UNDERSTANDING A MAJOR FOREST PEST: GENE FAMILY EVOLUTION IN WOOD-BORING BEETLES AND ASSEMBLY OF THE *DE NOVO* SOUTHERN PINE BEETLE (*DENDROCTONUS FRONTALIS*) TRANSCRIPTOME

A Thesis

by

SHELBY L. LANDA

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Claudio Casola
Committee Members,	Giridhar Athrey
	Heath Blackmon
Head of Department,	Kirk Winemiller

December 2021

Major Subject: Ecosystem Science and Management

ABSTRACT

The increasing availability of genomic data necessitates improved methods for whole-genome comparison. The study of complex phenotypes, especially, will benefit from comparative methods that are comprehensive across whole genomes. Here, I add to the growing amount of genomic insect pest data with a *de novo* assembly of the southern pine beetle (SPB), *Dendroctonus frontalis*, transcriptome. The SPB, a wood-boring beetle (WBB), is a major forest pest that is responsible for $1.5 billion of losses in natural and agricultural stands of Loblolly (*Pinus taeda L.*) and other yellow pine stands throughout the southeastern United States. I show that this transcriptome, assembled with RNAseq data from males, females, and larvae, contains 94.1% of the BUSCOs for the Endopterygota superorder, and indication of a nearly complete gene set. Next, I introduce methods for comparison of gene family changes across the genomes of multiple species of wood-boring beetles, which belong to independently evolved WBB lineages, and non-wood-boring beetles, with two goals: First, I aimed at testing hypotheses regarding the extent of convergent gene family changes associated with the independent evolution of WBBs; Second, I sought to identify candidate gene families contributing to the complex wood-boring phenotype. The comparisons of gene families did not show excess convergent loss or gain in the WBB gene family change pairwise comparisons. However, the methods to tests gene family convergence presented here can be applied to any system with convergent phenotypes, high-quality genomic resources and accurate phylogenetic inferences. Furthermore, I identified several gene families

with shared expansions and contractions along the three WBB branches of the beetle phylogeny. The gene families sharing expansions and contractions in WBBs include genes involved in chitin metabolism, organization of the cuticle, cuticle development, immunity, and hormone metabolism. These findings significantly expand the number of genes and biological processes that may contribute to the wood-boring phenotype, beyond well-known examples among chemosensorial families. The integration of available WBB genomic resources with novel data sets, including the SPB gene set, will further improve the ability to understand the genetic underpinning of the wood-boring habit. These efforts should accelerate the discovery of next-generating management strategies for tree-killing outbreaks based on genetic and genomic information.

ACKNOWLEDGEMENTS

First and foremost, I wish to extend my deepest gratitude and appreciation to my committee chair, Dr. Casola. He approached this project with patience and enthusiasm, and I made it to the end due to his daily encouragement. Next, I would like to thank my committee members, Dr. Athrey and Dr. Blackmon, for their expertise and upbeat guidance.

I also wish to convey gratitude to the funding sources that made the project possible. Thank you to the Ecosystem Science and Management Harold Maxwell Graduate Assistantship and the College of Agriculture and Life Sciences Excellence Fellowship for funding my time at Texas A&M. Also, thank you to the Eppley foundation for the funding that made RNA sequencing possible.

Thanks, also, to Dr. Dickens and the Texas A&M High Performance Research Computing Center, and Texas A&M AgriLife Genomics and Bioinformatics Service for the computational and sequencing resources. I would also like to acknowledge our collaborator, Dr. Rieske-Kinney, as the Southern Pine Beetle sequencing would not have been possible without the Rieske-Kinney lab.

Finally, thank you to my dad and my mom, Mark and Amy, for their endless love and support. Thank you to my friends for their patience, to my brother for his positivity, and to my boyfriend, Troy, for always having faith and for the many dinners delivered computer-side. Most finally, I must express eternal gratitude to my cat, Meera, for being my study buddy and my joy during many long, isolated months.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This project was supervised by a thesis committee consisting of Dr. Claudio Casola, committee chair, and Dr. Giridhar Athrey of the Department of Poultry Science and Dr. Heath Blackmon of the Department of Biology.

The data collection and assembly for sections 2.1.4 and 3.1.2 were completed by Terrence Sylvester and James Alfieri under the direction of Dr. Blackmon.

Funding Sources

Graduate study was supported by the Excellence Fellowship from the College of Agriculture and Life Science and the Harold Maxwell Graduate Assistantship from the department of Ecosystem Science and Management.

NOMENCLATURE

ALB	Asian longhorn beetle
CPB	Colorado potato beetle
EAB	Eastern ash borer
MPB	Mountain pine beetle
PCWDE	Plant cell-wall degrading enzyme
WBB	Wood-boring beetle

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Wood-boring beetles (WBBs) include major pests of natural and commercial forests worldwide. Current pest management strategies are often ineffective in containing outbreaks of WBBs, which are expected to increase in intensity and frequency as a result of anthropogenic influences, including changes in local climate. A better understanding of the biology of wood-boring beetles is therefore critical to the development of improved management strategies. While many aspects of the physiology and ecology of WBBs have been extensively studied, the genetic underpinnings of these and other traits shared by WBBs remain largely unknown. Genomic data provide fundamental resources to unveil changes at the level of gene content, sequences and expression that contribute to specify the wood-boring habit. In this project, I first aimed at developing novel genomic resources in bark beetles of the genus *Dendroctonus*, a taxon largely populated by species with the potential to generate large tree-killing outbreaks. Second, taking advantage of the convergent evolution of the wood-boring habit across several beetle lineages, I identified gene families that experienced parallel expansions and contractions in WBBs, and I tested if these changes occurred more frequently in WBBs compared to other beetles.

1.1. Novel transcriptomic and genomic resources in the southern pine beetle, *Dendroctonus frontalis*

Conifer-killing beetles are the subject of much research due to their ecological and economic impact. The genus *Dendroctonus* contains 19 species of wood-boring beetles, representing the majority of conifer-killing beetles in the world (Six and Bracewell 2015). However, while the i5k project (Evans et al. 2013) reflects 20 available annotated whole genomes belonging to the order Coleoptera, only a single species within the genus *Dendroctonus* (*D. ponderosae*) has been published (Keeling et al. 2013). Well-annotated genome assemblies are vital for further investigation into the molecular mechanisms underlying the wood-boring phenotype.

The Southern Pine Beetle (SPB) is a natural and agricultural forest pest that is responsible for over $1.5 billion of damage to forestland, particularly loblolly pine (*Pinus taeda L.*) and other yellow pines, in the southeastern United States, from Texas to New York, which makes up ~80% of forestland in the Southeastern United States (Huggett et al. 2013). The first goal of this project was to generate a robust annotation of the SPB genome. An additional, annotated wood-boring beetle genome will allow for comparative studies of genome architecture and gene family evolution in wood-boring beetles (WBBs). Furthermore, a well annotated genome will help to develop tools for improved management of outbreaks. For instance, RNAi approaches have been shown to work well in bark beetles, including SPB (Kyre et al. 2019). However, RNAi probes designed for SPB genes are known to induce mortality in other *Dendroctonus* species and might affect benign beetles. Using whole-genome data will allow collaborators to

design species-specific probes and thus to more effectively implement RNAi-based strategies in the field, as supported by preliminary analyses using a limited set of SPB ESTs (Casola et al. 2020).

A notable facet of the molecular biology of *Dendroctonus* is the diverse number of chromosomes, and sex chromosome configuration. The configuration of the *Dendroctonus* karyotype is highly variable, with $2n$=30 being the presumed ancestral chromosome number, and $2n$=12 being the smallest karyotype (Six and Bracewell 2015). A number of younger lineages in the genus are marked by fewer chromosome numbers, which indicates multiple chromosome fusions and evolution toward smaller chromosome numbers (Six and Bracewell 2015). The drivers behind the rapid chromosome evolution in *Dendroctonus*, however, remain unknown. An additional, annotated *Dendroctonus* genome, especially one with nearly chromosome-level contig assembly, will be vital to research into chromosome synteny and sex chromosome gene migration in the genus.

The SPB genome was assembled with Nanopore long-read sequences followed by chromosome-level scaffolding using the Dovetail Genomics HiRise scaffolding platform (Putnam et al. 2016). Genome annotation of a chromosome-level assembly will generate data for analyzing chromosome evolution as well as provide highly accurate gene models that can be used to investigate gene family evolution in order to understand the genomic basis of the tree-killing phenotype in *Dendroctonus* and other wood-boring beetles.

1.2. SPB annotation and tree-killing beetle genomes

At the times that the studies discussed below were published there were three available whole genome assemblies of WBBs from three superfamilies of Coleoptera: *Agrilus planipennis* (Buprestoidea), *Anoplophora glabripennis* (Phytophaga; Chrysomeloidea), and *Dendroctonus ponderosae* (Phytophaga; Curculionoidea) [Table 1] (Crook and Mastro 2010; Keeling et al. 2013; McKenna et al. 2016). A fourth, *Ips typographus* (Phytophaga; Curculionoidea), has recently been published (Powell et al. 2020). Thus, the addition of a *Dendroctonus frontalis* genome annotation to the growing number of tree-killing beetle genomes will amount to a 20% increase in the availability of genomic data for ecologically harmful wood-boring beetle pests.

1.3. Gene family evolution in bark-boring beetles

Wood-boring beetles must overcome a suite of physical and chemical obstacles, making the phenotype complex and subject to multiple biochemical pathways, especially those involved in detoxification, metabolism and location of suitable hosts. Gene family evolution has been shown to play a vital role in host preference, host range, and specialization of feeding habits across Coleoptera (Seppey et al. 2019). Given barriers such as a bark layer and host-specific defensive metabolites, the life cycle and feeding habits of bark-boring beetles are among the most specialized and preferential within Coleoptera.

While research into the chemosensory and metabolic genes of individual wood-boring species is extensive (Hu et al. 2009; Negrón and Fettig 2014; Villari et al. 2016),

very few studies have compared data across the three published genomes of WBB species. A recent assessment of the genomic content of chemosensory genes in *D. ponderosae*, *A. glabripennis* and *A. planipennis* found a correlation between host specificity and chemosensory gene number (Andersson et al. 2019), but there have been no genome-wide comparative studies to date to assess possible correlations between gene family changes and the emergence of the complex suite of traits associated with the wood-boring habit. The second goal of this project was to identify gene family gains and losses unique to wood-boring Coleopterans by analyzing gene family contractions and expansions across a number of Coleopteran whole genome assemblies. One of the expectations was that signals of elevated gene family expansions and contractions specific to wood-boring species will occur in those families whose molecular mechanisms are driving the tree-killing phenotype. Further, I aimed to test the hypothesis that the convergent evolution of the wood-boring habit was associated with higher levels of shared (convergent) gene family expansions and contractions in WBBs compared to beetles with different life histories.

The three WBB species with published and available whole genome assemblies and annotations – *A. glabripennis*, *D. ponderosae*, and *A. planipennis* – appear to have independently evolved the wood-boring phenotype. *A. glabripennis* and *D. ponderosae* belong to the superfamilies Chrysomeloidea and Curculionoidea, respectively, in the suborder Phytophaga (herbivorous beetles). The wood-boring habit evolved independently at least once in each of these two superfamilies. *A. planipennis* is a representative of the superfamily Buprestoidea (metallic wood-boring beetles).

Phytophaga and Buprestoidea are separated by 250 million years of evolution [Fig. 1.1]. Genomes of species with no wood-boring phenotype are available for each of these three superfamilies, allowing thorough comparison of evolutionary dynamics of gene families between WBBs and their close relatives [Fig. 1.1]. The three most well-characterized of such genomes belong to the species *L. decemlineata*, *S. oryzae* and *P. pyralis*. By determining the number of genes belonging to gene families across several beetle species (those included in Figure 1.1) I was able to compare gains and losses at each node for each family to all other nodes in the phylogeny. The analysis not only informs whether all WBBs or pairs of WBBs show higher levels of convergence across all families compared to other beetles, but also pinpoints specific families with convergence in WBBs. The biological networks to which those particular gene families belong were also investigated. After a functional annotation the pathways enriched in WBBs were also established.

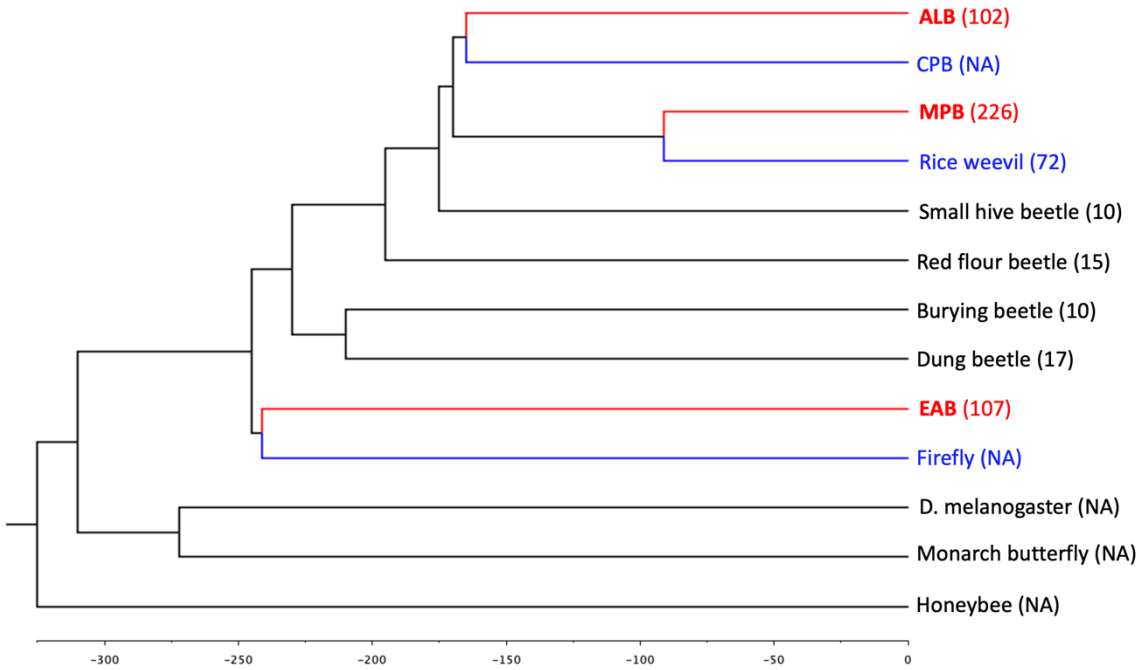

Figure 1.1. Wood-boring species labels represented in bold. Phytophaga; Curculionidae: Mountain Pine Beetle; MPB (*D. ponderosae*) and Asian Longhorn Beetle; ALB (*A. glabripennis*). Buprestoidea: Emerald Ash Borer; EAB (*A. planipennis*). Total number of PCWDE homologs associated with each species in parentheses.

1.4. Previous evidence of convergent evolution of gene families in wood-boring beetles

A survey of 154 Coleopteran transcriptomes and genomes correlates expansions in horizontally acquired plant cell wall-degrading enzymes (PCWDEs) with adaptive radiations and specialized herbivory (McKenna et al., 2019). Aside from clades GH1 and

GH9, which are present in nearly all animals, all other PCWDE genes and duplicates present in Coleoptera are thought to be the result of horizontal gene transfer (HGT) from bacteria or fungi. McKenna et al. 2019 report that the most extensive PCWDE family expansions were found in Phytophaga and Buprestoidea, the most species rich and specialized lineages within Coleoptera. Phylogenetic evidence indicates that ancestors of lineages within these coleopteran taxa experienced separate HGT events involving genes with homologous functional properties that may have facilitated specialization of plant-feeding habits. The presence of three wood-boring lineages with sequenced genomes within Phytophaga and Buprestoidea provides a unique opportunity to investigate the convergence of molecular mechanisms, beyond PCWDEs, that underlie the complex phenotype.

While the dataset for the McKenna and collaborators' study was large, only 18 gene sets of the 154 species studied came from annotated Coleopteran genomes. The rest came from *de novo* transcriptome assemblies. A poorly annotated genome, or a transcriptome assembly without a reference, can result in misleading analysis. For example, across families in the referenced study, those gene sets that came from annotated genomes contained overall higher numbers of PCWDE genes than the *de novo* transcriptome assemblies of their sister species. Novel and improved gene annotations in WBBs and their close relatives allowed me to test the hypothesis that the expansion of some PCWDEs was particularly prominent among wood-boring species.

2. METHODS

2.1. Generation of genomic resources for SPB

2.1.1. Specimens collection

SPB specimens were collected from infested loblolly pine trees in the Homochitto National Forest, MS, in September 2019.

