Molecular ecological characterization of a honey bee ectoparasitic mite, *Tropilaelaps mercedesae*.

Thesis submitted in accordance with the requirements of the

University of Liverpool for the degree of Doctor in Philosophy

By

Xiaofeng Dong

March 2016

Table of contents

Abstract

Tropilaelaps mercedesae (small mite) is one of two major honey bee ectoparasitic mite species responsible for the colony losses of *Apis mellifera* in Asia. Although *T. mercedesae* mites are still restricted in Asia (except Japan), they may diffuse all over the world due to the ever-increasing global trade of live honey bees (ex. *Varroa destructor*). Understanding the ecological characteristics of *T. mercedesae* at molecular level could potentially result in improving the management and control programs. However, molecular and genomic characterization of *T. mercedesae* remains poorly studied, and even no genes have been deposited in Genbank to date. Therefore, I conducted *T. mercedesae* genome and transcriptome sequencing. By comparing *T. mercedesae* genome with other arthropods, I have gained new insights into evolution of Parasitiformes and the evolutionary changes associated with specific habitats and life history of honey bee ectoparasitic mite that could potentially improve the control programs of *T. mercedesae*. Finally, characterization of *T. mercedesae* transient receptor potential channel, subfamily A, member 1 (TmTRPA1) would also help us to develop a novel control method for *T. mercedesae*.

Background

Apis mellifera, the western honey bee, is the major species used for apiculture around the globe. The value of its contribution to agriculture in the form of providing pollination services worth US$215 billion annually worldwide and the number of managed colonies has increased 45% since 1961 (Smith et al., 2013). Mite parasitization is one of the most important causes inducing colony losses of *A. mellifera*. In China, two species of honey bee ectoparasitic mites, *Tropilaelaps mercedesae* (small mite) and *Varroa destructor* (big mite), are prevalent (Luo et al., 2011). They are serious ecoparasites of honey bee causing direct impacts on bee health as well as indirect effects caused by vectoring viruses and other bee pathogens, affecting both developing brood and adult honey bees. Parasitisation by these mites causes abnormal brood development, death of both broods and adult bees, leading to colony decline and collapse, and causes the bees to abscond from the hive (Sammataro et al., 2000). As a consequence, this will impose negative impacts on ecosystem and agriculture through disruption to pollination services.

Tropilaelaps mites are found throughout Asia, such as Pakistan, Afghanistan, Papua New Guinea, Sri Lank, India, Nepal, Malaysia, Thailand, and Korea, except Japan (Anderson and Morgan 2007; Figure 0.1). Although *Tropilaelaps* mites have not been found in Oceania, America, and Europe, they may diffuse all over the world due to global trade of honey bee and honey bee products. The same already happened with *Varroa* mite, which infested *A. mellifera* colonies brought in Asia from the original host, *Apis cerana*, in the middle of 1950s, and now parasitizes *A. mellifera* colonies all over the world except Australia (Oldroyd, 1999). Thus, *Tropilaelaps* mite could become a major threat in not only Asia but also Europe and US apiculture industries.

Compared to *Varroa* mites, *Tropilaelaps* mites are smaller (approximately 1.0 mm long x 0.6 mm wide), flatter, and more oval-shaped. *Tropilaelaps* mites move fast on the hive frame and only feed the haemolymph of honey bee larva. To begin their reproduction, gravid female mites enter worker or drone brood cells before capping, feed on larval haemolymph for about 2 days or less, and then each female mite lays 1-4 eggs (but typically 3-4) about 1 day apart (Anderson and Morgan 2013). The egg, larva, protonymph, deutonymph, and young adult mite can appear in a single brood cell (Anderson and Morgan 2013). The developmental time from egg to adult takes about 6-7 days and the first egg will normally develop into a male (Anderson and Morgan 2013). By my experiences, only female mites (mother mite and her offsprings) will emerge from the brood cell with the adult bee. They have probably mated before emerging, and then enter other brood cells with no more than a 3-day phoretic phase (Anderson and Morgan 2013). The phoretic phase of *T. mercedesae* may be as short as 1-2 days that are much shorter than that of *Varroa* mites (Anderson and Morgan 2013). Therefore, *T. mercedesae* has quicker reproductive cycle than *Varroa* mite, and hence their population builds up much faster.

T. mercedesae has been controlled by the same methods as *V. destructor* in China to date, but these methods may not be fully effective against them, and the mites have been reported to develop the resistance against commonly used miticides. Molecular ecology was defined by Weiss (1950) as, "the entire continuum of biological interactions between the molecular, cellular, organismal levels to the environment" (Lambert 1995). In my research project, I sequenced *T. mercedesae* genome and transcriptomes in order to provide deep insights into the evolution of this parasite and analyze the genes associated with detoxification, chemoreception, and reproduction systems as well as identify potential target proteins for the control based on the KEGG annotated metabolic pathways. I also characterized the *T. mercedesae* transient receptor potential channel, subfamily A, member 1 (TmTRPA1) to ask if it can be used to develop a novel control method for *T. mercedesae*.

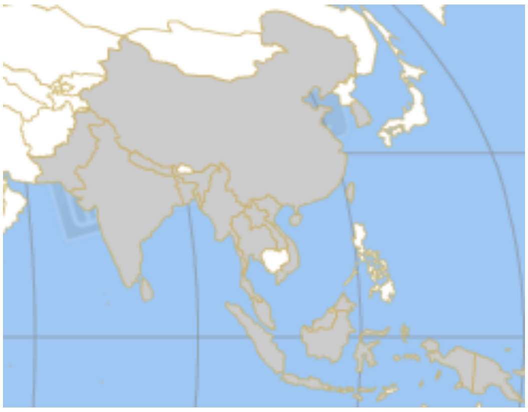


Figure 0.1 Distribution map of *Tropilaelaps* mites in Asia. *Tropilaelaps* mites have been reported in the countries shown by grey colour.

1. Genome and transcriptome sequencing

1.1 Introduction

Next-generation sequencing (NGS) platforms are commonly used to sequence a library constructed with short template DNA fragments, typically about 200-1000 bp, and each template contains forward and reverse primer-binding sites. The sequencing machine reads the DNA starting from both ends of the template and grows along each primer-binding sites, producing two reads that overlap or is separated by a short gap (insert) of approximately known length. This process is known as paired-end sequencing. In the later stages of the genome assembly process, these paired-end reads are used to combine contigs into larger scaffolds and employed as a measure for testing the quality of the assembled genome (El-Metwally et al., 2013).

Prior to the genome sequencing of *T. mercedesae* with NGS methods, an accurate estimation of the genome size is necessary to ensure sufficient sequencing coverage and to provide a firm reference for genome assembly. Propidium iodide staining with flow cytometry is a powerful method to robustly estimate eukaryotic nuclear DNA content before the whole-genome sequencing. Compared to the other most commonly used method, densitometry, flow cytometry has the advantage that is relatively fast, works with a wide variety of materials and provides information on a very large number of nuclei (Bennett et al, 2008). qPCR has been suggested as an alternative method to estimate the nuclear DNA content (Gao and Scott, 2006). Although it does not require nuclear isolation and can be conducted with scarce materials in a molecular laboratory, the genome size estimated by qPCR method is often smaller than those determined by densitometry and flow cytometry (Gao and Scott, 2006).

In the method of propidium iodide staining with flow cytometry, nuclei are stained with propidium iodide, which intercalates into the major groove of DNA and RNA without sequence preference. After RNase removes the RNA, unknown samples are typically co-stained with standard nuclei of a known genome size and ploidity, and the nuclear DNA content of the unknown can be calculated using the relative fluorescence of the DNA in the unknown and standard. If the ploidity information of target organism is known, the genome size can be then estimated.

Because there is not ploidity information of *T. mercedesae* to date, in this chapter, I would combine the flow cytometry with the k-mer analysis, which can be used to estimate the genome size without prior knowledge (Xu et al., 2011; Li et al., 2012; Li et al., 2014), to determine the genome size and ploidy of *T. mercedesae*.

1.2 Materials and Methods

1.2.1 Mite collection

Morphological differences between adult male and female *T. mercedesae* were previously described (Anderson and Morgan 2007). Adult males have a pair of long and attenuate sperm transfer organ (movable chela, functioning as spermatodactyl) with a spirally coiled apex, whereas, females have a small subapical tooth on the movable chela of chelicerae (Figure 1.1A and B). Adult females also have an overlap on the posterior margin of the anal plate with the apex of the epigynial plate. This overlap is an obvious characteristic can be used to quickly separate females from males (Figure 1.2A, B). But the gravid females can also show well-separated plates (Figure 1.2D), due to their swollen abdomen. As an alternative, the arrow shape (sharply pointed) epigynial plate of adult males can be the other important obvious characteristic to differentiate the adult male and female. The nymph stages are very easily recognized due to their plain white body color and invisible underdeveloped anal and epigynial plate (Figure 1.2C). Experientially, there are also some ethological differences between adult male and female *T. mercedesae*, for example, non gravid (or early gravid) females like to mount up brood cells and move very fast on frame, but interestingly, gravid females behaving more like males and prefer to stay in brood cells.

Based on these morphological and ethological characteristics, the male, female, and female nymph of *T. mercedesae* were identified and collected directly for flow cytometric analysis and Illumina (genome and transcriptome) sequencing (detailed below) on three occasions in the single mite-infested honey bee colony purchased from Wuxi, China. Although many methods have been tested by Luo (2011), laboratory culture of *Tropilaelaps* mites independent from honey bees is still not possible. The homozygosity of the sequenced mite samples could not be guaranteed by sibmating (i.e. Intraspecific inbreeding) in the lab, but according to the life cycle of *T. mercedesae* (see section 2.4; Background section; Terrestrial Animal Health Code 2010), the mites carry out sibmating naturally, except that more than one mother mite simultaneously colonizes a single brood cell.

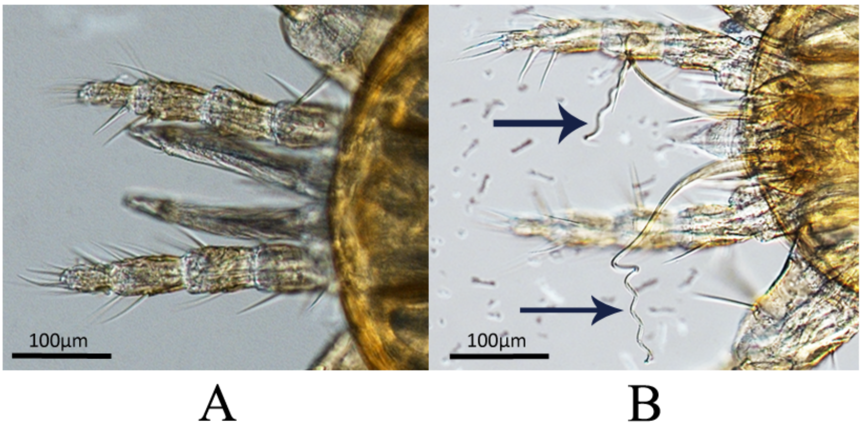


Figure 1.1 Mouthparts of *T. mercedesae*. A: Coiled apex of male chela spermatodactyl of *T. mercedesae*. B: Small subapical tooth on the movable chela of female chelicerae of *T. mercedesae*.

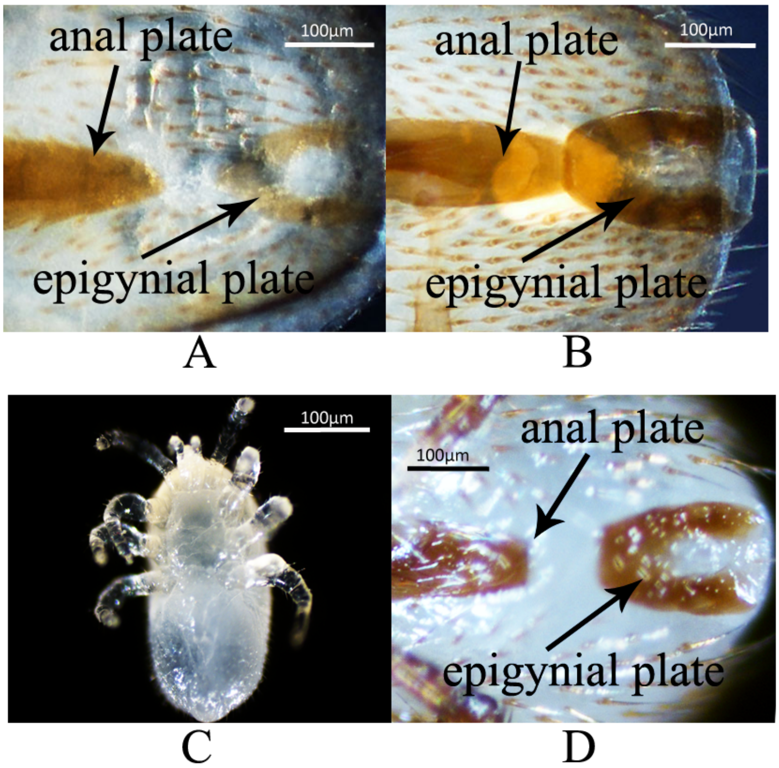


Figure 1.2 Morphological differences of nymph, adult male and female *T. mercedesae.* A: Separated anal and epigynial plates in a male *T. mercedesae*. B: Attached anal and epigynial plates in the female. C: Underdeveloped anal and epigynial plates with plain white body colour in the nymph. D: Separated anal and epigynial plates in the gravid female.

1.2.2 Flow cytometric analysis of *T. mercedesae* nuclear DNA content

Ten whole *T. mercedesae* males and ten females were identified and collected directly from the honey bee hive as described above. In order to analyze unreplicated nuclear DNA (G1 phase of the cell cycle), only the heads of ten adult *Drosophila melanogaster* females (1C=175Mb; Bennett, MD et al., 2003) and a diploid honey bee worker (1C=262 Mb; The Honey bee Genome Sequencing Consortium 2006) were collected as reference standards. These samples were then homogenized separately in 1 ml cold Galbraith buffer [30 mm sodium citrate, 18 mm MOPS (3-morpholinopropanesulfonic acid), 21 mm mgcl$_2$, 0.1% Triton X-100, 0.1% rnaase A] (Galbraith et al. 1983) using a loose pestle (stroke 15 times) and filtered through 20 μm nylon mesh to remove cellular debris. Propidium iodide (1 mg/ml; Sigma, St. Louis, MO) was added to the samples to a final concentration of 50 μg/ml and incubated under dark for 30 min at 4 °C. The stained male and female mite nuclei were analyzed separately with two reference standards using a BD Facscalibur flow cytometer (BD Biosciences, San Jose, CA). Both forward and side scatters were used to gate out anything except whole, intact nuclei, which scatter very little light. Because propidium iodide is excited at 488 nm and emits at 585 nm, DNA flow cytometry was set to count only red fluorescent nuclei in FL2 channel (564-606 nm). The gain of the instrument was adjusted so that the peak representing G1 nuclei of the first internal standard (*D. melanogaster*) was positioned on FL2-A (FL2 area signal) = 50. On average, approximately 10,000 nuclei were counted in each run. Peak means (peak position weighted by peak area) and histograms (Figure 1.3) were calculated and plotted using Modfit software (Verity Software House). Nuclear genome size was then calculated according to the following formula (Cires et al., 2011):

Sample nuclear DNA content = (sample G1 peak mean/standard G1 peak mean) × nuclear DNA content of standard

1.2.3 DNA library construction

Male and female *T. mercedesae* were identified and collected directly from the honey bee hive (methods as described above), stored in acetone at room temperature until use. Before DNA extraction, mite bodies were carefully washed twice with fresh acetone to remove any non-target organisms that might adhere on the mite surface. Subsequently, 20 male and 25 female mites were air dried (15 minutes) and individually triturated in 180 µL of lysozyme buffer, (1M Tris-hcl, 0.5M EDTA, 1.2% Triton X-100, and 0.02% lysozyme) with a tissuelyser II (Qiagen, Valencia, CA) using a 3 mm stainless steel bead at 25,000 motions per minute for 30 seconds. After incubation at 37 °C for 30 min, total DNA was extracted from each of the triturated samples with dneasy Blood and Tissue kit (Qiagen, Valencia, CA) by following the manufacturer's spin-column protocol for animal tissue. To maximize the yield of DNA extraction, two successive elution steps, each with 50 µl elution buffer, were performed. The DNA concentrations were determined by spectrophotometry, a sensitive and commonly used fluorescent dye assay (Qubit® dsdna

BR assay, Life Technologies Europe, Naerum, Denmark) according to the manufacturer's instructions.

Possible bacterial contamination in the extracted genome DNA was examined using PCR to amplify bacterial *16S rrna* gene with the conserved primers UFPL (5’-AGTTTGATCCTGGCTCAG-3’) and URPL (5’-GGTTACCTTGTTACGACTT-3’) (Coenye et al. 2002). The reaction mixture was prepared using 10 μL of Taq DNA polymerase mix (Mytaq Red Mix Bioline® Ltd), 1 μL of each primer (100 nmol), 1 μL of DNA template, and 7 μL distilled water. The PCR thermal cycling conditions were the followings; 95°C, 2 min followed by 30 or 35 cycles of 95°C, 15 sec, 55°C, 15 sec, and 72°C, 1 min.

Two paired-end Illumina DNA libraries were constructed with a bacteria-free male and a bacteria-contaminated female total genomic DNA samples (30ng each) using a Nextera DNA sample preparation kit (Illumina, Great Chesterford, United Kingdom; Table 1.1). Agilent high sensitivity DNA kit and the 2100 Bioanalyzer (Agilent Technologies) were used to check the quality and measure the insert size of the libraries. Both steps in this paragraph were conducted by following the manufacturer’s instructions.

1.2.4 Illumina DNA and RNA sequencing

The DNA libraries were then sequenced with Illumina Hiseq 2500 system in the Centre for Genomic Research (CGR) at the University of Liverpool. Male, female and nymph mites (each with 20~30; identified and collected as described above) were shipped to Beijing Genomics Institute (BGI) for total RNA extraction, mRNA (with poly-A) purification, cDNA library preparation, and Illumina Hiseq 2000 sequencing. The sequencing data are summarized in Table 1.1.

1.2.5 Genome size estimation by *k*-mer analysis

A *k*-mer refers to an artificial sequence division of sequencing reads into all the possible substrings with a length of *K* nucleotides. A raw sequence read with *L* bp contains (*L*-*K*+1) *k*-mers if the length of each *k*-mer is *K* bp. For example, the read of AGCTCGAGTG can produce 8 *k*-mers (AGC GCT CTC TCG CGA GAG AGT GTG) when K is equal to 3. Because the *K*-mer frequencies along the sequence depth gradient follow a Poisson distribution, the frequency of each *k*-mer can be calculated from the genome-sequence reads. The genome size was estimated by analyzing the occurrence and distribution of K-mers with following formula:

$$\text{Estimated genome size (bp)} = \text{K-mer number} / \text{K-mer depth},$$

Where the K-mer number is the total number of *k*-mer, and K-mer depth is the maximal frequency (Zhang et al, 2012). Here, I used command-line program of jellyfish (Marcais and Kingsford 2011) with a k-mer size of 31 and default setting to count k-mer frequency in both male and female genome sequencing data. The k-mer frequency distribution curves were plotted with the k-mer depth as the x-axis and the k-mer frequency as the y-axis to find the K-mer depths (Figure 1.4A and B).

1.3 Results and Discussion

1.3.1 Nuclear DNA content of *T. mercedesae*

During each sample run in the flow cytometry system, it is essential to confirm the linearity of fluorescence measurements. Because the nuclear DNA content measurement depends on the nuclear DNA content directly proportional to the measured fluorescence, the measurement quality could be confirmed by: (1) *D. melanogaster* and *A. Mellifera* 4C peaks were twice that of their 2C peaks in all three runs, confirming linearity of the fluorescence measurements; (2) The fluorescence peak mean ratio between *D. melanogaster* and *A. Mellifera* was about 1.5, which was same as their nuclear DNA content ratio, indicating that the reference standards worked.

Two reference standard samples were measured first to adjust parameters of flow cytometry and test if the existing artificial peaks (Figure 1.3A). The nuclear DNA contents of cells from *T. mercedesae* male and female were then calculated by averaging the sizes estimated by both *D. melanogaster* and *A. Mellifera* 2C standards, respectively (Figure 1.3B, C). As expected, the estimated nuclear DNA content of female mite (~1,287 Mb) was nearly double that of the male mite (~660 Mb). The *T. mercedesae* female nuclei showed a relatively broad peak, suggesting that the big nuclei become more heterogeneous by shearing the DNA during homogenization of the samples. This may suggest the nuclear DNA content estimated from the male is more accurate.

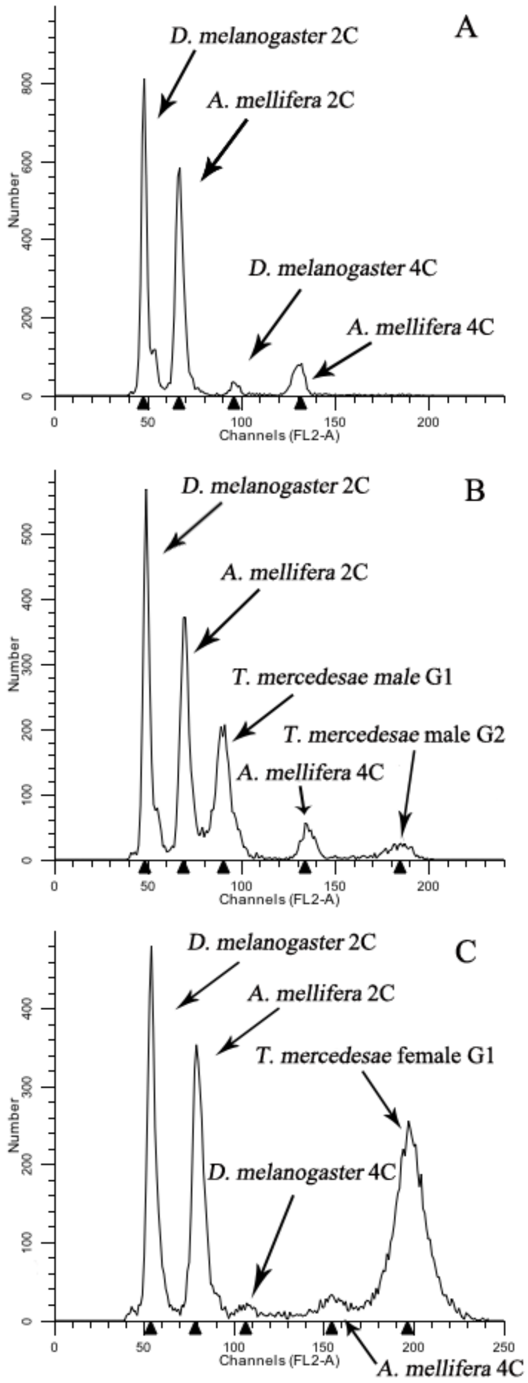


Figure 1.3 Histograms of counted nuclei number against red fluorescent (FL2-A) signals. Recorded by flow cytometry from propidium iodide stained *D. melanogaster* and *A. Mellifera* nuclei (A); *D. melanogaster*, honey bee and *A. Mellifera* male nuclei (B); *D. melanogaster*, *A. Mellifera*, *T. mercedesae* female nuclei (C).

1.3.2 Genome size and ploidy of *T. mercedesae*

The sex determination system of mites and ticks in Acari remains poorly studied, but the western orchard *Metaseiulus occidentalis* (Predatory mite) was suggested to have the sex determination system known as parahaploidy (Nelson-Rees et al., 1980). The haplodiploidy is known as paternal genome elimination, in which both males and females develop from diploid fertilized eggs. During embryogenesis in males, however, the paternal set of chromosomes is functionally inactivated by heterochromatinization and they develop to the functionally haploid adults (Nelson-Rees et al., 1980).

The flow cytometry results found the female (~1,287Mb) *T. mercedesae* has a double genome DNA content of the male (~660Mb). The K-mer (K=31) statistics estimated the *T. mercedesae* genome size as ~660Mb for male mite and ~628Mb for female mite (Figure 1.4A and B). These results together with the nuclear DNA contents estimated by flow cytometry suggested that *T. mercedesae* might follow a parahaploidy sex determination system, in which females are diploid and males are haploid. There is not sex chromosome in parahaploidy genetic system reported previously; sex is probably determined by a functional haploidy/diploidy mechanism. Thus, the smaller genome size of female compared to male by the k-mer analysis may be the result of technological bias occurred during the experimental procedures, for example, the isolation of genomic DNA, the construction of NGS libraries, and the high throughput sequencing step (Kim et al., 2014).

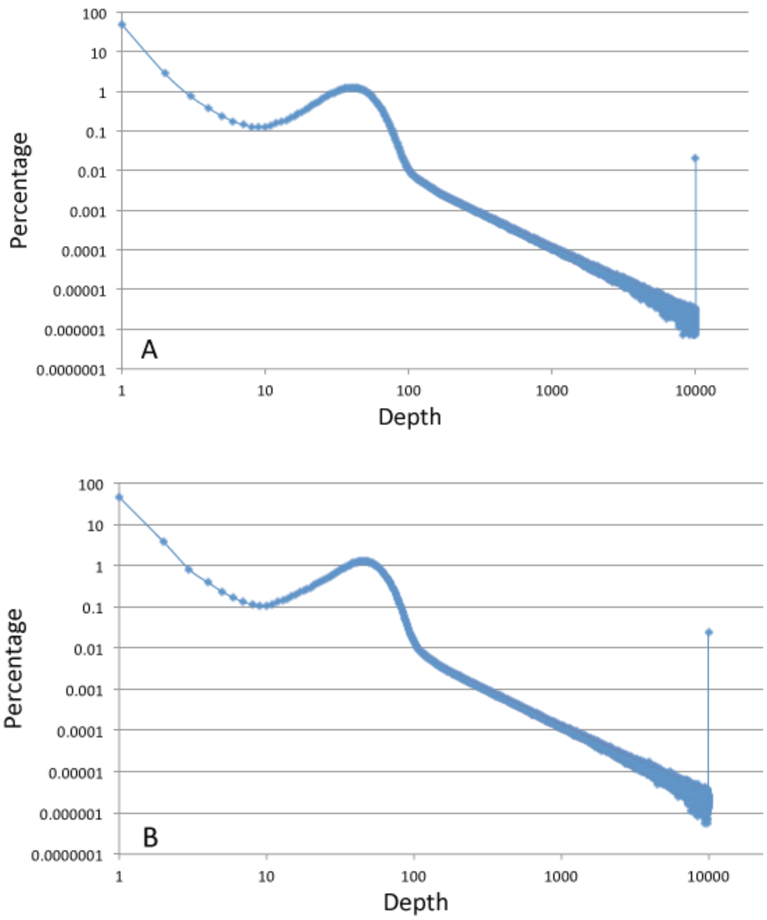


Figure 1.4 K-mer (K=31) analysis for estimating the genome size of male (A) and female (B) *T. mercedesae*. The depth of K-mers is plotted against the frequency at which they occur (both x-axis and y-axis were logarithmic scale). The upper left peaks represent the unique and rare k-mers produced by sequencing errors and are therefore not relevant to assess the genome size and architecture. The principal peaks at middle, representing in single copy of k-mers in the underlying genome, is proportional to sequencing coverage. The long tails following the principal peak primarily reflect the various classes of repetitive sequences that are present at different copy numbers in the mite genome (Moeller et al., 2014). The total K-mer numbers are 26,403,666,231 and 28,906,528,626, and the maximal frequencies are 40 and 46 for male and female *T. mercedesae*, respectively. The genome sizes are estimated by (equation 2.2), and 660 Mb for male and 628 Mb for female. A high proportion (49.8% for both male and female) of 31-mer sequences with a depth higher than 200 x suggests that the genomes contain many repetitive sequences with high sequence similarity.

The genome size of *T. mercedesae* is 15% larger than that of another *A. Mellifera* ectoparasitic mite, *V. destructor* (565 Mb), but much larger than that of *M. occidentalis*, *M. occidentalis* (88-90Mb; Jeyaprakash and Hoy 2009), and markedly smaller than those of ticks, such as *Ixodes scapularis* (black-legged tick) (2.1 Gb) and *Rhipicephalus* *microplus* (7.1 Gb; Ullmann et al., 2005). The genome size differences are most likely due to the different content of non-coding sequences such as repetitive sequences. *T. mercedesae* genome size is larger than those of sequenced insect genomes, for example, human body lice *Pediculus humanus humanus* and *P. Humanus capitis* (105-108 Mb) (Johnston et al. 2007), several *Drosophila* species (125-184 Mb) (Gao and Scott 2006), the red flour beetle *Tribolium castaneum* (158 Mb), *A. Mellifera* (262 Mb; The Honey bee Genome Sequencing Consortium 2006), housefly *Malus domestica* (295 Mb) (Gao and Scott 2006), and the silkworm *Bombyx mori* (530 Mb) (Xia et al. 2004).

Table 1.1 Genome and transcriptome Illumina sequencing data.

	Library	**DNA/RNA source**	**Depth**	**Sequencing type**	**Insert size (bp)**	**Read length (bp)**	**Clean Reads**
DNA-seq	Male	Single mite	60X	Paired end	1,239	100	274,538,548
				Single end	-	100	120,682,056
	Female	Single mite	62X	Paired end	1,006	100	311,718,384
				Single end	-	100	139,104,931
RNA-seq	Male	20 to 30 mites	-	Paired end	Overlap	90	48,725,296
	Female	20 to 30 mites	-	Paired end	Overlap	90	48,503,476
	Nymph	20 to 30 mites	-	Paired end	Overlap	90	54,915,998
	Randomly collected mites	20 to 30 mites, female must be dominated but may contain male.	-	Paired end	Overlap	90	48,100,134

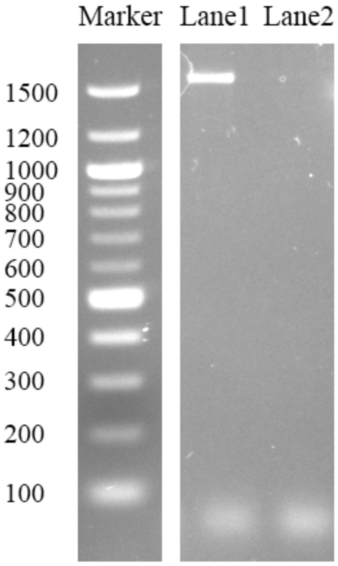


Figure 1.5 PCR amplified bacterial 16S rRNA gene from a female mite total DNA extraction (Lane1) and a male mite total DNA extraction (Lane2) total genomic DNA samples. Numbers on the left side are the molecular weights (Kb) of marker.

2. Genome and transcriptome assembly

2.1 Introduction

Both genome and transcriptome of *T. mercedesae* were sequenced using the Illumina short paired-end sequencing approach, which sheared the DNA sequences (genomic DNA or cDNA) into random short fragments and sequenced each fragment independently. The process of reconstructing the original sequence by computationally stitching these short sequences (reads) into contiguous DNA sequences is known as sequence assembly (Pop and Salzberg 2008). When the raw sequencing data are generated from genomic DNA, the process is called genome assembly, whereas assembling the cDNA reads is called transcriptome assembly.

Non-model organisms usually lack the pre-existing reference genomes, and thus the genome and transcriptome assemblies must be conducted *de novo*. A number of *de novo* assembly algorithms and applications have been developed; however, the reads produced by the NGS platform are much short than those generated by the Sanger sequencing method. This feature of NGS poses computational challenges to the *de novo* genome assembly strategies. The central challenge is to deal with the presence of repetitive sequences with low complexity in the higher eukaryotic genomes, which also contain the extensive duplications. If a read comes from a repeat and is shorter than the repeat, the original position of this read on the genome can be indistinguishable (Treangen and Salzberg 2012). Paired end NGS sequencing has been commonly used to efficiently resolve this problem. The secondary important challenge is to handle a massive amount of data generated by NGS. Because the short read length requiring a higher coverage sequencing to meet the overlap detection criteria of the genome *de novo* assemblers, the computational complexity is largely increased. Consequently, many assemblers use k-mer (defined in 1.2.5) based algorithms to manage the large NGS data within a computer's processing power (detailed blow).

The most commonly used approach for assembly of short read data is currently based on the formation of De Bruijn graphs of k-mers (Ekblom and Wolf 2014). The k-mers derived from original reads are generally catalogued into a hash table, and only a single copy of each k-mer is stored. This approach allows constant-time lookups during graph constriction under a lower memory requirement. These hashed k-mers instead of reads form the nodes of the graph and are linked when sharing a k-1 mer. After the complete de Bruijn graph has been created from the k-mers, there are normally several possible paths existing through the graph. In order to separate the true path from the multiple ones to built contigs, sequence information needs to be added back into the network. Theoretically, the k-mer based algorithms increases the computational requirements with genome size but not with the number of reads, because, without sequencing error, the reads using same k will construct same de Bruijn graph for a genome, regardless of the coverage (Schlebusch and Illing 2012). However, the k-mer based algorithms are less sensitive than the algorithms searching all-against-all overlap, resulting in the loss of some true overlap so that create links between two unrelated sequences (Miller et al., 2010). Therefore, it is important to choose a k such that most incorrect overlaps do not share a k-mer by chance and small enough to share 'k-1' nucleotides in the true overlap.

Several *de novo* genome assemblers such as EULER (Pevzner et al., 2001), ALLPATHS (Butler et al., 2008), Velvet (Zerbino and Birney 2008), abyss (Simpson et al., 2009), and soapdenovo (Luo et al., 2012) have been developed to assemble the large number of NGS shorts reads. Different assembly strategies or parameters often perform differently when dealing with various compositions of GC content, size, repetitive sequences and levels of polymorphism in the genomes of different species. A wide range of basic metrics has been used to assess the performance, but the precise behavior of assemblers on a given genome is hard to predict without a reference. A length-based statistics for scaffold/contig, N50, is the most widely used metrics for examining the quality of a *de novo* genome assembly (Yandell and Ence 2012). The scaffold/contig N50 is calculated by first ordering every scaffold/contig by length from the longest to shortest, and then the lengths of each contigs are summed from the longest contig until the sum equals to one-half of the total length of all scaffolds/contigs in the assembly. Despite it is generally agreed the longer N50 means the better assembly, poor assembly can still incorrectly join unrelated sequences to provide ostensibly large N50 value (Yandell and Ence 2012).

Sequence contaminations are also serious concerns to the quality of data used for downstream analysis, causing misassembly of sequence contigs and erroneous conclusions (Kumar et al., 2013). It is technically difficult to prepare the pure genomic DNA from a target species, which is often small, and cannot be isolated free from the living environment. The read contaminations may be introduced from the ingested food, the host of parasite/pathogen, or the commensal and symbiotic organisms attached to or within the individuals sampled (Kumar et al., 2013). Therefore, it is necessary to remove the data from non-target genome to optimize the quality of genome assembly, and re-assembly of the non-target genome may provide good opportunities to understand the host-parasite systems.

Similar to short-read genome assembly, transcriptome assembly involves either reference-based or *de novo* assembly strategy. The reference-based assembly generally performs well when a high-quality reference genome already exists. Reference-based assembly combined with gene prediction has become a powerful tool for comprehensive gene annotation in a newly sequenced genome (see section 3). By contrast, the *de novo* transcriptome assembly strategy does not depend on a reference genome; it works similarly as the *de novo* genome assembly to find overlaps between the k-mers and assembles them into transcripts via a De Bruijn graph-based approach. The difference here is that short repeats are not problematic for transcriptome assembly, as the repeats are often present in intergenic regions, which usually do not exist in the transcriptome. The genome assemblers cannot be directly applied to transcriptome assembly because, for example, some genome assemblers treat the abundant transcripts (with high sequencing coverage) as repetitive when DNA sequencing depth is expected to be constant across the genome (Martin and Wang 2011). As genomic DNA sequencing, contamination in transcriptome reads can not be avoided.

2.2 Materials and Methods

2.2.1 Estimation of insert size and *de novo* gnoeme assemble

I decided the average insert sizes of male and female mites in the respective short-read datasets using the average library insert sizes (Table 1.1) to minus length primer-binding sites (67bp for each) at both ends. Calculated sizes of the inserts for male and female *T. mercedesae* are 1105 bp and 939 bp, respectively. Then the gnoeme *de novo* was assembled by Velvet and soapdenovo.

2.2.2 Exploring contaminants in raw DNA-seq data

Blobology is a GC%-coverage based pipeline that visualizes contaminants in the preliminary assembly and then uses the visualisations to guide read separation and re-assembly (Kumar1 et al., 2013). This package provides two important tools for exploring the taxon-annotated GC-coverage plots.

2.2.3 Evaluation of genome assembly quality

As described in the introduction, the length-based statistic for scaffold/contig, N50, is the most widely used metrics to evaluate the quality of *de novo* assembled genome. In addition to N50, several extra metrics were also considered and measured using the QUAST assessment tool (Gurevich et al., 2013) and CEGMA tool (Parra et al., 2007; Parra et al., 2009) with the default settings.

QUAST: QUAST (version: 2.3) is a new assembly quality assessment tool. Here, I considered the following metrics measured by QUAST for the *T. mercedesae* genome assembly:

(1) Largest contig size: The length of the largest contig in an assembly.

(2) N50 size: An assembly (contig/scaffold) with high N50 value is obviously considered to be a high quality assembler.

(3) L50: The number of contigs/scaffolds with length larger than or equal to N50.

(4) Number of n's: In the genome assembly, mis-assemblies and gaps are commonly expressed as (N) resulting usually from repeats as well as secondary structures (Tsai et al., 2010). This value is high for low quality assemblies.

CEGMA: The CEGMA (version 2.5) is a classic approach for building a reliable set of gene annotations in the absence of experimental data. A subset of 458 highly conserved single-copy eukaryotic core genes (cegs) defined form a wide range of species were further refined into 248 genes that are generally present at low copy number in higher eukaryotes. Based on the average degree of conservation observed from each CEG, the work in (Parra et al., 2009) divides the 248 cegs into four groups (group 1 has the least conserved cegs while group 4 has the most conserved cegs). The cegs mapped in the new genome assembly can provide a rough approximation for the proportion of all known genes that may be present (Parra et al., 2009). Hence, the proportion of cegs in 248 and in each group present as complete (full-length) or partial can be taken as measure for the completeness of the gene content of a new assembly. The metrics of CEG percentages have been thought as a useful complement for the metric of N50 size to evaluate the genome assembly quality (Abbas et al., 2014).

2.2.4 Read mapping

Reads alignment against the assembled *T. mercedesae* genome was preformed using Bowtie2 (version: 2.2.1; Langmead and Salzberg, 2012) and Tophat (version: 2.011; Trapnell et al., 2009) aligners. Bowtie2 first indexed the reference genome and then aligned the DNA-seq reads. On the other hand, Tophat uses Bowtie2 to map reads from RNA-seq against the genome and identify splice junctions between exons. The alignment information of Bowtie was recorded in a standard SAM format document. Samtools (version: 0.1.19; Li et al., 2009) can read and convert the SAM format alignment into sorted BAM format for extracting either mapped or unmapped pair end reads using bam2fastq (version: 1.1.0; http://gsl.hudsonalpha.org/information/software/bam2fastq).

2.2.5 Masking and annotation of repeated sequences

I used the software repeatmasker (A.F.A. Smit and P.Green, unpublished). To identify known repeat sequences in the genomic scaffolds, while novel repeat sequences were predicted by repeatmodeler (A.F.A. Smit and P.Green, unpublished), which includes two *de novo* programs, RECON (Bao and Eddy 2002) and repeatscout (Price et al. 2005) (details see section 3.2.1).

2.2.6 Transcriptome assembly

***De novo* transcriptome assembly:** Trinity (version: 20131110; Grabherr et al., 2011) is a modular *de novo* transcriptome assembler that takes advantages from both greedy and De Bruijn graph based algorithms, containing three stages: Inchworm, Chrysalis, and Butterfly. The Inchworm step assembles transcript contigs using a greedy extension based on (*k*-1)-mer overlaps. The Chrysalis step clusters the transcript contigs into components corresponding to different splicing isoforms or closely related gene families, and constructs a De Bruijn graph for each cluster. The Butterfly step analyzes various paths on the De Bruijn graphs to differentiate and report the transcripts for splice isoforms and paralogous genes. Most notably, Trinity only provides a fixed size of k-mer (25bps).

Reference based transcriptome assembly: Tophat firstly aligned raw reads against a repeat masked genome to determine exon-exon splice junctions, and then Cufflinks (version: v0.8.2; Trapnell et al, 2010) was used to reconstruct reference based transcripts from the spliced alignments.

2.2.7 Transcript abundance estimation

RSEM (Li and Dewey 2011) software package automates the alignment of reads to reference transcripts using the Bowtie (version: 0.12.7; Langmead et al., 2009) alignment program for estimating gene and isoform expression levels (abundance) from RNA-Seq data. Because long transcript attracts more reads during the mapping step, RSEM is using a normalization method known as FPKM (Trapnell et al. 2010) for paired end reads. The FPKM is the number of RNA-seq fragments per kilobase of transcript effective length per million fragments mapped to all transcripts.

2.2.8 Real time RT-PCR for absolute quantification

The RNA extracts of male, female and nymph *T. mercedesae* (20~30 mites for each) were prepared and reverse transcribed into cDNA as described in Chapter 10. The cDNA products were then diluted 100-fold with DNA free water before used in Real-time PCR.

The quantitative PCR reactions were performed, each in triplicate, using an Applied Biosystems 7500 Fast Real-Time PCR System and 2X KAPA SYBR FAST qPCR Master Mix (KAPA Biosystems Woburn, MA). Each reaction mixture (10 μl) contained 5 μl 2× SYBR Master Mix, 200 nm each of forward and reverse primers, 1 μl of template cDNA and 0.2 μl ROX dye. The PCR amplification cycling conditions were: 95°C for 3 min, followed by 40 cycles of 30 sec at 95°C, Annealing Temperature 60°C for 30 sec, extension at 72°C for 30 sec, and a data collection window at 76-77°C for 30 sec.

Standard curves were constructed using RT-PCR product of target gene. The PCR product was cut off from gel and purified with the axyprep DNA Gel Extraction Kit (Axygen, China). The concentration was measured using Nanodrop 2000 spectrophotometer (Thermo Scientific, USA) for calculating original copy number by a formula:

Copies = DNA concentration (ng per μl) × 6.02 × 10^{23} (copies per mol) / length (bp) × 6.6 × 10^{11} (ng per mol), Where is 6.6×10^{11}ng mol^{-1} is the average molecular mass of a dsDNA bp, and 6.022×1023 copies mol−1 is the Avogadro's number (mckew and Smith 2012). Linear Standard curves were generated by duplicate amplifications of serial dilutions of target gene with 10^5, 10^6, 10^7, 10^8 and 10^9 copies of the gene per reaction, plotting the Ct values against copy number. Target gene copies in a sample could be estimated by comparing to its standard curve.

2.3 Results and discussion

2.3.1 Preliminary genome assembly

Different combinations of the paired end and single end reads (the honey bee genomic DNA contaminants in the female reads were removed prior to assembly; see below) for DNA-seq were assembled using Velvet assembler (Zerbino and Birney 2008) into five preliminary *de novo* assemblies with their best hash length (k-mer) as showed in Table 2.1.

Table 2.1. Statistics for *T. mercedesae* genome assemblies derived from different reads combinations. Assembly with all DNA-seq reads was out of computer cluster's memory.

Assembly	**Reads used**	**Best hash length (bp)**	**N50 (bp)**	**Contig Number**	**Size (Mb)**
Male 1	Paired end male	61	7,779	104,345	307
Male 2	Paired end male, single end male	67	8,680	104,854	309
Female 1	Paired end female	63	5,668	148,833	309
Female 2	Paired end female, single end female	67	5,994	154,540	311
Combined male-female 1	Paired end male, paired end female	71	4,563	197,351	313
Combined male-female 2	All reads for DNA seq	-	-		-

Table 2.1 showed both of the male genome assemblies have larger N50 contig lengths than each female assembly, suggesting the genomic DNA extracted from female was more heterogeneous. Because the ploidity of female is double of male, the female inevitably have more heterogeneities in each chromosome, and these must complicate the assembly (Kelley and Salzberg 2010). This can also be explained by *T. mercedesae* life cycle: *T. mercedesae* mother mite lays one male and several female eggs in a capped brood cell; mites develop and mate in the cell until the mother mite emerge from the brood with the adult bee and search for a new host (Terrestrial Animal Health Code 2010); although no previous report exists, the male mite should never get out of the brood cell as *Varroa* mites (Anderson and Morgan 2007). As a result, the male and female would mate as soon as they are fully developed. Although all the female mites I collected did not show pregnancy phase, they could be early gravid, probably containing both fertilized (developing to male) and unfertilized (developing to female) eggs. This increased genetic diversity of the sequenced female. In addition, I also found the genome assembly generated by combined male-female reads has a much smaller N50 (Table 2.1). It suggests there are significant genetic heterogeneities between male and female.

2.3.2 Exploring preliminary genome assemblies

1) Exploring DNA contamination in the preliminary genome assemblies

Many species are difficult to completely separate from their living environments, either because they cannot be independently cultured, or because they are very intimately involved with a host or their commensal and symbiotic organisms, such as closely associated bacteria. *T. mercedesae* is not a traditional, inbred, laboratory-reared model organism. They directly feed on honey bee brood and were reared in the hive. As a result, it is impossible that DNA and RNA extracted from whole-organism are free from its host and associated microbes, particularly gut microbes. Because most genome assemblers perform poorly with the mixed-organism DNA (Kumar1 et al., 2013), removal of the raw reads derived from non-target genomes will be crucial for improving genome assembly quality.

The host of *T. mercedesae,* honey bee*,* was genome sequenced and the data could be downloaded from a database (Beebase). I aligned all reads derived from both male and female libraries to the honey bee genome sequences using the Bowtie 2 short read aligner. All unaligned reads were then extracted by bam2fastq, and assembled by Velvet. A total of 37 and 70 contigs were assembled from the aligned male and female reads, respectively, but most of them were low-complexity (simple) repeats, which is because of the low-complexity reads of *T. mercedesae* mapped to the low-complexity regions of honey bee genome. After blasting (blastn with a low-complexity filter parameter) these aligned contig sequences to the honey bee reference scaffolds, I found one contig assembled using female reads was identified as honey bee DNA. The reads mapped to this contig (the female preliminary assembly) were then removed from the female raw reads. This DNA contaminant might be present in hemolymph of honey bee larvae, and get into female mite's body as food.

To identify DNA contaminated from the host was relatively simple because honey bee genome sequence is available. However, DNA contaminated from the associated microbes, especially gut microbes, can be complex and varied. Because of the relative molarity differences and distinct G-C contents between the mite and bacterial genomes, G-C content (% GC) and coverage based method can be used to separate the target genome from its bacterial contaminants. The same strategy was used to assemble *V. destructor* genomic contigs (optimized) (Cornman et al., 2011). Here, I used the improved GC%-coverage based pipeline, Blobology, to identify the contaminated DNA sequences in both male and female genome assemblies. As described in the pipeline, the male and female genomic contigs were first assembled (assemblies: male 2 and female 2; Table 2.1) using their best k-mer without any optimized parameters, and then passed to the core processing of Blobology, which was achieved by (1) calculating the average GC content of each contig, (2) extracting the node (k-mer) coverage directly derived from the assembler, and (3) using the megablast (E-value cutoff < 1e−05) to identify species with the highest similarity to each contig in NCBI nr database. By above three steps, GC%-coverage plots were generated for both male (Figure 2.1A) and female (Figure 2.1B) preliminary assemblies.

Without optimized parameter (expected coverage), Velvet assumes that all reads varying slightly from each other are derived from the different sources. Therefore, Velvet attempts to assemble each of them separately. This prevents the contaminating DNA assembled into the mite gnome. In addition, the best k-mer can ensure the accurate assembly (Zerbino 2010). The node (k-mer) coverage is directly proportional to the read coverage ($C_k = C * (L-k+1)/L$, where k is the k-mer (hash) length, C is the common read coverage, and L is the read length (Zerbino 2010) so that the node (k-mer) coverage could be used to plot directly. Originally, Blobology calculated the read coverage with BAM file which is produced by transformation of Bowtie 2 output using samtools, but the pipeline of Blobology does not provide a normalization method (without normalization, the longer contigs may show higher coverage because they can be aligned more reads than the shorter contigs). Although I also tried to use a RSEM pipeline, which was initially designed to normalize expression levels for RNA-seq, to cancel the length-associated biases, the I found GC%-coverage plot using the node coverage directly has a better performance (i.e. More close to the poisson distribution as described below).

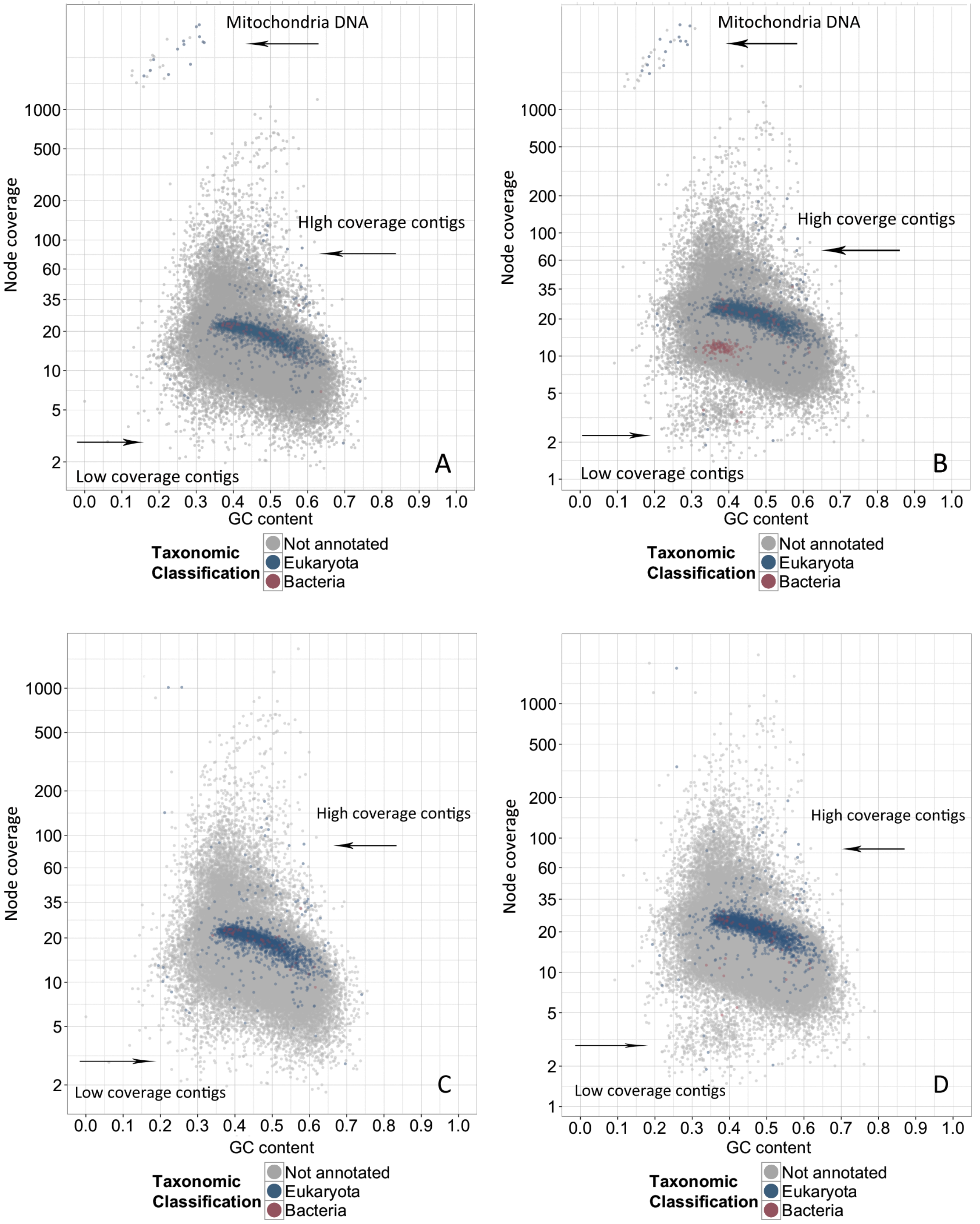


Figure 2.1 GC%-coverage plots for male 2 and female 2 preliminary assemblies. Individual contigs are plotted based on their GC content (x-axis) and their node coverage (y-axis; logarithmic scale). Contigs are colored according to the taxonomic order of their best megablast hit to the NCBI nr database (with E-value cutoff < 1e−5). Contigs without the annotation are in gray. (A) GC% plotted against node coverage for male 2 contigs. (B) GC% plotted against node coverage for female 2 contigs. (C) GC% plotted against node coverage for re-assembled male 2 contigs after removing the mitochondria DNA. (D) GC% plotted against node coverage for re-assembled female 2 contigs after removing the mitochondria and bacteria DNA.

The node (k-mer) coverage distribution of assembled genome contigs should normally obey poisson distribution, but the observed distribution differs in three ways (Zerbino 2010). First, because of cloning bias (i.e. sequencing bias typically associated with GC content, and the skewed blobs in Figure 2.1 were due to this bias), the variance of the observed distribution is larger than expected. Second, occasional random errors create a large number of words (K-mers) which are rarely observed. This creates very short contigs with low complexity and coverage. Third, eukaryotic genomes usually contain the repeated sequences. This produces a significant number of contigs with high coverage (The number of the high coverage contigs varies with the multiplicity of repeats in the different genomes). Figure 2.1 shows the distribution of GC%-node coverage plots is consistent with the probability (distribution) described by Zerbino (2010), suggesting the GC%-node coverage plots can preform well to classify the contigs and identify the contaminated DNA in the mite genome assemblies.

There is a distinct blob apart from other contigs with very high (>100-fold) coverage at the upper left corner of both male and female GC%-coverage plots (Figure 2.1A and B). I found that these contigs were enriched with the mitochondrial DNA by blastning (E-value<1e-5) their sequences against NCBI nr database. In addition, both plots show a major blue blob of contigs with ~20-fold node coverage with predominant taxonomic identification as Eukaryota. There are also some contigs annotated as Bacteria which spread over the major blue blob annotated as Eukaryota (Figure 2.1A and B). Moreover, a red blob of low coverage in the female plot was also identified as contaminating DNA from bacteria (Figure 2.1B).

Are the contigs in the blue blob identified as bacterial origin true? As described above, I examined the extracted DNA by PCR prior to sequencing, and found that no bacteria was detected in the sequenced male mite. I therefore extracted the contig sequences assigned as bacterial origin by megablast within the Blobology pipeline, and then applied blastn (E-value<1e-5) to search them in a NCBI nr database. The result showed that most of the contigs in the blue blobs of both female and male were shifted to match strongly to arthropod sequences (best hit), and merely few contigs were still mapped to bacteria (Figure 2.3A and B). However, the contigs in the low coverage red blob of female were still annotated as bacterial sequences with relatively long alignment (Figure 2.3A, the best hit). The miss annotation might be due to the fact that megablast is specifically designed to efficiently align very similar sequences. Blastn is more sensitive to detect the sequence similarity than megablast since it uses a shorter default word size. For example, if a contig has a short sequence that is very similar to a bacterial sequence and the remaining sequence does not have significant similarity to other arthropod sequences, the megablast can only recognize the short sequence with high similarity and report this contig as bacterial sequence. However, blastn uses short word size, which can extend to the remaining sequence well.

By blastn, seven and six contigs in the male and female blue blobs were identified to contain *Wolbachia* DNA sequences, and they have relative long alignment length than the contigs identified as other bacteria (Figure 2.1A and B; Figure 2.3A and B). However, I have checked the presence of *16S rRNA* gene of *Wolbachia* using PCR, no amplification was detected (Figure 2.2 A). *Wolbachia* genes in the mite genomic contigs might be due to the remnants of horizontal gene transfer or artifact. It has been reported that horizontal gene transfer from *Wolbachia* endosymbionts was detected in ~33% of the sequenced arthropod genomes using a bioinformatic approach (Robinson et al., 2013). Gene annotations of the contigs using blastx (E-value<0.05) revealed that most of these contigs contained both mite and *Wolbachia* genes. The fragments of these *Wolbachia* genes show high similarity to Type IV secretion system protein virb6, transposase, phospho-N-acetylmuramoyl-pentapeptide-transferase, ATP-dependent Zn protease, cell division protein ftsh and NADH dehydrogenase subunit M. To confirm that these genes are integrated into *T. mercedesae* genomic DNA, and are not assembly artifacts by contaminating DNA, I designed two pairs of primers (one primer located in the mite gene, and the other in *Wolbachia* gene; Table 2.2) to test genomic DNA organization for two selected contigs in the male2 assembly (contig016872 and contig016313). I found both amplified genomic DNA sizes are same as the assembled genomic DNA sizes (Figure 2.2 B).

Wolbachia is often localized in the reproductive tissues of arthropods and it is responsible for the induction of a number of reproductive alterations such as feminization (the infected males develop as females or infertile pseudo females), parthenogenesis, male-killing (the infected males die during larval development) and cytoplasmic incompatibility (the inability of *Wolbachia*-infected males to successfully reproduce with uninfected females or females infected with another *Wolbachia* strain). Although the ratio of female to male is 5.5 to 1 with *Tropilaelaps* mites (Anderson and Morgan 2007), the blastx results indicated that these *Wolbachia* genes are only partial length. Thus, these genes are unlikely to function after the horizontal gene transfer event. The skewed sex ratios must be due to other reasons.

Table 2.2 Primers for testing genomic DNA organization.

Selected contigs	**Expected DNA size of amplification up to assembled contigs**	**Forward primer (5' -> 3')**	**Reverse primer (5' -> 3')**
Contig016872	~3000bp	GCGAACACACATTATCCCCTTCCGCGCA	TCAGATCAGACGCCATACTGAAGCTGAG
Contig016313	~52000bp	AACACGTATACTCGCACGTGAAGTACGG	ATGCAAGCAAACAATATGGGGAGTCAGC

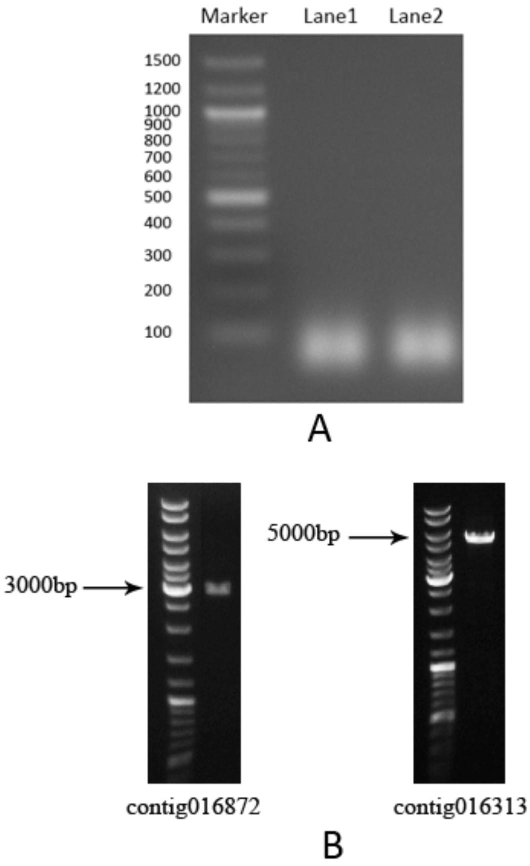


Figure 2.2 (A) PCR detection of the presence of *Wolbachia 16S rRNA* from a female mite total DNA extraction (Lane1) and a male mite total DNA extraction (Lane2) total genomic DNA samples. Numbers on the left side are the molecular weights (Kb) of marker. (B) The PCR tests were carried out with two selected contigs (contig016872 and contig016313) containing Wolbachia gene in the male2 assembly. The lengths of amplification are the same as those expected from the contig sequences.

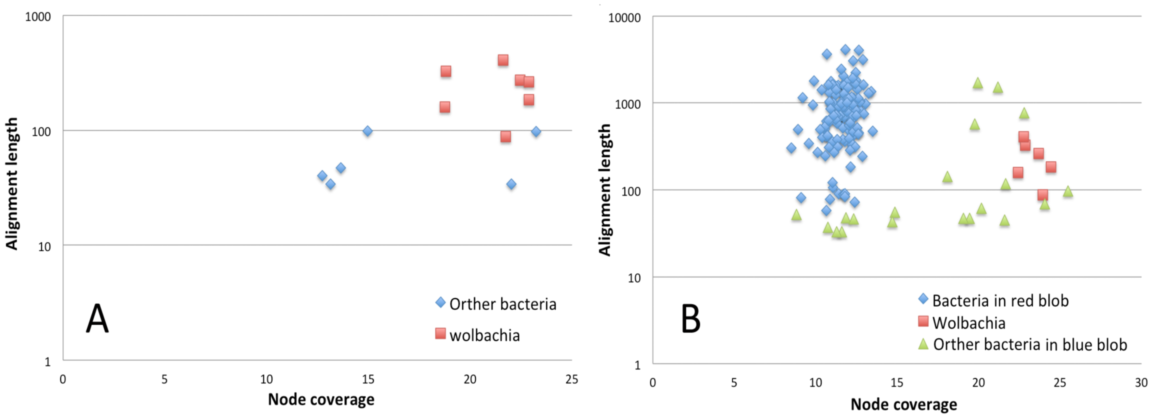


Figure 2.3 Plots for bacteria identified by blastn from the contig sequences initially annotated as bacteria by megablast in male 2 (A) and female 2 (B) preliminary assemblies. The alignment length (y-axis) derived from the best blastn hit is in logarithmic scale.

Megablast search for bacterial species with the contigs in the low coverage red blobs (Figure 2.1B) suggested that one or more gamma-proteobacteria species were particularly abundant in the sequenced female mite. To further identify the bacterial species, all reads mapped (by bowtie2) on these contigs were extracted and re-assembled into 96 contigs using spades genome assembler. Prokka (Seemann 2014), a rapid annotation tool for bacterial sequences, predicted 751 protein-coding genes from 81 contigs. The annotated proteins were used as queries for BLASTP (E-value<0.05) search against the genbank nr database to determine the originated bacteria. I found almost all of them are members of gamma-proteobacteria, in which 667 proteins show high similarity to *Rickettsiella grylli* proteins with an average identity of 79%. The genus *Rickettsiella* (Philip) are often considered as intracellular pathogenic bacteria associated with a wide range of arthropods that often proliferate in the fat body or hemolymph cells of the host (Cordaux et al., 2007). However, one study reported mutualistic interaction of a *Rickettsiella* bacterium with its host, *Acyrthosiphon pisum*, a pea aphid (Tsuchida et al., 2010). In naturally infected hosts, *Rickettsiella*-mediated disease are usually chronic and affect the larvae and adults with various symptoms in the different hosts, such as greenish blue discoloration observed by *R. Popilliae* infection in Japanese beetle (Dutky and Gooden 1952) larvae and behavioral fever (defined as an acute change in thermal preference driven by pathogen recognition) in the *R. Grylli*-infected crickets (Adamo 1998). A *Rickettsiella* related bacterium has also been observed with a bisexual laboratory colony of the hard tick (Leclerque and Kleespies 2012). Here, I also found *Rickettsiella grylli*-like bacteria in the sequenced female *T. mercedesae*, however, the roles on the host as well as the possibility of transmission to honey bee remain to be determined.

