linux operating system disc image hosted by Amazon. Software, data, and scripts are stored on EBS volumes and software installation is simplified by a script that unpacks and installs all of the packages required for this pipeline to a newly created 'bare' cloud instance. All functional aspects of the pipeline shown in Figure 3.1 are performed by a wrapper script which sequentially performs the assembly and analysis of sequence data before storing it remotely and terminating the instance to minimize computing cost which is calculated in hourly blocks based on instance type. The pipeline ran to completion in approximately 20 hours. Larger sequence data sets requiring more memory and computing time may benefit from separating memory-intensive assembly from processor-intensive downstream analysis as the cost of processing with cloud computing is much lower than reserving large blocks of memory and storage space.

Trimming and quality filtering reads

Prior to assembly, the reads are processed to remove adaptor sequences, low-quality reads and regions, and highly redundant sequences. The initial quality of the untrimmed sequence reads is assessed using FastQC, which also generates a list of over-represented sequences which may then be removed("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data" n.d.). The raw sequence reads are then converted to a standard format which is passed on to the FastX Toolkit which removes adaptor sequences using

trimming and clipping functions("FASTX-Toolkit" n.d.). The reads are subsequently run through the FastX quality filter which removes reads that fail to pass a quality check (80% of the bases having a phred score of 20 or higher, corresponding to a 1:100 base-calling error rate were used for the data presented here). The remaining reads are analyzed for redundancy by FastX and then collapsed into a single representative read. This removes large numbers of identical reads that may result from the amplification process prior to sequencing. Reducing the number of reads can dramatically reduce the amount of memory needed during the assembly process. It can also significantly reduce the amount of time required for assembly, which is an important consideration when generating multiple assemblies(Cahais et al. 2012).

Assembly

It has been shown that performance varies significantly between assemblers and data sets (Kumar and Blaxter 2010). This has prompted the development of a number of techniques, such as multiple-k approaches, to retrieve more contigs from the initial sequence reads (Martin et al. 2010, Gruenheit et al. 2012, Hornett and Wheat 2012, Mundry et al. 2012, Vijay et al. 2013).

To assemble the *T. biloba* sequence reads we have used a multiple k-mer length approach that creates a large number of assemblies, each of which contains potentially unique transcripts. Because many assembly programs can support multiple k-mer assembly after the addition of custom scripts, we compared the performance of four different assembly programs: Abyss, Newbler, Trinity and Velvet-Oasis, using a previously described protocol (Supplemental Table 1)("ABySS 1.3.5 — Canada's Michael Smith Genome Sciences Centre" n.d., "Oases: a transcriptome assembler for very short reads" n.d., "Velvet: a sequence assembler for very short reads" n.d., Kumar and Blaxter 2010, Grabherr et al. 2011, Henschel et al. 2012, O'Neil and Emrich 2013). *T. biloba* sequence reads from multiple life stages were pooled and assembled

with a k-mer length of 25 using each of the four assembly programs (Table 3.1). The resulting transcripts were then aligned to the *Drosophila melanogaster* transcriptome. A conservative cut-off value with a minimum aligned length of 400bp was used to create the distribution in Table 3.1. While Velvet-Oases produced the longest contigs, Trinity generated a larger number of contigs. A nucleotide BLAST of contigs in each assembly showed an increase in the number of contigs unique to one assembly in those produced by Trinity and Velvet-Oases. Based on these results, Velvet-Oases was selected for the length of the resulting transcripts and the ease of generating assemblies of different k-mer lengths, and a single Trinity assembly is included to provide isoform detection. The Velvet-Oases and Trinity *de novo* assembler algorithms have complementary strengths and weaknesses when comparing memory requirements and run-time.

Table 3.1. Comparison of assemblers and identification of unique transcripts

Assembler	Contigs	Contig n50	BLAST hits	Unique hits
Velvet-Oases	18960	296	5114	1817
Abyss	19664	127	5341	1566
Newbler	13398	208	4302	1509
Trinity	25144	244	6826	2194

The *T. biloba* sequence data was used to generate assemblies with k-mer lengths of 17, 19, 21, 23, 25, 27, 29, and 31 base pairs. To demonstrate that assemblies with different k-mer lengths recover unique transcripts, the stand-alone BLAST algorithm was used to align contigs from each assembly to a pool of contigs from all assemblies, with the resulting unaligned contigs representing those unique to one assembly (Figure 3.2). For example, to determine the number of contigs unique to the K17 assembly, the K17 contigs were blasted against the pooled contigs from all other assemblies. If a contig did not align, then it was unique to the k17 assembly. Contigs were discarded that were less than 200 base pairs. Next, BLAST was performed against *D. melanogaster* to annotate the unique contigs, and only those contigs with orthology to *D.*

melanogaster were reported (Table 3.2). After the initial analysis, the pooled assemblies were also annotated using the *D. melanoga*ster transcriptome to generate a total number of transcripts for the pool, to which the number of unique transcripts could be compared (Table 3.2). A significant number of transcripts were represented in only one of the single k-mer length assemblies (Table 3.2). In total, 2,296 transcripts were identified as unique to a specific assembly using BLAST analysis. For k-mer lengths 17-27, unique transcripts were approximately 2% of each assembly, and this percentage did not decrease with increasing k-mer length. However, at K29, unique transcripts decreased to only 0.8% of the total. The number of unique transcripts generated from this analysis is a low estimate because it contains only conserved *Drosophila* orthologs, and excludes transcripts unique to *T. biloba* and those too divergent to be identified by BLAST. Therefore, the number of unique transcripts recovered from different k-mer assemblies is likely higher. Our analysis confirms that restricting assemblies to only a single k-mer length limits the number of transcripts recovered, regardless of which k-mer length is chosen.

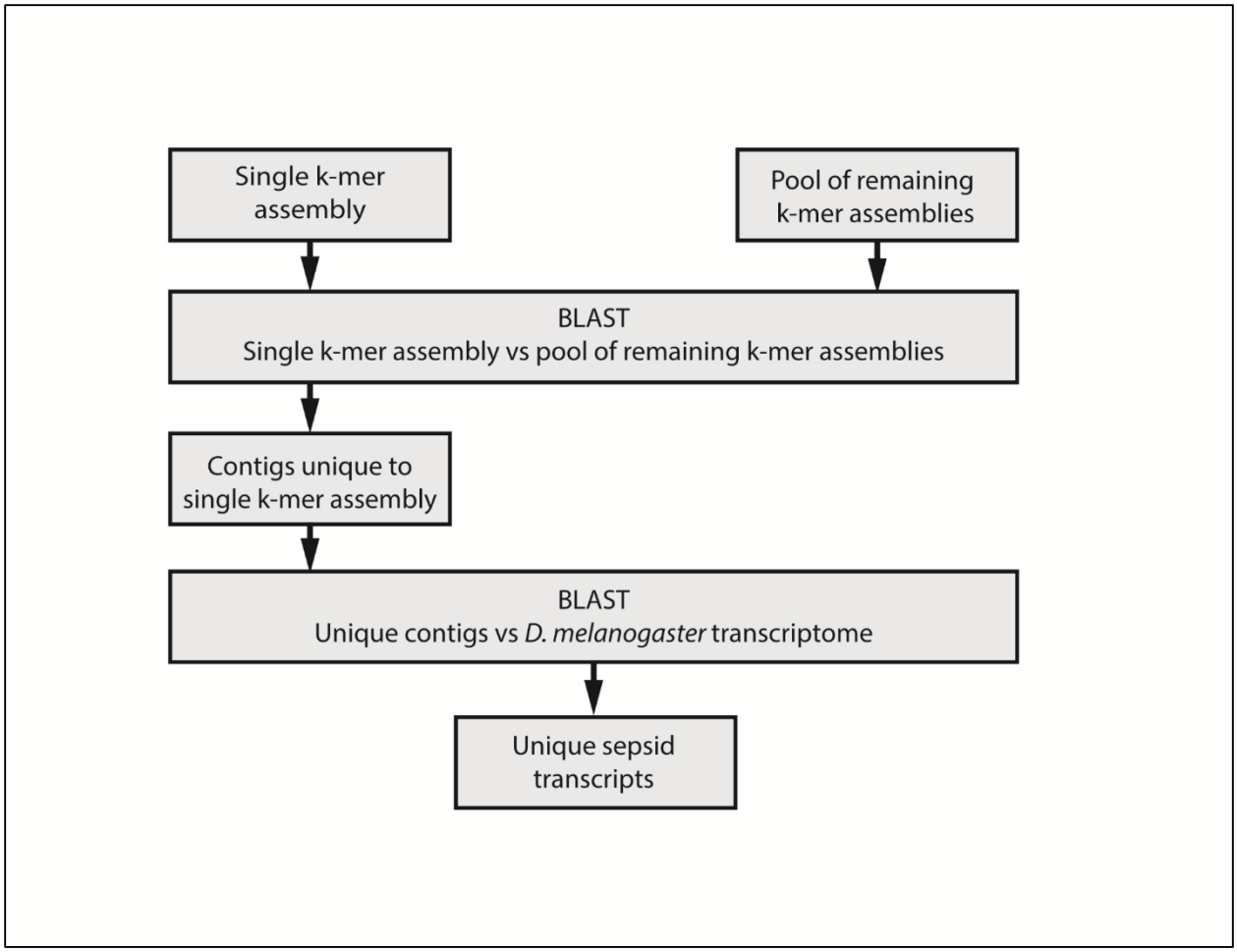

Figure 3.2. BLAST strategy to identify unique transcripts. Identification of unique transcripts in each individual assembly was performed by reserving contigs from one assembly and pooling all contigs from the remaining assemblies. The contigs from the single assembly were aligned to the pooled contigs. Contigs that fail to align were considered unique to that single assembly. The unique contigs were annotated by aligning to the *D. melanogaster* transcriptome.

Table 3.2. Unique transcripts per k-mer length in paired-end assemblies using Velvet-Oases

K-mer length	Total transcripts	Transcripts absent from one or more assemblies	Transcripts unique to one assembly	% unique transcripts
17	21296	2331	464	2.2
19	20080	2105	397	2.0
21	17950	1875	410	2.3
23	16668	1686	316	1.9
25	15894	1434	280	1.8
27	15496	2398	313	2.0
29	15138	3855	116	0.8
Total	122522	15684	2296	1.9

Meta-assembly

The assemblies generated with k-mer lengths of 23, 25, 27, and 29 base pairs were combined through meta-assembly which extends contigs found in multiple assemblies and retaining contigs found in only one. K-mer lengths shorter than 23 resulted in a large number of singletons and short contigs. Assemblies with a k-mer length larger than 29 required much larger memory allocations and computational time and were more conservative than other assemblies resulting in diminishing returns in which larger k-mer word sizes produce few novel transcripts not present in other assemblies.

The CAP3 software was used to construct the meta-assembly(Huang and Madan 1999). The CAP3 software removes the redundancy generated within and between assemblies of different k-mer lengths to consolidate the transcripts. Consolidating the results of all k-mer assemblies created a pool of 138,954 contigs. CAP3 clustered and assembled these sequences into a meta-assembly of 15,984 extended contigs and 8,511 singletons. The singletons represent sequences for which no overlap exists between assemblies and thus could not be extended by

CAP3. The final meta-assembly consisted of 24,495 contigs with a mean sequence length 1,403 base pairs, an increase of 372bp (34.1%) compared to the K25 assembly.

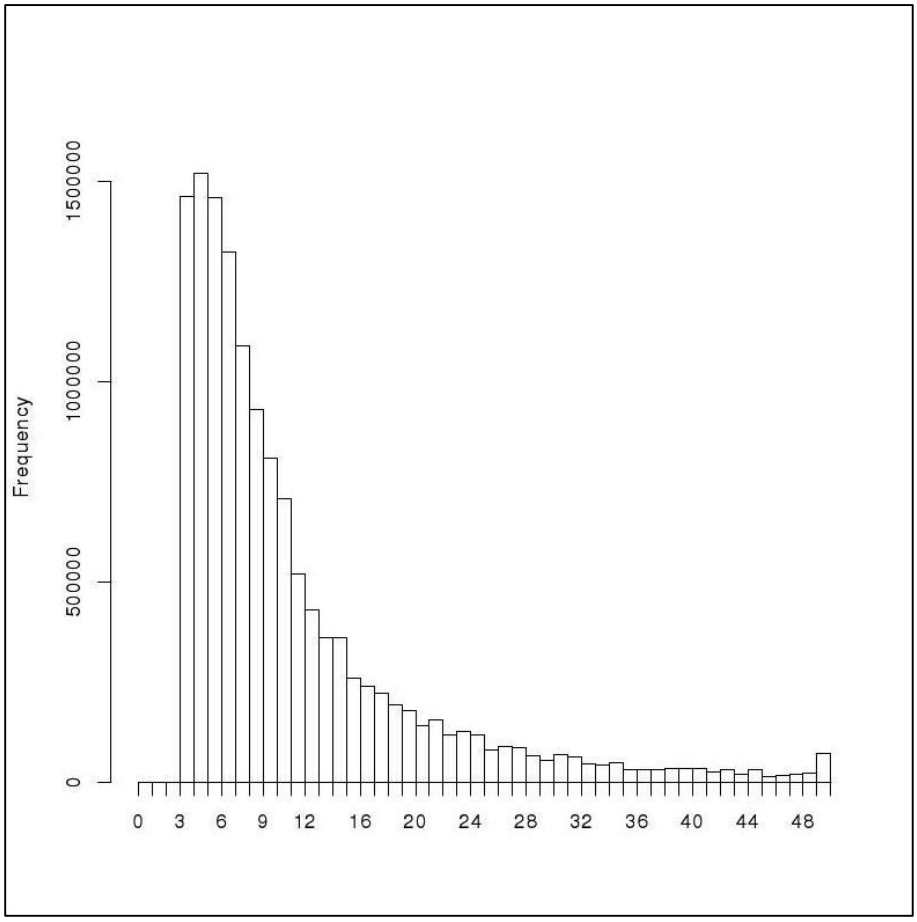

Figure 3.3. Average distribution of coverage of *T. biloba* contigs. Coverage estimates were generated using the Velvet software. Frequency indicates the number of times a k-mer is represented in the unassembled sequence reads.

Analysis of transcript length revealed that the total number of base pairs assembled improved significantly from 17.4Mb to 32.7Mb and the mean contig length increased by 310bp from 1,093bp to 1,403bp. A frequency distribution of the number of contigs of a given length (Figure 3.4) shows an increase in the number of longer contigs in the meta-assembly, compared to the single k-mer assemblies and the Trinity assembly. The single k-mer assemblies have a relatively high number of singletons (sequences of less than 500bp). The number of singletons was greatly reduced in the meta-assembly, indicating that meta-assembly was able to extend

contigs by incorporating singletons. To demonstrate that contigs from different k-mer assemblies were used to create extended consensus contigs, genes from a candidate list of transcription factors were tracked from the 454 reads through the assembly and meta-assembly process (Table 3.3). Transcription factors are generally low abundance transcripts, and therefore full-length sequences are less likely to be recovered in single k-mer assemblies. Five out of the seven transcripts were extended through CAP3 re-assembly (Table 3.3). Primers were designed for four sequences and PCR amplification using *T. biloba* cDNA produced bands of the expected size, indicating that these extended contigs are correctly assembled transcripts. To better visualize how meta-assembly extends transcript length, we examined in further detail how *extradenticle* contigs from different assemblies were meta-assembled (Figure 3.6). The meta-assembly recovered the entire length of the coding sequence of the *Tbil-exd* transcript, as compared to *Drosophila*. Assembling the full transcript required contigs from multiple assemblies, and only a subset of the individual assemblies contained sequences fragments for the middle of the transcript. Contigs from assemblies outside the 23-29 k-mer range show a reduction in coverage caused by fragmentation in assemblies with shorter k-mer lengths and conservative assembly with larger k-mer lengths. The *Tbil-exd* sequence contains several single nucleotide insertions within the region aligned to the *Drosophila* reference and 83% of the nucleotide identities are conserved.

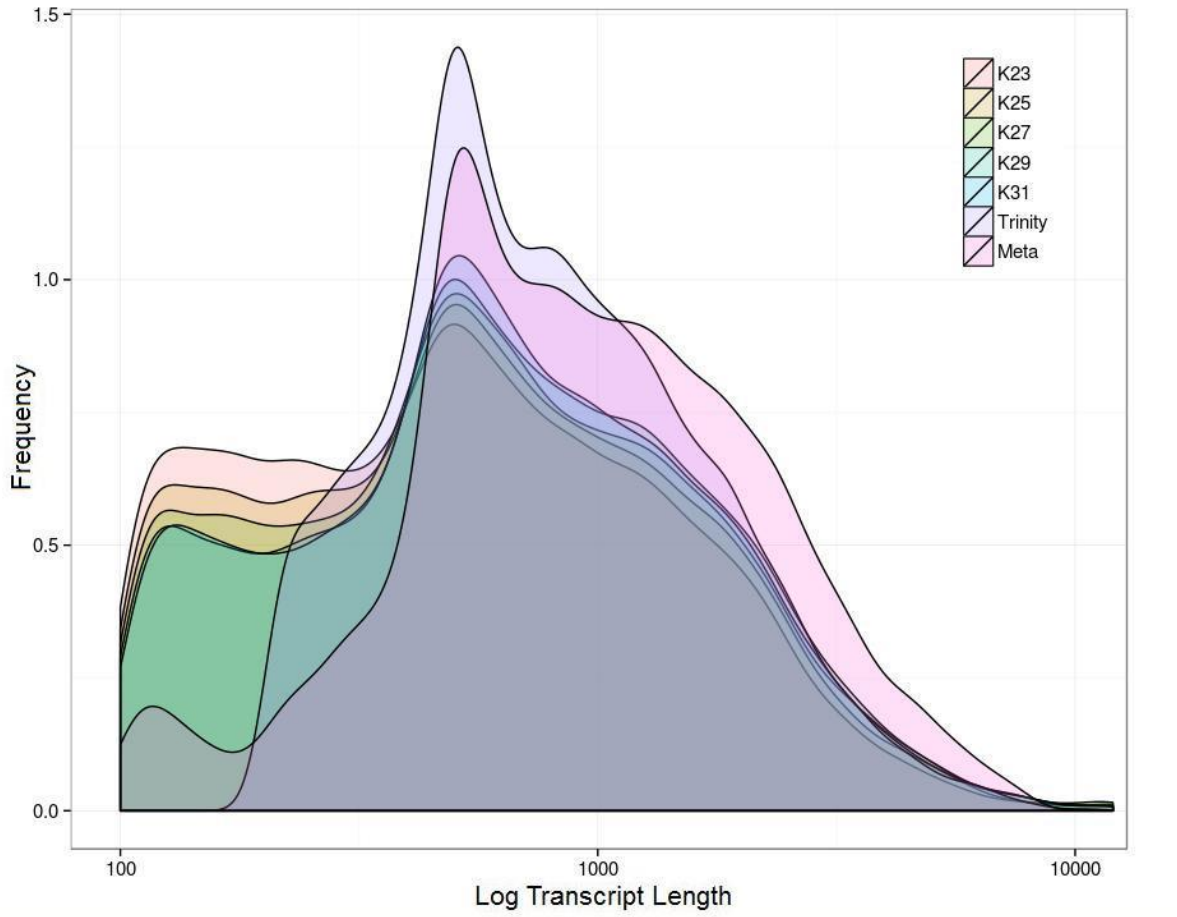

Figure 3.4. Frequency distribution of transcript lengths by assembly. A plot of the quantity of transcripts with a given length per assembly shows differences in assembly output and a pronounced peak representing the median transcript length. The meta-assembly was generated by the re-assembly of all k-mer lengths using CAP3. Meta-assembly improved transcript length, as indicated by the leading edge of the graph. Meta-assembly also reduced the number of short contigs, compared to the single k-mer assemblies. Trinity automatically removes contigs smaller than 200 base pairs.