2.1.2. Transcriptome data

RNA-seq data from 8 individuals, 3 female, 4 male and 1 larva, were used to assemble the SPB transcriptome. Quality assessment of the data was performed using FastQC (Andrews 2010). TrimGalore (Krueger 2015) was used to remove reads from the dataset that are under the phred threshold of 30 and under the length threshold of 20bp. After low-quality reads were removed, contaminant sequences were identified using FastqScreen. The small size of the organism necessitated extracting RNA from whole-body samples, and contaminant sequences from the gut microbiome or SPB symbionts may have been present.

After contaminants were identified, the RNAseq reads were mapped to contaminant genomes with the Burrows-Wheeler Alignment tool (Li and Durbin 2009) and filtered according to map quality [Fig. 2.1]. rRNA contamination was also removed by mapping the RNAseq reads to a comprehensive set of Coleopteran rRNA sequences

retrieved from the SILVA rRNA gene database (Quast et al. 2013). The remaining reads should represent only mRNA expressed by female, male, and larval SPB samples.

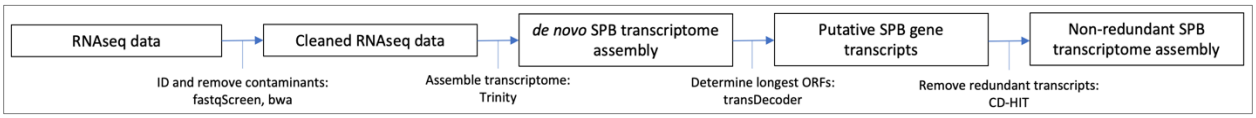

Figure 2.1. Workflow for de novo assembly of D. frontalis transcriptome.

2.1.3. Transcriptome Assembly

Transcriptome assembly was completed using methodology derived from numerous published transcriptome and genome assembly projects using the Trinity *de novo* assembly pipeline (Grabherr et al. 2011). Transcripts were subsequently clustered using the cd-hit-est tool available through the CD-HIT software package (Fu et al. 2012) to remove redundancy. The TransDecoder (Haas and Papanicolaou 2019) pipeline, using BLAST (Camacho et al. 2009) and Pfam (Finn et al. 2014) evidence, was used to identify transcripts in both the full and reduced assemblies that represent the longest open reading frame (ORF). The TransDecoder step filters out smaller isoforms and spurious or chimeric assemblies. The final assembly is a complete, non-redundant set of gene transcripts expressed by *D. frontalis* [Fig. 2.1].

2.1.4. Genome sequencing

To date, i5k reflects 22 whole-genome assemblies (two not annotated) in the order Coleoptera, but only five of which are the product of long-read sequencing technology. The *D. frontalis* genome assembly was carried out by collaborators using Nanopore long read sequencing data and Hi-C scaffolding performed at Dovetail Genomics.

2.1.5. Repeat annotation

Accurate gene prediction requires a library of repeat sequences occurring throughout the target genome for masking of repeats. Currently, there are no Coleopteran repeat libraries available through the open-source repetitive DNA element database, Dfam (Storer et al. 2021). I implemented the RepeatModeler2 (Flynn et al. 2020) and RepeatMasker (Smit et al. 2013-2015) programs in generating a SPB-specific repetitive element library. RepeatModeler2 scans the genome for tandem repeats and low-complexity regions of DNA, and RepeatMasker adds annotation of the identified repeats using Dfam homology. After repeat regions were identified, the output was filtered for proteins with known transposon homology and blasted against the *D. ponderosae* proteome to remove any additional repeats that occur in coding regions. The resulting library consists of non-coding repeats specific to the *D. frontalis* genome.

2.2. Gene family evolution

2.2.1. Sequence datasets

The complete gene sets of three wood-boring Coleopterans, seven non-wood-boring Coleopterans and three outgroup species belonging to the larger Insecta class were used to analyze gene family changes in WBBs [Table 2.1]. Files containing complete transcriptomes and protein translations for each species were obtained and polished before analysis.

Table 2.1. Species included in gene family evolution analysis. Wood-boring species indicated in bold.

Order	Family	Species	Common Name
Coleoptera	Chrysomelidae	*Leptinotarsa decemlineata*	Colorado Potato Beetle (CPB)
Coleoptera	Curculionidae	*Sitophilus oryzae*	Rice Weevil
Coleoptera	**Buprestidae**	***Agrilus planipennis***	**Emerald Ash Borer (EAB)**
Coleoptera	**Cerambycidae**	***Anoplophora glabripennis***	**Asian Longhorned Beetle (ALB)**
Coleoptera	**Curculionidae**	***Dendroctonus ponderosae***	**Mountain Pine Beetle (MPB)**
Coleoptera	Lampyridae	*Photinus pyralis*	Common Eastern Firefly
Coleoptera	Nitidulidae	*Aethina tumida*	small hive beetle
Coleoptera	Scarabaeidae	*Onthophagus taurus*	NA
Coleoptera	Silphidae	*Nicrophorus vespilloides*	NA

Table 2.1. Continued

Order	Family	Species	Common Name
Coleoptera	Tenebrionidae	*Tribolium castaneum*	Red Flour Beetle
Lepidoptera	Nymphalidae	*Danaus plexippus*	Monarch Butterfly
Diptera	Drosophilidae	*Drosophila melanogaster*	NA
Hymenoptera	Apidae	*Apis mellifera*	Western Honey Bee

2.2.2. Fasta sequences polishing

Accurate gene family size prediction requires that the input protein sequences are not redundant or alternatively spliced. Inclusion of these type of sequences, from here referred to as isoforms, in the dataset would artificially inflate the size of the gene family to which the isoform sequences belong. Protein and transcript fasta files and the gff file of each species were downloaded. The gff was used to identify the longest coding sequence/protein per locus and the corresponding transcript IDs, in order to avoid including multiple isoforms/proteins in loci with alternative transcript data [Table 2.2].

Protein sequence files were filtered accordingly in order to include only the longest protein for each gene using the in-house python scripts: isoform_ID.py, parse_aethina_tumida_ncbi.py, and rename_aethina.py [Appendix A, B, C].

Table 2.2. Protein sequence datasets. WBB species in bold.

Species	Total sequences	No isoforms	No fragments	CAFE input
Leptinotarsa decemlineata	19,038	14,000	13,439	10,764
Sitophilus oryzae	26,663	15,057	14,899	10,834
Agrilus planipennis	**15,497**	**15,497**	**14,193**	**9,575**
Anoplophora glabripennis	**22,343**	**22,253**	**20,572**	**11,537**
Dendroctonus ponderosae	**13,457**	**13,088**	**11,881**	**9,796**
Photinus pyralis	15,774	15,773	15,640	10,802
Aethina tumida	18,674	14,076	13,850	11,209
Onthophagus taurus	17,483	17,483	16,111	10,217
Nicrophorus vespilloides	13,516	13,516	12,728	9,303
Tribolium castaneum	16,645	16,626	16,228	10,110
Danaus plexippus	15,128	15,128	NA	9,942
Drosophila melanogaster	21,243	13,784	NA	9,303
Apis mellifera	15,314	15,314	NA	9,186

A separate file with the nucleotide sequences corresponding to the filtered protein sequences was created for each species using the in-house python scripts rename_seq.py, write_seq_file.py [Appendix D, E]. This step also ensured that protein and nucleotide sequences shared the same fasta header, thus simplifying further analyses. While the longest isoform sequence may not correspond to the most commonly

expressed protein, using longer isoforms increased the accuracy of gene family reconstruction.

2.2.3. Removal of incorrect gene duplications

A common gene annotation error results from the fragmentation of a given locus into multiple shorter gene models, which can be included in the same gene family, thus appearing as duplicated gene copies. To minimize this type of annotation error we performed BLASTp searches of proteins from each beetle species against the *Drosophila melanogaster*'s proteome. Genes of a target species whose proteins matched different or slightly overlapping (within 20 amino acids) portions of the same *D. melanogaster* gene were considered gene fragments. We then removed all fragments of a given gene except the one with the highest BLAST score.

2.2.4. Clustering of genes in gene families

The OrthoFinder software (v2.3.8, BLAST+) (Emms and Kelly 2019) was used to cluster the polished set of protein sequences for all species into 'orthogroups' (gene families) based on homology of input sequences. A tab-separated file containing orthogroup names (OG0001, OG0002, etc.) and the number of transcripts in each species belonging to that orthogroup [Fig. 2.2; Orthogroups table], included in the OrthoFinder output, was used as input for the next step in the gene family evolution pipeline – predicting gene family expansions and contractions.

Gene family expansions and contractions across the 13 species and 12 internal phylogenetic nodes [Fig. 1.1] were predicted using the CAFE (v5.0) software. The CAFE algorithm interprets size changes in gene families in a given dataset using a birth-death model while accounting for phylogeny. The ultrametric input tree was based on published phylogenies of beetles (Zhang et al. 2018; McKenna et al. 2019) and evolutionary distances obtained from TimeTree (Kumar et al. 2017). Default CAFE settings, using a single rate parameter, were used with the exception of the *--zero_root* flag. This setting allows inclusion of orthogroups predicted to have a count of zero at the root of the phylogeny in the output. The *--zero_root* option was chosen to retain data that would otherwise have been automatically removed by the CAFE algorithm. The CAFE output of interest is a table, much like the OrthoFinder output, containing the orthogroup names and the quantitative increases or decreases in the number of genes at each node—including the tips, i.e., the species—in the phylogeny compared to the node's most recent ancestor [Fig. 2.2; Expanding/contracting OGs table]. Only OrthoFinder gene families with at least 1 gene in minimum 7 species were used for CAFE analyses.

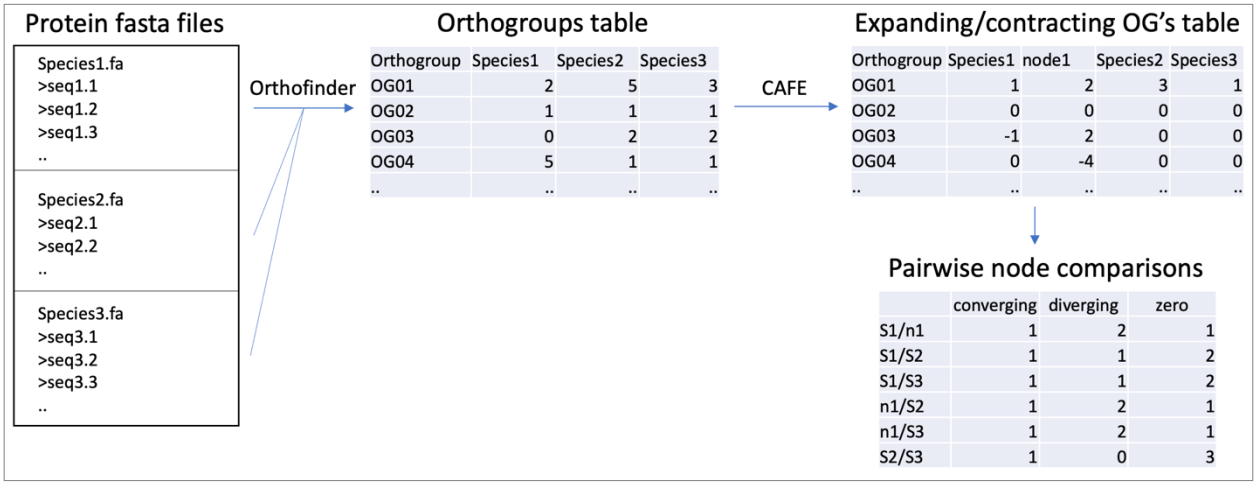

Figure 2.2. Gene family evolution data pipeline using truncated, simulated data. Protein fasta files: all protein coding sequences for each species; Orthogroups table: Orthofinder output detailing number of genes per species in each orthogroup; Expanding/contracting OGs table: CAFE output detailing increases, decreases, no change in orthogroup size in each species and at each internal node; Pairwise node comparisons: number of converging, diverging and no change orthogroups between all species and nodes.

Analysis of the CAFE output was adapted from the methods for assessing adaptive convergence laid out by Castoe and collaborators (Castoe et al. 2019) and Thomas and Hahn (Thomas and Hahn 2015). Their method of pairwise comparison between all pairs of species in the dataset creates a null model for statistically determining excess convergence. For our purposes, each node in the CAFE output –

including the tips – was compared to the others within a pairwise framework. A custom script, cafeGainLoss-expanded.R [Appendix F], was implemented in order to tally the number of *converging*, both nodes expanding or contracting, *diverging*, one node expanding while the other is contracting, or *zero*, neither node shows expansion or contraction, gene family changes between each pair. Similarly, I have applied a test based on the comparison of shared gene family expansions and contractions across any combination of three species, or triads.

2.2.5. PCWDEs analyses

All protein sequences obtained after filtering out gene fragments in each species were queried against the Conserved Domain NCBI database using the Web CD-search batch tool with default parameters (Marchler-Bauer et al. 2015). Sequences that contained a PCWDE domain belonging to the clades reported by McKenna et al. (2019) (Table 2.3) were used to identify gene families with PCWDE genes obtained by OrthoFinder. Some families contained more genes than the number of predicted genes encoding proteins in PCWDE domains. The non-PCWDE members of these families encoded proteins that contained one or multiple domains sharing sequence homology with one or multiple protein domain contained in the PCWDE members. I interpreted this as the result of misannotation of some PCWDE-encoding genes that were joined with another, non-homologous gene to form a chimeric locus. Thus, genes that did not encoded for proteins containing PCWDE domains were manually excluded from PCWDE families. A complete list of reannotated PCWDE is available at

https://github.com/CasolaLab/beetle_convergence_gene_families, file PCWDE gene families.xlsx.

Table 2.3. PCWDE domains retrieved from CD-search runs.

PCWDE clade	Short name	Accession	Superfamily
GH1	Glyco_hydro superfamily	cl23725	-
GH28	Glyco_hydro_28	pfam00295	cl37622
GH28	Glyco_hydro_28 superfamily	cl37622	-
GH32	SacC	COG1621	cl34321
GH32	scrB_fam superfamily	cl36871	-
GH45	Glyco_hydro_45 superfamily	cl03405	-
GH48	Glyco_hydro_48	pfam02011	cl20227
GH48	Glyco_hydro_48 superfamily	cl20227	-
CE8	PRK10531 superfamily	cl30603	-
PL4	RGL4_C	cd10317	cl15687
GH9	Glyco_hydro_9	pfam00759	cl02959
GH9	Glyco_hydro_9 superfamily	cl02959	-

2.2.6. Biological functions of gene families with changes shared by multiple species.

We performed functional enrichment analyses of the gene families that expanded, contracted or were entirely lost in the three wood-boring beetles using the STRING database (Szklarczyk et al. 2019). The analysis was performed using protein sequences from *D. melanogaster* as the target database and, for families with no gene in *D. melanogaster*, from *T. castaneum*, *A. tumida* and *D. plexippus*.

3. RESULTS

3.1. Generation of genomic material for SPB

3.1.1. Transcriptome data, assembly and verification

The raw SPB transcriptomic data, obtained using standard Illumina MiSeq protocol, consists of between 840k and 1.1M paired-end reads for each individual. After quality control and cleaning the reads were assembled into 73,853 mRNA transcripts by the Trinity pipeline. Clustering with CD-HIT reduced that number to 59,289 transcripts belonging to the *de novo* assembled SPB transcriptome. TransDecoder further reduced the transcript numbers to 61,830 and 42,796 for the Trinity and CD-HIT-reduced assemblies respectively [Table 3.1].

Completeness assessment of the de novo SPB transcriptome assembly before and after implementation of CD-HIT shows that the redundancy-reducing software removed nearly 15,000 redundant transcripts without significantly impacting transcriptome completeness. BUSCO (benchmarking universal single copy ortholog) results for the Trinity assembled transcripts and the reduced transcripts identified by CD-HIT are 94.6% and 94.3%, respectively [Table 3.1]. The BUSCO database used for comparison is a set of 2,124 single copy Endopterygota orthologs. Number of missing orthologs in each assembly can be found in Table 3.1 as well. The percentage of complete duplicate BUSCOs in the SPB assemblies is higher than the other coleopteran and outgroup gene

sets [Fig. 3.1]. The excess duplicates should be further reduced by transcriptome annotation using the SPB reference genome.

Table 3.1. SPB transcriptome assembly quality metrics. Trinity assembly is the full set of transcripts assembled by Trinity, and CD-HIT reduced is the assembly after removing redundant transcripts.