Based on the current data, the male mite genome we sequenced is not contaminated by DNA from the host (honey bee) and bacteria. Therefore, I removed the reads annotated as the mitochondria DNA sequences (some of the mitochondria DNA sequences were also detected in the high coverage region as described below), and then re-assembled the genome to check the completeness using GC%-node coverage plot (Figure 2.1 C). For the female mite genome, I removed the reads mapped to the contigs identified as bacterial DNA using blastn (E-value<1e-10; alignment length cutoff>100) and megablast as well as the reads corresponding to the mitochondria DNA sequences before re-assembly (Figure 2.1 D).

2) Exploring high coverage contigs in the assembled genome

Low coverage contigs with short length and low complexity, which are most likely to be errors (see above), can be effectively removed by setting the node coverage cut-off (an optimized parameter of velvet, Zerbino 2010). Thus, studying the high coverage contigs would be more interesting. These can represent either sequence repeats or mitochondrial DNA. Repeated DNA sequences in the genome include the tandem repeats (simple short sequences repeated in tandem) and the interspersed repeats (repetitive sequences distributed in the whole genome) (Do et al., 2008; Zhuang et al., 2012). For example, ribosomal RNA gene repeats in tandem arrays are essential housekeeping genes found in all organisms (Ide et al., 2010). New segmental or gene duplication in the genome could become important source for evolution since it would undergo gene divergence for novel functions. If the DNA-seq reads (only 100 bp in my project) are not long enough to span the repetitive region, the repetitive sequences will not be resolved (Zhuang et al., 2012). As a result, *de novo* assembly of sequenced reads will produce many compressed high coverage contigs by recruiting reads from the repeated sequences. Lack of segmental duplications is well known deficit of short read next generation sequencing and assembly strategies (Kelley and Salzberg 2010).

With the GC%-node coverage plots, I identified high coverage contigs (here, I consider the contigs apart from the major "blob" with node coverage larger than 35 as the high coverage contigs). In order to identify the types of repeat sequences in high coverage contigs (mentioned above), I first ran the program repeatmasker and repeatmodeler to screen DNA sequences for masking interspersed repeats and low complexity DNA sequences. A scatter plot (Figure 2.4) was made to demonstrate the relationship between masked percentage and node coverage for each contig in preliminary assembled male2 genome. About half of the contigs with very high coverage were almost completely masked (3975 contigs with normalized converage larger than 20; 1696 out of 3975 contigs were 80%-100% masked, Figure 2.4). More than 90% of the high coverage contig sequences were annotated as repeat including major simple repeats and unknown repeats (Table 2.3; annotation of the repeats was discussed in the next chapter). On the other hand, the high coverage contigs that could not be masked or partially masked were annotated as mitochondrial DNA, ribosomal RNA gene, or duplicated segments.

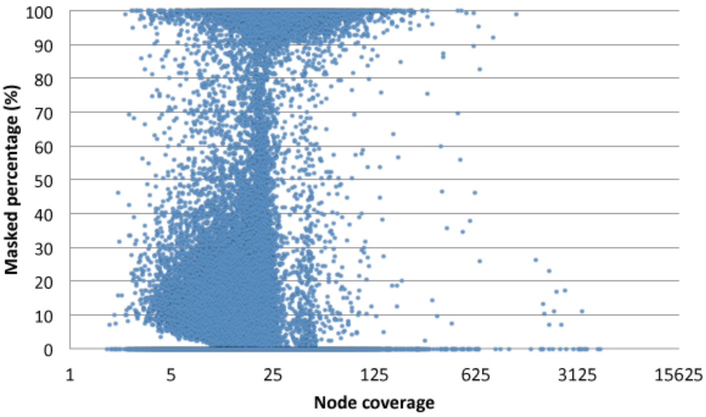


Figure 2.4 Individual contigs are plotted according to their normalized read coverage (x-axis; logarithmic scale) and their masked percentage (y-axis).

Table 2.3 Identification of repetitive sequences in the 3975 contigs with normalized converge larger than 20. The results integrate both homology-based and *de novo* predictions. The total lengths of the repetitive sequences and their corresponding percentages in the genome for different categories were calculated based on the high coverage contig size, i.e. 1.8Mb.

Repeat Type	**Length (bp)**	**% of high coverage contig**
DNA	7,206	0.43
LINE	201,485	12.09
SINE	14,165	0.85
LTR	2,524	0.15
Satellite	0	0
Simple repeat	395,443	23.74
Unknown	870,215	52.21
Small RNA	4,870	0.29
Low complexity	10,284	0.62
Total	15,06,192	90.37

In order to further study the un-masked and partially masked contigs, I used blastn (E-value < 1e−5) to annotate the male2 high coverage contigs obtained by repeatmasker and repeatmodeler (the masked sequences would not be searched by blastn) using NCBI nr database. A total of 171 contigs were annotated to contain known sequences, in which nine were mitochondrial DNA sequences and two were ribosomal RNA sequences. Most of the remaining contigs were annotated to have various functional genes. Here, I asked if these genes are present in high copy number and expressed in the mite. I therefore used Tophat to align the RNA-seq reads (derived from randomly collected mites) to the high coverage contigs obtained by repeatmasker and repeatmodeler. Then, Cufflinks could use this map generated by Tophat against these contigs to assemble the reads into transcripts. These transcripts were analyzed by blastx (E-value < 1e-5) against NCBI nr database. A total of 141 transcripts could be annotated and 115 transcripts (excluding isoforms) predicted from 108 contigs appear to have specific functions. In Appendix 1, the transcripts (excluding isoforms) with the same functional annotation were clustered together.

As described above, above genes in high coverage contigs may have been generated by the very recent repeated segmental or gene duplication. Gene duplication is one of the dominant driving forces in adaptive evolution of genome and genetic systems (Hittinger and Carroll 2007). The duplicated gene may lose the function by nonfunctionalization, evolve a new function by neofunctionalization, or be stably maintained as daughter copy partitioning of the ancestral gene function by subfunctionalization (Zhong et al., 2013). The following genes are particularly interesting in terms of the mite biology:

(1) Canalicular multispecific organic anion transporter 1 (cmoat), a member of the ATP-binding cassette transporter family involved in multi-drug resistance (Paulusma et al., 1999).

(2) Multidrug resistance-associated protein 1 (MRP1), another member of ATP-binding cassette transporter superfamily. There are evidences that the particular polymorphisms are significantly associated with a wide range of drug resistance. (Yin and Zhang 2011).

(3) Gamma-glutamyl hydrolase (GGH) was shown to significantly co-relate with a wide range of drug resistance (Schneider and Ryan 2006).

(4) Heat shock protein 70 family members were shown to be strongly upregulated by heat stress and toxic chemicals (Simpson and Alexander 2005).

Various chmical compounds have been applied to control honey bee pathogens and parasites. Notably, sulfur derivatives are commonly used as the miticide in China. Westerners introduced large numbers of honey bee colonies into Asia for commercial purposes just since 70 years ago (Crane 1988). This means the *T. mercedesae*s were treated by the mitecides no more than 70 years. The duplication of genes discussed above may be generated by adaptive evolution to toward the mitecides. The recently duplicated genes often retain the same function as the original copy, whereas the ancient gene tends to diverge and acquire novel functions (Denslow et al., 2006). Above duplicated genes may thus enhance the detoxification functions in *T. mercedesae*. Coincidentally, as a nice example, the natural selection operating on a new gene duplicate at the resistance to dieldrin (Rdl) has been reported in *D. melanogaster* (Remnant et al., 2013).

Heat and oxidative stress often damage proteins by inducing the partial unfolding and aggregation. Heat shock protein 70 (HSP70) genes have crucial roles in protecting cells against these stresses by preventing these partially denatured proteins from aggregating and fostering them to refold (Tutar and Tutar 2010). Hsp70 was also found to involve with either inhibiting viral infection or promoting viral replication (Li et al., 2011). The recent study with mosquitoes has suggested the protective response of HSP70 in warm blood-feeding arthropods (Benoit et al., 2011). In addition, association of HSP70 with heavy metal and drug stresses has been reported (Tutar and Tutar 2010). Moreover, the expression of hsps in *I. Scapularis* suggested that HSP70 is involved in the tick response to blood-feeding stress and immune response to pathogen infection (Busby et al., 2012). Thus, the HSP70 may also have the roles to protect the mite from DWV infection with very high titer (see below).

3) Post genome assembly and quality estimation

After analyzing the raw assemblies of mite genome (no cut-off or optimized parameters applied) and removing the bacterial and mitochondrial DNA contaminants using the modified Bloblogy pipeline, male 2 genomic reads were re-assembled and optimized up to scaffold level by both velvetoptimiser (https://github.com/Victorian-Bioinformatics-Consortium/VelvetOptimiser.git) and soapdenovo with their respective best k-mer sizes. Table 2.4 and Table 2.5 provide statistics to examine the quality of re-assembled male genome DNA.

Table 2.4 Statistics for the optimized genome assemblies of *T. mercedesae* male. All statistics are based on the scaffolds with their size >= 200 bp. The best k-mer sizes used by the assemblers are 67.

Tools	Metric	Soapdenovo	Velvet
QUAST	**Size (Mb)**	385	353
	Total Number	39,249	34,155
	Largest scaffold (kb)	244,654	327,111
	N50 size (kb)	31,094	28,807
	L50	3,700	3,638
	N's (per 100 kb)	14,977.08	7,593.87

Table 2.5 Quality check of *T. mercedesae* genome assembled by Velvet and soapdenovo by testing the presence of 248 cegma genes in comparison with the reference genomes (reference genomes detailed in section 4.2.1).

Species	Group	#Prots	%Completeness	#Total	Average	%Ortho
T. mercedesae **Velvet assembly**	**Complete**	228	91.94	290	1.27	21.93
	Group 1	57	86.36	76	1.33	26.32
	Group 2	51	91.07	65	1.27	23.53
	Group 3	59	96.72	79	1.34	27.12
	Group 4	61	93.85	70	1.15	11.48
	Partial	244	98.39	368	1.51	38.52
	Group 1	62	93.94	99	1.6	48.39
	Group 2	56	100	88	1.57	50
	Group 3	61	100	95	1.56	37.7
	Group 4	65	100	86	1.32	20
T. mercedesae **soapdenovo assembly**	**Complete**	218	87.9	260	1.19	16.97
	Group 1	51	77.27	59	1.16	13.73
	Group 2	50	89.29	60	1.2	18
	Group 3	58	95.08	70	1.21	18.97
	Group 4	59	90.77	71	1.2	16.95

	Partial	241	97.18	334	1.39	30.71
	Group 1	61	92.42	81	1.33	29.51
	Group 2	55	98.21	80	1.45	34.55
	Group 3	60	98.36	84	1.4	33.33
	Group 4	65	100	89	1.37	26.15
M. occidentalis	**Complete**	244	98.39	286	1.17	13.93
	Group 1	64	96.97	70	1.09	9.38
	Group 2	55	98.21	74	1.35	21.82
	Group 3	61	100	70	1.15	14.75
	Group 4	64	98.46	72	1.12	10.94
	Partial	246	99.19	312	1.27	21.14
	Group 1	65	98.48	77	1.18	18.46
	Group 2	55	98.21	83	1.51	30.91
	Group 3	61	100	74	1.21	18.03
	Group 4	65	100	78	1.2	18.46
I. Scapularis	**Complete**	113	45.56	131	1.16	13.27
	Group 1	22	33.33	24	1.09	4.55
	Group 2	29	51.79	35	1.21	20.69
	Group 3	26	42.62	28	1.08	7.69
	Group 4	36	55.38	44	1.22	16.67
	Partial	209	84.27	348	1.67	45.45
	Group 1	50	75.76	71	1.42	30
	Group 2	47	83.93	75	1.6	48.94
	Group 3	54	88.52	90	1.67	50
	Group 4	58	89.23	112	1.93	51.72
S. mimosarum	**Complete**	63	25.4	67	1.06	6.35
	Group 1	13	19.7	13	1	0

	Group 2	17	30.36	18	1.06	5.88
	Group 3	12	19.67	13	1.08	8.33
	Group 4	21	32.31	23	1.1	9.52
	Partial	174	70.16	241	1.39	29.31
	Group 1	40	60.61	45	1.12	12.5
	Group 2	37	66.07	50	1.35	29.73
	Group 3	45	73.77	61	1.36	28.89
	Group 4	52	80	85	1.63	42.31
T. urticae	**Complete**	247	99.6	291	1.18	14.98
	Group 1	65	98.48	71	1.09	9.23
	Group 2	56	100	69	1.23	17.86
	Group 3	61	100	74	1.21	16.39
	Group 4	65	100	77	1.18	16.92
	Partial	247	99.6	330	1.34	22.27
	Group 1	65	98.48	76	1.17	12.31
	Group 2	56	100	79	1.41	25
	Group 3	61	100	87	1.43	27.87
	Group 4	65	100	88	1.35	24.62

By either soapdenovo or Velvet, the size of *de novo* assembled *T. mercedesae* genome is substantially less than the estimated genome size of 660 Mbp (Table 2.4). Large eukaryotic genomes are known to comprise of the significant amount of non-coding repetitive DNA sequences (Pagel Van Zee et al., 2007), which can reduce the assembled genome size even with increased sequencing coverage (especially for the short read DNA-seq; see above). Indeed, a large number of high coverage contigs were recognized as repeat sequences (Table 2.3, Figure 2.4). Moreover, in k-mer based analysis of the *T. mercedesae* genome, 50% of 31-mers had at least two identical copies in the genome (the long tails in Figure 1.4 A and B). This can also reflect the abundance of repeated sequences in the mite genome. Coincidentally, in genomic survey project of *V. destructor* mite, the assembly of *V. destructor* genome resulted in 318 Mbp of sequence using, much smaller than the estimated genome size of 565 Mbp by flow cytometry (Cornman et al., 2010).

Fundamentally, k-mer analysis is based solely on the sequence contents of NGS reads. Thus, if the NGS reads represent contents of the whole genome without any bias during the experimental procedures, the k-mer output should give accurate estimate of the genome size (Kim et al., 2014). I can see the genome size of male mite estimated by the k-mer based method is almost identical to that estimated by the flow cytometry. Therefore, sampling and sequencing bias is not the reason of losing some sequences in the genome assembly. Moreover, I used Bowtie2 to map the cleaned paired end reads and the single end reads back to the male genome assembled by Velvet. The mapping rate was 94.1 %, in other words, up to 94.1 % of the reads has been used for the genome assembly. Therefore, the missing sequences in *T. mercedesae* genome are more likely due to the compressed repeats occurred during the assembly (high coverage contigs in Figure 2.1).

The longer the scaffold N50 is, the better the assembly is. However, there is no strict standard for good N50 scaffold length. Accurate annotation of a genome requires the contiguous genome assembly in order to avoid gene splitting across the scaffolds. If the scaffold N50 is around the median gene length, ~50% of the genes would be contained in a single scaffold where the complete genes together with fragments from the rest of the genome provide resources for the downstream analyses (Cantarel et al., 2008; Ye et al., 2011). The median gene length is roughly proportional to the genome size, and it could be calculated from Figure 2.5. Because the estimated genome size of *T. mercedesae* is ~660 Mb, the corresponding median gene length obtained from Figure 2.5 is about 4,466 bp. Both of N50 values calculated by soapdenovo and Velvet were much larger than the minimum N50 scaffold length for annotation.

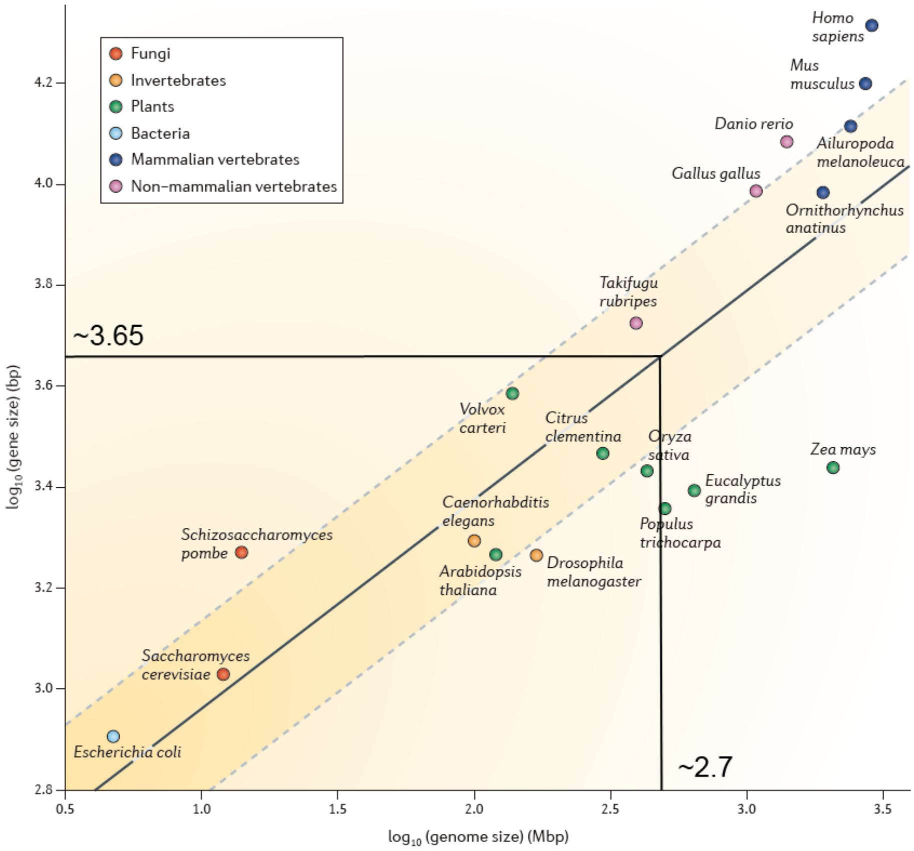


Figure 2.5 Relationship between genome size and gene length for a representative set of genomes (extracted from Yandell and Ence 2012). Average Gene length is plotted as a function of genome size for representative bacteria, fungi, plants, and animals. This figure illustrates a simple rule of thumb; in general, larger genomes have longer genes. Thus, accurate annotation of a large genome requires the contiguous genome assembly in order to avoid splitting the genes across scaffolds. The median gene length of *T. mercedesae* was calculated by this curve.

Table 2.4 shows the quality metrics of *T. mercedesae* male genome assembled by two methods. Soapdenovo generated a lager genome assembly and longer N50 value, but produced a huge number of gaps compared to Velvet. Based on the CEGs metrics (Table 2.5), Velvet assembly resulted in having more core genes than soapdenovo. Miss joining unrelated contigs might cause the longer N50 in soapdenovo assembly. However, to compare the performance of these two assemblers for the *T. mercedesae* genome should not just focus on the N50 size. The completeness and correctness of informative genes present in the assembly are more important for the down stream analysis (Bradnam et al., 2013). Although the size of male genome assembled by Velvet is much smaller than the estimated genome size with smaller N50 than soapdenovo, a total of 228 (92%) genes were predicted to have the complete gene model (CEGMA alignment length > 70%), and this proportion increased to 98.39% if the partial matches were incorporated. On the other hand, the quantity of *de novo* assembled transcripts mapped back to the genome assemblies could also assess the genome assembly quality. Blat is able to align a spliced sequence to the genome contigs. The transcripts derived from all tanscriptome samples could be efficiently aligned to the Velvet-assembled male genome with this tool. The alignment results also revealed that Velvet has assembled almost all genes that predicted by Trinity; 141440 out of 142174 contigs could be mapped to the optimized male2 contigs. Therefore, Velvet-assembled male genome has enough quality and was used for the further genomic analysis.

The GC content of *Tropilaelaps* mite genome was similar to those of the *I. Scapularis* and *D. melanogaster* genomes, but lower than that of *M. occidentalis* and higher than those of *T. urticae*, *A. Mellifera*, and *S. mimosarum*. Only a minor fraction (0.005%) of the *T. mercedesae* genome has the GC content < 20% and there is no region with the GC content > 80% (Figure 2.6). Thus, the *de novo* genomic assembly was not strongly affected by GC-biased non-random sampling.

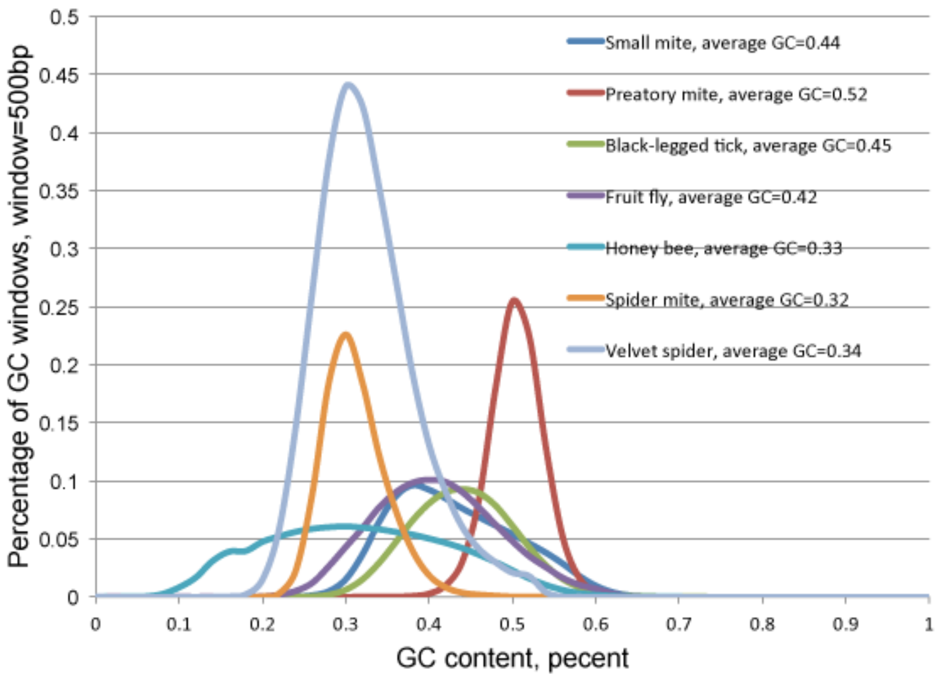


Figure 2.6 GC content distribution in the genomes of small mite (*T. mercedesae*), predatory mite (*M. occidentalis*), black-legged tick (*I. Scapularis*), fuit fly (*D. melanogaster*), hony bee (*A. Mellifera*), spider mite (*T. urticae*) and velvet spider (*S. mimosarum*). The GC content in 500 bp non-overlapping sliding windows along the genome were calculated.

2.3.3 Transcriptome assembly

1) Identification of contaminated RNA in RNA-seq data

It is impossible that RNA extracted from the mites is free from contamination by its host's RNA. I used Bowtie2 mapping method to filter the contaminated RNA sequences. The reads that mapped to honey bee genomic scaffolds were extracted, and then assembled using Trinity assembler. One transcript in male, two transcripts in nymph, 17 transcripts in female, and 43 transcripts in randomly collected mites (almost all are expected to be female) were assembled in total. These transcripts were then analyzed using blastx against NCBI nr database (e-value < 1e-5). The transcripts found in male and nymph could not be identified; however, four transcripts in the female and randomly collected mites were annotated as honey bee hexamerin mRNA which is primarily synthesized by the larval fat body and are massively stored in hemolymph as an amino acid source for development toward the adult (Burmester and Scheller 1999). This suggests that the male and nymph do not feed on honey bee larvae.

The *T. mercedesae* genome becomes a good reference to further explore contaminated RNA in the mite transcriptomes using the Tophat map technology. Compared to the mapping rate of RNA sequences from the randomly collected mites (62.9%), those from male, female, and nymph were only 12.4%, 21%, and 14.2%, respectively. I therefore extracted and *de novo* assembled the unmapped RNA reads for all RNA-seq samples. I found all samples were abundant with honey bee virus (Table 2.7: FPKM value). Especially for the male, female, and nymph samples, almost half of the reads corresponded to the Deformed wing virus (DWV) RNA. I afterwards confirmed this result by running qRT-PCR.

The qRT-PCR method developed for DWV detection allowed quantification of the copy number of virus and a mite 'housekeeping' gene (Elongation factor-1 alpha) mRNA present in the investigated samples (male, female, and nymph mites) using the respective calibration standard curvse. The primers designed were DWV-fwd (5'-GGGAATAAAACCTCACA-3') and DWV-rev (5'-ATGCCATAAAATCTAAGT-3') to target a highly conserved region of DWV RNA. The primers to detect the mite Elongation factor-1 alpha mRNA were EF-1ɑ-fwd (5'-ATTCCGGTAAGTCAACCACCAC-3') and EF-1ɑ-rev (5'-GCTCGGCCTTCAGTTTGTCCAA -3') in which the forward primer spans the splicing junction. Then, I compared the ratio of copy number to the ratio of FPKM calculated by RSEM between DWV and EF-1ɑ RNA for each sample (Table 2.6). I found the results of qRT-PCR agree with the expression levels (FPKM) calculated by RSEM. Thus, these results are consistent with that the low mapping rates of transcriptome sequences to the genome were due to very high DWV RNA expression.

Table 2.6 Comparison between the ratios of copy numbers of DWV RNA to Elongation factor-1 alpha mRNA determined by qrt-PCR and the ratios of DWV RNA FPKM to Elongation factor-1 alpha mRNA FPKM in the male, female and nymph transcriptomes. The FPKM values were obtained by RSEM (see section 2.2.7).

Sample	**Ratio of gene copy number**	**Ratio of FPKM**
Male	556	692
Female	193	131
Nymphy	212	226

Deformed wing virus is a major positive-stranded RNA virus (Lanzi et al. 2006), which could lead to honey bee colony collapse by the infection. The previous report suggested that both *T. mercedesae* and *V. destructor* are linked to DWV infection in honey bee colony (Forsgren et, al 2009). My study confirmed *T. mercedesae* is an important DWV virus reservoir. It has been found that DWV replicates in both honey bee and *Varroa* mite (Ongus et al. 2004; Yue and Genersch 2005). Similarly, DWV is likely to replicate in the cells of *Tropilaelaps* mite. Since the male and nymph mites do not feed honey bee larva, the horizontal transmission of DWV from the infected honey bee larva is unlikely. Instead, the vertical transmission of DWV from the infected mother mite would be the major infection pathway.

Interestingly, *T. mercedesae* appears to have the ability to survive with very high titer of DWV infection. The presence of some viruses could be relatively benign and even beneficial to the host (Roossinck 2011). Pathogen virulence may be altered by changes in pathogen-host assemblages (Rigaud et al., 2010). In the case of multi-host pathogens with interspecies transmission, the pathogen's virulence may increase following introduction of the second host when the constraint on pathogen virulence in a given host is omitted (Woolhouse et al., 2001). Indeed, DWV is often present in honey bee by the covert infection (De Miranda and Fries, 2008). DWV can be vertically transmitted to the worker bees from both male and queen bees, and this transmission route results in covert infection without the visible symptoms (De Miranda and Fries, 2008). Overt DWV infection was only characterized in colonies infested with *Varroa* mites to cause the adult wing and abdominal deformities (Bowen-Walker et al. 1999). Therefore, DWV (DWV-*T. mercedesae* Wuxi strain) could be less virulent or even beneficial in *T. mercedesae* before transmitted to honey bees.

Two plant viruses (Strawberry latent ringspot virus and Cycas necrotic stunt virus) identified in the randomly collected mites sample might be derived from pollen or other plant materials collected by honey bees.

Table 2.7 Expected counts and expression levels (FPKM) of transcripts derived from the sequence reads that could not be mapped to *T. mercedesae* genome. The e-value and sequence description are from the best hit by blastn.

Sample	**Transcript_id**	**Length (bp)**	**Expected Count**	**FPKM**	**E-value**	**Sequence description**
Male	Comp958_c0_seq1	10,139	18,978,220.27	98250.2	0	Kakugo virus genomic RNA, complete genome
Female	Comp1081_c0_seq1	10,143	16,525,134.53	96782.5	0	Kakugo virus genomic RNA, complete genome
	Comp1050_c0_seq1	217	20,674.66	7572.53	3E-98	Deformed wing virus isolate DWV_76 polyprotein gene
Nymph	Comp1078_c0_seq1	10,135	20,722,000.10	98058.95	0	Kakugo virus genomic RNA, complete genome
	Comp956_c0_seq1	237	17,021.51	4590.76	2E-99	Deformed Wing Virus gene for polyprotein, genomic RNA
Randomly collected mites	Comp3186_c0_seq1	9,953	5,775,088.37	88412.97	0	Deformed wing virus isolate PA, complete genome
	Comp2907_c0_seq1	239	40540.7	37208.93	2E-109	Kakugo virus gene for polyprotein, RNA dependent RNA polymerase region
	Comp4020_c0_seq1	364	6	3.13	3E-84	Strawberry latent ringspot

						virus NCGR MEN 454.001 segment RNA1
	Comp4325_c0_seq1	333	8	4.68	8E-32	Cycas necrotic stunt virus PP1 gene for polyprotien 1

2) Phylogenetic analysis of DWV

Because a huge number of the sequence reads represent DWV RNA, this must decrease the read coverage for mite transcriptome assembly, But it offered good opportunity to assemble the genome of DWV-*T. mercedesae* Wuxi strain. The major challenge for assembling NGS short read sequence to the consensus viral RNA sequence is that the extremely uneven read depth combined with the extensive viral population diversity (Hunt et, al 2015). No extra RNA-seq assembler was applied for the DWV genome assembly since Trinity has been used to assemble viral sequence because it can handle irregular read depth (Hunt et, al 2015), The first published DWV genome is 10140 nucleotides in length excluding the poly(A) tail and contains a single large open reading frame encoding a 328-kilo Dalton (kda) polyprotein (Lanzi et al., 2006). Table 2.7 showed the Trinity assemblies almost completed whole DWV-wuxi genome in all transcriptome samples. Subsequently, I preformed a phylogenetic analysis to study the relationships between the assembled DWV-wuxi genomes and other complete DWV genomes obtained from Genbank (Figure 2.7). As expected, the DWV-wuxi strains clustered with the Japan 'Kakugo' and Korean DWV strains, and Chilensis and Italia DWV genomes formed the other main group, but the U.K strain strand away all others. The phylogenetic tree reflects the geography and polymorphic relationships of DWV evolution. Previously, Forsgren et, al (2015) indicated the DWV sequences isolated from honey bee colony from Haikou in the southern island province (Hainan) of China, were largely clonal with the Italy sequences or with the Japan 'Kakugo' sequence. This could be a result of geographic isolation of DWV in the Hainan island, although phylogenetic analysis only based on the polymorphic DWV-Lp sequences.

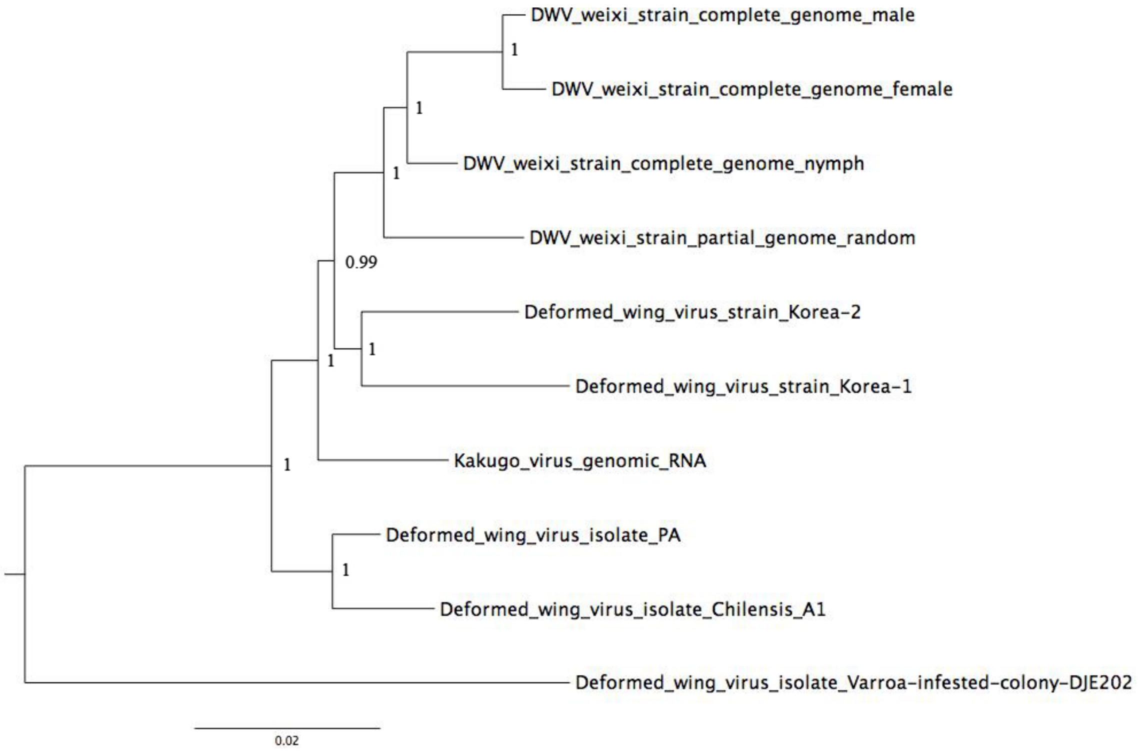

DWV_weixi_strain_complete_genome_male
1
DWV_weixi_strain_complete_genome_female
1
DWV_weixi_strain_complete_genome_nymph
1
DWV_weixi_strain_partial_genome_random
0.99
Deformed_wing_virus_strain_Korea-2
1
Deformed_wing_virus_strain_Korea-1
1
Kakugo_virus_genomic_RNA
1
Deformed_wing_virus_isolate_PA
1
Deformed_wing_virus_isolate_Chilensis_A1
Deformed_wing_virus_isolate_Varroa-infested-colony-DJE202
0.02

Figure 2.7. Phylogenetic analysis of the full genome sequences of various DWV isolates. The phylogenetic tree (rooted at middle point) was constructed using Mrbayers, but the poorly aligned 5' and 3' ends sequences were trimmed before constructing the phylogenetic tree. DWV-*T. mercedesae* Wuxi complete genome male, DWV-*T. mercedesae* Wuxi complete genome female, DWV-*T. mercedesae* Wuxi complete genome nymph, DWV-*T. mercedesae* Wuxi partial genome random correspond to the assembled transcripts comp958_c0_seq1, comp1081_c0_seq1, comp1078_c0_seq1 and comp3186_c0_seq1, respectively in Table 2.7. DWV genomes sequences are available in genbank: Kakugo virus (accession number: AB070959.1) was isolated in Japan, Deformed wing virus isolate PA (accession number: AY292384.1) was isolated in Italy, Deformed wing virus isolate Chilensis A1 (accession number: AY292384.1) was isolated in Chile, Deformed wing virus strains Korea-1 and -2 (accession number: JX878304.1 and JX878305.1) were isolated in South Korea. Deformed wing virus from *Varroa* infested colony DJE202 (accession number: KJ437447.1) was isolated in U.K.

3) Transcritpome assembly

The reads mapped to honey bee genome and virus genomes were removed before the assembly using both Trinity and Cufflinks. Trinity generated a single *de novo* transcript assembly based on combining all reads across male, female, nymph, and randomly collected mites. Cufflinks used the repeatmasker and repeatmodeler masked *T. mercedesae* genome as reference to produce a single genome reference-based transcript assembly where the exon/intron junctions were generated by Tophat mapping (default parameters except -r 20) in all transcriptome datasets. The assembly statistics were summarized in Table 2.8.

Table 2.8 Statistics for *T. mercedesae* trascriptome assemblies. All statistics are based on transcripts with the size >= 200 bp.

	Trinity	**Cufflinks**
Total number of transcripts	142,174	70,004,371
Total length of transcripts	274,714,415	70,156,311
GC (%)	45.98	46.98
N50 of transcripts	3,981	2,248

3. Genome annotation

3.1 Introduction

After successful genome assembly of *T. mercedesae*, the next step is gene annotation with the genome. A number of pipelines have been developed for genome-wide annotation of gene structures such as Maker (Cantarel et al., 2008; Holt and Yandell 2011), PASA (Haas et al., 2003), and Enembl (Curwen et al., 2004). Although these pipelines differ in their details, all each has three distinct phases in common.

In the first phase, 'low-complexity' sequences such as transposable (mobile) elements and repeats are identified and masked in the genome assembly. Eukaryotic genomes are very rich in the repeat sequences but these sequences are rarely present as the complete elements. The repeats can often insert within other repeats and only short fragments often exist within the other fragments. Unmasked repeats can seed millions of spurious blast alignments in the second phase and result in errors for gene annotation. Many transposons have the open reading frames (orfs) which are recognized as real protein-coding genes (exons) by gene predictors. Thus, masking repeat sequences in the genome is crucial step for the accurate annotation of protein-coding genes.

In the second phase, the exon-intron structures were predicted with *ab initio* gene predictions followed by the alignment with ESTs, RNA-seq data or homologous proteins. The *ab initio* gene predictors can provide fast and easy means to identify genes in assembled DNA sequences. Although these tools use mathematical models to determine the exon-intron structures rather than external evidences, accurate gene predictions can be made with enough pre-existing training data (Reese and Guigo, 2006). For example, hundreds of gene models already existed before sequencing the genome of *D. melanogaster*, *C. elegan*, and *Homo sapiens*. However, such data sets are rarely available for newly sequenced genomes. Prior to this study, there is not any *T. mercedesae* gene deposited in Genbank. In the absence of pre-existing reference gene models, RNA-seq data have been thought to improve the accuracy of gene annotations (Holt and Yandell

2011). Usually, the RNA-seq data are *de novo* or reference-based assembled into transcripts before using to train gene predictors. On the other hand, many pipeline annotators also align transcripts and homologous proteins using blast or blat as the further ‘polishing’ step to identify coding gene structures.

In the third phase, all gene structures generated by different *ab initio* gene predictors and alignment-based evidences are produced and compared. Then, the most representative predictions of the exon-intron structure are selected as the final sets of gene annotation. This annotation phase can be done automatically with a number of tools (also known as a ‘combiner’) such as EVM (Haas et al., 2008), Maker, and PASA using a ‘chooser algorithm’.

T. mercedesae genome annotation reported in this section was based on an improved Maker based pipeline. Non-coding RNAs (ncRNAs) annotation was also preformed for *T. mercedesae* genome.

3.2 Materials and Methods

3.2.1 Genome annotation tools

1) Masking the repeat sequences and gene annotation

Detecting and masking the repeat sequences was performed with combination of *de novo* and homology-based approaches. *De novo* prediction of repeats involved building a *de novo* repeat library using repeatmodeler and subsequently running repeatmasker to find and classify the repeat sequences in the genome assembly. Then, the homology-based prediction of repeats was achieved using repeatmasker to further search the genome against a known repeat library issued on 1-13-2014. For the non-interspersed repeat sequences, I ran repeatmasker with the ‘-noint’ option, which is specified for simple repeats, micro satellites, and low complexity repeats. Tandem repeats in the genome were scanned with the TRF (version: 4.04; Benson 1999) program. All programs in this section were used with default parameters.

2) *De novo* gene prediction

Three *ab initio* gene predictors, including Augustus (version: 3.0.3; Stanke et al. 2006), SNAP (version: 2013-11-29; Korf 2004), and genemarker (version: 2.3e; Lukashin and Borodovsky 1998) were used to obtain *de novo* predicted gene structures.

3) Gene combiner

Maker (version: 2.31.4) is a genome annotation pipeline which does not predict genes but leverage existing software tools such as the gene predictors in 4.2.1, and integrate their output to produce the most representative gene model for a given location based on evidence alignments.

4) Protein-coding prediction pipelines

The protein-coding prediction pipeline shown in Figure 3.1 is composed of three major phases:

In the first phase, repeatmasker and repeatmodeler were used to mask known repeat sequences and novel repeat sequences in the genomic scaffolds.

In the second phase, RNA-seq reads obtained from all samples were aligned to the masked genomic scaffolds to determine the exon-intron junctions using Tophat. Cufflinks used the spliced alignments to reconstruct 44,614 transcripts from which 12,298 transcripts with intact coding sequences were selected by transdecoder (version: r20140704; http://transdecoder.sourceforge.net/) to train *de novo* gene prediction softwares including Augustus, SNAP, and Genemarker. As a result, 32,561, 67,258, and 79,928 gene models were predicted by Augustus, SNAP, and Genemarker, respectively (Table 3.1). Invertebrate refseq protein sequences (downloaded on 2014-05-17 form NCBI) were aligned with the genome scaffolds using blastx. Sequences of assembled transcripts were aligned to the genome using blastn. Blast has no model for splice sites, and thus the ends of sequence alignments are only rough approximation of exon boundaries (Slater and Birney, 2005). Exonerate could ‘polish’ blast hits to determine the exact exon/intron boundary in the genome after matching highly similar transcripts and proteins to the input genomic sequence.

In the third phase, Maker integrated data from RNA-Seq, *de novo* gene prediction, and protein alignment with default parameters to produce an integrated gene set with high quality. Genes identified by *de novo* prediction, which did not overlap with any genes in the integrated gene set, were added to the gene set if they showed significant hit (BLASTP E-value < 1e-5) to Swissprot proteins or could be annotated by Interproscan (see below) with superfamily database.

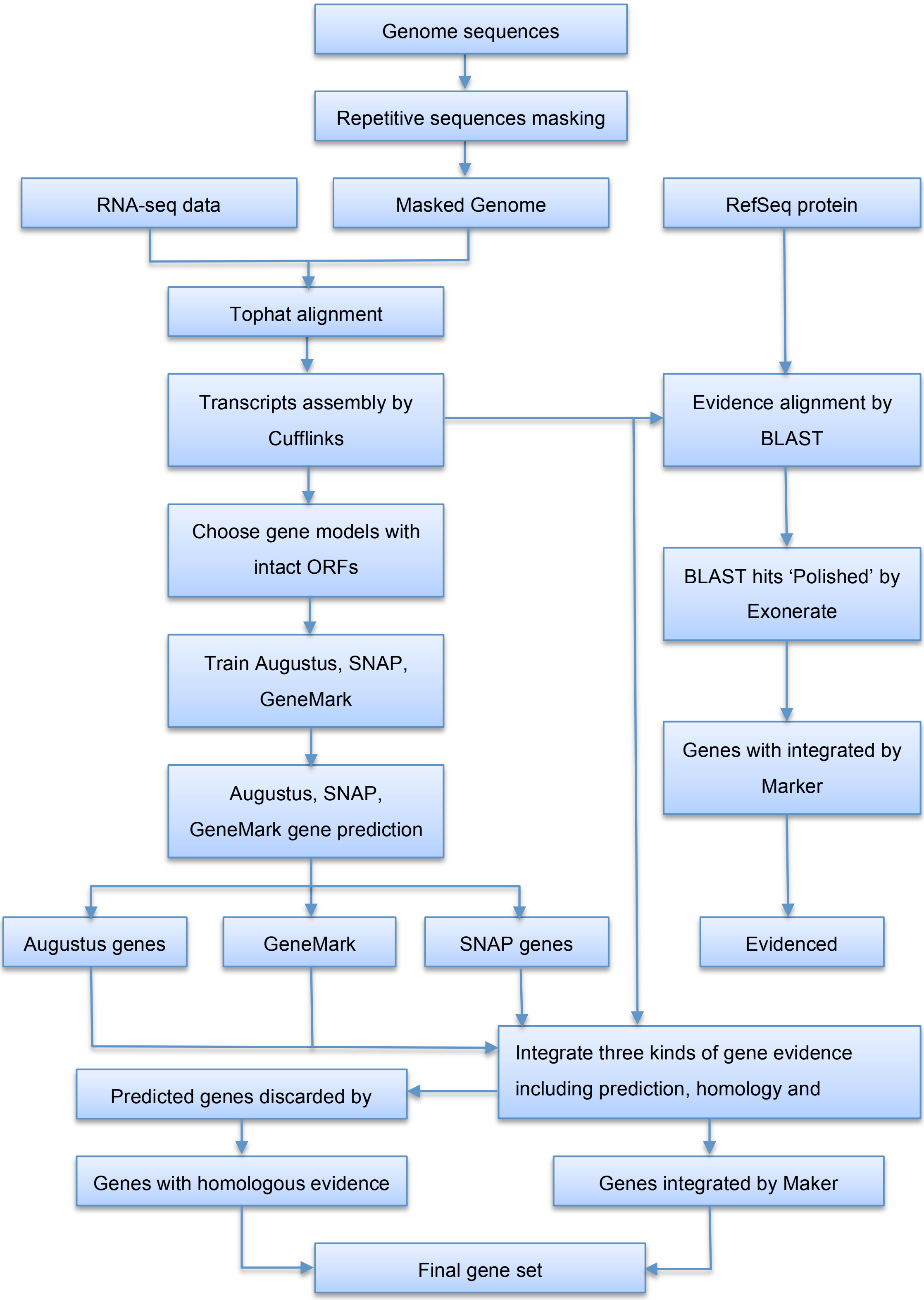

Genome sequences
Repetitive sequences masking
RNA-seq data
Masked Genome
RefSeq protein
Tophat alignment
Transcripts assembly by Cufflinks
Evidence alignment by BLAST
Choose gene models with intact ORFs
BLAST hits 'Polished' by Exonerate
Train Augustus, SNAP, GeneMark
Genes with integrated by Marker
Augustus, SNAP, GeneMark gene prediction
Augustus genes
GeneMark
SNAP genes
Evidenced
Integrate three kinds of gene evidence including prediction, homology and
Predicted genes discarded by
Genes with homologous evidence
Genes integrated by Maker
Final gene set

Figure 3.1 Maker based *T. mercedesae* genome annotation pipeline. The repeat sequences were removed by repeatmasker. Cufflinks was used to obtain a reliable gene set for training gene prediction software. Maker was used to integrate different types of genetic evidence and to generate a consensus gene set.

5) nc (non-coding) RNA annotation

In this analysis, four types of ncRNA were annotated: transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA, and small nuclear RNA. tRNA genes were predicted by trnascan-SE (version 1.3.1; Lowe and Eddy 1997) with eukaryote parameters. rRNA fragments were identified by aligning the rRNA template sequences from invertebrate animals (database: SILVA 119) to the *T. mercedesae* genome using blastn with an E-value cutoff of 1e-5. MicroRNA and small nuclear RNA genes were inferred by the Infernal software (version 1.1.1; Nawrocki and Eddy, 2013) using release 12 of the Rfam database.

6) Protein-coding gene annotation

Domains are the evolutionary units of protein-coding genes and their emergence and modular rearrangements are strongly associated with adaptive processes that are not always obvious at the gene level. Initial and principal domain annotation was performed using the Pfam database (vision: release 27) using HMMER hmmscan script with the default E-value cutoff at 10. Additional domains were assigned using Interproscan (version 4.8; Zdobnov and Apweiler, 2001) with superfamily, gene3d, tigrfams, smart, prosite, and prints domain models. The domian/motif based GO term was also obtained through interproscan search.

Blast2GO pipeline (version 2.5; Conesa et al., 2005) was used to further annotated proteins by Gene ontology (GO) terms. In the first step, a total of 17,508 protein sequences were searched against nr database with blastp. The E-value cutoff was set at 1e-6 and taking the best 20 hits for annotation. Base on the blast results, Blast2GO pipeline then predicted the functions of the sequences to assign GO terms, and merged the interproscan deduced domian/motif based GO terms into this Blast based annotations.

The metabolic pathway was constructed based on the KAAS (KEGG Automatic Annotation Server) online server (Kanehisa and Goto, 2000) using the recommended eukaryote set plus all other available insects and *I. scapularis*. The pathways in which each gene product might be involved were derived from the best KO hit with BBH (bi-directional best hit) method.

7) Testing the completeness of gene prediction

To demonstrate the completeness of *T. mercedesae* genome assembly and gene modeling, blastp was used to align *T. mercedesae* proteins with 458 core eukaryotic genes from CEGMA (see section 2.2.3) by setting the cutoff E-value < 1e-5.

8) Genome features

A set of custom perl scripts was used to study the *T. mercedesae* genome features.

3.3 Results and Discussion

3.3.1 Genome annotation and the quality assessment

Maker based *T. mercedesae* genome annotation pipeline (Figure 3.1) produced a gene set that contains 15,190 protein-coding genes (Table 3.1), in which 14,591 (96.1%) genes were supported by at least one RNA-seq read and 11,275 (74.2%) genes supported by at least 100 RNA-seq reads from all four transcriptomes (see Table 1.1). I found that 451 (98.4%) of the 458 CEGMA core genes are identified in the gene model, while 435 (94.9%) and 443 (96.7%) are supported by the cufflinks RNA-seq assembly and Trinity (*de novo*) RNA-seq assembly, respectively. With the Orthomcl gene clustering (see section 4.2.3), 9,344 out of 15,190 (61.5%) show similarity to proteins from species other than *T. mercedesae* (Table 3.2; Figure 4.4). In *M. occidentalis*, *I. Scapularis*, and *T. urticae* have, 86%, 58%, and 67% of the proteins have homologues from other species, respectively (Table 3.2).

In the current genome assembly, only 7.25% of the genome represents repetitive sequences, in which about half is unclassified repeats (3.43%) and the high proportion of common transposable elements such as satellite (2.55%) and simple repeat (2.12%) is also present (Table 3.3). Due to the high coverage of repeated sequences in the assembly, the calculated percentages of repetitive sequences must be lower than real values. 19 micrornas, 310 tRNAs, 129 rRNAs, and 123 snRNAs were also annotated in the genome of *T. mercedesae* (Table 3.4).

GO terms were assigned to the gene models for *T. mercedesae* using Blast2GO pipeline that integrates interproscan deduced domain/motif based GO terms, and this yielded 9,033 genes with at least one GO term. These genes were grouped into three main GO categories at level 2 (biological process, cellular component, and molecular function) and 49 subcategories (Figure 3.2). The most abundant Gene Ontology (GO) biological processes represented by small mite genes were cellular process (53.0%) and metabolic process (60.0%). The most abundant molecular processes were binding (61.2%) and catalytic activity (41.5%). The most abundant cellular components were cell (55.6%) and organelle (32.0%). Overall, the distribution of GO classifications of genes was very similar among *T. mercedesae*, *M. occidentalis* and *I. Scapularis* (Figure 3.2).

The interpro annotation showed that 5,203 distinct domains were present in 9,945 sequences, and 6,134 genes contained more than one interpro domain (Table 3.2). Pfam annotation yielded 95,125 predicted domains (7,542 distinct families found in 11,040 termite proteins; Table 3.2). KO-based annotation assigned 3,265 KEGG (Kyoto Encyclopedia of Genes and Genomes) orthology terms, 938 (24%) of which are involved in 175 metabolic pathways. A similar gene cluster and annotation pattern between *T. mercedesae* and its references suggests a good quality of predicted gene models in *T. mercedesae* (Table 3.2).

Table 3.1 General statistics of each gene set and the integrated prediction (Homolog based prediction was not included).

Method	Gene set	Number	Average exons per gene	Average length (bp)			
				Gene	CDS	Exon	Intron
De novo	Augustus	32,561	3	2,743	837	268	918
	Genemark	79,928	4	2,218	525	184	470
	SNAP	67,358	2	2,964	548	224	1,441
RNA-seq		37,857	5	6,603	996	473	934
Maker		14,423	5	5,323	1,138	381	792
Final gene set		15,190	5	5,298	1,117	407	788

Table 3.2 Comparison of statistics on the gene models for *T. mercedesae* and reference genomes.

Species	**Total proteins**	**Clustered by orthomcl (%)**	**Clustered with small mite (%)**	**With Pfam (%)**	**Unique Pfam**	**With interpro domain (%)**	**With more than one interpro domain (%)**	**Unique interpro domain**
T. mercedesae	15,190	9,344 (61.5%)	-	11,501 (75.7%)	7,542	9,945 (65.5%)	6,134 (40.4%)	5,203
M. occidentalis	11,738	10,048 (85.6%)	8,630 (73.5%)	10,598 (90.3%)	7,390	8,915 (75.9%)	5,985 (51.0%)	5,132
I. Scapularis	20,486	11,875 (58.0%)	7,015 (34.2%)	14,743 (72.0%)	8,190	12,442 (60.7%)	7,520 (36.7%)	5,544
S. mimosarum	27,135	15,546 (57.3%)	7,909 (29.1%)	20,278 (74.7%)	9,190	16,830 (62.0%)	10,127 (37.3%)	5,801
T. urticae	18,342	12,338 (67.3%)	6,730 (36.7%)	12,796 (69.8%)	7,862	10,312 (56.2%)	6,660 (36.3%)	5,120
D. melanogaster	13,918	10,434 (75.0%)	6,149 (44.2%)	12,285 (88.3%)	8,017	10,881 (78.2%)	7,227 (51.9%)	5,915
A. mellifera	15,314	9,117 (59.5%)	6,008 (39.2%)	11,057 (72.2%)	8,377	9,526 (62.2%)	6,219 (40.6%)	6,062

Table 3.3 Identification of the repetitive sequences in *T. mercedesae* genome. The results integrate both homology-based and *de novo* predictions. The total lengths of the repetitive sequences and their corresponding percentages in the genome for different categories were calculated based on the assembled genome size, i.e. 353Mb.

Repeat Type	**Length (bp)**	**% of genome**
DNA	1,160,558	0.329
LINE	3,659,707	1.037
SINE	8,458	0.002
LTR	762,578	0.216
Satellite	90	2.550
Simple repeat	7,483,749	2.120
Unknown	12,122,518	3.434
Small RNA	57,046	0.016
Low complexity	340,275	0.096
Total	25,594,979	7.250

Table 3.4 Genes for non-coding RNA, microrna (mirna), transfer RNA (trna), ribosomal RNA (rrna), and small nuclear RNA (snrna), in *T. mercedesae* genome. The rrnas were divided into four sub-classes based on their molecular weight (18S, 28S, 5.8S and 5S), whereas the snrnas contain the sub-classes of CD-box, HACA-box, and spliceosomal RNA. The total lengths of the Non-coding RNA genes and their corresponding percentages in the genome for different categories were calculated based on the assembled genome size, i.e. 353Mb.

Type		**Copy #**	**Average length (bp)**	**Total length (bp)**	**% of genome**
Mirna		19	79.1	1503	0.000425775
Trna		310	73.7	22850	0.006473027
Rrna	**18S**	10	204.1	2041	0.000578181
	28S	23	280.57	5453	0.001544745
	5.8S	1	154	154	4.36256E-05
	5S	95	116	11024	0.003122917
	Total	129	754.67	18672	0.005289468
Snrna	**CD-box**	8	165.5	1324	0.000375067
	HACA-box	1	135	135	3.82433E-05
	Splicing	114	128.73	15816	0.004480411
	Total	123	429.23	17275	0.004893721

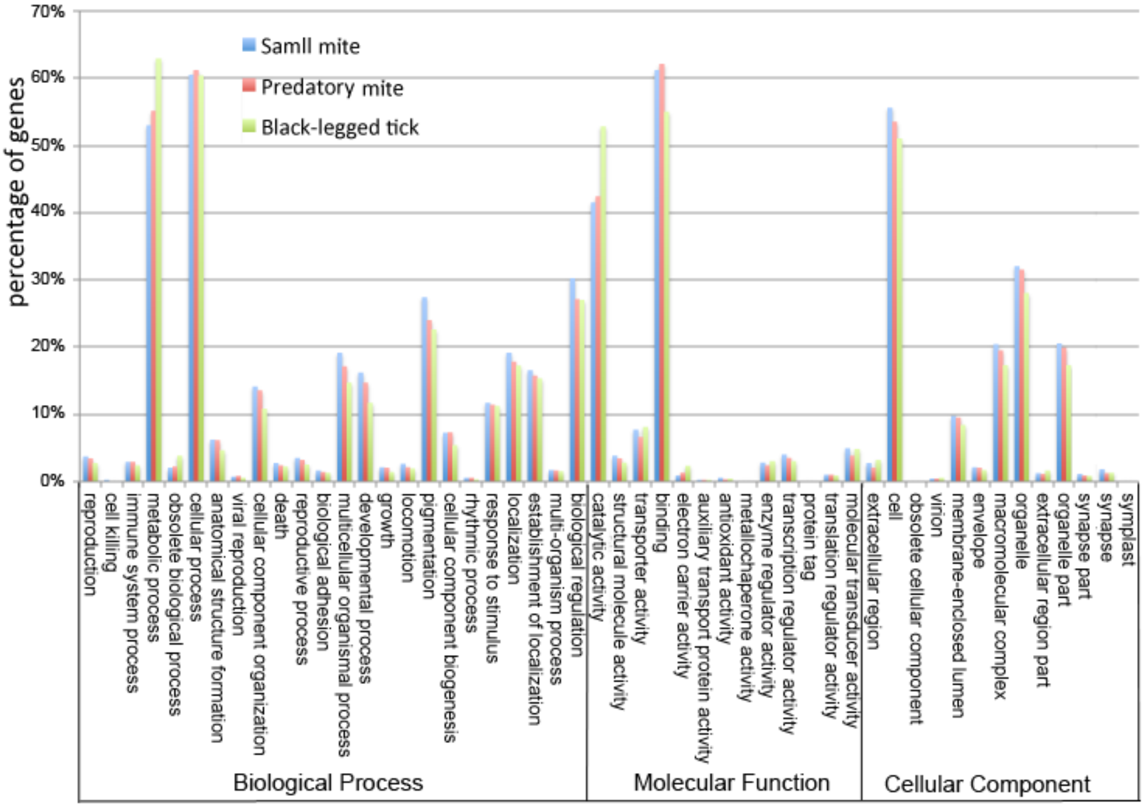


Figure 3.2 Gene ontology classification of small mite (*T. mercedesae*), predatory mite (*M. occidentalis*), and black-legged tick (*I. Scapularis*) protein-coding genes which are grouped into three main categories (biological processes, cellular components, and molecular function) and 50 subcategories based on GO level 2 classification. The X-axis represents the GO categories, whereas the **Y**-axis denotes the percentage of protein-coding genes.

3.3.2 Genomtic features

T. mercedesae has a lager genome size than *M. occidentalis* and many insects (see section 1.3.2). In order to understand the basis of large genome, the gene density (Figure 3.3), the size distribution of genes (Figure 3.4 A), the lengths of coding sequences (CDS) (Figure 3.4 B), exons (Figure 3.5 A) and introns (Figure 3.5 B), and number of exons/gene (Figure 3.6) for *T. mercedesae* genome were compared with those of six reference genomes (see section 4.2.1). Figure 3.3 shows that *T. mercedesae* has a low gene density (with larger intergenic region) as other two species with the large genome, *S. mimosarum* and *I. Scapularis*. Compared to the six reference genomes, *T. mercedesae* demonstrates a median gene length and CDS length distribution (Figure 3.4 A and B). No obvious difference in the exon size was seen for all genomes tested (Figure 3.5 A);

however, *T. mercedesae* genome contains fewer short introns than those of other species (Figure 3.5 B). Therefore, *T. mercedesae* has relatively large genome because of the significant amount of repeated sequences, large intergenic region, and fewer short introns.

In the six reference genomes, *S. mimosarum* and *I. Scapularis* also have the large genome size, only *T. mercedesae* shows lacking short introns. This may indicate that large *T. mercedesae* introns are likely to contain functional elements in addition to what might be regarded as 'normal' intron structure (Bradnam and Korf 2008). Large introns may promote alternative splicing via exon-skipping and exon turnover during evolution likely due to frequent errors in their removal from maturing mRNA, and can be a reservoir of genetic diversity because of their greater number of mutable sites than short introns (Kandul and Noor, 2009).

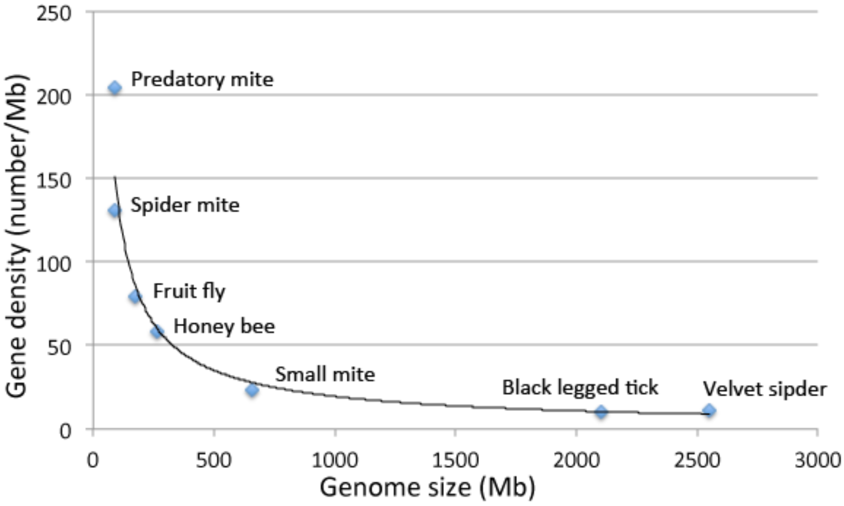


Figure 3.3 Genome sizes plot against gene densities of small mite (*T. mercedesae*) with six sequenced reference genomes, fitting a power relation.

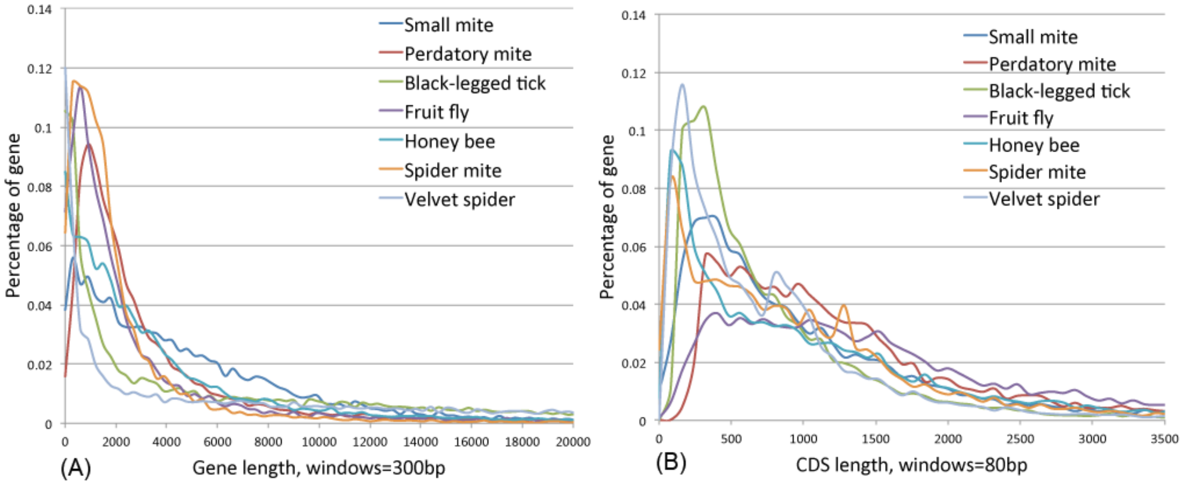


Figure 3.4 Comparison of the sizes of genes (A) and CDSs (B) among seven sequenced genomes including *T. mercedesae*. The x-axis indicates size (bp) of each genetic feature, and the y-axis indicates the percentage of genes that have the corresponding size.