Table 3.3. Transcripts of interest extended by meta-assembly

Identity	Meta-assembly	Individual assembly
engrailed*	1140	1140
escargot*	1244	782
evenskipped*	876	717
extradenticle	1143	574, 417, 138
hunchback	800	699, 472
Sex-combs reduced	232	281
Ultrabithorax	1084	526, 368, 370, 874

* Validated by PCR

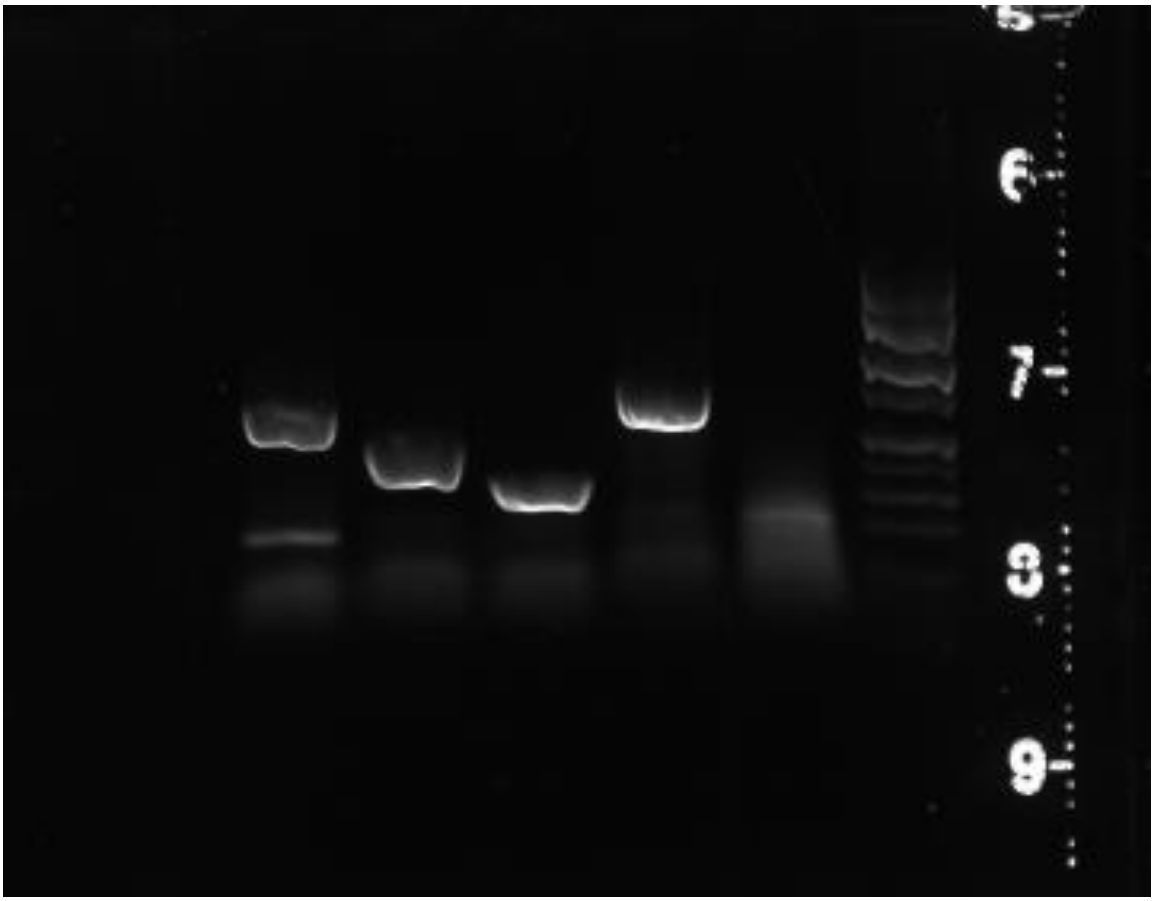

Figure 3.5. PCR validation of assembled contigs. Primers designed from bioinformatically generated contigs annotated using the *Drosophila* transcriptome produced the expected band sizes (from left to right) for *engrailed, escargot, and evenskipped.*

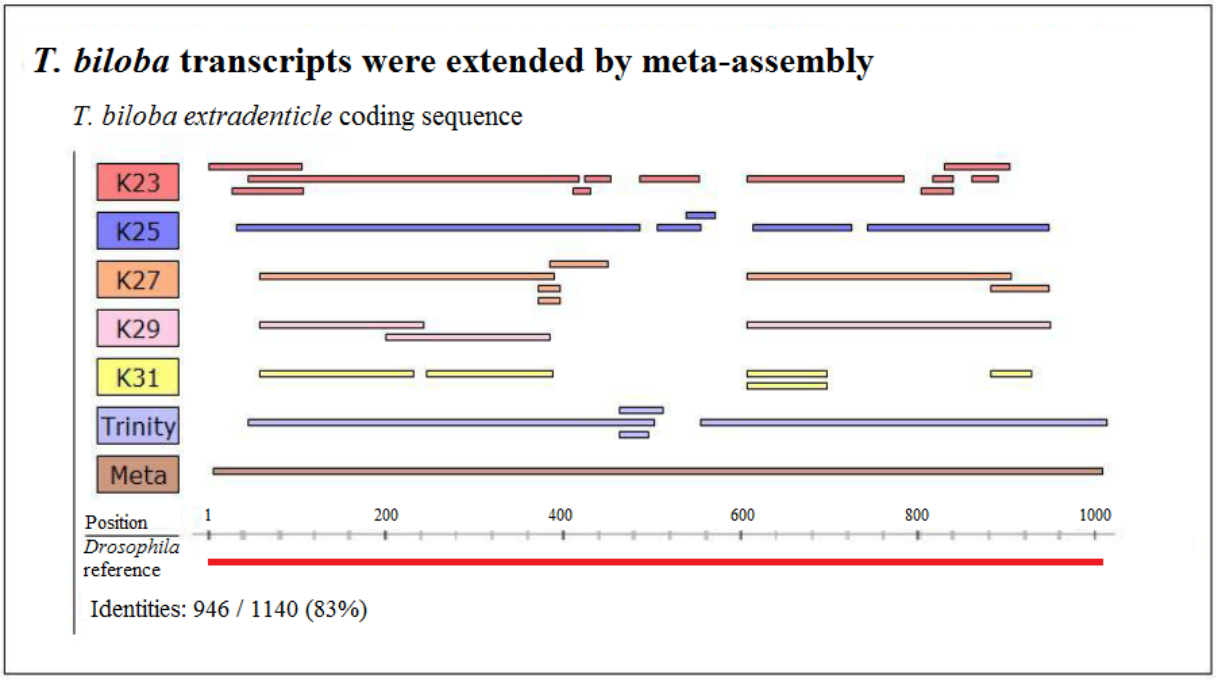

Figure 3.6. Extension of *extradenticle* sequence by meta-assembly. Contigs generated by multiple k-mer lengths were consolidated by meta-assembly to recover the entire coding sequence of the gene *extradenticle* from sequence fragments. Contigs from individual assemblies of multiple k-mer lengths are shown in alignment to the meta-assembly and the *Drosophila* transcript. The k-mer length 31 contigs were not included in the meta-assembly and show a reduction in coverage compared to other assemblies. Assemblies with shorter k-mer lengths also show a reduction in coverage but are not shown due to excessive fragmentation which results in a large number of short contigs that cannot be confidently aligned. The extended transcript aligns to the full length of the *Drosophila* reference sequence with 83% nucleotide sequence conservation.

To determine whether meta-assembly would improve transcriptome quality across taxa, the meta-assembly process was performed on three archived datasets (*Oncopeltus fasciatus*: SRR057573; *Silene vulgaris*: SRR245489; *Ictidomys tridecemlineatus*: SRR352220) using the same pipeline used to generate the *T. biloba* transcriptome. (Table 3.4; Figure 3.7). The meta-assemblies for each of the four datasets were compared to a single 25 k-mer length assembly.

Table 3.4. Single and multiple k-mer length meta assembly across 4 species

Assembly	Base-pairs	n	Median	Mean	n50	% reads used
I. tridecemlineatus						
25-mer assembly	25446725	33363	460	762	1328	49.97%
meta assembly	52328097	50869	608	1028	1708	70.26%
S. vulgaris						
25-mer assembly	21706584	34262	404	633	1124	66.31%
meta assembly	40068740	43475	815	921	1351	85.45%
O. fasciatus						
25-mer assembly	9487925	15886	421	597	894	69.12%
meta assembly	18283749	18106	797	1009	1312	81.80%
T. biloba						
25-mer assembly	20431185	22423	549	911	1571	58.87%
meta assembly	32887248	24495	887	1342	2010	64.01%

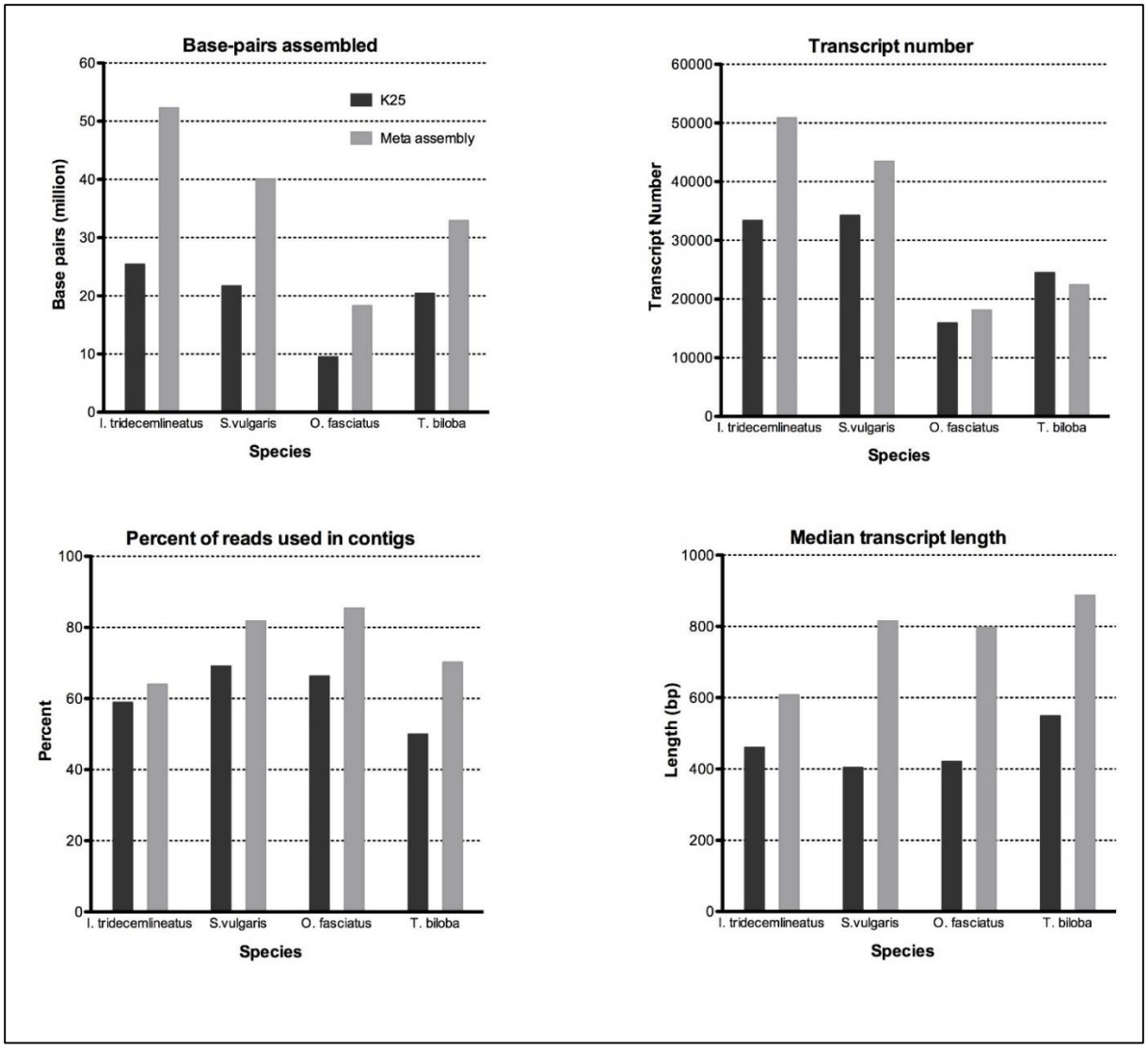

Figure 3.7. Performance of meta-assembly across species. A single assembly using Velvet-Oases with a K-mer length of 25 (light gray) was compared to the multiple k-mer length meta-assembly (black) for four species. Meta-assembly improved overall transcript length. The total assembled base-pairs (A), transcript number (B), percent of reads used in contigs (C), and median transcript length (D) show improvement in transcript assembly.

We used multiple metrics to compare transcription quality between the 25 k-mer length assembly and the meta-assembly including: number of base pairs assembled, number of contigs, percent of reads used in the contigs, and median contig length (Figure 3.7; Table 3.4). In all four datasets, the number of base pairs assembled was greater in the meta-assembly. The greatest increased was observed in *I. tridecemlineatus* in which the number of base pairs assembled

doubled with meta-assembly. Overall, the total number of assembled base pairs is 60.1% to 105.6% greater. The increase in base-pairs assembled was mirrored by an increase in contig length in all four species, as measured by mean contig length, median contig length, and n50 (Figure 3.7D; Table 3.4). The increase in length is presumably a result of incorporating more reads, because the percent of total reads that were assembled into contigs also increased with meta-assembly (Figure 3.7B). In addition to increasing contig length, the meta-assembly also increased contig number in the *I. tridecemlineatus, S. vulgaris,* and *O. faciatus,* data sets (Figure 3.7B). The increase in contig number is further evidence that meta-assembly recovers unique contigs from different k-mer length assemblies. The gain in contig number was likely even greater than the observed increase because the 25 k-mer assembly includes redundant contigs, whereas the meta-assembly does not. The same pre-processing steps were used to generate the filtered reads for both the 25 k-mer and meta-assemblies but the 25 k-mer assemblies did not undergo a secondary assembly to remove internal redundancy. When applied to a single Velvet-Oases assembly, CAP3 reduces the number of contigs by 5.5%. The only species to see a reduction in the number of contigs after meta-assembly was *T. biloba*. We hypothesize this reduction was due to either elimination of duplicates, consolidation of contigs, or both.

Alignment and annotation of the Themira biloba transcriptome

The *T. biloba* transcriptome was annotated using the *Drosophila melanogaster* transcriptome as a reference. The pipeline aligned the *T. biloba* transcripts to *D. melanogaster* using the standalone BLAST package and a reference database available from FlyBase (McQuilton et al. 2012). 11,008 transcripts from the meta-assembly were identified via BLAST as homologous to *Drosophila* sequences (44.9%). We found that the aligned *T. biloba* sequences were 82.3% conserved (mean sequence conservation taken from a subset of 500 BLAST hits)

indicating that BLAST may not be sufficient to identify some sequences. Therefore, sequence divergence between the two species could explain why over half the *T. biloba* contigs in the meta-assembly could be annotated based on *Drosophila*. However, contig mis-assembly could also cause low annotation rates. To determine whether sequence divergence or mis-assembly was the cause, we annotated the T. biloba transcriptome with a more closely related Dipteran.

Sepsidae is more closely related to Tephritidae than the drosophilids (Wiegmann et al. 2011), so it would be expected that higher sequence conservation exists between these two families, and that comparison to a tephritid would identify more transcripts. To determine whether such a comparison would identify more transcripts than *Drosophila,* a transcriptome was constructed using archived Illumina sequence reads from adult male and female *Bactrocera dorsalis* (SRR818498, SRR818496)("Bactrocera dorsalis (ID 167923) - BioProject - NCBI" n.d.). Bi-directional alignments were created using *T. biloba, B. dorsalis,* and *D. melanogaster*. Contrary to our prediction, the alignments between *T. biloba* and *B. dorsalis* did not show increased aligned contigs or even conserved sequence versus *Drosophila* (Table 5). On average, *B. dorsalis* had around the same sequence similarity to *T. biloba* that *Drosophila* did, and the number of matching transcripts actually decreased, as did the average length of the matching region. The decrease in number of matches may be due to the nature of the datasets. The *Drosophila* transcriptome includes multiple life stages and has a high level of coverage, whereas the *B. dorsalis* transcriptome only includes the adult stage [50]. Decreased representation could result in alignment of fewer genes even though the amount of sequence divergence is similar. In the end, annotation to *B. dorsalis* had the same limitations as Drosophila because of sequence divergence in the Sepsidae lineage.

Table 3.5. BLAST matches and percent identities

Query	Database	Matched	Unmatched	Mean length	Mean % conserved
T. biloba	*D. melanogaster*	11008	13487	1200	82.21%
T. biloba	*B. dorsalis*	6273	18222	729	82.46%
B. dorsalis	*D. melanogaster*	9334	41053	802	84.17%
B. dorsalis	*T. biloba*	6277	40273	726	85.99%
D. melanogaster	*T. biloba*	13544	23852	1336	82.37%
D. melanogaster	*B. dorsalis*	10333	26337	794	83.62%

To determine whether comparison with other more complete databases could increase the number of annotated contigs, the contigs from the *T. biloba* meta-assembly were compared to the SwissProt databases. SwissProt has the ability to compare translated contigs, thus reducing the problem posed by nucleotide divergence. Additional transcripts were annotated through BLASTx against the SwissProt database, which had not been annotated through the comparison with *D. melanogaster*. An expect-value cutoff of 0.00001 resulted in alignment of 16,705 (68.2%) of the translated sequences to sequences in the SwissProt database, which was a difference of 5,697 contigs (23.2%) compared to nucleotide BLAST against a single species. Analysis was performed to determine known protein domains in the Pfam database using the Trinity utility TransDecoder (Punta et al. 2011). An additional 221 contigs that had not been annotated were found to contain Pfam domains increasing the number of contigs identified by at least one searched database to 16,926 (69.1%). The number of annotated contigs compares favorably to other *de novo* assemblies (Schwartz et al. 2010, Wang et al. 2010, Bao and Xu 2011). The high percentage of annotated transcripts indicates that the contigs generated through meta-assembly are true transcripts, and not mis-assembled contigs. Further improvements in annotation likely require greater coverage through increased sequencing depth and a larger sequence data set. To determine ontology, *T. biloba* transcripts were submitted for KEGG

pathway analysis resulting in 5,080 contigs with identified functions. Many developmentally import pathways involved in cell signaling such as the notch pathway were near complete. Transcripts were assigned gene ontologies, which were then grouped by function (Figure 3.8) to determine whether the transcripts recovered from the meta-assembly were representative of the main cellular processes. A broad range of functional groups were present in the assembly, indicating that transcripts representing many different kinds of proteins were recovered. The distribution of contig gene ontologies is similar to those found in the distribution of GO terms found in the *Drosophila* transcriptome and other *de novo* transcriptome assembly efforts (Tweedie et al. 2009, Wang et al. 2010, Bao and Xu 2011, Ewen-Campen et al. 2011).

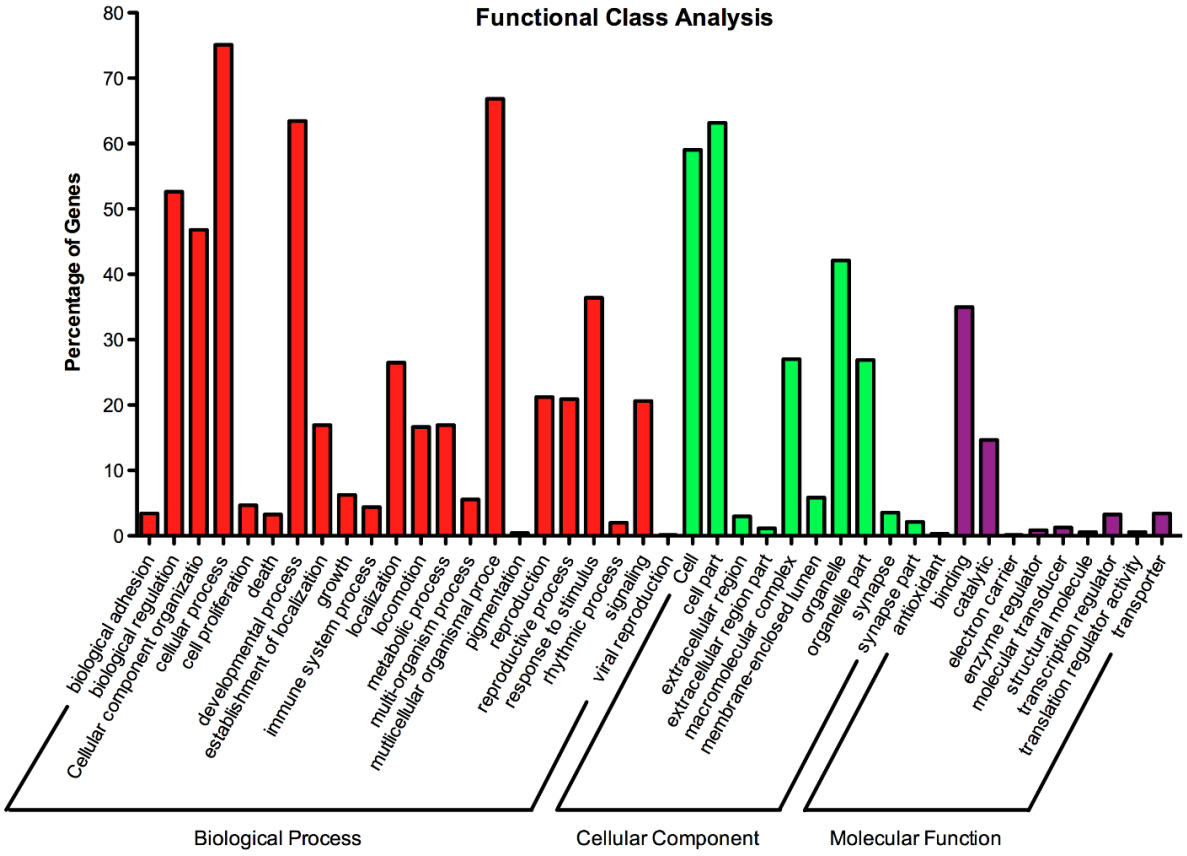

Figure 3.8. Gene Ontology classification of the *T. biloba* transcriptome. Gene Ontology (GO) was assigned to all contigs from the *T. biloba* meta-assembly. Gene ontologies were group into three main categories and 42 sub-categories. Contigs are grouped by the percentage of sequences that match a specific GO term within three major groups. The most abundant transcripts represent the sub-categories containing structural proteins and regulators of various cellular processes.

Discussion

Bioinformatics and data management

The *de novo* assembly of a transcriptome presents multiple challenges including computational requirements and accurate assembly of low abundance transcripts. Here we present a pipeline for *de novo* assembly that uses cloud computing and a multiple k-mer meta-assembly processes. The strength of a distributed, cloud-based approach to transcriptome assembly and sequence analysis is its versatility and the low initial investment in data processing(Sboner et al. 2011, Jourdren et al. 2012). We have found the primary advantage of hosting data analysis off-site is the ability to construct a low-cost, scalable network on demand with unrestricted access. The increased computing power is particularly important when generating multiple *de novo* assemblies, as is done in our meta-assembly processes. Meta-assembly processes that use a multiple k-mer length approach have been previously demonstrated to significantly improve the quality of transcriptomes (Surget-Groba and Montoya-Burgos 2010, Zhao et al. 2011).

The pipeline presented here incorporates an extensive and automated toolkit for parsing and trimming sequence reads prior to multiple k-mer assembly and the generation of a meta-assembly that best represents the transcripts available to be recovered. Automated sequence analysis tools are included to provide graphical views of read quality, transcript length and coverage per assembly, transcript extension, annotation information of sequence homologs from various databases, and the presence of unique sequences, and the assembly parameters used to recover the sequences.

Increasing transcriptome quality with meta-assembly

We validated our pipeline by assembling three previous published transcriptomes and the transcriptome of the sepsid fly *T. biloba,* which was sequenced as part of this project. Transcriptome quality was compared between our pipleline, which employs a meta-assembly process, and the standard practice of using a single 25bp k-mer length for assembly. In all four species, the meta-assembly increased the number of base pairs assembled, increased the length of contigs, increased the percentage of reads used in the contigs and recovered a greater number of transcripts than the 25 k-mer assembly. The increased quality of meta-assembly was further investigated in the *T. biloba* transcriptome by tracking the improvement in a candidate list of low abundance transcripts. For a subset of these transcripts, RT-PCR confirmed that meta-assembly increased the length of the transcripts by connecting fragments recovered from multiple k-mer length assemblies.

Conclusions

We have assembled transcript sequences from the complete life cycle of *T. biloba,* a sepsid fly which exhibits primary gain of a novel trait, and identified many developmentally important genes. These transcripts represent the first large-scale sequencing that has been performed within the family Sepsidae, a large and diverse family with over 250 species distributed globally. Sepsid flies have been used for taxonomic and behavioral studies and have diverse genital and appendage morphologies, but lack of sequence data has made genetic investigation of these traits difficult(Ang et al. 2008, Puniamoorthy et al. 2008, 2009, 2012, Eberhard 2001a, Bowsher et al. 2013). While many orthologous genes retain their functions between dipterans, large regions of gene sequence are often not conserved (Hare et al. 2008, Concha et al. 2010).