	Trinity Assembly	CD-HIT reduced
Number of transcripts	73,853	59,289
BUSCO %	94.6	94.3
Missing BUSCOs	90	99
Longest ORFs	61,830	42,796
Longest ORFs – BUSCO %	94.1	93.8
Longest ORFs – missing BUSCOs	99	103

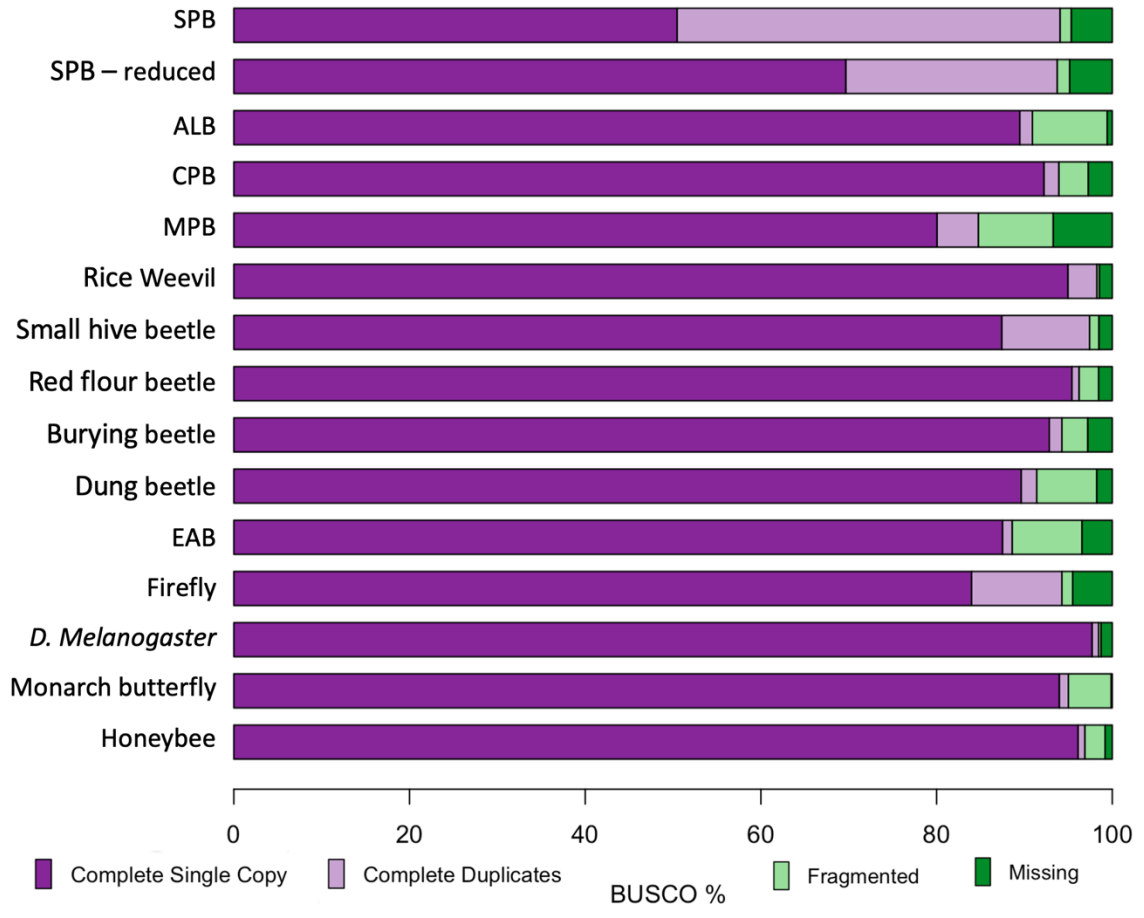

Figure 3.1. BUSCO (Benchmarking universal single-copy ortholog) percentage results for SPB (Southern pine beetle; *D. frontalis*) transcripts assembled by trinity and reduced by CD-HIT as well as all other gene sets used for gene family evolution analysis. (ALB: Asian Longhorn Beetle; CPB: Colorado potato beetle; MPB: Mountain pine beetle; EAB: Emerald ash borer).

3.1.2. Genome verification

The SPB genome assembly was carried out by collaborators in the Blackmon Lab, Texas A&M University, using Nanopore long-read sequencing, the Flye (Kolmogorov et al. 2019) genome assembly pipeline and CANU (Koren et al. 2017)

genome assembly. A final, combined assembly including most Flye contigs and a select number of CANU contigs, putatively containing transcript sequences not represented in the Flye assembly, has been used to produce a chromosome-level scaffolding of the SPB genome through Hi-C sequencing and scaffolding by Dovetail Genomics.

After the initial genome was assembled and cleaned of contaminants, a nucleotide BLAST was run with the transcriptome as the query and the genome as the target. A total of 2,427 out of the 73,853 Trinity transcripts had no hits under the e-value threshold of 10^{-5} against the assembled genome, and 2,381 of the 59,289 reduced sequences had no hits. The combined BUSCO and BLAST results provide high confidence that the collected RNAseq data was sufficient for producing a robust SPB transcriptome assembly, and that the CD-HIT reduction of redundant sequences did not sacrifice transcriptome completeness.

The 'No hit' blast results were not removed from the transcriptome assembly because the transcripts were compiled from male, female and larval samples while the genome assembly was produced from only female samples. The Blackmon lab is currently working on a Y chromosome assembly, and the 'no hit' transcripts will be queried against the potential Y chromosome contig(s).

3.1.3. Repeat library

The SPB repeat library is comprised of 957 repetitive elements. 219 sequences were identified by RepeatMasker as LINEs (Long Interspersed Nuclear Elements), 146 sequences belong to DNA transposons, 60 are long terminal repeats (LTRs), 6 sequences

are repeats found within rRNA coding sequences, 2 are rolling-copy (RC) Helitrons and a single simple repeat was also detected [Table 3.2]. The remaining 585 sequences of the library were not annotated by the initial RepeatModeler/RepeatMasker steps and were thus labelled 'Unknown'. Completeness of the repeat library would benefit from additional annotation of the unknown sequences.

Table 3.2. Summary of repeat element classification in the *de novo D. frontalis* repeat library, derived from the *D. frontalis* genome.

Classification	Number of Sequences
LINE	157
DNA Transposon	146
LTR	60
rRNA	6
RC – Helitron	2
Simple Repeat	1
Unknown	585
Total	957

3.2. Gene family evolution

3.2.1. Gene families

Genes from the ten beetles and three outgroup species were clustered in 14,880 gene families by OrthoFinder (file: Orthofinder.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). I analyzed 8,234 families – those present in seven or more species – using CAFE to determine patterns of gene gains and losses across the phylogeny (file: CAFE base change.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). I then applied the pairwise and triad approaches to determine if wood-boring beetles experienced higher rates of convergent expansions and contractions of gene families compared to other assortments of beetle species.

3.2.2. Pairwise test results

Species pairwise analyses of gene family expansions and contractions show a nearly eight-fold variation in convergence/divergence (C/D) ratios, from 0.29 between *Sitophilus oryzae* and *Dendroctonus ponderosae* (MPB) to 2.27 between *Nicrophorus vespilloides* and the honeybee [Fig. 3.1; Appendix G and H]. The three pairs of WBBs showed C/D values between 0.94 (MPB-ALB) and 1.48 (ALB-EAB) [Fig. 3.2]. C/D values were comparable in the corresponding pairs of species sister to WBBs [Fig. 3.2; Appendix G and H].

The same pattern was observed in families with convergent expansions alone [Appendix G, Fig. 3.3]. The highest number of convergent expansions of gene families was in the *Aethina-Photinus* pair with 236 families, whereas *Sitophilus-Aethina* paired showed the highest CE/D ratio of 1.08 [Appendix G, Fig. 3.3].

Convergent gene family contractions, normalized by divergence changes, were higher in WBBs compared to their sister species with the exception of MPB/ALB. However, contractions C/D values in WBB pairs were not higher than many other pairs of beetles [Appendix H, Fig. 3.4]. The *Drosophila*-honeybee pair showed the highest number of shared contracted gene families (375), whereas *Nicrophorus* and honeybee shared the highest proportion of convergent contractions per divergent changes (1.91) [Appendix H, Fig. 3.4].

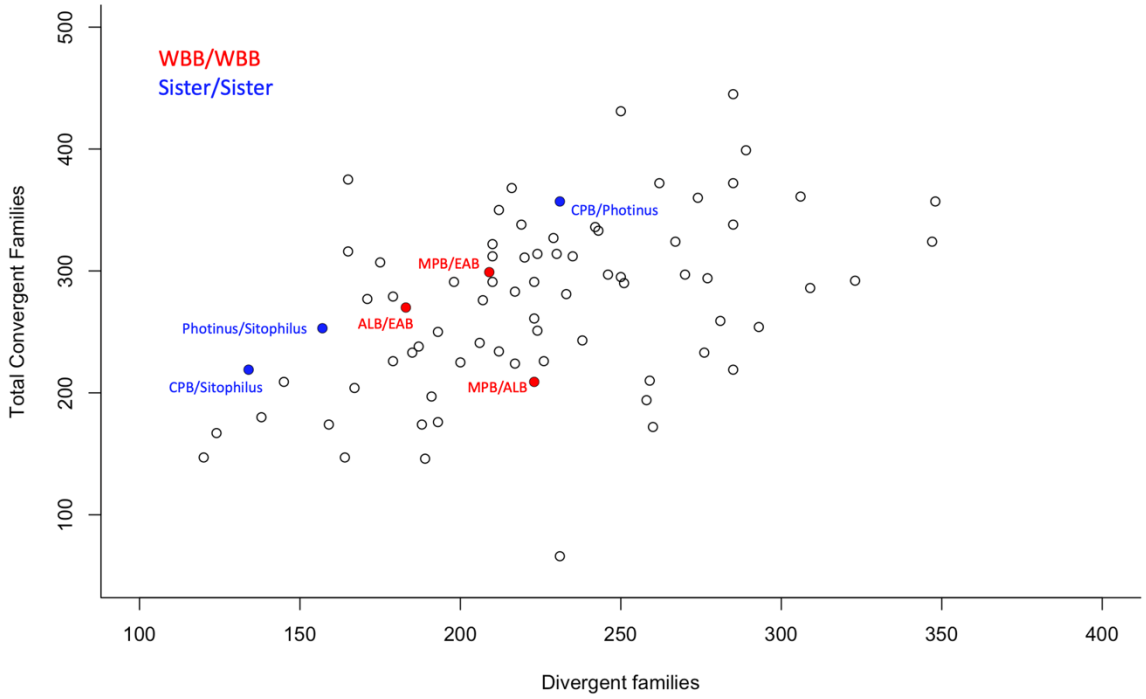

Figure 3.2. Pairwise total convergent gene families and divergent gene families (D) for all species. Comparisons between WBBs and between each sister species are highlighted.

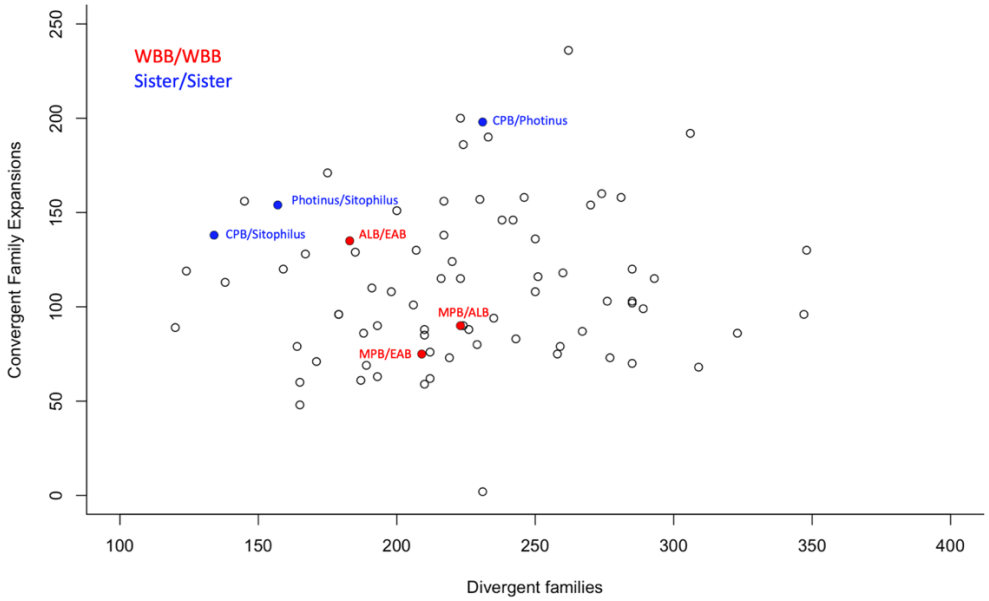

Figure 3.3. Pairwise convergent gene family expansions and divergent gene families for all species. Comparisons between WBBs and between each sister species are highlighted.

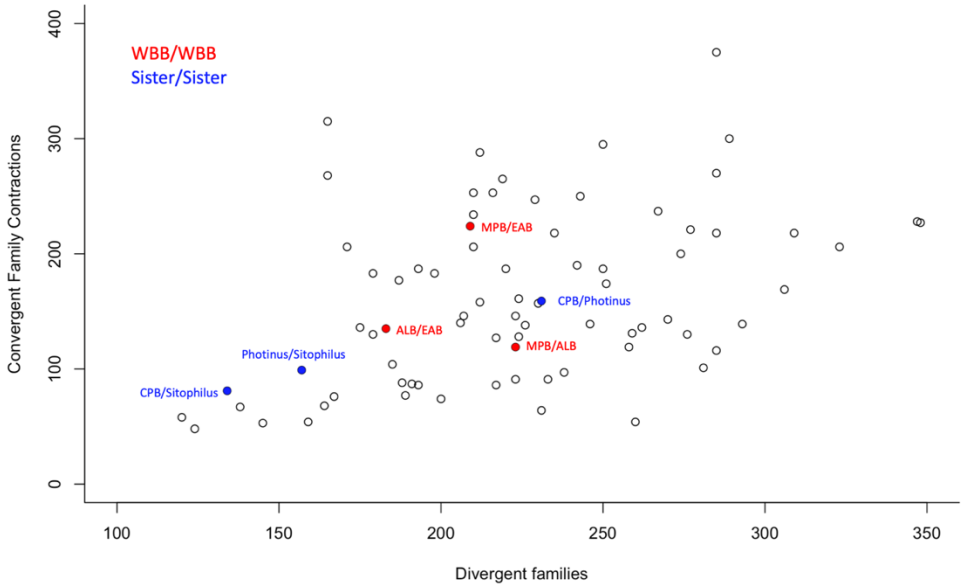

Figure 3.4. Pairwise convergent gene family contractions (CC) and divergent gene families (D) for all species. Comparisons between WBBs and between each sister species are highlighted.

3.2.2.1. Triad test results

The comparison triads of species revealed that WBBs share 16 and 40 gene family expansions and contractions, respectively. These values were comparable to species triads of other beetles (Appendix I). In sister species of WBBs, we identified 38 expansions and 37 contractions. To better compare these values, we normalized them by the average number of families with expansions and contractions in the three species of each triad (Appendix I). The normalized values for expansions and contractions in the WBBs triad fall towards the upper end of the distribution of other triads for the

convergent expansions, indicating that WBB experienced more convergent expansions per number of expansions than most other beetle triads (Fig. 3.5). However, the triad formed by sister species of WBBs showed higher convergent expansions per number of expansions than the three WBBs (Fig. 3.5). Conversely, WBBs convergent contractions per average contraction were lower than most other beetle triads (Fig. 3.5). The triad with the highest number of convergent expansions per number of expansions was formed by two WBB species, ALB and EAB, and the dung beetle *Onthophagus taurus*. The triad with the highest number of convergent contractions per number of contractions included the two outgroup species *D. melanogaster* and *A. mellifera* and the burying beetle *Nicrophorus vespilloides.*

The results of the pairwise and triad comparisons remain to be ranked using statistical methods. I am still exploring the correct way to quantify the level of convergence. The Jaccard similarity index, a statistic for measuring similarity between datasets, is an option for assessing the overlap between each species comparison in order to determine convergence level. A framework to test significance of pairs or triads based on the Jaccard similarity index has been recently developed and could be applied to the beetle datasets (Chung et al. 2019).