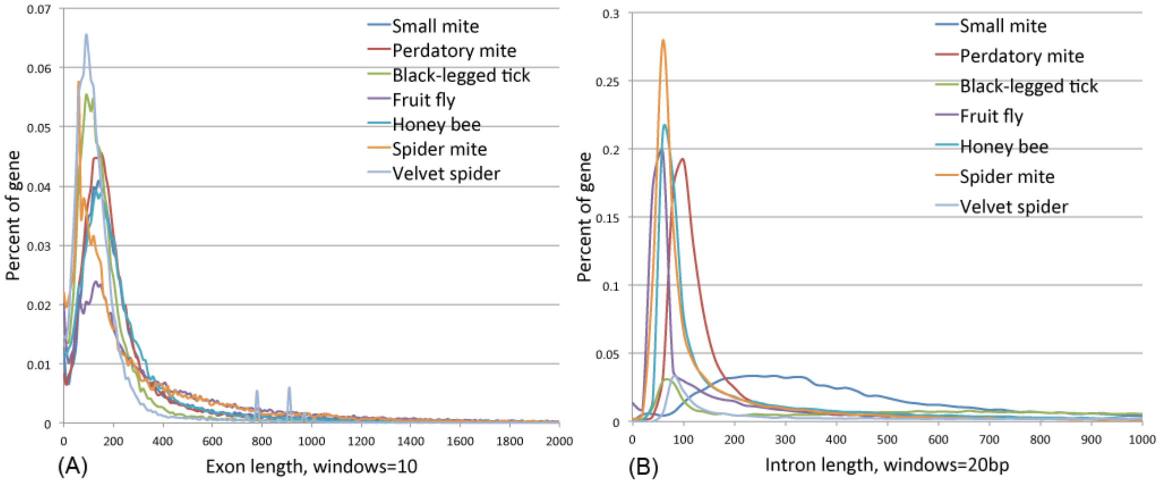


Figure 3.5 Comparison of the sizes of exons (A) and introns (B) among seven sequenced genomes including small mite (*T. mercedesae*). The x-axis indicates size (bp) of each genetic feature and the y-axis indicates the percentage of genes that have the corresponding size.

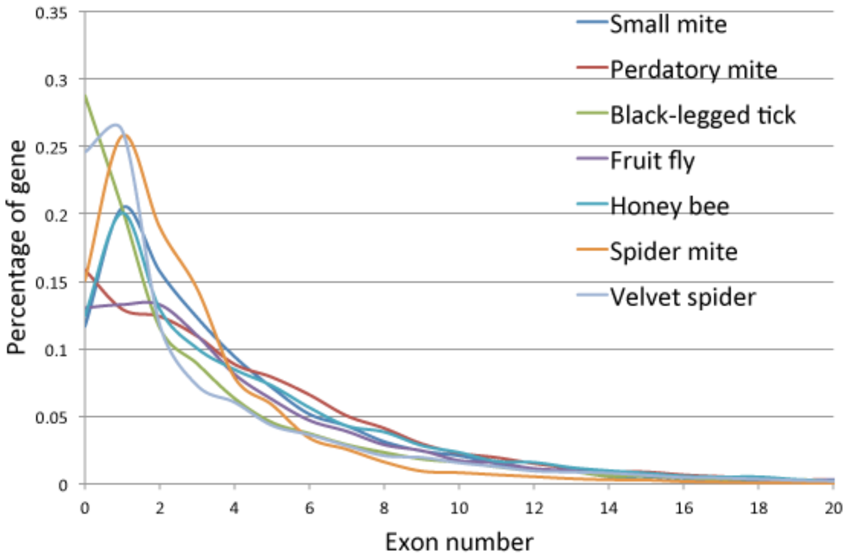


Figure 3.6 Comparison of the number of exons/gene among seven sequenced genomes including small mite (*T. mercedesae*). The x-axis indicates the number of exons/gene and the y-axis indicates the percentage of genes that have the corresponding number of exons/gene.

4. Evolutionary analysis

4.1 Introduction

Environmental changes are considered as the major driving forces for evolution of living organisms (Hoffman and Parsons 1991; Parsons 2005). Evolution can happen through actual environmental changes induced by either physical features of the environment such as climatic factors and pollutants, or the biotic environment such as host, parasites, competitors, and predators (Bijlsma et al., 1997). It can also occur through the displacement of organisms into new environments (Zeigler 2014). To enhance the survival or reproduction of organisms, they are likely to adapt to the environmental changes to fit better to new environments. The majority of adaptations are from three sources: modifications of existing characteristics (including exaptations), the addition of new characteristics, or the loss of former characteristics (Zeigler 2014). Genetic variations are the raw materials of these sources for the adaptation that can be supplied by three major mechanisms (Simpson 1952): conditions of growth affect the development; mechanism of sexual reproduction assures change from one generation to inherit to the next generation (genetic shuffling); mutations with selection as well as genetic drifts can produce changes in genes and more generally in chromosomes. Thus, natural selection actually selects the genetic variations that result in well adapting to their environments for the organisms.

A variety of genetic variations can happen during molecular evolution that involves point mutation, homologous recombination, gene duplication, and horizontal gene transfer (Chia and Guttenberg 2011). Adaptations of species due to environmental changes can be inferred by studying the molecular evolution with molecular sequence data from present-day organisms. The life cycle of *Tropilaelaps* mite was previously introduced in the background section. *Tropilaelaps* mites seem to spend whole life in a honey bee hive, and are usually inside the brood cell except female mite looks for the 5th instar larva for reproduction. *Tropilaelaps* mite has been inside a honey bee colony in association with Asian honey bees though the life history. The environment must be quite stable inside a honey bee colony, for example, both *A. Dorsata* and honey bee maintain the brood nest's temperature constant around 34^{o}c by thermoregulatory activities of the colony (Mardan and Kevan 2002). Under this living environment, what kinds of adaptations have happened on *T. mercedesae* during evolution? This question would be answered in this chapter.

4.2 Materials and Methods

4.2.1 Selection of the reference genomes

The following arthropod genomes were used for comparative analyses of *T. mercedesae* genome; *D. melanogaster* (fruit fly; GOS release: 6.03; Adams et al., 2000), *A. Mellifera* (honey bee; GOS release: 3.2; The Honey bee Genome Sequencing Consortium 2006), *Tetranychus urticae* (spider mite; GOS release: 20140320; Grbic et al., 2011) *Stegodyphus mimosarum* (velvet spider; GOS release: 1.0; Sanggaard et al., 2014), *I. Scapularis* (black-legged tick; GOS release: 1.4; *I. Scapularis* Genome Consortium, unpublished data), *M. occidentalis* (predatory mite; GOS release: 1.0; Hoy, et al., unpublished data). *Caenorhabditis elegans* (roundworm; GOS release: WS239; Coulson 1996) was used as outgroup. Protein data sets of above reference genomes are listed in Table 3.2. Domain, GO, and KEGG annotation of proteins in the reference species (if required) was conducted using the same approach as for *T. mercedesae* (see section 3.2).

4.2.2 Reciprocal blast

Orthologs are defined as homologous genes present after a speciation event, referred to as the “same genes” in different species (Fitch 2000). Reciprocal Best Hits (RBH) has been common proxy to identify the one to one orthology in comparative genomics. Here, I used reciprocal blastp approach (Proteinortho: version 5.05; Lechner et al., 2011) to compare the protein sequences and score the matches across *T. mercedesae* and all reference genomes.

4.2.3 Gene family clustering

Orthologous gene families were defined using Orthomcl (version: 1.4; Li et al., 2003) which generated graphical representation of sequence relationships, and then divided it into the sub-graphs using Markov Clustering Algorithm (MCL) from multiple eukaryotic genomes (Li et al., 2003). Orthomcl was run with default parameters for all proteins from *T. mercedesae* and all reference genomes including *C. elegan*.

4.2.4 Phylogenetic analysis

Phylogenetic analysis was conducted with several data sets of orthologous proteins, and the methods used for phylogenetic reconstruction include:

(1) Alignment

Orthologous protein sequences were aligned with Mafft (version: 7.012b; Katoh and Standley, 2013) or Kalign (version: 2.0; Lassmann et al., 2009). Both Mafft and Kalign are reliable multiple sequence alignment algorithms; however, compared to Mafft, Kalign performs more robustly when aligning large number of sequences or distant sequences in large-scale benchmark of generated alignments (Lassmann and Sonnhammer, 2005).

(2) Selection of conserved blocks

In phylogenetic analysis, poorly aligned positions and divergent regions in the aligned DNA or protein sequences can readily impact the phylogenetic reconstruction since they may not be homologous or may have been saturated by multiple substitutions (Castresana, 2000). In order to improve such phylogenetic reconstruction, Gblocks (version: 0.91b; Castresana, 2000) was used to automatically eliminate these divergent regions or gap positions prior to phylogenetic analysis. But, for big gene set, I trimmed the aligned sequences manually.

(3) Phylogenetic tree reconstruction with model-based methods

Phylogenetic trees were reconstructed using a maximum likelihood approach or

Bayesian approach. Phyml (version: 3.1; Guindon et al., 2010) estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences in the large number of substitution models coupled to various options to search the space of phylogenetic tree topologies. Mrbayes (version: 3.2.3; Ronquist and Huelsenbeck 2013) performs Bayesian inference of phylogeny by taking advantage of both codon-based and amino acid-based algorithms and adjusts them to the topology of the species tree. In order to make the best possible estimates of evolutionary distances among sequences on a tree, the best substitution models of amino acid substitution were always determined for alignment by prottest (version: 3.4; Darriba et al., 2011) with parameters set to "-all-matrices, -all-distributions, -AIC".

(4) Phylogenetic tree reconstruction with distance-matrix methods

Distance-based methods estimate genetic distances by calculating the pairwise distances between molecular sequences. The neighbor-joining method is used for building trees by distance-based methods. The neighbor-joining tree was reconstructed using MEGA software version 6.06 (Tamura et al., 2013).

4.2.5 MCMC species tree

PAML mcmctree (PAML version: 4.7; Yang 2007) was used to estimate divergence time with approximate likelihood calculation using the following fossil calibration times (Mya): tick-spider: 311-503 (oldest spider from coal, UK; Sanggaard et al., 2014), *T. urticae*-tick-spider: 395-503 (oldest Acari; Sanggaard et al., 2014), *A. Mellifera*-*D. melanogaster*: 238-307 (http://www.fossilrecord.net/) and nematode-arthropods: 521-581 (http://www.fossilrecord.net/). In the first step, rough substitution rate (rate per time unit) among species was computed by PAML codeml in the aligned super amino acid sequences with a species tree structure (see section 4.3.1): ((((((*M. occidentalis*, *T. mercedesae*), *I. Scapularis*), *S. mimosarum*), *T. urticae*), (*A. Mellifera*, *D. melanogaster*)), *C. elegan*) '@5.5', where '@5.5' tag was a simple fossil calibration of 450Ma ago for the root. In the second step, mcmctree was conducted with maximum likelihood estimates of the gradient and Hessian branch lengths for the species tree topology by setting ‘usedata = 3’ in mcmctree.ctl (control file). In the third step, the mcmctree process of PAML mcmctree was set to sample 10,000,000 times with the sample frequency set to 1,000 after a burn-in of 5,000,000 iterations. A rooted binary tree used by mcmctree was labeled with fossil calibrations using a lower and upper bound method: ((((((*M. occidentalis*, *T. mercedesae*), *I. Scapularis*), *S. mimosarum*) '>3.11<5.03', *T. urticae*) '>3.95<5.03', (*A. Mellifera*, *D. melanogaster*) '>2.385<3.072'), *C. elegan*), when the fossil calibration of 550Ma ago for the root age was given to mcmctree.ctl. The shape and scale parameters (rgene_gamma = 1 12.5; sigma2_gamma = 1 4.5) describing the gamma to the overall substitution rates were set according to the estimated substitution rate in the first step. Mcmctree was run based on a Hessian matrix calculated by codeml using WAG+Gamma after setting usedata = 2 in mcmctree.ctl. The fine-tuned parameter was automatically adjusted so that the acceptance proportions fall in the interval (20%, 40%). The other parameters were set at the default values. To check the convergence, mcmctree was run twice independently. After that, the posterior mean divergence times obtained from MCMC run1 analysis for each node in the species tree is plotted against the posterior mean divergence times obtained from MCMC run2 analysis for each node in the species tree.

4.2.6 Testing gene family expansion and contraction

CAFE (version: 3.1; De Bie et al., 2006) is a tool to employ a random birth and death model to study gene gain and loss in gene families across a specified phylogeny. The output of orthomcl (see section 4.2.3) and phylogenic ultrametric species tree created in the section 4.2.5 were processed by CAFE to compute the maximum likelihood value of the gene birth and death parameter (λ, probability of both gene gain and loss per gene per million years) over the whole tree (global parameter λ) or for user-specified subsets of branches (multi-parameter λ) in the tree. The global parameter λ describes the gene birth and death rates across branches of the tree for all gene families. Since different parts of the tree may evolve at different rates, the branches termed in different parts can specify the same of different lambda values. Additional λ parameters result in more complex birth and death models with a higher likelihood score. However, adding additional parameters does not always significantly improve a model to fit to a particular dataset. Because the more complex birth and death model differs from the simple model just by the addition of one or more λ parameters, I used the likelihood ratio test (LRT; LR = 2*(score of λ model 1 - score of λ model 2)) to assess the significance of models with additional λ parameters in the observed data. Because this LRT statistics approximately follow chi-square distribution (Huelsenbeck and Crandall 1997), the degree of freedom is equal to the number of additional λ parameters in the more complex model and the critical value of the test statistic can be determined from standard statistical tables. If the LR is greater than 95% ($P<0.05$) of the distribution (critical value), the model 2 was thought to be significantly better than the model 1. Based on the best λ parameter, CAFE computes an overall *P*-value for each gene family, and the families with overall *P*-values less than threshold (0.01) were considered as having an accelerated rate of gain or loss. Simultaneously, the branch specific *P*-values (0.001) were also calculated by CAFE. Branches with low *P*-values represent unusually large changes, either contraction or expansion, and are responsible for low overall *P*-values of significant families.

4.2.7 Maximum likelihood analysis for evolutionary rates of core genes

I employed codeml (PAML version: 4.7a; Yang 2007) in the PAML package to estimate the evolutionary rates of orthologous genes. The orthologous genes were defined as described in section 4.2.3 and aligned by Mafft based on the amino acid sequences using the '-auto' option. Pal2nal (version: 14; Mikita et al., 2006) was used to convert the amino acid sequence alignment and the corresponding DNA sequences into a codon alignment, and remove gaps using '-nogap' option. An unrooted concatenated phylogenetic tree previously generated by Phyml (see section 4.2.4; Figure 4.1A) was used as input data in PAML to analyze the selection pressure with the free-ratios model (one of branch models; nssites = 0; model = 1) to calculate different ω (dn/ds) values of nonsynonymous (dn) to synonymous (ds) substitution rate for each branch. Branches with dn/ds > 1 are considered under positive selection. The null model used for branch test was the one-ratio model (nssites = 0; model = 0) where ω was the same for all branches. Kappa and omega values were automatically estimated from the data, when clock was set to be entirely free to change among branches. P-value was determined twice using the log-likelihood difference between the two models compared to χ2 distribution with the difference in number of parameters between one-ratio and free-ratio models. To estimate significance with the P-value, likelihood-ratio test (LRT) was used to compare lnl values for each model and test if they are significantly different. The differences in log-likelihood values between two models were compared to chi-square distribution with degree of freedom equal to the difference in the number of parameters for two models. Measurement of ds was assessed for substitution saturation, and only ds values below 3.0 were maintained in the analysis for positive selection. Genes with high dn/ds (>10) were also discarded.

4.2.8 Enrichment analysis

GO enrichment analyses for genes were performed with Fisher's exact test embedded in the Blast2GO desktop version (version: 2.8). If not specifically stated, the P-values were corrected according to critical False Discovery Rate (FDR). The enrichments were tested by comparing the GO term complement against the background of the pooled set of GO terms from *T. mercedesae*.

GO enrichment analyses for interpro domains were performed with dcgor R package (Fang 2013) with dcenrichment () function based on the Fisher's exact test. P-values were adjusted with BH controlled FDR. All annotatable interpro domains in *T. mercedesae* genome are used as background.

4.3 Results and Discussion

4.3.1 Phylogenetic analysis of *T. mercedesae*

The phylum Arthropoda contains two major lineages, Myriochelata which includes the two subphyla, Myriapoda (e.g., millipede and centipede) and Chelicerata (e.g., spider, mite, and tick), and Pancrustacea which includes two subphyla, Crustacea (e.g., water flea and crab) and Insecta (insect) (Shultz and Regier 2000). Acari (mite and tick) is an exceptionally diverse group of Arachnida (Chelicerata), and currently divided into three major lineages: Acariformes, Parasitiformes, and Opilioacariformes (Dermauw et al, 2010). Argument about monophyly or diphyly of the Acari has not yet to be resolved. Traditionally, Acari was considered as monophyletic, but new evidences support that Acari is diphyletic. Van der Hammen (1989) first suggested that Acari is diphyletic, and Acariformes and Parasitiformes are most distantly related. The other phylogenetic study based on the 18S rRNA genes supported that Acariformes and Parasitiformes are two separated large monophyletic groups, and revealed two sister relationships: Acariformesa-Solifugae and Parasitiformes-Pseudoscorpionida (Dabert et al., 2010). Recently, phylogenetic analyses based on mitochondrial DNA have shown that Acari can be divided into two superorders, monophyletic Parasitiformes and Acariformes (Gu et al., 2014).

Because the recent study has stated that the fast evolution of Acariform mites may challenge the phylogenetic analysis due to the long-branch attraction (Dabert et al., 2010), I used a very strict e-value (1e-50) when preforming a reciprocal blastp to gate out the most variant orthologous genes across all genomes tested. The reciprocal blast search resulted in identifying a total of 926 highly conserved one-to-one orthologs in all eight genomes. Each of these proteins was aligned using Mafft in “-auto” option, and these alignments were trimmed using Gblocks with parameters of ‘-t=p’. After the trimmed protein sequences were concatenated into a “super-gene” for each species with a custom perl script, phylogenetic analysis was conducted with both Phyml and Mrbaye using the the best substitution model (LG) of amino acid determined by Prottest. Phyml estimated the maximum likelihood phylogeny for the concatenated alignment using default settings, and for 100 bootstrap resampling replicates. Mrbayes was run with the parameter set to 1,000,000 (1 sample / 100 generations) and the first 25% sample were burned-in, using *C. elegan* as outgroup. Both methods generated a tree with the same topology (Figure 4.1 A and B). I found that *T. urticae* (Acariform mite) outgroups the Parasitiformes (*T. mercedesae*, *M. occidentalis*, and blacklegged tick) and non-acarine *S. Mimosarum, S. mimosarum* (Figure 4.1 A and B). These phylogenetic trees confirmed the diphyly of Acari, and the Acariformes and Parasitiformes are distantly related.

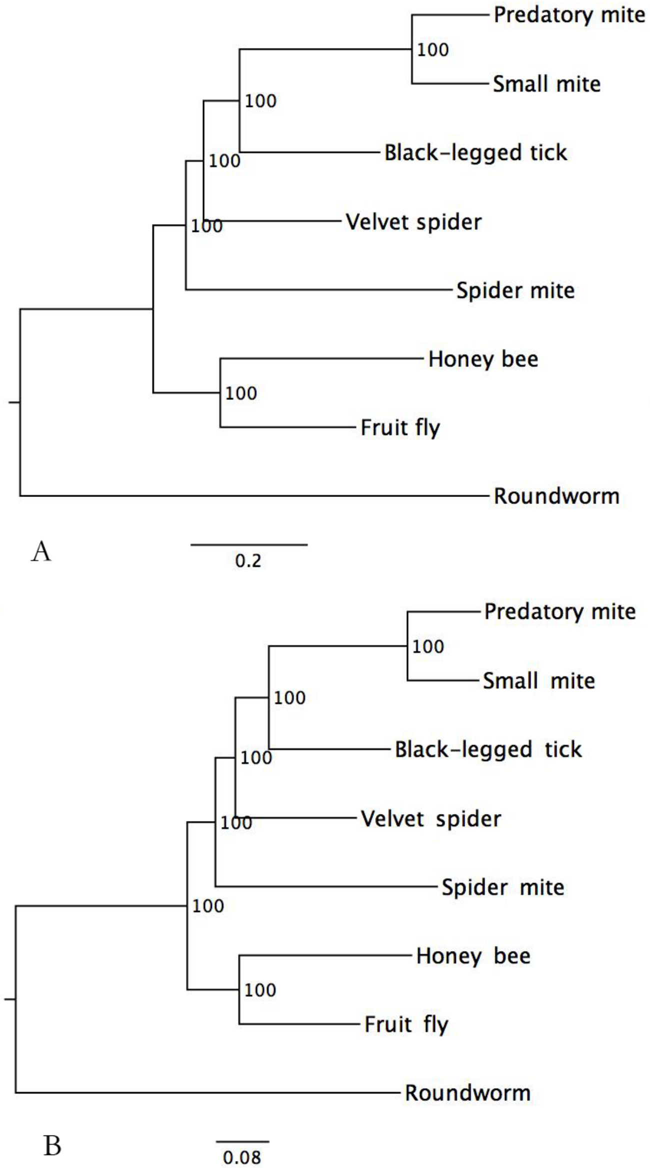


Figure 4.1 Phylogenetic trees based on the amino acid sequences of 926 one-to-one orthologous genes in eight species using Maximum likelihood (A) and Bayesian methods (B).

4.3.2 Estimating the divergence time

Molecular clock data can make robust contribution to phylogenetics. Based on topology defined by the phylogenetic trees (Figure 4.1 A and B), I estimated the divergence time of each species using the Bayesian MCMC method in PAML package together with several fossil records that can be used to calibrate the time points. Linear distribution showed in Figure 4.2 suggests that the convergence was achieved between two MCMC runs. In the Figure 4.3, *T. mercedesae* and *M. occidentalis* were clustered together representing Parasitiformic mites and shared the last common ancestor about 123 million years ago (Mya). More importantly, the separation of Parasitiformic mites and ticks was estimated to be approximately 302 Mya for the first time (Figure 4.3).

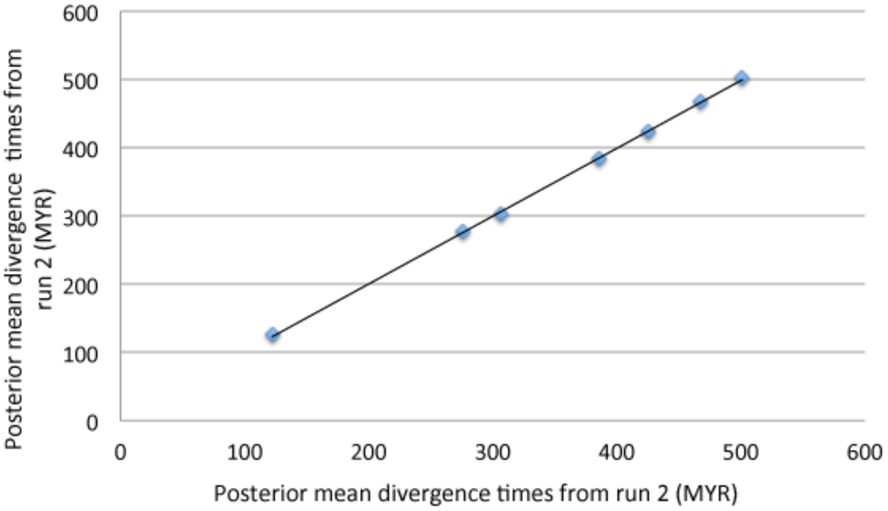


Figure 4.2 The posterior mean divergence times obtained by MCMC run1 analysis for each node in the species tree is plotted against the posterior mean divergence times obtained by MCMC run2 analysis for each node in the species tree.

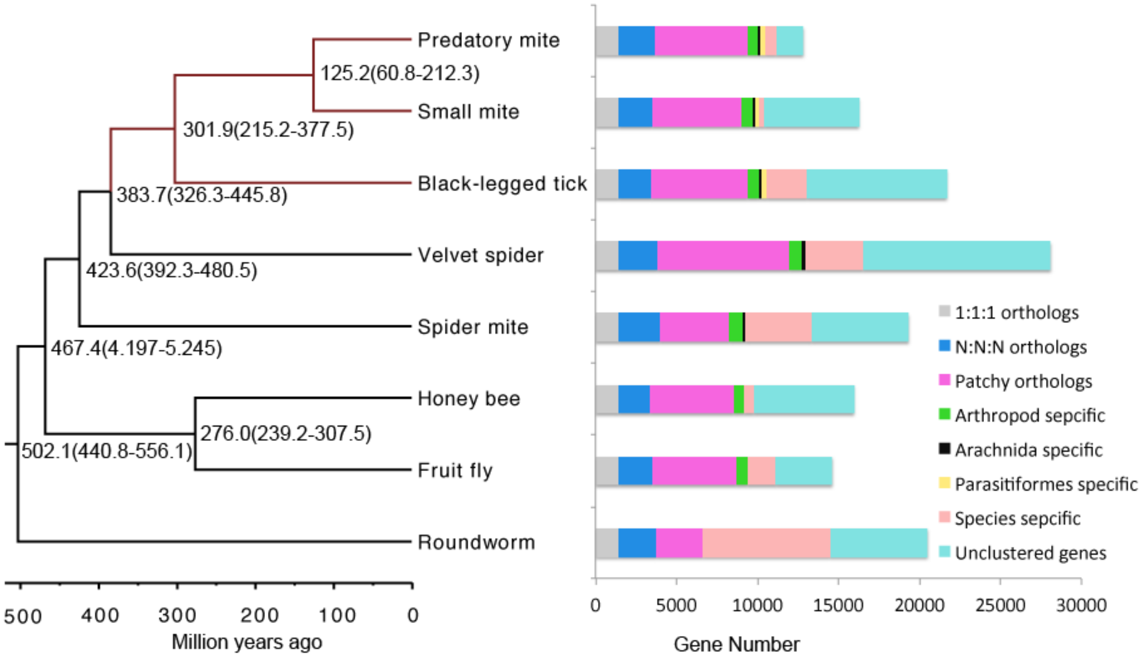


Figure 4.3 Estimated divergence times using a relaxed molecular clock with fossil calibration time and classification of protein-coding genes between eight species of Ecdysozoa. The red branches represent Parasitiforms. Roundworm (*C. elegan*) was used as the outgroup, and a bootstrap value was set as 10,000,000. 1:1:1 orthologs include the common orthologs with the same copy numbers in the different species, and N:N:N orthologs include the common orthologs with different copy numbers in the different species. Patchy orthologs are shared between more than one, but not all species (excluding those belongs to previous categories). Unclustered genes include those that cannot be clustered into gene families.

4.3.3 Gene family cluster analysis

In *T. mercedesae*, 15,190 protein-coding genes were predicted by the MARKER based pipeline (Table 3.1). Orthomcl clustered the predicted proteins of *T. mercedesae* together with proteins from other arthropods and classified them into a total of 15,506 gene families. When compared with other species, *T. mercedesae* has the similar numbers of 1:1:1 orthologs and N:N:N orthologs, but fewer species specific gene orthologs (Figure 4.3). There are 226 orthologous clusters specifically shared between *T. mercedesae*, *M. occidentalis*, and *I. Scapularis* (Figure 4.4). These clusters may represent Parasitiformes specific genes.

5,091 gene families are shared between all species analyzed in this study, while 119 gene families including 332 genes and 5,846 unclustered genes could be identified only in *T. mercedesae* but not in other reference genomes (Figure 4.3; Figure 4.4). These total 6,178 *T. mercedesae* specific genes could have important roles for the evolution and adaptation to make *T. mercedesae* unique. Among these genes, two GO terms, 'macromolecules transport' and 'nucleic acid processing (assembly and package)', which are required for cell growth and division, are highly enriched (FDR > 0.05) (Table 4.1). Other enriched GO terms include ‘peroxidase reaction’, ‘oxidoreductase activity acting on peroxide as acceptor’, ‘peroxidase activity’, and ‘antioxidant activity’ (Table 4.1) that would be related to antioxidant defense for protecting *T. mercedesae*s from the oxidative damages.

In eight species analyzed, *T. mercedesae* and *I. Scapularis* are only two blood feeding parasites. The *I. scapularis* is also an ectoparasite like *T. mercedesae* but feed on a variety of hosts (Pinger 2008). The tick larvae feed on a variety of mammals and birds but most frequently white-footed mouse. As the tick becomes a nymph and adult, it starts to feed on large mammals such as raccoons, deer, dog, cat, and human. The results of gene clustering in seven arthropod species showed that 135 gene families are specifically shared between *T. mercedesae* and *I. Scapularis* (Figure 4.4). *T. mercedesae* shares many genes with *M. occidentalis* as expected but the GO enrichment analysis in these specific gene families could reveal specific genes associated with blood feeding Parasitiforms. Table 4.2 shows that GO terms related to renal excretory system (renal tubule as the central part) and nervous system are highly enriched. Insect blood (hemolymph, 90% water and 10% other molecules) contains many inorganic salts (Na^+, Cl^-, K^+, Mg^{2+}, and Ca^{2+}), and organic compounds (carbohydrates, proteins, and lipids) (Nation 2008). By sucking hemolymph from honey bee larva, *T. mercedesae* faces the significant physiological challenges to osmoregulation and metabolism. The over representation of GO terms associated with renal tubule development and other related GO terms in renal excretory system (Table 4.2) suggests the enhanced ability for excretion by renal (malpighian) tubules, which may play critical roles for the initial processing of ingested hemolymph by excreting excess water and salts. Such role for malpighian tubules has been demonstrated with female mosquito (*Aedes albopictus*) during blood feeding (Williams et al., 1983). The GO terms enriched in the nervous system are also involved in motor neuron migration (GO:0097475, GO:0097476 and GO:0097477) and regulation of granule cell precursor cell proliferation (GO:0021936, GO:0021937 and GO:0021940) (Table 4.2).

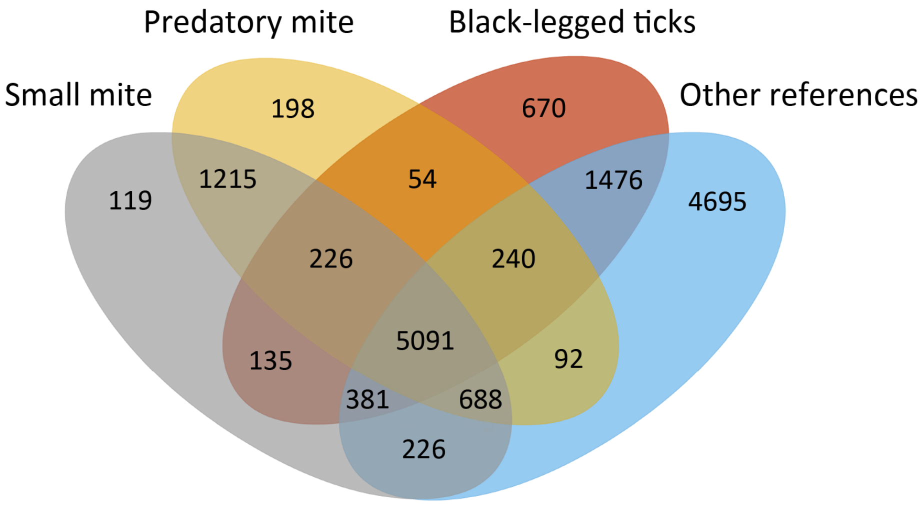


Figure 4.4 The number of gene families shared among small mite (*T. mercedesae*), predatory mite (*M. occidentalis*), black-legged tick (*I. Scapularis*) and other references including Velvet spider (*S. mimosarum*), spider mite (*T. urticae*), fruit fly (*D. melanogaster*), and honry bee (*A. Mellifera*) by Orthomcl classification algorithm.

Table 4.1 GO enrichment of *T. mercedesae* specific genes. The corrected FDR P-value (< 0.05) was calculated by Fisher's extract test embedded in the blast2go software desktop version.

GO Term	**Description**	**Type**	**FDR**
GO:0042302	Structural constituent of cuticle	F	1.30E-04
GO:0000786	Nucleosome	C	1.50E-04
GO:1990104	DNA bending complex	C	1.50E-04
GO:0031497	Chromatin assembly	P	7.20E-04
GO:0020037	Heme binding	F	7.20E-04
GO:0044815	DNA packaging complex	C	7.20E-04
GO:0046906	Tetrapyrrole binding	F	8.30E-04
GO:0006334	Nucleosome assembly	P	9.60E-04
GO:0005215	Transporter activity	F	9.60E-04
GO:0006820	Anion transport	P	1.30E-03
GO:0071944	Cell periphery	C	1.50E-03
GO:0006333	Chromatin assembly or disassembly	P	1.50E-03
GO:0005886	Plasma membrane	C	2.00E-03
GO:0032993	Protein-DNA complex	C	2.20E-03
GO:0006811	Ion transport	P	2.80E-03
GO:0005230	Extracellular ligand-gated ion channel activity	F	2.80E-03
GO:0034728	Nucleosome organization	P	3.00E-03
GO:0015698	Inorganic anion transport	P	4.20E-03
GO:0065004	Protein-DNA complex assembly	P	4.30E-03
GO:0007215	Glutamate receptor signaling pathway	P	4.40E-03
GO:0045202	Synapse	C	6.10E-03
GO:0055085	Transmembrane transport	P	7.20E-03
GO:0022857	Transmembrane transporter activity	F	7.20E-03
GO:0006323	DNA packaging	P	7.30E-03
GO:0006804	Peroxidase reaction	P	7.60E-03

GO:0016684	Oxidoreductase activity, acting on peroxide as acceptor	F	7.60E-03
GO:0004601	Peroxidase activity	F	7.60E-03
GO:0016209	Antioxidant activity	F	9.40E-03
GO:0022834	Ligand-gated channel activity	F	1.30E-02
GO:0015276	Ligand-gated ion channel activity	F	1.30E-02
GO:0071824	Protein-DNA complex subunit organization	P	1.60E-02
GO:0005576	Extracellular region	C	1.60E-02
GO:0030054	Cell junction	C	3.20E-02
GO:0071103	DNA conformation change	P	3.40E-02
GO:0034220	Ion transmembrane transport	P	4.70E-02

Table 4.2 GO enrichment of genes specifically shared between *T. mercedesae* and *I. Scapularis* The P-value (< 0.001) was calculated by Fisher's extract test embedded in the blast2go software desktop version.

GO Term	Description	Type	P-Value
GO:0061326	Renal tubule development	P	4.00E-06
GO:0021587	Cerebellum morphogenesis	P	8.60E-05
GO:0035295	Tube development	P	1.30E-04
GO:0060562	Epithelial tube morphogenesis	P	1.40E-04
GO:0072080	Nephron tubule development	P	2.00E-04
GO:0048793	Pronephros development	P	2.00E-04
GO:0035239	Tube morphogenesis	P	2.00E-04
GO:0039020	Pronephric nephron tubule development	P	2.50E-04
GO:0039019	Pronephric nephron development	P	2.50E-04
GO:0021549	Cerebellum development	P	2.50E-04
GO:0072009	Nephron epithelium development	P	2.50E-04
GO:0022037	Metencephalon development	P	3.10E-04
GO:0072073	Kidney epithelium development	P	3.10E-04
GO:0072001	Renal system development	P	3.30E-04
GO:0001655	Urogenital system development	P	3.80E-04
GO:2000768	Positive regulation of nephron tubule epithelial cell differentiation	P	4.90E-04
GO:2000744	Positive regulation of anterior head development	P	4.90E-04
GO:2000742	Regulation of anterior head development	P	4.90E-04
GO:2000698	Positive regulation of epithelial cell differentiation involved in kidney development	P	4.90E-04
GO:0072050	S-shaped body morphogenesis	P	4.90E-04
GO:0072049	Comma-shaped body morphogenesis	P	4.90E-04
GO:0060059	Embryonic retina morphogenesis in camera-type eye	P	4.90E-04
GO:0090009	Primitive streak formation	P	4.90E-04
GO:0097477	Lateral motor column neuron migration	P	4.90E-04

GO:0097476	Spinal cord motor neuron migration	P	4.90E-04
GO:0097475	Motor neuron migration	P	4.90E-04
GO:0021937	Cerebellar Purkinje cell-granule cell precursor cell signaling involved in regulation of granule cell precursor cell proliferation	P	4.90E-04
GO:0021575	Hindbrain morphogenesis	P	5.50E-04
GO:0072006	Nephron development	P	5.50E-04
GO:2000177	Regulation of neural precursor cell proliferation	P	5.50E-04
GO:0060429	Epithelium development	P	5.70E-04
GO:0008283	Cell proliferation	P	7.30E-04
GO:0001705	Ectoderm formation	P	8.20E-04
GO:0021940	Positive regulation of cerebellar granule cell precursor proliferation	P	8.20E-04
GO:0021936	Regulation of cerebellar granule cell precursor proliferation	P	8.20E-04
GO:0002009	Morphogenesis of an epithelium	P	8.90E-04

4.3.4 Gene family expansion and contraction

Both gene family expansion and contraction events played a major role to achieve gene evolution (Chen et al., 2010). Here, I used CAFE to infer the direction of change in gene family size of *T. mercedesae* against those of other six reference arthropod species (Figure 4.4). The ultrametric species tree used in CAFE analyses was created as described in section 4.2.4, but the outgroup was excluded following the (CAFE) author's suggestion. A global parameter λ and several multi-parameter λ along species tree were tested for selecting the best model to fit this dataset (Table 4.3). More specific gene family changes of *T. mercedesae* in the Parasitiform lineage (Figure 4.4) was also investigated with the same procedure (Table 4.4).

Table 4.3 Model with different number of λ is along a row, and the other is along a column, and the results are shown where a row and a column meet (critical value in the blue triangle and LR value in the gray triangle).

	Model 1 (1λ)	**Model 2 (1λ)**	**Model 3 (1λ)**	**Model 4 (1λ)**	**Model 5 (1λ)**	**Model 6 (1λ)**	**Model 7 (1λ)**
Model 1 (1λ)	-	328.19	1357.21	1411.07	1417.77	1597.98	2355.98
Model 2 (1λ)	3.84	-	1029.02	1082.88	1089.59	1269.79	2027.79
Model 3 (1λ)	5.99	3.84	-	53.86	60.56	240.77	998.77
Model 4 (1λ)	7.81	5.99	3.84	-	6.70	186.91	944.91
Model 5 (1λ)	9.49	7.81	5.99	3.84	-	180.21	938.21
Model 6 (1λ)	12.59	9.49	7.81	5.99	3.84	-	758.00
Model 7 (1λ)	14.07	12.59	9.49	7.81	5.99	3.84	-

In the species tree of (((((*M. occidentalis*, *T. mercedesae*), *I. Scapularis*), *S. mimosarum*), *T. urticae*), (*A. Mellifera*, *D. melanogaster*)), Model 1 has a single global parameter λ across all branches of the tree, Model 2 has one λ value for the arachnid branches and the second λ value for all other branches, Model 3 has one λ value for *M. occidentalis*, *T. mercedesae*, *I. Scapularis*, and *S. mimosarum* branches, the second λ value for *T. urticae*, and the third λ value for insect branches, Model 4 has one λ value for the Parasitiform branches, the second λ value for *S. mimosarum*, the third λ value for *T. urticae*, and the fourth λ for insect branches, Model 5 has one λ value for *M. occidentalis* and *T. mercedesae*, the second λ value for *I. Scapularis*, the third λ value for *S. mimosarum*, the fourth λ value for *T. urticae*, and the fifth λ value for insect branches, Model 6 has one λ value for *M. occidentalis*, the second λ value for *T. mercedesae*, the third λ value for *I. Scapularis*, the fourth λ value for *S. mimosarum*, the fifth λ value for *T. urticae*, and the sixth λ value for insect branches, and Model 7 specifies a particular λ value for each branch.

Table 4.4 Model with different number of λ is along a row, and the other is along a column, and the results are shown where a row and a column meet (critical value in the blue triangle and LR value in the gray triangle).

	Model 1 (1λ)	Model 2 (1λ)	Model 3 (1λ)
Model 1 (1λ)	-	2369.14	2686.66
Model 2 (1λ)	3.84	-	317.52
Model 3 (1λ)	5.99	3.84	-

In the species tree of ((*M. occidentalis*, *T. mercedesae*), *I. Scapularis*), Model 1 has a single global parameter λ across all branches of the tree, Model 2 has one λ value for the arachnid branches, and the second λ value for all other branches, Model 3 has one λ value for *M. occidentalis*, *T. mercedesae*, *I. Scapularis*, and *S. mimosarum* branches, the second λ value for *T. urticae*, and the third λ value for insect branches.

1) Estimation of gene gain and loss events in *T. mercedesae* and other arthropod genomes

Table 4.3 shows that Model 7 is the best one so that this was used for CAFE (default setting) to compute the rate of gene gain and loss of in the given gene families. The total number of gene families that have experienced expansion, contraction, or no change was summarized along with each branch of the species tree in Figure 4.5. Among arthropod species analyzed, *T. mercedesae* has the fewest gene families undergone expansion or contraction (Figure 4.5). The conservation of gene family size is good indicator of the slow evolution rate of *T. mercedesae* genome. Because *T. mercedesae* appears to live under the simple and stable environment associated with bees irrespective of the host species (see section 4.1), its genome tends to contain few variable gene families (Hoffman and Parson 1991; Parson 2005).

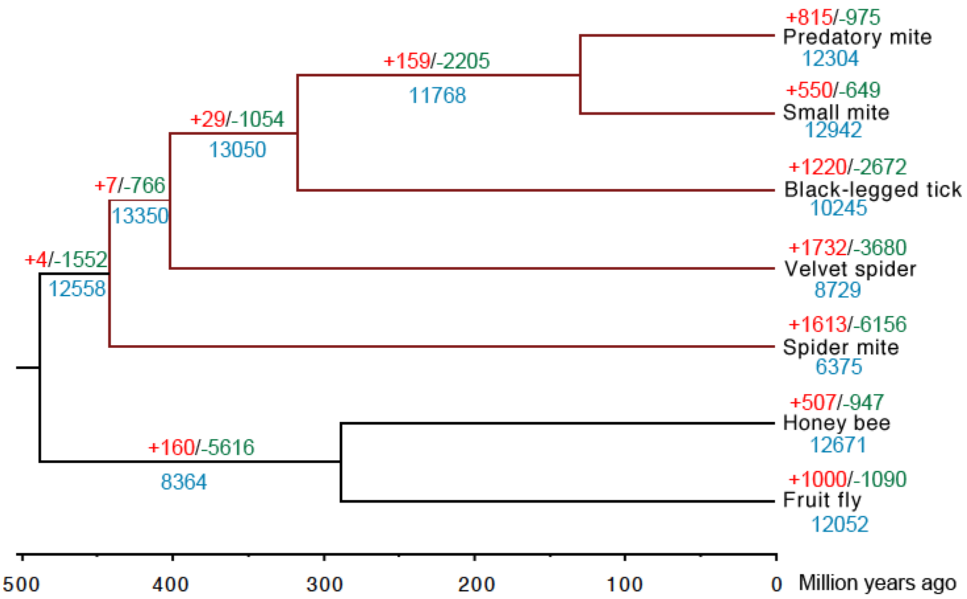


Figure 4.5 Gene family contraction and expansion in seven arthropod species. No outgroup was applied in Gene family analysis with CAFE. The red branches represent the Arachnids. The numbers of expanded, contacted, and stable gene families in each species and node are indicated by red, green, and blue, respectively.

For each family in the data file, CAFE computes probability (p-value) of observing the data given the average rate of gene gain and loss. P-value < 0.01 represents the significant change in gene family size. Branch-specific P-value < 0.001 indicates the large change either by contraction or expansion. According to these criteria, I found ten expanding and six contracting gene families in *T. mercedesae* compared to six other arthropod species (Table 4.5). The expanded gene families contain, for example, genes enconding DNA binding domains, play vital roles in DNA transcription, replication, packaging, repair, and rearrangement (Gao and Skolnick 2009). The most interesting contracted gene family of *T. mercedesae* is a heat shock 70 protein sub-family. As described in the section 2.3.2, heat shock 70 protiens function as stress resistance proteins.

Table 4.5 Changes of gene family size in *T. mercedesae* (Tm) in comparison with six reference arthropod species, *M. occidentalis* (Mo), *I. Scapularis* (Is), *S. mimosarum* (Sm), *T. urticae* (Tu), *A. Mellifera* (Am), and *D. melanogaster* (Dm). Gene familes in white and grey represent the ones expanded and contracted in *T. mercedesae*, respectively. Estimating changes in gene familiy size was performed by CAFE calculation with overall P-value < 0.01 and branch specific P-values < 0.001. Gene description was obtained by integration of top blast hit and GO and Pfam anotations.

Gene family	**Mo**	**Tm**	**Is**	**Sm**	**Tu**	**Am**	**Dm**	**P-value**	**Description**
ORTHOMCL158	0	21	0	0	0	0	0	0	Metal ion and DNA binding
ORTHOMCL261	0	16	0	0	0	0	0	0	Peptide cross-linking in cytoplasm
ORTHOMCL391	0	12	0	0	1	0	0	0	Metal ion, zinc ion and nucleic acid binding
ORTHOMCL467	0	12	0	0	0	0	0	0	Peptide cross-linking in cytoplasm; DNA binding in nucleus
ORTHOMCL571	0	10	1	0	0	0	0	0	Polyphosphate kinase: kinase activity for phosphorylation
ORTHOMCL711	0	10	0	0	0	0	0	0	Zinc finger protein: nucleic acid binding
ORTHOMCL177	2	10	6	2	0	0	0	8E-04	Reverse transcriptase : RNA-dependent DNA replication
ORTHOMCL980	1	8	0	0	0	0	0	3E-06	Zinc ion binding
ORTHOMCL2495	0	7	0	0	0	0	0	2.9E-05	Peptide cross-linking in cytoplasm
ORTHOMCL4818	0	6	0	0	0	0	0	6E-06	Unknown
ORTHOMCL8	120	0	1	1	0	0	0	0	Integrase and reverse transcriptase domains
ORTHOMCL24	39	1	1	27	0	0	0	5E-06	Integrase and reverse transcriptase domains
ORTHOMCL67	8	0	5	6	2	3	11	1E-04	Heat shock 70 kda protein cognate 4: response to heat stress
ORTHOMCL58	8	0	4	26	0	0	0	1E-04	DNA binding; DDE superfamily endonuclease
ORTHOMCL32	54	0	0	0	0	0	0	1E-04	Nucleic acid binding; protein dimerization activity; Ribonuclease H-like domain
ORTHOMCL18	70	2	2	0	0	0	0	1E-04	Integrase and reverse transcriptase domains

2) *T. mercedesae* against Parasitiforms

When I run CAFE pipeline to infer the gene family evolution in Parasitiforms, I found four expanded and nine contracted gene families in *T. mercedesae* (Table 4.6), compared to *M. occidentalis* and *I. Scapularis*. A number of orthology groups related to DNA, ion, and protein binding function are significantly expanded in *T. mercedesae* (Table 4.6). Based on phylogenetic analysis of *T. mercedesae* P450 genes (Figure 6.1), the contracted cytochrome P450 family (Table 4.6) could be identified as CYP 3 clan, which has important roles for xenobiotic metabolism (detailed in section 6.3). The contracted *T. mercedesae* P450 gene family includes four more genes than the Orthomcl constructed gene family because these four genes are short fragments (Table 6.1) that were repelled to be clustered when using blast-based scoring algorithm in orthomcl or were not identified by Maker based pipeline (but identified by manual annotation see section 6.3). The P450 gene family contraction may be the consequence of adapatoin of *T. mercedesae*s to simple and stable environment inside honey bee colony. Retinol dehydrogenase 12 is a primary enzyme responsible for the reduction of all-trans-retinal into retinol during recovery phase in the visual cycle. Gene family loss of retinol dehydrogenase 12 can cause vision loss or weakness (Janecke et al., 2004). But the other gene family (ORTHOMCL6204) including a *T. mercedesae* gene (Tm_05381) and a *M. occidentalis* gene (XP_003748476.1) was also annotated as Retinol dehydrogenase 12. This suggests *T. mercedesae* may still keep vision to sense light intensity and help parasitizing *A. Dorsata* open colony in which light still can penetrate. The *T. mercedesae* vision will be discussed later in the sections 8.3.

Table 4.6 Changes of gene family size in *T. mercedesae* (Tm) in comparison with two other Parasitiform, *M. occidentalis* (Mo) and *I. Scapularis* (Is). The gene families in white and grey represent the expanded and contracted ones, respectively. The gene families with extensive expansion or contraction compared to Table 4.5 are highlighted in blue. Estimating the changes in gene familiy size was performed by CAFE calculation with overall P-value < 0.01 and branch specific P-values < 0.001. Gene description was obtained by integration of top blast hit and GO and Pfam anotations.

Gene family	**Mo**	**Tm**	**Is**	**P-value**	**Description**
ORTHOMCL158	0	21	0	0	Metal ion and DNA binding
ORTHOMCL261	0	16	0	0	Peptide cross-linking in cytoplasm
ORTHOMCL391	0	12	0	0	Metal ion, zinc ion and nucleic acid binding
ORTHOMCL467	0	12	0	0	Peptide cross-linking in cytoplasm; DNA binding in nucleus
ORTHOMCL571	0	10	1	1E-05	Polyphosphate kinase: kinase activity for phosphorylation
ORTHOMCL711	0	10	0	1E-05	Zinc finger protein: nucleic acid binding
ORTHOMCL980	1	8	0	7E-04	Zinc ion binding
ORTHOMCL2495	0	7	0	1E-05	Multiple banded antigen
ORTHOMCL4818	0	6	0	1E-04	Unknown
ORTHOMCL7149	0	4	0	3E-4	Unknown
ORTHOMCL7160	0	4	0	3E-4	Ribonuclease
ORTHOMCL7161	0	4	0	3E-4	Zinc finger protein: nucleic acid binding
ORTHOMCL7163	0	4	0	3E-4	Serine-type endopeptidase activity for proteolysis
ORTHOMCL8	120	0	1	0	Integrase and reverse transcriptase domains
ORTHOMCL24	39	1	1	3E-05	Integrase and reverse transcriptase domains
ORTHOMCL67	8	0	5	5E-05	Heat shock 70 kda protein cognate 4: response to heat stress
ORTHOMCL58	8	0	4	9E-04	DNA binding; DDE superfamily endonuclease
ORTHOMCL32	54	0	0	0	Nucleic acid binding; protein dimerization activity; Ribonuclease H-like domain
ORTHOMCL18	70	2	2	0	Integrase and reverse transcriptase domains
ORTHOMCL40	44	0	0	3E-06	Telomere maintenance; DNA helicase activity; DNA repair
ORTHOMCL209	9	0	9	3E-06	DNA binding; DDE superfamily endonuclease

ORTHOMCL210	11	0	7	3E-06	Cytochrome P450
ORTHOMCL175	16	0	3	5E-05	ATP binding; telomere maintenance; DNA helicase activity; DNA repair
ORTHOMCL114	23	0	3	5E-05	Protein dimerization activity; nucleic acid binding
ORTHOMCL108	3	0	9	9E-04	Nuclease activity
ORTHOMCL476	5	0	6	9E-04	Retinol dehydrogenase 12-like; dehydrogenase
ORTHOMCL135	21	0	1	9E-04	Zinc finger domain
ORTHOMCL109	27	0	0	9E-04	Single-stranded DNA-binding heterotrimeric complex

4.3.5 Selection pressure

Darwin's theory of evolution suggests that, at the phenotypic level, traits in a population that enhance survival are known as positive selection, while traits that reduce fitness are known as negative selection. Selection pressure on particular gene has been assessed by several approaches, one of which relies on the mutations in a protein-coding DNA sequence. Mutations can involve almost any fraction of the genome; however, point mutations involving single nucleotide changes (substitutions) are the most common, and the other two mutations are deletion and insertion. However, for mutations involved in the coding region, the deletion and insertion often knock out the gene function. For example, deleting or inserting one base will cause frame shift in the downstream, which will often generate smaller protein without the functions. Hence, selection is usually involved in nucleotide substitution. A single nucleotide substitution affects only one base in the single codon that can be either synonymous or nonsynonymous. When a base substitution in the single codon does not result in the amino acid change, it is called synonymous substitution. This mutation is silent because the nucleotide substitution does not affect the amino acid sequence of protein, and thus traits of organism. In contrast, nonsynonymous substitution does change the amino acid sequence of protein, and may affect the traits of organism. Comparison of the rates of nonsynonymous substitution per nonsynonymous site (dn) versus synonymous substitution per synonymous site (ds) may reveal the evidence of either positive or negative selection. If dn is greater than ds, this implies that the amino acid substitutions are favored, suggesting the DNA sequence is under positive selection. Compared to the ancestral state, this gene could have gained a new function to, for example, to adapt to the new environmental under positive selection pressure. When ds is greater than dn, this implies that negative selection or purifying selection occurs in the DNA sequence. Negative selection occurs when there has been evolutionary pressure to limit change in the corresponding amino acid sequence to maintain a critical protein function. Thus, for faster evolving genes, the higher dn/ds ratios can be observed.

In order to understand the overall evolutionary state of *T. mercedesae*, I calculated dn/ds ratios for 1,865 one-to-one orthologs defined by Orthomcl in *T. mercedesae* and the six reference genomes, using codeml in the PAML package with the free-ratio model to estimate dn, ds, and calculate dn/ds ratios on the different branches. Genes with high dn/ds (>10) were discarded. It should be note that this test is aimed to generate the most accurate statistical value of dn/ds on each lineage, but is not a statistical test of significance of dn/ds > 1 when it is observed. Thus, the null model (one-ratio model) was not preformed. I found that the *T. mercedesae* genome has a significantly lower overall dn/ds ratio than other arthropods except *D. melanogaster* and *I. Scapularis* (mean omega values; *T. urticae*: 0.091; *D. melanogaster*: 0.026; *A. Mellifera*: 0.105; *S. mimosarum*: 0.098; *I. Scapularis*: 0.040; *M. occidentalis*: 0.087; *T. mercedesae*: 0.056) under paired Wilcoxon rank sum tests (P <0.0001; Figure 4.6; Table 4.7), indicating the relatively slow protein-coding gene evolution in *T. mercedesae*.

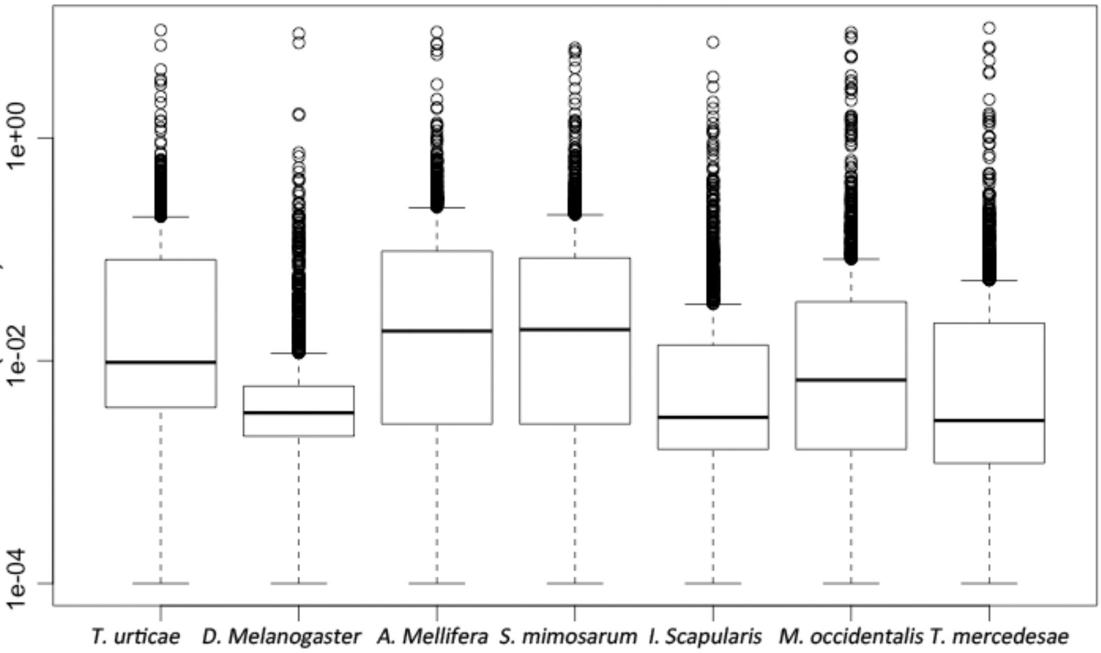


Figure 4.6 Comparison of the mean omega ratios of all single copy orthologous genes in seven srthropod species. The average omega ratio of *T. mercedesae* genome is significantly lower than other arthropods except *D. melanogaster* and *I. Scapularis* (paired Wilcoxon rank sum tests: P< 0.0001). Mean omega values are *T. urticae*: 0.091; *D. melanogaster*: 0.026; *A. Mellifera*: 0.105; *S. mimosarum*: 0.098; *I. Scapularis*: 0.040; *M. occidentalis*: 0.087; *T. mercedesae*: 0.056.

Table 4.7 P-value of paired Wilcoxon rank sum tests on the mean omega ratios of all single copy orthologous genes between *T. mercedesae* (dataset 1) and one of the reference genomes (dataset 2).

Dataset 1	**Dataset 2**	**P-value**
T. mercedesae	*T. urticae*	0
T. mercedesae	*D. melanogaster*	0.2022
T. mercedesae	*A. Mellifera*	0
T. mercedesae	*S. mimosarum*	0
T. mercedesae	*I. Scapularis*	0.0819
T. mercedesae	*M. occidentalis*	0

All of 91 genes with dn/ds > 1 in *T. mercedesae* that were further analyzed using the one ratio model (null model) to test the significance. The significant positively selected genes were listed Table 4.8, if they are under the criteria in section 4.2.7. All these genes in Tale 4.8 encode enzyme. Tm_10257 may play roles in development of *Drosophila* mushroom body (Nicolai et al., 2003) which is involved in olfactory learning and short-term memory (de Belle and Heisenberg, 1994; Davis, 1996; Heisenberg, 1998), short-term and long-term memory of courtship conditioning (mcbride et al., 1999), visual context generalization (Liu et al., 1999), and in long-term olfactory memory (Pascual and Preat, 2001). Tm_10257 is required for ovarian ring canal morphogenesis during *Drosophila* oogenesis (Dobson et al., 1998).

Table 4.8 Detailed information about four positively selected genes in *T. mercedesae*.

Orthomcl cluster	**Gene ID**	**Dn/ds Ratio**	**Putative function**
ORTHOMCL2609	Tm_01377	1.0194	Thioredoxin-related transmembrane protein 1-like; It has one thioredoxin-like domain with oxidoreductase activity, playing an important role in proper disulfide bond formation (Matsuo et al., 2001).
ORTHOMCL2604	Tm_10257	1.0212	Tyrosine-protein kinase src64b-like; May play a role in the development of *Drosophila* mushroom body (Nicolai et al., 2003) and oogenesis (Dobson et al., 1998).
ORTHOMCL4405	Tm_07523	3.9205	Cytosolic endo-beta-n-acetylglucosaminidase-like, involved in the processing of free oligosaccharides in the cytosol (Suzuki et al., 2002).
ORTHOMCL2644	Tm_07693	9.7219	Isocitrate dehydrogenase that catalyzes the oxidative decarboxylation of isocitrate, producing alpha-ketoglutarate (α-ketoglutarate) and CO_2.

5. Gene expression analysis with RNA-seq data

5.1 Introduction

High-throughput sequencing technology used for the DNA and RNA sequencing was introduced in the chapter 1. Since DNA-seq makes it possible to look at individual genome, RNA-seq is now the standard method for measuring RNA expression levels (Mortazavi et al., 2008). The expression level of each RNA can be measured by mapping millions of sequenced short reads to either reference genome or *de novo* assembled transcripts, whereby the number of mapped reads is expected to correlate directly with its abundance. The aim of mapping (or alignment) is to find the unique location where a short read is identical, or as similar as possible, to the reference sequence, however, in reality the reference is never perfect representation of the actual biological source of sequenced RNA due to the sample-specific attributes such as single nucleotide polymorphisms (snps) and insertions/deletions (indels). Therefore, the mapping procedure must be flexible enough to accommodate these attributes. Moreover, if the reference is genomic DNA, mapping procedure must adjust for splicing patterns as well. Ambiguity arises when the short reads align perfectly to multiple locations, and this should be also solved in the mapping procedure.

Most applications of RNA-seq are to identify the differentially expressed genes in two or more samples. The differential gene expression analysis generally consists of three components: normalization of counts, statistical modeling of gene expression and test for differential expression. Normalization is an essential step for the analysis of different expression from RNA-seq that enables accurate comparisons of expression levels between and within samples (Oshlack et al., 2010).

During the mapping procedure, longer transcripts can attract more fragments (reads) than shorter transcripts at the same expression level. To ensure that the expression levels for different transcripts can be compared within a sample, count for each transcript must be normalized with each transcript's length. The most commonly used normalization procedure is FPKM (Trapnell et al., 2010; fragments per kilobase of transcript per million mapped fragments), which is also known as RPKM (Nagalakshmi et al., 2010; reads per kilobase of transcript per million mapped reads). When comparing expression levels of an individual transcript between the samples, transcript length bias will be cancelled out because the transcript sequence used for mapping is the same between samples. However, sequencing runs between the different libraries may produce the different volumes of sequencing reads. Normalization by simply adjusting library sizes between the samples is, consequently, crucial for comparing expression levels of an individual transcript in the different libraries (Marioni et al., 2008). One method to normalize library size is the trimmed mean of M-values (TMM; Robinson and Oshlack 2010) that has been implemented in the edger (Robinson et al., 2010) Bioconductor package based on the hypothesis that most genes are not differentially expressed. TMM is computed as the weighted mean of log ratios between the test and the reference after excluding genes that exhibit high average read counts and genes that have large differences in expression.

Ideally, sequencing experiments are considered as random sampling of reads from a fixed pool of genes where the Poisson distribution can be used for modeling RNA-seq count data. However, in reality the variance of gene expression across multiple biological replicates is larger than its mean expression values (Rapaport et al., 2013). To address this over-dispersion problem for biological replicates, edger estimates an additional dispersion parameter in the negative binomial (NB) distribution model to account this biological and technical variability. As a result, Differential expression is assessed for each gene based on P-values generated by statistical modeling of NB distribution.

*T. mercedesae*s live in the honey bee colony for their entire lives; meanwhile, adult female mites exit the brood cell with adult honey bee and seek for new 5th instar larva in a phoretic period (1-2 days; see Background). Except female mites, male mites and nymph may not feed on honey bee larvae (see section 2.7.1). To understand the development of female nymph to adult and the sex differences with *T. mercedesae*, RNA was extracted from three different samples, male, female, and nymph (see section 1.2.1) for comparing the gene expression levels across and within these samples.

5.2 Materials and Methods

5.2.1 Gene expression analysis: comparison between libraries

Clean RNA-seq reads (see the section 2.3.3) were aligned to the *T. mercedesae* genome assembly using Tophat with default parameters except '-r 20' to adapt to the inner distance of transcriptome library. Tophat outputs the alignment map in BAM format. There will be reads that can be aligned to multiple locations (multi-reads) in the given reference genome. Htseq-count in the htseq Python package (version: 0.6.1; Anders et al., 2015) was performed on the BAM file to discard the reads ambiguously assigned to a gene, using the default union-counting mode with option '-a' to specify the minimum score for the alignment quality. The raw read count for each sample was then subject to further differential expression analysis using edger (version: 3.0) Bioconductor package. Expressed features without at least one count per million in one sample or replicates (low overall sum of counts) were removed, as this generally increases the statistical load of the differential expression analysis (Bourgon et al., 2010). Then, the library sizes of all samples were normalized according to the trimmed mean of M-values method.