The *T. biloba* transcriptome and many of the genes we have identified will be used for future RNA-Seq studies of comparative gene expression, knockdown, and *in situ* hybridization experiments. Sequence for many developmentally important genes and transcription factors of interest were obtained including members of the HOX family and those associated with embryonic and morphological development. In addition, many sequences for genes involved in cell signaling pathways such as notch and torso signaling were recovered. Sequence for the *T. biloba doublesex* ortholog as well as several transcripts associated with mating and courtship in *Drosophila* were also recovered which aids investigation of the sepsid sex allocation pathway and the genetic mechanisms behind behavioral traits associated with the sepsid novel appendage.

As more genomes become available, researchers using non-model organisms will have the opportunity to assemble RNA-seq reads to reference genomes of closely related species. Assembling to a reference, when available, yields a higher quality transcriptome than *de novo* assembly, and this result is robust to low-levels of genomic divergence between species(Hornett and Wheat 2012, Vijay et al. 2013). Although these findings are encouraging, those working with non-model organisms should proceed with caution(DeWoody et al. 2013). Based on *in silico* studies, assembling to a reference that has a sequence divergence greater than 15% decreases the number of transcripts recovered compared to *de novo* assembly (Vijay et al. 2013). In our case, assembling the *T. biloba* reads to the *Drosophila* genome would have been inappropriate because the 17 % sequence divergence between the two species would have resulted in decreased transcript recovery compared to *de novo* assembly. Choosing a closer relative based on phylogeny does not necessarily solve the problem, as our additional comparison to *B. dorsalis* revealed. Because the amount of sequence divergence between a non-model organism and its closely related reference species is rarely known prior to high-throughput

sequencing, *de novo* assembly remains a powerful tool for recovering transcripts in non-model organisms.

Materials and Methods

T. biloba colony

Cultures of *T. biloba* were maintained in an incubator at 25C with a 16:8 hour light-dark cycle in overlapping generations. Larvae were raised in Petri dishes and fed agar mixed with soy infant formula (ProSobee) covered with a 1.0cm layer of cow dung. Adults were fed honey mixed with water and provided with cow dung to facilitate mating and egg-laying.

Tissue collection and sequencing

Tissue was collected from embryos, 3rd instar larva, and 48-72 hour pupa. During collection all material was stored at -80□C in RNALater, prior to shipment to the sequencing facility. Embryos were collected regularly and washed several times with an egg wash solution of 0.12 M NaCl and 0.01% Triton X-100 to remove dung. The eggs were dechorionated using a 3% bleac solution. Third instar, wandering-phase larvae were everted in PEM buffer (100mM PIPES-disodium salt, 2.0mM EGTA, 1.0mM $MgSO_4$ anhydrous, pH 7.0) to facilitate RNA extraction. Prior to pupation, gut-purged larvae were allowed to wander on moistened filter paper to remove dung and particulates. Pupae were staged to 48-72 hours before collection. All samples were stored in RNALater overnight at 4□C and transferred to -80□C for storage prior to sequencing.

RNA isolation, library cDNA preparation, and 454 sequencing were performed by the University of Arizona Genetics Core (UAGC). Prior to sequencing, the cDNA was screened using a 2100 Bioanalyzer (Agilent Technologies). Sequencing was done on a GS FLX Titanium (454 Life Sciences). Embryos, larvae, and pupae were sequenced separately, creating 3 separate

pools of sequence. Approximately 1.48 million reads total with an average length of 400bp were generated.

Assembly and annotation

Pre-processing of the sequence reads generated from *T. biloba* was performed using the FastX Toolkit("FASTX-Toolkit" n.d.). Adaptor sequences were removed using the trimmer function. The quality filter removed sequences in which 80% of the base pairs had a Phred score of less than 20. The remaining 1.01 million reads were then converted to FASTA. The FastX collapsing tool was used to consolidate redundant sequences to reduce the amount of memory needed during the assembly process. An assembly was performed using the collapsed reads to determine the reduction in memory required for assembly. We determined that although collapsing the reads significantly reduced the memory requirements for assembly, it was not necessary for the data sets described in this publication and may lead to a reduction in coverage. FastQC (v0.10.1) was used to assess the quality of reads before and after pre-processing("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data" n.d.).

Paired-end assemblies with K-mer lengths of 19 to 29 were generated using Velvet-Oases with an insert size of 200bp("Oases: a transcriptome assembler for very short reads" n.d., "Velvet: a sequence assembler for very short reads" n.d.). Trinity was used to generate an additional paired-end assembly(Grabherr et al. 2011, Henschel et al. 2012). The resulting contigs were aligned to *Drosophila* using standalone BLAST to identify developmentally important transcripts. A BLAST alignment was then performed using each individual assembly as the query and the pooled contigs from all other assemblies as the database to identify contigs

unique to each assembly. The assemblies were then concatenated and the pool of 138,954 transcripts was re-assembled using CAP3(Huang and Madan 1999).

CHAPTER FOUR: GENE EXPRESSION OF A DEVELOPING NOVEL TRAIT

Abstract

Gene co-option is thought to play a vital role in the evolution of novel morphologies. Sepsid flies possess a novel abdominal appendage used by males during mating. The appendages are produced by histoblast nests rather than imaginal discs and have evolved multiple times within the Sepsidae family allowing for the comparison of the genetic basis of appendage evolution. The histoblast cells are likely specified to develop into appendages during embryogenesis while appendage development and patterning takes place during pupation. Using mRNA sequence data we profile gene expression in the tissue which generates the abdominal appendage. By sampling segments in males and females which do not produce the appendage we are able to remove sexually dimorphic genes and genes present in non-appendage producing segments to identify a list of candidates which may be implicated in appendage development. The candidate list contains many transcription factors with known developmental and cell signaling functions in *Drosophila*. Multiple genes known to be involved in the growth and proliferation of epidermal tissue were identified which is consistent with the unique developmental origin of the appendage from histoblast tissue. Genes involved in wing patterning were also differentially expressed which indicates that the appendage may share part of this pathway which is not homologous to other insect limbs.

Introduction

The evolutionary mechanisms that produce novel traits are many and are all well described (Muller and Wagner 1991, Ang et al. 2008). Novel traits may be the product of novel coding sequences (although rare), gene duplication and divergence of function, novel exon splicing patterns, protein-protein interactions, and changes in the location, duration, and

magnitude of gene expression during development through modification of cis-regulatory networks (Carroll 2005, 2008, Hoekstra and Coyne 2007). The diversification of novel traits are more complex, in that a trait is exposed to many potentially competing selective pressures, developmental and environmental constraints, and pleiotropic interactions. These factors not only shape the evolution of the trait, but the integration of a novel trait affects the ecology and behavior of the organism as a whole. While the contribution of each of these mechanisms to the evolution and diversification of novelties is debated, the mechanisms themselves are understood and have been shown sufficient to produce novelties in many systems (Hoekstra and Coyne 2007, Moczek 2008, Carroll 2008). The appearance of a novelty may be the product of one or more of these mechanisms; they are not mutually exclusive. Novelties are interesting and challenging, because although they can be broadly categorized, the specific genetic and developmental processes that produce them are often unique and unusual.

Sepsid flies (Diptera: Sepsidae) are an excellent system for investigating novel traits. Adult male sepsid flies have a novel abdominal appendage consisting of a cluster of bristles and a joint with a wide range of motion that is mounted on a modified sternite. The appendage has a unique developmental pathway that is not homologous to that of other insect appendages. The appendage develops from the ventral histoblast nest on the 4th abdominal segment (Bowsher and Nijhout 2007). This histoblast nest lacks the complex three-dimensional structure and organization of imaginal discs. The appendage is used during courtship and mating and appendage morphology and associated behaviors are highly diverse between species. While the developmental origin of the appendage is understood (Bowsher and Nijhout 2007, 2009, Bowsher et al. 2013), the genetic mechanisms responsible for specification and development of the appendage during embryogenesis and pupation are incompletely described. The sepsid

abdominal appendage is an ideal candidate for mRNA sequencing (RNA-Seq) and differential gene expression analysis (DE) to identify developmentally important transcription factors which may be involved in producing the appendage.

DE is a widely used tool to identify changes in gene expression in response to experimental conditions and is appropriate for a number of reasons. The transcriptome of the sepsid fly *T. biloba* has recently been sequenced and assembled and can be used as a reference and for the development of molecular tools to confirm gene candidates identified by DE (Melicher et al. 2014). The appendage-producing tissues and stage at which appendage development begins are easily identifiable, allowing for the collection and comparison of tissues of known stage which do and do not produce the appendage (Bowsher and Nijhout 2007). Sepsid appendages are sexually dimorphic allowing for an additional comparison of gene expression between sexes of appendage producing tissues.

Our goal was to identify genes involved in patterning the developing appendage can be identified by dissecting and sequencing the larval epidermis segments containing the ventral histoblast nests (Figure 4-1). Specification of the histoblast nests occurs during embryogenesis and segment-specific differences are detectable in third-instar larva (Bowsher et al. 2013). Development of imaginal tissues begins immediately prior to pupation during the larval gut-purged or wandering phase. Just prior to this stage histoblast nest cells begin to rapidly divide and proliferate to form the adult abdominal epidermis, which is completely intact 72-96 hours after pupation begins (Bowsher and Nijhout 2009). The abdominal appendages become apparent very early in pupation, approximately when sex-specific genital morphology can be observed. By making multiple comparisons between appendage producing segments and sexes, a set of candidate genes can be identified with an expression pattern unique to the 4^{th} male segment

(Figure 4-2). Comparing gene expression between the 3rd and 4th male segments will remove genes that are globally expressed in abdominal segments, genes that are common to epidermal tissue but unrelated to the appendages or histoblast nests, and genes that are specific to histoblast nests but not unique to the appendage-producing nest. Comparing gene expression between the 4th male and 4th female segments will remove genes that are not sexually dimorphic. Comparing gene expression between the 3rd male and 4th female segments will remove genes that are globally expressed and those that are sexually dimorphic but do not have expression restricted to the appendage producing nest. The resulting list of differentially expressed genes are highly likely to be specific to the 4th male histoblast nest and involved in appendage patterning during pupation.

This gene expression analysis identified that most differentially expressed genes are sexually dimorphic (92.3%), and that most of sexually dimorphic genes are down-regulated (90.75%). Many of the differentially expressed genes are involved in the growth and proliferation of epidermal tissue which is consistent with the developmental origin of the appendage. Many transcription factors involved in wing patterning were also identified which indicates that the sepsid abdominal appendage may have co-opted part of the wing development pathway, or that genes involved in wing patterning are highly versatile and capable of producing novel morphologies when expressed in new tissues and locations.

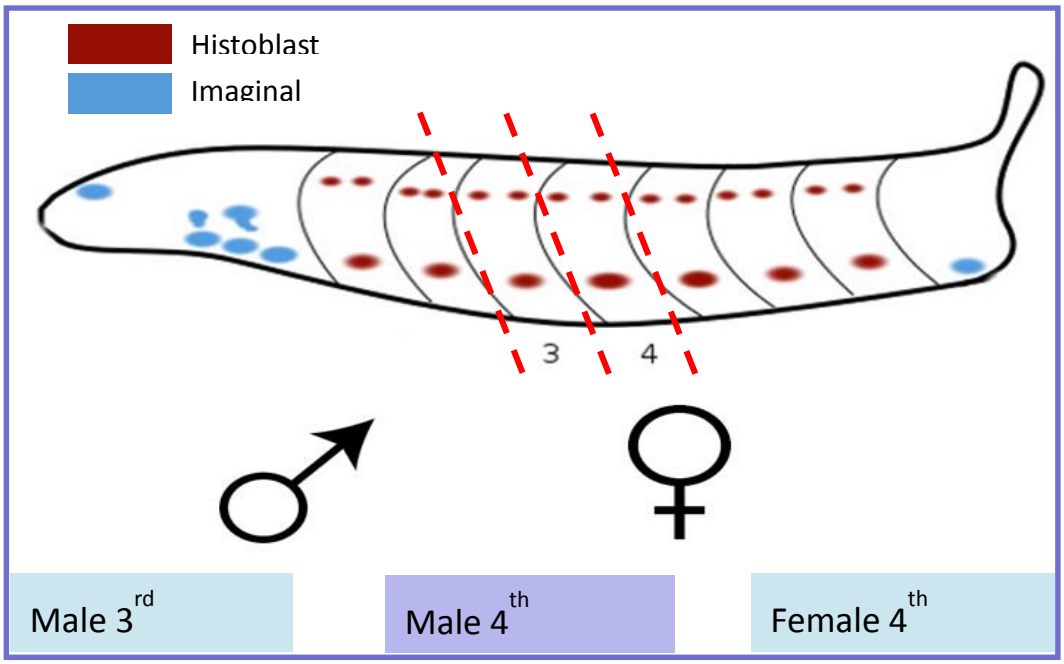

Figure 4.1. mRNA sequencing strategy. The 3rd male, 4th female, and the appendage-producing 4th male segments were dissected and sequenced.

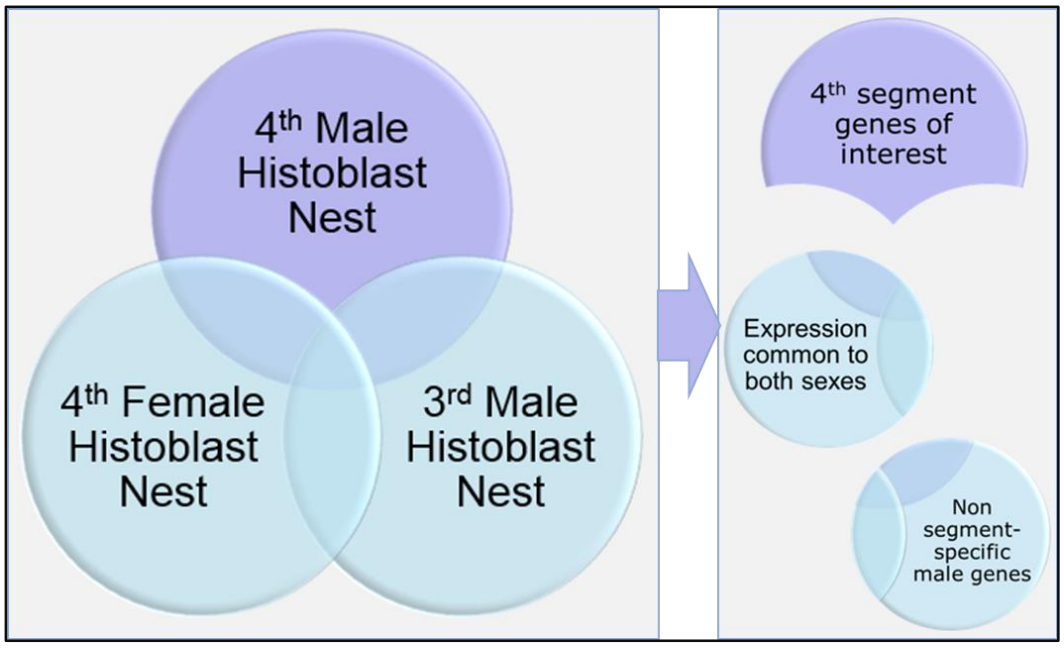

Figure 4.2. Candidate genes with expression patterns unique to the appendage-producing histoblast nest were identified by comparing expression in different segments and sexes to identify genes that are sexually dimorphic, segment-specific, and histoblast nest-specific.

Results

For a complete a list differentially expressed annotated *T. biloba* genes with *Drosophila melanogaster* homologs and Gene Ontology identities please see Appendix C.

Differences in gene expression in larval abdominal segments

Analysis of gene expression in the 4th male appendage-producing larval segment showed 499 genes that were differentially expressed relative to the 3rd male and 4th female segments. Of these, 51 genes were up-regulated, and 448 genes were down-regulated. Sexually dimorphic genes comprised the majority of differentially expressed genes. Comparisons of genes expressed in the 4th female segment to both male samples revealed 1079 (92.3%) genes that exhibit sexually dimorphic expression patterns. Comparing expression in the 3rd and 4th male segments identified 89 (7.62%) genes that are differentially expressed only in the 4th male segment (Table 4-1 and Figure 4-3). Among all comparisons, only a small minority of genes were up-regulated in male segments (9.24%) and only 13 genes were up-regulated in the 4th abdominal segment. We used the CummeRbund package for R to cluster genes based on expression and to visualize expression patterns (Figure 4-4) (Goff et al. 2012). Of the 9 expression patterns identified four were of interest to us because expression in the 4th male segment either up-regulated (Figure 4-4, panels 8-9) or down-regulated (Figure 4-4, panels 1, 4).

Table 4.1. Gene expression

Comparison	DE genes	Up-regulated	Down-regulated
4M x 3M	89 (7.62%)	13	76
4M x 4F	410 (35.1%)	38	372
3M x 4F	669 (57.28%)	57	612
Total	1168	108 (9.24%)	1060 (90.75%)

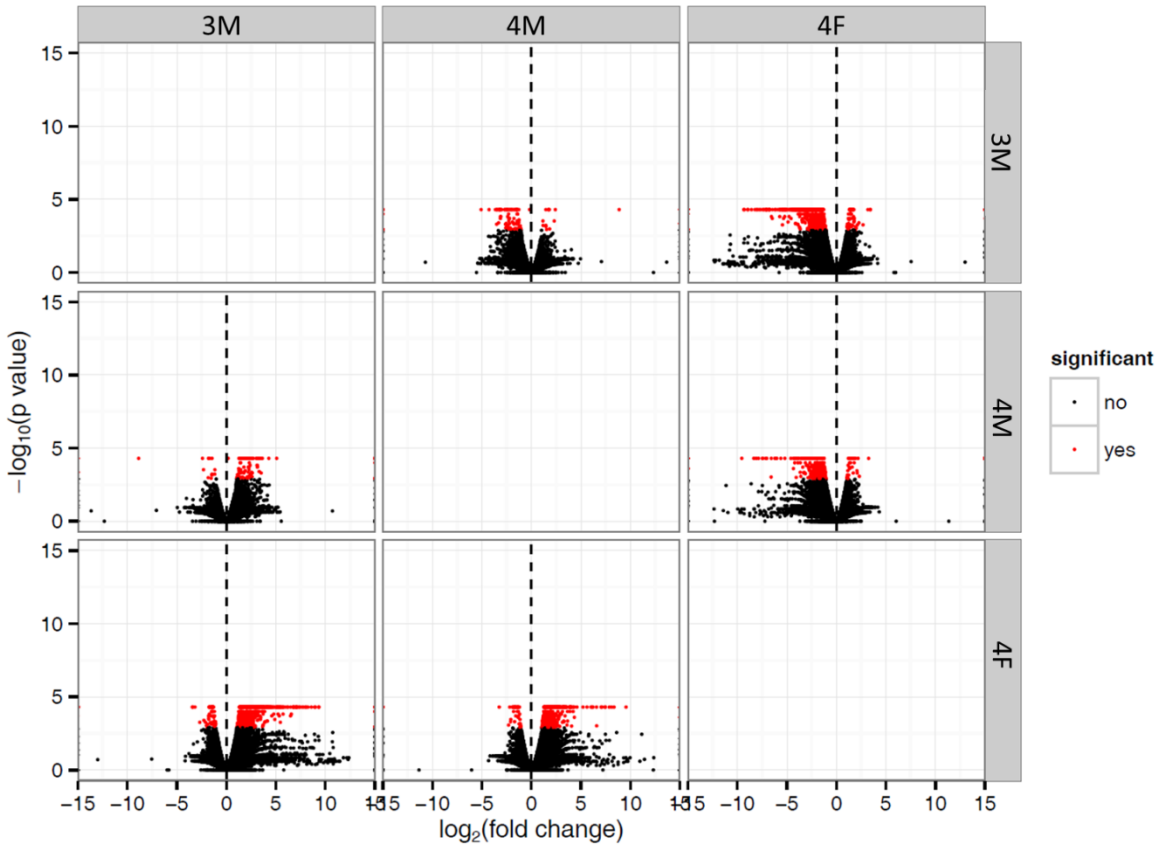

Figure 4.3. Multiple comparisons of log fold-change values show genes with significant differential expression between segments. Sexually dimorphic gene expression occurs at much high density than segment-specific gene expression4.

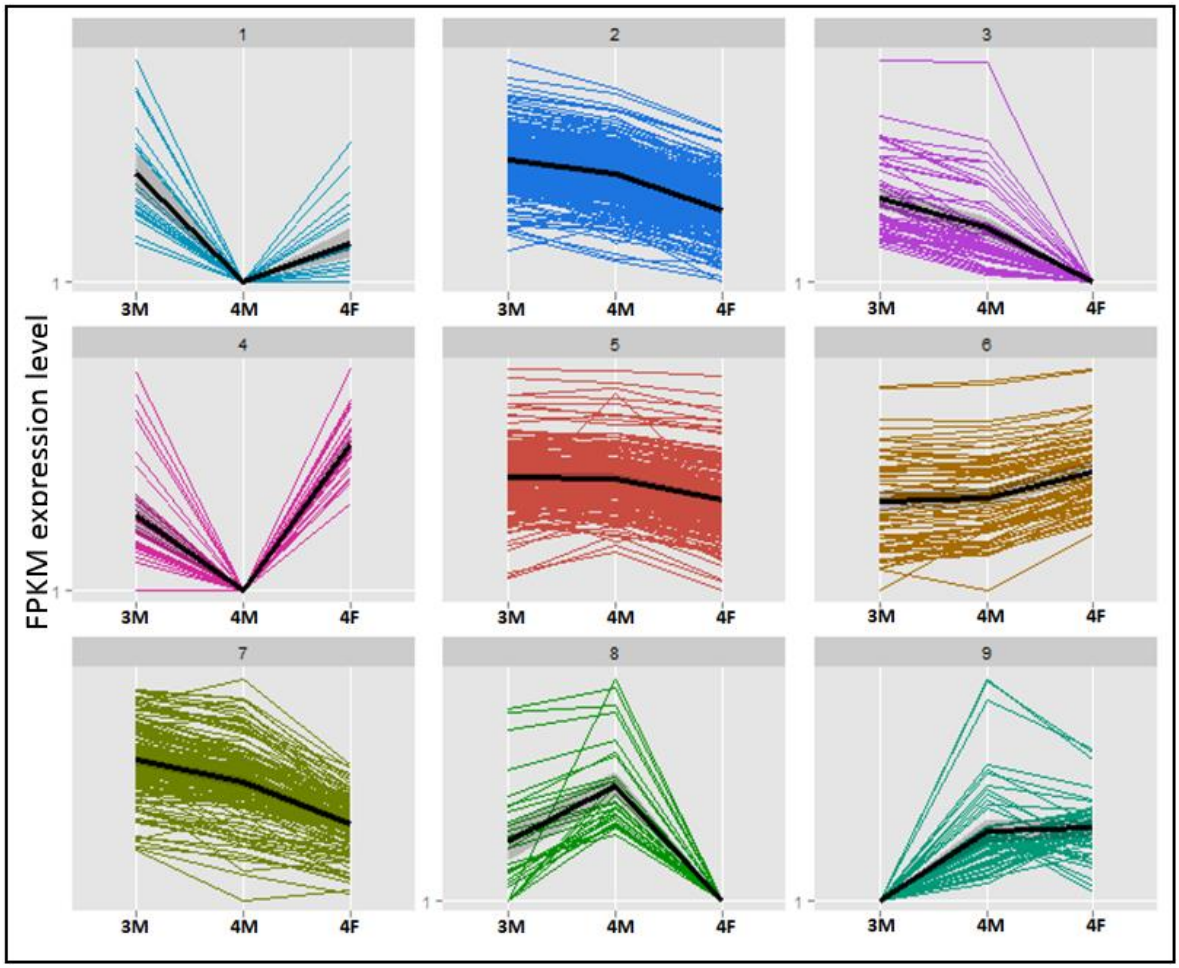

Figure 4.4. Genes grouped by expression pattern. Genes may be differentially expressed between sexes but not segments (3). Gene expression unique to the 4th male appendage-producing segment may show increased expression (8,9) or decreased expression (1,4) and may be directly involved in appendage patterning.