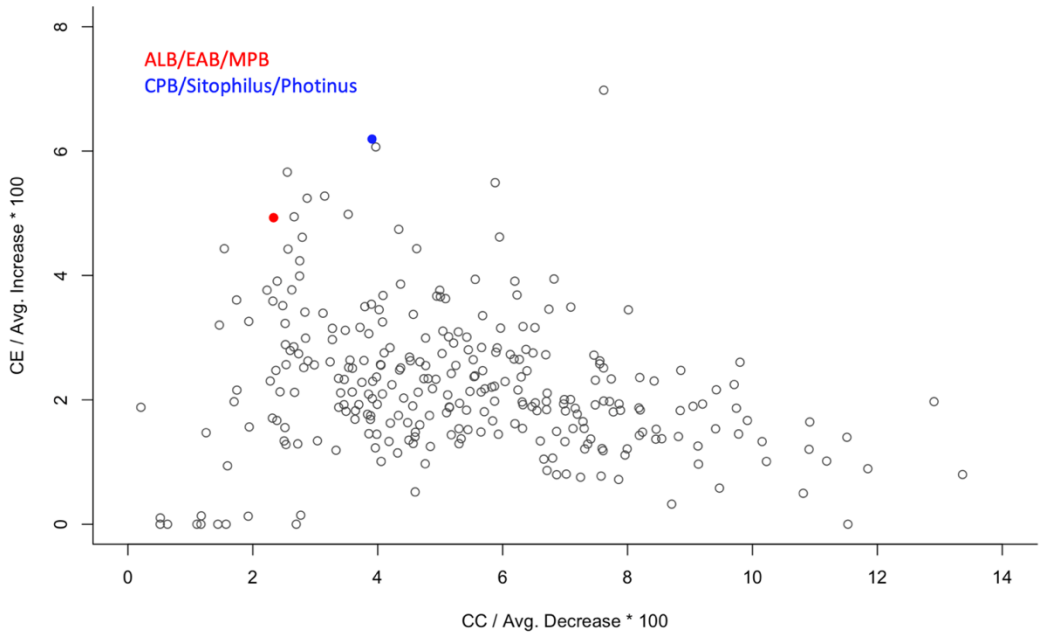

Figure 3.5. Normalized convergent gene family expansions (CE) and contractions (CC) for all species triads. The WBB and sister species triads are highlighted.

3.2.3. Convergent evolution of PCWDEs in beetles

Herbivorous beetles have been shown to contain more genes encoding for plant cell-wall degrading enzymes (PCWDEs), with particularly high numbers of genes in wood-boring species. (McKenna et al., 2019). One of the expectations for the present study was that WBBs should show signatures of convergent expansions of gene families encoding PCWDEs. We found 43 Orthofinder families associated with ten clades of PCWDEs. Nine of these families were widespread enough among the species in our dataset to be analyzed with CAFE. None of these nine families showed a shared

expansion across all WBBs (file PCWDE gene families.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). However, the average number of genes was higher in 8/10 PCWDEs clades in WBBS compared to their sister species, although three of these families occurred only in one WBB species. Among families shared by all WBBs, the gene gains were particularly striking for the GH1 and GH28 clades, and to a lower extent for the GH32 clade, which also represent the three clades with at least one gene in each WBB (Appendix J). Although GH1 genes were clustered in sixteen Orthofinder families, most of the expansion in this clade occurred in one family (OG0000003), which contained an average of 34 genes in WBBs compared to 16 genes in their sister species (Appendix J). In the GH28 clade, one family contained most genes for ALB and EAB, whereas MPB genes were scattered in six other families.

3.2.4. Biological functions of gene families with changes shared by wood-boring beetles

We performed functional annotation analyses of gene families that expanded, contracted or were entirely lost in the three wood-boring beetles using the STRING database (Szklarczyk et al. 2019). We used homologs from *D. melanogaster* in the corresponding gene families as proxies to determine the functional biology underlying the expanding and contracting families. The 16 gene families with gains in WBBs accounted for 73-102 genes in wood-boring beetles compared to 27-51 genes in their sister species (files STRING 16 fams EXP.xlsx and STRING expansions WBBs

function.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). Two of these gene families were associated with peptidase activity. The family OG0000037 encodes proteins with 'serine-type endopeptidase activity' and showed large expansions (3-22 genes) in WWBs. The family OG0000175 included proteins with 'metalloaminopeptidase activity'. The gene family OG0000114 is involved in 'secondary active sulfate transmembrane transporters' and includes the gene *epidermal stripes and patches* (*Esp*), which is implicated in female remating receptivity in *D. melanogaster* (Findlay et al. 2014).

Several genes duplicated in WBBs belonged to the family OG0000070 and are involved in 'dynein light intermediate chain binding'. Some of these genes are expressed in the nervous and the reproductive systems and encode components of the Axonemal complex necessary for motile cilia function. Two families with increased gene content in WBBs, OG0000055 and OG0000267, encode proteins implicated in chitin metabolism, including homologs to the *D. melanogaster Obst-G* gene, which belong to a gene family that plays an important role in the organization of the cuticle (Tajiri et al. 2017).

The 40 gene families with convergent contractions between wood-boring species shared only 56-70 genes in WBBs as opposed to 133-170 genes in their sister species (file STRING 40 fams CONTR.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). These families are implicated in a variety of biological processes (file STRING 40 fams CONTR.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). Notably,

some processes showed contractions in multiple gene families, a possible indication of adaptation via gene loss. Biological processes associated with the immune response, including 'Adaptive Immune System', 'Innate Immune System', 'innate immune response', 'defense response to Gram-positive bacterium' and 'positive regulation of Toll signaling pathway' appeared to be particularly affected by gene loss in WBBs. In the ten families with genes associated to these functions, WBBs species contained a total of 14-19 genes, compared to 31-38 genes found in sister species to WBBs (file STRING contraction immunity.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). Two of these gene families were entirely lost in WBBs. Four out of ten families associated with immune response are also involved in ‘Mitotic G1 phase and G1/S transition’ processes, together with a fifth family that has no known function in immunity. Two families with genes involved in ‘cuticle development’ also showed reduction to 8-9 genes in WBBs compared to 9-14 in their sister species (file STRING contraction immunity.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). Three families shared a functional role in hormone metabolism and regulation, including two families associated with immune response and cuticle development, respectively (file STRING contraction immunity.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families).

4. DISCUSSION

4.1. Southern pine beetle transcriptome assembly and annotation

The southern pine beetle (SPB), *Dendroctonus frontalis,* is a major pest that presents a threat to pine stands throughout the southeastern and eastern United States. Efforts to understand the biology of SPB are essential for developing the next-generation of pest management strategies. Accurate genome assembly and gene annotation represent key resources toward this goal but remain undeveloped in SPB and other *Dendroctonus* species. A major goal of this project was to initiate the gene annotation of SPB by analyzing the first suite of extensive transcriptomic data available in this species.

In the first part of this project, I produced a 94% complete transcript set for the SPB. The *de novo* SPB transcriptome assembly provides additional WBB and *Dendroctonus*-specific data that will be crucial in continuing to understand the wood-boring phenotype and pest behavior of *Dendroctonus*. Once annotated this set of SPB transcripts will add resolution to gene family studies within Coleoptera and, more specifically, to comparative studies within wood-boring species that aim to determine the underpinnings of this complex phenotype. Furthermore, a complete SPB gene set will accelerate the discovery of genes associated with host identification, escape from host defense mechanisms, mating behavior, and other traits that affect the ability of SPB to colonize conifer trees and initiate tree-killing outbreaks. For instance, transcriptomic studies will be facilitated by the availability of a reference genome and an accurate gene set. Specifically within *Dendroctonus*, the availability of a second pest transcriptome is

critical for identifying highly-specific RNAi targets designed to affect *D. frontalis* without acting on benign closely related species.

4.2. Gene family changes associated with convergent evolution of the wood-boring habit in beetles

The wood-boring phenotype is complex, involving overcoming physical and chemical barriers as well as locating suitable hosts. Notably, this phenotype has evolved numerous times across Coleoptera (McKenna et al., 2019). I expected, then, that a large suite of genes and genes families should underly such a phenotype and possibly show signatures of convergent evolution in the form of parallel gene gains and losses. Using full protein sets derived from beetle species with whole genome assemblies and three well-annotated outgroups, I applied a method adapted from previous works focused on amino acid changes (Castoe et al., 2009; Thomas and Hahn, 2015) for assessing convergent evolution of gene families across entire genomes. I show that this method comprehensively compares all possible pairwise and triple-wise combinations of the species and phylogenetic nodes included in the analysis. I also incorporate functional annotation in order to shed light on the biochemical pathways to which the gene families of note belong. It is evident from the pairwise and triad data that there is no strong indication of excess convergent gene family expansion or contraction in the WBB comparisons for this dataset. However, the three WBB species share a high level of convergent expansions per average number of expansions compared to other beetle triads. While this result does not necessarily imply that most convergent gene family

expansions in WBBs are associated with the wood-boring habit, it points to the potential of this approach to identify groups of species with high levels of convergence.

The lack of signatures of convergence in gene family changes across WBBs in both tests may be due to several factors. The three WBB lineages represented by the Mountain pine beetle, the Asian long-horned beetle and the emerald ash borer are separated by hundreds of million years of independent evolution (Zhang et al. 2018). Although it is possible that convergent gene losses and gains were rampant in the early stages of evolution of the wood-boring habit in the ancestors of WBBs, these could have been masked by subsequent changes in gene families associated with millions of years of lineage-specific adaptation. Alternatively, each wood-boring lineage might have achieved the ability to colonize bark tissues through genomic changes that affected different gene families in each taxon. It is important to notice that although the wood-boring habit is typically associated with specific phenotypic changes, the three wood-boring species analyzed in this project differ substantially in the realization of this habit.

It is also worth noting the factors that may lead to the loss of convergent signals, particularly signals of gene family expansion, associated with the methods presented. Foremost, I used a stringent approach with cleaning the data. Fragment and transcript redundancy removal reduced the number of transcripts available for analysis by an average of ~2200 sequences/species. The remaining sequences were used for OrthoFinder clustering and the resulting data was further culled by only keeping the orthogroups that contained seven or more represented species. There are consequences that result from the blanket removal of gene families in which six or fewer species are

represented by at least one gene. The most relevant of which is that, if they exist, those families that are *only* present in WBBs would be lost, including WBB-specific PCWDE families. Indeed, the PCWDE families that were included in the CAFE analysis included either lineages of the GH1 clade, which is widespread in metazoans, or other lineages with inflated gene counts in some species due to the misincorporation of the GH domain into genes with other functionalities. A final factor to consider in these analyses is the accuracy of the ultrametric tree used to guide the CAFE gain/loss predictions. An inaccurate phylogeny may have the potential to affect the output results from CAFE, thus impacting the pairwise and triad tests for gene family convergence. It will be important to simulate the impact of branch length on the CAFE gain and loss output.

While the WBB gene family convergence hypothesis was overall not supported by the evidence, the methods put forth are novel for the testing of evolutionary convergence in gene families across whole genomes. The pairwise and triad tests of gene family convergence I have developed for this project have general applicability to any system with convergent phenotypes, high-quality genomic resources and accurate phylogenetic inferences. The wealth of genomic data that are becoming increasingly available and the improved accuracy of gene annotation datasets should allow for application of this method broadly to determine if gene family expansions and contractions do play a role in convergent phenotypic evolution.

Although WBBs did not display evidence of extensive convergent gain and loss of genes, several gene families shared expansions and contractions along the three WBB branches of the beetle phylogeny. These families showed enrichment for specific

biological processes and functions that may play a role in the adaptation to a wood-boring habit. Two families with parallel expansions in WBBs contained genes associated with the chitin metabolism and the organization of the cuticle. These families may be important for the adaptation of WBB larvae to their particular environment during development beneath bark tissue. Other expanded families encoded for peptidases, sulfate transporters and dynein chain binding, all processes without an obvious association with adaptation to the wood-boring lifestyle. This could be the result of spurious changes (false positives), or evolutionary convergence for traits different from wood-boring. Future studies focusing on the expression pattern of these genes or knock-out experiments will be necessary to determine if any of these gene families are important for the specific adaptations of WBBs.

A more remarkable pattern emerged from the analysis of convergent gene losses in WBBs. These three beetles shared extensive loss of genes involved in immunity, hormone metabolism and cuticle development. Because many of these families were associated with multiple processes, it is difficult to determine if one or multiple functions were primarily affected. However, the loss in gene families involved in the immune response appear to be particularly widespread. One possibility is that a decreased efficiency in part of the immune response facilitate the numerous mutualistic relationships between WBBs and several kinds of bacteria and yeasts (Liu et al. 2020; Six and Elser 2020). This hypothesis is intriguing but at this point entirely speculative. Mutualistic relationships are indeed common among beetles (Estes et al. 2013; Florez et al. 2018). Alternatively, the loss of immune response components may be a consequence

of a larval life style mostly conducted in a relatively "sterile" environment represented by the bark tissue.

Among the 40 gene families with shared contractions among WBBs, ten families were entirely lost in the three wood-boring species, possibly as a result of adaptation via gene loss (Olson 1999; Albalat and Cañestro 2016). Gene loss has been associated with adaptation in several taxa, including mammals (Sharma et al. 2018), and could be represent an important factor shaping the evolution of wood-boring associated traits in beetles. However, any involvement of these 10 families with the emergence of WBB traits was not apparent.

One expectation of this project was to retrieve signatures of convergence evolution between WBBs in PCWDE gene families, as reported previously (McKenna et al., 2019). The numbers of PCWDE genes from different clades that I found in this analysis largely mirror those reported by McKenna and collaborators (2019). Nevertheless, a manual curation of genes from several gene families was necessary in order to remove genes that clustered with genuine PCWDE-encoding genes but do not express proteins that contain a PCWDE domain. This reannotation step revealed that eight families that were initially assumed to contain PCWDE genes in most species and were therefore included in the CAFE analysis shared only a few PCWDE genes (file PCWDE gene families.xlsx available at https://github.com/CasolaLab/beetle_convergence_gene_families). The overall pattern of the reannotated PCWDE clades in beetles suggest that gene gains in PCWDEs clades

occurred through either duplications within a single subclade, as in the case of GH1, or expansions in different subclades for different WBBs, like for GH28. Alternatively, subfamilies with a genes only in one species may be due to annotation artifacts wherein two or more genes are erroneously combined in a single locus that encode for a protein that clusters preferentially with non-GH proteins in Orthofinder.

CONCLUSION

The integration of extensive genomic resources and methods to discover the genetic basis of complex traits is key to improving our ability to control insect pests responsible for significant loss of agricultural and forest resources. The main goals of this project were to develop an accurate and nearly complete transcriptome for the southern pine beetle (SPB), and to identify gene families associated with the convergent evolution of the wood-boring habit across multiple beetle lineages. I used novel SPB RNA-seq data from adult females and males, larvae and pupae developmental stages to assemble 73,853 transcripts, which represent approximately 95% of the SPB protein-coding gene set. Using available gene complements from three other wood-boring beetles, seven non-WBB species and three non-Coleopteran well-annotated insects, I tested the hypothesis that WBBs exhibit a higher level of convergence in gene family changes than other groups of beetles. Although the hypothesis was rejected, I identified 16 and 40 gene families with parallel expansions and contractions in WBBs, respectively. These families encode proteins implicated in an array of processes, which in some cases are likely to be associated with the wood-boring habit. Additional expression or knock-out studies will be necessary for testing the families observed to be expanding or contracting along the three WBB branches. The comparative methods for gene family analyses across taxa with convergent traits can be readily applied in other systems with extensive genomic resources, phylogenetic information and phenotypes that evolved independently.

REFERENCES

Albalat, R. and C. Cañestro. 2016. Evolution by gene loss. Nature Reviews Genetics. 17: 379-391.

Andersson, M. N., C. I. Keeling, and R. F. Mitchell. 2019. Genomic content of chemosensory genes correlates with host range in wood-boring beetles (Dendroctonus ponderosae, Agrilus planipennis, and Anoplophora glabripennis). BMC Genomics 20:1–18.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:1–9.

Casola, C., S. Landa, C. Hjelmen, M. Jonika, B. Sullivan, B. Kyre, L.K. Rieske-Kinney, and H. Blackmon. "Using comparative genomics to identify species-specific targets for RNAi in Dendroctonus bark beetles". Annual Meeting of the Entomological Society of America, November 2020.

Castoe, T., A. P. J. de Koninga, H.M. Kima, W. Gua, B.P. Noonanb, G. Naylorc, Z.J. Jiangd, C.L. Parkinson, and D.D. Pollocka. " Evidence for an ancient adaptive episode of convergent molecular evolution". Proceedings of the National Academy of Sciences of the United States of America Vol.106, 22:8986-8991.

Chung, N., B.Z. Miasojedow, M. Startek, A. Gambin. " Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data". BMC Bioinformatics 20:Suppl 15, 1-11.

Crook, D. J., and V. C. Mastro. 2010. Chemical ecology of the emerald ash borer agrilus planipennis. Journal of Chemical Ecology 36:101–112.

Emms, D. M., and S. Kelly. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology 20:1–14.