In RNA-seq experiments, there is little use of technical replicates since the background is low and the variance can be better modeled (Marioni et al., 2008). However, the inherent biological variability needs to be eliminated using biological replicates. Thus, I estimated dispersion between the female and randomly collected sample (see section 1.2.1) using the quantile-adjusted conditional maximum likelihood method in the edger package, and then input the estimated value as fixed dispersion for analysis in the three unique samples without replicates (male, female, and nymph). After that, pairwise comparison of differential gene expression between above *T. mercedesae* samples was performed using the function of exacttest(), and the multiple testing of all samples was done with the functions of glmfit() and glmlrt(). As default setting, P-values were corrected using Benjamini-Hochberg correction (Benjamini and Hochberg, 2005) controlled false discovery rate (FDR). Because of the lack of biological replicates, I used very stringent threshold for significance (FDR < 0.01).

5.2.2 Gene expression analysis: comparison within a single library

The length of transcript is a crucial parameter for gene expression comparison within a single library. The expression values for each sample were calculated using RPKM calculation in edger Bioconductor package with rpkm() function with normalized library sizes.

5.2.3 Identification of over represented gene families in the differentially expressed genes

I identified gene families (defined by Orthomcl) in which the majority of members are over expressed in the pairwise comparisons between male and female and between nymph and female using one-tailed fisher's extract test. For the alternative/null hypothesis, I compared the proportion of differentially expressed genes in the family of interest among all genes in this family by building a 2X2 contingency table.

5.2.4 Differential expression of orphan genes

The proteins without any significant similarity to any proteins in NCBI non-redundant database (E-value < 1e-5) were thought as orphan genes. I identified such 1,624 orphan genes out of 11,060 expressed genes (at least one sample with RPKM>5; see the section 5.2.2). These orphan genes among the differentially expressed genes were discussed below.

5.2.5 Enrichment analysis

The same methods were used for gene and interpro domain enrichment analysis as described in the section 4.2.8.

5.3 Results and Discussion

Because the differentially expressed genes were identified in the three unique *T. mercedesae* samples, I estimated the dispersion from the female and randomly collected samples. Figure 5.1 shows the relationship between all samples, and the gene expression patterns of female and randomly collected mite samples coincide, suggesting that mites in the randomly collected sample are basically all female. The dispersion estimated from these two samples may represent the biological variability within a single *T. mercedesae* sample.

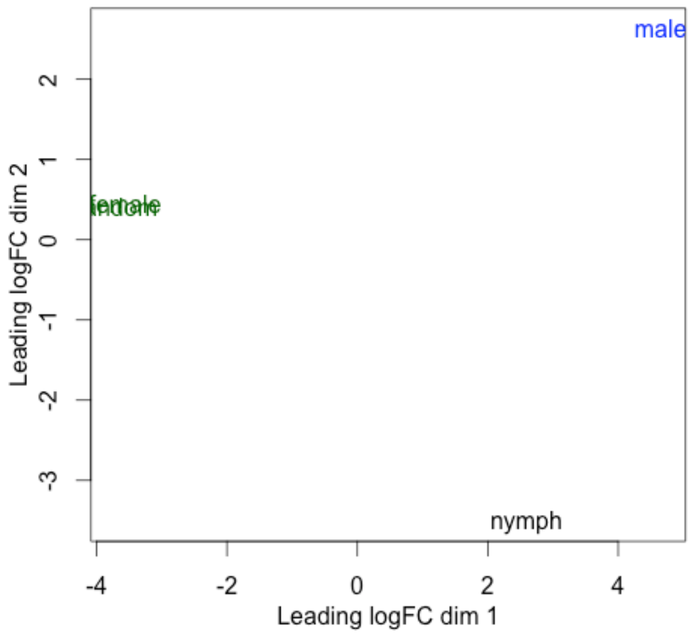


Figure 5.1 Relationship of the gene expression profiles between four mite samples. Using count-specific distance measure, edger's plotmds() function produces a multidimensional scaling plot showing the relationship between all samples, male mites (male), female mites (female), randomly collected mites (random), and mites at nymph stage (nymph).

5.3.1 Pair-wise comparison between male and female mites

Many animals exhibit sexually dimorphic morphology and behaviors that are linked to sexual reproduction (Kimura 2011). Differences of gene expression profiles between male and female *T. mercedesae* were analyzed using edger. I identified 867 and 279 genes highly expressed in female and male, respectively (FDR > 0.01; Table 5.1). Based on the 1,146 differentially expressed genes, I found 19 over represented gene families (P >

0.05, Table 5.2), and 11 are uncharacterized. Two gene families, vitellogenin and nanos (Becalska and Gavis, 2009), are shown to play important roles during oogenesis in *D. melanogaster*. Since both of them were enriched in the female transcriptome, they may be also important for the *T. mercedesae* oogenesis.

Genes up-regulated in the male trasncriptome include 425 orphan genes in which only 17 distinct interpro domains could be identified and enriched with GO terms of (FDR < 0.05) photoreceptor activity (GO:0009881), signaling receptor activity (GO:0038023), receptor activity (GO:0004872), molecular transducer activity (GO:0060089), signal transducer activity (GO:0004871), and ion binding (GO:0043167). The male biased genes are enriched with various enzyme activities to catalyze proteolysis and phosphorylation reactions (FDR > 0.01; Table 5.3).

Similar to male, 867 female biased genes contain a large number (101) of orphan genes. 16 distinct interpro domains could be identified, but no GO terms were significantly enriched. Except vitellogenin, innexin2 was previously shown to be essential for embryogenesis in *D. melanogaster* (Mukai et al., 2011), and thus it may function similarly in *T. mercedesae*. The female-biased genes are enriched with GO terms of tubulin binding, nucleic acid binding and expression regulation (P-value < 0.001, Table 5.4).

Table 5.1 Pair-wise comparisons of gene expression patterns between male and female, and between nymph and female.

		Male-female		**Nymph-female**	
		Male	**Female**	**Nymph**	**Female**
FDR < 0.01	**Differentiated gene number**	1146	1146	1047	1047
	Over expressed gene number (%)	867 (75.7%)	279 (24.3%)	692 (66.1%)	355 (33.9%)
	Down regulated gene number (%)	279 (24.3%)	867 (75.7%)	355 (33.9%)	692 (66.1%)
P-value < 0.05	**Differentiated gene number**	2160	2160	2548	2548
	Over-expressed gene number (%)	1489 (68.9%)	671 (31.1%)	1685 (66.1%)	863 (33.9%)
	Down regulated gene number (%)	671 (31.1%)	1489 (68.9%)	863 (33.9%)	1685 (66.1%)

Table 5.2 Gene families differentially expressed (DE) in the male and female samples (Fisher's extract test; P-value < 0.05).

Gene family	Gene family size	Number of DE proteins	Individuals with DE		P-value	Description
			Male	Female		
ORTHOMCL128	11	8	8	0	0.001	Pi-plc x domain-containing protein
ORTHOMCL146	10	6	6	0	0.018	Histone-lysine n-methyltransferase
ORTHOMCL162	9	5	5	0	0.045	Sodium-dependent phosphate transporter
ORTHOMCL257	6	5	0	5	0.004	Uncharacterized protein
ORTHOMCL259	6	4	0	4	0.035	Vitellogenin
ORTHOMCL264	6	4	4	0	0.035	N-acetyltransferase gcn5
ORTHOMCL336	5	4	4	0	0.014	Protein fam188b
ORTHOMCL338	5	5	4	1	0.001	Uncharacterized protein
ORTHOMCL442	4	4	0	4	0.004	Uncharacterized protein
ORTHOMCL443	4	3	3	0	0.048	Uncharacterized protein
ORTHOMCL454	4	3	0	3	0.048	Uncharacterized protein
ORTHOMCL456	4	4	4	0	0.004	Uncharacterized protein
ORTHOMCL684	3	3	3	0	0.015	Uncharacterized protein
ORTHOMCL686	3	3	3	0	0.015	Uncharacterized protein
ORTHOMCL691	3	3	3	0	0.015	Uncharacterized protein
ORTHOMCL694	3	3	3	0	0.015	Type 11 methyltransferase
ORTHOMCL701	3	3	0	3	0.015	Nanos-like protein
ORTHOMCL706	3	3	3	0	0.015	Uncharacterized protein
ORTHOMCL711	3	3	3	0	0.015	Uncharacterized protein

Table 5.3 GO functions enriched with the male biased genes. The corrected FDR P-value (< 0.05) was calculated by Fisher's extract test embedded in the blast2go software desktop version.

GO Term	Description	Type	FDR
GO:0004713	Protein tyrosine kinase activity	F	1.60E-04
GO:0004672	Protein kinase activity	F	4.90E-04
GO:0016773	Phosphotransferase activity, alcohol group as acceptor	F	1.90E-03
GO:0008233	Peptidase activity	F	1.90E-03
GO:0004252	Serine-type endopeptidase activity	F	6.50E-03
GO:0006468	Protein phosphorylation	P	8.40E-03
GO:0005524	ATP binding	F	9.90E-03
GO:0004175	Endopeptidase activity	F	1.10E-02
GO:0032559	Adenyl ribonucleotide binding	F	1.10E-02
GO:0030554	Adenyl nucleotide binding	F	1.30E-02
GO:0016301	Kinase activity	F	1.40E-02
GO:0005315	Inorganic phosphate transmembrane transporter activity	F	1.40E-02
GO:0017171	Serine hydrolase activity	F	1.40E-02
GO:0008236	Serine-type peptidase activity	F	1.40E-02
GO:0070011	Peptidase activity, acting on L-amino acid peptides	F	1.40E-02
GO:0006508	Proteolysis	P	1.70E-02
GO:0006817	Phosphate ion transport	P	3.00E-02
GO:1901677	Phosphate transmembrane transporter activity	F	3.00E-02
GO:0016787	Hydrolase activity	F	3.40E-02
GO:0032555	Purine ribonucleotide binding	F	3.90E-02
GO:0019538	Protein metabolic process	P	3.90E-02
GO:0032553	Ribonucleotide binding	F	4.60E-02
GO:0017076	Purine nucleotide binding	F	4.60E-02
GO:0035639	Purine ribonucleoside triphosphate binding	F	4.60E-02

Table 5.4 GO functions enriched with the female biased genes. The corrected P-value (< 0.001) was calculated by Fisher's extract test embedded in the blast2go software desktop version.

GO Term	Description	Type	P-value
GO:0010608	Posttranscriptional regulation of gene expression	P	4.60E-05
GO:0005488	Binding	F	4.70E-05
GO:0006417	Regulation of translation	P	2.10E-04
GO:0003723	RNA binding	F	2.20E-04
GO:0015631	Tubulin binding	F	3.00E-04
GO:0045298	Tubulin complex	C	4.10E-04
GO:0005179	Hormone activity	F	4.80E-04
GO:0010468	Regulation of gene expression	P	7.80E-04
GO:0008135	Translation factor activity, nucleic acid binding	F	8.10E-04
GO:0008017	Microtubule binding	F	9.90E-04

5.3.2 Pair-wise comparison between nymph and female

Differences in gene expression profiles between *T. mercedesae* nymph and female represent the developmental progression. Of the 1,047 differentially expressed genes, 692 (66.1%) and 355 (33.9%) were significantly overexpressed in nymph and adult female, respectively (according to edger; FDR < 0.01; Table 5.1). Based on the 1,146 differentially expressed genes, I identified 34 over represented gene families (P < 0.05; Table 5.5), and 23 of them are uncharacterized. Three over represented gene families are related to cuticular protein formation. The protocadherin fat genes are essential for growth, planar cell polarity (PCP) and proximodistal patterning during Drosophila development (Bosch et al., 2014). All members of protocadherin gene family were highly expressed in nymph, suggesting that the mite orthologs may perform similar functions as in *Drosophila*. Vitellogenin and Nanos were also found as over represented gene families highly expressed in adult females when compared to nymphs. Interestingly, chitinase gene family was over expressed in the nymph transcriptome, suggesting that these chitinases genes are necessary to reshape their own chitin (Sámi et al., 2001).

Genes up regulated in the nymph trasncriptome included 212 orphan genes in which only 10 distinct interpro domains could be identified and enriched with GO terms of photoreceptor activity (GO:0009881), signaling receptor activity (GO:0038023), receptor activity (GO:0004872), molecular transducer activity (GO:0060089), and signal transducer activity (GO:0004871) (FDR < 0.05). The enrichment of nymph biased genes suggests that nymph highly expresses genes associated with the molting (FDR < 0.05; Table 5.6).

By the pair-wise comparison, 355 female biased genes contain 131 orphan genes in which 19 distinct interpro domains could be identified and enriched with GO terms of: inositol 1,4,5 trisphosphate binding (GO:0070679), alcohol binding (GO:0043178), photoreceptor activity (GO:0009881), and small molecule binding (GO:0036094) (FDR < 0.05). The female-biased genes are enriched with GO categories of the tubulin binding and nucleic acid binding and expression regulation (P-value < 0.001; Table 5.7).

Table 5.5 Gene families differentially expressed (DE) in the nymph and female samples (Fisher's extract test; P-value < 0.05).

Gene family	Gene family size	Number of DE proteins	Individuals with DE		P-value	Description
			Female	Nymph		
ORTHOMCL2	125	44	8	36	0.000	Cuticle protein
ORTHOMCL26	29	11	0	11	0.022	Protocadherin fat
ORTHOMCL38	24	10	1	9	0.014	Chitin binding domain contain protein
ORTHOMCL57	19	11	5	6	0.000	Uncharacterized protein
ORTHOMCL67	17	8	4	4	0.012	Uncharacterized protein
ORTHOMCL72	16	7	0	7	0.030	Chitinase
ORTHOMCL85	14	8	0	8	0.003	Uncharacterized protein
ORTHOMCL106	12	8	0	8	0.001	Uncharacterized protein
ORTHOMCL112	12	7	6	1	0.004	Importin alpha
ORTHOMCL146	10	5	0	5	0.036	Histone-lysine n-methyltransferase
ORTHOMCL214	7	4	0	4	0.036	Uncharacterized protein
ORTHOMCL257	6	5	5	0	0.002	Uncharacterized protein
ORTHOMCL259	6	5	5	0	0.002	Vitellogenin
ORTHOMCL320	5	5	0	5	0.000	Uncharacterized protein
ORTHOMCL442	4	4	4	0	0.002	Uncharacterized protein
ORTHOMCL445	4	3	2	1	0.029	Uncharacterized protein
ORTHOMCL454	4	3	3	0	0.029	Uncharacterized protein
ORTHOMCL456	4	3	0	3	0.029	Uncharacterized protein
ORTHOMCL660	3	3	0	3	0.009	Uncharacterized protein
ORTHOMCL691	3	3	0	3	0.009	Uncharacterized protein
ORTHOMCL701	3	3	3	0	0.009	Nanos protein
ORTHOMCL702	3	3	3	0	0.009	Uncharacterized protein

ORTHOMCL706	3	3	0	3	0.009	Uncharacterized protein
ORTHOMCL760	2	2	0	2	0.042	Ganglioside gm2 activator
ORTHOMCL883	2	2	0	2	0.042	Uncharacterized protein
ORTHOMCL1057	2	2	2	0	0.042	Uncharacterized protein
ORTHOMCL1082	2	2	0	2	0.042	Zinc finger protein 512b
ORTHOMCL1112	2	2	0	2	0.042	Uncharacterized protein
ORTHOMCL1158	2	2	0	2	0.042	Uncharacterized protein
ORTHOMCL1160	2	2	2	0	0.042	Uncharacterized protein
ORTHOMCL1173	2	2	2	0	0.042	Protein lsm14
ORTHOMCL1177	2	2	2	0	0.042	Uncharacterized protein
ORTHOMCL1182	2	2	2	0	0.042	Uncharacterized protein
ORTHOMCL1189	2	2	0	2	0.042	Uncharacterized protein

Table 5.6 The enriched GO terms in genes highly expressed in nymph. The corrected FDR P-value (< 0.05) was calculated by Fisher's extract test embedded in the blast2go software desktop version.

GO Term	**Description**	**Type**	**FDR**
GO:0042302	Structural constituent of cuticle	F	4.30E-21
GO:0006030	Chitin metabolic process	P	4.30E-09
GO:1901071	Glucosamine-containing compound metabolic process	P	4.30E-09
GO:0008061	Chitin binding	F	4.30E-09
GO:0005198	Structural molecule activity	F	1.40E-08
GO:0006040	Amino sugar metabolic process	P	2.00E-08
GO:0006022	Aminoglycan metabolic process	P	3.80E-07
GO:0007156	Homophilic cell adhesion	P	5.40E-06
GO:0008233	Peptidase activity	F	3.20E-03
GO:0004252	Serine-type endopeptidase activity	F	3.20E-03
GO:0007469	Antennal development	P	4.10E-03
GO:0004175	Endopeptidase activity	F	1.10E-02
GO:0070011	Peptidase activity, acting on L-amino acid peptides	F	1.10E-02
GO:0030048	Actin filament-based movement	P	1.10E-02
GO:0017171	Serine hydrolase activity	F	1.10E-02
GO:0008236	Serine-type peptidase activity	F	1.10E-02
GO:0005509	Calcium ion binding	F	1.20E-02
GO:0016998	Cell wall macromolecule catabolic process	P	1.40E-02
GO:0016337	Cell-cell adhesion	P	1.70E-02
GO:0048800	Antennal morphogenesis	P	2.00E-02
GO:0008407	Chaeta morphogenesis	P	2.00E-02
GO:0006032	Chitin catabolic process	P	2.70E-02
GO:0004568	Chitinase activity	F	2.70E-02
GO:1901072	Glucosamine-containing compound catabolic process	P	3.50E-02

GO:0035214	Eye-antennal disc development	P	3.50E-02
GO:0046348	Amino sugar catabolic process	P	3.50E-02
GO:0005576	Extracellular region	C	3.70E-02
GO:0004104	Cholinesterase activity	F	4.90E-02

Table 5.7 The enriched GO terms in genes highly expressed in adult female. The corrected P-value (< 0.001) was calculated by Fisher's extract test embedded in the blast2go software desktop version.

GO Term	**Description**	**Type**	**P-value**
GO:0003723	RNA binding	F	8.70E-06
GO:0015631	Tubulin binding	F	9.70E-05
GO:0006417	Regulation of translation	P	1.20E-04
GO:0045298	Tubulin complex	C	1.40E-04
GO:0010608	Posttranscriptional regulation of gene expression	P	1.90E-04
GO:0051028	Mrna transport	P	2.00E-04
GO:0005319	Lipid transporter activity	F	2.40E-04
GO:0006446	Regulation of translational initiation	P	5.60E-04
GO:0031332	Rnai effector complex	C	7.50E-04
GO:0016442	RISC complex	C	7.50E-04
GO:0051236	Establishment of RNA localization	P	8.20E-04
GO:0050658	RNA transport	P	8.20E-04
GO:0050657	Nucleic acid transport	P	8.20E-04

5.3.3 Multiple comparisons of genes differentially expressed among male, female, and nymph

A total of 654 orphan genes were classified as differentially expressed (FDR < 0.01) in at least one of the three samples, male, female, and nymph. When the threshold was lowered to P-value < 0.05, 828 orphan genes could be classified as differentially expressed that account for about half of the orphan genes in *T. mercedesae*. This may suggest that many of young lineage-specific genes of *T. mercedesae* are associated with either development or sexual dimorphism.

Vitellogenin, one of glycolipoproteins, is expressed in female of nearly all oviparous species (Robinson 2008). It is a precursor of yolk protein, vitellin, critical for the yolk formation in oocyte and also a large lipid transfer protein involved in the assembly, secretion, and metabolism of lipoproteins (Babin et al., 1999). All vitellogenin genes were highly expressed in female compared to male and nymph. Two vitellogenin genes were expressed only in female, and three genes, especially Tm_03860, were also expressed in male and nymph, suggesting that they could be linked to lipid transport (Table 5.8).

Table 5.8 Normalized read counts for vitellogenin genes in male, female, and nymph with RPKM methods as described in the section 5.2.2.

Gene ID	**Male**	**Female**	**Nymph**
Tm_14630	0	5.69	0
Tm_03860	40.95	230.16	6.37
Tm_05542	1.63	22.74	0.70
Tm_10308	0.06	42.18	0.11
Tm_12804	0	54.05	0

6. Detoxification system in *T. mercedesae*

6.1 Introduction

Xenobiotic is an exogenous chemical substance to which an organism is exposed that is not naturally produced by or expected to be present within that organism (Nagarkatti and Nagarkatti 1987). At the biochemical level, resistance to xenobiotics typically involves decreases in target site sensitivity due to the point mutations as well as increases in the capabilities of detoxificative enzymes to metabolize the xenobiotic before it reaches to the target site (Li et al., 2007). Metabolic resistance usually occurs more common than target site mutations. Without the metabolism, many xenobiotics would accumulate to the toxic levels. Therefore, metabolism of xenobiotics is essential for an organism to survive under environments that contain toxic substances including both naturally occurring and man-made chemicals. The molecular aspects of metabolic resistance are not yet well characterized in mite but, in insect, three major groups of enzyme play the major roles in detoxifying the environmental chemicals (Li et al., 2007): cytochrome P450s (P450s), glutathione-S-transferases (GSTs), and Carboxyl/cholinesterases (CCEs).

The overall purpose of metabolizing xenobiotics is to increase their water solubility (polarity) so that to excrete them from the body. This conversion is called biotransformation which are normally associated with two phases; phase I - functionalisation reactions are thought to act as preparation of the xenobiotics for the phase II - biosynthetic (conjugation) reactions (Gibson and Skett 2001; Ionescu and Caira 2005). In the phase I reactions, chemically reactive functional groups are introduced on the parent compound by oxidation, reduction, hydrolysis, hydration reactions or other rarer miscellaneous reactions. Often, the resulting biologically active metabolites are more polar than the parent drug, and have the reduced ability to penetrate tissues and less adsorption by renal tubules than the parent drug. In the phase II reactions, conjugation reactions occur by the formation of covalent linkage between a functional group on the primary metabolite in phase I (or parent compound) and endogenously derived glucuronic acid, sulphate, glutathione, amino acid, or acetate. These very hydrophilic (polar) conjugates are generally inactive, and subsequently are much more easily to be excreted. However, very hydrophobic xenobiotics may persist in adipose tissue almost indefinitely if they were not converted to more polar forms.

Many of the various synthetic chemical acaricides used to control *Varroa* mites are also effective against *Tropilaelaps* mites (Pichai, et al., 2008). Sulphur, formic acid, and thymol have also proved satisfactory (Atwal and Goyal, 1971; Raffique et al., 2012). However, the emergence of drug-resistant *Varroa* mites has been reviewed (Van Leeuwen et al., 2010). Indeed, bee mites are able to develop resistance to these compounds rapidly due to their short life cycle. Similar to insects, enhanced enzymatic detoxification and target site insensitivity are also considered as the major causes of development of xenobiotics resistance in Acari (Van Leeuwen et al., 2010). *A. mellifera*, the host of *T. mercedesae*, has far fewer genes associated with xenobiotic metabolism (The Honey bee Genome Sequencing Consortium, 2006). The reduction in the *A. mellifera* may be related to its specialized eusocial behavior and homeostasis of the nest environment (Werren et al., 2010). *T. mercedesae* spends whole life inside the bee colony, and so the xenobiotics they exposed may be different and fewer than the open living organisms. In this chapter, the genes associated with xenobiotic metabolism in *T. mercedesae* would be explored based on discussion of P450s, GSTs, and CCEs.

6.2 Materials and Methods

The gene families of P450s, GSTs and CCEs in *T. mercedesae* were manually annotated. Phylogenetic analyses were conducted as described in the section 4.2.4. The constructed phylogenetic trees were viewed and graphically edited with figtree (version: 1.4.2; tree.bio.ed.ac.uk).

6.3 Cytochrome P450s (P450s)

Cytochrome P450 is a very large and diverse group of enzymes that can be found in all domains of life. P450 is best known for the monooxygenase role (a phase I reaction) involved in metabolism of a wide range of endogenous and exogenous chemical substances such as hormones, pheromones, pesticides, and plant secondary compounds (Scott 1999; Feyereisen 1999; Iga and Kataoka 2012). Insect P450s have been subdivided into four major clans: CYP2, CYP3, CYP4, and mitochondrial P450s (Nelson 1998). Based on the degree of amino acid sequence similarity, each clade has been further subdivided into various CYP families (Scott 1999).

The putative P450s were identified based on the Interproscan domain annotation first. All proteins of *T. mercedesae* and *M. occidentalis* assigned with the intoerpro entry for P450 (IPR001128) were considered as P450 candidates. Then, all candidate P450s of *T. mercedesae* and *M. occidentalis* together with insect (*Anopheles gambiae*, *D. melanogaster*, *D. Pseudoobscura,* and *B. mori)* P450 protein sequences downloaded from the Cytochrome P450 Homepage (http://drnelson.uthsc.edu/cytochromeP450.html; Nelson 2009) were used as queries to manually search for putative *T. mercedesae* P450 (TmP450) gene models (E-value cutoff = 1e-3) by tblastn. TmP450 gene models were confirmed or refined when the transcript sequences were available in the Trinity *de novo* assembly based on combining reads from all samples (male, female, nymph, and randomly collected mites) as inputs. As a result, I manually annotated 56 TmP450s. There are 4 genes contain one or more stop codons in the open reading frame, and 7 genes are just short fragments (length < 200 amino acids), and one gene has an internal deletion (Table 6.1). In addition, none of these gene models were supported by RNA-seq data. Thus, they are more likely to be pseudogenes. Among the remaining 44 TmP450 genes, 23 have at least one complete P450 domain, and most of other 21 genes have the incomplete N-terminal amino acid sequences. The average length of TmP450s with intact P450 domain is about 500 amino acids, which is comparable to the average gene size in insects (Ai et al., 2011).

All TmP450s, including the short gene fragments and pseudogenes, were aligned with *M. occidentalis* P450s and *D. melanogaster* P450s by Kalign using default setting.

Poorly aligned N-terminal, C-terminal, and internal regions with variable length were excluded for phylogenetic analysis. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because removing these regions can potentially mislead relationships between subfamilies. They may provide important information for relationships within subfamilies. Based on the trimmed alignment, a PhyML tree was constructed using the ProtTest suggested best-fit substitution model of LG with invariant sites (I) and gamma (G) distributed rates. Here, bootstrap analysis with 100 repetitions was performed to assess the significance of phylogenetic clustering. The phylogenetic tree resolved the four major clans found in insects, which include CYP2, CYP3, CYP4, and mitochondrial clans, encompassed all P450s in *T. mercedesae* (Figure 6.1). *T. mercedesae* CYP clans were classified based on *D. melanogaster* P450s, but only three TmP450 genes (Tm11277, Tm11316, and Tm10252) could be clearly assigned as *D. melanogaster* P450 orthologs (Figure 6.1).

In mammals, gene products in CYP2 clan are generally specialized for the metabolism of exogenous compounds (Nebert and Russell 2002). Although the functions of many insect CYP2 clan P450s are still unknown, some are involved in synthesis of ecdysone (Baldwin et al., 2010) which is the major steroid hormone in insects and plays essential roles for coordinating developmental transitions such as larval molting and metamorphosis through its active metabolite 20-hydroxyecdysone (Truman and Riddiford 2002). Tm11277 gene is orthologous to *D. melanogaster* CYP307A1 and CYP307A2 in the CYP 2 clan (Figure 6.1). *Drosophila* Halloween genes (i.e. A set of genes identified in *Drosophila* that influence embryonic development), CYP307A1 and its paralog CYP307A2, play essential roles for the early steps of ecdysone synthesis from cholesterol (Ono et al., 2006; Rewitz et al., 2007). Although it was proposed these genes have the equivalent biochemical roles, *Drosophila* seems to use these two CYP307A genes in the distinct tissues at different stage during development. *Drosophila* CY307A1 is expressed only in the embryonic amnioserosa and in the adult ovary, whereas CYP307A2 is expressed in the prothoracic cells of the ring gland from late embryogenesis to late larval stages (Ono et al., 2006; Rewitz et al., 2007). Unlike closely related *M. occidentalis, T. mercedesae* contains only one CYP307 gene (Tm11277) similar to *Daphnia pulex* (Baldwin et al., 2009). Tm11277 gene may have similar functions as *Drosophila* CYP307 gene; however, there are no significant differences of the gene expression among male, female, and nymph (Table 6.3; Figure 6.2). Three other *T. mercedesae* CYP2 genes (Tm0064, Tm1694 and Tm1596) are only clustered with *M. occidentalis* genes, suggesting that they may be specific for Parasitiform mites. Tm00041 is highly expressed in the nymph, and thus it could be important for the early development of *T. mercedesae* (Table 6.3; Figure 6.2). Interestingly, CYP2 genes in *M. occidentalis* and *T. urticae* were largely expanded relative to either *T. mercedesae* or *D. melanogaster* (Table 6.2).

Members in CYP3 clan are involved in metabolizing both endogenous and exogenous compounds in arthropods (Baldwin et al., 2009). CYP6 gene family in CYP3 clan is considered to be essential for a wide range of insecticide resistance since it is the most abundant among insect P450 genes (Feyereisen 2012). Figure 6.1 shows that all mite P450 genes in this group were phylogenetically separated from *D. melanogaster* P450s. Thus, P450 genes related to pesticides or insecticides metabolism must have evolved differently in Parasitiformes and insects. CYP3 clan of *T. mercedesae* contains 19 genes but seven genes including five short fragment genes, one gene with internal deletion, and one gene with internal stop codon were not expressed (Table 6.1). In addition, *T. mercedesae* CYP3 clan members have contracted compared to those in *M. occidentalis* and *I. Scapularis* (see section 4.3.4; Figure 6.1). Thus, *T. mercedesae* has lost some ancestral CYP3 clan genes in Parasitiform.

Amino acid sequences of CYP4 clan members are highly diverse, and some genes are clearly inducible by xenobiotics (Feyereisen 2012). However, members of the CYP4 clan are also known to be associated with hormone metabolism (Bradfield et al., 1991; Davies et al., 2006) and sensory perception in insects as they are found in the antennae (Ono et al., 2005). Similar to the CYP3 clan; all *T. mercedesae* P450 genes in this group are phylogenetically separated from *D. melanogaster* P450s. In the CYP4 clan, six TmP450 genes were pseudogenes, and the expression of complete Tm_01048 gene was not detected either (Table 6.1). Three male and female *T. mercedesae* biased genes (Tm06697 Tm01436 Tm02528) may preform essential functions in adults (Table 6.3; Figure 6.2).

Previous studies have shown that mitochondrial P450s have essential physiological functions such as the *Drosophila* Halloween genes (CYP302A1, CYP315A1, and CYP314A1) that produce the pulses of cyclic ecdysteroid to trigger moulting (Rewitz and Gilbert 2008). The recent study has also found that CYP301A1 is involved in the formation of adult cuticle and ecdysone regulation in *D. melanogaster* (Sztal et al., 2012). Insect mitochondrial P450 genes associating with xenobiotic metabolism has also been reported (Ai et al., 2011). The mitochondrial P450s of vertebrates and insects (as well as other Metazoa; Feyereisen, 2011) are monophyletic. This monophyletic character of mitochondrial P450s is also shared with *T. mercedesae* and *M. occidentalis* (Figure 6.1).

Feyereisen (2011) proposed that main driver of P450s evolution is also due to gene duplication. The duplicated copy can be defined as tandem if it is adjacent to the original copy or as dispersed if it is scattered somewhere else in the genome. As shown in Table 6.1, 56 TmP450s are located on 52 scaffolds in which 44 of them are found to be individually located on the single scaffold, while the rest of 12 TmP450s (about 25%) are on six scaffolds with each containing at least two TmP450s (Table 6.1). These 12 TmP450 genes are tandem duplicates in five clusters and classified to CYP 3 and CYP 4 clans (Table 6.1). In contrast, a total of 32 genes from 14 clusters are tandem duplicates (approximately 43% of 75 *M. occidentalis* P450s), and 21 of them are classified to CYP 3 and CYP 4 clans. *M. occidentalis* possesses more P450 tandem duplicates than *T. mercedesae* probably because they have more chances to be exposed to xenobiotics than honey bee nest protected *T. mercedesae.* As discussed in the section 2.3.2, the duplicated gene may lose gene function by nonfunctionalization, evolve a new beneficial function by neofunctionalization, or stably be maintained as daughter copy to take a part of the ancestral gene functions by subfunctionalization. The dispersed duplicated copies could also evolve novel functions by recruiting new regulatory elements (e.g., Wang et al. 2002). However, the tandemly duplicated genes are likely to share the same regulatory elements (e.g., Li et al. 2006; Ponce and Hartl 2006; Arisue et al. 2007). In addition, CYP 3 and CYP 4 clans are thought to be important for xenobiotic metabolism (Oakeshott et al., 2010). Therefore, these tandemly duplicated TmP450 genes in CYP 3 and CYP 4 clans may enhance the P450 gene functions for detoxification after they were fixed on the genome during evolution. But, Tm_03629 in the tandem cluster 5 and Tm_15195 in the tandem cluster 6 become pseudogenes due to the internal deletion or internal stop codon. Furthermore, the gene fragments of Tm_15004 in the cluster 4 and Tm_03924 are also thought unexpressed pseudogenes (Table 6.1).

Reviewing the numbers of genes in each clan in different species, they vary considerably among arthropod species (Table 6.2). Copy number variations of gene are emerging as an important factor in genome evolution, and these are targets of selection (Emerson et al., 2008). All of TmP450 genes can be tracked at least one ortholog in *M. occidentalis* genome (Figure 6.1). Although expression of *M. occidentalis* P450 genes was not characterized, *M. occidentalis* has 19 P450 genes more than *T. mercedesae*, and only 6 genes encode proteins shoter than 200 amino acids. Table 6.1 shows that a large number (18) of TmP450 genes have lost their protein-coding capacity due to the deletion or internal stop codon. The losses of TmP450 genes may suggest this is the result of adaptation of *T. mercedesae* to honey bee colony environment. It is undeniable that the NGS technologies have limitation on gene annotation. Cytochrome P450 genes in Parasitiformes are still not well studied. These results can provide basis for further genetic research on the biological functions of P450s in Parasitiformes, particularly in relation to miticide resistance.

Table 6.1 Properties of *T. mercedesae* P450 genes classified to four Clans. If a P450 gene contains at least one count per million in one RNA-seq sample or the replicates, the gene is considered to be expressed and shown with 'yes'. Tandemly duplicated cytochrome P450 genes are not supposed to be separated by other genes. In this Table, 'fragment' indicates a P450 gene coding the protein shorter than 200 amino acids (aa), and 'complete' means the corresponding P450 gene has an intact P450 domain. The unexpressed fragmented *T. mercedesae* P450 genes, or *T. mercedesae* P450 genes with internal deletion or stop codon are considered as pseudogenes which are highlighted with gray background.

Scaffold ID	**Tandem cluster**	**Gene ID**	**CYP clans**	**Express.**	**Length (aa)**	**Note**
Scaffold0000000258	-	Tm_00041	CYP2 clan	Yes	612	Complete
Scaffold0000001158	-	Tm_15196	CYP2 clan	No	62	Fragment
Scaffold0000003393	-	Tm_15193	CYP2 clan	No	73	Internal stop codon
Scaffold0000005240	-	Tm_00664	CYP2 clan	Yes	464	Complete
Scaffold0000005442	-	Tm_08470	CYP2 clan	Yes	607	Complete
Scaffold0000006696	-	Tm_15194	CYP2 clan	No	117	Internal stop codon
Scaffold0000015432	-	Tm_11277	CYP2 clan	Yes	489	Complete
Scaffold0000008201	Cluster 2	Tm_03434	CYP3 clan	Yes	547	Complete
Scaffold0000008201	Cluster 2	Tm_09150	CYP3 clan	Yes	550	Complete
Scaffold0000008201	Cluster 2	Tm_15190	CYP3 clan	Yes	324	Complete
Scaffold0000010685	Cluster 4	Tm_15004	CYP3 clan	No	118	Incompleted N-terminal and C-terminal; fragment
Scaffold0000010685	Cluster 4	Tm_10021	CYP3 clan	Yes	532	Complete
Scaffold0000010685	Cluster 4	Tm_10022	CYP3 clan	Yes	284	Incompleted N-terminal
Scaffold0000010339	Cluster 5	Tm_09891	CYP3 clan	Yes	513	Complete
Scaffold0000010339	Cluster 5	Tm_03629	CYP3 clan	No	157	Internal deleation
Scaffold0000013932	Cluster 6	Tm_03924	CYP3 clan	No	106	Incompleted N-terminal; fragment
Scaffold0000013932	Cluster 6	Tm_15195	CYP3 clan	No	145	Internal stop codon
Scaffold0000002746	-	Tm_07004	CYP3 clan	Yes	494	Complete

Scaffold0000004790	-	Tm_07853	CYP3 clan	Yes	239	Incompleted C-terminal
Scaffold0000011341	-	Tm_01372	CYP3 clan	Yes	501	Complete
Scaffold0000015709	-	Tm_12854	CYP3 clan	Yes	273	Incompleted N-terminal
Scaffold0000018584	-	Tm_14705	CYP3 clan	Yes	139	Incompleted N-terminal and C-terminal; fragment
Scaffold0000020747	-	Tm_14091	CYP3 clan	Yes	542	Complete
Scaffold0000025786	-	Tm_04218	CYP3 clan	No	206	Incompleted N-terminal
Scaffold0000027228	-	Tm_08589	CYP3 clan	No	101	Incompleted N-terminal; fragment
Scaffold0000027778	-	Tm_15044	CYP3 clan	No	103	Incompleted N-terminal and C-terminal; fragment
Scaffold0000000842	Cluster 1	Tm_02528	CYP4 clan	Yes	356	Incompleted N-terminal
Scaffold0000000842	Cluster 1	Tm_02530	CYP4 clan	Yes	405	Incompleted C-terminal
Scaffold0000010068	Cluster 3	Tm_01157	CYP4 clan	Yes	276	Incompleted N-terminal
Scaffold0000010068	Cluster 3	Tm_09780	CYP4 clan	No	64	Incompleted C-terminal; fragment
Scaffold0000000108	-	Tm_05131	CYP4 clan	Yes	523	Complete
Scaffold0000001225	-	Tm_05802	CYP4 clan	No	157	Incompleted C-terminal; fragment
Scaffold0000002195	-	Tm_00300	CYP4 clan	Yes	522	Complete
Scaffold0000002651	-	Tm_06642	CYP4 clan	Yes	145	Incompleted N-terminal fragment
Scaffold0000006476	-	Tm_03283	CYP4 clan	Yes	217	Incompleted N-terminal and C-terminal
Scaffold0000006609	-	Tm_08626	CYP4 clan	No	58	Incompleted C-terminal; fragment
Scaffold0000006794	-	Tm_01048	CYP4 clan	No	550	Complete
Scaffold0000009676	-	Tm_15198	CYP4 clan	No	65	Incompleted N-terminal and

						C-terminal; fragment
Scaffold0000012018	-	Tm_01415	CYP4 clan	Yes	381	Complete
Scaffold0000012454	-	Tm_15199	CYP4 clan	No	69	Incompleted N-terminal and C-terminal; fragment
Scaffold0000013246	-	Tm_01436	CYP4 clan	Yes	524	Complete
Scaffold0000013315	-	Tm_11223	CYP4 clan	Yes	540	Complete
Scaffold0000013604	-	Tm_01583	CYP4 clan	Yes	490	Complete
Scaffold0000024534	-	Tm_15200	CYP4 clan	No	58	Incompleted N-terminal fragment
Scaffold0000026310	-	Tm_06697	CYP4 clan	Yes	491	Complete
Scaffold0000032351	-	Tm_15191	CYP4 clan	Yes	535	Complete
Scaffold0000000576	-	Tm_15192	Mito. Clan	No	68	Internal stop codon
Scaffold0000002181	-	Tm_15197	Mito. Clan	No	60	Incompleted N-terminal and C-terminal; fragment
Scaffold0000004081	-	Tm_07765	Mito. Clan	Yes	312	Incompleted C-terminal
Scaffold0000004970	-	Tm_07959	Mito. Clan	No	224	Incompleted N-terminal
Scaffold0000006918	-	Tm_01118	Mito. Clan	No	462	Complete
Scaffold0000008953	-	Tm_10252	Mito. Clan	Yes	266	Incompleted C-terminal
Scaffold0000015613	-	Tm_11295	Mito. Clan	Yes	289	Incompleted N-terminal
Scaffold0000015651	-	Tm_11316	Mito. Clan	Yes	232	Incompleted N-terminal
Scaffold0000015784	-	Tm_13084	Mito. Clan	Yes	515	Complete
Scaffold0000018584	-	Tm_13131	Mito. Clan	No	194	Incompleted N-terminal; fragment

Table 6.2 Comparison of the number of members in CYP2, 3, 4, and Mitochondrial clans in Insecta, Crustacea, and Acari. These numbers were derived from (Feyereisen 2012) and this study.

	Total	**CYP2 clan**	**CYP3 clan**	**CYP4 clan**	**Mitochondrial clan**
Insecta					
Drosophila melanogaster	88	7	11	32	36
Anopheles gambiae	105	10	9	46	40
Aedes aegypti	160	12	9	57	82
Bombyx mori	85	7	12	36	30
Apis mellifera	46	8	6	4	28
Nasonia vitripennis	92	7	7	30	48
Tribolium castaneum	134	8	9	45	72
Acyrthosiphon pisum	64	10	8	23	23
Pediculus humanus humanus	36	8	8	9	11
Crustacea					
Daphnia pulex	75	20	6	37	12
Acari					
Tropilaelapse mercedesae	56	7	19	20	10
Metaseiulus occidentalis	75	19	32	19	5
Tetranychus urticae	86	48	5	23	10

Table 6.3 Pair-wise comparisons of expression levels of *T. mercedesae* P450 genes between female, male, and nymph. The results are listed as the FDR value. Gray background highlights the FDR values to show the corresponding genes are statistically overrepresented in pair-wise comparisons (FDR < 0.05).

Gene ID	**CYP Clans**	**Female vs. Male**	**Female vs. Nymph**	**Male vs. Nymph**
Tm_00041	CYP2 clan	2.86E-01	9.13E-06	3.38E-03
Tm_10252	Mito. Clan	-	1.84E-03	8.03E-03
Tm_10021	CYP3 clan	6.02E-01	3.03E-02	2.01E-03
Tm_10022	CYP3 clan	8.43E-01	9.17E-02	1.75E-02
Tm_01157	CYP4 clan	1.04E-01	2.56E-01	9.92E-01
Tm_14091	CYP3 clan	9.62E-01	5.74E-01	9.74E-01
Tm_06697	CYP4 clan	9.88E-01	1.63E-02	2.46E-02
Tm_12854	CYP3 clan	1.00E+00	8.44E-01	9.91E-01
Tm_00664	CYP2 clan	1.00E+00	2.43E-01	2.15E-01
Tm_01372	CYP3 clan	9.88E-01	6.64E-01	9.81E-01
Tm_11223	CYP4 clan	1.00E+00	3.47E-01	5.78E-01
Tm_08470	CYP2 clan	9.60E-01	3.31E-01	1.55E-01
Tm_05131	CYP4 clan	5.34E-01	9.00E-01	8.58E-01
Tm_07853	CYP3 clan	9.07E-01	7.31E-02	2.06E-02
Tm_09891	CYP3 clan	9.88E-01	2.68E-01	1.58E-01
Tm_01415	CYP4 clan	1.00E+00	4.72E-02	1.06E-01
Tm_01436	CYP4 clan	8.72E-01	3.64E-03	2.34E-03
Tm_07004	CYP3 clan	1.07E-01	5.34E-01	7.46E-01
Tm_11316	Mito. Clan	-	9.75E-04	4.60E-03
Tm_11277	CYP2 clan	8.11E-01	8.86E-01	1.00E+00
Tm_13084	Mito. Clan	9.97E-01	1.00E+00	1.00E+00
Tm_06642	CYP4 clan	1.00E+00	1.66E-01	2.53E-01
Tm_03283	CYP4 clan	7.55E-01	2.50E-01	9.35E-01

Tm_11295	Mito. Clan	1.00E+00	9.89E-01	1.00E+00
Tm_03434	CYP3 clan	1.00E+00	8.86E-01	9.91E-01
Tm_09150	CYP3 clan	8.97E-01	2.11E-01	7.31E-01
Tm_02528	CYP4 clan	9.92E-01	8.45E-03	1.28E-02
Tm_02530	CYP4 clan	1.00E+00	4.07E-01	4.06E-01
Tm_07765	Mito. Clan	2.49E-03	4.81E-01	1.12E-01
Tm_01583	CYP4 clan	9.94E-01	4.58E-02	1.26E-01
Tm_00300	CYP4 clan	9.24E-01	5.74E-01	2.89E-01
Tm_14705	CYP3 clan	9.92E-01	8.90E-01	8.52E-01
Tm_15190	CYP3 clan	1.00E+00	8.90E-01	9.96E-01
Tm_15191	CYP4 clan	9.07E-01	1.55E-01	5.98E-01

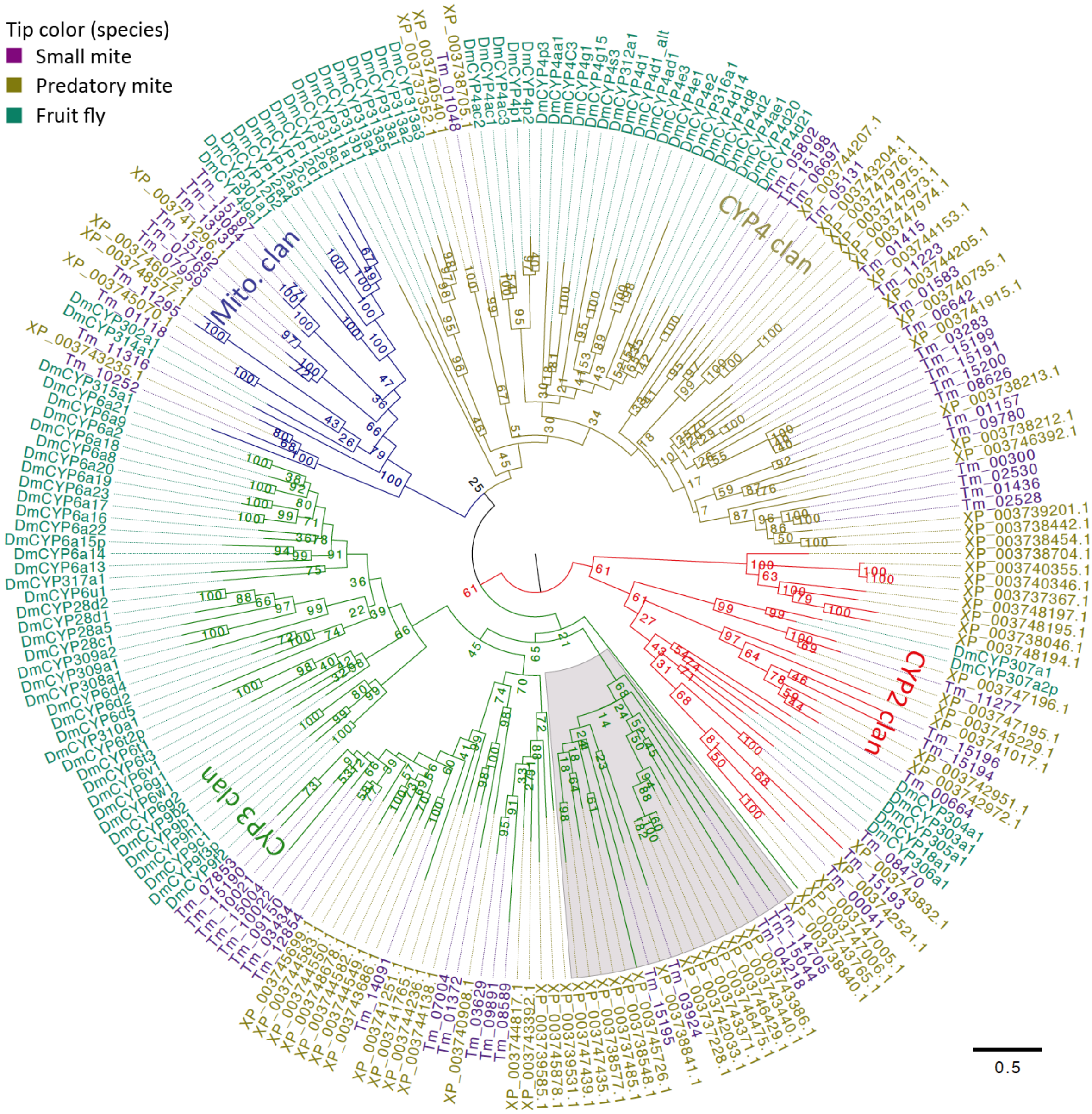


Figure 6.1 Maximum likelihood tree rooted at middle point to show the phylogenetic relationships of cytochorme P450s of *T. mercedesae*, *M. occidentalis,* and *D. melanogaster*. The phylogenetic tree was divided into four P450 clans, and the red, green, blue and yellow branches represent CYP2, CYP3, CYP4, and mitochondrial clan, respectively. Three species are distinguished by different colors; purple for *T. mercedesae*, yellow for *M. occidentalis*, and dark green for *D. melanogaster*. *D. melanogaster* P450s are named as DmCYP with the first two letters representing the acronym of scientific name, whereas the gene ids are used for *T. mercedesae* and *M. occidentalis*. The clade involved in constructing cytochrome P450 gene family of *T. mercedesae* (see the section 4.3.4) is highlighted by grey background.

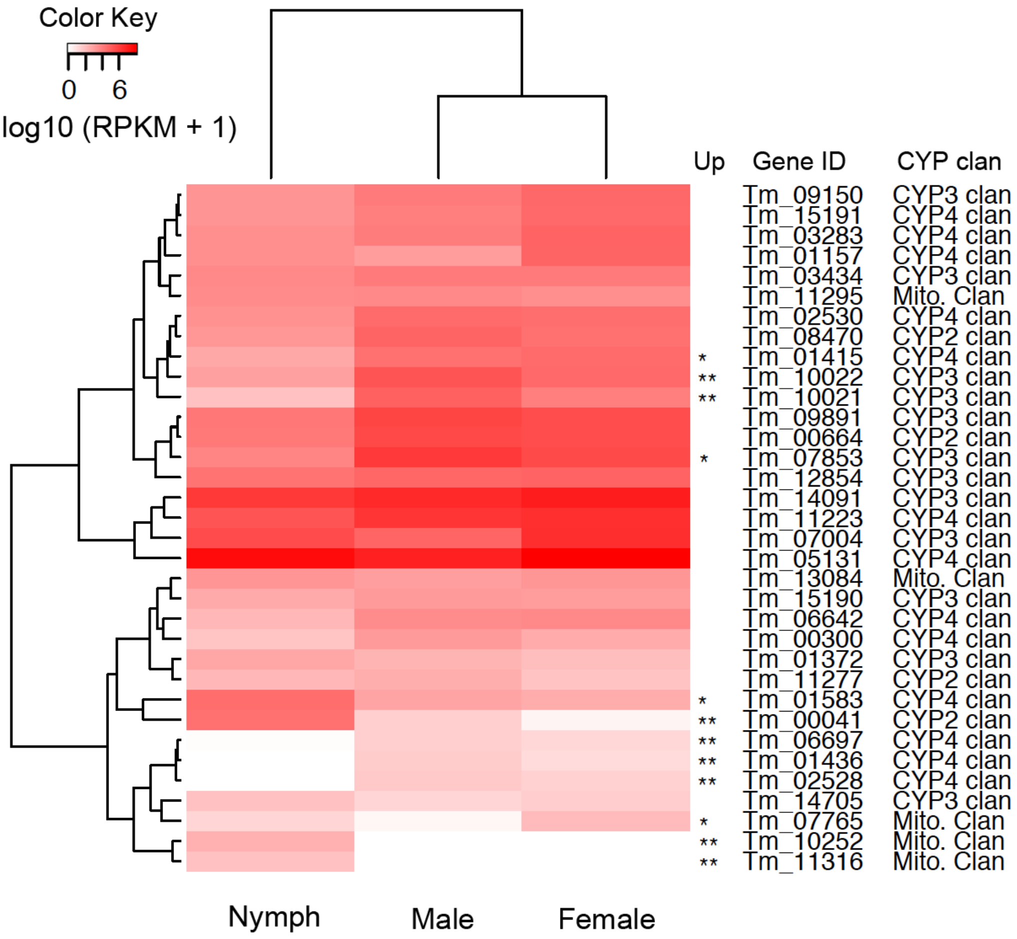


Figure 6.2 Expression profiling of cytochrome P450 genes in male, female, and nymph of *T. mercedesae*. The expression levels were measured in reads per kilobase of exon per million mapped sequence reads (RPKM) using normalized library sizes. Here, the 'expressed' was defined as a gene with a feature of at least one count per million in one sample or replicates. One asterisk (*) refers to significant transcriptional upregulation (FDR < 0.05) was detected by one particular pairwise comparison, and two asterisks (**) indicates that significant transcriptional upregulation (FDR < 0.05) was detected by two particular pairwise comparisons (see Table 6.3).

6.4 Glutathione S-transferase (GST)

Glutathione S-transferases (GSTs) comprising a family of eukaryotic and prokaryotic phase II metabolic enzymes (Liska 1998) are best known for their ability to catalyze conjugating reduced glutathione to electrophilic centers of various compounds (Li et al., 2007). It has been reported that GSTs play important roles for the development of pesticide resistance, disease pathogenesis, and cellular stress responses in mites and ticks (Liao et al., 2013). Based on their cellular localizations, all

GSTs have been classified into four major families, namely cytosolic GSTs, mitochondrial (kappa) GSTs, microsomal GSTs, and bacterial-fosfomycin-resistance proteins (Allocati et al., 2009). Based on the sequence homology, cytosolic GSTs have been further classified to mu, alpha, pi, theta, sigma, zeta, and omega as well as organism-specific classes of nu (nematode), lambda, phi, and tau (plants), beta (prokaryotes), delta, epsilon, iota, and chi (bacteria and insects) (Pandey et al., 2015).

GSTs candidates were identified using *T. mercedesae* and *M. occidentalis* total proteins as queries to search against the non-redundant (Nr) protein database of NCBI and further confirmed by Pfam, where the integrated GST proteins containing GST N-terminal domain (PFAM domain PF02798) or GST C-terminal domain (PFAM domain PF00043) were identified (sequence length > 200 amino acids). Then, all candidate GSTs of *T. mercedesae* and *M. occidentalis* together with *D. melanogaster,* and *T. urticae* GST sequences (based on Shi et al., 2012 and Grbić et al., 2011) were used as queries to manually search for putative *T. mercedesae* GSTs (TmGST) gene models by tblastn (E-value cutoff = 10^{-3}). TmGST gene models were confirmed or refined when the transcript sequences were available in the single Trinity *de novo* assembly by combining the reads across all samples (male, female, nymph, and randomly collected mites). Here, I manually annotated 15 TmGSTs in total. There are four genes containing one or more stop codons in the open reading frame, and two unexpressed short fragment genes (sequences length < 100 amino acids, half of the average size of GST proteins), suggesting they are probably pseudogenes (Table 6.5). Tandemly duplicated genes were not found.

All TmGSTs including above pseudogenes were aligned with *M. occidentalis* GSTs and selected *D. melanogaster* and *T. urticae* GSTs by Kalign using default setting. Poorly aligned sequences and N-terminal, C-terminal, and internal regions with variable length were excluded for the phylogenetic analysis. Other regions of potentially uncertain alignment between these highly divergent proteins were retained because they may provide important information for relationships within subfamilies. Based on the trimmed alignment, a PhyML tree was constructed using the ProtTest suggesting the best-fit substitution model of LG with invariant sites (I) and gamma (G) distributed rates (LG+I+G). SH-like local supports method was used to assess the significance of phylogenetic clustering. Based on the reference data sets (*D. melanogaster* and *T. urticae* GSTs), the phylogenetic analysis of TmGSTs has revealed the presence of four different families and an unclassified TmGST gene (Table 6.4; Figure 6.3).

Members of the Delta and Epsilon subclasses have been implicated for resistance to pesticides, for example, organophosphates, organochlorines, and pyrethroids (Enayati et al., 2005). The recent report has indicated that inhibition of mu class GST led to increased susceptibility of *Rhipicephalus sanguineus sensu lato* tick to permethrin (Duscher et al., 2014). Other GST subgroups of the Omega, Theta, and Zeta also appear to be involved in other cellular processes such as protection against oxidative stress (Roncalli et al., 2015). Sigma GSTs were found to play an important role for lipid peroxidation as well as detoxification (Gawande et al., 2014). Theta and Sigma class of GSTs are present in insects, human, and plants but absent in *T. mercedesae* and *I. Scapularis;* however, Theta gene is also present in *T. urticae* (Table 6.4). *T. mercedesae* also lacks Epsilon GSTs (Table 6.4). As a result, Delta and mu GSTs subclasses would be important for the pesticide resistance in *T. mercedesae* (Table 6.4). Interestingly, except kappa TmGST, all pseudogenes were assigned to above two GSTs subclasses but their full length orthologs (if existing) are present in *M. occidentalis* (Table 6.5; Figure 6.3), suggesting that Delta and mu GSTs subclasses have undergone constriction in *T. mercedesae*. Looking at the expression level, all TmGSTs are equally expressed in all three transcriptomes (Table 6.6).

Table 6.4 The number of GST subfamilies in different species. The numbers for each species are derived from Hayes et al. (2005), Oakeshott et al. (2010), Grbić et al., (2011), Reddy et al., (2011), and this study.

GST subfamily	**alpha**	**delta**	**epsilon**	**mu**	**pi**	**omega**	**sigma**	**theta**	**zeta**	**Kappa**	**unknown**	**Total**
T. mercedesae	0	5	0	6	0	2	0	0	0	1	1	15
M. occidentalis	0	3	0	6	0	3	0	0	0	2	1	12
I. scapularis	0	7	5	14	0	3	0	0	3	2	0	34
T. urticae	0	16	0	12	0	2	0	1	0	1	0	32
D. melanogaster	0	11	14	0	0	5	1	4	2	0	0	37
A. gambiae	0	12	8	0	0	1	1	2	1	0	3	28
A. mellifera	0	1	0	0	0	1	4	1	1	0	0	8
N. vitripennis	0	5	0	0	0	2	8	3	1	0	0	19
T. castaneum	0	3	19	0	0	4	7	1	1	0	0	35
B. mori	0	4	8	0	0	4	2	1	2	0	2	23
H. sapiens	5	0	0	5	1	2	1	2	1	1	0	18

Table 6.5 Properties of *T. mercedesae* GST genes classified by subfamily. If a GST gene contains at least one count per million in one transcriptome sample or the replicates, it is considered to be expressed and labeled 'yes', and if not, labeled 'no'. The 'fragment' indicates a GST gene encoding the protein shorter than 100 amino acids (aa), and 'complete' indicates a GST gene encoding the protein with both intact N-terminal and C-terminal domains. The unexpressed and fragmented GST genes and those with stop codons are treated as pseudogenes highlighted with gray background.

Gene ID	**GST subfamily**	**Length (aa)**	**Expression**	**Notes**
Tm_15207	Delta	50	no	Internal stop codon
Tm_15206	Delta	80	no	Short fragment
Tm_15204	Delta	214	yes	Complete
Tm_09167	Delta	180	no	Incompleted C-terminal
Tm_07535	Delta	218	yes	Complete
Tm_12349	kappa	224	yes	Complete
Tm_15205	mu	223	yes	Complete
Tm_15203	mu	116	no	Internal stop codon
Tm_15202	mu	199	no	Internal stop codon
Tm_15201	mu	139	no	Internal stop codon
Tm_07777	mu	72	yes	Complete
Tm_03887	mu	238	yes	Complete
Tm_06475	Omega	238	yes	Complete
Tm_06358	Omega	241	yes	Complete
Tm_05455	Unknown	84	no	Short fragment

Table 6.6 Pairwise comparisons of expression levels of *T. mercedesae* GST mRNAs between female and male, female and nymph, and male and nymph. The results are shown as the FDR value.

Gene ID	**GST family**	**Pairwise comparisons of expression**		
		female vs. male	**female vs. nymph**	**male vs. nymph**
Tm_15207	Delta	-	-	-
Tm_15206	Delta	-	-	-
Tm_15204	Delta	4.86E-01	3.88E-02	6.34E-01
Tm_09167	Delta	-	-	-
Tm_07535	Delta	8.06E-01	7.66E-02	4.94E-01
Tm_12349	kappa	1.00E+00	9.87E-01	1.00E+00
Tm_15205	mu	1.00E+00	6.34E-01	7.72E-01
Tm_15203	mu	-	-	-
Tm_15202	mu	-	-	-
Tm_15201	mu	-	-	-
Tm_07777	mu	2.58E-01	5.98E-01	-
Tm_03887	mu	1.00E+00	1.00E+00	1.00E+00
Tm_06475	Omega	1.00E+00	9.75E-01	1.00E+00
Tm_06358	Omega	9.30E-01	9.84E-01	8.90E-01
Tm_05455	Unknow	-	-	-

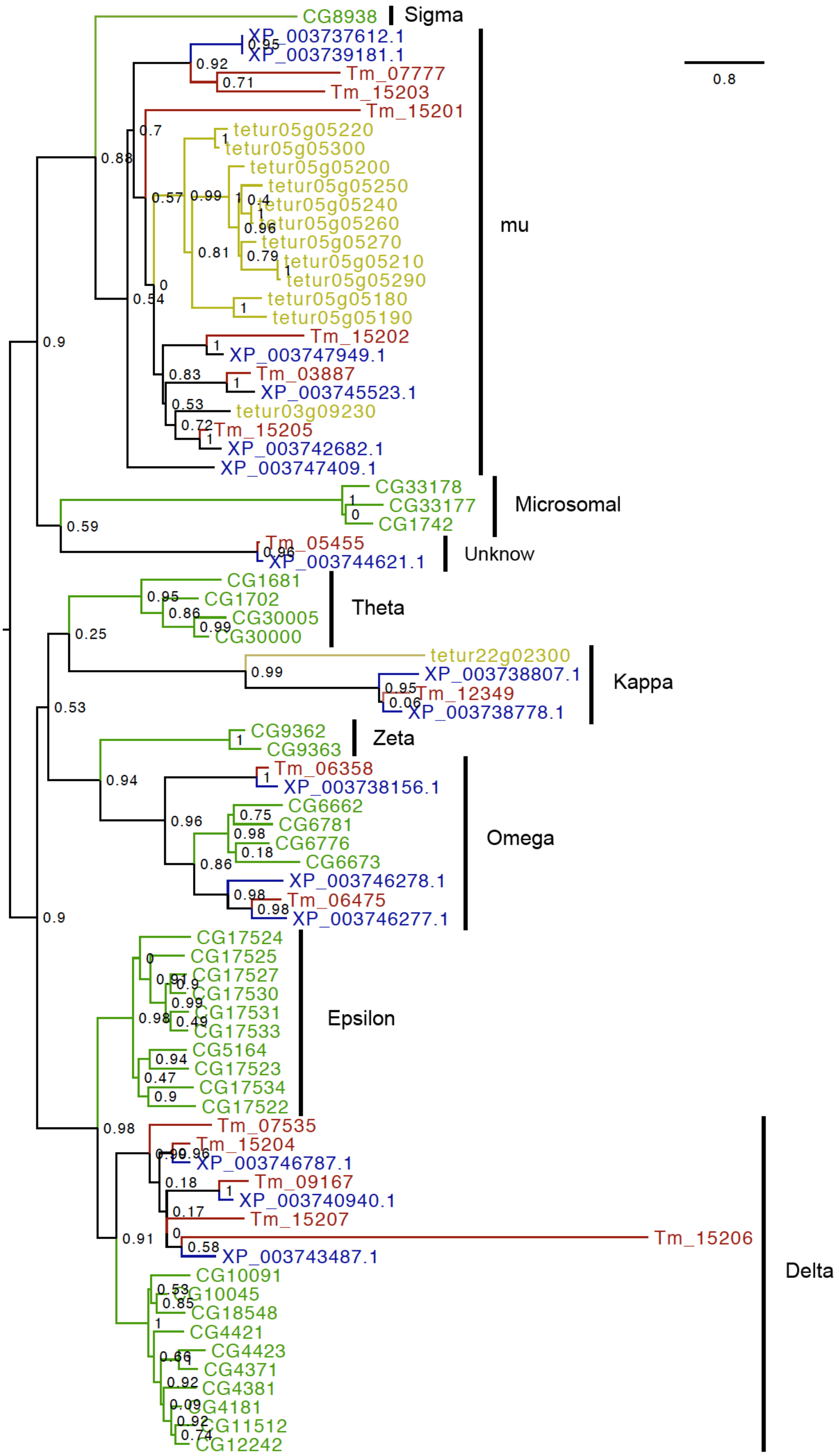


Figure 6.3 Maximum likelihood tree shows the phylogenetic relationships of GST proteins in *T. mercedesae* (red), *M. occidentalis* (blue), fuit fly (green), and *I. Scapularis* (yellow). The number at each branch node represents the bootstrap probability. The phylogenetically distinct clusters (subfamilies) and their names are shown on the right side of the tree.