Identification of differentially expressed transcripts and biological functions

The subset of differentially expressed genes was then annotated using the *T. biloba* transcriptome and translated using BLASTx to identify known biological functions of genes that may be involved in appendage patterning. Gene ontology terms and functional characteristics were obtained for genes that were successfully identified through BLAST (Table 4.2).

Table 4.2. Selected genes with differential expression in the 4th segment and gene ontology derived biological function

Drosophila homolog	Biological function
tau	microtubule formation
derailed	muscle attachment, Wnt signaling
Roc1a	cell proliferation, smoothed signaling
fused	segment identity, smoothed signaling
stardust	epithelial morphogenesis
Larval cuticle p.	cuticle development
flapwing	imaginal disc wing morphogenesis
combgap	imaginal disc wing morphogenesis
broad	male genetalia morphogenesis, muscle development
slowdown	regulation of wing imaginal disc size, muscle attachment
CG9932	wing disc development, chaeta development
mastermind	wing surface imaginal disc and chaeta morphogenesis, Notch signaling
stubble	wing disc development, morphogenesis of larval imaginal disc epithelium
discs large 1	male courtship and mating behavior, morphogenesis of epithelium
little imaginal discs	larval muscle development
empty spiricles	HOX gene, mandibular segment identity, multiple developmental and morphogenic functions
white	male courtship behavior

Discussion

Sepsid abdominal appendages are an evolutionarily novel structure and are not homologous to other insect appendages in their morphology, location, or developmental pathway. Other insects lack abdominal appendages other than genitalia. Histoblast nests are not known to produce complex structures with three-dimensional organization in other species (Bowsher and Nijhout 2007). Previous research has shown through immunohistochemistry that some genes involved in appendage patterning such as *engrailed*, *Notch,* and *extradenticle* are expressed in the developing appendage, while others such as *distal-less* (*Dll*) are not (Bowsher and Nijhout 2009). Genes associated with the production of insect appendages such as *Dll* have been shown to produce novel beetle horns, treehopper helmets, and butterfly wing eyespots

(Nijhout 1991, Emlen et al. 2007, Prud'homme et al. 2011, Yoshizawa 2012). *Dll* expression in appendages and developing imaginal tissues shows that genes may be recruited to produce novelties with similar developmental pathways, but the recruitment of *Dll* to produce wing eyespots is a unique functionality that has led to the development of multiple competing models some of which are models in search of a homology (Carroll et al. 1994, Held 2013). This illustrates many incomplete aspects to our understanding of gene function. The classification of genes based on known function can also be problematic in that they are often associated with a long history of investigation that focuses on a limited aspect when other functions are unknown. Developmentally important transcription factors may be highly useful, versatile, and their functionality as logical regulatory elements may allow for interchangeability. However it is likely that a complex trait that is produced by gene co-option would involve multiple genes in a co-opted developmental pathway so genes may be versatile but limited by functional association within a network. The similarity of genes within gene families and the duplication and divergence of genes also allows for redundancy and some degree of interchangeability. Many broadly expressed, developmentally important transcription factors function as a form of logical regulation of the expression of genes that pattern structures. We might expect these regulatory transcription factors to be both highly interchangeable and also necessary for the development of a trait. Transcription factors are also modular and can evolve quickly without negative effect (Wagner and Lynch 2008). Many of the genes identified in this study are transcription factors and in the sepsid system a central question regarding co-option concerns the mobility of regulatory genes that are buried in developmental pathways and whether an existing homologous pathway produces the appendage or whether the pathway itself is novel.

The presence of gene expression associated with segment identity and cell-signaling during limb development indicates that it is likely that other genes involved in the patterning and developmental regulation of insect appendages are involved in these processes in the sepsid abdominal appendage during pupation (Bowsher and Nijhout 2009). However, because the abdominal appendage develops from histoblast nests instead of imaginal discs, genes associated with the organization of imaginal discs may be absent. Identification of other genes involved in appendage patterning was critical to understanding whether or not the genetic mechanisms that pattern the abdominal appendage are homologous to other insect appendages. Our objective was to identify these components of the regulatory network that pattern the developing appendage during morphogenesis using mRNA-Sequencing and gene expression analysis which allowed us to broadly characterize gene expression rather than investigating individual candidates.

Many of the genes that were differentially expressed are developmentally important transcription factors. While genes involved in appendage patterning found in previous studies were absent in our expression data, many genes that were identified fall in to several distinct categories. Genes associated with cell signaling, development and proliferation of epidermal tissue, and apoptosis such as *tau*, *fused*, *stardust*, *Larval cuticle protein*, and *Pez* were differentially expressed. Also found were *Wnt* and *smoothed* signaling pathway elements *derailed and* Roc1a.

It appears from our data that parts of the wing patterning pathway are involved in appendage development but many elements were not found. Many genes expressed in the appendage-producing segment are associated with wing patterning, the development of appendage-associated musculature, and innervation. This is encouraging as insect wings have a novel developmental pathway of unknown evolutionary origin that is not homologous to other

appendages. Commonalities in gene expression may indicate that part of the wing-development pathway may be co-opted to produce sepsid abdominal appendages. Wing-associated genes include as *flapwing, combgap, broad,* and *slowdown*. The genes *CG9932* and *mastermind* are involved in wing disc and chaeta development. *Mastermind* is also part of the *Notch* signaling pathway. In addition to wing patterning, *combgap, slowdown, flapwing, stubble, discs-large,* and *little imaginal discs* are also active in patterning imaginal discs. Expression of these genes in the developing histoblast nests may provide a necessary functionality allows the development of histoblast nest tissue in to a complex appendage. Some genes associated with male courtship behavior such as *white* are also expressed in the appendage-producing segment. Finally, the presence of the HOX gene *empty spiricles,* which is associated with antenna and mandibular segments, is also curious as there is speculation that the abdominal appendages may share a developmental pathway with mandibles (Bowsher and Nijhout 2009).

The expression of many genes in the appendage-producing segment involved in wing patterning and imaginal disc development, especially those involved specifically in the wing disc, along with genes involved in bristle development are an exciting development in our understanding of evolutionary novelties. The sepsid abdominal appendage may have co-opted elements of both the insect limb and wing patterning pathways. Specific to sepsids, this may indicate that the part of abdominal appendage developmental pathway is homologous to insect wings. Research on v*estigial* and *apterous* expression has shown that these genes are necessary to wing development and that *vestigial* expression in abdominal segments results in the gain of serially homologous bristle-like appendages (Clark-Hachtel and Tomoyasu 2016). While *vestigial* and *apterous* did not appear in our expression analysis, the upstream function in wing patterning and their ability to produce abdominal appendages makes them possible candidates for

specification of the histoblast nest. More broadly, the presence of multiple partially co-opted pathways would be made much more exciting by their capability of restructuring histoblast nest tissue into a novel complex, sexually dimorphic appendage under strong sexual selective pressure with diversity of associated behaviors. It will be necessary to localize gene expression to the histoblast nest itself using immunohistochemistry and the use of CRISPR deletions of non-lethal genes will allow us to better understand the role of these genes in developing appendage.

Methods

T. biloba colony

Cultures of *T. biloba* were maintained in an incubator at 25C with a 16:8 hour light-dark cycle in overlapping generations. Larvae were raised in Petri dishes and fed agar mixed with soy infant formula (ProSobee) covered with a 1.0cm layer of cow dung. Adults were fed honey mixed with water and provided with cow dung to facilitate mating and egg-laying.

Tissue collection and sequencing

Tissue was collected using a protocol adapted from Chapter 2 of this dissertation. 3rd instar wandering phase larva were identified by their pale yellow coloration. The larva were sexed based on the presence or absence of testes which appear as two large, clear ovoid masses between the 4th and 6th abdominal segments. Males and females were placed in separate collection dishes. Larva were sacrificed by immersion into water heated to 55C. This method results in a loss of muscle tone which aids in dissection and very consistent appearance of the larval epidermis. The larva were dissected in an isotonic phosphate-buffered saline (PBS) solution (see Appendix A). Lateral cuts at the location of the anterior and posterior tracheal cross-branches remove the head and open the larval posterior while leaving the abdominal segments 1-8 intact. A longitudinal incision was then made along the dorsal line between the

two main tracheal trunks. The tracheae, organ systems, and fat body were removed leaving a flat fillet of epidermal tissue.

The 3rd and 4th male segments and the 3rd female segment were dissected out by cutting along the denticle band that borders each segment. RNA was extracted according to TriZol protocol (see Appendix A) and quantified using Nanodrop. RNA was shipped on dry ice to the Beijing Genomics Institute (BGI) sequencing facility where quality was assessed using a BioAnalyzer prior to Illumina sequencing. Sequencing of the three samples generated 90 base pair reads per sample after adaptor trimming. The FastQC toolkit and BioPython were used to determine read quality and for the identification and trimming of over-represented sequences.

Expression analysis

Gene expression analysis and the identification of differentially expressed genes was performed using the Broad Institute Tuxedo Suite pipeline for read mapping and expression analysis. It was performed locally on a PC running an Ubuntu Linux 14.04 image using VirtualBox. Read mapping to the sepsid transcriptome was performed using Bowtie and Tophat. Differentially expressed transcripts were detected using Cufflinks. Three pairwise comparisons between each experimental sample were performed to identify expression levels in the corresponding tissue. Individual genes and expression patterns of interest were identified with CummeRbund (Trapnell et al. 2009, 2010, 2012b, 2012a, Roberts et al. 2011b, 2011a, Langmead and Salzberg 2012). Genes of interest were identified using GO terms and known functional associations using the *Drosophila* transcriptome. Gene function was further investigated using Flybase to prioritize genes based on previous research.

CHAPTER FIVE: CONCLUSIONS

This dissertation has several objectives. The first objective was to examine the evolutionary history of the abdominal appendage to identify patterns in histoblast nest morphology within and between species to determine if distinct character states exist or if transitional morphologies exist. Second, it was necessary to create high-coverage transcriptome sequences for *T. biloba* that may be used as a reference for gene expression studies and to allow for the investigation of the appendages using molecular tools designed for this species. The third objective was to identify genes expressed in the developing appendage using mRNA sequencing and expression analysis of specific tissues.

It appears that histoblast nest morphology does indeed fall into distinct character states represented by the evolutionary history of gains and losses of the abdominal appendage (Chapter 2). Gain of the appendage results in an enlargement of the histoblast nest in males which peaks in the appendage-producing segment. This sexually dimorphic histoblast nest morphology is shared by some primary gain species but is absent in others indicating that nest size may be correlated to the degree of sternite modification in the adult. The ancestral species *O. luctuosum* and the species that lost the histoblast nest both also lose the pattern of enlarged histoblast nest enlargement. Variation exists between primary gain species in individual histoblast nest size, but all of these species share a segment-specific and sexually dimorphic pattern not seen in species lacking the appendage. Finally, *P. dikowi* which recovered the appendage after the initial loss shows an enlarged female histoblast nest which represents a loss of sexual dimorphism in the developing larva although adult females still lack the appendage. This supports previous research which identified these patterns as distinct character states (Bowsher et al. 2013) and the increased sampling of closely related species for this dissertation indicates that while variation between species exists, transitional states do not. The histoblast nest data collected for this

research represents a valuable resource for future research and would be improved by measurements of the adult abdominal appendage morphology. It may be possible to correlate bristle number and length and the degree of sternite modification to histoblast nest size.

The development of the first sepsid transcriptome for *T. biloba* is a resource that will continue to benefit this system (Chapter 3). The custom bioinformatic pipeline developed for the assembly and annotation of this transcriptome improved the overall coverage and completeness (Melicher et al. 2014). It combines published methods for generating high-quality *de novo* assembly (Earl et al. 2011, Bradnam et al. 2013, Wilson et al. 2014) with a strategy of merging multiple conservative assemblies and multiple metrics for assessing over quality to retrieve a much higher amount of transcript sequence from a limited sequencing data set. Similar sophisticated approaches have been developed since the publication of the *T. biloba* transcriptome and pipeline to perform *de novo* assemblies in complex systems such as metagenomes of soil microbiomes (Crusoe et al. 2014, Howe et al. 2014). The *T. biloba* assembly and annotation were performed entirely using cloud computing resources. This demonstrates that cloud computing removes barriers to performing computationally expensive analyses and allows super-user access to computational resources similar in scale to large university servers traditionally used to analyze large bioinformatic data sets. The ability to save the state of a current analysis using a "snapshot" also allowed for easy troubleshooting and recovery and off-site archiving of data. The cloud resources available to individual users have rapidly increased over the previous five years and are expected to continue to grow exponentially. The ability to share snapshots and virtual machine images that are functional bioinformatics platforms with others using cloud resources has benefited this research through the use of tutorials and sample analyses which are valuable learning tools.

The *T. biloba* transcriptome was used to perform the first gene expression analysis on this species and identified several developmentally important transcription factors (Chapter 4). Gene Ontology analysis and a literature search of these genes identified a number of them that are involved in wing patterning (Clark-Hachtel and Tomoyasu 2016). While many of these genes are expressed during development of the wing disc, the genes *vestigial* and *apterous* which upstream regulators of wing development and necessary for wing formation were not differentially expressed in the appendage-producing segment. These gene sequences are present in the transcriptome, but may be expressed at too low of a level to be detected by size of our gene expression data set. The abdominal appendage also has co-opted part of the limb patterning pathway (Bowsher and Nijhout 2009) which may indicate that co-option of the wing patterning pathway may also be incomplete. This system will benefit from the application of CRISPR. Showing that these genes are necessary for normal appendage development and morphology using CRISPR deletions to modify gene expression will allow us to link gene expression to specific appendage morphologies. The development of transgenic lines to explore this trait will also be a valuable resource that will increase the visibility and viability of sepsids.

The sepsid system continues to be an ideal candidate for the study of evolutionary development and the evolution of novelty. The rich evolutionary history of Sepsidae and the presence of a novel sexually dimorphic and sexually selected trait linked to highly variable and complex courtship behaviors allow many opportunities for investigating the evolution of novelty within and between sepsid species. Sepsid novelty is complex the appendage, the developmental pathway, and the imaginal tissue from which the appendage develops are novel in Diptera. The same complexity that makes this system unique and interesting also make it challenging, but the

tools to investigate this system are rapidly improving and sepsids will continue to improve our understanding of how organisms evolve new traits.

REFERENCES

Abbasi, A. A. 2010. Unraveling ancient segmental duplication events in human genome by phylogenetic analysis of multigene families residing on HOX-cluster paralogons. Molecular Phylogenetics and Evolution 57:836–848.

Abràmoff, M. D., P. J. Magalhães, and S. J. Ram. 2004. Image processing with ImageJ. Biophotonics international 11:36–42.

ABySS 1.3.5 — Canada's Michael Smith Genome Sciences Centre. (n.d.). . http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.5.

Ang, Y., N. Puniamoorthy, and R. Meier. 2008. Secondarily reduced foreleg armature in Perochaeta dikowi sp.n. (Diptera: Cyclorrhapha: Sepsidae) due to a novel mounting technique. Systematic Entomology 33:552–559.

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (n.d.). . http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bactrocera dorsalis (ID 167923) - BioProject - NCBI. (n.d.). . http://www.ncbi.nlm.nih.gov/bioproject?LinkName=sra_bioproject&from_uid=366392.

Baena, M. L., and W. G. Eberhard. 2007. Appearances deceive: female "resistance" behaviour in a sepsid fly is not a test of male ability to hold on. Ethology Ecology & Evolution 19:27–50.

Bao, B., and W.-H. Xu. 2011. Identification of gene expression changes associated with the initiation of diapause in the brain of the cotton bollworm, Helicoverpa armigera. BMC Genomics 12:224.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67.

Blanckenhorn, W. U., U. R. S. Kraushaar, Y. Teuschl, and C. Reim. 2004. Sexual selection on morphological and physiological traits and fluctuating asymmetry in the black scavenger fly Sepsis cynipsea. Journal of Evolutionary Biology 17:629–641.

Blankenberg, D., G. V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. 2010. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *in* F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, editors. Current Protocols in Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Bowsher, J. H., Y. Ang, T. Ferderer, and R. Meier. 2013. Deciphering the evolutionary history and developmental mechanisms of a complex sexual ornament: the abdominal appendages of sepsidae (diptera). Evolution 67:1069–1080.

Bowsher, J. H., and H. F. Nijhout. 2007. Evolution of novel abdominal appendages in a sepsid fly from histoblasts, not imaginal discs. Evolution & development 9:347–354.

Bowsher, J. H., and H. F. Nijhout. 2009. Partial co-option of the appendage patterning pathway in the development of abdominal appendages in the sepsid fly Themira biloba. Development genes and evolution 219:577–587.

Bowsher, J. H., and H. F. Nijhout. 2010. Partial co-option of the appendage patterning pathway in the development of abdominal appendages in the sepsid fly Themira biloba. Development Genes and Evolution 219:577–587.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, and R. Chikhi. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:1–31.

Cahais, V., P. Gayral, G. Tsagkogeorga, J. Melo-Ferreira, M. Ballenghien, L. Weinert, Y. Chiari, K. Belkhir, V. Ranwez, and N. Galtier. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data: DE NOVO NGS-BASED TRANSCRIPTOME ASSEMBLY. Molecular Ecology Resources 12:834–845.

Carroll, S. B. 2005. Evolution at Two Levels: On Genes and Form. PLoS Biology 3:e245.

Carroll, S. B. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25–36.

Carroll, S. B., J. Gates, D. N. Keys, S. W. Paddock, G. E. Panganiban, J. E. Selegue, and J. A. Williams. 1994. Pattern formation and eyespot determination in butterfly wings. Science 265:109–114.

Casewell, N. R., W. Wüster, F. J. Vonk, R. A. Harrison, and B. G. Fry. 2013. Complex cocktails: the evolutionary novelty of venoms. Trends in ecology & evolution 28:219–229.

Clark, D. P., S. Durell, W. L. Maloy, and M. Zasloff. 1994. Ranalexin. A novel antimicrobial peptide from bullfrog (Rana catesbeiana) skin, structurally related to the bacterial antibiotic, polymyxin. Journal of Biological Chemistry 269:10849–10855.

Clark-Hachtel, C. M., and Y. Tomoyasu. 2016. Exploring the origin of insect wings from an evo-devo perspective. Current Opinion in Insect Science 13:77–85.

Collin, R., and M. P. Miglietta. 2008. Reversing opinions on Dollo's Law. Trends in Ecology & Evolution 23:602–609.

Concha, C., F. Li, and M. J. Scott. 2010. Conservation and sex-specific splicing of the doublesex gene in the economically important pest species Lucilia cuprina. Journal of Genetics 89:279–285.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics (Oxford, England) 21:3674–3676.

Crusoe, M. R., G. Edvenson, J. Fish, A. Howe, E. McDonald, J. Nahum, K. Nanlohy, H. Ortiz-Zuazaga, J. Pell, and J. Simpson. 2014. The khmer software package: enabling efficient sequence analysis. URL http://dx. doi. org/10.6084/m9. figshare 979190.

Darwin, C. 1859. On the origin of species. Murray, London 360.

DeWoody, J. A., K. C. Abts, A. L. Fahey, Y. Ji, S. J. A. Kimble, N. J. Marra, B. K. Wijayawardena, and J. R. Willoughby. 2013. Of contigs and quagmires: next-generation sequencing pitfalls associated with transcriptomic studies. Molecular Ecology Resources 13:551–558.

Dobzhansky, T. 1963. Animal species and evolution. by Ernst Mayr. XVI + 797 pp. Harvard University Press, Cambridge, 1963. American Journal of Physical Anthropology 21:387–389.

Domes, K., R. A. Norton, M. Maraun, and S. Scheu. 2007. Reevolution of sexuality breaks Dollo's law. Proceedings of the National Academy of Sciences 104:7139–7144.

Earl, D., K. Bradnam, J. S. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, and M. Diekhans. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome research 21:2224–2241.

Eberhard, W. G. 2001a. Species-specific genitalic copulatory courtship in sepsid flies (Diptera, Sepsidae, Microsepsis) and theories of genitalic evolution. Evolution; international journal of organic evolution 55:93–102.

Eberhard, W. G. 2001b. Multiple origins of a major novelty: moveable abdominal lobes in male sepsid flies (Diptera: Sepsidae), and the question of developmental constraints. Evolution & development 3:206–222.

Eberhard, W. G. 2005. Evolutionary Conflicts of Interest: Are Female Sexual Decisions Different? The American Naturalist 165:S19–S25.

Eberhard, W. G. 2012. Sexual behavior and morphology of Themira minor (Diptera: Sepsidae) males and the evolution of male sternal lobes and genitalic surstyli. The Canadian Entomologist 135:569–581.

Ekblom, R., and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15.

Emlen, D. J., L. C. Lavine, and B. Ewen-Campen. 2007. On the origin and evolutionary diversification of beetle horns. Proceedings of the National Academy of Sciences 104:8661–8668.

Ewen-Campen, B., N. Shaner, K. A. Panfilio, Y. Suzuki, S. Roth, and C. G. Extavour. 2011. The maternal and early embryonic transcriptome of the milkweed bug Oncopeltus fasciatus. BMC Genomics 12:61.

FASTX-Toolkit. (n.d.). . http://hannonlab.cshl.edu/fastx_toolkit/.

Gay, N. J., S. J. Poole, and T. B. Kornberg. 1988. The *Drosophila engrailed* protein is phosphorylated by a serine-specific protein kinase. Nucleic Acids Research 16:6637–6647.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko. 2005.

Galaxy: a platform for interactive large-scale genome analysis. Genome research 15:1451–1455.

Goecks, J., A. Nekrutenko, J. Taylor, and T. Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11:R86.