Estes, D., David J. Hearn, E.C. Snell-Rood, M. Feindler , K. Feeser , T, Abebe, J.C. Dunning Hotopp, and A.P. Moczek. 2013. Brood ball-mediated transmission of microbiome members in the dung beetle, Onthophagus taurus (Coleoptera: Scarabaeidae). PLoS ONE. 8: 1-15.

Evans, J. D., S. J. Brown, K. J. J. Hackett, G. Robinson, S. Richards, D. Lawson, C.

Elsik, J. Coddington, O. Edwards, S. Emrich, T. Gabaldon, M. Goldsmith, G. Hanes, B. Misof, M. Muñoz-Torres, O. Niehuis, A. Papanicolaou, M. Pfrender, M. Poelchau, M. Purcell-Miramontes, H. M. Robertson, O. Ryder, D. Tagu, T. Torres, E. Zdobnov, G. Zhang, and X. Zhou. 2013. The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. Journal of Heredity 104:595–600.

Findlay, G., J.L. Sitnik, W. Wang, C.F. Aquadro, N.L. Clark, and M.F. Wolfner. 2014. Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for Drosophila melanogaster Female Post-Mating Responses. PLoS Genetics. 10.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. 2014. Pfam: The protein families database. Nucleic Acids Research 42:222–230.

Florez, L., K. Scherlach, I.J. Miller, A. Rodrigues , J.C. Kwan ,
C. Hertweck, and M. Kaltenpoth. 2018. An antifungal polyketide associated with horizontally acquired genes supports symbiont-mediated defense in Lagria villosa beetles. Nature Communications. 9.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Acadamy of Sciences of the U. S. A. 117:9451–9457.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

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.

Haas, BJ and Papanicolaou, A. 2019. TransDecoder 5.5.0 [Online]. Available online at: https://www.github.com/TransDecoder/TransDecoder/wiki

Han, M. V., G. W. C. Thomas, J. Lugo-Martinez, and M. W. Hahn. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular Biology and Evolution 30:1987–1997.

Holt, C., and M. Yandell. 2011. MAKER2: An annotation pipeline and genome-database

management tool for second-generation genome projects. BMC Bioinformatics 12.

Hu, J., S. Angeli, S. Schuetz, Y. Luo, and A. E. Hajek. 2009. Ecology and management of exotic and endemic Asian longhorned beetle Anoplophora glabripennis. Agricultural and Forest Entomology 11:359–375.

Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. Von Mering, and P. Bork. 2019. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47:D309–D314.

Huggett R, Wear DN, Li R, Coulston J, Liu S. 2013. Forest forecasts. . In: Wear DN, Greis JG, editors. The Southern Forest Future Project: technical report. Asheville, N.C.: USDA Forest Service, Southern Research Station.

Jones, P., D. Binns, H. Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez, and S. Hunter. 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30:1236–1240.

Keeling, C. I., M. M. S. Yuen, N. Y. Liao, T. R. Docking, K. S. S. K. Chan, G. A. Taylor, D. L. Palmquist, S. D. Jackman, A. Nguyen, M. Li, H. Henderson, J. K. Janes, Y. Zhao, P. Pandoh, R. Moore, F. A. H. Sperling, D. P. W. Huber, I. Birol, S. J. M. Jones, and J. Bohlmann. 2013. Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. Genome Biology 14.

Kolmogorov, M., J. Yuan, Y. Lin, and P. A. Pevzner. 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37:540–546. Springer US.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. 2017. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Research 27:722–736.

Krueger, F. 2015. TrimGalore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files [Online]. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kumar, Sudhir, G. Stecher, M. Suleski, and S.B. Hedges. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution 34: 1812-1819.

Kyre, B. R., T. B. Rodrigues, and L. K. Rieske. 2019. RNA interference and validation of reference genes for gene expression analyses using qPCR in southern pine beetle, Dendroctonus frontalis. Scientific Reports 9:1–8.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Liu, F., J.D. Wickham, Q. Cao, M. Lu, and J. Sun. 2020. An invasive beetle–fungus complex is maintained by fungal nutritional-compensation mediated by bacterial volatiles. ISME Journal. 14:2829-2842.

Marchler-Bauer, A., , M.K. Derbyshire, N R. Gonzales, S. Lu, F. Chitsaz, L.Y. Geer, R.C. Geer, J.He, M.Gwadz, D.I. Hurwitz, C.J. Lanczycki, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng and S.H. Bryant. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Research 43: D222-D226.

McKenna, D. D., E. D. Scully, Y. Pauchet, K. Hoover, R. Kirsch, S. M. Geib, R. F. Mitchell, R. M. Waterhouse, S. J. Ahn, D. Arsala, J. B. Benoit, H. Blackmon, T. Bledsoe, J. H. Bowsher, A. Busch, B. Calla, H. Chao, A. K. Childers, C. Childers, D. J. Clarke, L. Cohen, J. P. Demuth, H. Dinh, H. V. Doddapaneni, A. Dolan, J. J. Duan, S. Dugan, M. Friedrich, K. M. Glastad, M. A. D. Goodisman, S. Haddad, Y. Han, D. S. T. Hughes, P. Ioannidis, J. S. Johnston, J. W. Jones, L. A. Kuhn, D. R. Lance, C. Y. Lee, S. L. Lee, H. Lin, J. A. Lynch, A. P. Moczek, S. C. Murali, D. M. Muzny, D. R. Nelson, S. R. Palli, K. A. Panfilio, D. Pers, M. F. Poelchau, H. Quan, J. Qu, A. M. Ray, J. P. Rinehart, H. M. Robertson, R. Roehrdanz, A. J. Rosendale, S. Shin, C. Silva, A. S. Torson, I. M. V. Jentzsch, J. H. Werren, K. C. Worley, G. Yocum, E. M. Zdobnov, R. A. Gibbs, and S. Richards. 2016. Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. Genome Biology 17:1–18.

McKenna, D. D., S. Shin, D. Ahrens, M. Balke, C. Beza-Beza, D. J. Clarke, A. Donath, H. E. Escalona, F. Friedrich, H. Letsch, S. Liu, D. Maddison, C. Mayer, B. Misof, P. J. Murin, O. Niehuis, R. S. Peters, L. Podsiadlowski, H. Pohl, E. D. Scully, E. V. Yan, X. Zhou, A. Ślipiński, and R. G. Beutel. 2019. The evolution and genomic basis of beetle diversity. Proceedings of the National Academy of Sciences of the U. S. A. 116:24729–24737.

Negrón, J. F., and C. J. Fettig. 2014. Mountain pine beetle, a major disturbance agent in us western coniferous forests: A synthesis of the state of knowledge. Forest Science 60:409–413.

Powell, D., E. Groβe-wilde, P. Krokene, A. Roy, and A. Chakraborty. 2020. A highly contiguous genome assembly of a major forest pest, the Eurasian spruce bark beetle Ips typographus. bioRxiv. doi: https://doi.org/10.1101/2020.11.28.401976

Putnam, N. H., B. O. Connell, J. C. Stites, B. J. Rice, P. D. Hartley, C. W. Sugnet, D. Haussler, and D. S. Rokhsar. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. 2015. Genome Research 26:342–350.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41:590–596.

Seppey, M., P. Ioannidis, B. C. Emerson, C. Pitteloud, M. Robinson-Rechavi, J. Roux, H. E. Escalona, D. D. McKenna, B. Misof, S. Shin, X. Zhou, R. M. Waterhouse, and N. Alvarez. 2019. Genomic signatures accompanying the dietary shift to phytophagy in polyphagan beetles. Genome Biology 20:1–14.

Sharma, V., N. Hecker, J.G. Roscito, L. Foerster, B.E. Langer and M. Hiller. 2018. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. Nature Communications 9:1-9.

Six, D. L., and R. Bracewell. 2015. Dendroctonus. Smit, AFA, Hubley, R & Green, P. 2013-2015. *RepeatMasker Open-4.0*. [Online]. Available online at: http://www.repeatmasker.org

Six, D. and J.J. Esler. 2020. Mutualism is not restricted to tree-killing bark beetles and fungi: the ecological stoichiometry of secondary bark beetles, fungi, and a scavenger. Ecological Entomology. 45: 1134-1145.

Szklarczyk, D., A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N.T. Doncheva, J.H. Morris, P. Bork, L.J. Jensen, and C. Von Mering. 2019. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 47: D607-D613.

Storer, J., R. Hubley, J. Rosen, T. J. Wheeler, and A. F. Smit. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12:1–14.

Tajiri, R., N. Ogawa, H. Fujiwara, and T. Kojima. 2017. Mechanical Control of Whole Body Shape by a Single Cuticular Protein Obstructor-E in Drosophila melanogaster. PLoS genetics 13:1-26.

Thomas, G. W. C., and M. W. Hahn. 2015. Determining the null model for detecting

adaptive convergence from genomic data: A case study using echolocating mammals. Molecular Biology and Evolution 32:1232–1236.

Villari, C., D. A. Herms, J. G. A. Whitehill, D. Cipollini, and P. Bonello. 2016. Progress and gaps in understanding mechanisms of ash tree resistance to emerald ash borer, a model for wood-boring insects that kill angiosperms. New Phytologist 209:63–79.

Zhang, S., L.H. Che, Y. Li, D. Liang, H. Pang, A. Ślipiński & P. Zhang. 2018. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. Nature Communications 9:1–11.

APPENDIX A

ISOFORM_ID.PY : CREATES ARRAY OF ALL FASTA HEADERS AND ASSOCIATED SEQUENCES. LENGTH OF SEQUENCES ASSOCIATED WITH IDENTICAL HEADERS ARE COMPARED, HEADER WITH THE LONGEST SEQUENCE IS WRITTEN TO A 'NO ISOFORM' FILE.

```
# Shelby Landa
# 5 Jan 2020
# isoform_ID.py

#from rename_seq import cds_ID_list
import numpy as np
from Bio import SeqIO
# re for
import re
# list_duplicates function depends on dictionary listing
from collections import defaultdict
#cds_IDs = cds_ID_list()

#input file
file = input("file w/ isoforms: ")

# sp_cds_ID variable (below) will contain a list of all gene IDs without isoform tag (-RA, -RB, -RC, etc)

# code for list_duplicates copied and modified from :
https://stackoverflow.com/questions/5419204/index-of-duplicates-items-in-a-python-list
# somewhat misleading function name because I modified to return all keys -- even those with only one instance in the list
# better name would be list_dictionary
def list_duplicates(seq):
   # defaultdict from collections package creates dictionary in form of list
   tally = defaultdict(list)
   for i, item in enumerate(seq):
      tally[item].append(i)
   # commented line below returns key ( gene in this case ) and position of all duplicates ( isoforms in this case )
      # return ((key, locs) for key, locs in tally.items() if len(locs) > 1)
   # active line (below) returns key and position of all genes ( used this way so that lines can be written to file in the for loop below)
   return ((key, locs) for key, locs in tally.items())
```

```
    #return (key for (key, locs) in tally.items() if len(locs) > 1)

# might not need this
# could just loop through fasta_array ( below ) and create a list of gene IDs without
isoform tags, but then wouldnt know how big to make the array
with open(file, 'r') as sp_cds:
    # define list for all gene IDs
    sp_cds_ID = []
    for line in sp_cds:
        # searching for gene ID without the isoform tag ( -RA, -RB, -RC, etc)
        if re.search(r'>\w+', line):
            # re.search returns "match" object
            current_ID = re.search(r'>\w+', line)
            # current_ID.group(0) isolates match string
            current_ID = current_ID.group(0)
            # append ID to list of protein IDs for current species
            sp_cds_ID.append(current_ID)

# create parse object using SeqIO (imported above)
fasta_seq = SeqIO.parse(file,'fasta')
# define empty array for name, sequence info in fasta_seq
fasta_array = np.empty((len(sp_cds_ID), 2), dtype=object)
# create counter variable for filling array
array_counter = 0
# loop through each fasta object in fasta_seq
for fasta in fasta_seq:
    # save name and sequence into respective variables
    name, seq = fasta.id, str(fasta.seq)
    # save name, seq into array columns
    fasta_array[array_counter, 0] = name
    fasta_array[array_counter, 1] = seq
    print (fasta_array[array_counter, 0])
    # update counter
    array_counter += 1

# create new file to be written: will contain only longest isoform of genes with multiple
sequences
writeFile = input(str("New file name: "))
writeFile = open(writeFile, 'w')
# define list for duplicates -- dont think this actually gets used
dup_list = []
# dup in list duplicates returns format of ('key',[1,2,3])
for dup in list_duplicates(sp_cds_ID):
    # create list of genes with isoforms
```

```
    dup_list.append(dup)
    print (dup)
    # geneSeq_list will contain the sequences of each isoform--will be used to find longest
sequence
    # resets for each key
    geneSeq_list = []
    # i in dup[1] will iterate through the positions of each isoform and pull the sequences
from fasta_array
    for i in dup[1]:
        geneSeq_list.append(fasta_array[i, 1])
    # if there is only one gene (no isoforms) write the name of the gene and the sequence
to the new file
    if len(geneSeq_list) == 1:
        #print (dup[1])
        writeFile.write('>' + fasta_array[dup[1][0], 0] + '\n')
        writeFile.write(fasta_array[dup[1][0], 1] + '\n')
    # need to find longest sequence for genes with > 1 sequences (isoforms)
    else:
        # reset/initialize longestSeq variable
        longestSeq = ''
        for x in range(0, len(geneSeq_list)):
            currentSeq = geneSeq_list[x]
            # update longestSeq if currentSeq is longer
            if len(currentSeq) > len(longestSeq):
                longestSeq = currentSeq
                # save longestSeq index in geneSeq_list because it will correspond with the
position in the dup[1] list
                longestSeqIndex = x
        # dup[1] contains all positions for current key -- dup[1][longestSeqIndex] will
retrieve the position (row) of the correct gene
            # in the fasta_array
        writeFile.write('>' + fasta_array[dup[1][longestSeqIndex], 0] + '\n')
        writeFile.write(fasta_array[dup[1][longestSeqIndex], 1] + '\n')

writeFile.close()
```

APPENDIX B

PARSE_AETHINA_TUMINDA_NCBI.PY : PARSES NUCLEOTIDE AND PROTEIN SEQUENCE IDS FROM A. TUMIDA GENBANK ACCESSION DATA AND PLACES THEM IN AN ARRAY TO BE IMPORTED BY RENAME_AETHINA.PY

```
# Parse aethina_tumida_ncbi.txt file for protein and cds fasta headers
# Shelby Landa
# 31 Oct 2019
# requires re and numpy packages
# parse_aethina_tumida_ncbi.py

######################
# define function to call
def parse_aethina():
      # re module for regular expressions; numpy module for assembling final array
   import re
   import numpy as np

   # define regular expression to search lines for XM labels for cds
   xm = re.compile(r'/coded_by="X')

   # open text file as data -- not read all at once. Will be read line-by-line
   with open('aethina_tumida_ncbi.txt', 'r') as data:
   # define protein and cds lists -- will contain matching protein and
   #    cds labels by position
      proteins = []
      cds = []
      # iterate through lines in data (all file contents)
      for line in data:
         # re.search (VERSION XP_1234) isolates protein ID (XP_1234) a single time
         # could use any line containing 'XP' and another distinct string
         #   ex: 'LOCUS' and 'XP_'
         if re.search('VERSION', line) and re.search('XP_', line):
            # split('\n') splits string by newline resulting in an empty element
            # strip() removes whitespace (empty elements)
            line = line.strip().split('\n')
            # append ID to proteins list
            ### HARD CODED FOR THIS FILE (aethina_tumida_ncbi.txt)
            proteins.append(line[0][12:26])
         # use regular expression defined above to search each line for cds ID
```

```
            #   associated with protein
            elif xm.search(line):
                line = line.strip().split('\n')
                ### HARD CODED FOR THIS FILE (aethina_tumida_ncbi.txt)
                cds.append(line[0][11:25])
        # concatenate proteins list with cds list so that protein ID and corresponding cds ID
are in the same row
        aethina_array = np.column_stack((cds, proteins))
        #print(aethina_array[0])
        #print(aethina_array[0][0])
        return (aethina_array)

    print(aethina_array)
    # hard-coded, but works for this file
```