6.5 Carboxyl/cholinesterases (CCEs)

The CCE superfamily is comprised of functionally diverse proteins that catalyze the hydrolysis of carboxylic esters to their component alcohols and acids (Yu et al., 2009). Insect esterases can be partitioned into 14 major clades (A to N), and ascribed into three big groups with the functions of dietary detoxification (A-C), hormone and pheromone degradation (D-H), and neurodevelopment (I-N) (Yu et al., 2009).

CCEs candidates were identified using *T. mercedesae* and *M. occidentalis* total proteins as queries to search against the non-redundant (Nr) protein database of NCBI and further confirmed by Pfam (PFAM COesterase domain PF00135). Then, all candidate GSTs of *T. mercedesae* and *M. occidentalis* together with *D. melanogaster* (based on Yu et al., 2009) were used as queries to manually search for putative *T. mercedesae* CCEs (TmCCE) gene models by tblastn (E-value cutoff =1e-3. TmCCE gene models were confirmed or refined when the transcript sequences were available in the single Trinity *de novo* assembly by combining the reads across all samples (male, female, nymph, and randomly collected mites). Here, I manually annotated 50 TmGSTs in total. There are three genes containing one or more stop codons in the open reading frame, and eight unexpressed short fragment genes (sequences length < 150 amino acids, half of the average size of GST proteins), suggesting they are probably pseudogenes (Table 6.8).

All TmCCEs including above pseudogenes were aligned with *M. occidentalis* CCEs and selected *D. melanogaster* CCEs and *B. mori* GSTs by Kalign using default setting. Poorly aligned sequences and N-terminal, C-terminal, and internal regions with variable length were excluded for the phylogenetic analysis. Other regions of potentially uncertain alignment between these highly divergent proteins were retained because they may provide important information for relationships within subfamilies. Based on the trimmed alignment, a PhyML tree was constructed using the ProtTest suggesting the best-fit substitution model of LG with invariant sites (I) and gamma (G) distributed rates (LG+I+G). SH-like local supports method was used to assess the significance of phylogenetic clustering. Based on the reference data sets (*D. melanogaster* and *B. mori*

CCEs), the phylogenetic analysis of TmCCEs has revealed the presence of three different families and two unclassified TmCCE genes (Table 6.7; Figure 6.4).

In insects, esterases from clades A-C are involved in broad substrate specificities and more general dietary and/or detoxification functions (Ramsey et al., 2010), but mites do not have this group (Table 6.7). Members of the second group (clades D-G) are virtually all secreted enzymes, which include pheromone degrading esterases (PDEs), juvenile hormone (JH), juvenile hormone esterase (JHE) and β-esterases (Ramsey et al., 2010). Only one JHE gene was found in *T. mercedesae* (Table 6.7). The third group (clades I-N), except for acetylcholinesterases that hydrolyze the neurotransmitter acetylcholine, tends to be non-catalytic and are involved in cell-cell interactions (Ramsey et al., 2010). In compared to insects, mites show an extensive acetylcholinesterase gene duplication (Table 6.7). Interestingly, There are two acetylcholinesterase genes (Tm_00699 and Tm_10626) are adult *T. mercedesae* specific expressed genes and two (Tm_10626 and Tm_09507) are nymph specific (Table 6.9).

Table 6.7 The number of CCEs subfamilies in different species. The numbers for each species are derived from Yu et al., (2009).

	T. mercedesae	***M. occidentalis***	***D. melanogaster***	**A. gambiae**	***A. mellifera***	***B. mori***
Dietary class						
Clade A	0	0	1	0	0	42
Clade B	0	0	12	0	8	13
Clade C	0	0	0	16	0	0
Hormone/semiochemical processing						
Clade D (integument esterases)	0	0	3	4	1	4
Clade E (secreted β-Esterases)	0	0	2	0	1	2

Clade F (dipteran JHE)	0	0	3	5	0	0
Clade G (lepidopteran JhE)	1	1	0	5	3	2
Neuro/developmental						
Clade H (glutactin)	0	0	1	1	1	1
Clade I (uncharacterized clade)	0	0	4	9	0	0
Clade J (acetylcholinesterases)	29	36	1	2	2	2
Clade K (gliotactin)	0	0	1	1	1	1
Clade L (neuroligins)	1	1	1	1	1	1
Clade M (neurotactins)	18	6	4	5	5	6
Clade N (neurotactin)	0	0	2	2	1	2
Other	2	3	0	0	0	0
CCEs total	50	47	35	51	24	76

Table 6.8 Properties of *T. mercedesae* CCEs genes classified by subfamily. If a CCE gene contains at least one count per million in one transcriptome sample or the replicates, it is considered to be expressed and labeled 'yes', and if not, labeled 'no'. The 'fragment' indicates a CCE gene encoding the protein shorter than 150 amino acids (aa), and 'complete' indicates a CCE gene encoding a complete COesterase domain (PF00135). The unexpressed and fragmented CCE genes and those with stop codons are treated as pseudogenes highlighted with gray background.

Gene ID	**CCE subfamily**	**Length (aa)**	**Expression**	**Notes**
Tm_00126	Clade G	594	yes	Complete
Tm_00699	Clade J	346	yes	Incompleted N-terminal
Tm_00745	Clade J	327	yes	Incompleted C-terminal
Tm_01590	Clade J	475	no	Complete
Tm_02203	Clade J	234	yes	Incompleted C-terminal
Tm_02327	Clade J	591	yes	Complete
Tm_02390	Clade J	516	yes	Complete
Tm_04344	Clade J	184	no	Incompleted C-terminal
Tm_05109	Clade J	295	yes	Incompleted N-terminal
Tm_06412	Clade J	1503	yes	Complete
Tm_08305	Clade J	689	yes	Complete
Tm_08415	Clade J	166	no	Short fragment
Tm_09468	Clade J	677	yes	Complete
Tm_09507	Clade J	438	yes	Incompleted C-terminal
Tm_10069	Clade J	563	yes	Complete
Tm_10626	Clade J	627	yes	Complete
Tm_10790	Clade J	321	yes	Incompleted N-terminal
Tm_11238	Clade J	577	yes	Complete
Tm_11273	Clade J	599	yes	Complete
Tm_11965	Clade J	468	yes	Complete
Tm_12088	Clade J	286	yes	Incompleted C-terminal

Tm_12621	Clade J	286	yes	Incompleted C-terminal
Tm_12914	Clade J	561	yes	Complete
Tm_13989	Clade J	548	yes	Complete
Tm_13990	Clade J	548	yes	Complete
Tm_13991	Clade J	551	yes	Complete
Tm_15253	Clade J	50	no	Short fragment
Tm_15254	Clade J	69	no	Short fragment
Tm_15250	Clade J	124	no	Internal stop codon
Tm_15251	Clade J	135	no	Internal stop codon
Tm_15252	Clade J	210	no	Internal stop codon
Tm_05721	Clade L	773	yes	Complete
Tm_01355	Clade M	650	yes	Complete
Tm_01458	Clade M	1053	yes	Complete
Tm_01522	Clade M	619	yes	Complete
Tm_01778	Clade M	176	yes	Incompleted C-terminal
Tm_01875	Clade M	199	yes	Incompleted C-terminal
Tm_01913	Clade M	168	no	Incompleted C-terminal
Tm_01918	Clade M	145	no	Short fragment
Tm_04413	Clade M	104	no	Short fragment
Tm_05618	Clade M	695	yes	Complete
Tm_06192	Clade M	597	yes	Complete
Tm_11821	Clade M	238	yes	Incompleted N-terminal
Tm_11953	Clade M	51	no	Short fragment
Tm_12394	Clade M	221	no	Incompleted N-terminal
Tm_12528	Clade M	1036	yes	Complete
Tm_12928	Clade M	772	yes	Incompleted N-terminal
Tm_13352	Clade M	890	yes	Complete

Tm_14715	Clade M	158	no	Incompleted N-terminal
Tm_14639	Clade M	108	no	Short fragment
Tm_04770	Undetermined	72	no	Short fragment

Table 6.9 Pairwise comparisons of expression levels of *T. mercedesae* CCE mRNAs between female and male, female and nymph, and male and nymph. The results are shown as the FDR value, and the upregulated sample was indicated in the bracket.

		Pairwise comparisons of expression		
Gene ID	**GST family**	**female vs. male**	**female vs. nymph**	**male vs. nymph**
Tm_00126	Clade G	9.26E-01	4.18E-01	9.43E-01
Tm_00699	Clade J	9.88E-01	1.71E-03 (female)	2.45E-02 (male)
Tm_00745	Clade J	3.43E-01	9.85E-01	2.73E-01
Tm_02203	Clade J	1.89E-05 (male)	2.32E-03 (nymph)	3.45E-01
Tm_02327	Clade J	5.23E-01	9.69E-01	7.80E-01
Tm_02390	Clade J	-	5.78E-07 (nymph)	3.14E-06 (nymph)
Tm_05109	Clade J	9.41E-01	1.74E-01	5.88E-02
Tm_06412	Clade J	8.48E-01	6.65E-01	1.00E+00
Tm_08305	Clade J	9.64E-01	7.89E-01	1.00E+00
Tm_09468	Clade J	8.23E-01	9.89E-01	9.27E-01
Tm_09507	Clade J	-	7.40E-06 (nymph)	1.91E-04 (nymph)
Tm_10069	Clade J	9.78E-01	1.46E-01	4.72E-01
Tm_10626	Clade J	3.54E-01	1.30E-02 (female)	2.73E-04 (male)
Tm_10790	Clade J	4.46E-06 (male)	1.14E-04 (nymph)	7.12E-01
Tm_11238	Clade J	9.94E-01	7.66E-02	5.56E-02
Tm_11273	Clade J	6.31E-01	1.35E-01	8.62E-01
Tm_11965	Clade J	-	1.55E-02(nymph)	5.13E-02
Tm_12088	Clade J	1.00E+00	1.00E+00	1.00E+00
Tm_12621	Clade J	1.00E+00	1.00E+00	1.00E+00
Tm_12914	Clade J	9.26E-01	1.93E-01	6.33E-02
Tm_13989	Clade J	9.09E-01	4.88E-01	9.78E-01
Tm_13990	Clade J	9.78E-01	6.40E-01	-
Tm_13991	Clade J	7.74E-01	8.91E-01	4.46E-01

Tm_05721	Clade L	9.09E-01	9.25E-01	7.10E-01
Tm_01355	Clade M	1.00E+00	9.67E-01	1.00E+00
Tm_01458	Clade M	9.92E-01	6.49E-01	9.54E-01
Tm_01522	Clade M	2.36E-01	1.14E-04(nymph)	1.51E-06(nymph)
Tm_01778	Clade M	-	6.40E-01	1.00E+00
Tm_01875	Clade M	9.08E-01	5.75E-01	1.00E+00
Tm_05618	Clade M	7.65E-01	5.09E-01	1.00E+00
Tm_06192	Clade M	1.03E-01	3.24E-01	9.60E-01
Tm_11821	Clade M	9.45E-01	7.06E-01	1.00E+00
Tm_12528	Clade M	6.12E-01	1.02E-01	8.17E-01
Tm_12928	Clade M	1.00E+00	6.46E-01	7.88E-01
Tm_13352	Clade M	3.15E-02 (male)	1.50E-01	1.78E-04 (male)

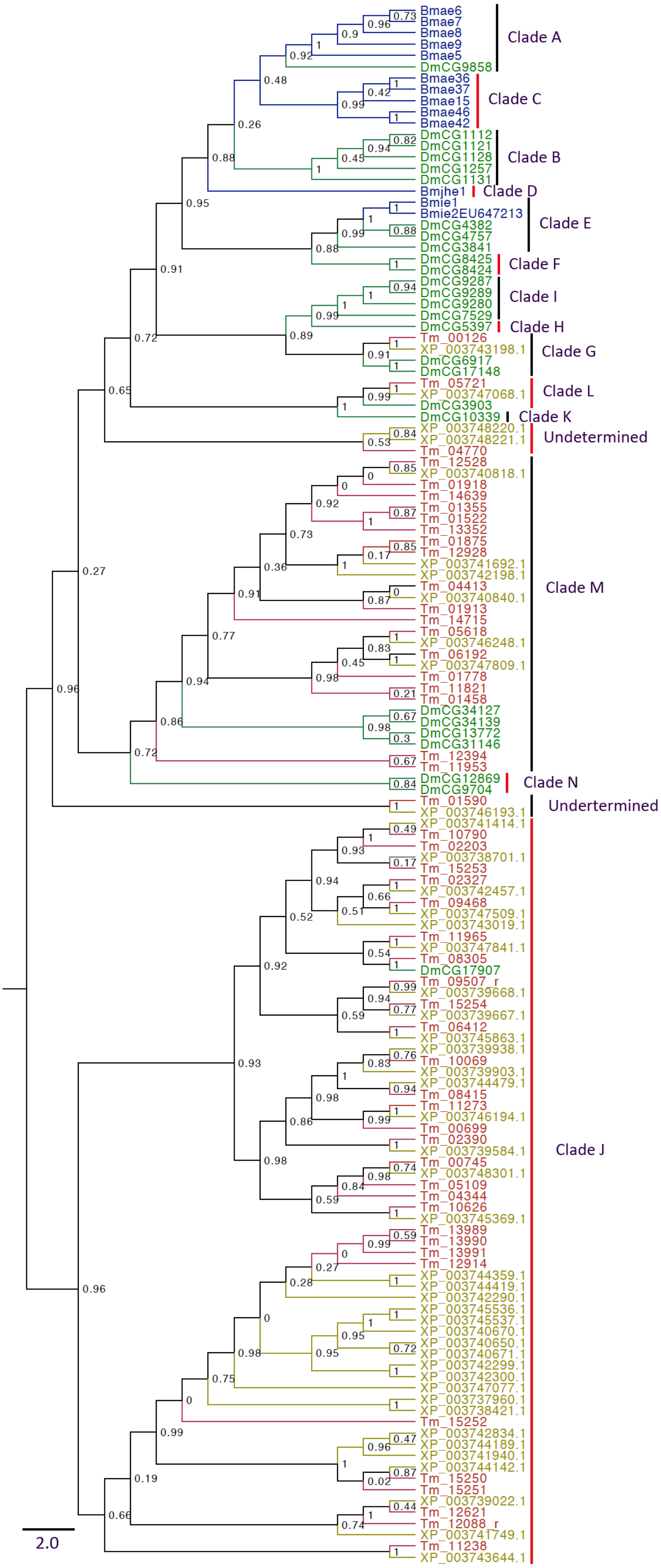

Bmae6
Bmae7
Bmae8
Bmae9
Bmae5
DmCG9858
Clade A
Bmae36
Bmae37
Bmae15
Bmae46
Bmae42
Clade C
DmCG1112
DmCG1121
DmCG1128
DmCG1257
DmCG1131
Clade B
Bmjhe1
Clade D
Bmie1
Bmie2EU647213
DmCG4382
DmCG4757
DmCG3841
Clade E
DmCG8425
DmCG8424
Clade F
DmCG9287
DmCG9289
DmCG9280
DmCG7529
Clade I
DmCG5397
Clade H
Tm_00126
XP_003743198.1
DmCG6917
DmCG17148
Clade G
Tm_05721
XP_003747068.1
DmCG3903
Clade L
DmCG10339
Clade K
XP_003748220.1
XP_003748221.1
Tm_04770
Undetermined
Tm_12528
XP_003740818.1
Tm_01918
Tm_14639
Tm_01355
Tm_01522
Tm_13352
Tm_01875
Tm_12928
XP_003741692.1
XP_003742198.1
Tm_04413
XP_003740840.1
Tm_01913
Tm_14715
Tm_05618
XP_003746248.1
Tm_06192
XP_003747809.1
Tm_01778
Tm_11821
Tm_01458
DmCG34127
DmCG34139
DmCG13772
DmCG31146
Tm_12394
Tm_11953
Clade M
DmCG12869
DmCG9704
Clade N
Tm_01590
XP_003746193.1
Undertermined
XP_003741414.1
Tm_10790
Tm_02203
XP_003738701.1
Tm_15253
Tm_02327
XP_003742457.1
Tm_09468
XP_003747509.1
XP_003743019.1
Tm_11965
XP_003747841.1
Tm_08305
DmCG17907
Tm_09507_r
XP_003739668.1
Tm_15254
XP_003739667.1
Tm_06412
XP_003745863.1
XP_003739938.1
Tm_10069
XP_003739903.1
XP_003744479.1
Tm_08415
Tm_11273
XP_003746194.1
Tm_00699
Tm_02390
XP_003739584.1
Tm_00745
XP_003748301.1
Tm_05109
Tm_04344
Tm_10626
XP_003745369.1
Tm_13989
Tm_13990
Tm_13991
Tm_12914
XP_003744359.1
XP_003744419.1
XP_003742290.1
XP_003745536.1
XP_003745537.1
XP_003740670.1
XP_003740650.1
XP_003740671.1
XP_003742299.1
XP_003742300.1
XP_003747077.1
XP_003737960.1
XP_003738421.1
Tm_15252
XP_003742834.1
XP_003744189.1
XP_003741940.1
XP_003744142.1
Tm_15250
Tm_15251
XP_003739022.1
Tm_12621
Tm_12088_r
XP_003741749.1
Tm_11238
XP_003743644.1
Clade J
2.0

Figure 6.4 Maximum likelihood tree shows the phylogenetic relationships of CCE proteins in *T. mercedesae* (red), *M. occidentalis* (yellow), fuit fly (green), and silkworm (blue). The number at each branch node represents the bootstrap probability. The phylogenetically distinct clusters (subfamilies) and their names are shown on the right side of the tree.

7. Sensory systems in *T. mercedesae*

7.1 Introduction

Successful parasites require efficient access and allocation of the hosts necessary for their survival. The parasites are able to use one of three basic strategies to find hosts: active transmission, passive transmission, transmission facilitated by another organism such as host manipulation, intermediate hosts, and vectors (Lewis et al., 2002). The host seeking behaviors is an essential prerequisite for the parasites in the active transmission, and general thought as an intrinsic response to sensory stimuli emanating from host (Haas 2003). The sensory stimulus can be either a physical or a chemical nature. Parasites need to able to detect these stimuli, indicating the presence of host within a highly complex environment of very different stimuli. In addition, the sensory system also helps animals to locate shelter, mates and offspring, and to avoid environmental danger (e.g. fires and noxious chemicals). Therefore, a better understanding of the sensory cues, especially chemosensory cues, that attracts and repels these parasites is also great interest for the development of specific and effective traps and repellents.

After emerging with adult bee, adult *T. mercedesae*s need to search new host (see Background) relying on their sensory system. In this chapter, I would focus on the genes referring to vision and chemoperception for exploring sensory world of *T. mercedesae*.

7.2 Materials and Methods

7.2.1 Phylogenetic analysis

Phylogenetic analyses were conducted as described in section 4.2.4. The constructed phylogenetic trees were viewed and graphically edited with FigTree.

7.2.2 Real time RT-PCR for relative quantification

Total RNAs were isolated from female *T. mercedesae* first legs, 2nd-4th legs, and whole bodies without legs using Trizol reagent (Thermo Scientific). The extracted RNA was then reverse transcribed to cDNA as described in Chapter 10. The cDNA products were then diluted 10-fold with DNase-free water. The primer pairs used for real time RT-PCR were designed based on gene structures of *T. mercedesae* genome assembly. Each primer pair spans at least one intron to eliminate the unexpected amplification from genomic DNA. The primer sequences and tested genes are summarized in Table 7.1.

Quantitative PCR reactions were performed using an Applied Biosystems 7500 Fast Real-Time PCR System and 2X KAPA SYBR FAST qPCR Master Mix (KAPA Biosystems Woburn, MA). Each reaction mixture (10 µl) contained 5 µl of 2× SYBR Master Mix, 200 nM each of forward and reverse primers, 1 µl of template cDNA, and 0.2 µl ROX dye. The PCR amplification cycling condition was; 95°C for 3 min, followed by 40 cycles of 30 sec at 95°C, 30 sec at 60°C, and 30 sec at 72°C, and a data collection window at 76-77°C for 30 sec. Relative quantification of cDNA was carried out by $2^{-\Delta\Delta Ct}$ method with Elongation factor-1alpha mRNA as an internal reference. Each experiment was repeated three times.

Table 7.1 Primer sets for real time RT-PCR validation.

Gene ID	Function protein	Forward primer (5' -> 3')	Reverse primer (5' -> 3')
Internal reference			
Tm_14054	Elongation factor-1 alpha	ATTCCGGTAAGTCAACCACCAC	GCTCGGCCTTCAGTTTGTCCAA
Tested genes			
Tm_08036	Peropsin	CCTGGCTGGACGGTGGCTCTACGGT	GATAACGTTGTGAGGGCGCAGCAT
Tm_15229	Ionotropic glutamate receptor	TTCACTTGGGATACTTGGCTAGC	GCACCGAAGCAGTACCACTCGCA
Tm_15231	Ionotropic glutamate receptor	GGAGACCTCGTCAAAAAGAAAGCGG	TGTCTCGAGCACTGTCAGGAACTTG
Tm_15244	Unclassified glutamate receptor	AGTTACGGCTGTATGCCAGTTGATG	GCCCGATATTGAAGCTGAGAATGAA

7.3 Visual system in *T. mercedesae*

Vision is important for activation and orientation of many blood-feeding insects (Thompson 1976). Hunting behavior is stimulated by visual images of host, and this has been shown with the tick, *Hyalomma asiaticum*, which has developed eyes (Waladde and Rice 1982). Not surprisingly, vision is most widely used by diurnal insects living in open habitats. However, visual information seems to be less important for host searching behavior of *T. mercedesae* because they spend whole life inisde the completely dark honey bee hive. Noticeably, *T. mercedesae* massively switched hosts from its native Asian open-nesting species, *A. dorsata* to honey bee, probably just since 70 years ago, when westerners introduced large numbers of honey bee colonies into Asia for commercial purposes (Crane 1988). This means, before they arrived honey bee's hive in the recent decades, there probably was light stimulus to *T. mercedesae*s for the past long years. Therefore vision could be also important for *T. mercedesae* in the search for the host. There is no previous report about the visual system of *T. mercedesae*, but I made interesting observation during the mite collection. The female mites move toward my white rubber gloves immediately when I open the capped brood cell. It is hard to directly investigate whether *T. mercedesae* has developed eyes like organ due to its tiny size, whereas opsin genes have been found in many animals and their apparent functions are related to vision and light-guided behaviors (Eriksson et al., 2013).

Different animals have different opsins. In the past thirty years, more than 2000 opsin genes have been identified (Terakita et al., 2012), and these can be phylogenetically divided into eight subfamilies (Terakita 2005; Figure 7.2). Members of four opsin subfamilies are known to act as light-sensing G protein coupled receptors (GPCRs) associating with Gt, Go, Gs, and Gq trimeric G proteins (Koyanagi et al., 2008). The Gt-, Go-, and Gs-coupled opsins are expressed in ciliary photoreceptor cells. Gt-coupled opsin subfamily comprises visual and non-visual opsins in vertebrates, but Go- and Gs-coupled opsins absent in higher vertebrates are found in molluscs and chordate amphioxus, and in cnidarians, respectively (reviewed by Shichida and Matsuyama 2009). Gq-coupled opsins expressed in rhabdomeric photoreceptor cells are the major opsins responsible for vision in the most of invertebrates such as cephalopods and arthropods (reviewed by Shichida and Matsuyama 2009). These G protein coupled opsins have seven transmembrane helical domains with the N-terminus in the extracellular side and the C-terminus in the cytosolic side. They are able to bind vitamin A-based chromophores (11-*cis* retinal) with a lysine residue at the seventh membrane-spanning domain in formation of photosensitive pigments, rhodopsins. Absorption of light by the 11-*cis* retinal-bearing rhodopsins causes 11-*cis* to all-*trans* isomerization of the chromophore. This change in the molecular structure of rhodopsin allows it to bind to and activate G protein by catalyzing the exchange of GDP to GTP, which mediates an enzymatic signaling cascade and eventually generates an electrical response in the photoreceptor cell by ion channels. Different opsins engage in the diverse enzymatic signaling pathways by coupling with different G proteins. The opsins transmitting light signals through the invertebrate Gq subfamily are involved with phospholipase C (PLC) (Terakita et al. 1993; Lee et al. 1994). The GTP-bound Gqα dissociates from Gβγ, and exposes its active site to bind and activate the effector enzyme, PLCβ. This PLC mediated enzymatic reaction results in visual signaling cascade to activate invertebrate visual TRP (transient receptor potential) channel in response to light stimuli.

The Gq signaling pathway in *D. melanogaster* has been illustrated in Figure 7.1. Our current KEGG annotation includes all necessary genes in this *D. melanogaster* phototransduction pathway with the exception of rhodopsin (Figure 7.1). Therefore, I further searched the current annotated genes in *T. mercedesae* genome, and found two genes (Tm_08036 and Tm_10503) belonging to opsin gene family. According to NCBI database, five opsin-like sequences are also found in *M. occidentalis* (XP_003744578.1 and XP_003744590.1) and *I. Scapularis* (ISCW004568, ISCW005498 and ISCW006006). These opsin genes were classified by phylogenetic analysis with neighbor-joining method and 1000 bootstrapping. The multiple alignment of the amino acid sequences was carried out using Mafft with '--auto' option. The gaps deletion of the alignment was set to 25%. Accession numbers of other opsins used for the analysis are listed in Table 7.2. Phylogenetic analysis of opsins indicated that the opsin-like sequences from *T. mercedesae* and *M. occidentalis* are clustered with peropsin subfamily members. Although the probability was too low to conclude that the mite opsins are peropsin subfamily members, they also contain the amino acid sequences conserved among deuterostome peropsins (Figure 7.3). This supports the idea that the mite opsins are peropsin homologues. In contrast, tick *I. Scapularis* appears to contain a Gq-coupled opsin and a photoisomerase opsin, but they are fragments suggested as incomplete annotation or pseudogenes (Eriksson et al., 2013). It should be noted that eyes was considered to be absent in *I. Scapularis* (Waladde and Rice 1982).

Peropsins and retinal photoisomerase opsins are known as all-*trans* retinal-bearing rhodopsins (Nagata et al., 2010). These two opsin subfamilies serve as retinal-photoisomerases, which bind all-*trans* retinal by a lysine residue in the seventh membrane-spanning domain to generate 11-*cis*-retinal (Hara and Hara 1967, 1968; Terakita et al. 1989). Unlike retinal photoisomerase opsins which do not couple to a G protein and thus do not generate a signaling cascade, peropsins have the conserved 'NPXXY' motif at the seventh transmembrane domain that plays important roles for coupling with G protein (Shichida and Matsuyama 2009). *T. mercedesae* and *M. occidentalis* peropsins also contain the highly conserved lysine residue and 'NPXXY' motif in the seventh transmembrane domain (Figure 7.3). Nagata et al., (2010) have proved that the jumping spider peropsin can form a photosensitive pigment in the retina, but localized in non-visual cells. If the peropsin functions as a photoisomerase, a retinal transport system with a retinal-binding protein that carries *11-cis* retinal to visual cells is necessary (Nagata et al., 2010). Vision in jumping spider depends on Gq-coupled opsin, but, obviously, the only conserved opsin genes in *T. mercedesae* are peropsins. Between two *T. mercedesae* peropsin genes, only Tm_08036 is expressed (containing at least one count per million in one sample or replicates). Both RNA-seq and qRT-PCR analyses consistently showed that Tm_08036 gene are expressed in the adult female (Figure 7.4A and B). Interestingly, this preopsin gene is predominantly expressed in the female body without legs (Figure 7.4C). These results suggest that the preopsin gene has important roles for adult female *T. mercedesae* in the body part.

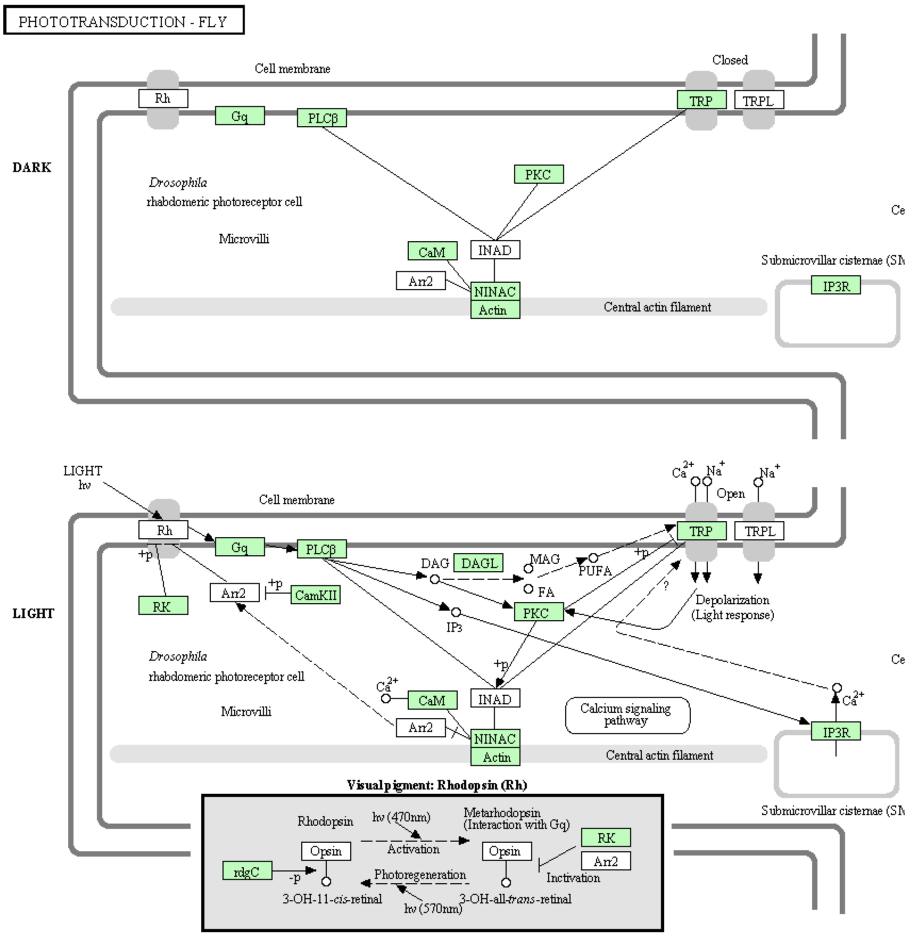


Figure 7.1 Phototransduction pathway (*D. melanogaster*). Green boxes indicate the genes present in *T. mercedesae* genome, whereas those in white boxes are absent.

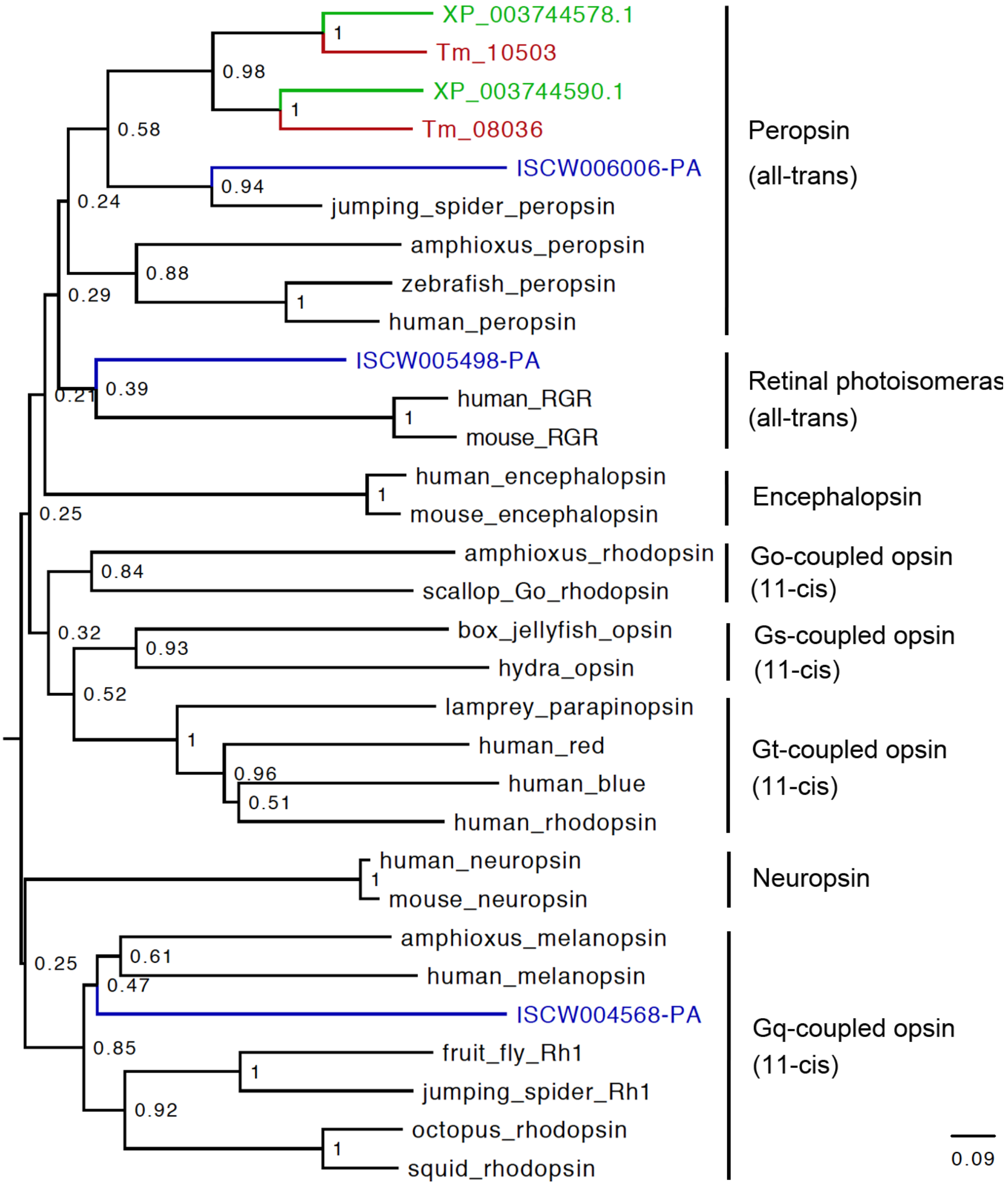
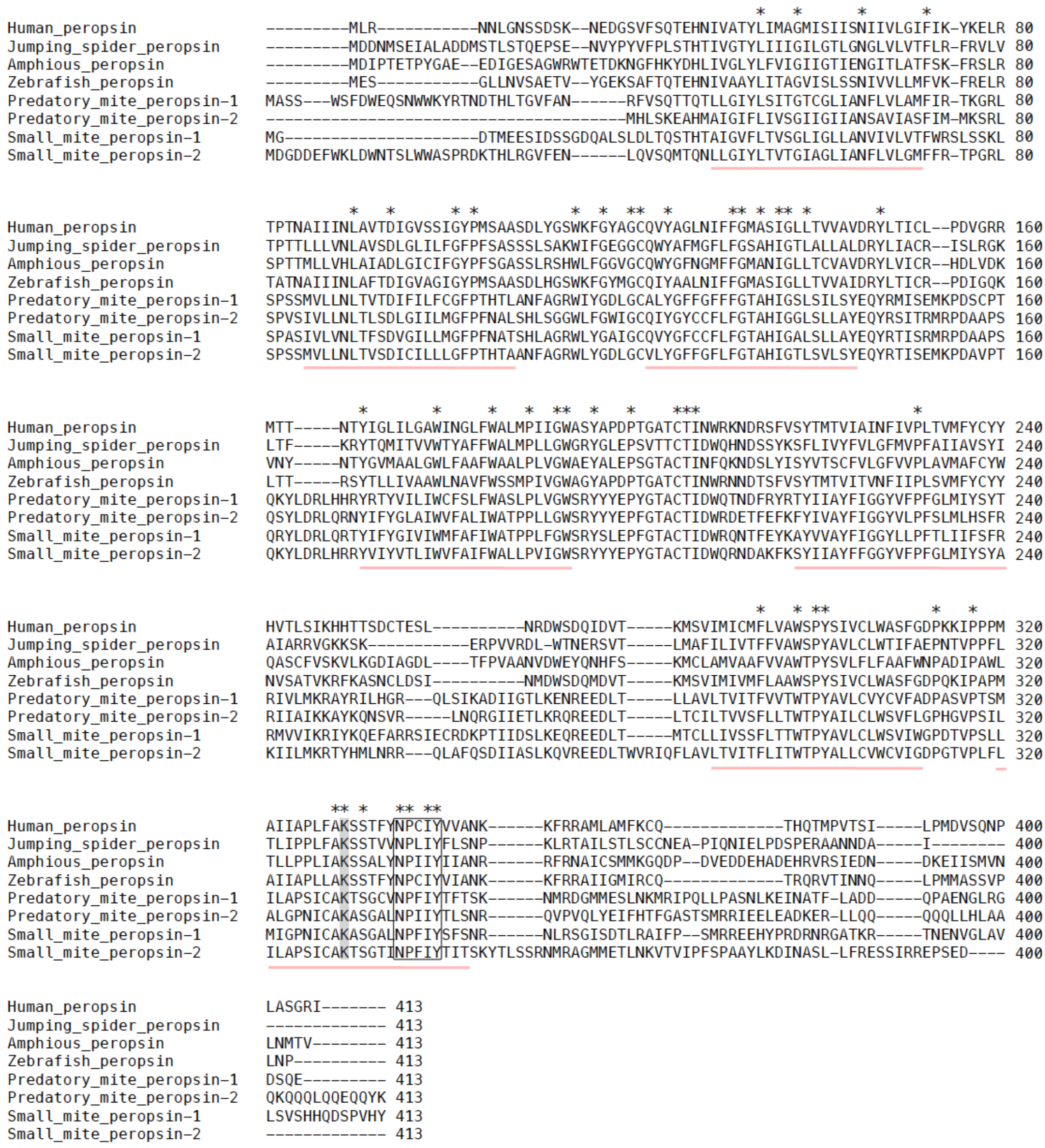


Figure 7.2 Phylogenetic positions of the *T. mercedesae* (red), *M. occidentalis* (green) and *I. Scapularis* (blue) opsins in the opsin family. The phylogenetic tree of opsin family was inferred by the neighbor-joining method. The number at each branch node represents the bootstrap probability (bootstrap=1000). The phylogenetically distinct clusters (subfamilies) and their names are shown on the right side of the tree. Known chromophore configuration in the dark state of members of each subfamily is indicated in parentheses.



```
                                                                                 *   *      *      *
Human_peropsin              ---------MLR-----------NNLGNSSDSK--NEDGSVFSQTEHNIVATYLIMAGMISIISNIIVLGIFIK-YKELR 80
Jumping_spider_peropsin     ---------MDDNMSEIALADDMSTLSTQEPSE--NVYPYVFPLSTHTIVGTYLIIIGILGTLGNGLVLVTFLR-FRVLV 80
Amphious_peropsin           ---------MDIPTETPYGAE--EDIGESAGWRWTETDKNGFHKYDHLIVGLYLFVIGIIGTIENGITLATFSK-FRSLR 80
Zebrafish_peropsin          ---------MES-----------GLLNVSAETV--YGEKSAFTQTEHNIVAAYLITAGVISLSSNIVVLLMFVK-FRELR 80
Predatory_mite_peropsin-1   MASS---WSFDWEQSNWWKYRTNDTHLTGVFAN------RFVSQTTQTLLGIYLSITGTCGLIANFLVLAMFIR-TKGRL 80
Predatory_mite_peropsin-2   --------------------------------------MHLSKEAHMAIGIFLIVSGIIGIIANSAVIASFIM-MKSRL 80
Small_mite_peropsin-1       MG---------------------DTMEESIDSSGDQALSLDLTQSTHTAIGVFLTVSGLIGLLANVIVLVTFWRSLSSKL 80
Small_mite_peropsin-2       MDGDDEFWKLDWNTSLWWASPRDKTHLRGVFEN------LQVSQMTQNLLGIYLTVTGIAGLIANFLVLGMFFR-TPGRL 80

                                     *   *      * *          *  *  **  *      ** * ** *       *
Human_peropsin              TPTNAIIINLAVTDIGVSSIGYPMSAASDLYGSWKFGYAGCQVYAGLNIFFGMASIGLLTVVAVDRYLTICL--PDVGRR 160
Jumping_spider_peropsin     TPTTLLLVNLAVSDLGLILFGFPFSASSSLSAKWIFGEGGCQWYAFMGFLFGSAHIGTLALLALDRYLIACR--ISLRGK 160
Amphious_peropsin           SPTTMLLVHLAIADLGICIFGYPFSGASSLRSHWLFGGVGCQWYGFNGMFFGMANIGLLTCVAVDRYLVICR--HDLVDK 160
Zebrafish_peropsin          TATNAIIINLAFTDIGVAGIGYPMSAASDLHGSWKFGYMGCQIYAALNIFFGMASIGLLTVVAIDRYLTICR--PDIGQK 160
Predatory_mite_peropsin-1   SPSSMVLLNLTVTDIFILFCGFPTHTLANFAGRWIYGDLGCALYGFFGFFFGTAHIGSLSILSYEQYRMISEMKPDSCPT 160
Predatory_mite_peropsin-2   SPVSIVLLNLTLSDLGIILMGFPFNALSHLSGGWLFGWIGCQIYGYCCFLFGTAHIGGLSLLAYEQYRSITRMRPDAAPS 160
Small_mite_peropsin-1       SPASIVLVNLTFSDVGILLMGFPFNATSHLAGRWLYGAIGCQVYGFCCFLFGTAHIGALSLLAYEQYRTISRMRPDAAPS 160
Small_mite_peropsin-2       SPSSMVLLNLTVSDICILLLGFPTHTAANFAGRWLYGDLGCVLYGFFGFLFGTAHIGTLSVLSYEQYRTISEMKPDAVPT 160

                                      *       *     *   *  **  *   *    ***                       *
Human_peropsin              MTT-----NTYIGLILGAWINGLFWALMPIIGWASYAPDPTGATCTINWRKNDRSFVSYTMTVIAINFIVPLTVMFYCYY 240
Jumping_spider_peropsin     LTF-----KRYTQMITVVWTYAFFWALMPLLGWGRYGLEPSVTTCTIDWQHNDSSYKSFLIVYFVLGFMVPFAIIAVSYI 240
Amphious_peropsin           VNY-----NTYGVMAALGWLFAAFWAALPLVGWAEYALEPSGTACTINFQKNDSLYISYVTSCFVLGFVVPLAVMAFCYW 240
Zebrafish_peropsin          LTT-----RSYTLLIVAAWLNAVFWSSMPIVGWAGYAPDPTGATCTINWRNNDTSFVSYTMTVITVNFIIPLSVMFYCYY 240
Predatory_mite_peropsin-1   QKYLDRLHHRYRTYVILIWCFSLFWASLPLVGWSRYYYEPYGTACTIDWQTNDFRYRTYIIAYFIGGYVFPFGLMIYSYT 240
Predatory_mite_peropsin-2   QSYLDRLQRNYIFYGLAIWVFALIWATPPLLGWSRYYYEPFGTACTIDWRDETFEFKFYIVAYFIGGYVLPFSLMLHSFR 240
Small_mite_peropsin-1       QRYLDRLQRTYIFYGIVIWMFAFIWATPPLFGWSRYSLEPFGTACTIDWRQNTFEYKAYVVAYFIGGYLLPFTLIIFSFR 240
Small_mite_peropsin-2       QKYLDRLHRRYVIYVTLIWVFAIFWALLPVIGWSRYYYEPYGTACTIDWQRNDAKFKSYIIAYFFGGYVFPFGLMIYSYA 240

                                                                                 *   * **           *   *
Human_peropsin              HVTLSIKHHTTSDCTESL----------NRDWSDQIDVT-----KMSVIMICMFLVAWSPYSIVCLWASFGDPKKIPPPM 320
Jumping_spider_peropsin     AIARRVGKKSK-----------ERPVVRDL-WTNERSVT-----LMAFILIVTFFVAWSPYAVLCLWTIFAEPNTVPPFL 320
Amphious_peropsin           QASCFVSKVLKGDIAGDL----TFPVAANVDWEYQNHFS-----KMCLAMVAAFVVAWTPYSVLFLFAAFWNPADIPAWL 320
Zebrafish_peropsin          NVSATVKRFKASNCLDSI----------NMDWSDQMDVT-----KMSVIMIVMFLAAWSPYSIVCLWASFGDPQKIPAPM 320
Predatory_mite_peropsin-1   RIVLMKRAYRILHGR---QLSIKADIIGTLKENREEDLT-----LLAVLTVITFVVTWTPYAVLCVYCVFADPASVPTSM 320
Predatory_mite_peropsin-2   RIIAIKKAYKQNSVR-----LNQRGIIETLKRQREEDLT-----LTCILTVVSFLLTWTPYAILCLWSVFLGPHGVPSIL 320
Small_mite_peropsin-1       RMVVIKRIYKQEFARRSIECRDKPTIIDSLKEQREEDLT-----MTCLLIVSSFLTTWTPYAVLCLWSVIWGPDTVPSLL 320
Small_mite_peropsin-2       KIILMKRTYHMLNRR---QLAFQSDIIASLKQVREEDLTWVRIQFLAVLTVITFLITWTPYALLCVWCVIGDPGTVPLFL 320

                                   ** *   ** **
Human_peropsin              AIIAPLFAKSSTFYNPCIYVVANK------KFRRAMLAMFKCQ-------------THQTMPVTSI-----LPMDVSQNP 400
Jumping_spider_peropsin     TLIPPLFAKSSTVVNPLIYFLSNP------KLRTAILSTLSCCNEA-PIQNIELPDSPERAANNDA-----I-------- 400
Amphious_peropsin           TLLPPLIAKSSALYNPIIYIIANR------RFRNAICSMMKGQDP--DVEDDEHADEHRVRSIEDN-----DKEIISMVN 400
Zebrafish_peropsin          AIIAPLLAKSSTFYNPCIYVIANK------KFRRAIIGMIRCQ-------------TRQRVTINNQ-----LPMMASSVP 400
Predatory_mite_peropsin-1   ILAPSICAKTSGCVNPFIYTFTSK------NMRDGMMESLNKMRIPQLLPASNLKEINATF-LADD-----QPAENGLRG 400
Predatory_mite_peropsin-2   ALGPNICAKASGALNPIIYTLSNR------QVPVQLYEIFHTFGASTSMRRIEELEADKER-LLQQ-----QQQLLHLAA 400
Small_mite_peropsin-1       MIGPNICAKASGALNPFIYSFSNR------NLRSGISDTLRAIFP--SMRREEHYPRDRNRGATKR-----TNENVGLAV 400
Small_mite_peropsin-2       ILAPSICAKTSGTINPFIYTITSKYTLSSRNMRAGMMETLNKVTVIPFSPAAYLKDINASL-LFRESSIRREPSED---- 400

Human_peropsin              LASGRI------- 413
Jumping_spider_peropsin     ------------- 413
Amphious_peropsin           LNMTV-------- 413
Zebrafish_peropsin          LNP---------- 413
Predatory_mite_peropsin-1   DSQE--------- 413
Predatory_mite_peropsin-2   QKQQQLQQEQQYK 413
Small_mite_peropsin-1       LSVSHHQDSPVHY 413
Small_mite_peropsin-2       ------------- 413
```

Figure 7.3 Comparison of the amino acid sequences between *T. mercedesae* peropsins and peropsins from other species (human, jumping spider, amphioxus, zebrafish, and *M. occidentalis*). The amino acid sequence alignment was preformed using Mafft in "-auto" option. The lysine (K) residue which binds to the retinal chromospheres, is highlighted in gray color. Amino acid residues conserved in all peropsins are indicated by asterisks (*). The'NPXXY' motif at the seventh transmembrane domain which is a highly conserved motif among G protein coupled receptors is boxed. It shows the remarkable conservation of the motif among peropsins. The transmembrane segments in peropsin amino acid sequences were analyzed using the online TMHMM Server v2.0 (Krogh et al., 2001). Seven transmembrane segments are underlined with red.

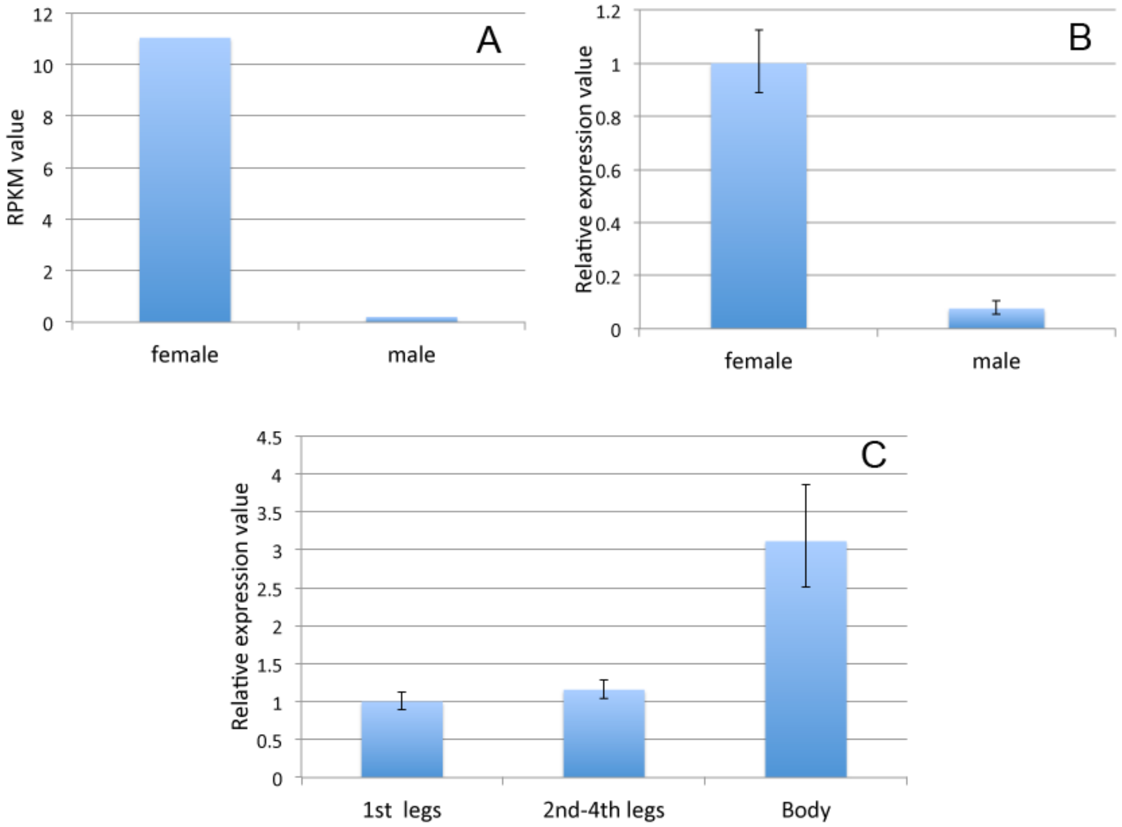


Figure 7.4 *T. mercedesae* peropsin mRNA (Tm_08036) expression analyzed by RNA-seq and qRT-PCR. Tm_08036 mRNA levels in male and female shown by RPKM value (A) and relative expression value (B). The relative expression levels of Tm_08036 mRNA in the first legs, 2nd-4th legs, and whole bodies without legs are also shown (C). Error bars indicate the standard deviations.

Table 7.2 Species name, opsin type, accession number, and the retrieved database of the sequences used for the phylogenetic analysis shown in Figure 7.2.

Species	**Opsin type**	**Accession number**	**Database**
Branchiostoma belcheri (amphioxus)	melanopsin	BAE00065.1	NCBI
Branchiostoma belcheri (amphioxus)	peropsin	BAC76023.1	NCBI
Branchiostoma belcheri (amphioxus)	rhodopsin	BAC76019.1	NCBI
Carybdea rastonii (box jellyfish)	opsin	BAG80696.1	NCBI
D. melanogaster (fuit fly)	fruit_fly_Rh1	CG4550	flybase
Homo sapiens (human)	RGR	AAB92384.1	NCBI
Homo sapiens (human)	blue	AAB05207.1	NCBI
Homo sapiens (human)	encephalopsin	AAD32671.1	NCBI
Homo sapiens (human)	melanopsin	AAF24978.1	NCBI
Homo sapiens (human)	neuropsin	AAR21109.1	NCBI
Homo sapiens (human)	peropsin	AAC51757.1	NCBI
Homo sapiens (human)	red	CAA92342.1	NCBI
Homo sapiens (human)	rhodopsin	AAC31763.1	NCBI
Hydra magnipapillata (hydra)	hydra_opsin	Hma2.233963	http://www.metazome.net
Hasarius adansoni (jumping spider)	Rh1	BAG14330.1	NCBI
Hasarius adansoni (jumping spider)	peropsin	BAJ22674.1	NCBI
Lethenteron camtschaticum (lamprey)	parapinopsin	BAD13381.1	NCBI
Mus musculus (mouse)	RGR	NP_067315.1	NCBI
Mus musculus (mouse)	encephalopsin	NP_034228.1	NCBI
Mus musculus (mouse)	neuropsin	NP_861418.2	NCBI
Enteroctopus dofleini (octopus)	rhodopsin	CAA30644.1	NCBI
Mizuhopecten yessoensis (scallop)	Go_rhodopsin	BAA22218.1	NCBI
Todarodes pacificus (squid)	rhodopsin	P31356.2	NCBI
Danio rerio (zebrafish)	peropsin	AAH81524.1	NCBI

7.4 Chemosensory system in *T. mercedesae*

Vision, for most instances, is usually integrated with information from other senses to search for the host. Lehane (2005) reviewed that host searching is initiated by host odor, and the visual information is only used at the final stages of orientation after getting close to the host. Inside the completely dark honey bee hive, chemosensory would be more important for *T. mercedesae* to locate the specific chemical signals emanating from the honey bee larvae. However, chemosensory receptors have been rarely studied with mite and tick lineage. Insect chemosensory receptor proteins at the plasma membrane of olfactory sensory neurons can bind odor ligands and shift this chemical signal into electrical signal (de Bruyne et al 2008; Benton et al., 2009; Sato et al., 2009). Three major chemosensory receptors in insects are odorant receptors (OR), ionotropic receptors (IR), and gustatory receptors (GR) (Vieira and Rozas 2011). To understand the gene composition of the chemosensory toolbox of *T. mercedesae*, I examined these three gene families as well as two major protein families involved in the perireceptor events of the insect chemosensory system (Vieira and Rozas 2011), odorant binding protein (OBP) and chemosensory protein (CSP) families.

7.4.1 Odorant binding protein

OBPs are small (generally 135-220 amino acids), globular, and soluble proteins with six α-helical domains joined by either two or three disulfide bonds (Sánchez-Gracia et al., 2009; Vogt 2003). They are abundantly expressed in the sensillar lymph of insects (Sánchez-Gracia et al., 2009), generally considered to bind different odorant molecules (Plettner et al. 2000) owing to their high divergence within the family. They transport the odorant molecules to their respective chemosensory receptors initiating the olfaction (Pelosi and Maida 1995). The Recent evidence has indicated the direct involvement of OBPs in the recognition of different volatile chemicals in insects (reviewed by Mulla et al., 2015). The arthropod OBPs constitute the specific large multi-gene family have been identified in a number of insect species (Table 7.3).

Search for *T. mercedesae* OBPs was initially preformed with tblastn search (e-value < 1e-3) against *T. mercedesae* genome assembly using *D. melanogaster*, *D. mojavensis*, *Anopheles gambiae*, *B, mori*, *Tribolium castaneum*, *A. Mellifera*, *P. humanus humanus*, and *Acyrthosyphon pisum* OBPs (identified by Vieira and Rozas 2011) as queries. No OBPs were found in this *T. mercedesae* genome assembly. Because OBPs are very divergent in terms of the amino acid sequences within the family, and the sequence identities between the family members from the different species could drop as low as 8% (Vieira and Rozas 2011), tblastn search would be ineffective to identify these genes. The OBPs search was therefore preformed again with blastp (e-value < 1e-3) to search both automated protein prediction from the genome assembly and *de novo* transcriptome assembly. As a result, the OBPs could not been found in *T. mercedesae*. Similarly, OBP family is also absent in the tick, *I. Scapularis* (Vieira and Rozas 2011), centipede, *Strigamia maritime* (Chipman et al., 2014), and water flea*, Daphnia pulex* (Peñalva-Arana et al., 2009).

7.4.2 Chemosensory protein

Chemosensory proteins (CSPs) are a large family of soluble polypeptides constituting another class of small ligand-binding proteins. The four conserved cysteines in CSPs form two disulfide bridges (Wanner et al., 2004). Compared to OBPs, CSPs (10-15 kDa) are smaller and show high amino acid identity between diverse insect species (Gu et al., 2012). They are also proposed to play an important role for in the insect chemoreception by capturing and transporting hydrophobic chemicals from the environment to the chemosensory receptors (Ishida et al., 2002; Calvello et al., 2003; Ozaki et al., 2005; González et al., 2009). However, except chemosensory organs such as antennae, maxillary palps, labial palps, and proboscis (Gu et al., 2012), CSPs are also expressed in various non-sensory tissues (Wanner et al., 2005). They seem to play roles in development, moulting (Wanner et al., 2005), and leg regeneration (Kitabayashi et al., 1998) that are unrelated to chemical communication.

Identification of *T. mercedesae* CSPs was preformed using the same methods as the OBPs. The queries sequences were also based on Vieira and Rozas' (2011) work, *D. melanogaster, D. mojavensis, A. gambiae, B. mori, T. castaneum, A. Mellifera, P. humanu humanus, A. pisum, I. Scapularis,* and *D. pulex*. Like OBPs, CSPs could not be identified in *T. mercedesae* either but they are putatively present in *I. Scapularis* (Vieira and Rozas 2011), centipede (Chipman et al., 2014), and water flea (Peñalva-Arana et al., 2009).

7.4.3 Odorant receptors

Insect ORs are seven transmembrane domain proteins (Clyne et al., 1999) unrelated to the vertebrate ORs that are G-protein-coupled receptors (GPCR). In contrast to the vertebrate ORs, they have the reversed membrane topology (with the intracellular N-terminus and the extracellular C-terminus) (Benton et al., 2006). They are responsible for detecting volatile compounds, and are a relatively recently evolved family within the insect chemosensory superfamily of ligand-gated ion channels (Su et al., 2009). ORs have been identified in a wide range of insects (Table 7.3) but, in general, show little sequence homology between the family members (Andersson et al., 2013).

Both tblastn and blastp searches were performed with the assembled *T. mercedesae* genome and transcritptome using *D. melanogaster* and *A. Mellifera* ORs (identified by Nozawa and Nei (2007) and Robertson and Wanner (2006), respectively) as queries. Although very relaxed e-value threshold (1e-3) was used, no positive hits were found. Therefore, ORs are also absent in *T. mercedesae* similar to other non-insect arthropods such as *I. Scapularis* (Vieira and Rozas 2011), *T. urticae* (Cao 2014), *Strigamia maritime* (centipede) (Chipman et al., 2014), and *D. pulex* (Peñalva-Arana et al., 2009). ORs and OBPs do not appear to exit in non-insect arthropods and may have emerged after speciation of insect.

7.4.4 Gustatory receptors

The GR family is referred to mediate most of insect gustation as well as some aspects of olfaction (Touhara and Vosshall 2009) including the carbon dioxide receptors in *D. melanogaster* (Jones et al., 2007) and *A. aegypti* (Erdelyan et al., 2012). *D. melanogaster, A. aegypti, B. mori,* and *A. pisum* have 50-100 GRs but *A. Mellifera* and *Pediculus humanus humanus* only contain 10 and 6 GRs, respectively (Table 7.3). GR family is more ancient than OR family since it is found in *D. pulex* (Peñalva-Arana et al., 2009), *I. Scapularis* (Hugh M. Robertson, unpublished), *S. maritime* (Chipman et al., 2014), and many other animals (Hugh M. Roberson, unpublished).

T. mercedesae GR family was manually annotated according to tblastn (E-value < 1e-3) search against *T. mercedesae* genome using all *I. Scapularis* GRs (indentified by Hugh M. Robertson, unpublished) as queries. Iterative search was also conducted with termite GRs as queries until no new genes were identified in each major subfamily or lineage. GR gene models were confirmed or refined when the transcript sequences were available in the single Trinity *de novo* assembly combining all reads from male, female, nymph, and randomly collected mite samples. I manually annotated five *T. mercedesae* GRs (Tm_09829, Tm_03548, Tm_09509, Tm_05586, and Tm_15249), in which one gene (Tm_15249) contains internal stop codons in the open reading frame, and two genes (Tm_03548 and Tm_09509) were supported by RNA-seq data.