Goff, L., C. Trapnell, and D. Kelley. 2012. cummeRbund: analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2.

Goff, S. A., M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W. H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S.-J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S. M. Welch, K. A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D. J. Kleibenstein, J. W. White, J. Leebens-Mack, M. J. Donoghue, E. P. Spalding, T. J. Vision, C. R. Myers, D. Lowenthal, B. J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein, and D. Stanzione. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Frontiers in Plant Science 2.

Gompel, N., B. Prud'homme, P. J. Wittkopp, V. A. Kassner, and S. B. Carroll. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433:481–487.

Gould, S. J. 1970. Dollo on Dollo's law: irreversibility and the status of evolutionary laws. Journal of the History of Biology 3:189–212.

Goyal, N., J. K. Gupta, and S. K. Soni. 2005. A novel raw starch digesting thermostable α-amylase from Bacillus sp. I-3 and its use in the direct hydrolysis of raw potato starch. Enzyme and Microbial Technology 37:723–734.

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

Graveley, B. R. 2001. Alternative splicing: increasing diversity in the proteomic world. TRENDS in Genetics 17:100–107.

Gruenheit, N., O. Deusch, C. Esser, M. Becker, C. Voelckel, and P. Lockhart. 2012. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics 13:92.

Hampton, M., R. G. Melvin, A. H. Kendall, B. R. Kirkpatrick, N. Peterson, and M. T. Andrews. 2011. Deep Sequencing the Transcriptome Reveals Seasonal Adaptive Mechanisms in a Hibernating Mammal. PLoS ONE 6:e27021.

Hare, E. E., B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen. 2008. Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. PLoS Genetics 4:e1000106.

Held, L. I. 2013. Rethinking Butterfly Eyespots. Evolutionary Biology 40:158–168.

Henschel, R., P. M. Nista, M. Lieber, B. J. Haas, L.-S. Wu, and R. D. LeDuc. 2012. Trinity RNA-Seq assembler performance optimization. Pages 2–7. ACM Press.

Hoekstra, H. E., and J. A. Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution; international journal of organic evolution 61:995–1016.

Hornett, E. A., and C. W. Wheat. 2012. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. BMC Genomics 13:361.

Howe, A. C., J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. Proceedings of the National Academy of Sciences 111:4904–4909.

Hsu, J.-C., T.-Y. Chien, C.-C. Hu, M.-J. M. Chen, W.-J. Wu, H.-T. Feng, D. S. Haymer, and C.-Y. Chen. 2012. Discovery of Genes Related to Insecticide Resistance in Bactrocera dorsalis by Functional Genomic Analysis of a De Novo Assembled Transcriptome. PLoS ONE 7:e40950.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome research 9:868–877.

Ingram, K. K., T. Laamanen, N. Puniamoorthy, and R. Meier. 2008. Lack of morphological coevolution between male forelegs and female wings in Themira (Sepsidae: Diptera: Insecta). Biological Journal of the Linnean Society 93:227–238.

Iwasa, M., and T. H. Thinh. 2012. Taxonomic and faunistic studies of the Sepsidae (Diptera) from Vietnam, with descriptions of six new species: Sepsidae from Vietnam. Entomological Science 15:99–114.

Jones, G., and E. Teeling. 2006. The evolution of echolocation in bats. Trends in Ecology & Evolution 21:149–156.

Jourdren, L., M. Bernard, M.-A. Dillies, and S. Le Crom. 2012. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics (Oxford, England) 28:1542–1543.

Kumar, S., and M. L. Blaxter. 2010. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11:571.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.

Marshall, C. R., E. C. Raff, and R. A. Raff. 1994. Dollo's law and the death and resurrection of genes. Proceedings of the National Academy of Sciences 91:12283–12287.

Martin, J., V. M. Bruno, Z. Fang, X. Meng, M. Blow, T. Zhang, G. Sherlock, M. Snyder, and Z. Wang. 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11:663.

Martin, O. Y., and D. J. Hosken. 2003. The evolution of reproductive isolation through sexual conflict. Nature 423:979–982.

McQuilton, P., S. E. St Pierre, J. Thurmond, and FlyBase Consortium. 2012. FlyBase 101--the basics of navigating FlyBase. Nucleic acids research 40:D706–714.

Melicher, D., A. S. Torson, I. Dworkin, and J. H. Bowsher. 2014. A pipeline for the de novo assembly of the Themira biloba (Sepsidae: Diptera) transcriptome using a multiple k-mer length approach. BMC Genomics 15:188.

Moczek, A. P. 2005. The evolution and development of novel traits, or how beetles got their horns. BioScience 55:937–951.

Moczek, A. P. 2006. Integrating micro-and macroevolution of development through the study of horned beetles. Heredity 97:168–178.

Moczek, A. P. 2008. On the origins of novelty in development and evolution. BioEssays 30:432–447.

Muller, G. B., and G. P. Wagner. 1991. Novelty in Evolution: Restructuring the Concept. Annual Review of Ecology and Systematics 22:229–256.

Mundry, M., E. Bornberg-Bauer, M. Sammeth, and P. G. D. Feulner. 2012. Evaluating Characteristics of De Novo Assembly Software on 454 Transcriptome Data: A Simulation Approach. PLoS ONE 7:e31410.

Nijhout, H. F. 1991. The development and evolution of butterfly wing patterns. Smithsonian series in comparative evolutionary biology (USA).

Nirmala, X., M. F. Schetelig, F. Yu, and A. M. Handler. 2013. An EST database of the Caribbean fruit fly, Anastrepha suspensa (Diptera: Tephritidae). Gene 517:212–217.

Oases: a transcriptome assembler for very short reads. (n.d.). . http://www.ebi.ac.uk/~zerbino/oases/.

O'Neil, S. T., and S. J. Emrich. 2013. Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14:465.

Oshlack, A., M. D. Robinson, and M. D. Young. 2010. From RNA-seq reads to differential expression results. Genome Biology 11:220.

Park, C. B., M. S. Kim, and S. C. Kim. 1996. A Novel Antimicrobial Peptide fromBufo bufo gargarizans. Biochemical and biophysical research communications 218:408–413.

Pont, A. C. 2002. The Sepsidae (Diptera) of Europe. Brill, Leiden ; Boston.

Pont, A., and R. Meier. 2002. The Sepsidae (Diptera) of Europe. Brill.

Prud'homme, B., N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams, S.-D. Yeh, J. R. True, and S. B. Carroll. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440:1050–1053.

Prud'homme, B., C. Minervino, M. Hocine, J. D. Cande, A. Aouane, H. D. Dufour, V. A. Kassner, and N. Gompel. 2011. Body plan innovation in treehoppers through the evolution of an extra wing-like appendage. Nature 473:83–86.

Puniamoorthy, N., M. R. B. Ismail, D. S. H. Tan, and R. Meier. 2009. From kissing to belly stridulation: comparative analysis reveals surprising diversity, rapid evolution, and much homoplasy in the mating behaviour of 27 species of sepsid flies (Diptera: Sepsidae). Journal of Evolutionary Biology 22:2146–2156.

Puniamoorthy, N., M. A. Schäfer, and W. U. Blanckenhorn. 2012. Sexual selection accounts for the geographic reversal of sexual size dimorphism in the dung fly, sepsis punctum (Diptera: Sepsidae). Evolution; international journal of organic evolution 66:2117–2126.

Puniamoorthy, N., K. Su, and R. Meier. 2008. Bending for love: losses and gains of sexual dimorphisms are strictly correlated with changes in the mounting position of sepsid flies (Sepsidae: Diptera). BMC Evolutionary Biology 8:155.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. 2011. The Pfam protein families database. Nucleic Acids Research 40:D290–D301.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. Methods in Ecology and Evolution 3:217–223.

Rideout, E. J., J.-C. Billeter, and S. F. Goodwin. 2007. The Sex-Determination Genes fruitless and doublesex Specify a Neural Substrate Required for Courtship Song. Current Biology 17:1473–1478.

Roberts, A., H. Pimentel, C. Trapnell, and L. Pachter. 2011a. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325–2329.

Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. 2011b. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 12:R22.

Sanders, L. E., and M. N. Arbeitman. 2008. Doublesex establishes sexual dimorphism in the Drosophila central nervous system in an isoform-dependent manner by directing cell number. Developmental Biology 320:378–390.

Sboner, A., X. J. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein. 2011. The real cost of sequencing: higher than you think! Genome biology 12:125.

Schwartz, T. S., H. Tae, Y. Yang, K. Mockaitis, J. L. Van Hemert, S. R. Proulx, J.-H. Choi, and A. M. Bronikowski. 2010. A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. BMC Genomics 11:694.

Schwarz, D., H. M. Robertson, J. L. Feder, K. Varala, M. E. Hudson, G. J. Ragland, D. A. Hahn, and S. H. Berlocher. 2009. Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot Rhagoletis pomonella. BMC Genomics 10:633.

Simpson, G. G. 1955. Major features of evolution.

Sloan, D. B., S. R. Keller, A. E. Berardi, B. J. Sanderson, J. F. Karpovich, and D. R. Taylor. 2012. De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). Molecular Ecology Resources 12:333–343.

Soshnikova, N., R. Dewaele, P. Janvier, R. Krumlauf, and D. Duboule. 2013. Duplications of hox gene clusters and the emergence of vertebrates. Developmental Biology 378:194–199.

Stern, D. L., and V. Orgogozo. 2009. Is genetic evolution predictable? Science 323:746–751.

Surget-Groba, Y., and J. I. Montoya-Burgos. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome research 20:1432–1440.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. 2012a. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology 31:46–53.

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

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. 2012b. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7:562–578.

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

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, H. Zhang, and The FlyBase Consortium. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Research 37:D555–D559.

Velvet: a sequence assembler for very short reads. (n.d.). . http://www.ebi.ac.uk/~zerbino/velvet/.

Vijay, N., J. W. Poelstra, A. Künstner, and J. B. W. Wolf. 2013. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. Molecular Ecology 22:620–634.

Vogel, H., and C. W. Wheat. 2012. Accessing the Transcriptome: How to Normalize mRNA Pools. Pages 105–128 *in* V. Orgogozo and M. V. Rockman, editors. Molecular Methods for Evolutionary Genetics. Humana Press, Totowa, NJ.

Wagner, G. P., and V. J. Lynch. 2008. The gene regulatory logic of transcription factor evolution. Trends in Ecology & Evolution 23:377–385.

Wagner, G. P., and V. J. Lynch. 2010. Evolutionary novelties. Current Biology 20:R48–R52.

Wang, X.-W., J.-B. Luan, J.-M. Li, Y.-Y. Bao, C.-X. Zhang, and S.-S. Liu. 2010. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11:400.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10:57–63.

Wheat, C. W., and H. Vogel. 2012. Transcriptome Sequencing Goals, Assembly, and Assessment. Pages 129–144 *in* V. Orgogozo and M. V. Rockman, editors. Molecular Methods for Evolutionary Genetics. Humana Press, Totowa, NJ.

Wiegmann, B. M., M. D. Trautwein, I. S. Winkler, N. B. Barr, J.-W. Kim, C. Lambkin, M. A. Bertone, B. K. Cassel, K. M. Bayless, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, T. Pape, B. J. Sinclair, J. H. Skevington, V. Blagoderov, J. Caravas, S. N. Kutty, U. Schmidt-Ott, G. E. Kampmeier, F. C. Thompson, D. A. Grimaldi, A. T. Beckenbach, G.

W. Courtney, M. Friedrich, R. Meier, and D. K. Yeates. 2011. Episodic radiations in the fly tree of life. Proceedings of the National Academy of Sciences 108:5690–5695.

Wiegmann, B. M., D. K. Yeates, J. L. Thorne, and H. Kishino. 2003. Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. Systematic Biology 52:745–756.

Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson. 2014. Best Practices for Scientific Computing. PLoS Biology 12:e1001745.

Yoshizawa, K. 2012. The treehopper's helmet is not homologous with wings (Hemiptera: Membracidae). Systematic Entomology 37:2–6.

Zhao, Q.-Y., Y. Wang, Y.-M. Kong, D. Luo, X. Li, and P. Hao. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12:S2.

Zheng, W., T. Peng, W. He, and H. Zhang. 2012. High-Throughput Sequencing to Reveal Genes Involved in Reproduction and Development in Bactrocera dorsalis (Diptera: Tephritidae). PLoS ONE 7:e36463.

APPENDIX A: PROTOCOLS AND REAGENTS

Ringer's Solution (dissection buffer)

Combine the following in 900mL of ddH2O and stir well to dissolve. Adjust the pH to 7.2 using 1N HCl. Bring the volume to 1L by adding ddH2O. Autoclave if desired, although it is not necessary in most cases.

13.6g KCl
2.7g NaCl
0.33g CaCl2 * 2H2O
1.21g Tris base

PBS buffer 500mL 10x

Combine the following in 300mL of ddH2O and stir well to dissolve. Autoclave if desired.

To 300ml DEPC water add:
51.1g NaCl
5.97g Na2HPO4
1.28g NaH2PO4

Adjust pH pH to 7.4 with NaOH
Raise volume to 500ml with DEPC water
Store at RT

PEM pH 7.0

To 500mL ddH2O add:
15.12g PIPES-Disodium salt (final concentration 100mM)
0.380g EGTA (final concentration 2.0mM)
0.060g anhydrous MgSO4 (final concentration 1.0mM)

Mix for 20 min
Adjust pH to 7.0 with concentrated HCl
Store in a 500ml glass bottle in refrigerator
Will keep for one year

Kahle's Fixative

Combine the following:

12mL formalin
32mL ethanol (100% absolute)
2mL glacial acetic acid
60mL ddH2O

Store at room temperature.

Schiff's Feulgen reagent (basic fuschin stain)

1. Add 1g of basic fuschin to a 500mL Erlenmeyer flask.
2. Add 200mL of boiling water to dissolve the stain. Do not add the stain to boiling water or it will boil over and your workspace will be forever pink.
3. Swirl to stir for 5 minutes.
4. Insert thermometer and cap with aluminum foil.
5. Prepare a vacuum filter bottle by wrapping it in aluminum foil to keep light out.
6. Filter into the vacuum bottle when the temperature reaches exactly 50C.
7. Add 20mL of 1N HCl to the filtrate.
8. Insert the thermometer and cover with aluminum foil.
9. Wait until the temperature reaches 25C. This will take much longer, although you can speed it with the refrigerator if you keep a close watch.
10. At 25C add 1g of sodium metabisulfite and swirl to stir for 5 minutes.
11. Store at room temperature in a darkened cabinet for 24 hours. The solution should become pale, yellow-pink and somewhat clear. This may take up to 72 hours.
12. Add 1g of activated charcoal and swirl to stir for 2 minutes.
13. Filter into a foil-wrapped vacuum bottle. Filtrate should appear clear and yellow.
14. Store the solution at 4C.

Notes:

A drop of the clear, yellow solution on a paper towel should slowly become strongly pink as the solution dries and with light exposure.

It is very difficult to remove the pink color from any surface including metal or glass. The color may take some time to develop on clothing and skin, but is more or less permanent so be cautious of dripping and work carefully. If used solution is disposed of in a sink, the sink will become stained pink and a strong sour smell of hydrochloric acid will persist for quite some time.

It is recommended to bring the solution to room temperature before use, but in almost all cases this is not necessary because of the small quantities used. 2mL will raise to room temperature quickly enough not to matter.

Too pink:

Make sure the tissue is flat. The tissue may roll into a ball or wrinkle if not enough time is allowed during the dehydration/rehydration cycles. If the tissue is rolled into a ball rather than being flat it will become solidly magenta and be useless. Some samples may curl, but if this happens often enough to be a problem simply add more time and add steps to the ethanol gradient during dehydration/rehydration.

Not pink enough:

If a darker stain is desired, expose the tissue to more light during the first rinse. Bright light will develop the stain to a more intense pink. This must be done during the first rinse, as the color will stabilize after repeated water changes and the ethanol dehydration prior to storage or mounting. Pay close attention so it doesn't over-develop and replace in a shaded area for subsequent water changes.

Feulgen staining for whole-mount larval integuments

Adapted from:
M. M. Madhavan and K. Madhavan. (2004) "Analysis of Histoblasts." In Methods in Molecular Biology: Drosophila Cytogenetics Protocols, vol 247. D. S. Henderson, ed. Humana Press, Totowa, New Jersey.
Materials:

1. Dissection buffer
    a. Drosophila Ringer's solution (or Lepidopteran Ringers)
    b. PBS is acceptable

   Kahle's fixative
   Schiff's Feulgen reagent (basic fuschin stain)
   6*N* HCl
   Clearing agent (HemoDe, Histoclear, Xylene)
   Mounting medium (toluene-based, Permount, DPX)

Dissection and fixation:

1. Collect larva of appropriate age or stage in a Petri dish lined with moist filter paper.
2. Sex them if required by identifying the presence or absence of testes. They often appear ⅓ of the way from the posterior end and appear as two clear spheres displacing the fat bodies. When poked with a forceps, the sphere will persist, sliding to either side rather than collapsing (which just indicates a space between fat bodies).
3. Heat an eppendorf tube (or other appropriate container) filled with dissection buffer to 55C in a hot block.
4. Transfer larva to to the heated dissection buffer. Within 30 seconds the heat-killed larva will straighten. If not, increase the temperature to 60C.
5. Transfer the larva to a Petri dish filled at least half-full with dissection buffer. The larva should sink to the bottom.
6. Under a dissection scope hold the center of the larva with a forceps. Lift the anterior (head) end into the scissors and make a transverse cut. Flip the larva and make a transverse cut across the posterior end removing the spiracles. Do not remove the anus. This should leave an opening large enough to insert the scissors.
7. Hold the larva dorsal-side up and insert the scissors into the posterior end. Make a longitudinal cut down the dorsal midline. Use the trachea as a guide if necessary, cutting between them. Rather than moving the scissors, slide the larva farther up the lower blade for each cut using the forceps. Be cautious not to let the larva roll off-center. Once the longitudinal cut is complete, the larva should open into a wide, flat fillet exposing the internal organs and fat bodies.

8. Use a dull forceps to hold the fillet down while removing the internal organs without damaging the integument or tearing the attached muscles. A gentle scraping technique can be used to roll away the internal organs from anterior to posterior.
9. Once the tissue appears clean, transfer it to a dish containing fresh dissection buffer to rinse. The dissection dish tends to become full of adipose tissue and bits which may become a problem during staining. It may be necessary to change dissection buffer after around 5-10 dissections depending on skill.
10. If the tissue becomes curled, flatten it gently against the bottom of the dissection dish. Transfer the epidermis from the buffer to an eppendorf tube containing Kahle's fixative. Make sure all tissue samples are fully submerged. Leave for 18-24 hours.
11. Store the tissue in ethanol or proceed to staining.
12. Storage:
    a. Dehydrate the tissue using increasing concentrations of ethanol from 50%, 70%, 90%, to 100% incubating for at least 5 minutes in each solution. If the tissue becomes curled, increase incubation time.
    b. Tissue can be stored at room temperature in 100% ethanol overnight or for several days as dissections continue prior to staining, but should be stained as soon as possible.
    c. Prior to staining, rehydrate with 90%, 70%, 50%, 30%, and finally ddH2O incubating in each for at least 5 minutes.

Staining:

1. The following steps may be performed easily using a cell-culture plate. Multiple samples may be stained using more than one well, but to keep incubation times constant it is recommended not to use more than 4-6 wells depending on pipetting speed.
2. From fixative, transfer the tissue into decreasing concentrations of ethanol starting with 70%, to 50%, 30%, and finally to ddh2O incubating for 5 minutes in each. If the tissue becomes curled, skip the 70% (the fixative is 30% ethanol).
3. Remove the distilled water and add 2mL of 6N HCl for hydrolysis. Incubate for 10 minutes. Remove the acid and rinse with distilled water, paying special attention to rinse the acid from the sides of the well.
4. Quickly remove the water and add 2mL of Schiff's reagent making sure all tissue samples are submerged. Cover the dish and place in a light-proof box for 90 minutes at room temperature.
5. Remove the Schiff's reagent and rinse with distilled water. The pink color will develop and darken with light exposure during this first rinse. If a darker color is desired increase light exposure, but pay close attention so the tissue does not over-develop. This may take practice.
6. Perform 2-4 additional rinses by changing the distilled water quickly. Then continue rinsing 3-5 more times with a minimum 3 minute incubation between rinses. As long as the tissue is shaded, it is okay to walk away. The color will be stable after this step and the rinse water should no longer be pink. Perform a final 30 minute rinse in distilled water.
7. Dehydrate the tissue with ethanol from 30%, 50%, 70%, 90%, 100% incubating for at least 5 minutes per stage to prevent the tissue from curling.
8. Store in 100% ethanol or proceed to mounting.

Mounting

1. Transfer the tissue to a GLASS dish containing 50% ethanol and 50% clearing agent and incubate for 5 minutes. HemoDe and Histoclear have both been used for this protocol with good results- both are based on a citrus extract d-limonene mixed with a hydrocarbon, xylene may work but has not been tested with the mounting agent. The clearing agent will dissolve some plastics, including cell culture dishes.
2. Transfer to fresh 100% clearing agent and incubate for 5 minutes, then repeat.
3. Place a drop of mounting medium on a slide. This slide will be used to remove the clearing agent from the tissue and can be used repeatedly. The mounting medium smells powerfully bad and permanently adheres to any surface so a garbage slide is the most efficient way to do this.
4. Using forceps, carefully remove an epidermis and swish it gently through the mounting medium to remove the clearing agent. The clearing agent will dilute the mounting medium and increase drying time from days to weeks unless this is done.
5. Place the epidermis on a clean slide, cuticle side up. Flatten the tissue carefully and position as desired.
6. Place two large drops of mounting medium over the tissue and apply a coverslip. Use the back of the forceps to gently force the coverslip down. A small amount of mounting medium should leak out the sides. A little is desired as it will dry quickly and hold the coverslip in place while the rest dries, while too much will make a big mess.
7. Place a small weight over the tissue and ensure that the mounting medium spreads to completely fill the space under the coverslip. If there is a gap, simply apply a drop to the edge of the coverslip and it will spread to fill it. The mounting medium will continue to flow around for several days, and if too little is applied it may crawl away from the tissue and ruin the sample before drying.
8. Store flat for at least 2-3 days, longer if possible. The slides can be viewed within this time, but the mounting medium will not be dry so they cannot be stored on edge. Thinner tissues will dry faster and have less risk of air bubbles.