APPENDIX C

RENAME_AETHINA.PY : IMPORTS NUCLEOTIDE/PROTEIN ARRAY FROM PARSE_AETHINA_TUMIDA_NCBI.PY. USES REGULAR EXPRESSIONS TO FIND MATCHING PROTEIN/NUCLEOTIDE IDS AND WRITE NEW FASTA HEADERS TO FILE.

```
# rename fasta headers to match between cds and protein files for each species
# Shelby Landa
# rename_aethina.py
import re
from parse_aethina_tumida_ncbi import parse_aethina

renameFile = input("file to rename: ")
newFileName = str(input("new file name (species + prt or cds): "))

readFile = open(renameFile)
newFileName = open(newFileName, 'w')

lines = readFile.readlines()
# save list of gene IDs
aethina_list = parse_aethina()
print(len(aethina_list))

# iterate through each line of prt file
for line in lines:
   # search for PROTEIN id with regular expression – regex may need to be changed
depending on formatting of fasta header
   if re.search(r'\w+_\d+.\d', line):
      ID = re.search(r'\w+_\d+.\d', line)
      ID = ID.group(0)
      for i in range(0, len(aethina_list)):
         if ID == aethina_list[i][0]:
            newFileName.write( ">" + str(aethina_list[i][1]) + '\n')
         elif ID == aethina_list[i][1]:
            newFileName.write(">" + str(aethina_list[i][1]) + '\n')
   else:
      newFileName.write(line)
```

APPENDIX D

RENAME_SEQ.PY : CREATES LIBRARY OF ALL FASTA HEADERS TO BE IMPORTED BY WRITE_SEQ.PY

```
# Rename amino acid and nucleotide sequences to match each other
# Shelby
# 31 Oct 2019

#from parse_aethina_tumida_ncbi import aethina_prt_cds

#print(aethina_prt_cds())
import re
#import numpy as np

# retrieve gene name from cds
def cds_ID_list():
    # open input file as read only
    # input cds file to retrieve list of gene names
    input_file = input("cds file: ")
    with open(input_file, 'r') as sp_cds:
        # define list for all gene IDs
        sp_cds_ID = []
        for line in sp_cds:
            # regular expression searches for a string beginning with '>' followed by
characters or digits with the option
            # of being followed by a decimal
            # Anoplophora glab regex: >\w+-\w+
            # D melanogaster regex: [A-Z]+\d+-\w+
            # T castaneum: >[A-Z]*\d+
            # general regex: >\w+[.\d]*
            # ldec regex: XP_\d+.\d+
            if re.search(r'XP_\d+.\d+', line):
                # re.search returns "match" object
                current_ID = re.search(r'XP_\d+.\d+', line)
                # current_ID.group(0) isolates match string
                current_ID = current_ID.group(0)
                # append ID to list of protein IDs for current species
                sp_cds_ID.append(current_ID)
    return sp_cds_ID
```

APPENDIX E

WRITE_SEQ_FILE.PY : IMPORTS LIBRARY OF CDS IDS FROM RENAME_SEQ.PY TO SEARCH, USING REGULAR EXPRESSIONS, FOR IDENTICAL IDS IN THE FILE TO BE RENAMED. IDS WITH A MATCH ARE PLACED INTO AN ARRAY CONTAINING A SIMPLIFIED ID IN ONE COLUMN AND THE FASTA SEQUENCE IN A SECOND COLUMN

```
# Shelby Landa
# 19 Nov 2019
# Write new IDs to sequences

# import regex module
import re
# import list of all cds gene IDs from function defined/run through rename_seq.py file
from rename_seq import cds_ID_list

# prompts command line for file name for renaming
renameFile = input("file to rename: ")
# prompts command line for new name to write renamed sequences to
newFileName = str(input("new file name (ex: Dmelanogaster_new_prt): "))

# open file with 'old' gene IDs (original file)
readFile = open(renameFile)
# open file to write new gene IDs to
newFileName = open(newFileName, 'w')

# readFile.readlines opens original file line by line
lines = readFile.readlines()
# cds_IDs is list of all regex gene ID matches (essentially the new gene IDs)
cds_IDs = cds_ID_list()
# print(len(cds_IDs))          # debugging/troubleshooting line
# print(cds_IDs)               # debugging/troubleshooting line

# iterate through each line of prt file
for line in lines:
    ### search for PROTEIN id with regular expression
    # anoplophora glab regex: \w+-\w+
    # for anoplophora glab, need to paste '>' onto beginning of string
    # D melanogaster regex: [A-Z]+\d+-\w+
    # need to paste '>' for Dmel as well as L. decemlineata
    if re.search(r'XP_\d+.\d+', line):
```

```
        # save
        ID = re.search(r'XP_\d+.\d+', line)
        ID = ID.group(0)
        for i in range(0, len(cds_IDs)):
            if ID == cds_IDs[i]:
                newFileName.write('>' + cds_IDs[i] + '\n')
    else:
        newFileName.write(line)

newFileName.close()
readFile.close()
```

APPENDIX F

CAFEGAINLOSS-EXPANDED.R : COMPARES GAINS AND LOSSES DETAILED IN THE CAFE FILE 'BASE_CHANGE.TAB' BETWEEN EACH NODE (INCLUDING TIPS) OF THE INPUT TREE IN ORDER TO DETERMINE THE CONVERGENT AND DIVERGENGT CHANGES BETWEEN FAMILIES ACROSS NODES AND SPECIES.

```
# Shelby Landa
# 17 Aug 2021
# parse cafe output (base_change.tab) to determine total number of gain and loss events between each node as predicted by cafe

setwd("~/Desktop/Dendroctonus_project/CAFE/Orthofinder_noCBB/CAFE_noCBB_zeroRoot/")
# read in base change table output from CAFE
change_tab <- read.delim("Base_change_zeroRoot.tab")

# name rows, remove first column from dataframe
rownames(change_tab) <- change_tab[,1]
change_tab <- change_tab[,2:ncol(change_tab)]

# list of new column names -- node numbers start at 0
nodeNames <- c(paste0("node_", seq(from = 0, to = (ncol(change_tab) - 1))))
# rename columns
colnames(change_tab) <- nodeNames
# convert factors to numeric
change_tab <- as.matrix(change_tab)

# create empty matrix with same dimensions, row/column names as original matrix
new_changeTab <- matrix(nrow = nrow(change_tab), ncol = ncol(change_tab))
colnames(new_changeTab) <- colnames(change_tab)
rownames(new_changeTab) <- rownames(change_tab)
# convert positive, negative values to symbols
# formula replaces any cases satisfying each boolean statement in the original matrix (change_tab) with '-', '+' or '0'in the new matrix (new_changeTab)
new_changeTab[change_tab < 0] <- '-'
new_changeTab[change_tab > 0] <- '+'
new_changeTab[change_tab == 0] <- 0

####
#
```

```
# compare gain and loss events between each node
#
####
# create empty matrix to store convergent ('+'), divergent ('-'), or no change ('0') between
each pairwise comparison
# ncol determined by number of pairwise comparisons
compare_matrix <- matrix(nrow = nrow(new_changeTab), ncol =
sum(1:ncol(new_changeTab) - 1))
# name rows with same orthogroups as original matrix
rownames(compare_matrix) <- rownames(new_changeTab)

# pre-define list of column names -- each column will represent an individual
comparison between nodes (0/1,/0/2,0/3 etc)
compare_colNames <- c()
# for loop generates column names
# start count variable at 0 because we have a 0 node, go through 24 because the 24th
node will the the last to be compared (to node 24)
for (i in 0:(ncol(new_changeTab) - 2)){
  # paste 0 statement concatenates the count variable from the for loop with a sequence
of numbers beginning one above the count variable
  # list results in pairwise labels ('0/1, 0/2, 0/3) for all node comparisons
  compare_colNames <- c(compare_colNames, paste0(i, '/', seq(from = i+1, to =
ncol(new_changeTab) - 1)))
  # print statement to check that code works as planned
  # print (paste0(i, '/', seq(from = i+1, to = 24)))
}
# apply names to compare_matrix
colnames(compare_matrix) <- compare_colNames

# pairwise comparison of gain/loss events
# convergent gains/losses marked by a +, divergent gains/losses marked by a -, 0/0
comparison marked by 0

# define global variable that should count through the number of columns in the
compare_matrix matrix
pairwiseNum <- 1

# first loop needs to be the length of the count data minus 1 because the last column does
not need to be compared to anything further
for (i in 1:(ncol(new_changeTab) - 1)) {
  # compareNodes will be a list of column numbers that baseNode will be compared to
  compareNodes <- c(seq(from = i+1, to = ncol(new_changeTab)))
  # baseNode is node to be compared, pairwise, to the rest of the count data
  baseNode <- new_changeTab[,i]
```

```
  # loop through each individual column that the baseNode is being compared to
  for (j in 1:length(compareNodes)) {
    print (pairwiseNum)
    # the following code generates a boolean list -- where the list is 'TRUE' the given
symbol will be placed in the corresponding coordinates in compare_matrix
    # if baseNode is equal to (+/+ or -/-) its pairwise comparison, that is a convergence
and is marked by a '+' in compare_matrix
    compare_matrix[baseNode == new_changeTab[,compareNodes[j]], pairwiseNum] <-
'+'
    # if baseNode is not equal to its pairwise comparison (+/- , -/+ , 0/+ , 0/-), that is a
divergence and is marked by a '-' in compare matrix
    compare_matrix[baseNode != new_changeTab[,compareNodes[j]], pairwiseNum] <-
'-'
    # a 0/0 comparison will be recognized by the code above as a '+' designation
    # the for loop below scans each data point in the current column of compare_matrix
    for (k in 1:nrow(compare_matrix)){

      # if a '+' has been placed in the compare_matrix row AND the datapoint in baseNode
is a 0 then the datapoint in compare_matrix needs to be updated to a 0 rather than a '+'
      if (compare_matrix[k,pairwiseNum] == '+' && new_changeTab[k, i] == '0') {
        compare_matrix[k, pairwiseNum] <- '0'
        # if a '+' has been placed in the compare_matrix row AND the datapoint in
baseNode is a + then this is an expanding convergence
      } else if (compare_matrix[k,pairwiseNum] == '+' && new_changeTab[k, i] == '+') {
        compare_matrix[k, pairwiseNum] <- '+/+'
        # if a '+' has been placed in the compare_matrix row AND the datapoint in
baseNode is a - then this is a contracting convergence
      } else if (compare_matrix[k,pairwiseNum] == '+' && new_changeTab[k, i] == '-') {
        compare_matrix[k, pairwiseNum] <- '+/-'
      }
        # when evaluating divergence, 0/- and 0/+ comparisons are deemed 'divergent'
        # the following 'else if' code further parses the divergence
        # for divergences where the base node has 0 change, this block determines whether
the divergence was a gain or a loss in the comparison node
      else if (compare_matrix[k,pairwiseNum] == '-' && new_changeTab[k, i] == '0') {

        # if the comparison node has a gain, pairwise comparison will be marked 0/+
        if (new_changeTab[k, i+1] == '+') {
          compare_matrix[k, pairwiseNum] <- '0/+'

          # if the comparison node has a gain, pairwise comparison will be marked 0/-
        } else if (new_changeTab[k, i+1] == '-'){
          compare_matrix[k, pairwiseNum] <- '0/-'
        }
```

```
    }
     # for divergences where the base node has a '+' or a '-' change and the comparison
node has 0 change, this block determines whether the divergence was a gain or loss
    else if (compare_matrix[k,pairwiseNum] == '-' && new_changeTab[k, i+1] == '0') {
     # if the base node has a gain, and comparison node has no change, pairwise
comparison will be marked +/0
     if (new_changeTab[k, i] == '+') {
      compare_matrix[k, pairwiseNum] <- '+/0'
      # if the base node has a loss, and comparison node has no change, pairwise
comparison will be marked -/0
     } else if (new_changeTab[k, i] == '-') {
      compare_matrix[k, pairwiseNum] <- '-/0'
     }
    }
   }
   # update pairwiseNum for each comparison so that data will be added to the correct
column in compare_matrix
   pairwiseNum <- pairwiseNum + 1
  }
}

# matrix to record number of convergence, divergence, and 0 for each pairwise
comparison
summary_matrix <- matrix(nrow = ncol(compare_matrix), ncol = 8)
# sum total convergence, divergence and zero change for each pairwise comparison
colnames(summary_matrix) <- c('convGain', 'convLoss', 'div', '0/+', '0/-', '+/0', '-/0',
'zero')
rownames(summary_matrix) <- colnames(compare_matrix)
for (i in 1:ncol(compare_matrix)) {
  # generate boolean vector where TRUE indicates convergent gains, '+/+', in those
comparisons
  TF_convGain <- compare_matrix[, i] == "+/+"
  summary_matrix[i,1] <- sum(TF_convGain, na.rm = TRUE)
  # generate boolean vector where TRUE indicates convergent losses, '+/-', in those
comparisons
  TF_convLoss <- compare_matrix[, i] == "+/-"
  summary_matrix[i,2] <- sum(TF_convLoss, na.rm = TRUE)
  # generate boolean vector where TRUE indicates divergences, '-', in those comparisons
  TF_div <- compare_matrix[, i] == '-'
  summary_matrix[i,3] <- sum(TF_div, na.rm = TRUE)
  # generate boolean vector where TRUE indicates 0/+ divergences, '0/+', in those
comparisons
  TF_0plus <- compare_matrix[, i] == '0/+'
  summary_matrix[i,4] <- sum(TF_0plus, na.rm = TRUE)
```

```
  # generate boolean vector where TRUE indicates 0/- divergences, '0/-', in those
comparisons
  TF_0minus <- compare_matrix[, i] == '0/-'
  summary_matrix[i,5] <- sum(TF_0minus, na.rm = TRUE)
  # generate boolean vector where TRUE indicates +/0 divergences, '+/0', in those
comparisons
  TF_plus0 <- compare_matrix[, i] == '+/0'
  summary_matrix[i,6] <- sum(TF_plus0, na.rm = TRUE)
  # generate boolean vector where TRUE indicates -/0 divergences, '-/0', in those
comparisons
  TF_minus0 <- compare_matrix[, i] == '-/0'
  summary_matrix[i,7] <- sum(TF_minus0, na.rm = TRUE)
  # generate boolean vector where TRUE indicates no change, '0', in those comparisons
  TF_zero <- compare_matrix[, i] == '0'
  summary_matrix[i,8] <- sum(TF_zero, na.rm = TRUE)
}

write.table(summary_matrix,
            file = "expanded_conv_div_noCBB_zeroRoot.tsv",
            quote = FALSE,
            sep = "\t")
```

APPENDIX G

PAIRWISE CONVERGENT GENE FAMILY EXPANSIONS (CE) AND DIVERGENT GENE FAMILIES (D). PAIRS WERE ORDERED BY DECREASING CE/D VALUES. RED ROWS: WBBS (ALB: ASIAN LONGHORN BEETLE (*A. GLABRIPENNIS*); EAB: EMERALD ASH BORER (*A. PLANIPENNIS*); MPB: MOUNTAIN PINE BEETLE (*D. PONDEROSAE*)). BLUE ROWS: SISTER SPECIES OF WBBS (CPB: COLORADO POTATO BEETLE (*L. DECEMLINEATA*)).