In the phylogenetic analysis, five *T. mercedesae* GRs (TmGr) were aligned with *I. Scapularis* GRs (IsGrs) and *D. melanogaster* GRs (DmGrs) by Kalign with default setting. Poorly aligned and variable N-terminal and C-terminal regions as well as several internal regions of highly variable sequences were excluded for the phylogenetic analysis. Other regions of potentially uncertain alignment between these highly divergent proteins were retained because removing these regions can potentially mislead the relationships between subfamilies when they may provide important relationship information within subfamilies. Based on the trimmed alignment, a PhyML tree was constructed using the best-fit substitution model of LG with invariant sites (I) and gamma (G) distributed rates suggested by ProtTest. Here, SH-like local supports method was used to assess the significance of phylogenetic clustering. The phylogenetic tree demonstrated that none of TmGrs and IsGrs has orthologues in *D. melanogaster*. Except IsGr47, TmGrs and IsGrs cluster to two clades with four major subfamilies (Figure 7.5), indicating the separate expansion of GR repertoire in Parasitiformes lineage. All TmGrs are assigned to two of these subfamilies in a clade (Figure 7.5).

Both *T. mercedesae* and *I. Scapularis* are blood feeding Parasitiformes. *I. Scapularis* feeds a variety of animal hosts (Pinger 2008) and as a result, more GRs would be required for responding to various blood from the different hosts. Therefore, it is not surprising *T. mercedesae* and *P. humanus humanus* (human body louse) with very limited host only keep a small number of GR genes. GRs are important to mediate carbon dioxide response in both *D. melanogaster* and *A. aegypti*. However, Phylogenetic analysis indicates that *T. mercedesae* dose not have these two insect carbon dioxide GR genes.

Because no 1:1 orthology was found between TmGrs and well-characterized DmGrs, the functions of TmGrs are difficult to interpret. However, based on the phylogentic tree, GR genes expanded in the mite and tick shared the common ancestor with *D. melanogaster* sugar Grs, suggesting that they may have similar testate or olfactory functions. Among five TmGRs, Tm_03548 and Tm_09509 were supported by RNA-seq data. Tm_09509 mRNA is up-regulated in adult female (infective stage) and Tm_03548 mRNA is only expressed in male at low level (Figure 7.6). Interestingly, Tm_09509 has weak similarity to a trehalose receptor (by blastp), Gr5a in many insects. Indeed, trehalose is the major carbohydrate in honey bee larval hemolymph (Woodring et al., 1993).

Table 7.3 List of the numbers of genes associated with chemosensory system in *T. mercedesae* and other arthropods.

Species	**GR**	**OR**	**IR**	**OBP**	**CSP**
T. mercedesae	5	0	8	0	0
S. maritima	77	0	60	0	2
I. Scapularis	60	0	22	0	1
D. pulex	53	0	85	0	3
D. melanogaster	73	62	66	51	4
A. mellifera	10	163	10	21	6
B. mori	56	48	18	44	18
A. pisum	53	48	11	15	13
P. humanus humanus	8	10	12	5	7
A. aegypti	56	100	95	66	18

Data referred to references: Sa´nchez-Gracia et al., (2011), Chipman et al., (2014), Zhou J.J. (unpublished) and this study.

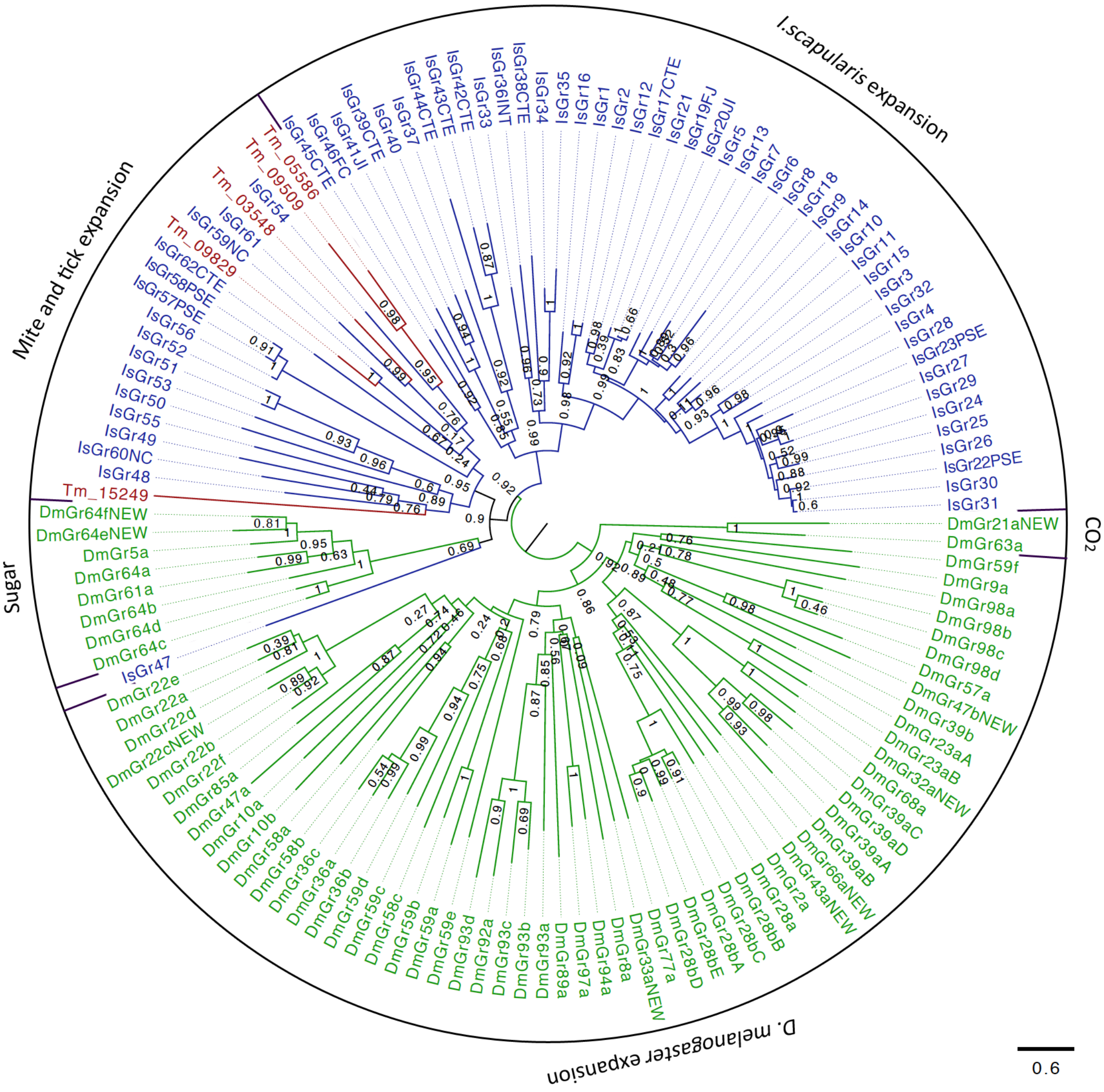


Figure 7.5 Maximum likelihood phylogenetic relationships of all gustatory receptors from *T. mercedesae* (red), *I. Scapularis* (Blue), and *D. melanogaster* (green). The tree was rooted at the middle point.

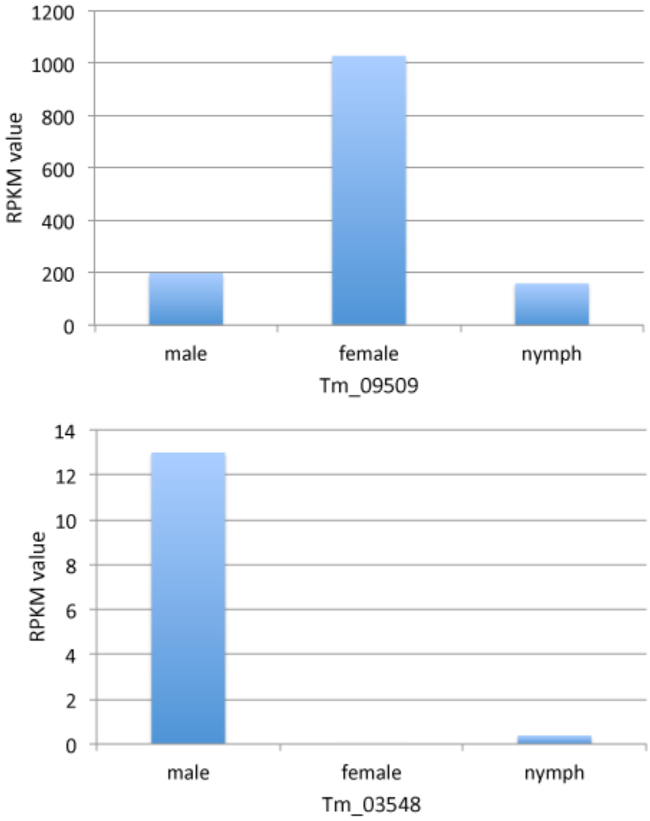


Figure 7.6 Analyses of two TmGrs (Tm_03548 and Tm_09509) mRNA shown with RPKM values.

7.4.5 Ionotropic receptors

In addition to ORs and GRs, IRs have been characterized as the second insect olfactory receptors (Benton et al., 2009), which are not related to insect ORs but rather have evolved from ionotropic glutamate receptors (iGluRs), a conserved family of ligand-gated ion channels involved in synaptic transmission (Croset et al., 2010). They include three subfamilies named after their major agonists: a-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), kainite, and N-methyl-D-aspartate (NMDA) receptors. IRs have been identified throughout protostome lineages, arthropod, mollusk, annelid, and nematode, and thus constitute a far more ancient group of receptors than ORs (Croset et al., 2010). In insects, the IRs are divided into two major groups: the “antennal IRs” that have olfactory functions and are conserved across the insect orders, and the species-specific “divergent IRs”, some of which may also play roles for taste detection (Croset et al., 2010). Antennal IRs in *D.*

melanogaster have the different odor specificity compared to the ORs and can respond to nitrogen-containing compounds, acids, and aromatics (Abuin et al., 2011).

All *T. mercedesae* iGluRs and IRs were manually annotated according to tblastn (E-value cutoff < 1e-3) search against *T. mercedesae* genome using all iGluRs and IRs identified by Croset et al., (2010) across vertebrates and invertebrates. Iterative searches were also conducted with termite protein as query until no new genes were identified in each major subfamily or lineage. The iGluR and IR gene models were confirmed or refined when the transcript sequences were available in the single Trinity *de novo* assembly based on combining all reads from male, female, nymph and randomly collected mites samples. I manually annotated 41 *T. mercedesae* IR (TmIR) and iGluR (TmiGluR) genes. The IGluRs and IRs in *M. occidentalis* were also identified as the complements for phylogenetic analysis.

In the phylogenetic analysis, all manually annotated *T. mercedesae* and *M. occidentalis* IRs and iGluRs were aligned with *D. melanogaster* IRs and iGluRs by Kalign with default setting. Poorly aligned and variable length N-terminal and C-terminal regions as well as several internal regions of highly length-variable sequences were excluded for phylogenetic analysis. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because remove of these regions can potentially mislead for relationships between subfamilies when they may provide important information for relationships within subfamilies. Based on the trimmed alignment, a PhyML tree was constructed using the ProtTest suggested best-fit substitution model of LG with invariant sites (I) and gamma (G) distributed rates. Here, SH-like local supports method was used to assess the significance of phylogenetic clustering. In order to do an analysis of the phylogeny of this gene family rooted with the more conserved iGluRs from which they evolved.

In 41 TmIRs and TmiGluRs, 34 proteins belong to iGluR subfamily that is not involved in chemosensation but highly conserved among animals in general. In the remaining 8 TmIRs, Tm_15231 and Tm_15229 are orthologs of DmIR25a and DmIR93a, respectively (Figure 7.7), that have been shown to express in the olfactory sensory neurons of *D. melanogaster* antennae (Rytz et al., 2013). The first two legs to serve as antennae-like sensory organs has been suggested for many mite species (see section 10.3.2). The results of qRT-PCR revealed that these two genes are highly expressed in the first legs of *T. mercedesae*s, suggesting that they may play an important role in the *T. mercedesae* olfactory sensory system (Figure 7.8 A and B).

In conclusion, *T. mercedesae* do not conserved the major olfactory receptor of OR and lost the carbon dioxide perception GR, but the antennal IRs identified in the *T. mercedesae* first two legs may play an important role in the olfactory sensory system within host searching behaviors. In compared to other species in Table 7.3, GR and IR gene families have contracted in *P. humanus humanus* (human body louse) and *T. mercedesae* probably because both of them are strict parasites with the specific hosts.

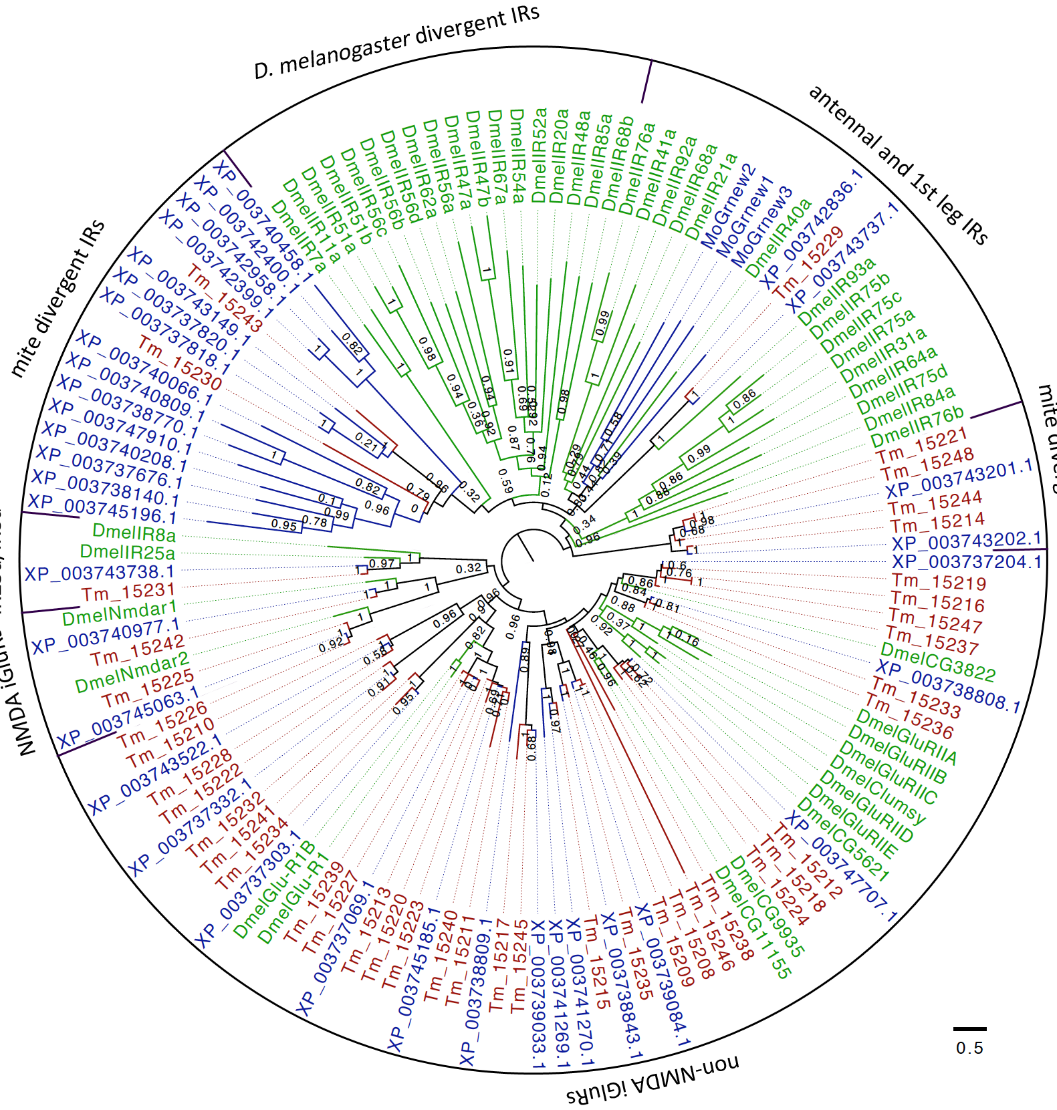


Figure 7.7 Phylogenetic tree of IRs and iGluRs (Ionotropic glutamate receptors) of *T. mercedesae* (red), *M. occidentalis* (blue), and *D. melanogaster* (green).

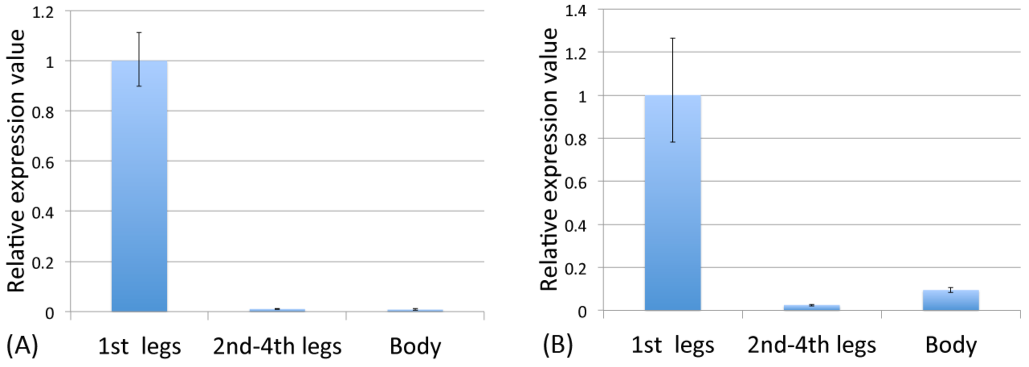


Figure 7.8 Relative mRNA expression levels of Tm_15229 (A) and Tm15231 (B) in the 1st legs, 2nd-4th legs, and main body of *T. mercedesae*.

8. Sex determination in *T. mercedesae*

Surprisingly diverse sex determination mechanisms have been studied in a variety of insect species. Genotype, especially ploidy level, is important for sex determination with the majority of insects. Many dipteran insects (flies and mosquitoes) have male heterogametic sex chromosome system, and thus females are XX and males are XY. While many lepidopteran insects (butterflies and moths) have female heterogametic sex chromosome system, and thus females are ZW and males are ZZ (Gempe and Bey 2011). Other insect species such as wasps, ants, and bees (hymenopteran insects) do not have visible sex chromosomes (Gempe and Bey 2011). Instead, they have haplodiploid sex-determination system in which sex is strictly determined according to the alleles at a single or series of loci. The fertilized diploid (2n) eggs become females and unfertilized haploid (n) eggs develop into males (Gempe and Bey 2011). *M. occidentalis* have an unusual genetic system to determine the sex known as parahaploidy in which functional elimination of the paternal set of chromosomes during early embryogenesis results in male develop (Nelson-Rees et al., 1980).

All of above-mentioned examples are for initiation of sex determination pathway. In *D. melanogaster*, the master-switch gene *Sex-lethal* (*Sxl*) regulates the female-specific splicing of pre-mRNA of *transformer* (*Tra*) (Bell et al., 1988). Two X chromosomes (high dose of X-linked genes) in *D. melanogaster* female activate *Sxl*, and result in the production of active Tra protein which interacts with protein product of the *transformer-2* (*Tra-2*) gene to produce female-specific splicing of doublesex (*dsx*) pre-mRNA (Erickson and Quintero 2007). In *D. melanogaster* male, the absence of functional Sxl and Tra proteins due to the low dose of X-linked genes results in the male-specific *dsx* pre-mRNA splicing to produce the unfunctional protein (Erickson and Quintero 2007). Dsx belongs to doublesex- and mab-3-related transcription factor (dmrt) family that appears to function in all animals as tissue-specific transcription factors involved in sex determination as well as many other developmental processes (Bellefroid et al., 2013). The intersex (ix) gene acts at the last step of the sex-determination cascade in *D. melanogaster* and is required for somatic female sexual development (Chase and Baker, 1995). For example, Ix protein interacts with female-specific *dsx* protein to regulate the expression of yolk protein genes in *D. melanogaster* (Garrett-Engele et al., 2002).

Understanding the molecular basis of sex determination in *T. mercedesae* could result in developing the improved genetic-control programs. Although possible molecular mechanisms for sex determination in *M. occidentalis* have been recently reported (Pomerantz et al., 2015), the sex determination genes were investigated without expression data in the male and female. In addition, only a few small-scale changes in existing and/or duplicated genes are sufficient to generate large differences in sex determination systems (Gempe and Bey 2011). Moreover, sex determination mechanisms can vary substantially between closely related species and even within a single species (Gempe and Bey 2011). Therefore, it would be essential to study sex-determination mechanisms in *T. mercedesae*.

Sex-determining genes of *D. melanogaste* (*Sex-lethal*, *transformer*, *transformer-2*, *dmrt,* and *intersex*) in *T. mercedesae* were manually annotated based on tblastn and blastp (E-value cutoff < 10^{-3}) searches against *T. mercedesae* genome using the arthropod sex-determining genes collected by Pomerantz et al., (2015), Zhang et al., (2014), and Kato et al., (2010) as queries. Dmrt proteins were further studied by phylogenetic analysis using neighbor-joining method with 1000 bootstrap. The multiple alignment of the amino acid sequences was calculated using Kalign with default setting. Gaps in the alignment were partially deleted (90%).

Similar to *M. occidentalis*, *T. mercedesae* does not contain upstream sex-determining genes (*Sxl* and *Tra*), but have the homologs of downstream sex-determining genes, *tra-2*, *dsx,* and *ix*. As most of arthropods, there is only one *tra-2* gene found in *T. mercedesae* genome (Table 8.2). The expression level of *T. mercedesae* tra-2 mRNA did not show sexual bias. Without *tra* genes*, T. mercedesae tra-2* apparently has no sex-determining function. Alternatively a novel upstream switch protein may work with *T. mercedesae tra-2* in the mite sex-determination. Wilkins (1995) proposed the hypothesis that sex-determination pathways have evolved by bottom up; the later a gene acts in the pathway, the more likely it will be conserved. Therefore, based on our current knowledge, *dsx* may be at the bottom of mite and tick sex-determining pathway. The number of *dmrt* genes in *T. mercedesae* is the most among arthropods in Table 8.2. It should be noted that *T. mercedesae* has two more *dsx* genes than *M. occidentalis* (Figure 8.2). The fourth *dmrt* group named dmrt93B is present in *T. mercedesae* (Tm_07872) but no apparent ortholog was found in *M. occidentalis* (Figure 8.2). Interestingly, Tm_07872 is the only mRNA highly expressed in the male among all sex-determination genes identified in *T. mercedesae* (Table 8.1; Figure 8.1). Other *T. mercedesae dmrt* mRNAs show very low expression level or are undetectable in both adult and nymph stages (Table 8.1; Figure 8.1). Highly conserved arthropod *ix* gene was also found in *T. mercedesae*, but no positive signal was detected with any *T. mercedesae* transcriptome sequences.

Table 8.1 Pairwise comparisons of mRNA expression levels of *T. mercedesae* sex determination genes between female and male, female and nymph, and male and nymph. If the gene contains at least one count per million in one sample or replicates when testing with all transcriptome samples, this gene was considered to be expressed and labeled with 'yes'. If no, labeled with 'no'. The results are listed as the FDR value.

Gene family	**Gene ID**	**Express**	**Pairwise comparisons of expression**		
			female vs. male	**female vs. nymph**	**male vs. nymph**
Transformer-2	Tm_09923	yes	0.973	0.845	0.653
Dmrt	Tm_02277	yes	0.740	-	0.670
	Tm_14784	no	-	-	-
	Tm_04084	yes	0.403	-	0.332
	Tm_05831	yes	0.305	-	0.088
	Tm_07872	yes	0.003	0.696	0.053
	Tm_08561	no	-	-	-
	Tm_08581	yes	0.023	-	0.027
Intersex	Tm_06007	no	-	-	-

Table 8.2 The numbers of *transformer-2*, *dmrt,* and *intersex* genes among arthropods.

Species	***Transformer-2***	***Dmrt***	***Intersex***
T. mercedesae	1	7	1
M. occidentalis	1	4	1
I. Scapularis	1	5	1
D. melanogaster	1	4	1
D. Pseudoobscura	1	4	1
Ceratitis capitata	1	4	1
Musca domestica	1	4	1
B. mori	1	4	1
A. mellifera	1	4	1

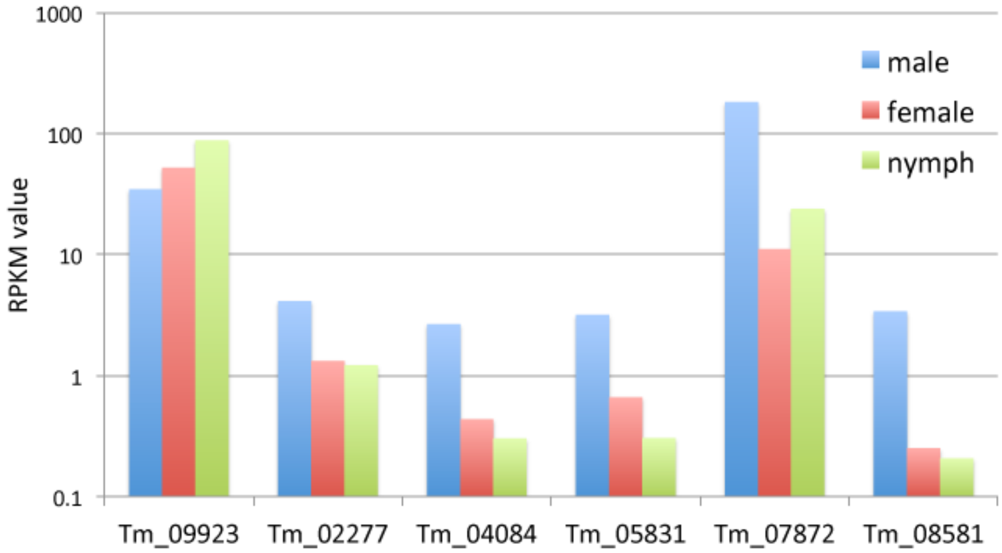


Figure 8.1 Expression analyses of *T. mercedesae Transformer-2* (Tm_09923) and *Doublesex* (Tm_02277, Tm_04084, Tm_05831, Tm_07872, and Tm_08581) mRNAs with RPKM value.

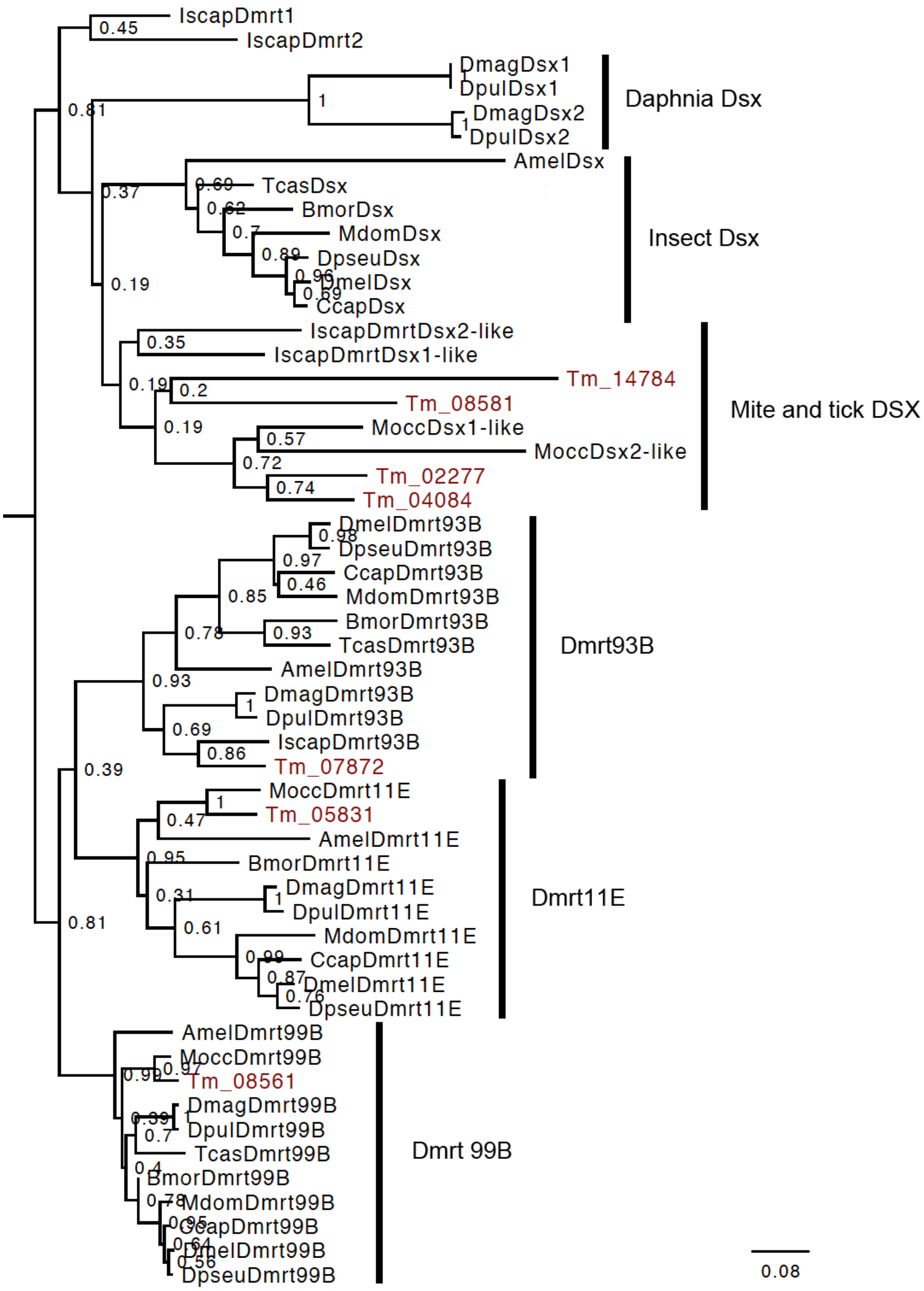


Figure 8.2 Neighbor-joining tree of the Dmrt proteins in *T. mercedesae* (Tm) and other arthropods (identified by Pomerantz et al., 2015), *M. occidentalis* (Mocc), *I. scapularis* (Iscap), *D. melanogaster* (Dmel), *D. pseudoobscura* (Dpseu), *C. capitata* (Ccap), *M. domestica* (Mdom), *T. castaneum* (Tcas), silkworm (Bmor), *A. Mellifera* (Amel), *D. magna* (Dmag), and water flea (Dpul). *T. mercedesae* proteins are highlighted with red. The number at each branch node represents the bootstrap probability (bootstrap=1000).

9. KEGG based analysis of novel drug targets

Although apicultural miticide plays major roles to control mite infestation, there have been no studies on the targets in *T. mercedesae* for the chemical control. Discovery of chemical control agents that selectively kill an arthropod pest of an arthropod host poses a unique pharmacological challenge. Studying metabolic pathways, particularly chokepoint reactions within the particular pathways has provided a systematic way of identifying novel potential targets for the chemical control of pathogenic organisms (Taylor et al., 2013). A chokepoint reaction is defined as a reaction that either consumes a unique substrate or produces a unique product. If the enzyme catalyzing such reaction can be inhibited, the entire pathway will be blocked, leading to accumulation of the unique substrate or the organism being starved of unique product (Yeh et al., 2014).

The chokepoint reactions and corresponding chokepoint enzymes were identified as descried by Taylor et al., (2013). Because the KEGG reaction database is no longer free for download, the reaction equations with directionalities (reversible or irreversible) were retrieved from the KEGG database via its new representational state transfer application-programming interface (REST API) using custom Perl script. The reactions were placed into a [compound × reaction number] matrix by parsing an intermediate file that contained the directionality and all the products and reactants for the reaction within the matrix. In this scheme, -1 indicated the compound was consumed (i.e. The compound was listed on the left side of the equation), +1 indicated the compound was produced (i.e. the compound was listed on the right side of the equation), 2 indicated the reaction was reversible, and 0 indicated the compound did not take part in the reaction. By parsing the matrix, a compound was produced or consumed in a single reaction (i.e. only a single 1 or -1 would be present across the entire compound row within the matrix) that can be identified as a chokepoint reaction. In some cases, a compound was uniquely produced or uniquely consumed, but the reaction they participated can be reversible. If this reversible reaction was the only reaction in which the compound participated, this was also called a chokepoint (i.e. a single 2 would be present within a row).

Prioritization of chokepoint reactions and targets was performed by assigning a point to meet each of the following criteria, then ranked based on the number of points; (1) transcript based expression (RPKM ≥ 10) found in adult female *T. mercedesae*, (2) chokepoint enzyme functioning in two or more pathways, (3) chokepoint enzyme involved in nucleic acid metabolism, and (4) chokepoint being a hydrolase based on their enrichment (classification as EC 3, enzyme commission number) as the targets in drug databases. The chokepoints have to satisfy at least one of the following two criteria, (5) all genes encoding chokepoint enzymes must have less than 30% sequence identity to those of *A. Mellifera* over half of the sequence length and (6) *T. mercedesae* specific chokepoint (not an *A. Mellifera* chokepoint). If the chokepoints satisfy the chokepoint criteria (5) and (6), the corresponding proteins are blastp searched against protein targets in DrugBank to identify targets with known inhibitors or approved drugs (e-value cutoff < 1e-5; 50% or greater identity over 80% of sequence length). To prioritize the drugs, drugs were given one point to meet for each of the following criteria; (1) molecular weight ≤ 500, (2) 0 < number of rotatable bonds ≤ 10, (3) hydrogen-bond donors ≤ 5, (4) hydrogen-bond acceptors ≤ 10, and (5) logP ≤ 5 (Lipinski et al., 2001).

T. mercedesae has 795 metabolic pathway enzymes with 275 (35%) classified as a chokepoint whereas there are 294 (33%) *A. Mellifera* chokepoint identified out of 761 metabolic pathway enzymes. The identified *A. Mellifera* chokepoints and their potential drugs were prioritized using the methods above, and prioritized chokepoints, and corresponding genes and drugs are listed in Appendix 2, Appendix 3, and Appendix 4.

10. Characterization of Transient receptor potential channel, subfamily A, member 1 (TRPA1)

10.1 Introduction

Ion channels in the transient receptor potential (TRP) superfamily share six common transmembrane segments that form sensor and pore domains and confer cation permeability. However, TRP channels are unusual among the various ion channels in that they display diverse cation selectivities and activation mechanisms (Venkatachalam and Montell 2007). TRP channels play major roles in a variety of sensory modalities such as vision, thermosensation, olfaction, hearing, taste sensation, and mechanosensation by functioning as primary signal integrators that allow animals to perceive the external environment (Damann et al., 2008). TRP channels also enable individual cells to detect changes in temperature, osmolarity, and fluid flow in their local environment (Venkatachalam and Montell 2007; Nilius and Owsianik 2011). TRP channels thus play essential roles in several physiological processes, including sensory, homeostatic, and motile functions.

The metazoan TRP superfamily is classified into seven subfamilies-TRPA, TRPC, TRPM, TRPML, TRPN, TRPP, and TRPV-based on their amino acid sequences and domains (Venkatachalam and Montell 2007; Nilius and Owsianik 2011). Among them, TRPA1 specifically contains 14-15 ankyrin repeats (ARs) in the N-terminus. ARs are 33 residue motifs consisting of pairs of antiparallel α-helices connected by β-hairpin motifs. ARs appear to be necessary for the sensitivity to various stimuli and interaction with protein partners (Nilius et al., 2012; Paulsen et al., 2015). Molecular structure of TRPA1 is similar to that of TRPV1 by forming a homotetramer to contain the pore with two gates (Paulsen et al., 2015). The physiological functions of TRPA1 have been characterized using genetically tractable model organisms, mouse, zebra fish, fruit fly, and nematode. TRPA1 is activated by nociceptive thermal (either heat or cold) and chemical stimuli, demonstrating that it plays major roles in nociception and inflammatory pain (Nilius et al., 2012). In addition, the roles in temperature entrainment of circadian clock, promoting longevity at cold temperatures, and induction of embryonic diapause in progeny have been recently reported using *D. melanogaster*, *C. elegans*, and *B. mori*, respectively (Lee and Montell 2013; Xiao et al., 2013; Sato et al., 2014).

In this study, I identified and characterized *T. mercedesae* TRPA1 (TmTRPA1) and discuss the isoform-specific modulation of chemical sensitivity, and the potential use of TmTRPA1-activating compounds to control honey bee ectoparasitic mites in the apiculture industry.

10.2 Materials and Methods

10.2.1 5' and 3'RACE of TmTRPA1

Based on VdTRPA1 and TRPA1 sequences of two mite/tick species, *M. occidentalis* (XM_003748080.1) and *I. scapularis* (XM_002405489.1), we first designed degenerate primers (5'-YGARGCKGCSAARAAYGCKTCSGCYAACGC-3' and 5'-GAACATGSCVGCGCARTGSARTGGCGTCAT-3' for the 1st PCR; 5'-ACVGAKCGRCCYTCYTTRTCYGTDGC-3' and 5'-GAACATGSCVGCGCARTGSARTGGCGTCAT-3' for the 2nd PCR) to amplify a partial *TmTRPA1* cDNA at 5' end using *T. mercedesae* total RNA by nested RT-PCR. The sequence of amplified band was highly similar to that of TRPA1 from other species, and then we designed the primer sets for 5' and 3' RACE using this cDNA fragment. *T. mercedesae* total RNA and two primers, 5'-CGCGGATCCACAGCCTACTGATGATCAGTCGATG-3' (for the 1st PCR) and 5'-TCGCAGAACTCGAGCAGAGCCCTCA-3' (for the 2nd PCR) were used for 5'RACE with SMARTer™ RACE cDNA Amplification Kit (Clontech Laboratories). 3'RACE was carried out as above except the following two primers were used; 5'-GTCCACTCGCATCTACTCGACCAG-3' (for the 1st PCR) and 5'-AAGAACCGCTCGAGTACCGCAGTG-3' (for the 2nd PCR). To fully extend 5' end of TmTRPA1 cDNA, additional two primers, 5'-ACGTTCGTCAGCTGCCAGTAACCGTTG-3' (for the 1st PCR) and

5'-ACTCTGCCATATTGCCCTTCTCCGCAG-3' (for the 2nd PCR) were used with 5'-Full RACE Kit (TAKARA).

10.2.2 Construction of TmTRPA1-expressing vector for mammalian cells

I isolated full-length *TmTRPA1a*, *TmTRPA1b*, and *TmTRPAc* cDNAs by nested RT-PCR with *T. mercedesae* total RNA and two primers for the 1st PCR, either 5′-AAAATCGTCGAAGGCTACTGCCTC-3′ (for *TmTRPA1a*), 5'-AATCCGCGTTCTACCCTTGACCGT-3' (for *TmTRPA1b*), or 5'-GAGCAGATCTCCATCATCCAGTCG-3' (for *TmTRPA1c*) and 5′-GTCCGAAGCCACCAAGGACGTAATAGG-3′. The second PCR was then carried out using the 1st PCR product as a template and the following two primers; either 5'-AATTTGCGGCCGCACC**ATG**GCGCGAAGGCTCATGAGAGATGCTTC-3' (for *TmTRPA1a*), 5'-AATTTGCGGCCGCACC**ATG**CGTCAACAAACCGTGTCTTGG-3' (for *TmTRPA1b*), or 5'-AATTTGCGGCCGCACC**ATG**GACCACGGAACGGTTCACACG-3' (for *TmTRPA1c*) and 5'-TTTCTAGACTCTTGACATCCGTTTGTCCATCTAGCTGTTC-3'. The 2nd PCR products were digested with Not I and Xba I, and then cloned in pAc5.1/V5-His B vector (Life Technologies) in which *D. melanogaster actin 5C* promoter was replaced with CMV promoter. The TmTRPA1 protein expressed by this construct was tagged with a V5-epitope at the C-terminus, and this was used for verifying the expression and the cellular localization in HEK 293 cells by western blot and immunofluorescence with rabbit anti-V5-epitope antibody (Sigma-Aldrich), respectively. The staining patterns of V5-epitope-tagged TmTRPA1 channels were compared with FITC-WGA which specifically labels the plasma membrane (Chazotte et al., 2011). A construct expressing untagged TmTRPA1 protein was then prepared using the above DNA construct as a template, the above primer with the initiation codon, and the primer 5′-TTTCTAGACTCTACTTGACATCCGTTTGTCCATC-3′. This DNA construct was used for all of the experiments described in the text.

10.2.3 RT-PCR

Total RNA was isolated from the *Tropilaelaps* mite first legs, 2nd-4th legs, and whole bodies without legs and mouth parts using Trizol reagent (Life Technologies). 0.1 μg of total RNA was used for the reverse transcription reaction using random primer and ReverTra Ace reverse transcriptase (TOYOBO). The RT products were then used for the 1st PCR with the following primers; either 5'-GCTGATGTCCCTCGCAACTGTGTT-3' (for *TmTRPA1a*), 5'-TGGCAGAAGGAAAAGCCCGTAGGA-3' (for *TmTRPA1b*), or 5'-GCCTGACTGACTGCATGAGAAGTC-3' (for *TmTRPA1c*) and 5'-CGAAACTGCATCCGACGTTCGTCA-3'. The second PCR was then carried out using the 1st PCR products as templates and the two primers, either 5'-AAAATCGTCGAAGGCTACTGCCTC-3' (for *TmTRPA1a*), 5'-AATCCGCGTTCTACCCTTGACCGT-3' (for *TmTRPA1b*), or 5'-GAGCAGATCTCCATCATCCAGTCG-3' (for *TmTRPA1c*) and 5'-CCAGTAACCGTTGAAACTCTGCCA-3'. The resulting PCR products were sequenced to verify their identities.

10.2.4 Ca^{2+}-imaging method (HEK293 cells)

For Ca^{2+}-imaging, 1 μg of TmTRPA1 expression vector and 0.1 μg of pCMV-DsRed expression vector were transfected to HEK293 cells in OPTI-MEM medium (Life Technologies) using Lipofectamine Plus reagents (Life Technologies). After incubating for 3-4 h at 37 ^{o}C, cells were reseeded on cover glasses and further incubated at 33 ^{o}C. The cells were used for the experiments at 20-40 h after transfection. Transfected HEK293 cells on a cover glass were incubated in culture media containing 5 μM fura-2 AM (Life Technologies) at 33 °C for 1-2 h. The cover glass was washed and fura-2 fluorescence was measured in a standard bath solution containing (in mM) 140 NaCl, 5 KCl, 2 $MgCl_2$, 2 $CaCl_2$, 10 HEPES, and 10 glucose, pH 7.4 adjusted with NaOH. Calcium chloride was omitted and 5 mM EGTA was added in the calcium-free bath solution. A cover glass was mounted in a chamber (RC-26G, Warner Instruments Inc.)

connected to a gravity flow system to deliver hot bath solution and bath solution containing various compounds. The concentration of each compound was 1 mM except for carvacrol, geranylacetone, nerol (0.5 mM), menthol (3 mM), and creosote (0.1 %). The emitted fluorescence (510 nm) by 340 and 380 nm were measured by CCD camera (CoolSNAP ES, Roper scientific photometrics). Data were acquired and analyzed by IPlab software (Scanalytics Inc.).

10.3 Results

10.3.1 Identification of three TmTRPA1 mRNA isoforms

I amplified a partial *TmTRPA1* cDNA at 5' end by nested RT-PCR with degenerate primers designed based on VdTRPA1 and TRPA1 sequences of two mite/tick species, *M. occidentalis* (XM_003748080.1) and *I. scapularis* (XM_002405489.1). I obtained the RT-PCR product with an expected size of ~500 bp, and the sequence had high similarity to TRPA1 sequences of other species. I then designed the primer sets for 5' and 3' RACE using this cDNA fragment, and 1.3 kb 5' RACE and 2.3 kb 3' RACE products were obtained and sequenced. Although only single sequence represented the 3' RACE product, I identified three different sequences in the 5' RACE product, demonstrating that *TmTRPA1* mRNA has at least three isoforms. In fact, three full length *TmTRPA1* cDNAs (*TmTRPA1a*, *TmTRPA1b,* and *TmTRPA1c*) which differ at 5' ends were finally isolated (Figure 10.1A). According to the assembled genomic sequence of *T. mercedesae*, I found that these three mRNAs are encoded by 23 (*TmTRPA1a*) and 26 (*TmTRPA1b* and *TmTRPA1c*) exons, and 22 exons are shared among them. They are likely to be generated by transcription from the different initiation sites (Figure 1A). Because the translational start codon is located in the unique 5' end exon of each isoform, three TmTRPA1 variants contain the different N-terminal sequences. TmTRPA1b and TmTRPA1c share the most of N-terminal sequence, and TmTRPA1a has the shortest N-terminal sequence (Figure 10.1B). Nevertheless, all isoforms contain 15 ARs and six transmembrane segments (S1-S6) forming the ion-transport domain.

10.3.2. Expression profiles of three TmTRPA1 mRNA isoforms in *T. mercedesae*

I examined the expression of *TmTRPA1a*, *TmTRPA1b,* and *TmTRPA1c* mRNAs in the first legs, second to fourth legs, and the whole body (excluding the mouthparts and legs) of *T. mercedesae* using RT-PCR. The first legs are much longer and thinner than the other legs (Figure 10.1C) and not used for mite movement, and instead, always raised in the air. This demonstrates that the first legs correspond to the insect antennae as suggested for many mite species (Cruz et al., 2005). *TmTRPA1a* and *TmTRPA1b* mRNAs were present in all of the above body parts; however, higher levels of *TmTRPA1a* and *TmTRPA1b* mRNAs were detected in the whole body and the first legs, respectively (Figure 10.1D). *TmTRPA1c* mRNA was only detected in the whole body but not in any legs by our assay (Figure 10.1D). Because of the preferential expression of TmTRPA1b in the first legs of *T. mercedesae*, this isoform is likely to have major roles for sensory perception by the sensory pit organ in the first legs of mites.

10.3.3 TmTRPA1 protein expression in HEK293 cells

Prior to conducting calcium imaging of HEK293 cells expressing one of three TmTRPA1 isoforms, we characterized the protein expression and cellular localization of V5-epitope-tagged TmTRPA1 isoforms in HEK293 cells. The proteins with expected molecular weights (134, 143, and 144 kD for TmTRPA1a, TmTRPA1b, and TmTRPA1c, respectively) were specifically detected using western blot (Figure 10.1E). I then tested the cellular localizations of TmTRPA1a, TmTRPA1b, and TmTRPA1c by staining the cells expressing the channels with FITC-WGA (wheat germ agglutinin) and anti-V5-epitope antibody. As shown in Figure 10.1F, the fractions of three isoform proteins co-localized with FITC-WGA, indicating that some of these proteins are present at the plasma membrane. TmTRPA1b isoform is highly expressed in the first legs of *Tropilaelaps* mite and localized at the plasma membrane of the transfected HEK293 cells, it is likely to play major roles in the sensory perception. I thus focused on characterizing

TmTRPA1b in this study.

10.3.4 Heat activation of TmTRPA1

We used calcium imaging technique with Fura-2 to measure the relative changes of intracellular calcium levels of HEK293 cells expressing TmTRPA1 by temperature fluctuations. Activation of the channel is expected to increase the intracellular calcium levels by influx of extracellular calcium. As shown in Figure 10.2, increased temperature elevated the relative intracellular calcium levels of cells expressing either TmTRPA1a, TmTRPA1b, or TmTRPA1c but not mock transfected cells. Low temperature did not increase the relative intracellular calcium levels of cells expressing TmTRPA1 (Figure 10.2). These results demonstrate that all of TmTRPA1 isoforms are heat-activated.

10.3.5 Differential activation of TmTRPA1 isoforms by chemical compounds

Previous reports have shown that mammalian TRPA1 and DmTRPA1 can be activated by a variety of compounds, including electrophilic compounds, which covalently modify the cysteine residues, and by other compounds (for example, menthol and nifedipine), which do not covalently bind with the channel (Nilius et al., 2012). Nevertheless, it is not known how far the above chemical activation profiles can be extended to TRPA1 in other species. I therefore tested the activation of three TmTRPA1 isoforms by various chemical compounds, and particularly focused on plant-derived compounds with tick-repellent activity (Bissinger and Roe 2010). We tested 39 compounds using HEK293 cells expressing TmTRPA1b by calcium-imaging technique, and found 27 of them activated it. The list of positive and negative compounds to activate TmTRPA1b was shown in Appendix 5. We observed robust activation of TmTRPA1b by eight representative plant-derived compounds (1,8-cineole, geranylacetone, 2-undecanone, β-citronellol, nerol, methyl jasmonate, carvacrol, and α-terpineol) as shown in Appendix 5A. Since 1,8-cineole inhibits human TRPA1 activity (Takaishi et al., 2012), its effect on TmTRPA1b appears to be the opposite. The compound 2-undecanone is already used as a major ingredient of commercially available natural arthropod repellents (Bissinger et al., 2009). The chemical structures of TmTRPA1b-activating compounds are diverse (Figure 10.3 and Appendix 5B). These results suggest that at least some of plant-derived tick/mite repellents activate the TRPA1 channels. Similar to the activation of TRPA1 of other species, for example, DmTRPA1, electrophilic compounds such as allyl isothiocyanate (AITC), cinnamaldehyde (CA), and diallyl disulfide also activate TmTRPA1b (Figure 10.4).

Although chemical activation profile of TmTRPA1c was identical to that of TmTRPA1b (Appendix 5A and 5C), we found that only six of the compounds activated TmTRPA1a. These include nerol, 2-undecanone, carvacrol, geranylacetone, eugenol, and terpinen-4-ol (Figure 10.5 and Appendix 5D). However, many other TmTRPA1b/c-activating compounds such as 1,8-cineole, methyl jasmonate, α-terpineol, and AITC can not activate TmTRPA1a (Figure 10.5 and Appendix 5A).

10.4 Discussion

10.4.1 *T. mercedesae* expresses three TmTRPA1 isoforms

T. mercedesae expresses three *TmTRPA1* mRNA isoforms, *TmTRPA1a, TmTRPA1b,* and *TmTRPA1c*, by transcription from the different initiation sites (Figure 10.1A). All isoforms contain the same number of ARs but unique N-terminal amino acid sequences. Similar isoforms are also present for *DmTRPA1*, for example, *TrpA1-RG* and *TrpA1-RI* (FlyBase, http://flybase.org/). Although DmTRPA1 isoform with short N-terminal sequence is sensitive to both heat and chemical stimulations, the isoform with long N-terminal sequence is only activated by chemical stimulation and is preferentially expressed in the chemosensory neurons. Thus, this could be a mechanism to discriminate heat and chemical stimuli by using a single *TRPA1* gene in *D. melanogaster* (Kang et al., 2012; Zhong et al., 2012). Similarly in *V. destructor*, two TRPA1 isoforms (VdTRPA1L and VdTRPA1S) with the unique N-terminal sequences and different numbers of ARs are present. In contrast to VdTRPA1L, VdTRPA1S containing fewer ARs is not activated by heat or chemical stimulation when expressed in HEK293 cells and *D. melanogaster* (Peng et al., 2015). Only VdTRPA1L appears to be a direct sensor involved in heat- and chemo-reception with its exclusive expression in the front legs of *V. destructor*. VdTRPA1S was proposed to be a downstream component of signaling pathways activated by certain sensory stimuli (Peng et al., 2015). Three TmTRPA1 isoforms are different from above DmTRPA1 and VdTRPA1 isoforms since all of them can be activated by heat (Figure 10.2) with the same number of ARs. However, as shown in Figures 10.3-10.5 and Appendix 5A-D, TmTRPA1a and TmTRPA1b/c have the different chemical sensitivities for activation. The long TmTRPA1b/c is activated by more compounds than the short TmTRPA1a. These results demonstrate that the N-terminal amino acid sequence of TmTRPA1 facing the cytosol is critical determinant for the chemical activation.

TmTRPA1c is exclusively present in the body but not in any legs of *Tropilaelaps* mite by our method, suggesting that it may be specifically expressed in the brain and/or internal organs of body (Figure 10.1C and D). TmTRPA1a and TmTRPA1b are present throughout the whole body but at higher level in the body and the first legs, respectively (Figure 10.1C and D). This may suggest that TmTRPA1a and TmTRPA1b are expressed in the different types of sensory neurons enriched in the body and the first legs, respectively. The first legs of *Tropilaelaps* mite morphologically differentiate from the other legs (Figure 10.1C) and are always held upright, demonstrating that they function as the sensory organs like insect antennae (Cruz et al., 2005). TmTRPA1b is likely to be a direct sensor involved in heat- and chemo-reception by the major sensory organ (the front legs) of *Tropilaelaps* mite.

10.4.2 Potential use of TmTRPA1-activating compounds to control *T. mercedesae* in apiculture industry

Commonly used miticides have the negative effects on honey bees, and have been resisted by the mites. Therefore, the effective and safe methods are required to control honey bee mites. TRPA1 functions as a nocisensor to induce avoidance behavior against noxious stimuli. If a natural compound capable of specially activating the TmTRPA1 but not AmHsTRPA (a functional counterpart of TRPA1 in honey bee; Kohno et al., 2010), it can be used as potential repellents to develop novel control method for *T. mercedesae*. Although most of the plant-derived TmTRPA1b-activating compounds also stimulated AmHsTRPA, a honey bee nocisensitive TRPA channel (Kohno et al., 2010), a few compounds such as α-terpineol and carvacrol described above were inactive to open AmHsTRPA channel (Peng et al., 2015). Thus, they may strongly repel *Tropilaelaps* mites rather than honey bees.

Figure 10.1 Three TmTRPA1 isoforms and their expression in HEK293 cells. (A) Exon-Intron structures of three *TmTRPA1* mRNA isoforms, *TmTRPA1a*, *TmTRPA1b*, and *TmTRPA1c* predicted from the assembled genomic sequence of *Tropilaelaps mercedesae*. Three isoforms share the same 184 bp exon (yellow rectangle) and the downstream exons (not shown) but contain the different upstream exons. *TmTRPA1b* and *TmTRPA1c* mRNAs share the same 92, 162, and 153 bp exons (blue rectangles). Purple, red, and green rectangles represent the exons unique to each isoform. The position of translational initiation codon (ATG) is also indicated for each isoform. The numbers indicate the sizes of exons and introns (lines), and they are not in scale. The most upstream 115 bp exon (red rectangle) of *TmTRPA1c* is missing in the assembled genomic sequence, and thus the size of the following intron sequence is not known (dashed line). (B) Alignment of the N-terminal amino acid sequences of three TmTRPA1 isoforms. The shared amino acids are highlighted with yellow, and a part of the amino acid sequence of the first ankyrin repeat (AR1) is indicated with purple. (C) Ventral view of *T. mercedesae*. Black and white arrows indicate the first legs and the second to fourth legs, respectively. (D) Detection of *TmTRPA1a, TmTRPA1b, TmTRPA1c,* and *β-actin* mRNA in the whole body without mouthparts and legs (Body), the first legs, and second to fourth legs by RT-PCR. The position of 200-600 bp DNA molecular weight marker (MW) is shown at the left. (E) Proteins expressed in HEK293 cells transfected with empty vector (Mock), TmTRPA1a-, TmTRPA1b-, and TmTRPA1c-expressing constructs were analyzed by western blot. The size (kD) of protein molecular weight marker (MW) is at the left. (F) Localizations of plasma membrane-bound FITC-WGA and either TmTRPA1a, TmTRPA1b, or TmTRPA1c tagged with V5-epitope in the transfected HEK293 cells by immunofluorescence. The merged images are also shown.

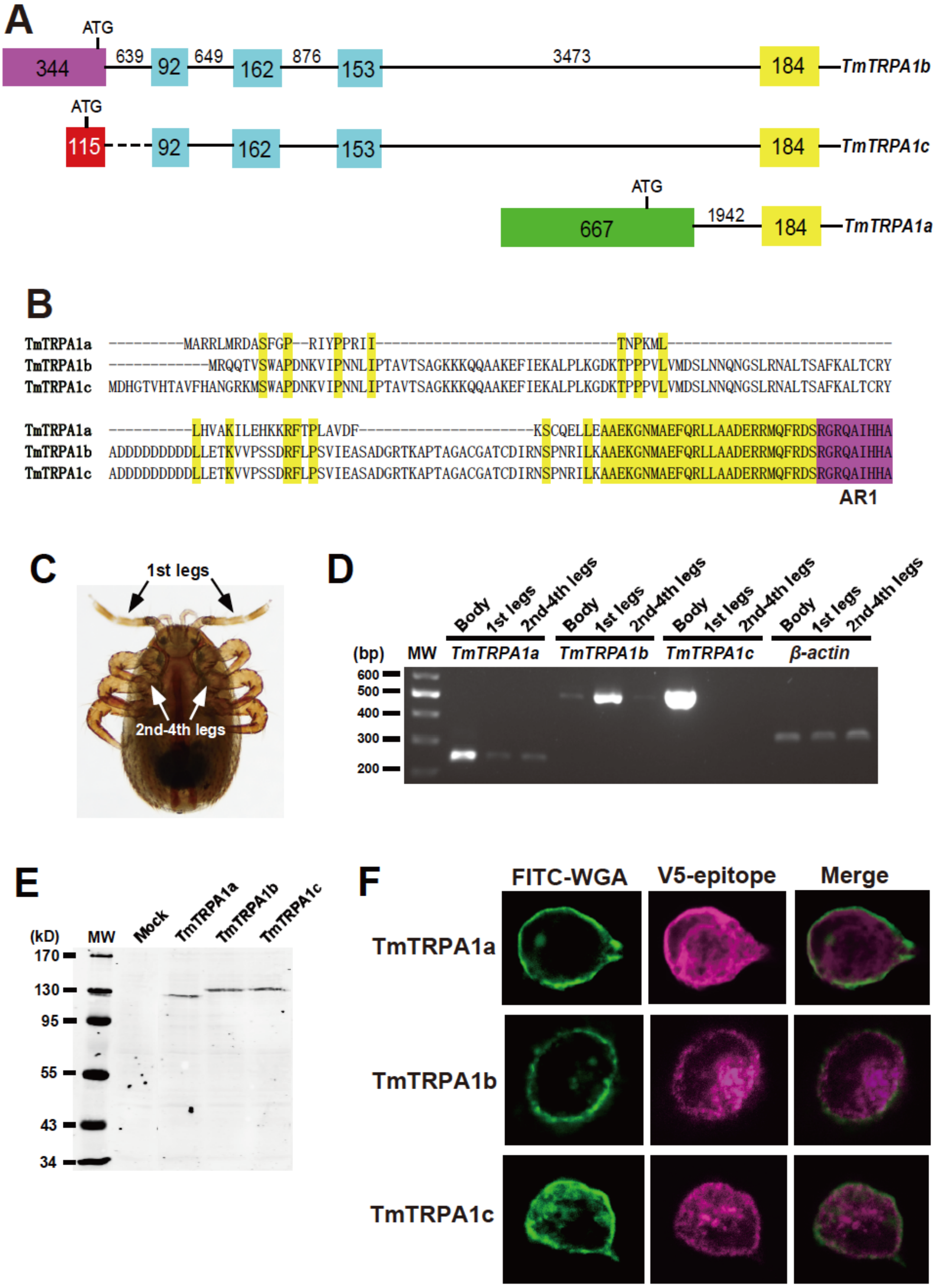

A
ATG
344
639
92
649
162
876
153
3473
184
TmTRPA1b
ATG
115
92
162
153
184
TmTRPA1c
ATG
667
1942
184
TmTRPA1a
B
TmTRPA1a ---------MARRLMRDASFGP--RIYPPRII-----------------------------TNPKML---------------------------
TmTRPA1b ------------MRQQTVSWAPDNKVIPNNLIPTAVTSAGKKKQQAAKEFIEKALPLKGDKTPPPVLVMDSLNNQNGSLRNALTSAFKALTCRY
TmTRPA1c MDHGTVHTAVFHANGRKMSWAPDNKVIPNNLIPTAVTSAGKKKQQAAKEFIEKALPLKGDKTPPPVLVMDSLNNQNGSLRNALTSAFKALTCRY
TmTRPA1a ----------LHVAKILEHKKRFTPLAVDF---------------------KSCQELLEAAEKGNMAEFQRLLAADERRMQFRDSRGRQAIHHA
TmTRPA1b ADDDDDDDDDLLETKVVPSSDRFLPSVIEASADGRTKAPTAGACGATCDIRNSPNRILKAAEKGNMAEFQRLLAADERRMQFRDSRGRQAIHHA
TmTRPA1c ADDDDDDDDDLLETKVVPSSDRFLPSVIEASADGRTKAPTAGACGATCDIRNSPNRILKAAEKGNMAEFQRLLAADERRMQFRDSRGRQAIHHA
AR1
C
1st legs
2nd-4th legs
D
(bp)
MW
Body
1st legs
2nd-4th legs
Body
1st legs
2nd-4th legs
Body
1st legs
2nd-4th legs
Body
1st legs
2nd-4th legs
TmTRPA1a
TmTRPA1b
TmTRPA1c
β-actin
600
500
400
300
200
E
(kD)
MW
Mock
TmTRPA1a
TmTRPA1b
TmTRPA1c
170
130
95
55
43
34
F
FITC-WGA
V5-epitope
Merge
TmTRPA1a
TmTRPA1b
TmTRPA1c

Figure 10.2 Heat activation of three TmTRPA1 isoforms. The upper traces indicate the changes of Fura-2 ratio (intracellular calcium level) in TmTRPA1a-, TmTRPA1b-, TmTRPA1c-, or Mock-transfected cells on temperature fluctuation in the presence of extracellular calcium. Each line represents the Fura-2 ratio in the individual cell measured by calcium imaging. The arrows show the time points when ionomycin was added. The lower traces show the changes of bath temperature by time (sec, second).

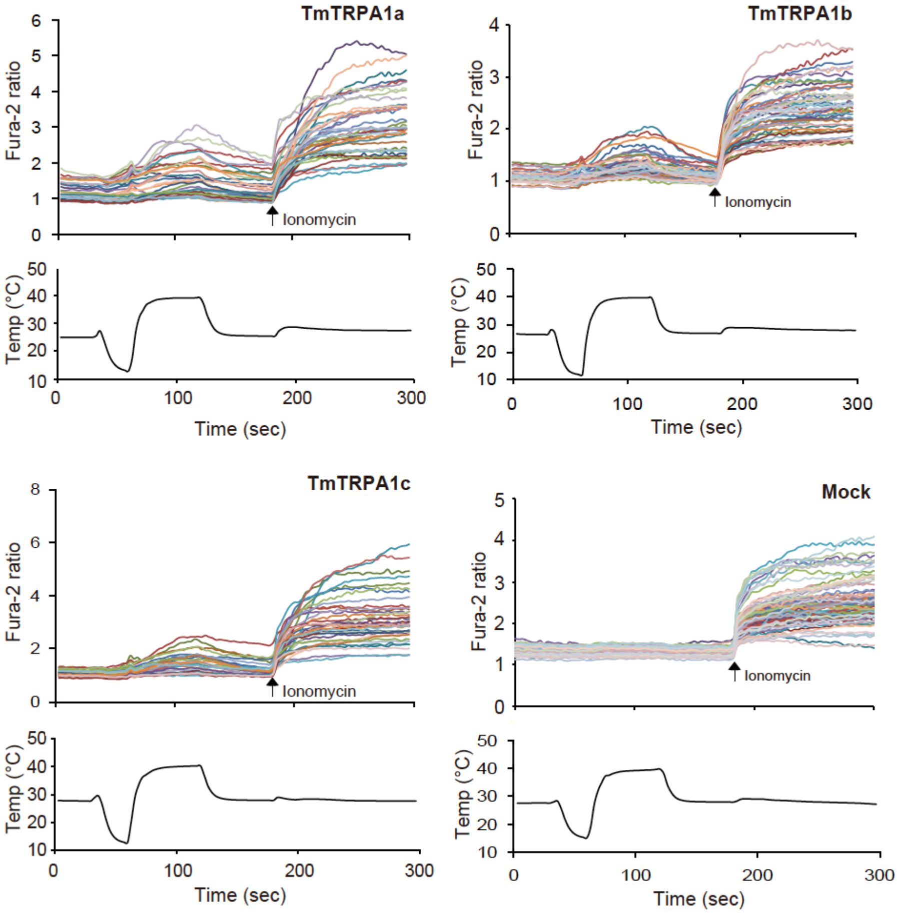


Figure 10.3 Many plant oil-derived tick repellents activate TmTRPA1b. Activation of TmTRPA1b by representative eight plant oil-derived tick repellents analyzed by calcium imaging. Red bars show the period when each compound was added, and then washed off. Arrows show the time points when ionomycin was added. Chemical structure of each compound is also shown. The concentration of each compound was 1 mM except for geranylacetone, nerol, and carvacrol (0.5 mM).