APPENDIX B: HISTOBLAST NEST AND LARVAL MEASUREMENT DATA

Statistical analysis using Lme4 and lmerTest

```
> LMM = lmer(scale(Cell_counts)~Sex*Segment+(as.numeric(Segment)|ID),Data_final)
> summary(LMM)
Linear mixed model fit by REML
t-tests use  Satterthwaite approximations to degrees of freedom ['lmerMod']
Formula: scale(Cell_counts) ~ Sex * Segment + (as.numeric(Segment) | ID)
   Data: Data_final

REML criterion at convergence: 100.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9535 -0.6208 -0.0120  0.5132  3.5232

Random effects:
 Groups   Name                Variance  Std.Dev. Corr
 ID       (Intercept)         0.0128314 0.11328
          as.numeric(Segment) 0.0001119 0.01058  -1.00
 Residual                     0.1666769 0.40826
Number of obs: 84, groups:  ID, 12

Fixed effects:
                Estimate Std. Error      df t value Pr(>|t|)
(Intercept)      -1.0397     0.1719 45.7600  -6.050 2.49e-07 ***
SexM              0.5820     0.2431 45.7600   2.395 0.020788 *
SegmentSeg2       0.5996     0.2357 60.2100   2.543 0.013565 *
SegmentSeg3       0.5319     0.2359 60.7300   2.255 0.027757 *
SegmentSeg4       0.6372     0.2361 61.2500   2.699 0.008968 **
SegmentSeg5       0.4491     0.2363 61.2200   1.900 0.062135 .
SegmentSeg6       0.4441     0.2367 59.9200   1.876 0.065523 .
SegmentSeg7       0.5068     0.2371 56.7000   2.137 0.036916 *
SexM:SegmentSeg2 -0.4390     0.3334 60.2100  -1.317 0.192883
SexM:SegmentSeg3  0.2609     0.3336 60.7300   0.782 0.437145
SexM:SegmentSeg4  2.4787     0.3338 61.2500   7.425 4.22e-10 ***
SexM:SegmentSeg5  1.2017     0.3342 61.2200   3.595 0.000648 ***
SexM:SegmentSeg6  0.4340     0.3347 59.9200   1.297 0.199747
SexM:SegmentSeg7  0.2082     0.3354 56.7000   0.621 0.537138
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) SexM   SgmnS2 SgmnS3 SgmnS4 SgmnS5 SgmnS6 SgmnS7 SM:SS2 SM:SS3 SM:SS4 SM:SS5
SM:SS6
SexM        -0.707
SegmentSeg2 -0.690  0.488
SegmentSeg3 -0.694  0.491  0.500
SegmentSeg4 -0.698  0.494  0.500  0.501
SegmentSeg5 -0.702  0.496  0.500  0.501  0.502
SegmentSeg6 -0.705  0.499  0.500  0.501  0.502  0.503
SegmentSeg7 -0.708  0.501  0.499  0.501  0.502  0.504  0.505
SxM:SgmntS2  0.488 -0.690 -0.707 -0.354 -0.354 -0.353 -0.353 -0.353
SxM:SgmntS3  0.491 -0.694 -0.354 -0.707 -0.354 -0.354 -0.354 -0.354  0.500
SxM:SgmntS4  0.494 -0.698 -0.354 -0.354 -0.707 -0.355 -0.355 -0.355  0.500  0.501
SxM:SgmntS5  0.496 -0.702 -0.353 -0.354 -0.355 -0.707 -0.356 -0.356  0.500  0.501  0.502
SxM:SgmntS6  0.499 -0.705 -0.353 -0.354 -0.355 -0.356 -0.707 -0.357  0.500  0.501  0.502  0.503
SxM:SgmntS7  0.501 -0.708 -0.353 -0.354 -0.355 -0.356 -0.357 -0.707  0.499  0.501  0.502  0.504
0.505
> plot(allEffects(LMM), main="Themira biloba\n Sex * Segment")
> allEffects(LMM)
 model: scale(Cell_counts) ~ Sex * Segment

 Sex*Segment effect
```

```
   Segment
Sex       Seg1       Seg2       Seg3       Seg4       Seg5       Seg6       Seg7
  F -1.0397028 -0.4401086 -0.5078452 -0.4024772 -0.5906343 -0.5956518 -0.5329328
  M -0.4576699 -0.2971091  0.3350990  2.6582128  1.1930957  0.4203969  0.2573273

> #### 4. pvalues ####
> # For the regression coefficients
> library(lmerTest)
>
> # For the main effects: Likelihood ratio tests
> #drop1(LMM,test="Chisq")
> LMM.1 = lmer(scale(Cell_counts)~Sex+Segment+(1|ID),Data_final)
> drop1(LMM.1,test="Chisq")
Single term deletions

Model:
scale(Cell_counts) ~ Sex + Segment + (1 | ID)
        Df    AIC    LRT   Pr(Chi)
<none>     168.85
Sex      1 197.03 30.188 3.920e-08 ***
Segment  6 209.35 52.502 1.478e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table B.1. Linear mixed model output segment effects of histoblast nest cell counts

Species	Sex	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7
Arch	F	-0.75446	-0.21536	-0.03566	-0.01569	0.044212	0.243882	0.243882
Arch	M	-0.16211	0.086809	-0.20604	0.174663	0.203948	0.057524	0.203948
Asep	F	-0.44063	0.320226	0.554431	0.034489	-0.14364	-0.87585	-0.36197
Asep	M	-0.60405	0.258836	0.523554	0.434128	-0.08459	-0.81539	-0.21401
Dicra	F	-0.13139	0.38583	0.320204	-0.28778	-0.73106	-0.58035	-0.53628
Dicra	M	0.398508	0.838354	0.737351	0.436755	-0.2308	-0.48091	0.164651
Malb	F	-0.41099	-0.01742	0.060693	-0.05948	0.159837	-0.03545	-0.06249
Malb	M	-0.26528	0.022012	0.001733	0.667575	0.130169	0.10651	-0.25176
Marm	F	-0.69365	0.731807	0.099905	-0.16461	-0.43648	-0.48056	-0.65691
Marm	M	-0.1591	1.181854	0.89713	0.566484	-0.0397	-0.23655	-0.25699
Nnit	F	-0.52542	-0.44608	-0.46591	-0.456	-0.53782	-0.55765	-0.83534
Nnit	M	-0.23286	-0.24587	0.284087	2.667977	0.789874	0.092558	-0.48761
Oluc	F	-0.69292	-0.2882	-0.16697	0.139834	-0.58089	0.270535	0.610359
Oluc	M	-0.6891	-0.02494	0.27567	0.220122	-0.07069	0.256142	0.818081
Pdik	F	-0.87654	-0.34989	0.202775	0.573384	0.04673	-0.18734	-0.29137
Pdik	M	-0.85573	-0.01379	0.949932	0.843642	0.026224	-0.15978	-0.42991
Slat	F	-0.63911	0.040596	-0.35368	-0.4378	-0.06124	-0.38346	-0.16957
Slat	M	-0.28824	0.11136	0.129941	0.438828	0.800122	0.499528	-0.0322
Spun	F	-0.14439	0.870486	0.204995	0.071897	0.204995	0.021985	-0.50209
Spun	M	-0.4189	0.47951	-0.05288	0.130127	-0.20262	-0.2858	-0.37731
Tbil	F	-1.0397	-0.44011	-0.50785	-0.40248	-0.59063	-0.59565	-0.53293
Tbil	M	-0.45767	-0.29711	0.335099	2.658213	1.193096	0.420397	0.257327
Tfla	F	-0.5487	-0.25933	-0.06774	-0.21543	-0.30998	-0.25301	-0.50989
Tfla	M	-0.47179	-0.24945	0.462307	1.991847	0.437837	-0.11835	-0.54739
Tluc	F	-0.12148	-0.22757	-0.24952	-0.08125	-0.48729	-0.45802	-0.47632
Tluc	M	-0.11173	0.457301	0.213431	1.79696	0.229689	-0.32308	-0.39462
Tmin	F	-0.57007	-0.11376	-0.09266	-0.29031	-0.48023	-0.46169	-0.72069
Tmin	M	-0.6508	0.046605	0.651934	2.698509	0.379597	-0.24437	-0.40769