Pair of species	**CE**	**Divergent**	**CE/D**
Sitophilus/Aethina	156	145	1.08
Sitophilus/CPB	**138**	**134**	**1.03**
Sitophilus/Photinus	**154**	**157**	**0.98**
ALB/Onthophagus	171	175	0.98
Sitophilus/ALB	119	124	0.96
Aethina/Photinus	236	262	0.90
CPB/Aethina	200	223	0.90
CPB/Photinus	**198**	**231**	**0.86**
CPB/Danaus	186	224	0.83
Sitophilus/Onthophagus	113	138	0.82
Aethina/Danaus	190	233	0.82
ALB/Tribolium	128	167	0.77
CPB/ALB	151	200	0.76
Sitophilus/Danaus	120	159	0.75
Sitophilus/Tribolium	89	120	0.74
ALB/EAB	**135**	**183**	**0.74**
CPB/Onthophagus	156	217	0.72
Aethina/Tribolium	129	185	0.70
Aethina/HoneyBee	157	230	0.68

MPB/Aethina	158	246	0.64
ALB/Danaus	138	217	0.64
Tribolium/Photinus	130	207	0.63
Photinus/Danaus	192	306	0.63
ALB/Photinus	146	238	0.61
Onthophagus/Danaus	146	242	0.60
Onthophagus/Photinus	160	274	0.58
CPB/Tribolium	110	191	0.58
Aethina/Drosophila	154	270	0.57
MPB/Danaus	124	220	0.56
Aethina/Onthophagus	158	281	0.56
Aethina/Nicrophorus	108	198	0.55
Danaus/HoneyBee	136	250	0.54
Tribolium/Danaus	96	179	0.54
Tribolium/Drosophila	96	179	0.54
Onthophagus/EAB	115	216	0.53
CPB/EAB	115	223	0.52
Onthophagus/Tribolium	101	206	0.49
Sitophilus/EAB	79	164	0.48
Sitophilus/Drosophila	90	193	0.47
CPB/HoneyBee	116	251	0.46
Sitophilus/HoneyBee	86	188	0.46
ALB/Aethina	118	260	0.45
EAB/Danaus	108	250	0.43
MPB/Photinus	120	285	0.42
MPB/HoneyBee	88	210	0.42

MPB/Tribolium	71	171	0.42
Nicrophorus/Danaus	85	210	0.40
MPB/ALB	**90**	**223**	**0.40**
CPB/Nicrophorus	90	224	0.40
MPB/Onthophagus	94	235	0.40
CPB/Drosophila	115	293	0.39
ALB/Nicrophorus	88	226	0.39
Photinus/Drosophila	130	348	0.37
ALB/Drosophila	103	276	0.37
Sitophilus/Nicrophorus	69	189	0.37
Nicrophorus/HoneyBee	60	165	0.36
Aethina/EAB	103	285	0.36
MPB/EAB	**75**	**209**	**0.36**
Tribolium/EAB	76	212	0.36
Onthophagus/Drosophila	102	285	0.36
Nicrophorus/Onthophagus	80	229	0.35
Danaus/Drosophila	99	289	0.34
MPB/Drosophila	83	243	0.34
EAB/HoneyBee	73	219	0.33
Tribolium/HoneyBee	63	193	0.33
Nicrophorus/Tribolium	61	187	0.33
Onthophagus/HoneyBee	87	267	0.33
MPB/CPB	79	259	0.31
Nicrophorus/Drosophila	62	212	0.29
MPB/Nicrophorus	48	165	0.29
ALB/HoneyBee	75	258	0.29

Nicrophorus/EAB	59	210	0.28
Photinus/HoneyBee	96	347	0.28
Photinus/EAB	86	323	0.27
Nicrophorus/Photinus	73	277	0.26
Drosophila/HoneyBee	70	285	0.25
EAB/Drosophila	68	309	0.22
Sitophilus/MPB	2	231	0.01

APPENDIX H

PAIRWISE CONVERGENT GENE FAMILY CONTRACTIONS (CC) AND DIVERGENT GENE FAMILIES (D). PAIRS WERE ORDERED BY DECREASING CC/D VALUES. RED ROWS: WBBS (ALB: ASIAN LONGHORN BEETLE (*A. GLABRIPENNIS*); EAB: EMERALD ASH BORER (*A. PLANIPENNIS*); MPB: MOUNTAIN PINE BEETLE (*D. PONDEROSAE*)).. BLUE ROWS: SISTER SPECIES OF WBBS (CPB: COLORADO POTATO BEETLE (*L. DECEMLINEATA*)).

Pair of species	**CC**	**Divergent**	**CC/D**
Nicrophorus/HoneyBee	315	165	1.91
MPB/Nicrophorus	268	165	1.62
Nicrophorus/Drosophila	288	212	1.36
Drosophila/HoneyBee	375	285	1.32
EAB/HoneyBee	265	219	1.21
Nicrophorus/EAB	253	210	1.20
MPB/Tribolium	206	171	1.20
Danaus/HoneyBee	295	250	1.18
Onthophagus/EAB	253	216	1.17
MPB/HoneyBee	234	210	1.11
Nicrophorus/Onthophagus	247	229	1.08
MPB/EAB	**224**	**209**	**1.07**
Danaus/Drosophila	300	289	1.04
MPB/Drosophila	250	243	1.03
Tribolium/Drosophila	183	179	1.02
Nicrophorus/Danaus	206	210	0.98
Tribolium/HoneyBee	187	193	0.97
Onthophagus/Drosophila	270	285	0.95
Nicrophorus/Tribolium	177	187	0.95

MPB/Onthophagus	218	235	0.93
Aethina/Nicrophorus	183	198	0.92
Onthophagus/HoneyBee	237	267	0.89
MPB/Danaus	187	220	0.85
Nicrophorus/Photinus	221	277	0.80
Onthophagus/Danaus	190	242	0.79
ALB/Onthophagus	136	175	0.78
MPB/Photinus	218	285	0.76
EAB/Danaus	187	250	0.75
Tribolium/EAB	158	212	0.75
ALB/EAB	**135**	**183**	**0.74**
Onthophagus/Photinus	200	274	0.73
Tribolium/Danaus	130	179	0.73
CPB/Nicrophorus	161	224	0.72
EAB/Drosophila	218	309	0.71
Tribolium/Photinus	146	207	0.71
CPB/HoneyBee	174	251	0.69
CPB/Photinus	**159**	**231**	**0.69**
Aethina/HoneyBee	157	230	0.68
Onthophagus/Tribolium	140	206	0.68
Photinus/HoneyBee	228	347	0.66
CPB/EAB	146	223	0.65
Photinus/Drosophila	227	348	0.65
Photinus/EAB	206	323	0.64
Sitophilus/Photinus	**99**	**157**	**0.63**
ALB/Nicrophorus	138	226	0.61

Sitophilus/CPB	**81**	**134**	**0.60**
CPB/Onthophagus	127	217	0.59
CPB/Danaus	128	224	0.57
MPB/Aethina	139	246	0.57
Aethina/Tribolium	104	185	0.56
Photinus/Danaus	169	306	0.55
MPB/ALB	**119**	**223**	**0.53**
Aethina/Drosophila	143	270	0.53
Aethina/Photinus	136	262	0.52
MPB/CPB	131	259	0.51
Sitophilus/Onthophagus	67	138	0.49
Sitophilus/Tribolium	58	120	0.48
CPB/Drosophila	139	293	0.47
ALB/Drosophila	130	276	0.47
Sitophilus/HoneyBee	88	188	0.47
ALB/HoneyBee	119	258	0.46
CPB/Tribolium	87	191	0.46
ALB/Tribolium	76	167	0.46
Sitophilus/Drosophila	86	193	0.45
Sitophilus/EAB	68	164	0.41
CPB/Aethina	91	223	0.41
ALB/Photinus	97	238	0.41
Sitophilus/Nicrophorus	77	189	0.41
Aethina/EAB	116	285	0.41
ALB/Danaus	86	217	0.40
Aethina/Danaus	91	233	0.39

Sitophilus/ALB	48	124	0.39
CPB/ALB	74	200	0.37
Sitophilus/Aethina	53	145	0.37
Aethina/Onthophagus	101	281	0.36
Sitophilus/Danaus	54	159	0.34
Sitophilus/MPB	64	231	0.28
ALB/Aethina	54	260	1

APPENDIX I

TRI-WISE CONVERGENT GENE FAMILY EXPANSIONS (CC) AND CONTRACTIONS (CC), AVERAGE INCREASE BY TRIAD (AVG INCREASE), AVERAGE DECREASE BY TRIAD (AVG DECREASE), PERCENT CE/AVG INCREASE AND PERCENT CC/AVG DECREASE. RED ROW: WBBS (ALB: ASIAN LONGHORN BEETLE (*A. GLABRIPENNIS*); EAB: EMERALD ASH BORER (*A. PLANIPENNIS*); MPB: MOUNTAIN PINE BEETLE (*D. PONDEROSAE*)).. BLUE ROW: SISTER SPECIES OF WBBS (CPB: COLORADO POTATO BEETLE (*L. DECEMLINEATA*)).

Triad	CE	CC	Avg Increase	Avg Decrease	CE/AVG Increase *100	CC/AVG Decrease *100
ALB/Onthophagus/EAB	47	58	673.3	761.3	6.98	7.618
CPB/ALB/Onthophagus	49	25	807.7	630	6.067	3.968
Sitophilus/ALB/Onthophagus	41	14	724	548.7	5.663	2.552
ALB/Onthophagus/Tribolium	36	36	655.3	612.3	5.493	5.879
CPB/ALB/EAB	39	21	739	666.7	5.277	3.15
ALB/Tribolium/Photinus	42	18	801	627.3	5.243	2.869
ALB/Tribolium/Drosophila	31	26	622	737.3	4.984	3.526
Sitophilus/ALB/Photinus	43	15	869.7	563.7	4.944	2.661
Sitophilus/Onthophagus/Photinus	43	30	906.7	692	4.743	4.335
ALB/Onthophagus/Photinus	41	44	887.7	739.7	4.619	5.949
CPB/ALB/Photinus	44	18	953.3	645	4.615	2.791
Sitophilus/CPB/ALB	35	7	789.7	454	4.432	1.542
ALB/Tribolium/EAB	26	30	586.7	649	4.432	4.622
Sitophilus/ALB/EAB	29	15	655.3	585.3	4.425	2.563
Sitophilus/ALB/Tribolium	27	12	637.3	436.3	4.236	2.75
Sitophilus/CPB/Onthophagus	33	16	826.7	582.3	3.992	2.748
Aethina/Tribolium/Drosophila	34	49	862	718.3	3.944	6.821

CPB/Onthophagus/Photinus	39	43	990.3	773.3	3.938	5.56
CPB/ALB/Danaus	35	15	895.3	627.7	3.909	2.39
Sitophilus/CPB/Photinus	**38**	**37**	**972.3**	**597.3**	**3.908**	**6.194**
CPB/Onthophagus/Danaus	36	33	932.3	756	3.861	4.365
ALB/Tribolium/Danaus	28	16	743	610	3.769	2.623
Sitophilus/ALB/Drosophila	26	15	690.7	673.7	3.764	2.227
CPB/Tribolium/Photinus	34	33	903.7	661	3.762	4.992
CPB/Aethina/Photinus	44	39	1193.3	626	3.687	6.23
ALB/EAB/Danaus	28	31	761	759	3.679	4.084
ALB/Onthophagus/Drosophila	26	42	708.7	849.7	3.669	4.943
Sitophilus/Tribolium/Photinus	30	29	820	579.7	3.659	5.003
ALB/Nicrophorus/Tribolium	20	33	551.3	649	3.628	5.085
CPB/ALB/Tribolium	26	9	721	517.7	3.606	1.739
Sitophilus/Tribolium/Drosophila	23	16	641	689.7	3.588	2.32
Aethina/Photinus/Danaus	43	28	1215.3	718.3	3.538	3.898
ALB/Danaus/Drosophila	28	21	796.3	847.3	3.516	2.478
Sitophilus/CPB/Danaus	32	22	914.3	580	3.5	3.793
CPB/Danaus/honeybee	32	57	916.3	804	3.492	7.09
Aethina/Tribolium/Photinus	36	41	1041	608.3	3.458	6.74
CPB/Tribolium/Drosophila	25	31	724.7	771	3.45	4.021
ALB/Nicrophorus/Onthophagus	22	61	638	761.3	3.448	8.012
Sitophilus/Onthophagus/Tribolium	23	16	674.3	564.7	3.411	2.834
ALB/Photinus/Drosophila	29	27	854.3	864.7	3.394	3.123
ALB/Onthophagus/Danaus	28	33	829.7	722.3	3.375	4.569
Tribolium/Photinus/Drosophila	27	50	804.7	880.7	3.355	5.678
ALB/Aethina/Tribolium	28	9	858.3	465	3.262	1.935
CPB/Aethina/Drosophila	33	30	1014.3	736	3.253	4.076
CPB/ALB/Drosophila	25	19	774.3	755	3.229	2.517
Sitophilus/ALB/Danaus	26	8	811.7	546.3	3.203	1.464

MPB/ALB/Onthophagus	24	49	755	774.7	3.179	6.325
CPB/Onthophagus/Tribolium	24	24	758	646	3.166	3.715
ALB/Nicrophorus/EAB	18	52	569.3	798	3.162	6.516
CPB/Photinus/Danaus	34	46	1078	771	3.154	5.966
CPB/Aethina/Onthophagus	33	20	1047.7	611	3.15	3.273
Sitophilus/Photinus/Danaus	31	24	994.3	689.7	3.118	3.48
Aethina/Onthophagus/honeybee	30	38	966	753.7	3.106	5.042
Tribolium/EAB/Danaus	22	41	711.3	775	3.093	5.29
Sitophilus/Aethina/Photinus	34	21	1109.7	544.7	3.064	3.856
Aethina/Onthophagus/Photinus	34	37	1127.7	720.7	3.015	5.134
CPB/EAB/Danaus	26	43	863.7	792.7	3.01	5.425
CPB/Aethina/Danaus	34	29	1135.3	608.7	2.995	4.765
Aethina/Onthophagus/Danaus	32	20	1069.7	703.3	2.992	2.844
ALB/EAB/Drosophila	19	29	640	886.3	2.969	3.272
CPB/Aethina/Tribolium	28	26	961	498.7	2.914	5.214
Sitophilus/Onthophagus/EAB	20	18	692.3	713.7	2.889	2.522
Sitophilus/Aethina/Danaus	30	14	1051.7	527.3	2.853	2.655
MPB/Onthophagus/Danaus	25	51	879.7	900.7	2.842	5.662
CPB/Tribolium/Danaus	24	27	845.7	643.7	2.838	4.195
CPB/Onthophagus/EAB	22	47	776	795	2.835	5.912
CPB/Aethina/honeybee	29	42	1031.7	659	2.811	6.373
MPB/Aethina/Photinus	32	42	1140.7	770.7	2.805	5.45
Sitophilus/Aethina/Drosophila	26	17	930.7	654.7	2.794	2.597
Onthophagus/Photinus/Danaus	28	51	1012.3	865.7	2.766	5.891
CPB/Tribolium/EAB	19	28	689.3	682.7	2.756	4.102
CPB/Aethina/Nicrophorus	26	42	943.7	647.7	2.755	6.485
Onthophagus/Tribolium/Photinus	23	38	838	755.7	2.745	5.029
Sitophilus/ALB/Nicrophorus	17	16	620	585.3	2.742	2.733
Onthophagus/Tribolium/Drosophila	18	53	659	865.7	2.731	6.122

Onthophagus/Tribolium/EAB	17	52	623.7	777.3	2.726	6.69
Nicrophorus/Onthophagus/Tribolium	16	58	588.3	777.3	2.72	7.461
ALB/Photinus/EAB	22	35	819	776.3	2.686	4.508
CPB/Nicrophorus/Danaus	22	49	828.3	792.7	2.656	6.182
Aethina/Photinus/Drosophila	29	53	1094.3	845.7	2.65	6.267
ALB/Nicrophorus/Drosophila	16	49	604.7	886.3	2.646	5.528
Sitophilus/CPB/EAB	20	22	758	619	2.639	3.554
MPB/ALB/Drosophila	19	34	721.7	899.7	2.633	3.779
Sitophilus/CPB/Nicrophorus	19	28	722.7	619	2.629	4.523
MPB/Onthophagus/EAB	19	71	723.3	939.7	2.627	7.556
Sitophilus/Photinus/EAB	22	21	838	728.7	2.625	2.882
CPB/Photinus/Drosophila	25	42	957	898.3	2.612	4.675
MPB/ALB/Danaus	22	25	842.7	772.3	2.611	3.237
Aethina/Nicrophorus/Drosophila	22	85	844.7	867.3	2.605	9.8
Onthophagus/Photinus/Drosophila	23	75	891.3	993	2.58	7.553
Sitophilus/CPB/Tribolium	19	19	740	470	2.568	4.043
ALB/Aethina/Photinus	28	15	1090.7	592.3	2.567	2.532
ALB/Photinus/Danaus	25	22	975.3	737.3	2.563	2.984
CPB/ALB/Nicrophorus	18	27	703.7	666.7	2.558	4.05
Photinus/Danaus/Drosophila	25	52	979	990.7	2.554	5.249
ALB/Nicrophorus/Photinus	20	37	783.7	776.3	2.552	4.766
Sitophilus/CPB/Drosophila	20	25	793.3	707.3	2.521	3.534
Sitophilus/Photinus/Drosophila	22	23	873.3	817	2.519	2.815
MPB/Tribolium/EAB	16	63	636.7	827.3	2.513	7.615
MPB/Aethina/Onthophagus	25	33	995	755.7	2.513	4.367
Sitophilus/Aethina/Tribolium	22	15	877.3	417.3	2.508	3.594
ALB/Nicrophorus/Danaus	18	33	725.7	759	2.48	4.348
Nicrophorus/Onthophagus/EAB	15	82	606.3	926.3	2.474	8.852
Sitophilus/Onthophagus/Drosophila	18	19	727.7	802	2.474	2.369