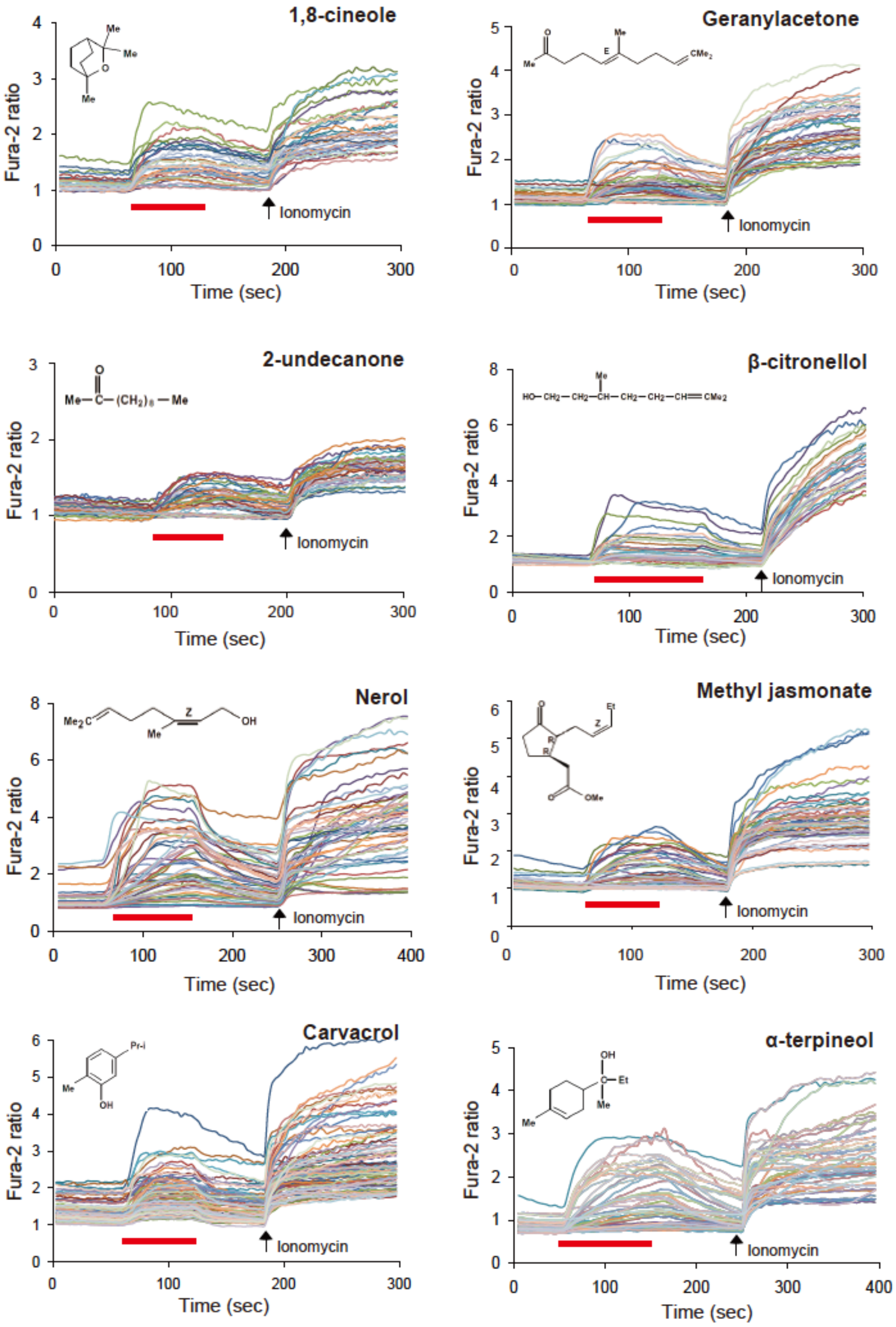


Figure 10.4 Conserved TmTRPA1b activation by electrophilic compounds. Activation of TmTRPA1b by three electrophiles, allyl isothiocyanate, cinnamaldehyde, and dially disulfide analyzed by calcium imaging. Red bars show the period when each compound was added, and then washed off. Arrows show the time points when ionomycin was added. Chemical structure of each compound is also shown. The concentration of each compound was 1 mM.

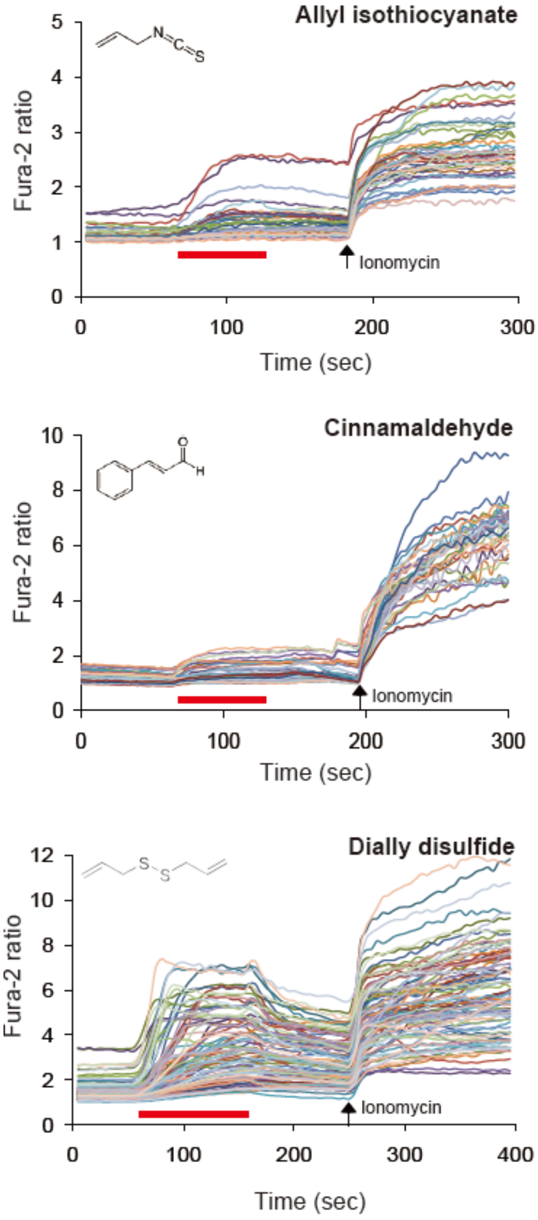
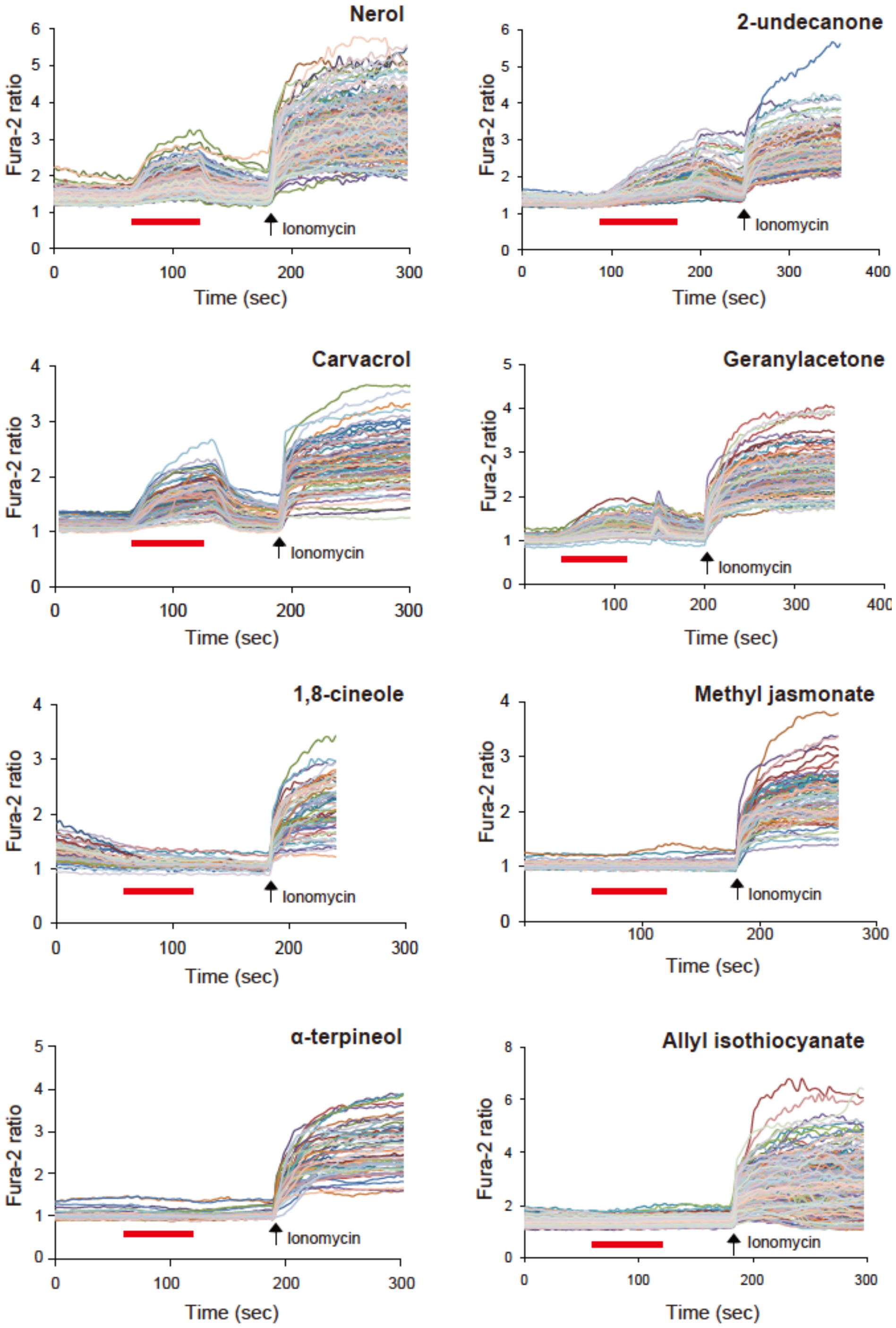


Figure 10.5 Responses of TmTRPA1a to TmTRPA1b/c activating compounds. Responses of TmTRPA1a to representative eight TmTRPA1b/c activating compounds analyzed by calcium imaging. Red bars show the period when each compound was added, and then washed off. Arrows show the time points when ionomycin was added. The concentration of each compound was 1 mM except for carvacrol and geranylacetone (0.5 mM).



Conclusions

Comparison of the small mite genome with other arthropods has provided new insights into the evolution of Parasitiformes and the evolutionary changes associated with specific habitats and life history of honey bee ectoparasitic mite that could potentially improve the control programs of small mite. The study of TmTRPA1 has developed a novel control method for small mite.

References

Abbas, M.M., Qutaibah M.M., Ponnuraman B. (2014) Assessment of de novo assemblers for draft genomes: a case study with fungal genomes. BMC genomics 15, 1-12.

Abuin L., Bargeton B., Ulbrich M.H., Isacoff E.Y., Kellenberger S., Benton R. (2011) Functional architecture of olfactory ionotropic glutamate receptors. Neuron 69, 44-60.

Adamo S. A. (1998) The specificity of behavioral fever in the cricket Acheta domesticus. Journal of Parasitology 84, 529-33.

Allocati N., Federici L., Masulli M., Di Ilio C., (2009) Glutathione transferases in bacteria. The FEBS journal. 276, 58-75.

Ai J., Zhu Y., Duan J., Yu Q., Zhang G., Wan F., Xiang Z.H. (2011) Genome-wide analysis of cytochrome P450 monooxygenase genes in the silkworm, Bombyx mori. Gene 480, 42-50.

Anders S., Pyl P.T., Huber W. (2015) HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-9.

Anderson, D.L., Morgan, M.J. (2007) Genetic and morphological variation of bee-parasitic Tropilaelaps mites (Acari: Laelapidae): new and re-defined species, Experimental and Applied Acarology. 43, 1-24.

Anderson, D.L., Morgan, M.J. (2013) Standard methods for Tropilaelaps mites research, Journal of Apicultural Research 52:21.

Andersson M.N., Grosse-Wilde E., Keeling C.I., Bengtsson J.M., Yuen M.M., Li M., Hillbur Y., Bohlmann J., Hansson B.S., Schlyter F. (2013) Antennal transcriptome analysis of the chemosensory gene families in the tree killing bark beetles, Ips typographus and Dendroctonus ponderosae (Coleoptera: Curculionidae: Scolytinae) BMC Genomics 21, 14-198.

Arisue N., Hirai M., Arai M., Matsuoka H., Horii T. (2007) Phylogeny and evolution of the SERA multigene family in the Genus Plasmodium. Journal of Molecular Evolution 65, 82-91.

Atwal, A.A., Goyal. N.P. (1971) Infestation of honey bee colonies with Tropilaelaps, and its control. Journal of Apicultural Research 10, 137-142.

Cires, E., Cuesta, C., Fernández prieto, J. A. (2011) Notes on genome size in the hybrid Ranunculus x luizetii (Ranunculaceae) and its parents by flow cytometry, Collectanea Botanica.

30, 97-99.

Baldwin W., Marko P.B., Nelson D. (2009) The cytochrome P450 (CYP) gene superfamily in Daphnia pulex. BMC Genomics 10, 169-181.

Babin P.J., Bogerd J., Kooiman F.P., Van Marrewijk W.J., Van der Horst D.J. (1999) Apolipophorin II/I, apolipoprotein B, vitellogenin, and microsomal triglyceride transfer protein genes are derived from a common ancestor. Journal of Molecular Evolution 49,150-60.

Bao Z. and Eddy S.R. (2002) Automated de novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Research 12, 1269-1276.

Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27, 573-580.

Bell L. R., Maine E. M., Schedl P., Cline T. W. (1988). Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell 55, 1037-1046.

Bellefroid E. J., Leclere L., Saulnier A., Keruzore M., Sirakov M., Vervoort M., De Clercq S. (2013). Expanding roles for the evolutionarily conserved Dmrt sex transcriptional regulators during embryogenesis. Cellular and Molecular Life Sciences 70, 3829-3845.

Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289-300.

Benton R., Vannice K.S., Gomez-Diaz C., Vosshall L.B. (2009) Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. Cell 136, 149-162.

Bennett, M.D., Leitch, I.J., Price, H.J., Johnston, J.S. (2003) Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb, Annals of biology 91, 547-57.

Bennett M.D., Price H.J., Johnston J.S. (2008) Anthocyanin inhibits propidium iodide DNA fluorescence in Euphorbia pulcherrima: implications for genome size variation and flow cytometry. Annals of Botany 101, 777-790.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19, 455-477.

Becalska A.N., Gavis E.R. (2009) Lighting up mRNA localization in Drosophila oogenesis. Development 136, 2493-2503.

Benton R., Sachse S., Michnick S.W., Vosshall L.B. (2006) Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo. PLoS Biology 4, 240-257.

Benton R., Vannice K.S., Gomez-Diaz C., Vosshall L.B. (2009) Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. Cell 136, 149-162.

Benoit J.B., Lopez-Martinez G., Patrick K.R., Phillips Z.P., Krause T.B., Denlinger D.L. (2011) Drinking a hot blood meal elicits a protective heat shock response in mosquitoes. PNAS 08, 8026-8029.

Bijlsma R., Bundgaard J., Boerema A.C. (1997) Genetic and environmental stress and the persistence of populations. Pages 193-207 in Bijlsma R., Loeschcke V., eds. Environmental Stress, Adaptation and Evolution. Basel (Switzerland): Birkhauser.

Bissinger B. W., Roe R. M. (2010) Tick repellents: Past, present, and future. Pesticide Biochemistry and Physiology 96, 63-79.

Bissinger B. W., Apperson, C. S., Sonenshine,D. E., Watson D. W., Roe R. M. (2009) Efficacy of the new repellent BioUD(A (R)) against three species of ixodid ticks. Experimental and Applied Acarology 48, 239-250.

Bosch J.A., Sumabat T.M., Hafezi Y., Pellock B.J., Gandhi K.D., Hariharan I.K. (2014) The Drosophila F-box protein Fbxl7 binds to the protocadherin Fat and regulates Dachs localization and Hippo signaling 3:e03383.

Bourgon R., Gentleman R., Huber W. (2010) Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences 107, 9546-9551.

Bowen-Walker P.L., Martin S.J., Gunn A. (1999) The transmission of deformed wing virus between honey bees (Apis mellifera L.) by the ectoparasitic mite Varroa jacobsoni Oud. Journal of Invertebrate Pathology 73, 101-106.

Bradfield J.Y., Lee Y.H., Keeley L.L. (1991) Cytochrome P450 family 4 in a cockroach: molecular cloning and regulation by regulation by hypertrehalosemic hormone. Proceedings of the National Academy of Sciences 88, 4558-4562.

Bradnam K.R. and Korf I. (2008) Longer first introns are a general property of eukaryotic gene structure. PLoS ONE 3, e3093.

Bradnam K.R., Fass J.N. et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species, GigaScience 2:10.

Busby A.T., Ayllón N., Kocan K.M., Blouin E.F., de la Fuente G., Galindo R.C., Villar M., de la Fuente J. (2012) Expression of heat shock proteins and subolesin affects stress responses, Anaplasma phagocytophilum infection and questing behaviour in the tick, Ixodes scapularis. Medical and Veterinary Entomology 26, 92-102.

Burmester T., Scheller K. (1999) Ligands and receptors: common theme in insect storage protein transport. Naturwissenschaften 86, 468-474.

Butler J., MacCallum I., Kleber M., Shlyakhter I.A., Belmonte M.K., Lander E.S., Nusbaum C., Jaffe D.B. (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Reseash. 18, 810-820.

Cantarel B.L., Korf I., Robb S.M., Parra G., Ross E., Moore B., Holt C., Sánchez Alvarado A., Yandell M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18, 188-96.

Cao, T. N. P. (2014) Genome Annotation and Evolution of Chemosensory Receptors in Spider Mites. Ghent, Belgium: Ghent University. Faculty of Sciences.

Carey, A.F. and Carlson, J.R. (2011) Insect olfaction from model systems to disease control. Proceedings of the National Academy of Sciences 108, 12987-12995.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17, 540-552.

Chaisson K.E., Hallem E.A. (2012) Hallem Chemosensory behaviors of parasites, Trends in Parasitology 28, 427-36.

Chase B. A., Baker B. S. (1995). A genetic analysis of intersex, a gene regulating sexual differentiation in Drosophila melanogaster females. Genetics 139, 1649-1661.

Chazotte, B. (2011) Labeling membrane glycoproteins or glycolipids with fluorescent wheat germ agglutinin. Cold Spring Harbor protocols pdb prot5623.

Chipman A.D., Ferrier D.E., Brena C., Qu J., Hughes D.S., Schröder R., Torres-Oliva M., Znassi N., Jiang H., Almeida F.C., Alonso C.R., Apostolou Z., Aqrawi P., Arthur W., Barna J.C., Blankenburg K.P., Brites D., Capella-Gutiérrez S., Coyle M., Dearden P.K., Du Pasquier L., Duncan E.J., Ebert D., Eibner C., Erikson G., Evans P.D., Extavour C.G., Francisco L., Gabaldón T., Gillis W.J., Goodwin-Horn E.A., Green J.E., Griffiths-Jones S., Grimmelikhuijzen C.J., Gubbala S., Guigó R., Han Y., Hauser F., Havlak P., Hayden L., Helbing S., Holder M., Hui J.H., Hunn J.P., Hunnekuhl V.S., Jackson L., Javaid M., Jhangiani S.N., Jiggins F.M., Jones T.E., Kaiser T.S,. Kalra D., Kenny N.J., Korchina V., Kovar C.L., Kraus F.B., Lapraz F., Lee S.L., Lv J., Mandapat C., Manning G., Mariotti M., Mata R., Mathew T., Neumann T., Newsham I., Ngo D.N., Ninova M., Okwuonu G., Ongeri F., Palmer W.J., Patil S., Patraquim P., Pham C., Pu L.L., Putman N.H., Rabouille C., Ramos O.M., Rhodes A.C., Robertson H.E., Robertson H.M., Ronshaugen M., Rozas J., Saada N., Sánchez-Gracia A., Scherer S.E., Schurko A.M., Siggens K.W., Simmons D., Stief A., Stolle E., Telford M.J., Tessmar-Raible K., Thornton R., van der Zee M., von Haeseler A., Williams J.M., Willis J.H., Wu Y., Zou X., Lawson D., Muzny D.M., Worley K.C., Gibbs R.A., Akam M., Richards S. (2014) The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede Strigamia maritima. PLoS Biology 12:e1002005.

Chia N., Guttenberg N. (2011) Dynamics of gene duplication and transposons in microbial genomes following a sudden environmental change. Mobile Genetic Elements 1, 221-224.

Chen F.C., Chen C.J., Li W.H., Chuang T.J. (2010) Gene Family Size Conservation Is a Good Indicator of Evolutionary Rates. Molecular Biology and Evolution 27, 1750-1758.

Clyne P., Warr C., Freeman M., Lessing D., Kim J.C. Carlson J.R. (1999) A novel family of divergent seven-transmembrane proteins: candidate odorant receptors in Drosophila. Neuron

22, 327-338.

Coenye T., Goris J., Spilker T., Vandamme P., LiPuma J.J. (2002) Characterization of unusual bacteria isolated from respiratory secretions of cystic fibrosis patients and description of Inquilinus limosus gen. nov., sp. nov. Journal of Clinical Microbiology 40, 2062-2069.

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

Cordaux R., Paces-Fessy M., Raimond M., Michel-Salzat A., Zimmer M., Bouchon D. (2007) Molecular characterization and evolution of arthropod pathogenic Rickettsiella bacteria. Appl Environm Microbiol 73, 5045-5047.

Cornman, R.S.; Schatz, M.C., Johnston, J.S., Chen, Y-P, Pettis, J., Hunt, G., Bourgeois, L, elsik, C., Anderson, D. (2010) Genomic survey of the ectoparasitic mite Varroa destructor, a major pest of the honey bee Apis mellifera, BMC Genomics 11, 602.

Coulson, A. (1996) The Caenorhabditis elegans genome project. Biochemical Society Transactions 24, 289.

Crane E. (1988) Africanized bee, and mites parasitic on bees, in relation to world beekeeping. In: Needham GR, Page RE Jr, Delfinado-Baker M, Bowman CE editor, Africanized honey bees and bee mites. Ellis Horwood, Chichester, UK., pp 1-9.

Croset V., Rytz R., Cummins S.F., Budd A., Brawand D., Kaessmann H., Gibson T.J., Benton R. (2010) Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genetics 6:e1001064.

Cruz M. D. S., Robles M. C. V., Jespersen J. B., Kilpinen O., Birkett M., Dewhirst S., Pickett J. (2005) Scanning electron microscopy of foreleg tarsal sense organs of the poultry red mite, Dermanyssus gallinae (DeGeer) (Acari : Dermanyssidae). Micron 36, 415-421.

Curwen V., Eyras E., Andrews T.D., Clarke L., Mongin E., Searle S.M., Clamp M. (2004) The Ensembl automatic gene annotation system. Genome Research 14, 942-950.

Dainat B, Tan K, Berthoud H, Neumann P. (2009) The ectoparasitic mite Tropilaelaps mercedesae (Acari, Laelapidae) as a vector of honey bee viruses. Insectes Sociaux 56: 40-43.

Dabert M., Witalinski W., Kazmierski A., Olszanowski Z., Dabert J. (2010) Molecular phylogeny of acariform mites (Acari, Arachnida): Strong conflict between phylogenetic signal and long-branch attraction artifacts. Molecular Phylogenetics and Evolution 56, 222-41.

Damann N., Voets T., Nilius, B. 2008 TRPs in our senses. Current Biology 18, R880-R889.

Darriba D., Taboada G.L., Doallo R., Posada D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27,1164-1165.

Davis RL. 1996. Physiology and biochemistry of Drosophila learning mutants. Physiological Reviews 76, 299 -317.

Davies L., Williams D.R., Aguiar-Santana I.A., Pedersen J., Turner P.C., Rees H.H. (2006) Expression and down-regulation of cytochrome P450 genes of the CYP4 family by ecdysteroid agonists in Spodoptera littoralis and Drosophila melanogaster. Insect Biochemistry and Molecular Biology 36, 801-807.

de Belle J.S., Heisenberg M. (1994) Associative odor learning in Drosophila abolished by chemical ablation of mushroom bodies. Science 263, 692- 695.

De Bie T., Cristianini N., Demuth J.P., Hahn M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22, 1269-1271.

Denslow N.D., Colbourne J.K., Dix D., Freedman J., Helbing C., Kennedy S., Williams P. (2006) Selection of surrogate animal species for comparative toxicogenomics. In Genomic Approaches for Cross-Species Extrapolation in Toxicology. Edited by Benson W.H., Di Giulio R.T. Portland, Oregon, USA: CRC Press, Chapter 4: 103-142.

Dermauw W., Van Leeuwen T., Vanholme B., Tirry L. (2009) The complete mitochondrial genome of the house dust mite Dermatophagoides pteronyssinus (Trouessart): a novel gene arrangement among arthropods. BMC Genomics 10: 107.

de Miranda J.R. and Fries I. (2008). Venereal and vertical transmission of deformed wing virus in honey bees (Apis mellifera L.) Journal of Invertebrate Pathology 98, 184-189.

de Jong D., Morse, R.A. & Eickwort, G.C. (1982). Mite pests of honey bees. Annual Review of Entomology 27, 229-252.

de Bruyne M., Baker T.C. (2008) Odor detection in insects: volatile codes. Journal of Chemical

Ecology 34, 882-897.

Dutky, S.R., Gooden E.L. (1952). Coxiella popilliae, n. sp., a Rickettsia causing blue disease of Japanese beetle larvae. Journal of Bacteriology 63, 743-750.

Do H.H., Choi K.P., Preparata F.P., Sung W.K., Zhang L. (2008) Spectrum-Based De Novo Repeat Detection in Genomic Sequences, Journal of Computational Biology 15, 469-487.

Dodson GS, Guarnieri DJ, Simon MA. 1998. Src64 is required for ovarian ring canal morphogenesis during Drosophila oogenesis. Development 125, 2883-2892.

Duscher G.G., Galindo R.C., Tichy A., Hummel K., Kocan K.M., de la Fuente J. (2014) Glutathione S-transferase affects permethrin detoxification in the brown dog tick, Rhipicephalus sanguineus. Ticks and Tick-borne Diseases 5, 225-233.

Engel P., Moran N. A. (2013) The gut microbiota of insects - diversity in structure and function. FEMS Microbiology Reviews 37, 699-735.

Ekblom R., Wolf J.B. (2014) A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 7, 1026-42.

Emerson J. J., Cardoso-Moreira M., Borevitz J.O., Long, M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320, 1629-1631.

Enayati A., Ranson H., Hemingway J. (2005) Insect glutathione transferases and insecticide resistance. Insect Molecular Biology 14, 3-8.

Erdelyan C.N., Mahood T.H., Bader T.S., Whyard S. (2012) Functional validation of the carbon dioxide receptor genes in Aedes aegypti mosquitoes using RNA interference. Insect Molecular Biology 21, 119-27.

Erickson J.W., Quintero J.J. (2007) Indirect effects of ploidy suggest X chromosome dose, not the X:A ratio, signals sex in Drosophila. PLoS Biol 5: e332.

Eriksson B.J., Fredman D., Steiner G., Schmid A. (2013) Characterisation and localisation of the opsin protein repertoire in the brain and retinas of a spider and an onychophoran. BMC Evolutionary Biology 13:186.

El-Metwally S., Hamza T., Zakaria M., Helmy M. Next-Generation Sequence Assembly: Four

Stages of Data Processing and Computational Challenges, PLOS Computational Biology 9: e1003345.

Falcon S., Gentleman R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics 23, 257-8.

Fang H. (2014) dcGOR: an R package for analysing ontologies and protein domain annotations. PLOS Computational Biology 10:e1003929.

Feyereisen R. (1999) Insect P450 enzymes, Annu. The Annual Review of Entomology 44, 507-533.

Feyereisen, R. (2011). Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Biochimica et Biophysica Acta 1814, 19-28.

Feyereisen R. (2012) Insect CYP genes and P450 enzymes. In: Gilbert L.I. editor, Insect Molecular Biology and Biochemistry. Academic Press: London, Chapter 8, 236-316.

Forsgren E., De miranda J.R., Isaksson M., Wei S., Fries I. (2009) Deformed wing virus associated with Tropilaelaps mercedesae infesting European honey bees (Apis mellifera). Experimental and Applied Acarology 47, 87-97.

Fitch W. M. (2000) Homology a personal view on some of the problems. Trends in Genetics 16, 227-231.

Gao J, Scott J.G. (2006) Use of quantitative real-time polymerase chain reaction to estimate the size of the house-fly Musca domestica genome. Insect Molecular Biology 15, 835-837.

Gao M., Skolnick J. (2009) A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLOS Computational Biology 5:e1000567.

Garrett-Engele, C. M., Siegal, M. L., Manoli, D. S., Williams, B. C., Li, H., & Baker, B. S. (2002). Intersex, a gene required for female sexual development in Drosophila, is expressed in both sexes and functions together with doublesex to regulate terminal differentiation. Development 129, 4661-4675.

Gawande N.D., Subashini S., Murugan M., Subbarayalu M. (2014) Molecular screening of insecticides with sigma glutathione S-transferases (GST) in cotton aphid Aphis gossypii using docking. Bioinformation 10, 679-683.

Gempe T., Beye M. (2011) Function and evolution of sex determination mechanisms, genes and pathways in insects. Bioessays 33, 52-60.

Gibson G.G., Skett P. (2001). Introduction to Drug Metabolism, second edition. Blackie Academic & Professional, An Imprint of Chapman & Hall, London, U.K. Chapter 1.

Glenn T.C. (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 5, 759-769.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644-652.

Grozinger C.M., Evans J.D. (2010) Genomic survey of the ectoparasitic mite Varroa destructor, a major pest of the honey bee Apis mellifera. BMC Genomics 11, 1471-2164.

Grbić M., Van Leeuwen T., Clark R. M., Rombauts S., Rouzé P., Grbić V., Osborne E. J., Dermauw W., Ngoc P. C., Ortego F., Hernández-Crespo P., Diaz I., Martinez M., Navajas M., Sucena É., Magalhães S., Nagy L., Pace R. M., Djuranović S., Smagghe G., Iga M., Christiaens O., Veenstra J. A., Ewer J., Villalobos R. M., Hutter J. L., Hudson S. D., Velez M., Yi S. V., Zeng J., Pires-daSilva A., Roch F., Cazaux M., Navarro M., Zhurov V., Acevedo G., Bjelica A., Fawcett J. A., Bonnet E., Martens C., Baele G., Wissler L., Sanchez-Rodriguez A., Tirry L., Blais C., Demeestere K., Henz S. R., Gregory T. R., Mathieu J., Verdon L., Farinelli L., Schmutz J., Lindquist E., Feyereisen R., Van de Peer Y.. (2011) The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature 479, 487-92.

Geraci N.S., Johnston J.S., Robinson J.P., Wikel S.K., Hill C.A. (2007) Variation in genome size of argasid and ixodid ticks. Insect Biochemistry and Molecular Biology 37, 399-408.

Gu S.H., Wang S.Y., Zhang X.Y., Ji P., Liu J.T., Wang G.R., Wu K.M., Guo Y.Y., Zhou J.J., Zhang Y.J. (2012) Functional characterizations of chemosensory proteins of the alfalfa plant bug Adelphocoris lineolatus indicate their involvement in host recognition. PLoS One 7:e42871.

Gu X.B., Liu G.H., Song H.Q., Liu T.Y., Yang G.Y., Zhu X.Q. (2014) The complete mitochondrial genome of the scab mite Psoroptes cuniculi (Arthropoda: Arachnida) provides insights into Acari phylogeny. Parasites & Vectors 7:340.

Gurevich A., Saveliev V., Vyahhi N., Tesler G. (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072-1075.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology 59, 307-321.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research 31, 5654-5666.

Haas W. (2003) Parasitic worms: strategies of host finding, recognition and invasion. Zoology (Jena) 106, 349-64.

Haas B.J., Salzberg S.L., Zhu W., Pertea M., Allen J.E., Orvis J., White O., Buell C.R., Wortman J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology 9: R7.

Hara T., Hara R. (1967) Rhodopsin and retinochrome in the squid retina. Nature 214, 573-575.

Hara T., Hara R. (1968) Regeneration of squid retinochrome. Nature 219, 450-454.

Hayes J.D., Flanagan J.U., Jowsey I.R. (2005) Glutathione transferases. Annual Review of Pharmacology and Toxicology 45, 51-88.

Heisenberg M. 1998. What do the mushroom bodies do for the insect brain? An introduction. Learning & Memory 5, 1-10.

Holt C. and Yandell M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.

Hoffman AA, Parsons PA. 1991. Evolutionary genetics and environmental stress. University Press, Oxford, Oxford, U.K.

Huelsenbeck J. P., Crandall K. A. (1997) Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood. Annual Review of Ecology and Systematics 28, 437-466.

Hunt M., Gall A., Ong S.H., Brener J., Ferns B., Goulder P., Nastouli E., Keane J.A., Kellam P., Otto T.D. (2015) IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics pii:

btv120.

Hittinger G. H., Carroll S. B. (2007) Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449, 677-681.

Iga M., Kataoka, H. (2012) Recent studies on insect hormone metabolic pathways mediated by cytochrome P450 enzymes. Biological and Pharmaceutical Bulletin 35, 838-843.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

Ionescu C., Caira M. R. (2005). Drug Metabolism: Current Concepts. Dordrecht: Springer, U.K., Chapter 2.

Ide S., Miyazaki T., Maki H., Kobayashi T. (2010) Abundance of ribosomal RNA gene copies maintains genome integrity. Science 327, 693-696.

Janecke A. R., Thompson D. A., Utermann G., Becker C., Hubner C. A., Schmid E., McHenry C. L., Nair A. R., Ruschendorf F., Heckenlively J., Wissinger B., Nurnberg P., and Gal A. (2004) Mutations in RDH12 encoding a photoreceptor cell retinol dehydrogenase cause childhood-onset severe retinal dystrophy. Nature genetics 36, 850-854.

Jeyaprakash A., Hoy M.A. (2009) The nuclear genome of the phytoseiid Metaseiulus occidentalis (Acari: Phytoseiidae) is among the smallest known in arthropods. Experimental and Applied Acarology 47, 263-273.

Johnston J.S., Yoon K.S., Strycharz J.P., Pittendrigh B.R., Clark J.M. (2007) Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. Journal of Medical Entomology 44, 1009-1012.

Jones, W.D., Cayirlioglu P., Kadow I.G., Vosshall L.B. (2007) Two chemosensory receptors. together mediate carbon dioxide detection in Drosophila. Nature 4, 86-90.

Kang J.S., Koh Y.H., Moon Y.S., Lee S.H. (2012) Molecular properties of a venom allergen-like protein suggest a parasitic function in the pinewood nematode Bursaphelenchus xylophilus. International Journal for Parasitology 42, 63-70.

Kang K., Panzano V. C., Chang E. C., Ni L., Dainis A. M., Jenkins A. M., Regna K., Muskavitch M. A. T., Garrity P. A. (2012) Modulation of TRPA1 thermal sensitivity enables sensory discrimination in Drosophila. Nature 481, 76-82.

Kato Y., Kobayashi K., Oda S., Tatarazako N., Watanabe H., Iguchi T. (2010) Sequence divergence and expression of a transformer gene in the branchiopod crustacean, Daphnia magna. Genomics 95, 160-5.

Katoh K., Standley D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30, 772-80.

Kandul N.P., Noor M.AF., (2009) Large introns in relation to alternative splicing and gene evolution: a case study of Drosophila bruno-3. BMC Genetics, 10:67.

Kanehisa M., Goto S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27-30.

Kelley DR, Salzberg SL. 2010. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biology 11: R28.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L., 2010 Quake: quality-aware detection and correction of sequencing errors. Genome Biology 11:R116.

Kumar S., Jones, M. Koutsovoulos, G. Clarke, M. Blaxter, M. (2013) Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. Frontiers in Genetics 4:273.

Kelley D.R. and Salzberg S.L. (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biology 11:R28

Kitabayashi AN, Arai T, Kubo T, Natori S (1998) Molecular cloning of cDNA for p10, a novel protein that increases in the regenerating legs of Periplaneta americana (American cockroach). Insect Biochemistry and Molecular Biology 28, 785-790.

Kim J.H., Roh J.Y., Kwon D.H., Kim Y.H., Yoon K.A., Yoo S., Noh S., Park J., Shin E., Park M., Si Leecorresponding H. (2014) Estimation of the genome sizes of the chigger mites Leptotrombidium pallidum and Leptotrombidium scutellare based on quantitative PCR and k-mer analysis. Parasit Vectors 7:279.

Kimura K. (2011) Role of cell death in the formation of sexual dimorphism in the Drosophila central nervous system. Development Growth and Differentiation 53, 236-244.

Korf I. (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59.

Koyanagi M., Takano K., Tsukamoto H., Ohtsu K., Tokunaga F., Terakita A. (2008) Jelly**W**sh vision starts with cAMP signaling mediated by opsin-Gs cascade. Proceedings of the National Academy of Sciences 105, 15576-15580.

Kohno K., Sokabe T., Tominaga M., Kadowaki T. (2010) Honey Bee Thermal/Chemical Sensor, AmHsTRPA, Reveals Neofunctionalization and Loss of Transient Receptor Potential Channel Genes. Journal of Neuroscience 30, 12219-12229.

Krogh A., Larsson B., von Heijne G., Sonnhammer E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology 305, 567-580.

Lanzi G., de Miranda J.R., Boniotti M.B., Cameron C.E., Lavazza A., Capucci L., Camazine S.M., Rossi C. (2006) Molecular and biological characterization of deformed wing virus of honey bees (Apis mellifera L.). Journal of Virology 80, 4998-5009.

Lambert, D.M., (1995). The new science of molecular ecology. New Zealand Journal of Ecology 9, 93-96.

Langmead B., Cole T., Mihai P., Steven L.S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.

Langmead, B., and Salzberg, S.L. (2012) Fastgapped-read alignment with Bowtie2. Natuare Methods 9, 357-359.

Lassmann T., Sonnhammer E.L. (2005) Kalign - an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6:298.

Lassmann T., Frings O., Sonnhammer E.L. (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic acids research 37, 858-65.

Lehane, M. J. (2005) The Biology of Blood-sucking Insects. cambridge university Press, cambridge, U.K. Chapter 4, pp 27 -52.

Lechner M., Findeiß S., Steiner L., Marz M., Stadler P.F., Prohaska S.J. (2011) Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis. BMC Bioinformatics 28: 124.

Leclerque A., Kleespies R.G. (2012) A Rickettsiella bacterium from the hard tick, Ixodes woodi: molecular taxonomy combining multilocus sequence typing (MLST) with significance testing. PLoS One 7:e38062.

Lee Y. J., Shah S., Suzuki E., Zars T., O'Day P. M., Hyde D. R. (1994) The Drosophila dgq gene encodes a G alpha protein that mediates phototransduction. Neuron 13, 1143-1157.

Lee Y., Montell C. (2013) Drosophila TRPA1 Functions in Temperature Control of Circadian Rhythm in Pacemaker Neurons. Journal of Neuroscience 33, 6716-6725.

Lewis E.E., Campbell J.F., Sukhdeo M.V.K. (2002) Parasite Behavioural Ecology in a Field of Diverse Perspectives. In: Lewis, E.E., Campbell J.F., Sukhdeo M.V.K. editor, The Behavioural Ecology of Parasites. CABI Publishing, London, UK., Chapter 16, 337-346.

Liao C.Y., Zhang K., Niu J.Z., Ding T.B., Zhong R., Xia W.K., Dou W., Wang J.J. (2013) Identification and characterization of seven glutathione S-transferase genes from citrus red mite, Panonychus citri (McGregor). International Journal of Molecular Sciences 14, 24255-24270.

Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews 46, 3-26.

Liska, D. J. (1998). The detoxification enzyme system. Alternative Medicine Review 3, 187-198.

Li X., Duan X., Jiang H., Sun Y., Tang Y., Yuan Z., Guo J., Liang W., Chen L., Yin J., Ma H., Wang J., Zhang D. (2006) Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. Plant Physiology 141, 1167-1184.

Li B., Dewey C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics 12:323.

Li F., Fan G., Wang K., Sun F., Yuan Y., Song G., Li Q., Ma Z., Lu C., Zou C., Chen W., Liang X., Shang H., Liu W., Shi C., Xiao G., Gou C., Ye W., Xu X., Zhang X., Wei H., Li Z., Zhang G., Wang J., Liu K., Kohel R.J., Percy R.G., Yu J.Z., Zhu Y.X., Wang J., Yu S. (2014) Genome sequence of the cultivated cotton Gossypium arboreum. Nature Genetic 46, 567-572.

Li G, Zhang J, Tong X, Liu W, Ye X. (2011) Heat shock protein 70 inhibits the activity of

Influenza A virus ribonucleoprotein and blocks the replication of virus in vitro and in vivo. PLOS ONE 6: e16546.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25, 2078-9.

Li L., Stoeckert CJ. Jr., Roos D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13, 2178-2189.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y et al, (2012) The sequence and de novo assembly of the giant panda genome. Nature 463, 311-317.

Li X.C., Schuler M.A., Berenbaum M.R., (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annual Review of Entomology 52, 231-253.

Liu L., Wolf R., Ernst R., Heisenberg M. (1999) Context generalization in Drosophila visual learning requires the mushroom bodies. Nature 400, 753-756.

Liu J., Xiao H., Huang S., Li F. (2014) OMIGA: Optimized Maker−Based Insect Genome Annotation, Molecular Genetics and Genomics 289, 567-573.

Li X., Schuler M.A., Berenbaum M.R. (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annual Review of Entomology 52, 231-253.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

Luo, Q.H. (2011) Ph.D. Thesis: Study on the neutral strains, epidemiology and parasitology of Tropilaelaps mites (Acari: Laelipdea) in China.

Loman N.J., Misra R.V., Dallman T.J., Constantinidou C., Gharbia S.E., et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology 30, 434-439.

Lowe T.M., Eddy S.R. (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Research 25, 955-964.

Lukashin A. V., Borodovsky M. (1998) GeneMark.hmm: new solutions for gene finding.

Nucleic Acids Research 26, 1107-1115.

Marcais G. and Kingsford C., (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764-770.

Marioni J.C., Mason C.E., Mane S.M., Stephens M., Gilad Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Reserach 18, 1509-1517.

Mardan M., Kevan P.G. (2002) Critical temperatures for survival of brood and adult workers of the giant honeybee, Apis dorsata (Hymenoptera: Apidae). Apidologie 33, 295-301.

Martin J.A., Wang Z. (2011) Next-generation transcriptome assembly. Nature Reviews Genetics 12, 671-682.

Matsuo Y., Akiyama N., Nakamura H., Yodoi J., Noda M., Kizaka-Kondoh S. (2001) Identification of a Novel Thioredoxin-related Transmembrane Protein. The Journal of Biological Chemistry 276, 10032-10038.

McKew B.A., Smith C.J. (2010) Real-Time PCR Approaches for analysis of hydrocarbon-degrading bacterial communities. In: Timmis KN, editor. Handbook of hydrocarbon and lipid microbiology. Springer-Verlag Berlin Heidelberg, 3995-4009.

McBride S.M., Giuliani G., Choi C., Krause P., Correale D., Watson K., Baker G., Siwicki K.K. (1999) Mushroom body ablation impairs short-term memory and long-term memory of courtship conditioning in Drosophila melanogaster. Neuron 24, 967-977.

Metzker M.L. (2010) Sequencing technologies - the next generation. Nature Reviews Genetics 11, 31-46.

Miller J.M., Malenfant R.M., Moore S.S., Coltman D.W. (2012) Short reads, circular genome: skimming solid sequence to construct the bighorn sheep mitochondrial genome. Journal of Heredity 103, 140-146.

Miller J.R., Koren S., Sutton G. (2010) Assembly algorithms for next-generation sequencing data. Genomics 95, 315-327.

Mikita S., David T., Peer B. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research 34, W609-W612.

Moeller, A. H., Y. Li, E. Mpoudi Ngole, S. Ahuka-Mundeke, E. V. Lonsdorf, A. E. Pusey, M. Peeters, B. H. Hahn, and H. Ochman (2014). Rapid changes in the gut microbiome during human evolution. Proceedings of the National Academy of Sciences. 111, 16431-16435.

Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nataure Methods 5, 621-628.

Mukai M., Kato H., Hira S., Nakamura K., Kita H., Kobayashi S. (2011) Innexin2 gap junctions in somatic support cells are required for cyst formation and for egg chamber formation in Drosophila. Mechanisms of Development 128, 510-523.

Mulla M.Y., Tuccori E., Magliulo M., Lattanzi G., Palazzo G., Persaud K., Torsi L. (2015) Capacitance-modulated transistor detects odorant binding protein chiral interactions. Nature Communication 16, 6006-6010.

Nation J.L. (2008) Insect Physiology and Biochemistry, Second Edition. CRC Press, Boca Raton.

Nagarkatti, P.S., Nagarkatti M. (1987) Immunotoxicology: Modulation of the immune system by xenobiotics, Defence Science Journal 37, 235-244.

Nawrocki E. P., Eddy S. R. (2013) Infernal 1.1: 100-fold faster RNA homology searches, Bioinformatics 29, 2933-2935.

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., Snyder M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344-1349.

Nagata T., Koyanagi M., Tsukamoto H., Terakita A. (2010) Identification and characterization of a protostome homologue of peropsin from a jumping spider. Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology 196, 51-59.

Nelson D.R. (1998) Metazoan cytochrome P450 evolution, Comparative Biochemistry and Physiology 121, 15-22.

Nebert D.W., Russell D.W. (2002) Clinical importance of the cytochromes P450. The Lancet 360, 1155-1162.

Nelson-Rees W. A., Hoy M. A., Roush R. T. (1980) Heterochromatization, chromatin elimination and haploidization in the parahaploid mite Metaseiulus occidentalis (Nesbitt) (Acarina: Phytoseiidae). Chromosoma 77, 263-276.

Nelson, D. R. (2009) The cytochrome p450 homepage. Human Genomics 4, 59-65.

Nelson-Rees, W. A., Hoy, M. A., Roush, R. T. (1980). Heterochromatization, chromatin elimination and haploidization in the parahaploid mite Metaseiulus occidentalis (Nesbitt) (Acarina: Phytoseiidae). Chromosoma 77, 263-276.

Nicolai M., Lasbleiz C., Dura J.M. (2003) Gain-of-function screen identifies a role of the Src64 oncogene in Drosophila mushroom body development. Journal of Neurobiology 57, 291−302.

Nilius B., Owsianik, G. 2011 The transient receptor potential family of ion channels. Genome Biology 12:218.

Nilius B., Appendino G., Owsianik G. (2012) The transient receptor potential channel TRPA1: from gene to pathophysiology. Pflugers Archiv-European Journal of Physiology 464, 425-458.

Nozawa M., Nei M. (2007) Evolutionary dynamics of olfactory receptor genes in Drosophila species. Proceedings of the National Academy of Sciences 104, 7122-7.

Terrestrial Animal Health Code 2010, tropilaelaps infestation of honey bees, chapter 2.2.6.

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

Oldroyd BP (1999) Coevolution while you wait: Varroa jacobsoni, a new parasite of western honey bees. Trends in Ecology & Evolution 14, 312-315.

Oakeshott J.G., Johnson R.M., Berenbaum M.R., Ranson H., Cristino A.S., Claudianos C. (2010) Metabolic enzymes associated with xenobiotic and chemosensory responses in Nasonia vitripennis. Insect Molecular Biology 19, 147-163.

Ongus J.R., Peters D., Bonmatin J-M., Bengsch E., Vlak J.M., van Oers M.M. (2004) Complete sequence of a picorna-like virus of the genus Iflavirus replicating in the mite Varroa destructor. Journal of General Virology 85, 3747-3755.

Ono H., Ozaka K., Yoshikawa H. (2005) Identification of cytochrome P450 and glutathione S-transferase genes preferentially expressed in chemosensory organs of the swallowtail butterflyl, Pailio xuthus L. Insect Biochemistry and Molecular Biology 35, 837-846.

Ono H., Rewitz K.F., Shinoda T., Itoyama K., Petryk A., Rybczynski R., Jarcho M., Warren J.T., Marques G., Shimell M.J., Gilbert L.I., O'Connor M.B. (2006) Spook and Spookier code for stage-specific components of the ecdysone biosynthetic pathway in Diptera. Developmental Biology 298, 555-570.

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

Pandey T., Singh S.K., Chhetri G., Tripathi T., Singh A.K. (2015) Characterization of a Highly pH Stable Chi-Class Glutathione S-Transferase from Synechocystis PCC 6803. PLoS One 10:e0126811.

Parsons P.A., (2005) Environments and evolution: interactions between stress, resource inadequacy, and energetic efficiency. Biological reviews of the Cambridge Philosophical Society Cambridge Philosophical Society 80, 589-610.

Pagel Van Zee J., Geraci N.S., Guerrero F.D., Wikel S.K., Stuart J.J., Nene V.M., Hill C.A. (2007) Tick genomics: The Ixodes genome project and beyond, International Journal for Parasitology 37, 1297-1305.

Paulusma C. C., Geer M .A., Evers R., Heijn M., Ottenhoff R., Borst P., Oude Elferink R. P. (1999) Canalicular multispecific organic anion transporter/multidrug resistance protein 2 mediates low-affinity transport of reduced glutathione. Biochemical Journal 338, 393-401.

Paulsen C. E., Armache J. P., Gao Y., Cheng Y., Julius, D. (2015) Structure of the TRPA1 ion channel suggests regulatory mechanisms. Nature 520, 511-517.

Parra G., Bradnam K., Ning Z., Keane T., Korf I. (2009) Assessing the gene space in draft genomes. Nucleic Acids Research 37, 289-297.

Parra G., Bradnam K., Korf I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061-1067.

Pascual A., Preat T. (2001) Localization of long-term memory within the Drosophila mushroom body. Science 294, 1115-1117.

Pelosi P, Maida R. (1995) Odorant-binding proteins in insects. Comp Biochem Physiol B Biochem Journal of Molecular Biology 111, 503-514.

Peñalva-Arana D.C., Lynch M., Robertson H.M. (2009) The chemoreceptor genes of the waterflea Daphnia pulex: many Grs but no Ors. BMC Evolutionary Biology 21, 9-79.

Peng Z, Zhao Z., Nath N., Froula J.L., Clum A., Zhang T., Cheng J.F., Copeland A.C., Pennacchio L.A., Chen F. (2012) Generation of long insert pairs using a Cre-LoxP inverse PCR approach. PLoS ONE 7: e29437.

Peng, G., Kashio, M., Morimoto, T., Li, T., Zhu, J., Tominaga, M., Kadowaki, T. (2015) Plant-Derived Tick Repellents Activate the Honey Bee Ectoparasitic Mite TRPA1. Cell Reports 12, 190-202.

Pepato A. R., da Rocha C. E. F., Dunlop J. A. (2010) Phylogenetic position of the acariform mites: sensitivity to homology assessment under total evidence. BMC Evolutionary Biology 10: 235.

Pevzner P.A., Tang H., Waterman M.S. (2001) An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences 98, 9748-9753.

Plettner E., Lazar J., Prestwich E.G., Prestwich G.D. (2000) Discrimination of pheromone enantiomers by two pheromone binding proteins from the gypsy moth Lymantria dispar. Biochemistry 39, 8953-8962.

Pinger R. R., 2008: The blacklegged tick (Ixodes scapularis) in Indiana: a review. Proceedings of the Indian Academy of Science 117, 159-166.

Pichai K., Polgar, G., Heine J. (2008) The efficacy of Bayvarol® and CheckMite+® in the control of Tropilaelaps mercedesae in the European honey bee (Apis mellifera) in Thailand. Apiacta 43, 12-16.

Pomerantz A.F., Hoy M.A., Kawahara A.Y. (2015) Molecular characterization and evolutionary insights into potential sex-determination genes in the western orchard predatory mite

Metaseiulus occidentalis (Chelicerata: Arachnida: Acari: Phytoseiidae). Journal of Biomolecular Structure and Dynamics 33, 1239-53.

Ponce R., Hartl D.L. (2006) The evolution of the novel Sdic gene cluster in Drosophila melanogaster. Gene 376, 174-183.

Pop M., Salzberg S.L. (2008) Bioinformatics challenges of new sequencing technology. Trends in Genetics 24, 142-149.

Pop M. (2009) Genome assembly reborn: Recent computational challenges. Briefings in Bioinformatics 10, 354-366.

Price A.L., Jones N.C. and Pevzner P.A. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21, 351-8.

Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P., Gu Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Rapaport F., Khanin R., Liang Y., Pirun M., Krek A., Zumbo P., Mason C.E., Socci N.D., Betel D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology 14: R95.

Raffique M.K., Mahmood R., Aslam M., Sarwar G. (2012) Control of Tropilaelaps clareae mite by using formic acid and thymol in honey bee Apis mellifera colonies. Pakistan Journal of Zoology 44, 1129-1135.

Ramsey J.S., Rider D.S., Walsh T.K., De Vos M., Gordon K.H., Ponnala L., Macmil S.L., Roe B.A., Jander G. (2010) Comparative analysis of detoxification enzymes in Acyrthosiphon pisum and Myzus persicae. Insect Molecular Biology 2, 155-64.

Rebers J.E., Willis J.H. (2001) A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochemistry and Molecular Biology 31, 1083-1093.

Reddy B., Prasad G., Raghavendra K. (2011) In silico analysis of glutathione S-transferase supergene family revealed hitherto unreported insect specific δ-and ε-GSTs and mammalian specific μ-GSTs in Ixodes scapularis (Acari: Ixodidae). Computational Biology and Chemistry 35, 114-120.

Remnant E.J., Good R.T., Schmidt J.M., Lumb C., Robin C., Daborn P.J., Batterham P. (2013) Gene duplication in the major insecticide target site, Rdl, in Drosophila melanogaster. Proceedings of the National Academy of Sciences 110, 14705-14710.

Rewitz K.F., O'Connor M.B., Gilbert L.I. (2007) Molecular evolution of the insect Halloween family of cytochrome P450s: phylogeny, gene organization and functional conservation. Insect Biochemistry and Molecular Biology 37, 741-753.

Rewitz K. F. and Gilbert L. I. (2008) Daphnia Halloween genes that encode cytochrome P450s mediating the synthesis of the arthropod molting hormone: evolutionary implications. BMC Evolutionary Biology 8: 60.

Reese M. G, Guigo, R. (2006) EGASP: Introduction. Genome Biology 7, 1-3.

Robinson R. (2008) For Mammals, Loss of Yolk and Gain of Milk Went Hand in Hand. PLoS Biology 6:e77.

Robinson M.D., McCarthy D.J., Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-40.

Robertson H.M., Wanner K.W. (2006) The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family. Genome Research 16, 1395-403.

Robinson M.D., Oshlack A., (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11:R25.

Ronquist F., Huelsenbeck J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Rosenkranz P., Aumeier P., Ziegelmann B. (2010) Biology and control of Varroa destructor. Journal of Invertebrate Pathology 103, S96-S119.

Robinson K. M., Sieber K. B, Hotopp J. C. D. (2013) A Review of Bacteria-Animal Lateral Gene Transfer May Inform Our Understanding of Diseases like Cancer, PLOS Genetics 9:e1003877.

Roossinck, M.J. (2011) The good viruses: viral mutualistic symbioses. Nature Reviews Microbiology 9, 99-108.

Rigaud T., Perrot-Minnot M-J., Brown M.J.F. (2010) Parasite and host assemblages: embracing the reality will improve our knowledge of parasite transmission and virulence. Proceedings of the Royal Society B-Biological Sciences 277, 3693-3702.

Roncalli V., Cieslak M.C., Passamaneck Y., Christie A.E., Lenz P.H. (2015) Glutathione S-Transferase (GST) Gene Diversity in the Crustacean Calanus finmarchicus--Contributors to Cellular Detoxification. PLoS One. 10:e0123322.

Rytz R., Croset V., Benton R. (2013) Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in Drosophila and beyond. Insect Biochemistry and Molecular Biology 43, 888-897.

Sámi L., Pusztahelyi T., Emri T., Varecza Z., Fekete A., Grallert A., Karanyi Z., Kiss L., Pócsi I. (2001) Autolysis and aging of Penicillium chrysogenum cultures under carbon starvation: Chitinase production and antifungal effect of allosamidin. The Journal of General and Applied Microbiology 47, 201-211.

Sanger F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463-5467.

Salzberg S.L., Phillippy A.M., Zimin A., Puiu D., Magoc T., Koren S., Treangen T.J., Schatz M.C., Delcher A.L., Roberts M. (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Resesrach 22, 557-567.

Sanggaard K.W., Bechsgaard J.S., Fang X., Duan J., Dyrlund T.F., Gupta V., Jiang X., Cheng L., Fan D., Feng Y., Han L., Huang Z., Wu Z., Liao L., Settepani V., Thøgersen I.B., Vanthournout B., Wang T., Zhu Y., Funch P., Enghild J.J., Schauser L., Andersen S.U., Villesen P., Schierup M.H., Bilde T., Wang J. (2014) Spider genomes provide insight into composition and evolution of venom and silk, Nature Communications 5:3765.

Sánchez-Gracia A., Vieira F.G., Rozas J. (2009) Molecular evolution of the major chemosensory gene families in insects. Heredity 103, 208-216.

Sa´nchez-Gracia, A., Vieira F.G., Almeida F.C., Rozas J (2011) Comparative Genomics of the Major Chemosensory Gene Families in Arthropods. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.

Sato K., Touhara K. (2009) Insect olfaction: receptors, signal transduction, and behavior. Results and Problems in Cell Differentiation 47, 121-138.

Sato A., Sokabe T., Kashio M., Yasukochi Y., Tominaga M., Shiomi K. (2014) Embryonic thermosensitive TRPA1 determines transgenerational diapause phenotype of the silkworm, Bombyx mori. Proceedings of the National Academy of Sciences of the United States of America 111, 1249-1255.

Scott J.G. (1999) Cytochromes P450 and insecticide resistance, Insect Biochemistry and Molecular Biology 29, 757-777.

Schneider E., Ryan T. J. (2006) Gamma-glutamyl hydrolase and drug resistance, Science Direct 374, 25-32.

Schlebusch S., Illing N. (2012) Next generation shotgun sequencing and the challenges of de novo genome assembly. south african journal of science 108, 11-12.

Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30, 2068-2069.

Simpson, G. G. (1952) The Meaning of Evolution: A Study of the History of Life and of Its Significance for Man. Yale University Press, New Haven.

Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. (2009) ABySS: a parallel assembler for short read sequence data. Genome Research 19, 1117-1123.

Simpson S. A., Alexander D.J. (2005) Reed C.J. Induction of heat shock protein 70 in rat olfactory epithelium by toxic chemicals: in vitro and in vivo studies. Archives of Toxicology 79, 224-230.

Stanke M., Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research 33, 465-467.

Slater G.S., Birney E. (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Sammataro D., Gerson U., Needham G. (2000) Parasitic mites of honey bees: life history, implications, and impact. Annual Review of Entomology 45, 519-548.

Shichida Y., Matsuyama T. (2009) Evolution of opsins and phototransduction Philosophical transactions of the Royal Society of London. Series B, Biological sciences 364, 2881-2895.

Shi H., Pei L., Gu S., Zhu S., Wang Y., Zhang Y., Li B. (2012) Glutathione S-transferase (GST) genes in the red flour beetle, Tribolium castaneum, and comparative analysis with five additional insects. Genomics 100, 327-35.

Shultz J.W., Regier J.C. (2000) Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. Proceedings of the Royal Society of London B Biological Sciences 267,1011-1019.

Smith K.M., Loh E.H., Rostal M.K., Zambrana-Torrelio C.M., Mendiola L., Daszak P. (2013) Pathogens, pests, and economics: drivers of honey bee colony declines and losses. Ecohealth. 10, 434-45.

Sonenshine, D. E., 2013: Biology of tick. Volume 1. Oxford University press, Oxford: university press.

Sojka D., Franta Z., Horn M., Caffrey C.R., Mareš M., Kopáček P. (2013) New insights into the machinery of blood digestion by ticks. Trends in Parasitology 29, 276-85.

Su C.Y., Menuz K., Carlson J.R. (2009) Olfactory perception: receptors, cells, and circuits. Cell 139, 45-59.

Suzuki T., Yano K., Sugimoto S., Kitajima K., Lennarz W.J., Inoue S., Inoue Y., Emori Y. (2002) Proceedings of the National Academy of Sciences. Endo-beta-N-acetylglucosaminidase, an enzyme involved in processing of free oligosaccharides in the cytosol. 99, 9691-9696.

Sztal T., Chung H., Berger S., Currie P.D., Batterham P., Daborn P.J. (2012) A cytochrome p450 conserved in insects is involved in cuticle formation. PLoS One 7: e36544.

Takaishi M., Fujita F., Uchida, K., Yamamoto S., Sawada M., Hatai C., Shimizu M., Tominaga M. (2012) 1,8-cineole, a TRPM8 agonist, is a novel natural antagonist of human TRPA1. Molecular Pain 8:86.

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution 30, 2725-9.

Taylor C.M., Wang Q., Rosa B.A., Huang S.C., Powell K., Schedl T., Pearce E.J., Abubucker S., Mitreva M. (2013) Drug targets Discovery of anthelmintic drug targets and drugs using chokepoints in nematode metabolic pathways. PLOS Pathogens 9:e1003505.

Terakita A., Hara R., Hara T. (1989) Retinal-binding protein as a shuttle for retinal in the rhodopsin retinochrome system of the squid visual cells. Vision Research 29, 639-652.

Terakita A., Hariyama T., Tsukahara, Y., Katsukura Y., Tashiro H. (1993) Interaction of GTP-binding protein Gq with photoactivated rhodopsin in the photoreceptor membranes of crayfish. FEBS Letter 330, 197-200.

Terakita A. (2005) The opsins. Genome Biology 6:213.