Tput	F	-0.07635	-0.23519	-0.11258	-0.35501	-0.01226	-0.45811	-0.49713
Tput	M	-0.37521	-0.2021	0.0271	1.794837	0.373332	0.019785	-0.10944

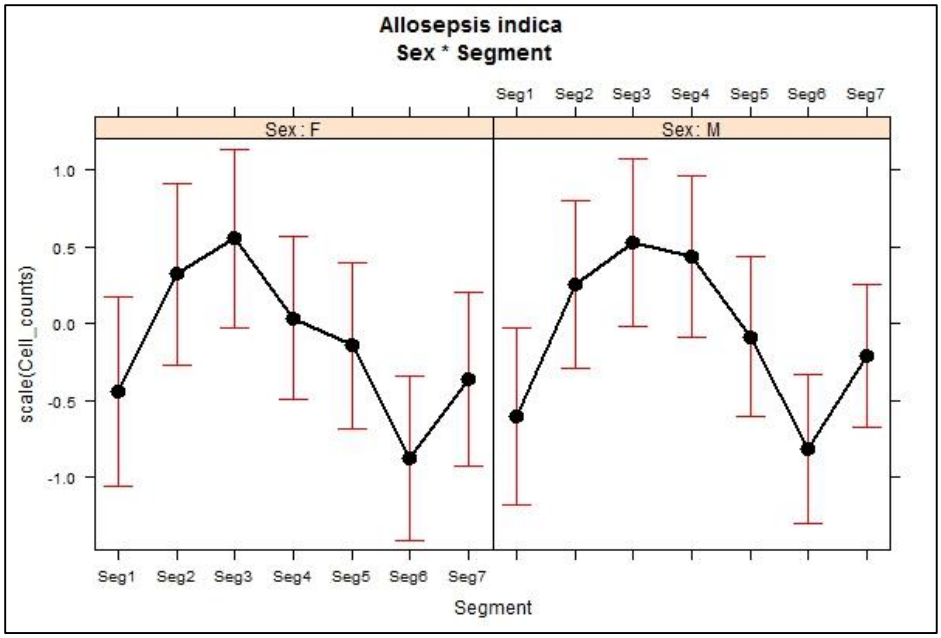

Figure B.1. Sex by segment effects for *A. indica*

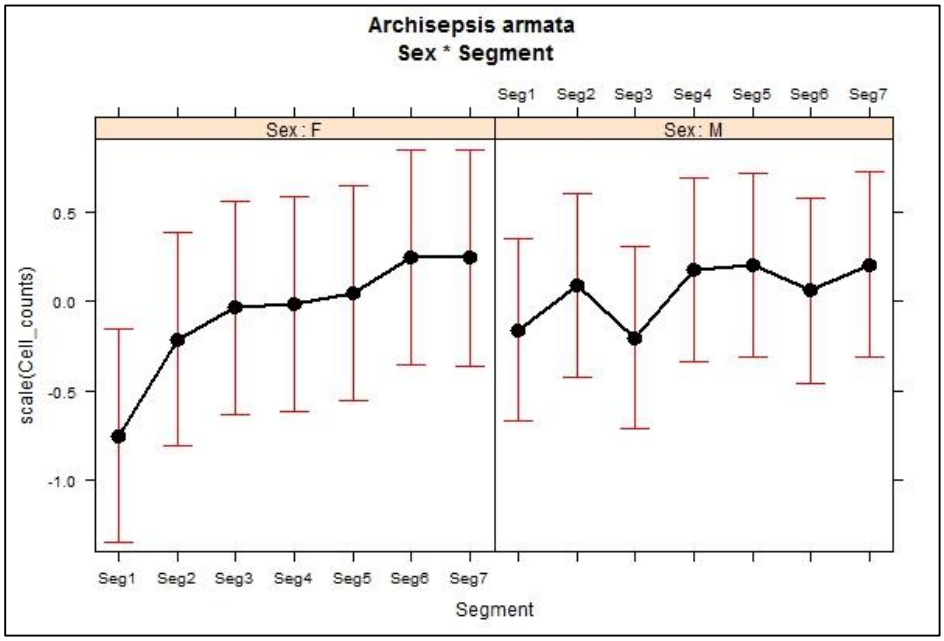

Figure B.2. Sex by segment effects for *A. armata*

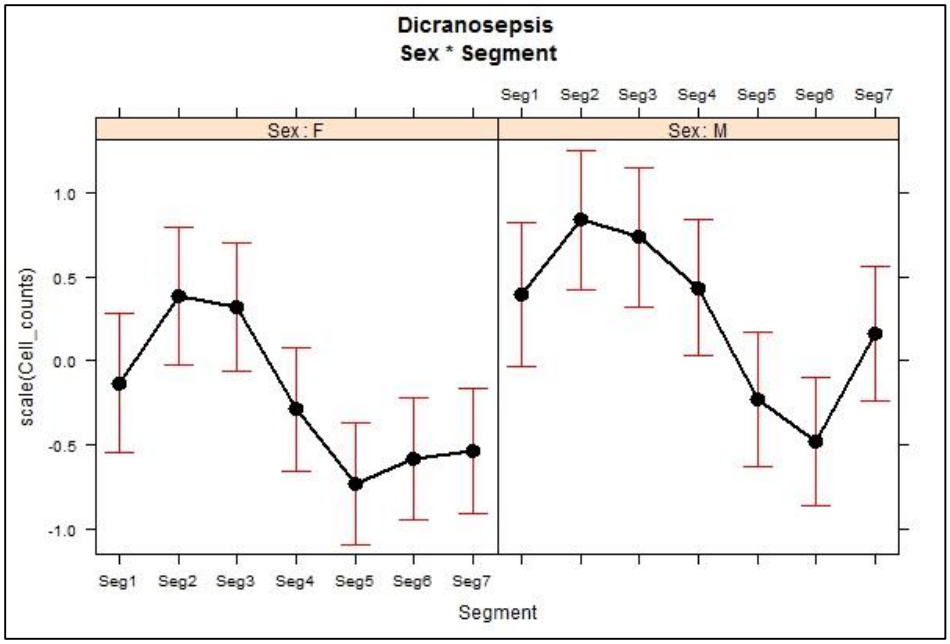

Figure B.3. Sex by segment effects for *Dicranosepsis sp.*

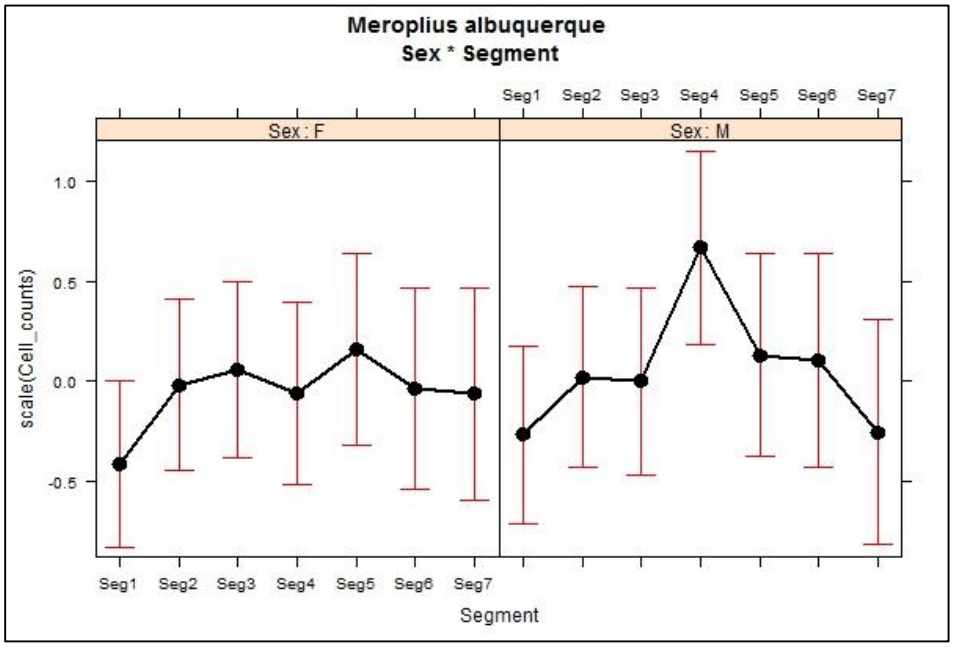

Figure B.4. Sex by segment effects for *M. albuquerque*

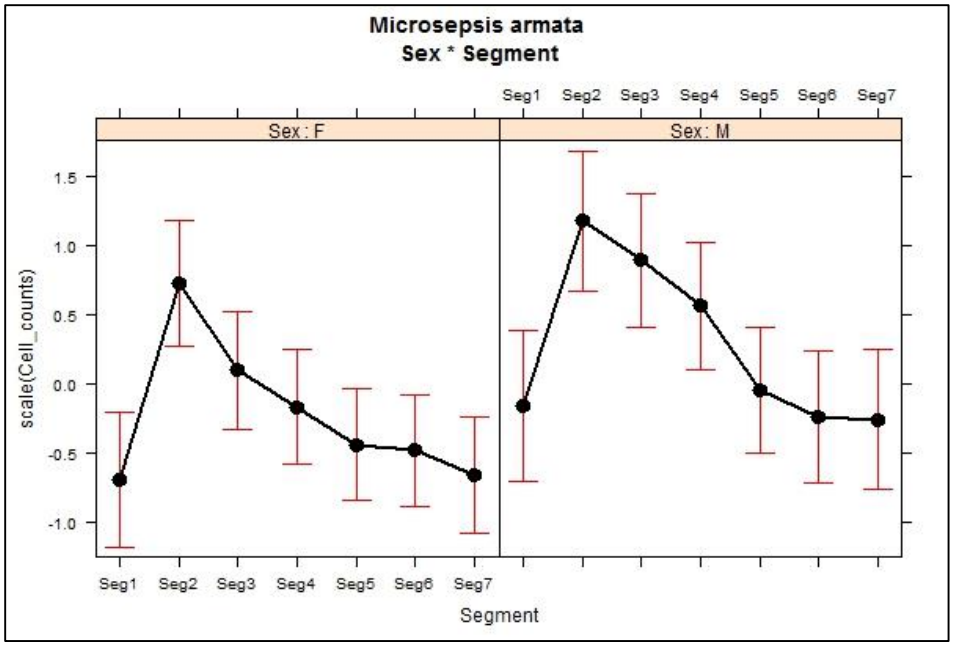

Figure B.5. Sex by segment effects for *M. armata*

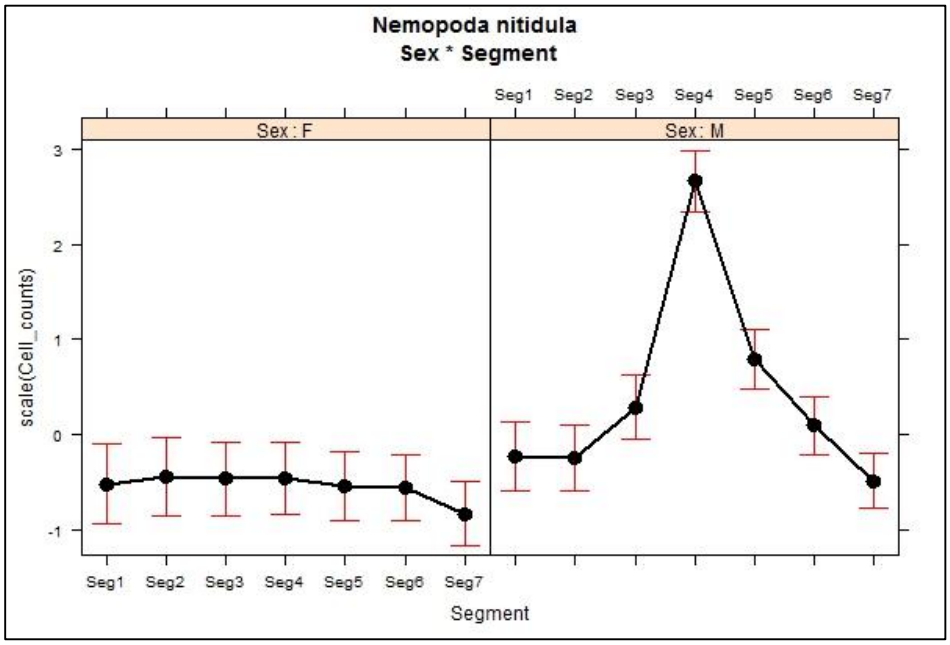

Figure B.6. Sex by segment effects for *N. nitidula*

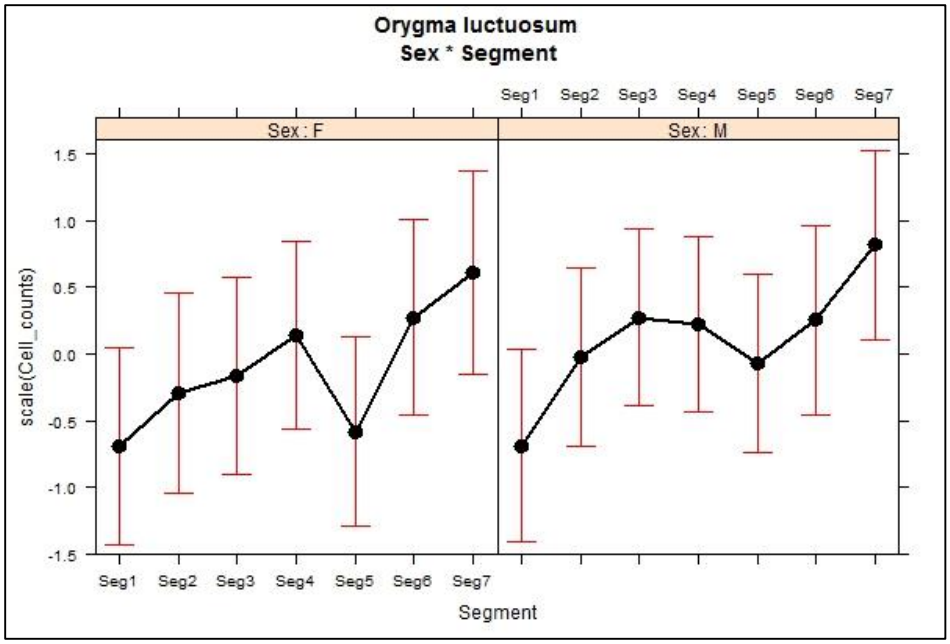

Figure B.7. Sex by segment effects for *O. luctuosum*

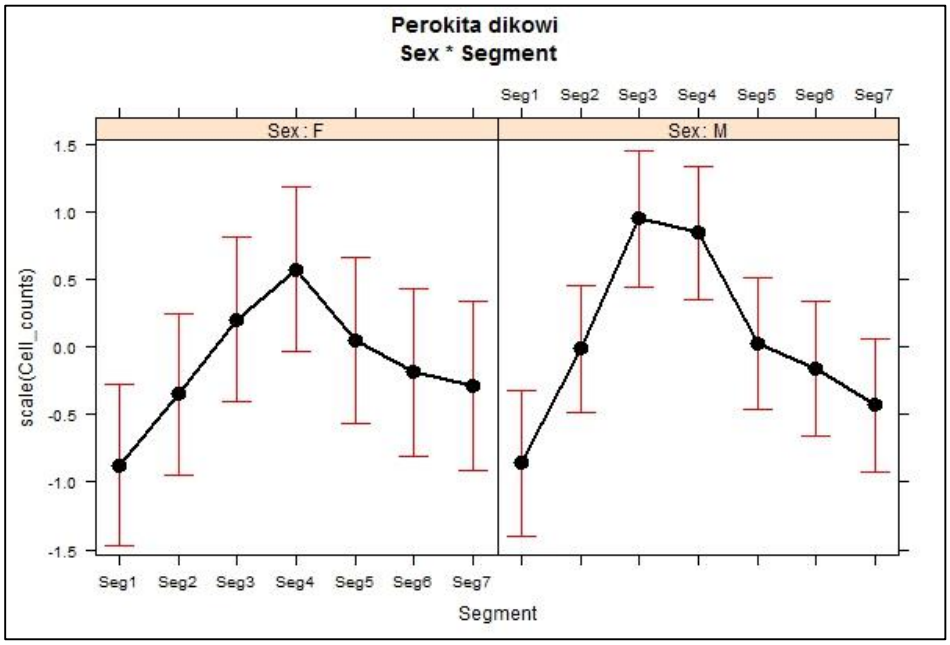

Figure B.8. Sex by segment effects for *P. dikowi*

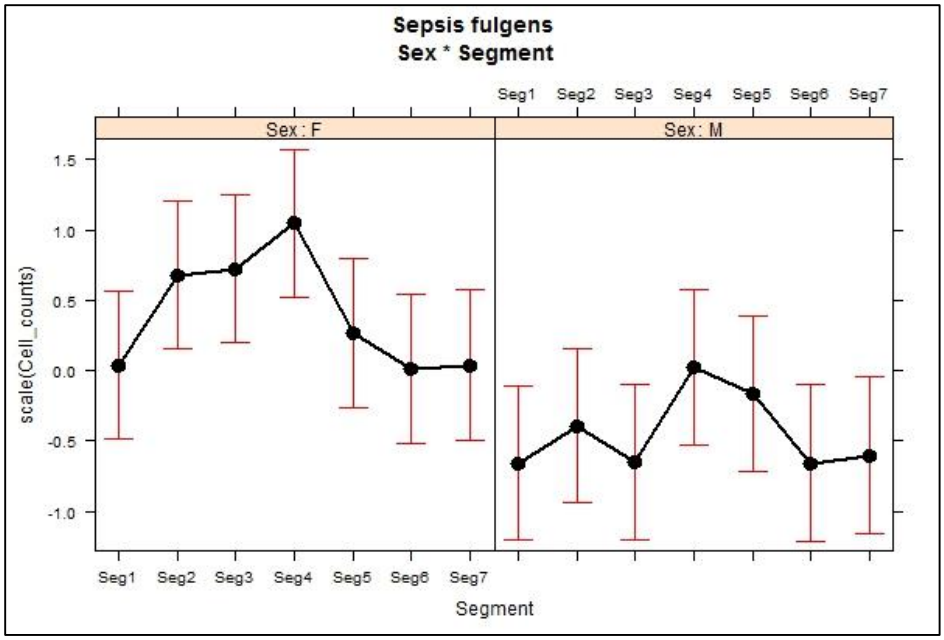

Figure B.9. Sex by segment effects for *S. fulgens*

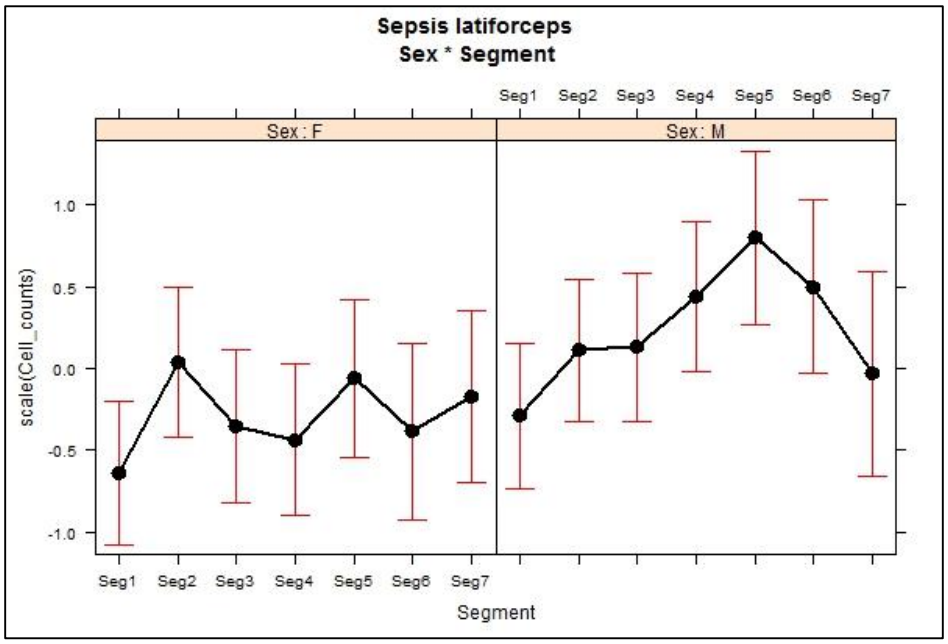

Figure B.10. Sex by segment effects for *S. latiforceps*

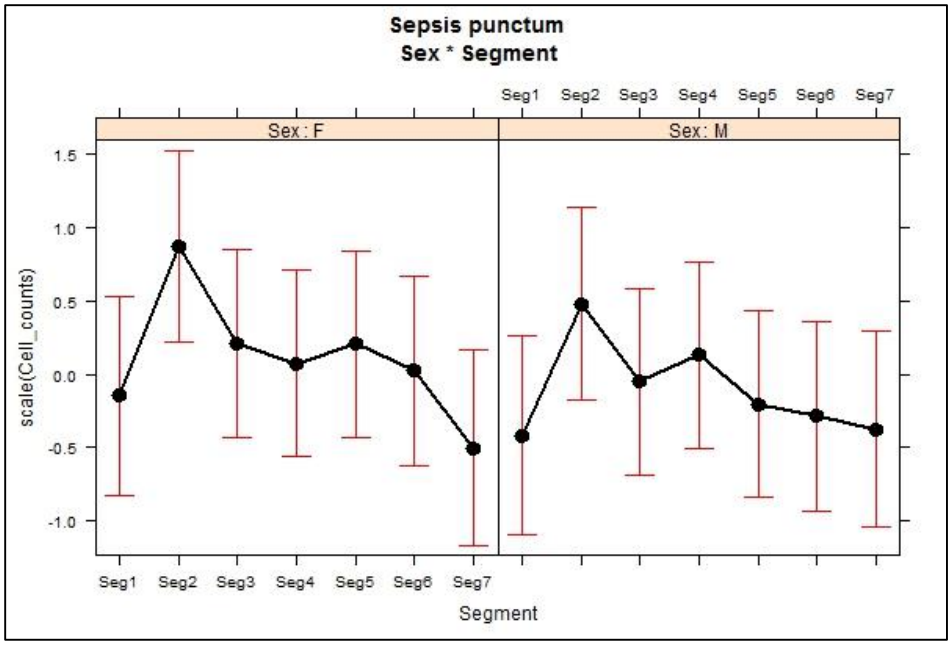

Figure B.11. Sex by segment effects for *S. punctum*

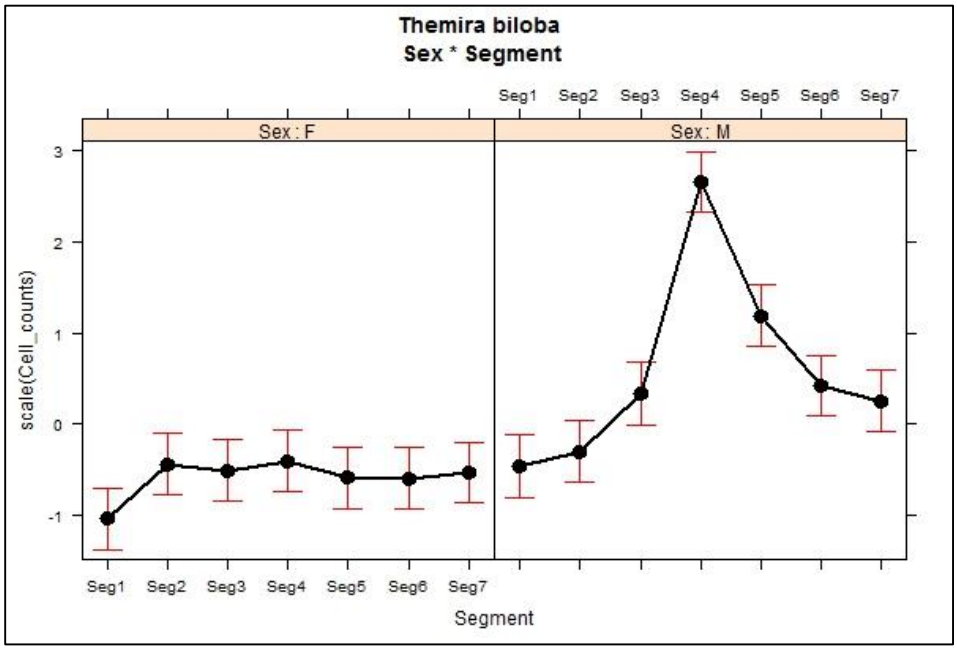

Figure B.12. Sex by segment effects for *T. biloba*

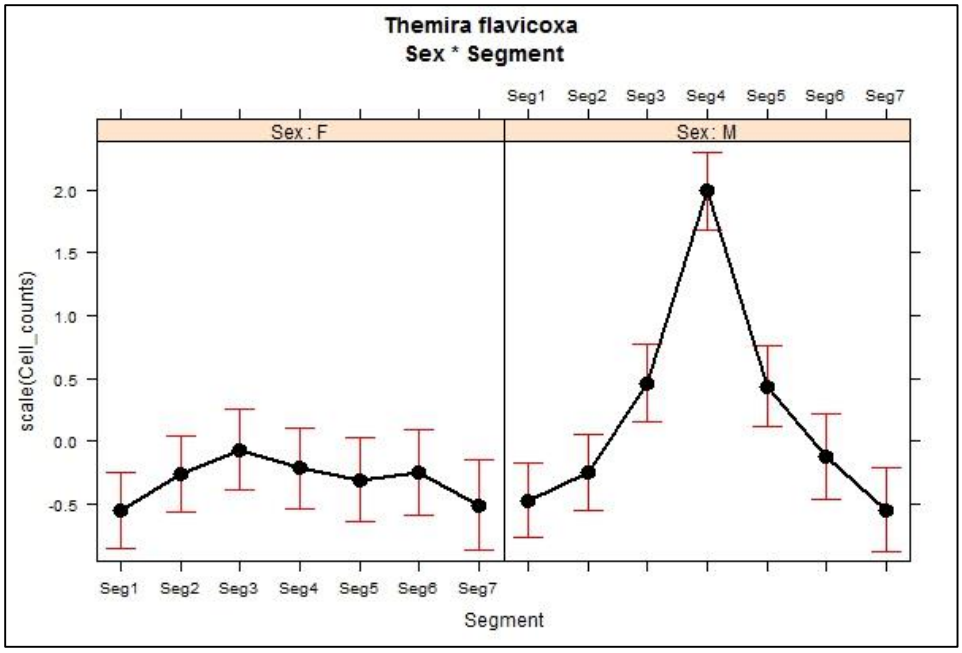

Figure B.13. Sex by segment effects for *T. flavicoxa*

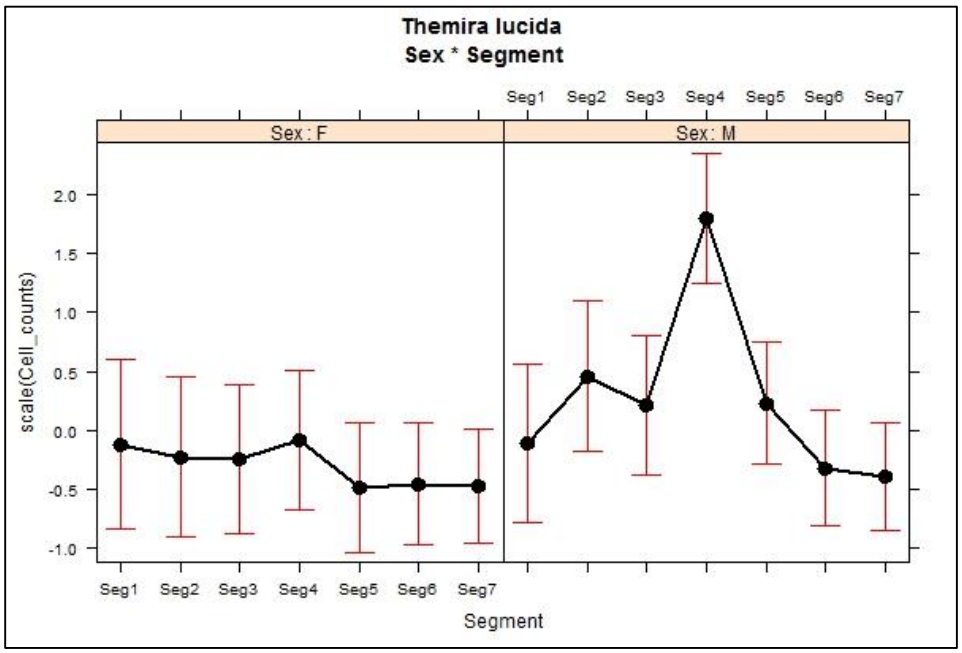

Figure B.14. Sex by segment effects for *T. lucida*

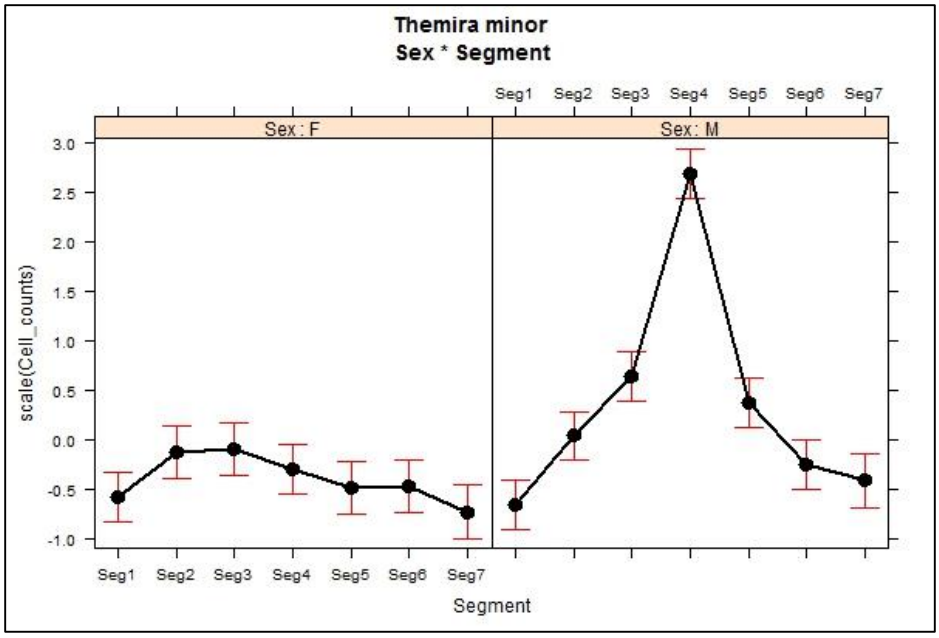

Figure B.15. Sex by segment effects for *T. minor*

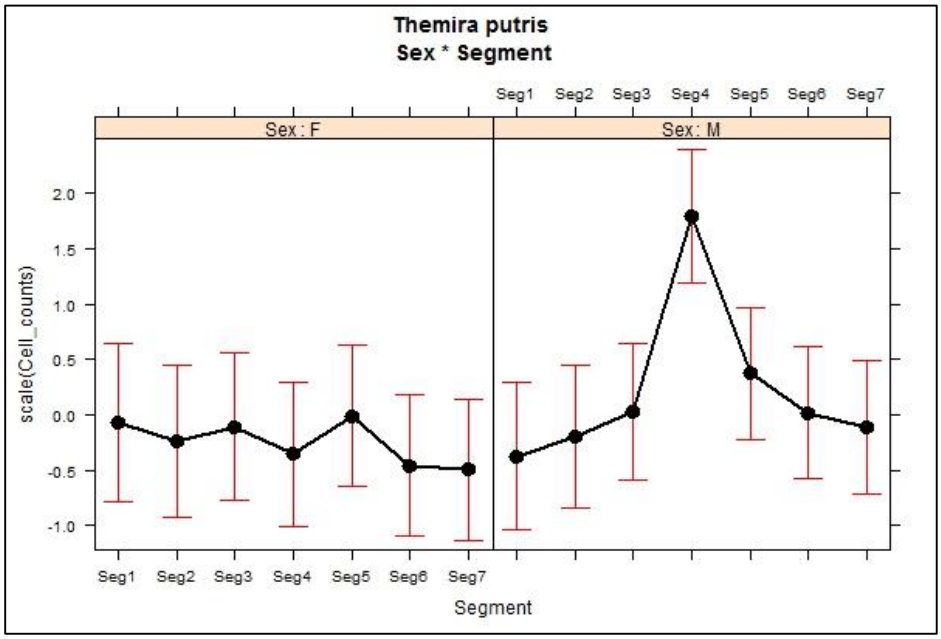

Figure B.16. Sex by segment effects for *T. putris*

APPENDIX C: DIFFERENTIALLY EXPRESSED GENES

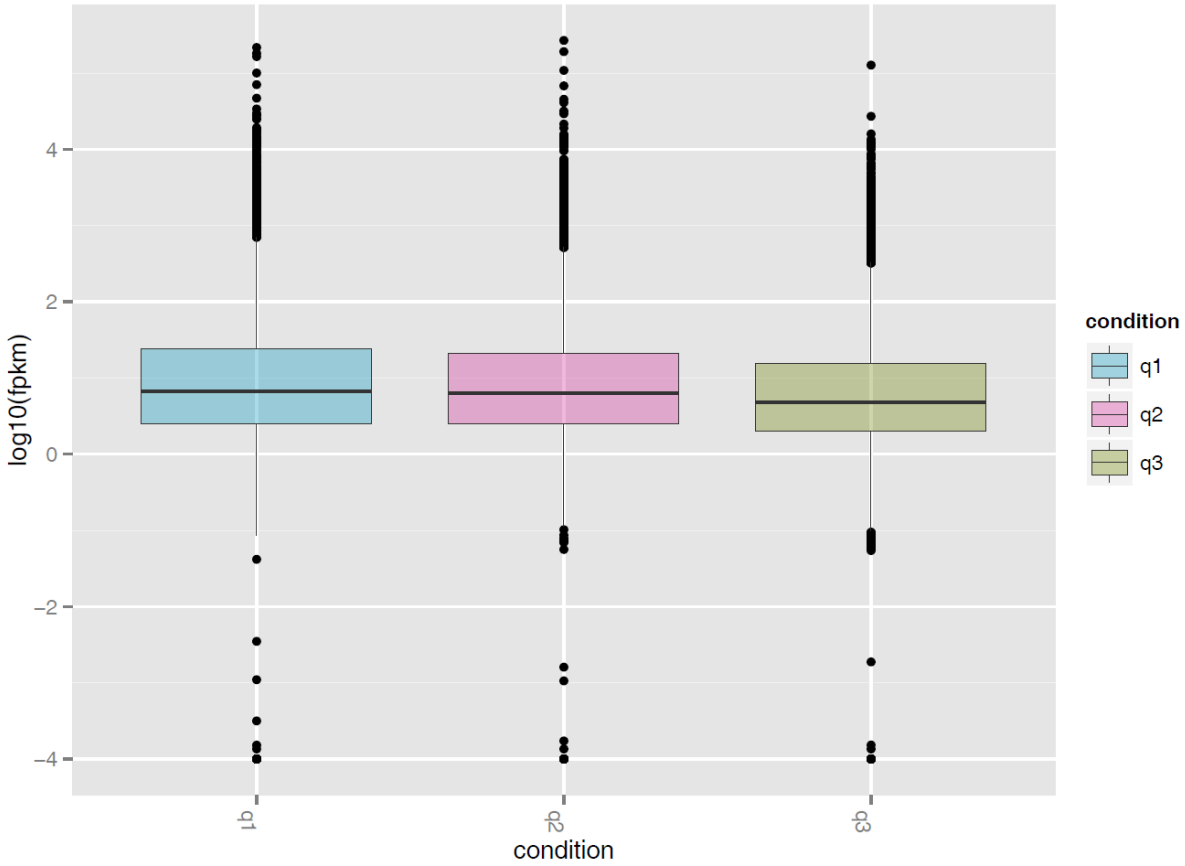

Figure C.1. Boxplot distribution of FPKM values across experimental conditions.

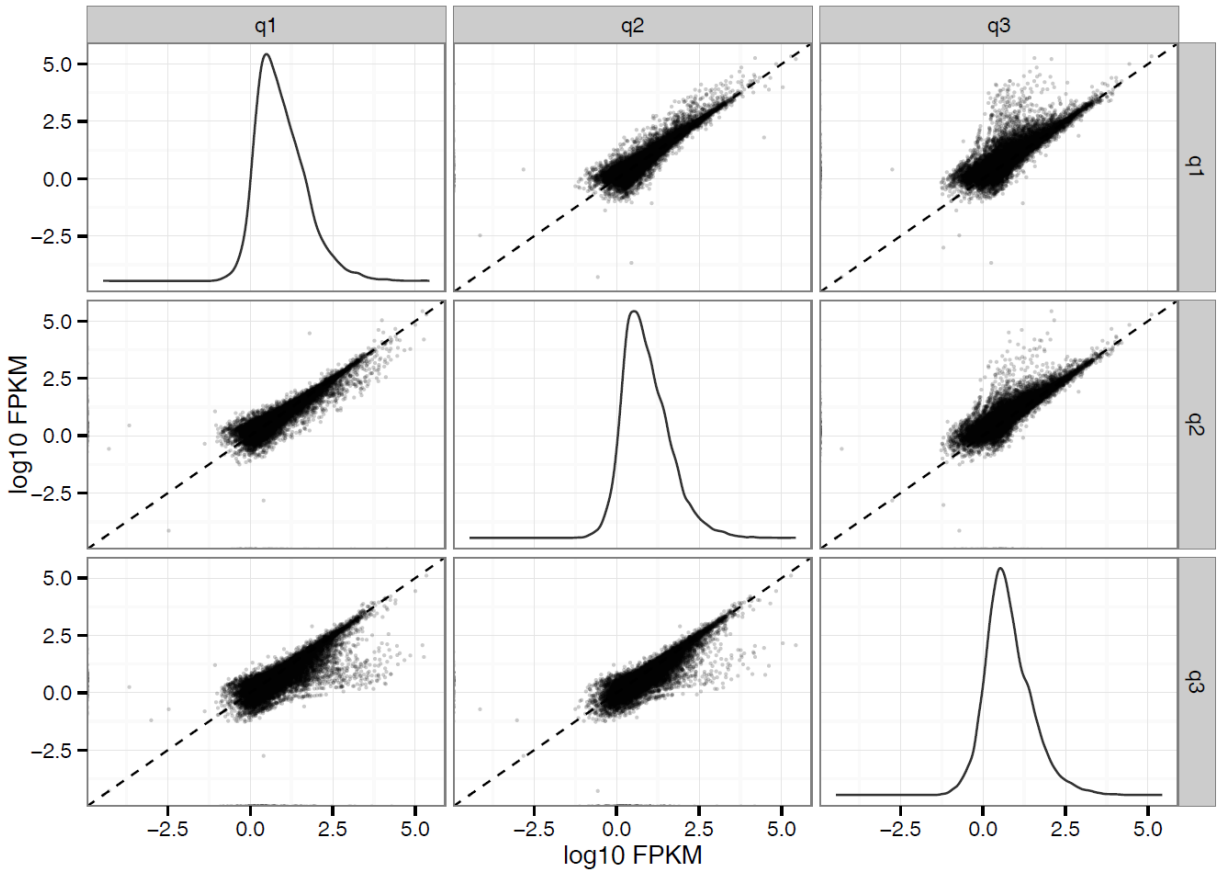

Figure C.2. Distribution of FPKM values between experimental conditions.

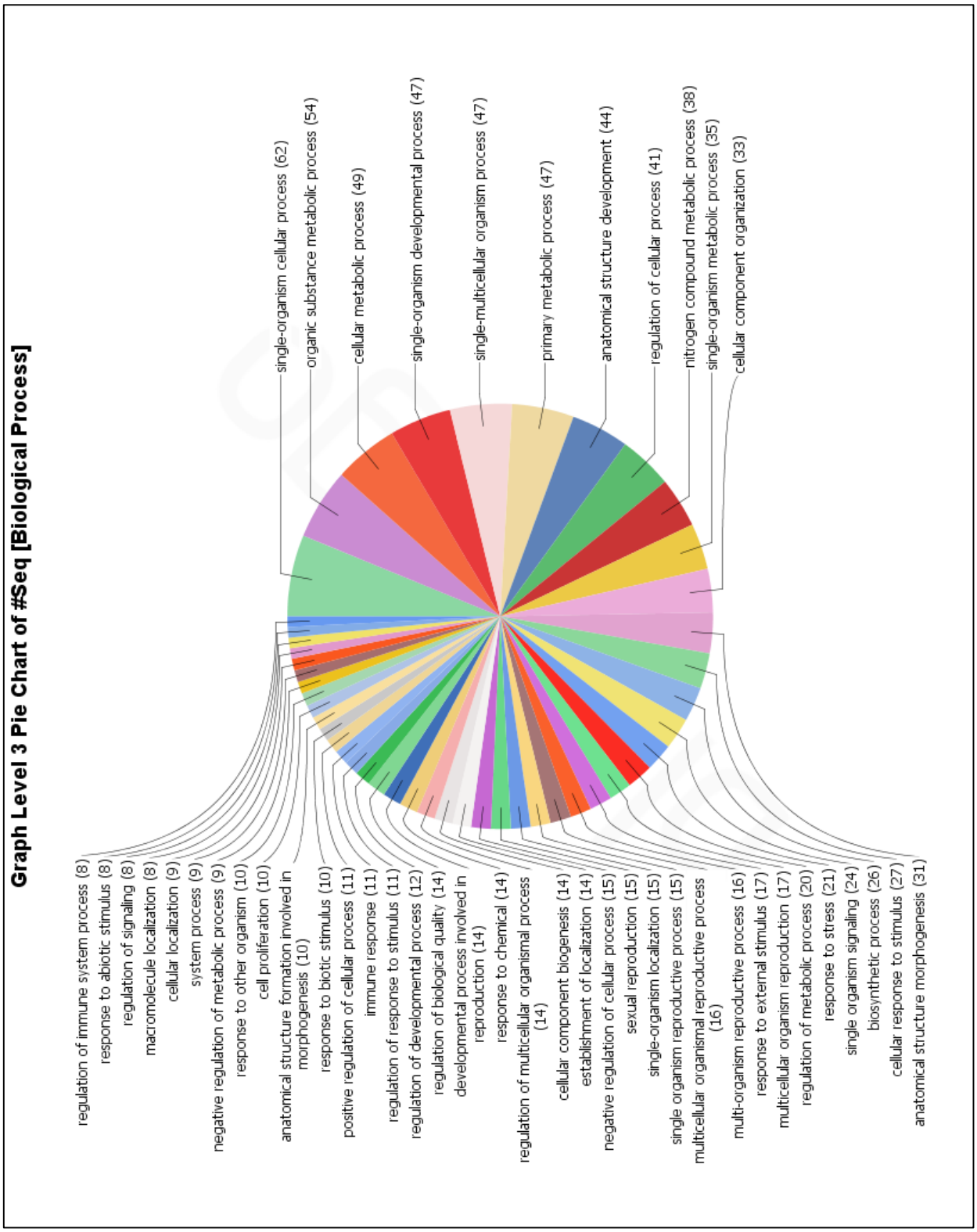

Figure C.3. Distribution of level 3 biological process Gene Ontology classifications of differentially expressed genes in all experimental samples.

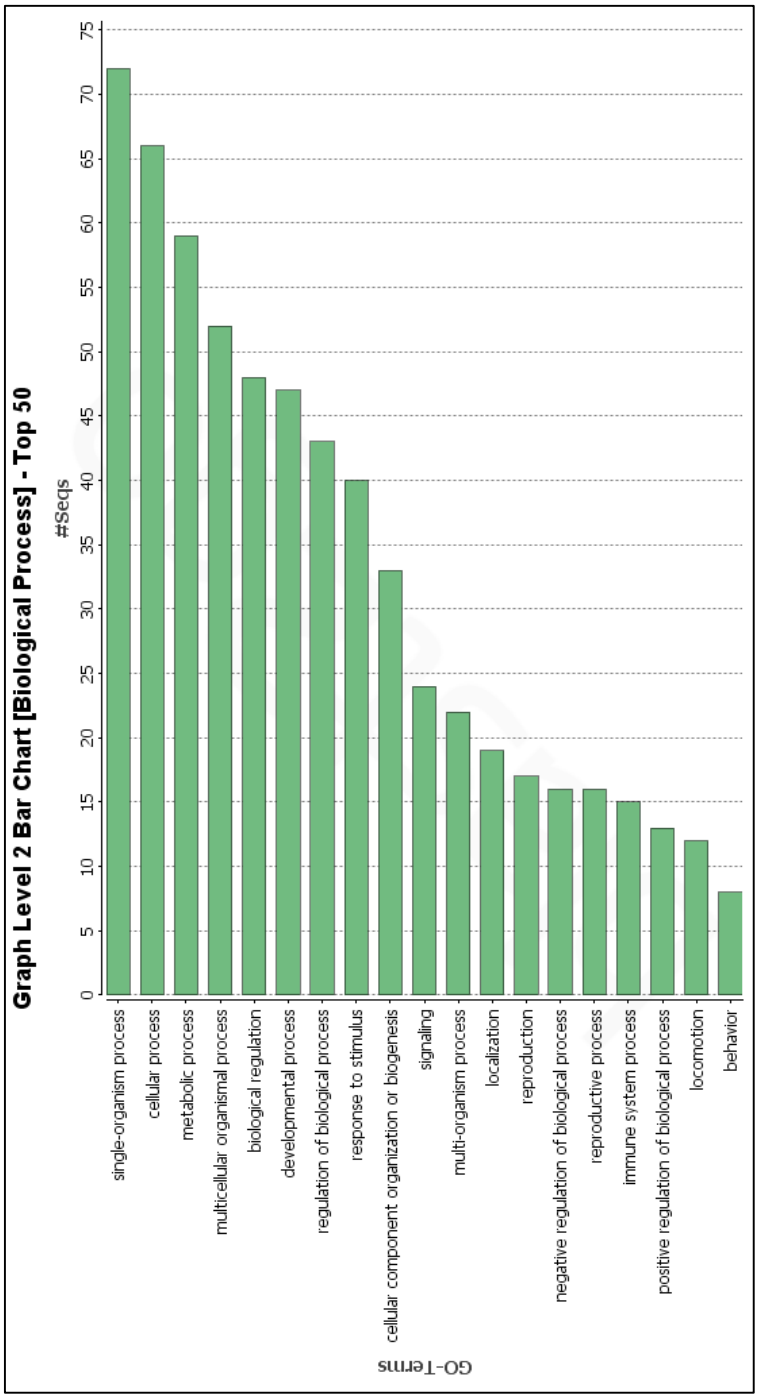

Figure C.4. Level 2 Gene Ontology biological process classifications for differentially expressed genes.

Table C.1. Gene Ontology term scores

GO-term	Score
macromolecule localization	3.5884799999999997
response to organic substance	4.440959999999999
cellular component assembly	6.6768
carbohydrate derivative metabolic process	2.3198976
anatomical structure formation involved in morphogenesis	5.64576
mitotic cell cycle	18.705599999999997
regulation of signaling	3.290112
protein complex subunit organization	3.6576
wing disc development	13.96
phosphorylation	24.215999999999998
proteolysis	11.33856
imaginal disc-derived appendage development	4.056
cellular response to stress	9.451199999999998
organic substance transport	5.1743999999999994
regulation of biological quality	3.5371100159999993
sensory organ development	4.659839999999999
intracellular signal transduction	19.8192
single-organism transport	8.949119999999999
appendage morphogenesis	11.84
negative regulation of cellular metabolic process	2.351232
system process	2.4773759999999996
single-organism organelle organization	15.820799999999998
cuticle development	13.8
embryonic morphogenesis	5.4
regulation of immune system process	3.10434816
regulation of multicellular organismal development	3.0519935999999994
regulation of response to stimulus	5.46806016
neuron projection morphogenesis	18.8896
single-organism biosynthetic process	3.0417791999999997
locomotion	14.839488
movement of cell or subcellular component	2.3656319999999997
defense response	22.696732160000003
oogenesis	19.577915392
regulation of cell communication	3.650112
negative regulation of macromolecule metabolic process	2.3225472
cell morphogenesis involved in neuron differentiation	8.8896
immune response	8.730908160000004
positive regulation of cellular process	4.5100416

Table C.1. Gene Ontology term scores (continued)

GO-term	Score
behavior	2.70432
imaginal disc morphogenesis	6.360000000000001
oxidation-reduction process	10.85248
organonitrogen compound biosynthetic process	3.7157759999999995
cell proliferation	15.171520000000001
cell surface receptor signaling pathway	12.399359999999996
alpha-amino acid metabolic process	4.608
cytoskeleton organization	6.537599999999999
response to abiotic stimulus	3.46752
response to other organism	9.0336
regulation of transcription, DNA-templated	12.168
cell cycle process	8.1456
cellular localization	2.5349759999999995
cell fate commitment	13.08096
lipid metabolic process	10.1312
cellular protein modification process	10.494719999999997
regulation of cellular component organization	4.747199999999999
pattern specification process	7.479359999999999

Table C.2. List of differentially expressed genes

Contig ID	FlyBase ID	BLAST evalue	Common name
comp17949_c1_seq5	FBpp0070557	2.00E-126	Rala-PB
comp18879_c0_seq1	FBpp0070894	8.00E-50	Ubi-p5E-PA
comp6392_c1_seq1	FBpp0071296	0	Hex-A-PA
comp24805_c0_seq1	FBpp0071381	5.00E-50	flw-PA
comp13161_c0_seq1	FBpp0071653	6.00E-46	CG4377-PA
comp14189_c0_seq1	FBpp0071674	1.00E-25	CG30281-PA
comp18352_c0_seq1	FBpp0071694	3.00E-166	CG3264-PA
comp16962_c0_seq1	FBpp0072365	0	Lsp1gamma-PA
comp30826_c0_seq1	FBpp0072687	5.00E-29	RpL23A-PA
comp34511_c0_seq1	FBpp0072687	1.00E-43	slow-PB
comp26375_c0_seq1	FBpp0073148	6.00E-144	Cpr12A-PA
comp12319_c0_seq1	FBpp0073615	4.00E-43	CG14205-PA
comp31625_c0_seq1	FBpp0074499	8.00E-54	Eip75B-PA
comp25248_c0_seq1	FBpp0074916	1.00E-07	spd-2-PA
comp4869_c0_seq1	FBpp0075122	4.00E-09	CG7924-PA