Tribolium/Photinus/EAB	19	45	769.3	792.3	2.47	5.679
Aethina/Danaus/honeybee	26	48	1053.7	751.3	2.468	6.389
Aethina/Onthophagus/Drosophila	23	43	948.7	830.7	2.424	5.177
CPB/Photinus/EAB	22	45	921.7	810	2.387	5.556
Tribolium/EAB/Drosophila	14	50	590.3	902.3	2.372	5.541
CPB/Nicrophorus/Photinus	21	51	886.3	810	2.369	6.296
Sitophilus/Nicrophorus/Photinus	19	29	802.7	728.7	2.367	3.98
Aethina/Drosophila/honeybee	22	72	932.7	878.7	2.359	8.194
Aethina/Onthophagus/Tribolium	21	20	895.3	593.3	2.345	3.371
MPB/Photinus/Danaus	24	44	1025.3	915.7	2.341	4.805
Aethina/Tribolium/Danaus	23	28	983	591	2.34	4.738
Onthophagus/Photinus/EAB	20	70	856	904.7	2.336	7.738
MPB/ALB/EAB	**16**	**40**	**686.3**	**811.3**	**2.331**	**4.93**
ALB/Aethina/Onthophagus	22	20	945	577.3	2.328	3.464
MPB/Danaus/honeybee	20	71	863.7	948.7	2.316	7.484
Onthophagus/EAB/honeybee	16	79	694.3	937.7	2.304	8.425
ALB/Aethina/Drosophila	21	16	911.7	702.3	2.303	2.278
Aethina/EAB/Danaus	23	29	1001	740	2.298	3.919
CPB/Nicrophorus/Onthophagus	17	48	740.7	795	2.295	6.038
ALB/Aethina/EAB	20	23	876.3	614	2.282	3.746
MPB/Aethina/honeybee	22	78	979	803.7	2.247	9.706
MPB/ALB/Tribolium	15	28	668.3	662.3	2.244	4.227
MPB/EAB/Danaus	18	55	811	937.3	2.219	5.868
ALB/Onthophagus/honeybee	16	45	726	772.7	2.204	5.824
Onthophagus/Tribolium/Danaus	17	36	780	738.3	2.179	4.876
Aethina/Tribolium/EAB	18	36	826.7	630	2.177	5.714
Nicrophorus/Tribolium/Drosophila	12	85	555	902.3	2.162	9.42
Aethina/Photinus/honeybee	24	48	1111.7	768.7	2.159	6.245
Sitophilus/ALB/Aethina	20	7	927	401.3	2.157	1.744

CPB/Drosophila/honeybee	17	51	795.3	931.3	2.137	5.476
ALB/Aethina/Danaus	22	14	1032.7	575	2.13	2.435
MPB/Danaus/Drosophila	18	58	846.3	1025.7	2.127	5.655
Sitophilus/Tribolium/honeybee	14	22	658.3	612.7	2.127	3.591
MPB/Aethina/Danaus	23	35	1082.7	753.3	2.124	4.646
Sitophilus/Onthophagus/Danaus	18	18	848.7	674.7	2.121	2.668
CPB/Danaus/Drosophila	19	30	899	881	2.113	3.405
MPB/Onthophagus/Drosophila	16	69	758.7	1028	2.109	6.712
CPB/Onthophagus/Drosophila	17	36	811.3	883.3	2.095	4.075
EAB/Danaus/Drosophila	16	39	764.7	1012.3	2.092	3.852
Aethina/Onthophagus/EAB	19	38	913.3	742.3	2.08	5.119
MPB/EAB/Drosophila	14	47	690	1064.7	2.029	4.415
CPB/EAB/Drosophila	15	36	742.7	920	2.02	3.913
Aethina/EAB/honeybee	18	56	897.3	790.3	2.006	7.086
Onthophagus/EAB/Danaus	16	62	798	887.3	2.005	6.987
MPB/Aethina/Tribolium	18	49	908.3	643.3	1.982	7.617
CPB/Nicrophorus/Drosophila	14	54	707.3	920	1.979	5.87
MPB/Aethina/Drosophila	19	59	961.7	880.7	1.976	6.699
CPB/EAB/honeybee	15	65	760	843	1.974	7.711
Aethina/Nicrophorus/honeybee	17	102	862	790.3	1.972	12.906
Sitophilus/Aethina/Onthophagus	19	9	964	529.7	1.971	1.699
Nicrophorus/Onthophagus/Danaus	15	56	762.7	887.3	1.967	6.311
MPB/Photinus/honeybee	18	63	921.7	966	1.953	6.522
Tribolium/Photinus/Danaus	18	40	925.7	753.3	1.945	5.31
MPB/Tribolium/Drosophila	13	72	672	915.7	1.935	7.863
CPB/Nicrophorus/EAB	13	58	672	831.7	1.935	6.974
Aethina/Tribolium/honeybee	17	59	879.3	641.3	1.933	9.2
Sitophilus/Nicrophorus/Tribolium	11	24	570.3	601.3	1.929	3.991
Sitophilus/Nicrophorus/Drosophila	12	31	623.7	838.7	1.924	3.696

Nicrophorus/Tribolium/Danaus	13	49	676	775	1.923	6.323
Sitophilus/Danaus/honeybee	16	25	832.7	722.7	1.922	3.459
Onthophagus/Danaus/Drosophila	16	73	833.3	975.7	1.92	7.482
ALB/Aethina/Nicrophorus	16	28	841	614	1.902	4.56
Aethina/Nicrophorus/Tribolium	15	57	791.3	630	1.896	9.048
MPB/Tribolium/Danaus	15	51	793	788.3	1.892	6.469
MPB/CPB/Aethina	20	34	1060.7	661	1.886	5.144
Sitophilus/Nicrophorus/Danaus	14	24	744.7	711.3	1.88	3.374
CPB/ALB/Aethina	19	1	1010.7	482.7	1.88	0.207
ALB/Tribolium/honeybee	12	34	639.3	660.3	1.877	5.149
Nicrophorus/Onthophagus/Drosophila	12	83	641.7	1014.7	1.87	8.18
Tribolium/Drosophila/honeybee	12	89	643	913.7	1.866	9.741
Aethina/Nicrophorus/Danaus	18	53	965.7	740	1.864	7.162
MPB/Onthophagus/Tribolium	13	53	705.3	790.7	1.843	6.703
MPB/EAB/honeybee	13	81	707.3	987.7	1.838	8.201
CPB/Nicrophorus/Tribolium	12	37	654	682.7	1.835	5.42
Tribolium/Danaus/honeybee	14	62	764	786.3	1.832	7.885
Nicrophorus/Onthophagus/Photinus	15	80	820.7	904.7	1.828	8.843
Sitophilus/Nicrophorus/Onthophagus	12	26	657	713.7	1.826	3.643
ALB/EAB/honeybee	12	53	657.3	809.3	1.826	6.549
Aethina/Nicrophorus/Onthophagus	16	52	878	742.3	1.822	7.005
Sitophilus/Tribolium/EAB	11	21	605.7	601.3	1.816	3.492
CPB/Onthophagus/honeybee	15	46	828.7	806.3	1.81	5.705
Photinus/Danaus/honeybee	18	71	996.3	913.7	1.807	7.771
MPB/CPB/Photinus	18	42	1003.3	823.3	1.794	5.101
Nicrophorus/Tribolium/Photinus	13	57	734	792.3	1.771	7.194
CPB/ALB/honeybee	14	26	791.7	678	1.768	3.835
MPB/CPB/Onthophagus	15	35	857.7	808.3	1.749	4.33
Sitophilus/CPB/Aethina	18	21	1029.7	435	1.748	4.828

Sitophilus/Aethina/Nicrophorus	15	22	860	566.3	1.744	3.885
Sitophilus/Tribolium/Danaus	13	13	762	562.3	1.706	2.312
Sitophilus/Aethina/honeybee	16	21	948	577.7	1.688	3.635
Sitophilus/Drosophila/honeybee	12	33	711.7	850	1.686	3.882
Nicrophorus/Onthophagus/honeybee	11	93	659	937.7	1.669	9.918
Sitophilus/EAB/Drosophila	11	20	659	838.7	1.669	2.385
MPB/CPB/honeybee	14	50	841.7	856.3	1.663	5.839
Nicrophorus/Photinus/Drosophila	13	75	787.3	1029.7	1.651	7.284
Onthophagus/Drosophila/honeybee	12	112	729.7	1026	1.645	10.916
CPB/Aethina/EAB	16	29	979	647.7	1.634	4.478
Sitophilus/EAB/honeybee	11	32	676.3	761.7	1.626	4.201
CPB/Tribolium/honeybee	12	43	742	694	1.617	6.196
ALB/Danaus/honeybee	13	36	813.7	770.3	1.598	4.673
Sitophilus/Aethina/EAB	14	11	895.3	566.3	1.564	1.942
Sitophilus/ALB/honeybee	11	15	708	596.7	1.554	2.514
Nicrophorus/Photinus/Danaus	14	57	908.3	902.3	1.541	6.317
CPB/Photinus/honeybee	15	60	974.3	821.3	1.54	7.305
Nicrophorus/Tribolium/EAB	8	58	519.7	814	1.539	7.125
MPB/ALB/Nicrophorus	10	43	651	811.3	1.536	5.3
EAB/Danaus/honeybee	12	88	782	935.3	1.535	9.408
Onthophagus/Danaus/honeybee	13	76	850.7	898.7	1.528	8.457
MPB/CPB/EAB	12	46	789	845	1.521	5.444
MPB/Onthophagus/Photinus	14	63	937.7	918	1.493	6.863
Photinus/EAB/Danaus	14	51	943.7	902.3	1.484	5.652
Tribolium/EAB/honeybee	9	68	607.7	825.3	1.481	8.239
Sitophilus/CPB/honeybee	12	29	810.7	630.3	1.48	4.601
Sitophilus/Danaus/Drosophila	12	10	815.3	799.7	1.472	1.251
MPB/CPB/Drosophila	12	36	824.3	933.3	1.456	3.857
MPB/Tribolium/honeybee	10	82	689.3	838.7	1.451	9.777

Aethina/Danaus/Drosophila	15	33	1036.3	828.3	1.447	3.984
ALB/Nicrophorus/honeybee	9	48	622	809.3	1.447	5.931
MPB/Photinus/Drosophila	13	54	904.3	1043	1.438	5.177
Photinus/EAB/honeybee	12	78	840	952.7	1.429	8.188
MPB/Tribolium/Photinus	12	71	851	805.7	1.41	8.813
Sitophilus/Nicrophorus/honeybee	9	35	641	761.7	1.404	4.595
Nicrophorus/Tribolium/honeybee	8	95	572.3	825.3	1.398	11.511
MPB/CPB/Danaus	13	43	945.3	806	1.375	5.335
MPB/Nicrophorus/Drosophila	9	91	654.7	1064.7	1.375	8.547
Nicrophorus/Danaus/Drosophila	10	75	729.3	1012.3	1.371	7.409
Aethina/Nicrophorus/Photinus	14	64	1023.7	757.3	1.368	8.451
MPB/ALB/honeybee	10	37	739	822.7	1.353	4.498
Sitophilus/Onthophagus/honeybee	10	22	745	725	1.342	3.034
MPB/CPB/ALB	11	17	820.7	680	1.34	2.5
Tribolium/Danaus/Drosophila	10	57	746.7	863.3	1.339	6.602
MPB/ALB/Photinus	12	33	900.7	789.7	1.332	4.179
MPB/Nicrophorus/Tribolium	8	84	601.3	827.3	1.33	10.153
Onthophagus/EAB/Drosophila	9	71	677	1014.7	1.329	6.997
ALB/Drosophila/honeybee	9	41	692.7	897.7	1.299	4.567
MPB/Aethina/EAB	12	42	926.3	792.3	1.295	5.301
ALB/Aethina/honeybee	12	17	929	625.3	1.292	2.719
MPB/Onthophagus/honeybee	10	70	776	951	1.289	7.361
Sitophilus/EAB/Danaus	10	18	780	711.3	1.282	2.53
Photinus/Drosophila/honeybee	11	95	875.3	1041	1.257	9.126
Aethina/EAB/Drosophila	11	42	880	867.3	1.25	4.842
Aethina/Photinus/EAB	13	30	1059	757.3	1.228	3.961
Tribolium/Photinus/honeybee	10	61	822	803.7	1.217	7.59
MPB/Drosophila/honeybee	9	86	742.7	1076	1.212	7.993
Onthophagus/Photinus/honeybee	11	67	908.7	916	1.211	7.314

Nicrophorus/Danaus/honeybee	9	102	746.7	935.3	1.205	10.905
Sitophilus/Nicrophorus/EAB	7	25	588.3	750.3	1.19	3.332
Onthophagus/Tribolium/honeybee	8	60	676.3	788.7	1.183	7.608
ALB/Photinus/honeybee	10	34	871.7	787.7	1.147	4.317
Aethina/Nicrophorus/EAB	9	62	809.3	779	1.112	7.959
Nicrophorus/Photinus/EAB	8	64	752	941.3	1.064	6.799
Nicrophorus/EAB/Drosophila	6	70	573	1051.3	1.047	6.658
Nicrophorus/EAB/honeybee	6	109	590.3	974.3	1.016	11.187
Sitophilus/Photinus/honeybee	9	30	890.7	740	1.01	4.054
MPB/Aethina/Nicrophorus	9	81	891	792.3	1.01	10.223
Photinus/EAB/Drosophila	8	49	822.7	1029.7	0.972	4.759
CPB/Nicrophorus/honeybee	7	77	724.7	843	0.966	9.134
MPB/ALB/Aethina	9	10	958	627.3	0.939	1.594
MPB/Nicrophorus/honeybee	6	117	672	987.7	0.893	11.846
Nicrophorus/EAB/Danaus	6	62	694	924	0.865	6.71
MPB/Photinus/EAB	7	67	869	954.7	0.806	7.018
Nicrophorus/Drosophila/honeybee	5	142	625.7	1062.7	0.799	13.363
MPB/CPB/Nicrophorus	6	58	753.7	845	0.796	6.864
MPB/Nicrophorus/Danaus	6	71	775.7	937.3	0.774	7.575
EAB/Drosophila/honeybee	5	77	661	1062.7	0.756	7.246
MPB/Nicrophorus/Photinus	6	75	833.7	954.7	0.72	7.856
MPB/Nicrophorus/Onthophagus	4	89	688	939.7	0.581	9.471
MPB/CPB/Tribolium	4	32	771	696	0.519	4.598
Nicrophorus/Photinus/honeybee	4	103	804.7	952.7	0.497	10.812
MPB/Nicrophorus/EAB	2	85	619.3	976.3	0.323	8.706
Sitophilus/MPB/Tribolium	1	17	687.3	614.7	0.145	2.766
Sitophilus/MPB/Drosophila	1	10	740.7	852	0.135	1.174
Sitophilus/MPB/Onthophagus	1	14	774	727	0.129	1.926
Sitophilus/MPB/Aethina	1	3	977	579.7	0.102	0.518

Danaus/Drosophila/honeybee	0	118	817.3	1023.7	0	11.527
Sitophilus/MPB/Photinus	0	20	919.7	742	0	2.695
Sitophilus/MPB/EAB	0	12	705.3	763.7	0	1.571
Sitophilus/MPB/Nicrophorus	0	11	670	763.7	0	1.44
Sitophilus/MPB/ALB	0	7	737	598.7	0	1.169
Sitophilus/MPB/Danaus	0	8	861.7	724.7	0	1.104
Sitophilus/MPB/CPB	0	4	839.7	632.3	0	0.633
Sitophilus/MPB/honeybee	0	4	758	775	0	0.516

APPENDIX J

SUMMARY OF GENE NUMBERS BY PCWDE GROUPS FOR EACH SPECIES AND AVERAGES BY PHENOTYPIC CATEGORIES. RED COLUMNS: WBBS. BLUE COLUMNS: SISTER SPECIES OF WBBS.

PCWDE	MPB	ALB	EAB	Sor	CPB	Ppy	Atu	Nve	Ota	Tca	Ame	Dme	Dpl	WBBs	WBBs-SS	Other Beetles	Outgroups
GH1	22	60	36	16	33	11	9	6	12	10	2	2	24	39.3	20.0	9.3	9.3
GH28	16	19	9	6	13	0	0	0	0	0	0	0	0	14.7	6.3	0.0	0.0
GH32	3	1	4	2	0	0	2	0	0	0	1	0	3	2.7	0.7	0.5	1.3
GH45	9	2	0	8	11	0	0	0	0	0	0	0	0	3.7	6.3	0.0	0.0
GH48	5	1	0	2	2	0	0	0	0	0	0	0	0	2.0	1.3	0.0	0.0
CE8	13	0	0	5	0	0	0	0	0	0	0	0	0	4.3	1.7	0.0	0.0
PL4	8	0	11	0	0	0	0	0	0	0	0	0	0	6.3	0.0	0.0	0.0
GH43	0	0	1	0	0	0	0	0	0	0	0	0	0	0.3	0.0	0.0	0.0
GH44	0	0	4	0	0	0	0	0	0	0	0	0	0	1.3	0.0	0.0	0.0
GH9	0	1	1	0	0	1	0	1	0	1	1	0	0	0.7	0.3	0.5	0.3