comp5147_c0_seq1	FBpp0075482	8.00E-15	Fbp1-PA
comp10856_c0_seq1	FBpp0075491	1.00E-48	Hml-PA
comp5944_c0_seq1	FBpp0075495	4.00E-169	Gcn5-PA
comp25492_c0_seq1	FBpp0075701	0	GNBP3-PA
comp21097_c0_seq1	FBpp0076237	5.00E-115	CG13676-PA
comp12080_c0_seq1	FBpp0076455	2.00E-09	CG7409-PA
comp16099_c0_seq2	FBpp0076700	1.00E-34	Lcp65Ac-PA
comp12145_c0_seq1	FBpp0076732	2.00E-36	Cpr65Av-PA
comp18740_c0_seq1	FBpp0076765	2.00E-36	Lcp65Af-PA
comp4948_c0_seq1	FBpp0076770	6.00E-45	l(3)mbn-PB
comp14222_c0_seq1	FBpp0076836	0	spo-PA
comp14058_c0_seq1	FBpp0077234	4.00E-28	CG3604-PA
comp15063_c0_seq1	FBpp0077849	3.00E-116	zye-PA
comp5092_c0_seq1	FBpp0077964	8.00E-128	Eip78C-PB
comp14302_c0_seq1	FBpp0078266	8.00E-53	Obp83g-PA
comp14031_c0_seq1	FBpp0078268	2.00E-175	Gasp-PA
comp12072_c0_seq1	FBpp0078795	1.00E-118	obst-E-PB
comp18294_c2_seq1	FBpp0079465	5.00E-14	IP3K1-PA
comp16331_c0_seq1	FBpp0080200	2.00E-136	Rab14-PB
comp18554_c0_seq1	FBpp0080596	0	kel-PB
comp3407_c0_seq1	FBpp0081055	2.00E-12	Mio-PE
comp5614_c0_seq1	FBpp0081565	2.00E-12	alphaTub85E-PA
comp11189_c1_seq1	FBpp0081757	2.00E-26	CG14687-PA
comp14019_c0_seq1	FBpp0082040	6.00E-150	CG4115-PA

Table C.2. List of differentially expressed genes (continued)

Contig ID	FlyBase ID	BLAST evalue	Common name
comp6061_c0_seq1	FBpp0082326	4.00E-81	ems-PA
comp21685_c0_seq1	FBpp0083513	2.00E-15	CG31176-PA
comp5407_c0_seq1	FBpp0084044	0	Ppox-PA
comp5610_c0_seq1	FBpp0084482	4.00E-78	grass-PB
comp31286_c0_seq1	FBpp0084585	4.00E-17	CG5590-PA
comp33526_c0_seq1	FBpp0085802	2.00E-13	Dpt-PA
comp18163_c0_seq2	FBpp0085803	1.00E-21	DptB-PA
comp11582_c0_seq1	FBpp0085951	4.00E-43	CG5726-PA
comp26067_c0_seq1	FBpp0086054	3.00E-55	CG10936-PA
comp33072_c0_seq1	FBpp0086054	3.00E-28	
comp34317_c0_seq1	FBpp0086054	2.00E-06	
comp38059_c0_seq1	FBpp0086160	2.00E-58	CG30460-PB
comp5347_c0_seq1	FBpp0086643	8.00E-44	Cpr51A-PA
comp13251_c1_seq1	FBpp0086657	4.00E-41	cg-PC
comp20445_c1_seq4	FBpp0086659	1.00E-29	CG30069-PA
comp20445_c1_seq5	FBpp0086659	2.00E-39	
comp17369_c0_seq1	FBpp0087094	2.00E-50	SmD3-PA
comp26205_c0_seq1	FBpp0087138	3.00E-166	CG13192-PA
comp18718_c0_seq2	FBpp0087178	0	Tret1-1-PA
comp22551_c0_seq1	FBpp0087518	1.00E-12	Def-PA
comp16346_c0_seq1	FBpp0087842	0	CG2121-PA
comp18536_c0_seq2	FBpp0088120	1.00E-62	Gadd45-PA
comp9644_c0_seq1	FBpp0088362	8.00E-23	CG10638-PA
comp19950_c0_seq1	FBpp0088679	2.00E-37	CG18619-PA
comp15338_c0_seq4	FBpp0088895	6.00E-28	CG9932-PA
comp21297_c0_seq1	FBpp0088899	8.00E-09	Tm1-PA
comp22782_c0_seq1	FBpp0088899	2.00E-38	
comp13591_c0_seq2	FBpp0089363	7.00E-11	bl-PC
comp16732_c0_seq4	FBpp0099646	2.00E-134	GstS1-PC
comp18312_c0_seq1	FBpp0111307	1.00E-11	CG34199-PA
comp39705_c0_seq1	FBpp0111536	1.00E-06	CG34383-PE
comp12896_c0_seq1	FBpp0111664	7.00E-28	CG34448-PA
comp40066_c0_seq1	FBpp0111713	1.00E-47	ec-PB
comp42135_c0_seq1	FBpp0111714	1.00E-10	ec-PC
comp17170_c0_seq2	FBpp0111740	6.00E-24	sdt-PG
comp10416_c0_seq2	FBpp0112047	5.00E-22	dar1-PB
comp18672_c0_seq1	FBpp0271892	1.00E-172	CG8213-PB
comp16646_c0_seq1	FBpp0271945	0	Cubn-PB
comp6677_c0_seq1	FBpp0288391	4.00E-44	alpha-Est10-PB
comp20576_c1_seq1	FBpp0288885	0	CG42269-PE

Table C.2. List of differentially expressed genes (continued)

Contig ID	FlyBase ID	BLAST evalue	Common name
comp14241_c0_seq1	FBpp0289423	7.00E-18	ITP-PE
comp12693_c0_seq2	FBpp0290353	2.00E-37	Tfb5-PB
comp37813_c0_seq1	FBpp0290679	6.00E-10	Mio-PM
comp5373_c0_seq1	FBpp0291850	2.00E-16	CG42673-PC
comp19936_c1_seq8	FBpp0292595	6.00E-106	br-PL
comp19631_c1_seq1	FBpp0297101	0	CG17374-PC
comp5126_c0_seq1	FBpp0297136	2.00E-58	l(2)06225-PD
comp20385_c0_seq19	FBpp0297484	7.00E-36	CG8086-PH
comp19677_c2_seq1	FBpp0297663	2.00E-133	obst-A-PB
comp5323_c1_seq1	FBpp0301053	7.00E-25	Unr-PC
comp17074_c0_seq1	FBpp0301282	6.00E-56	Hsp23-PB
comp13409_c0_seq1	FBpp0301738	8.00E-09	CG4297-PD
comp31831_c0_seq1	FBpp0303232	2.00E-08	CG13784-PF
comp12199_c0_seq1	FBpp0303242	0	drl-PB
comp17802_c0_seq2	FBpp0303668	2.00E-172	Pez-PB
comp21092_c1_seq1	FBpp0304115	1.00E-20	CG12111-PB
comp6109_c0_seq1	FBpp0304323	3.00E-06	crc-PE
comp16703_c0_seq1	FBpp0304441	0	CG42255-PB
comp18757_c0_seq2	FBpp0304646	2.00E-122	nrv1-PB
comp17970_c0_seq2	FBpp0304874	3.00E-09	spirit-PD
comp19236_c0_seq1	FBpp0304919	1.00E-44	Cpr62Bc-PB
comp18684_c0_seq1	FBpp0304923	7.00E-08	sls-PS
comp20476_c0_seq1	FBpp0304924	6.00E-65	sls-PT
comp19494_c0_seq1	FBpp0304925	7.00E-74	sls-PU
comp6238_c0_seq2	FBpp0305084	6.00E-23	W-PB
comp24184_c0_seq1	FBpp0305169	2.00E-06	nocte-PD
comp19887_c0_seq2	FBpp0305337	0	AcCoAS-PD
comp34610_c0_seq1	FBpp0305406	1.00E-20	CkIIbeta-PK
comp26313_c0_seq1	FBpp0305460	7.00E-40	dlg1-PT
comp20517_c0_seq1	FBpp0305733	6.00E-29	Lcp4-PB
comp23286_c0_seq1	FBpp0305758	9.00E-47	Lcp65Ad-PB
comp11676_c0_seq2	FBpp0306622	0	Gprk2-PB
comp18258_c0_seq1	FBpp0306886	9.00E-08	CG44085-PO
comp23132_c0_seq1	FBpp0307599	2.00E-28	CG9932-PD
comp14282_c0_seq1	FBpp0308230	2.00E-18	CG44242-PB
comp5120_c1_seq1	FBpp0308978	2.00E-09	mam-PD
comp19530_c0_seq1	FBpp0309085	4.00E-70	Pkcdelta-PE
comp5808_c0_seq1	FBpp0309103	2.00E-29	Sec16-PH
comp15985_c0_seq1	FBpp0309462	8.00E-17	CG15152-PB
comp32339_c0_seq1	FBpp0309961	5.00E-07	CG6106-PB

Table C.2. List of differentially expressed genes (continued)

Contig ID	FlyBase ID	BLAST evalue	Common name
comp18774_c0_seq2	FBpp0310038	4.00E-11	CG12560-PC
comp19622_c0_seq1	FBpp0310074	1.00E-64	Fbp2-PC
comp20173_c1_seq2	FBpp0310165	4.00E-72	Roc1a-PD
comp20173_c1_seq3	FBpp0310165	2.00E-07	
comp15688_c0_seq1	FBpp0310459	4.00E-148	Sb-PB
comp22324_c0_seq1	FBpp0311086	1.00E-39	ATPsyn-beta-PD
comp18245_c0_seq1	FBpp0311256	0	CG13907-PB
comp19036_c0_seq1	FBpp0311298	2.00E-79	CG34417-PW
comp24852_c0_seq1	FBpp0311376	1.00E-07	Hsc70-3-PE
comp15338_c0_seq1	FBpp0311492	1.00E-12	lid-PF