Terakita A., Kawano-Yamashita E., Koyanagi M. (2012) Evolution and diversity of opsins. WIREs Membrane Transport and Signaling 1, 104-111.

The Honey bee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honey bee Apis mellifera. Nature 443, 931-949.

Thompson, B.H. (1976) Studies on the attraction of Simulium damnosum (Diptera: Simuliidae) to its hosts. I. the relative importance of sight, exhaled breath and smell. Tropenmed Parasitol 27, 455-473.

Touhara, K., Vosshall, L. B. (2009) Sensing odorants and pheromones with chemosensory receptors. Annual Review of Physiology 71, 307-332.

Tutar L., Tutar Y. (2010) Heat Shock Proteins; An Overview, Current Pharmaceutical Biotechnology 11, 216-222.

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

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

Treangen T.J., Salzberg S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13, 36-46.

Truman, J. W., and Riddiford L. M. (2002) Endocrine insights into the evolution of metamorphosis in insects. Annual Review of Entomology 47, 467-500.

Tsai I., Otto T. (2010) Berriman M: Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology 11:R41.

Tsuchida T., Koga R., Horikawa M., Tsunoda T., Maoka T., Maoka T., Matsumoto T., Simon J.C., Fukatsu T. (2010) Symbiotic Bacterium Modifies Aphid Body Color. Science 330, 1102-1104.

Yandell M., Ence D. (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 2012, 13, 329-342.

Zerbino D.R., Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 18, 821-829.

Zhong Y, Jia Y, Gao Y, Tian D, Yang S, Zhang X. (2013) Functional requirements driving the gene duplication in 12 Drosophila species. BMC Genomics 14:555.

Zerbino D.R., Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821-829.

Zerbino D.R. (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Nature Protocol Chpter-11:Unit-11.5.

Zhuang X., Yang C., Fevolden S.E., Cheng C. (2012) Protein genes in repetitive sequence—antifreeze glycoproteins in Atlantic cod genome, BMC Genomics 13:293.

Zhang G., Fang X., Guo X., Li L., Luo R., Xu F., Yang P., Zhang L., Wang X., Qi H., Xiong Z., Que H., Xie Y., Holland P., Paps J., Zhu Y., Wu F., Chen Y., Wang J., Peng C., Meng J., Yang L., Liu J., Wen B., Zhang N., Huang Z., Zhu Q., Feng Y., Mount A., Hedgecock D., Xu Z., Liu Y., Domazet-Lošo T., Du Y., Sun X., Zhang S., Liu B., Cheng P., Jiang X., Li J., Fan D., Wang W., Fu W., Wang T., Wang B., Zhang J., Peng Z., Li Y., Li N., Wang J., Chen M., He Y., Tan F., Song X., Zheng Q., Huang R., Yang H., Du X., Chen L., Yang M., Gaffney PM., Wang S., Luo L., She Z., Ming Y., Huang W., Zhang S., Huang B., Zhang Y., Qu T., Ni P., Miao G., Wang J., Wang Q., Steinberg CE., Wang H., Li N., Qian L., Zhang G., Li Y., Yang H., Liu X., Wang J., Yin Y., Wang J. (2012) The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490, 49-54.

Ullmann A.J., Lima C.M.R., Guerrero F.D., Piesman J., Black W.C. (2005) Genome size and organization in the blacklegged tick, Ixodes scapularis and the Southern cattle tick. Boophilus microplus. Insect molecular biology 14, 217-222.

van Heesch S., Kloosterman W.P., Lansu N., Ruzius F.P., Levandowsky E., Lee C.C., Zhou S., Goldstein S., Schwartz D.C., Harkins T.T., Guryev V., Cuppen E. (2013) Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC Genomics 16, 14-257.

van Nieuwerburgh F., Thompson R.C., Ledesma J., Deforce D., Gaasterland T., Ordoukhanian P., Head S.R. (2011) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Research 40: e24.

van der Hammen. L. 1989. An Introduction to Comparative Arachnology. SPB Academic Publishing, The Hague.

van Leeuwen T., Vontas J., Tsagkarakou A., Dermauw W., Tirry L. (2010) Acaricide resistance mechanisms in the two-spotted spider mite Tetranychus urticae and other important Acari: a review Insect Biochem. Journal of Molecular Biology 40, 563-572.

van Wijk M., Wadman W.J., Sabelis M.W. (2006) Morphology of the olfactory system in the predatory mite Phytoseiulus persimilis. Experimental and Applied Acarology 40, 217-29.

Venkatachalam K., Montell C. 2007 TRP channels. Annual Review of Biochemistry 76, 387-417.

Vieira F.G., Rozas J. (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biology and Evolution 3, 476-490.

Vogt R.G. (2003) Biochemical diversity of odor detection: OBPs, ODEs and SNMPs. In Insect pheromone biochemistry and molecular biology. Edited by Blomquist G, Vogt RG. Academic Press, San Diego, U.S. pp 391-445.

Wang W., Brunet F.G, Nevo E., Long M. (2002) Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proceedings of the National Academy of Sciences. 99, 4448-4453.

Wanner K.W., Willis L.G., Theilmann D.A., Isman M.B., Feng Q., Plettner E. (2004) Analysis of the insect os-d-like gene family. Journal of Chemical Ecology 30, 889-911.

Wanner K.W., Isman M.B., Feng Q., Plettner E., Theilmann D.A. (2005) Developmental expression patterns of four chemosensory protein genes from the Eastern spruce budworm, Chroistoneura fumiferana. Insect Molecular Biology 14, 289-300.

Waladde S. M., Rice M. J. (1982) The sensory basis of tick feeding behavior, In F. Obenchain D. and Galun R. editor, Physiology of ticks. Pergamon, New York, 71-118.

Wetzel J., Kingsford C., Pop M. (2011) Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. BMC Bioinformatics 13, 12-95.

Weiss P. (1950). Perspectives in the field of morphogenesis. The Quarterly Review of Biology 25, 177-198.

Werren J.H., Richards S., Desjardins C.A., Niehuis O., G.adau J., Colbourne J.K.; Nasonia Genome Working Group (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327, 343-348.

Williams J.C., Hagedorn H.H., Beyenbach K.W. (1983) Dynamic changes in flow rate and composition of urine during the post blood meal diuresis in Aedes aegypti. Journal of Comparative Physiology [B] 153, 257-266.

Wilkins, A.S., (1995). Moving up the hierarchy: a hypothesis on the evolution of a genetic sex determination pathway. Bioessays 17, 71-77.

Woolhouse M.E.J., Taylor L.H., Haydon D.T. (2001) Population biology of multihost pathogens. Science 292, 1109-1112.

Woodring J., Boulden M., Das S., Gäde G. (1993) Studies on blood sugar homeostasis in the honeybee (Apis mellifera, L.). Journal of Insect Physiology 39, 89-97.

Xu X., Pan S., Cheng S., Zhang B., Mu D., Ni P., Zhang G., Yang S., Li R., Wang J. et al, (2011) Genome sequence and analysis of the tuber crop potato. Nature 475, 189-195.

Xia Q.Y., Zhou Z.Y., Lu C., Cheng D.J., Dai F.Y., Li B. et al (2004) A draft sequence for the genome of the domesticated silkworm (Bombyx mori). Science 306, 1937-1940.

Xiao R., Zhang B., Dong Y., Gong J., Xu T., Liu J., Xu X. Z. S. (2013) A Genetic Program Promotes C. elegans Longevity at Cold Temperatures via a Thermosensitive TRP Channel. Cell 152, 806-817.

Yang Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24, 1586-1591.

Yandell M., Ence D. (2012) A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 8, 329-42.

Ye L., Hillier L.W., Minx P., Thane N., Locke D.P., Martin J.C., Chen L., Mitreva M., Miller J.R., Haub K.V., Dooling D.J., Mardis E.R., Wilson R.K., Weinstock G.M., Warren W.C. (2011) A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biology 12:R31.

Yeh I., Hanekamp T., Tsoka S., Karp P.D., Altman R.B. (2004) Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. Genome Reserach 14, 917-924.

Yin J., Zhang J. (2011) Multidrug resistance-associated protein 1 (MRP1/ABCC1) polymorphism: from discovery to clinical application. Zhong Nan Da Xue Xue Bao Yi Xue Ban 36, 927-38.

Yu Q.Y., Lu C., Li W.L., Xiang Z.H., Zhang Z. (2009) Annotation and expression of carboxylesterases in the silkworm, Bombyx mori. BMC Genomics 24, 10-553.

Yue C., Genersch E. (2005) RT-PCR analysis of deformed wing virus in honey bees (Apis mellifera) and mites (Varroa destructor). Journal of General Virology 86, 3419-3424.

Zdobnov E.M., Apweiler R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-848.

Zeigler D. (2014) Evolution: Components and Mechanisms. Waltham (U.S.): Academic Press.

Zhang Z, Klein J, Nei M. (2014) Evolution of the sex-lethal gene in insects and origin of the sex-determination system in Drosophila. Journal of Molecular Evolution 78, 50-65.

Zhong L., Bellemer A., Yan H., Honjo K., Robertson J., Hwang R. Y., Pitt G. S., Tracey W. D. (2012) Thermosensory and Nonthermosensory Isoforms of Drosophila melanogaster TRPA1 Reveal Heat-Sensor Domains of a ThermoTRP Channel. Cell Reports 1, 43-55.

Zhou X., Ren L., Meng Q., Li Y., Yu Y., Yu J. (2010) The next-generation sequencing technology and application. Protein and Cell 1, 520-536.

Zhu F., Moural T.W., Shah K., Palli S.R. (2013) Integrated analysis of cytochrome P450 gene superfamily in the red flour beetle, Tribolium castaneum. BMC Genomics 14:174.

Appendix 1 List of transcripts identified by blastx sorted with the functional annotations.

Transcript ID	Contig ID	Node coverage	Sequence description	NCBI best hit	E-Value
CUFF.469.1	contig055076	55.58	Acetolactate synthase, putative	XP_002400242	2.19E-51
CUFF.73.1	contig076798	78.77	Acetolactate synthase-like	XP_003744242	4.96E-67
CUFF.125.1	contig082816	51.89	Acetolactate synthase-like	XP_003744242	5.50E-46
CUFF.74.1	contig077756	51.77	Acid phosphatase 5-like	XP_003748526	8.15E-06
CUFF.548.1	contig064398	44.76	Arylsulfatase b-like	XP_003737944	4.06E-24
CUFF.552.1	contig065101	43.50	Autophagy-related protein 9a-like	XP_001946903	4.56E-15
CUFF.422.1	contig047569	38.72	Autophagy-related protein 9a-like	XP_003743577	1.84E-51
CUFF.123.1	contig082903	43.26	Bifunctional atp-dependent dihydroxyacetone kinase fad-amp lyase -like	XP_003747813	1.90E-13
CUFF.113.1	contig080809	45.50	Canalicular multispecific organic anion transporter 1-like	XP_003740191	3.80E-22
CUFF.498.1	contig058801	40.25	Canalicular multispecific organic anion transporter 1-like	XP_003740191	3.65E-92
CUFF.508.1	contig059937	39.51	Canalicular multispecific organic anion transporter 1-like	XP_003740191	6.75E-26
CUFF.387.1	contig040332	49.23	Canalicular multispecific organic anion transporter partial	XP_003738712	8.33E-16
CUFF.580.1	contig067404	46.31	Canalicular multispecific organic anion transporter partial	XP_003738712	1.75E-22
CUFF.72.1	contig077592	45.36	Canalicular multispecific organic anion transporter partial	XP_003738712	2.92E-18
CUFF.543.1	contig063961	40.46	Canalicular multispecific organic anion transporter partial	XP_003738712	9.05E-29
CUFF.408.1	contig046276	45.47	Cd81 antigen-like	XP_003737077	4.52E-11

CUFF.545.1	contig060877	44.60	Cdgsh iron-sulfur domain-containing protein mitochondrial-like	XP_003742579	9.34E-22
CUFF.483.1	contig056199	45.05	Cdk-activating kinase assembly factor mat1-like	XP_003742096	2.68E-55
CUFF.536.1	contig061920	42.10	Cdk-activating kinase assembly factor mat1-like	XP_003742096	2.74E-30
CUFF.536.2	contig061920	42.10	Cdk-activating kinase assembly factor mat1-like	XP_003742096	3.14E-30
CUFF.436.1	contig005360	42.95	Cell division control protein 42 homolog isoform x1	XP_003452737	3.21E-44
CUFF.195.1	contig091792	38.31	Coatomer subunit delta-like	XP_003739134	4.79E-11
CUFF.177.1	contig090301	49.25	Conserved hypothetical protein	XP_002435289	4.84E-19
CUFF.433.1	contig050752	45.96	Cytochrome p450 4c3-like	XP_003740735	1.38E-12
CUFF.135.1	contig084659	39.68	Cytoplasmic- partial	XP_005315620	3.78E-43
CUFF.525.1	contig062028	49.10	Dna-directed rna polymerases and iii subunit rpabc1-like	XP_003745798	4.13E-43
CUFF.327.1	contig030270	44.44	E3 ubiquitin-protein ligase rnf25-like	XP_003746922	5.36E-15
CUFF.327.2	contig030270	44.44	E3 ubiquitin-protein ligase rnf25-like	XP_003746922	3.37E-22
CUFF.328.1	contig016686	42.36	Endoribonuclease dcr-1-like	XP_003739924	2.32E-26
CUFF.54.1	contig073875	42.27	Eukaryotic initiation factor 4a-iii	XP_003748087	1.79E-115
CUFF.58.1	contig076373	39.25	Eukaryotic initiation factor 4a-iii	XP_003748087	1.30E-82
CUFF.242.1	contig099310	36.78	Fatty acid synthase-like	XP_003739635	9.24E-57
CUFF.16.1	contig070844	38.34	Folylpolyglutamate mitochondrial-like	XP_003737634	1.08E-43
CUFF.488.1	contig055194	37.50	Folylpolyglutamate mitochondrial-like	XP_003737634	5.83E-158
CUFF.61.1	contig076702	37.53	Forkhead box protein k2- partial	XP_003745820	3.22E-54
CUFF.521.1	contig060516	35.94	Forkhead box protein k2- partial	XP_003745820	1.21E-93
CUFF.330.1	contig028409	45.33	Fructose- -bisphosphatase 1-like	XP_003744558	9.65E-81
CUFF.70.1	contig077384	58.91	Gamma-glutamyl hydrolase a-like	XP_003747812	7.73E-13

CUFF.505.1	contig058882	50.51	Gamma-glutamyl hydrolase a-like	XP_003747812	2.27E-15
CUFF.92.1	contig079602	44.47	General transcription factor iie subunit 1-like	XP_003738556	2.95E-21
CUFF.475.1	contig040216	65.11	Gh16126p	ACT88134	2.01E-36
CUFF.320.1	contig028467	40.55	Glutamine synthetase 2	BAN20347	5.49E-29
CUFF.191.1	contig086588	44.86	Glutamine synthetase 2 cytoplasmic-like	XP_003741674	1.00E-48
CUFF.316.1	contig028467	40.55	Glutamine synthetase 2 cytoplasmic-like	XP_003741674	8.85E-31
CUFF.122.1	contig081757	40.47	Glutamine synthetase 2 cytoplasmic-like	XP_003741674	5.98E-19
CUFF.97.1	contig079191	37.41	Glycine cleavage system h mitochondrial-like	XP_003741719	9.57E-19
CUFF.146.1	contig012647	37.38	Glycoprotein 3-alpha-l-fucosyltransferase a-like	XP_003742256	4.30E-169
CUFF.381.1	contig040305	47.75	Heat shock 70 kda protein 6-like	XP_001362950	5.56E-15
CUFF.301.1	contig028325	35.43	Heat shock protein 70	ACN54681	9.88E-38
CUFF.375.1	contig038961	113.07	Heat shock protein partial	AFN52221	2.71E-96
CUFF.550.1	contig064433	99.19	Heat shock protein partial	AGJ98022	1.86E-64
CUFF.501.1	contig058972	129.31	Hsp70	AAO44919	7.47E-67
CUFF.438.1	contig049190	109.28	Hsp70 protein	CBX24529	2.78E-81
CUFF.403.1	contig045051	86.42	Hsp70, partial	AET43984	6.63E-16
CUFF.283.1	contig015115	41.97	Hermansky-pudlak syndrome 5 protein homolog	XP_003742832	2.07E-83
CUFF.285.1	contig015115	41.97	Hermansky-pudlak syndrome 5 protein homolog	XP_003742832	1.03E-156
CUFF.314.1	contig031684	37.41	Histone -like	XP_004065634	7.64E-84
CUFF.342.1	contig034369	47.42	Hypothetical protein	XP_001615647	5.37E-09

CUFF.558.1	contig065217	42.06	Hypothetical protein DAPPUDRAFT_315759	EFX83495	2.52E-13
CUFF.553.1	contig065111	56.68	Hypothetical protein iscw_ISCW007056	XP_002404428	5.85E-24
CUFF.86.1	contig078758	39.28	Hypothetical protein iscw_ISCW017104	XP_002403261	3.00E-08
CUFF.116.1	contig081866	44.28	Hypothetical protein tcasga2_TC001508	EFA12484	1.89E-28
CUFF.52.1	contig010414	39.18	Low quality protein: intraflagellar transport protein 52 homolog	XP_003743334	4.81E-49
CUFF.120.1	contig080447	49.40	Low-density lipoprotein receptor-related protein 6	XP_003742906	1.57E-25
CUFF.282.1	contig025490	46.51	Low-density lipoprotein receptor-related protein 6	XP_003742906	8.55E-13
CUFF.163.1	contig079944	45.49	Mannose-p-dolichol utilization defect 1	XP_003742114	4.18E-66
CUFF.17.1	contig069273	42.46	Mediator of rna polymerase ii transcription subunit 6-like	XP_003741388	8.47E-25
CUFF.559.1	contig064604	41.71	Mediator of rna polymerase ii transcription subunit 6-like	XP_003741388	1.85E-45
CUFF.540.1	contig063362	47.89	Metallo-beta-lactamase domain-containing protein 1-like	XP_003746687	5.21E-11
CUFF.147.1	contig083824	47.10	Metallo-beta-lactamase domain-containing protein 1-like	XP_003746687	5.32E-21
CUFF.303.1	contig027732	42.94	Metallo-beta-lactamase domain-containing protein 1-like	XP_003746687	7.41E-17
CUFF.270.1	contig021115	45.29	Microtubule-associated protein tau-like	XP_003744844	1.03E-11
CUFF.329.1	contig032706	48.20	Mosc domain-containing protein mitochondrial-like	XP_003743333	3.51E-36
CUFF.450.1	contig053059	47.68	Multidrug resistance-associated protein 1- partial	XP_003738711	9.67E-27

CUFF.226.1	contig016568	39.85	Multidrug resistance-associated protein 1- partial	XP_003738711	5.55E-108
CUFF.402.1	contig040237	43.83	Myosin heavy muscle-like isoform 2	XP_003747493	1.03E-26
CUFF.427.1	contig046789	43.50	N-alpha-acetyltransferase 20-like	XP_003747157	1.46E-32
CUFF.427.2	contig046789	43.50	N-alpha-acetyltransferase 20-like	XP_003747157	1.30E-32
CUFF.524.1	contig061858	42.43	N-alpha-acetyltransferase 20-like	XP_003747157	1.80E-34
CUFF.359.1	contig032860	81.81	Neuronal acetylcholine receptor subunit alpha-7-like	XP_003740161	4.85E-36
CUFF.11.1	contig070839	107.02	Pab-dependent poly -specific ribonuclease subunit 3-like	XP_003737338	4.79E-12
CUFF.156.1	contig086678	44.72	Pab-dependent poly -specific ribonuclease subunit 3-like	XP_003737338	1.63E-29
CUFF.129.1	contig083324	46.80	Phospholipase a-2-activating	XP_003737169	1.17E-31
CUFF.581.1	contig064256	46.51	Phospholipase a-2-activating	XP_003737169	1.73E-29
CUFF.582.1	contig064256	46.51	Phospholipase a-2-activating	XP_003737169	1.20E-126
CUFF.582.2	contig064256	46.51	Phospholipase a-2-activating	XP_003737169	1.88E-128
CUFF.152.1	contig011429	43.15	Poly -specific ribonuclease parn-like	XP_003745685	3.69E-149
CUFF.152.2	contig011429	43.15	Poly -specific ribonuclease parn-like	XP_003745685	2.56E-147
CUFF.153.1	contig011429	43.15	Poly -specific ribonuclease parn-like	XP_003745685	1.12E-147
CUFF.539.1	contig008046	39.25	PREDICTED: neural-cadherin-like	XP_003743540	0
CUFF.53.1	contig076258	35.28	PREDICTED: semaphorin-1A-like	XP_003742005	1.34E-38
CUFF.182.1	contig090373	55.48	PREDICTED: TATA-box-binding protein-like	XP_003742533	5.68E-37
CUFF.373.1	contig037927	47.62	PREDICTED: TATA-box-binding protein-like	XP_003742533	5.63E-17
CUFF.574.1	contig066782	47.53	PREDICTED: uncharacterized protein c9orf114-like	XP_003744780	1.45E-18
CUFF.218.1	contig095208	43.99	PREDICTED: uncharacterized protein	XP_003744780	7.96E-24

			c9orf114-like		
CUFF.231.1	contig098041	43.77	PREDICTED: uncharacterized protein c9orf114-like	XP_003744780	2.20E-18
CUFF.276.1	contig103490	37.09	PREDICTED: uncharacterized protein LOC100898247	XP_003739177	1.41E-06
CUFF.160.1	contig087380	42.50	PREDICTED: uncharacterized protein LOC100899188	XP_003743623	5.24E-07
CUFF.33.2	contig070282	66.48	PREDICTED: uncharacterized protein LOC100899580	XP_003747600	5.49E-19
CUFF.7.1	contig069308	45.07	PREDICTED: uncharacterized protein LOC100899700	XP_003743788	2.98E-41
CUFF.111.1	contig080652	37.32	PREDICTED: uncharacterized protein LOC100899779	XP_003740529	9.34E-37
CUFF.358.1	contig036816	49.78	PREDICTED: uncharacterized protein LOC100899942	XP_003748478	4.87E-19
CUFF.60.1	contig076697	45.94	PREDICTED: uncharacterized protein LOC100899942	XP_003748478	2.14E-30
CUFF.56.1	contig076374	35.15	PREDICTED: uncharacterized protein LOC100899942	XP_003748478	2.25E-21
CUFF.449.1	contig053260	40.67	PREDICTED: uncharacterized protein LOC100900364	XP_003747605	3.71E-24
CUFF.238.1	contig099314	42.84	PREDICTED: uncharacterized protein LOC100901218	XP_003746037	2.19E-21
CUFF.81.1	contig078063	41.74	PREDICTED: uncharacterized protein LOC100901218	XP_003746037	1.68E-33
CUFF.434.1	contig051541	47.77	PREDICTED: uncharacterized protein LOC100901399	XP_003743404	1.54E-07
CUFF.390.1	contig040361	42.52	PREDICTED: uncharacterized protein	XP_003743404	2.89E-13

			LOC100901399		
CUFF.115.1	contig011722	41.96	PREDICTED: uncharacterized protein LOC100903560	XP_003743730	4.65E-23
CUFF.150.1	contig086153	39.52	PREDICTED: uncharacterized protein LOC100903560	XP_003743730	1.05E-74
CUFF.546.1	contig062207	39.21	PREDICTED: uncharacterized protein LOC100904087 isoform 2	XP_003743575	2.08E-85
CUFF.465.1	contig042688	42.23	PREDICTED: uncharacterized protein LOC100904146	XP_003746699	6.83E-113
CUFF.465.2	contig042688	42.23	PREDICTED: uncharacterized protein LOC100904146	XP_003746699	8.86E-115
CUFF.551.1	contig064673	38.27	PREDICTED: uncharacterized protein LOC100906374	XP_003747030	3.09E-08
CUFF.592.1	contig037016	39.93	PREDICTED: uncharacterized protein LOC100906870	XP_003739519	4.34E-125
CUFF.495.1	contig058298	36.79	PREDICTED: uncharacterized protein LOC100908780	XP_003741930	5.06E-25
CUFF.280.1	contig017492	42.57	PREDICTED: uncharacterized protein LOC100908922	XP_003742425	9.77E-42
CUFF.280.2	contig017492	42.57	PREDICTED: uncharacterized protein LOC100908922	XP_003742425	3.74E-42
CUFF.13.1	contig071205	36.50	PREDICTED: uncharacterized protein LOC100909062	XP_003740433	5.86E-11
CUFF.224.1	contig095435	39.67	Probable 4-coumarate-- ligase 5-like	XP_003746405	2.82E-07
CUFF.173.1	contig090081	44.15	Probable cytochrome p450 49a1-like	XP_003741296	1.21E-11
CUFF.413.1	contig046437	47.14	Probable phospholipid hydroperoxide glutathione peroxidase mitochondrial-like	XP_003742990	1.26E-12

CUFF.486.1	contig057667	45.56	Probable phospholipid hydroperoxide glutathione peroxidase mitochondrial-like	XP_003742990	1.12E-11
CUFF.76.1	contig077830	40.71	Probable phospholipid hydroperoxide glutathione peroxidase mitochondrial-like	XP_003742990	1.54E-16
CUFF.295.1	contig026037	50.04	Proteasome inhibitor pi31 subunit-like	XP_003747136	6.09E-24
CUFF.33.1	contig070282	66.48	Protein asteroid homolog 1-like	XP_003737199	1.58E-13
CUFF.187.1	contig086593	39.92	Protein crumbs-like	XP_003742957	5.07E-83
CUFF.187.2	contig086593	39.92	Protein crumbs-like	XP_003742957	1.07E-112
CUFF.187.3	contig086593	39.92	Protein crumbs-like	XP_003742957	5.36E-111
CUFF.26.1	contig072360	39.03	Protein crumbs-like	XP_003742957	4.25E-27
CUFF.589.1	contig067703	35.32	Retinal dehydrogenase 1-like	XP_003744198	4.92E-49
CUFF.464.3	contig053386	44.36	Sec1 family domain-containing protein 2-like	XP_006630198	1.63E-09
CUFF.464.4	contig053386	44.36	Sec1 family domain-containing protein 2-like	XP_006630198	2.47E-14
CUFF.131.1	contig082914	44.77	Serine threonine-protein kinase haspin homolog	XP_004524206	3.79E-22
CUFF.447.1	contig052419	45.77	Serine threonine-protein kinase plk1-partial	XP_004622712	9.10E-20
CUFF.456.1	contig053179	37.45	Serine threonine-protein kinase plk1-like	XP_003739726	1.14E-22
CUFF.432.1	contig051007	61.66	Sodium-dependent phosphate transporter 1-a-like	XP_003742817	6.29E-23
CUFF.35.1	contig073968	41.42	Sodium-dependent phosphate transporter 1-b-like	XP_003748145	2.12E-58
CUFF.503.1	contig059201	38.29	Sodium-dependent phosphate	XP_003748145	1.22E-36

			transporter 1-b-like		
CUFF.575.1	contig067317	39.11	Sphingomyelin phosphodiesterase-like	XP_003737231	5.52E-29
CUFF.28.1	contig073192	69.70	Sterile alpha and tir motif-containing protein 1-like	XP_003725234	1.07E-12
CUFF.309.1	contig028798	46.05	Tetratricopeptide repeat protein 1-like	XP_003739107	4.61E-25
CUFF.407.1	contig041823	45.28	Transcription elongation factor spt5-like	XP_003739477	0
CUFF.493.1	contig056404	35.20	Transcription elongation factor spt5-like	XP_003739477	2.99E-72
CUFF.263.1	contig015784	39.47	Uncharacterized protein CPUR_02104	CCE35173	2.05E-06
CUFF.272.1	contig104048	46.13	Unnamed protein product	BAK38646	5.19E-14
CUFF.71.1	contig071795	39.44	Upf0451 protein c17orf61 homolog	XP_003741439	5.95E-07
CUFF.185.1	contig090930	35.60	Upf0466 protein mitochondrial-like	XP_003741718	1.58E-11
CUFF.132.1	contig083442	44.88	Upf0488 protein c8orf33 homolog	XP_003741440	2.18E-06
CUFF.502.1	contig057818	35.76	Upf0488 protein c8orf33 homolog	XP_003741440	1.31E-11
CUFF.214.1	contig093247	59.18	Wd repeat-containing protein 46-like	XP_003743608	1.83E-19
CUFF.165.1	contig087998	45.06	Wd repeat-containing protein 46-like	XP_003743608	2.31E-14
CUFF.248.1	contig100393	37.31	Wd sam and u-box domain-containing protein 1-like	XP_003742971	4.20E-24
CUFF.235.1	contig098723	79.78	Zinc finger protein 28-like	XP_003746708	2.86E-23
CUFF.286.1	contig025610	38.45	Zinc finger protein partial	EHB01232	1.09E-13
CUFF.384.1	contig004539	44.84	Zinc transporter, putative	XP_002405991	1.06E-13

Appendix 2 Prioritized list of *T. mercedesae* chokepoints as potential drug targets.

Chokepoint (EC)	**Name**	**Criteria 1**	**Criteria 2**	**Criteria 3**	**Criteria 4**	**Score**	**Criteria 5**	**Criteria 6**
ec:3.2.1.20	alpha-D-glucoside glucohydrolase	1	1	0	1	3	no	yes
ec:3.2.1.3	4-alpha-D-glucan glucohydrolase	1	1	0	1	3	no	yes
ec:3.5.1.9	aryl-formylamine amidohydrolase	1	1	0	1	3	yes	yes
ec:3.7.1.5	3-acylpyruvate acylhydrolase	1	1	0	1	3	no	yes
ec:3.7.1.2	4-fumarylacetoacetate fumarylhydrolase	1	1	0	1	3	no	yes
ec:3.5.1.4	acylamide amidohydrolase	1	1	0	1	3	no	yes
ec:3.3.2.6	(7E,9E,11Z,14Z)-(5S,6S)-5,6-epoxyicosa-7,9,11,14-tetraenoate hydrolase	1	1	0	1	3	no	yes
ec:3.6.1.11	polyphosphate phosphohydrolase	1	0	1	1	3	yes	yes
ec:2.7.7.4	ATP:sulfate adenylyltransferase	1	1	1	0	3	yes	yes
ec:2.7.1.21	ATP:thymidine 5'-phosphotransferase	1	1	1	0	3	yes	yes
ec:2.4.2.8	IMP:diphosphate phospho-D-ribosyltransferase	1	1	1	0	3	yes	yes
ec:3.1.3.80	2,3-bisphospho-D-glycerate 3-phosphohydrolase	1	1	0	1	3	yes	yes

ec:2.4.1.25	(1->4)-alpha-D-glucan:(1->4)-alpha-D-glucan 4-alpha-D-glycosyltransferase	1	1	0	0	2	no	yes
ec:3.1.2.6	S-(2-hydroxyacyl)glutathione hydrolase	1	0	0	1	2	no	yes
ec:3.5.1.3	omega-amidodicarboxylate amidohydrolase	1	0	0	1	2	no	yes
ec:4.1.1.20	meso-2,6-diaminoheptanedioate carboxy-lyase (L-lysine-forming)	1	1	0	0	2	yes	yes
ec:1.8.3.1	sulfite:oxygen oxidoreductase	1	1	0	0	2	no	yes
ec:3.5.1.14	N-acyl-aliphatic-L-amino acid amidohydrolase (carboxylate-forming)	0	1	0	1	2	yes	no
ec:1.1.1.1	alcohol:NAD+ oxidoreductase	1	1	0	0	2	yes	yes
ec:2.7.7.12	UDP-alpha-D-glucose:alpha-D-galactose-1-phosphate uridylyltransferase	1	1	0	0	2	no	yes
ec:3.7.1.3	L-kynurenine hydrolase	0	1	0	1	2	yes	yes
ec:2.7.1.29	ATP:glycerone phosphotransferase	1	1	0	0	2	yes	no
ec:2.7.1.15	ATP:D-ribose 5-phosphotransferase	1	1	0	0	2	no	yes
ec:1.1.1.18	myo-inositol:NAD+ 2-oxidoreductase	1	1	0	0	2	no	yes

ec:2.1.1.1	S-adenosyl-L-methionine: nicotinamide N-methyltransferase	1	1	0	0	2	yes	yes
ec:5.3.2.1	phenylpyruvate keto---enol-isomerase	1	1	0	0	2	yes	yes
ec:4.1.2.10	(R)-mandelonitrile benzaldehyde-lyase (cyanide-forming)	1	1	0	0	2	no	yes
ec:2.7.8.5	CDP-diacylglycerol:sn-glycerol-3-phosphate 3-phosphatidyltransferase	1	1	0	0	2	no	yes
ec:2.7.1.14	ATP:sedoheptulose 7-phosphotransferase	1	1	0	0	2	yes	yes
ec:6.4.1.3	propanoyl-CoA:carbon-dioxide ligase (ADP-forming)	1	1	0	0	2	no	yes
ec:1.1.1.10	xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)	1	1	0	0	2	no	yes
ec:2.7.1.1	ATP:D-hexose 6-phosphotransferase	1	1	0	0	2	yes	yes
ec:1.1.1.145	3beta-hydroxy-Delta5-steroid:NAD+ 3-oxidoreductase	1	1	0	0	2	yes	yes
ec:5.3.3.1	3-oxosteroid Delta5-Delta4-isomerase	1	1	0	0	2	yes	yes
ec:5.3.99.5	(5Z,13E)-(15S)-9alpha,11alpha-epidioxy-15-hydrox	1	1	0	0	2	yes	yes

	yprosta-5,13-dienoate isomerase							
ec:2.1.2.5	5-formimidoyltetrahydrofolate:L-glutamate N-formimidoyltransferase	1	1	0	0	2	yes	yes
ec:4.3.1.4	5-formimidoyltetrahydrofolate ammonia-lyase (cyclizing; 5,10-methenyltetrahydrofolate-forming)	1	1	0	0	2	yes	yes
ec:4.4.1.1	L-cystathionine cysteine-lyase (deaminating; 2-oxobutanoate-forming)	1	1	0	0	2	no	yes
ec:3.5.1.92	(R)-pantetheine amidohydrolase	1	0	0	1	2	no	yes
ec:1.13.11.5	homogentisate:oxygen 1,2-oxidoreductase (ring-opening)	1	1	0	0	2	no	yes
ec:1.1.1.184	secondary-alcohol:NADP+ oxidoreductase	1	1	0	0	2	no	yes
ec:1.1.1.189	(5Z,13E)-(15S)-9alpha,11alpha,15-trihydroxyprosta-5,13-dienoate:NADP+ 9-oxidoreductase	1	1	0	0	2	no	yes
ec:4.2.1.49	3-(5-oxo-4,5-dihydro-3H-imidazol-4-yl)propanoate hydro-lyase	1	1	0	0	2	yes	yes

	(urocanate-forming)							
ec:2.7.1.52	ATP:beta-L-fucose 1-phosphotransferase	1	1	0	0	2	yes	yes
ec:2.7.1.81	GTP:5-hydroxy-L-lysine O-phosphotransferase	1	1	0	0	2	no	yes
ec:4.1.1.45	2-amino-3-(3-oxoprop-1-en-1-yl)but-2-enedioate carboxy-lyase (2-aminomuconate-semialdehyde-forming)	1	1	0	0	2	yes	yes
ec:4.2.1.107	(24R,25R)-3alpha,7alpha,12alpha,24-tetrahydroxy-5beta-cholestanoyl-CoA hydro-lyase [(24E)-3alpha,7alpha,12alpha-trihydroxy-5beta-cholest-24-enoyl-CoA-forming]	1	1	0	0	2	no	yes
ec:2.7.1.158	ATP:1D-myo-inositol 1,3,4,5,6-pentakisphosphate 2-phosphotransferase	1	1	0	0	2	no	yes
ec:1.3.1.70	4,4-dimethyl-5alpha-cholesta-8,24-dien-3beta-ol:NADP+ Delta14-oxidoreductase	1	1	0	0	2	yes	yes
ec:2.4.1.141	UDP-N-acetyl-D-glucosamine:N-acetyl-D-glucosa	1	1	0	0	2	no	yes

	minyl-diphosphodolichol N-acetyl-D-glucosaminyltransferase							
ec:2.4.1.132	GDP-D-mannose:D-Man-beta-(1->4)-D-GlcNAc-beta-(1->4)-D-GlcNAc-diphosphodolichol 3-alpha-mannosyltransferase	1	1	0	0	2	no	yes
ec:2.4.1.257	GDP-D-mannose:D-Man-alpha-(1->3)-D-Man-beta-(1->4)-D-GlcNAc-beta-(1->4)-D-GlcNAc-diphosphodolichol alpha-6-mannosyltransferase	1	1	0	0	2	no	yes
ec:2.4.99.1	CMP-N-acetylneuraminate:beta-D-galactosyl-(1->4)-N-acetyl-beta-D-glucosamine alpha-(2->6)-N-acetylneuraminyltransferase	1	1	0	0	2	yes	yes
ec:1.14.14.1	substrate,reduced-flavoprotein:oxygen oxidoreductase (RH-hydroxylating or -epoxidizing)	1	1	0	0	2	no	yes
ec:1.1.1.100	(3R)-3-hydroxyacyl-[acyl-	1	1	0	0	2	no	yes

	carrier protein]:NADP+ oxidoreductase							
ec:1.1.1.330	(3R)-3-hydroxyacyl-CoA: NADP+ oxidoreductase	1	1	0	0	2	no	yes
ec:1.1.1.62	17beta-estradiol:NAD(P)+ 17-oxidoreductase	1	1	0	0	2	no	yes
ec:2.3.1.78	acetyl-CoA:heparan-alpha -D-glucosaminide N-acetyltransferase	1	1	0	0	2	no	yes
ec:3.1.6.12	N-acetyl-D-galactosamine -4-sulfate 4-sulfohydrolase	0	1	0	1	2	no	yes
ec:1.14.13.8	N,N-dimethylaniline,NAD PH:oxygen oxidoreductase (N-oxide-forming)	1	1	0	0	2	no	yes
ec:3.1.3.62	1D-myo-inositol-hexakisp hosphate 5-phosphohydrolase	1	0	0	1	2	yes	no
ec:1.16.3.1	Fe(II):oxygen oxidoreductase	1	0	0	0	1	no	yes
ec:1.1.2.4	(R)-lactate:cytochrome-c 2-oxidoreductase	1	0	0	0	1	yes	yes
ec:2.7.7.83	UTP:N-acetyl-alpha-D-gal actosamine-1-phosphate uridylyltransferase	1	0	0	0	1	no	yes
ec:1.5.3.16	spermidine:oxygen oxidoreductase	0	1	0	0	1	yes	yes

	(spermidine-forming)							
ec:1.1.99.1	choline:acceptor 1-oxidoreductase	0	1	0	0	1	no	yes
ec:2.7.7.43	CTP:N-acylneuraminate cytidylyltransferase	0	1	0	0	1	yes	yes
ec:1.1.1.30	(R)-3-hydroxybutanoate:NAD+ oxidoreductase	0	1	0	0	1	no	yes
ec:2.6.1.64	L-glutamine:phenylpyruvate aminotransferase	1	0	0	0	1	no	yes
ec:4.4.1.13	L-cysteine-S-conjugate thiol-lyase (deaminating; pyruvate-forming)	1	0	0	0	1	no	yes
ec:1.14.11.1	4-trimethylammoniobutanoate,2-oxoglutarate:oxygen oxidoreductase (3-hydroxylating)	1	0	0	0	1	yes	yes
ec:1.14.17.1	dopamine,ascorbate:oxygen oxidoreductase (beta-hydroxylating)	0	1	0	0	1	no	yes
ec:1.1.1.197	(13E)-(15S)-11alpha,15-dihydroxy-9-oxoprost-13-enoate:NADP+ 15-oxidoreductase	1	0	0	0	1	no	yes
ec:1.17.4.4	phylloquinone:oxidized-dithiothreitol oxidoreductase	0	1	0	0	1	no	yes
ec:2.8.2.11	3'-phosphoadenylyl-sulfate:galactosylceramide 3'-sulfotransferase	0	1	0	0	1	yes	yes

ec:2.5.1.56	phosphoenolpyruvate:N-acetyl-D-mannosamine C-(1-carboxyvinyl)transferase (phosphate-hydrolysing, 2-carboxy-2-oxoethyl-forming)	0	1	0	0	1	yes	yes
ec:2.5.1.57	phosphoenolpyruvate:N-acyl-D-mannosamine-6-phosphate 1-(2-carboxy-2-oxoethyl)transferase	0	1	0	0	1	yes	yes
ec:2.4.1.223	UDP-N-acetyl-D-glucosamine:beta-D-glucuronosyl-(1->3)-beta-D-galactosyl-(1->3)-beta-D-galactosyl-(1->4)-beta-D-xylosyl-proteoglycan 4IV-alpha-N-acetyl-D-glucosaminyltransferase	0	1	0	0	1	no	yes
ec:1.3.1.38	acyl-CoA:NADP+ trans-2-oxidoreductase	0	1	0	0	1	yes	no
ec:2.7.1.164	ATP:L-seryl-tRNASec O-phosphotransferase	0	1	0	0	1	yes	yes
ec:2.1.1.60	S-adenosyl-L-methionine:calmodulin-L-lysine N6-methyltransferase	0	0	0	0	0	no	yes
ec:1.1.1.300	retinol:NADP+	0	0	0	0	0	no	yes

	oxidoreductase							

Appendix 3 *T. mercedesae* chokepoints with corresponding gene and potential drug inhibitors.

Chokepoint (EC)	**Gene ID**	**Drug Bank ID (Prioritized score)**
ec:3.2.1.20	Tm_00843	DB00284 (2); DB00491 (5); DB05200 (0)
	Tm_05918	
ec:3.2.1.3	Tm_05918	
ec:3.5.1.9	Tm_12621	
ec:3.7.1.5	Tm_06425	
ec:3.7.1.2	Tm_09880	
ec:3.5.1.4	Tm_02267	
	Tm_03753	
ec:3.3.2.6	Tm_00789	DB01197 (5); DB02062 (5); DB02352 (5); DB03424 (5); DB05177 (0); DB05745 (0); DB06828 (5); DB06851 (5); DB06917 (5); DB07094 (5); DB07099 (5); DB07102 (5); DB07104 (5); DB07196 (5); DB07237 (5); DB07258 (5); DB07259 (5); DB07260 (5); DB07292 (5); DB08040 (5); DB08466 (5)
	Tm_11048	
ec:3.6.1.11	Tm_10637	
ec:2.7.7.4	Tm_02408	DB02661 (3); DB03708 (4); DB04077 (5)
ec:2.7.1.21	Tm_08821	DB01692 (5); DB02452 (3)
ec:2.4.2.8	Tm_01158	DB00993 (5); DB01033 (4); DB01632 (3); DB02309 (4); DB03153 (3); DB04356 (4)
ec:3.1.3.80	Tm_07781	
	Tm_08070	
ec:3.4.24.84	Tm_05569	
ec:2.4.1.25	Tm_13046	
ec:3.1.2.6	Tm_11056	DB00143 (4); DB03889 (2)
ec:3.5.1.3	Tm_07649	
ec:4.1.1.20	Tm_12236	

ec:1.8.3.1	Tm_02189	DB03983 (2)
ec:3.5.1.14	Tm_00516	
ec:1.1.1.1	Tm_00801	DB00157 (1); DB03704 (4); DB04153 (3)
ec:2.7.7.12	Tm_00753	DB01861 (2); DB03435 (4); DB03685 (5)
ec:3.7.1.3	Tm_10963	DB00114 (5); DB00160 (5); DB07069 (5)
ec:2.7.1.29	Tm_01422	
ec:2.7.1.15	Tm_02191	
	Tm_14118	
ec:1.1.1.18	Tm_03586	
ec:2.1.1.1	Tm_02047	
ec:5.3.2.1	Tm_09349	DB01880 (5); DB02728 (5); DB04272 (5); DB07718 (5); DB07888 (5); DB08333 (5); DB08334 (5); DB08335 (5); DB08765 (5)
ec:4.1.2.10	Tm_07601	
ec:2.7.8.5	Tm_13458	
ec:2.7.1.14	Tm_10741	
ec:6.4.1.3	Tm_05483	DB00121 (5)
	Tm_06093	DB00121 (5); DB00161 (5)
ec:1.1.1.10	Tm_12146	
ec:2.7.1.1	Tm_00258	DB02007 (4); DB02379 (5); DB04395 (2)
	Tm_08929	DB02007 (4); DB02379 (5); DB04395 (2)
ec:1.1.1.145	Tm_12506	
ec:5.3.3.1	Tm_12506	
ec:5.3.99.5	Tm_03434	
	Tm_09150	
ec:2.1.2.5	Tm_05675	DB00114 (5); DB00116 (3); DB00142 (5); DB03256 (3)
ec:4.3.1.4	Tm_05675	DB00114 (5); DB00116 (3); DB00142 (5); DB03256 (3)
ec:4.4.1.1	Tm_03687	DB00114 (5); DB00151 (5); DB02328 (3); DB03928 (5); DB04217 (5)

ec:3.5.1.92	Tm_10750	
ec:1.13.11.5	Tm_05162	
ec:1.1.1.184	Tm_10113	DB03556 (4); DB04463 (5)
ec:1.1.1.189	Tm_10113	DB03556 (4); DB04463 (5)
ec:4.2.1.49	Tm_04128	
ec:2.7.1.52	Tm_12948	
ec:2.7.1.81	Tm_04508	
ec:4.1.1.45	Tm_02358	
ec:4.2.1.107	Tm_06130	DB00157 (1); DB03192 (1)
ec:2.7.1.158	Tm_06668	
ec:1.3.1.70	Tm_09169	
ec:2.4.1.141	Tm_10591	
	Tm_11129	
ec:2.4.1.132	Tm_13702	
ec:2.4.1.257	Tm_13702	
ec:2.4.99.1	Tm_10598	
ec:1.14.14.1	Tm_07004	
ec:1.1.1.100	Tm_04522	DB00157 (1); DB03461 (1)
	Tm_11136	
ec:1.1.1.330	Tm_06870	
ec:1.1.1.62	Tm_06870	
ec:2.3.1.78	Tm_07647	
ec:3.1.6.12	Tm_07869	
	Tm_10384	
ec:1.14.13.8	Tm_01831	
	Tm_04862	
ec:3.1.3.62	Tm_07781	

	Tm_08070	
ec:1.16.3.1	Tm_06139	
ec:1.1.2.4	Tm_07194	
ec:2.7.7.83	Tm_02786	
ec:1.5.3.16	Tm_14357	
ec:1.1.99.1	Tm_12396	
ec:2.7.7.43	Tm_13939	
ec:1.1.1.30	Tm_06091	
	Tm_06558	
	Tm_11978	
ec:2.6.1.64	Tm_00253	DB00114 (5); DB00142 (5)
	Tm_06471	DB00114 (5); DB02142 (5); DB02556 (5); DB04083 (3); DB07950 (5)
ec:4.4.1.13	Tm_00253	DB00114 (5); DB00142 (5)
	Tm_06471	DB00114 (5); DB02142 (5); DB02556 (5); DB04083 (3); DB07950 (5)
ec:1.14.11.1	Tm_11619	
ec:1.14.17.1	Tm_13232	DB00126 (5); DB00822 (5); DB00988 (5)
ec:1.1.1.197	Tm_10113	DB03556 (4); DB04463 (5)
ec:1.17.4.4	Tm_14733	
ec:2.8.2.11	Tm_05445	
ec:2.5.1.56	Tm_13706	
ec:2.5.1.57	Tm_13706	
ec:2.4.1.223	Tm_02354	
ec:1.3.1.38	Tm_10424	DB00173 (4)
ec:2.7.1.164	Tm_06580	
ec:2.1.1.60	Tm_01858	
ec:1.1.1.300	Tm_01780	DB00162 (4)

Appendix 4 Compounds from DrugBank prioritized as chokepoints inhibitors.

DB ID	**Name**	**Criteria 1**	**Criteria 2**	**Criteria 3**	**Criteria 4**	**Criteria 5**	**Score**
DB00284	Acarbose	0	1	0	0	1	2
DB00491	Miglitol	1	1	1	1	1	5
DB05200	AT2220	0	0	0	0	0	0
DB01197	Captopril	1	1	1	1	1	5
DB02062	N-[3-[(1-Aminoethyl)(Hydroxy)Phosphoryl]-2-(1,1'-Biphenyl-4-Ylmethyl)Propanoyl]Alanine	1	1	1	1	1	5
DB02352	3-(Benzyloxy)Pyridin-2-Amine	1	1	1	1	1	5
DB03424	Bestatin	1	1	1	1	1	5
DB05177	DG051	0	0	0	0	0	0
DB05745	CHR-2797	0	0	0	0	0	0
DB06828	5-[2-(1H-pyrrol-1-yl)ethoxy]-1H-indole	1	1	1	1	1	5
DB06851	N-(pyridin-3-ylmethyl)aniline	1	1	1	1	1	5
DB06917	(4-fluorophenyl)(pyridin-4-yl)methanone	1	1	1	1	1	5
DB07094	1-(2,2'-bithiophen-5-yl)methanamine	1	1	1	1	1	5
DB07099	N-[4-(benzyloxy)phenyl]glycinamide	1	1	1	1	1	5
DB07102	(2S)-2-amino-5-oxo-5-[(4-phenylmethoxyphenyl)amino]pentanoic acid	1	1	1	1	1	5
DB07104	4-amino-N-[4-(benzyloxy)phenyl]butanamide	1	1	1	1	1	5
DB07196	N-methyl-1-(2-thiophen-2-ylphenyl)methanamine	1	1	1	1	1	5

DB07237	{4-[(2R)-pyrrolidin-2-ylmethoxy]phenyl}(4-thiophen-3-ylphenyl)methanone	1	1	1	1	1	5
DB07258	(R)-pyridin-4-yl[4-(2-pyrrolidin-1-ylethoxy)phenyl]methanol	1	1	1	1	1	5
DB07259	1-(4-thiophen-2-ylphenyl)methanamine	1	1	1	1	1	5
DB07260	N-benzyl-4-[(2R)-pyrrolidin-2-ylmethoxy]aniline	1	1	1	1	1	5
DB07292	4-(2-amino-1,3-thiazol-4-yl)phenol	1	1	1	1	1	5
DB08040	N-[(2R)-2-benzyl-4-(hydroxyamino)-4-oxobutanoyl]-L-alanine	1	1	1	1	1	5
DB08466	5-[2-(4-hydroxyphenyl)ethyl]benzene-1,3-diol	1	1	1	1	1	5
DB02661	Adenosine-5'-Diphosphate-2',3'-Vanadate	0	1	1	0	1	3
DB03708	Adenosine-5'-Phosphosulfate	1	1	1	0	1	4
DB04077	Glycerol	1	1	1	1	1	5
DB01692	Dithioerythritol	1	1	1	1	1	5
DB02452	Thymidine-5'-Triphosphate	1	1	0	0	1	3
DB00993	Azathioprine	1	1	1	1	1	5
DB01033	Mercaptopurine	1	0	1	1	1	4
DB01632	Alpha-Phosphoribosylpyrophosphoric Acid	1	1	0	0	1	3
DB02309	5--Monophosphate-9-Beta-D-Ribofuranosyl Xanthine	1	1	0	1	1	4
DB03153	3h-Pyrazolo[4,3-D]Pyrimidin-7-Ol	1	0	1	1	0	3
DB04356	9-Deazaguanine	1	0	1	1	1	4
DB00143	Glutathione	1	1	0	1	1	4

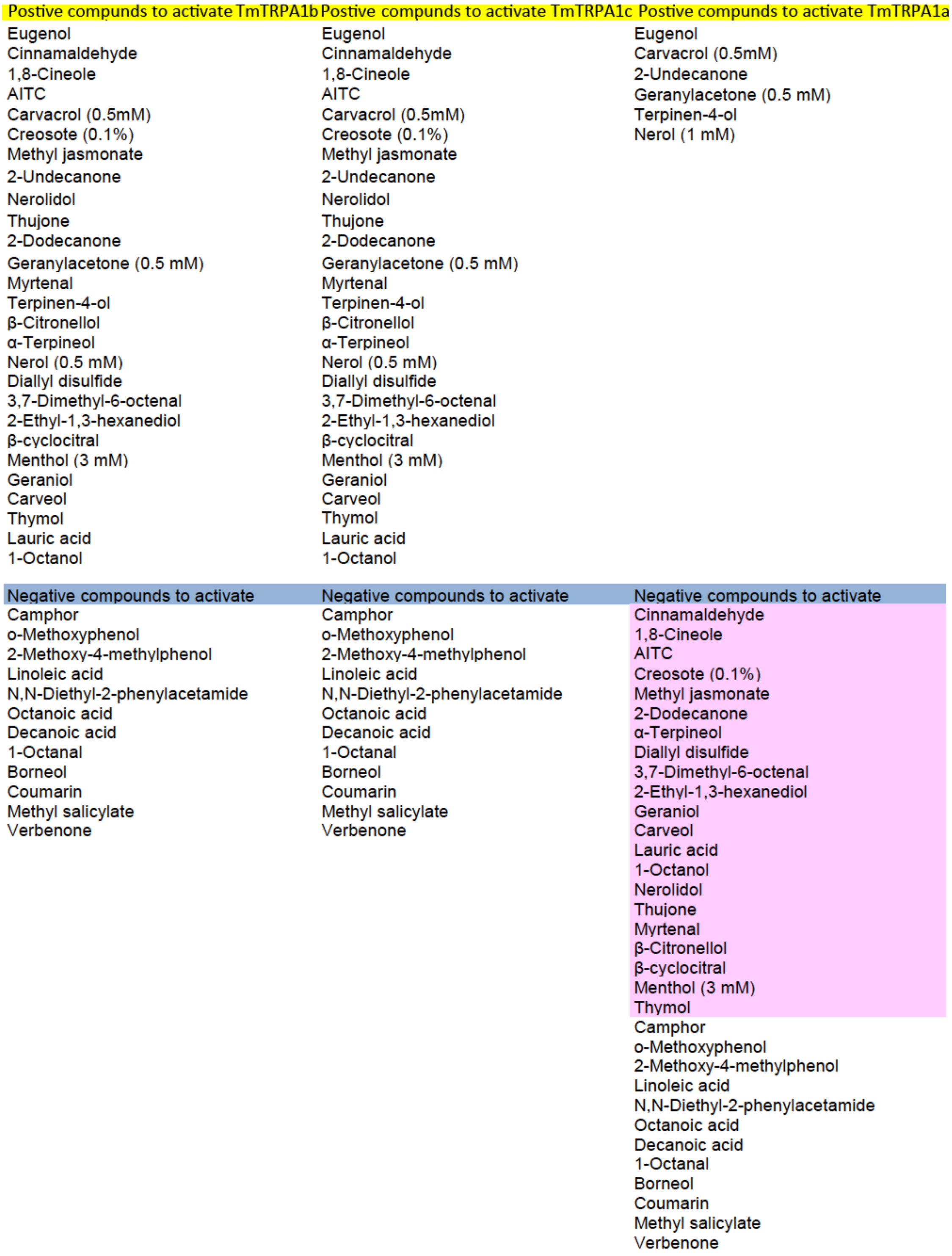
DB03889	S-(N-Hydroxy-N-Bromophenylcarbamoyl)Glutathione	0	0	0	1	1	2
DB03983	(Molybdopterin-S,S)-Dioxo-Thio-Molybdenum(V)	0	1	0	0	1	2
DB00157	NADH	0	0	0	0	1	1
DB03704	12-Hydroxydodecanoic Acid	1	0	1	1	1	4
DB04153	S-Hydroxymethyl Glutathione	1	0	0	1	1	3
DB01861	Glucose-Uridine-C1,5'-Diphosphate	0	1	0	0	1	2
DB03435	Uridine-5'-Diphosphate	1	1	0	1	1	4
DB03685	Uridine-5'-Monophosphate	1	1	1	1	1	5
DB00114	Pyridoxal Phosphate	1	1	1	1	1	5
DB00160	L-Alanine	1	1	1	1	1	5
DB07069	3-Hydroxyhippuric acid	1	1	1	1	1	5
DB01880	3,4-Dihydroxycinnamic Acid	1	1	1	1	1	5
DB02728	7-Hydroxy-2-Oxo-Chromene-3-Carboxylic Acid Ethyl Ester	1	1	1	1	1	5
DB04272	Citric Acid	1	1	1	1	1	5
DB07718	3-(4-HYDROXY-PHENYL)PYRUVIC ACID	1	1	1	1	1	5
DB07888	3-(4-HYDROXYPHENYL)-4,5-DIHYDRO-5-ISOXAZOLE-ACETIC ACID METHYL ESTER	1	1	1	1	1	5
DB08333	4-HYDROXYBENZALDEHYDE O-(CYCLOHEXYLCARBONYL)OXIME	1	1	1	1	1	5
DB08334	3-FLUORO-4-HYDROXYBENZALDEHYDE O-(CYCLOHEXYLCARBONYL)OX	1	1	1	1	1	5

	IME						
DB08335	4-HYDROXYBENZALDEHYDE O-(3,3-DIMETHYLBUTANOYL)OXIME	1	1	1	1	1	5
DB08765	6-HYDROXY-1,3-BENZOTHIAZOLE-2-SULFONAMIDE	1	1	1	1	1	5
DB00121	Biotin	1	1	1	1	1	5
DB00161	L-Valine	1	1	1	1	1	5
DB02007	Alpha-D-Glucose-6-Phosphate	1	1	0	1	1	4
DB02379	Beta-D-Glucose	1	1	1	1	1	5
DB04395	Phosphoaminophosphonic Acid-Adenylate Ester	0	1	0	0	1	2
DB00116	Tetrahydrofolic acid	1	1	0	0	1	3
DB00142	L-Glutamic Acid	1	1	1	1	1	5
DB03256	5-Formyl-5,6,7,8-Tetrahydrofolate	1	1	0	0	1	3
DB00151	L-Cysteine	1	1	1	1	1	5
DB02328	2-[(3-Hydroxy-2-Methyl-5-Phosphonooxymethyl-Pyridin-4-Ylmethyl)-Imino]-5-Phosphono-Pent-3-Enoic Acid	1	1	0	0	1	3
DB03928	Carboxymethylthio-3-(3-Chlorophenyl)-1,2,4-Oxadiazol	1	1	1	1	1	5
DB04217	L-2-amino-3-butynoic acid	1	1	1	1	1	5
DB03556	2-(2-{2-[2-(2-{2-[2-(2-Ethoxy-Ethoxy)-Ethoxy]-Ethoxy}-Ethoxy)-Ethoxy]-Ethoxy}-Ethoxy)-Ethanol, Polyethyleneglycol Peg400	1	0	1	1	1	4
DB04463	3-(4-Amino-1-Tert-Butyl-1h-Pyrazolo	1	1	1	1	1	5

	[3,4-D]Pyrimidin-3-Yl)Phenol						
DB03192	3r-Hydroxydecanoyl-Coa	0	0	0	0	1	1
DB03461	2'-Monophosphoadenosine 5'-Diphosphoribose	0	0	0	0	1	1
DB02142	Pyridoxamine-5'-Phosphate	1	1	1	1	1	5
DB02556	D-Phenylalanine	1	1	1	1	1	5
DB04083	N'-Pyridoxyl-Lysine-5'-Monophosphate	1	0	0	1	1	3
DB07950	1H-INDOL-3-YLACETIC ACID	1	1	1	1	1	5
DB00126	Vitamin C	1	1	1	1	1	5
DB00822	Disulfiram	1	1	1	1	1	5
DB00988	Dopamine	1	1	1	1	1	5
DB00173	Adenine	1	0	1	1	1	4
DB00162	Vitamin A	1	1	1	1	0	4

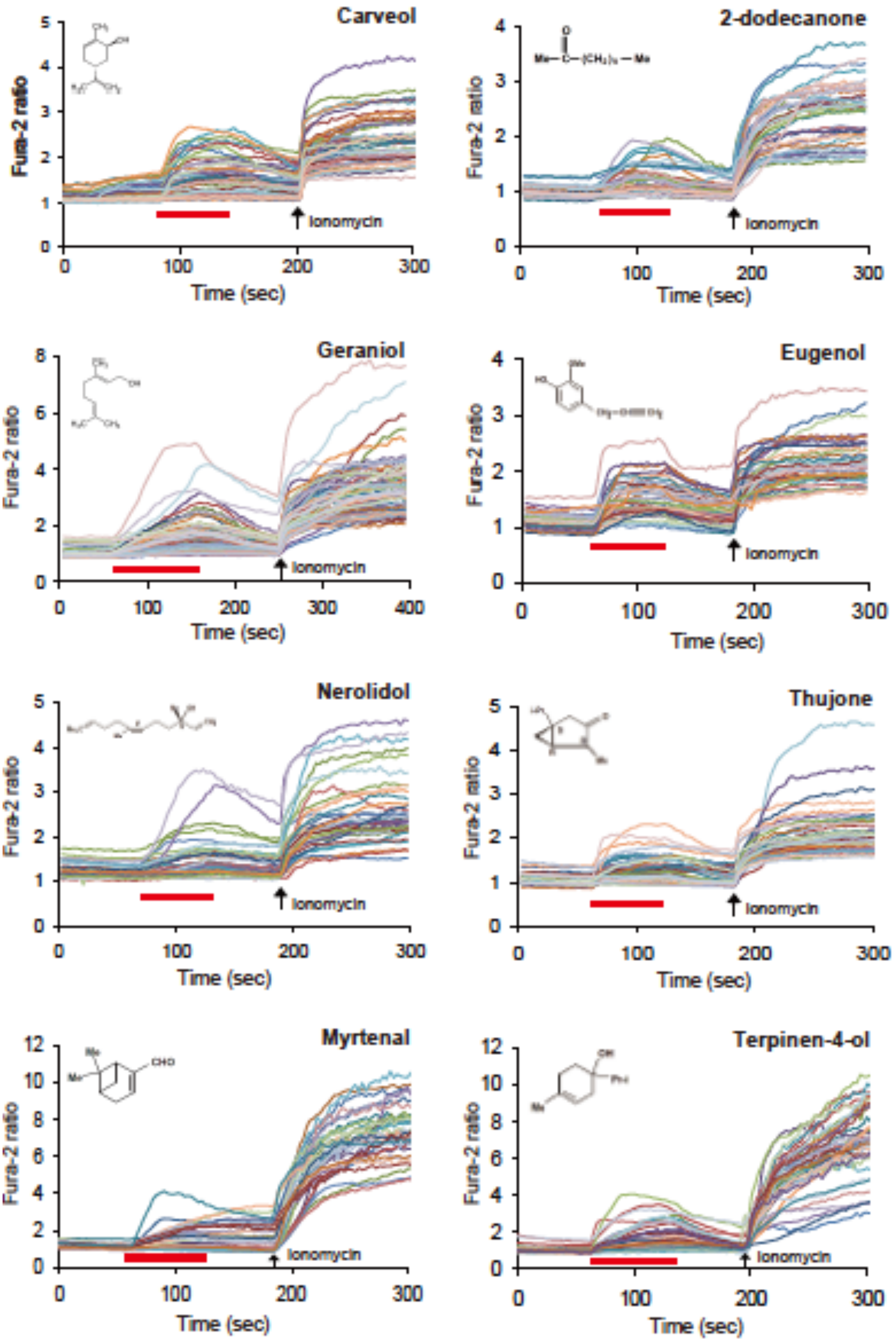
Appendix 5

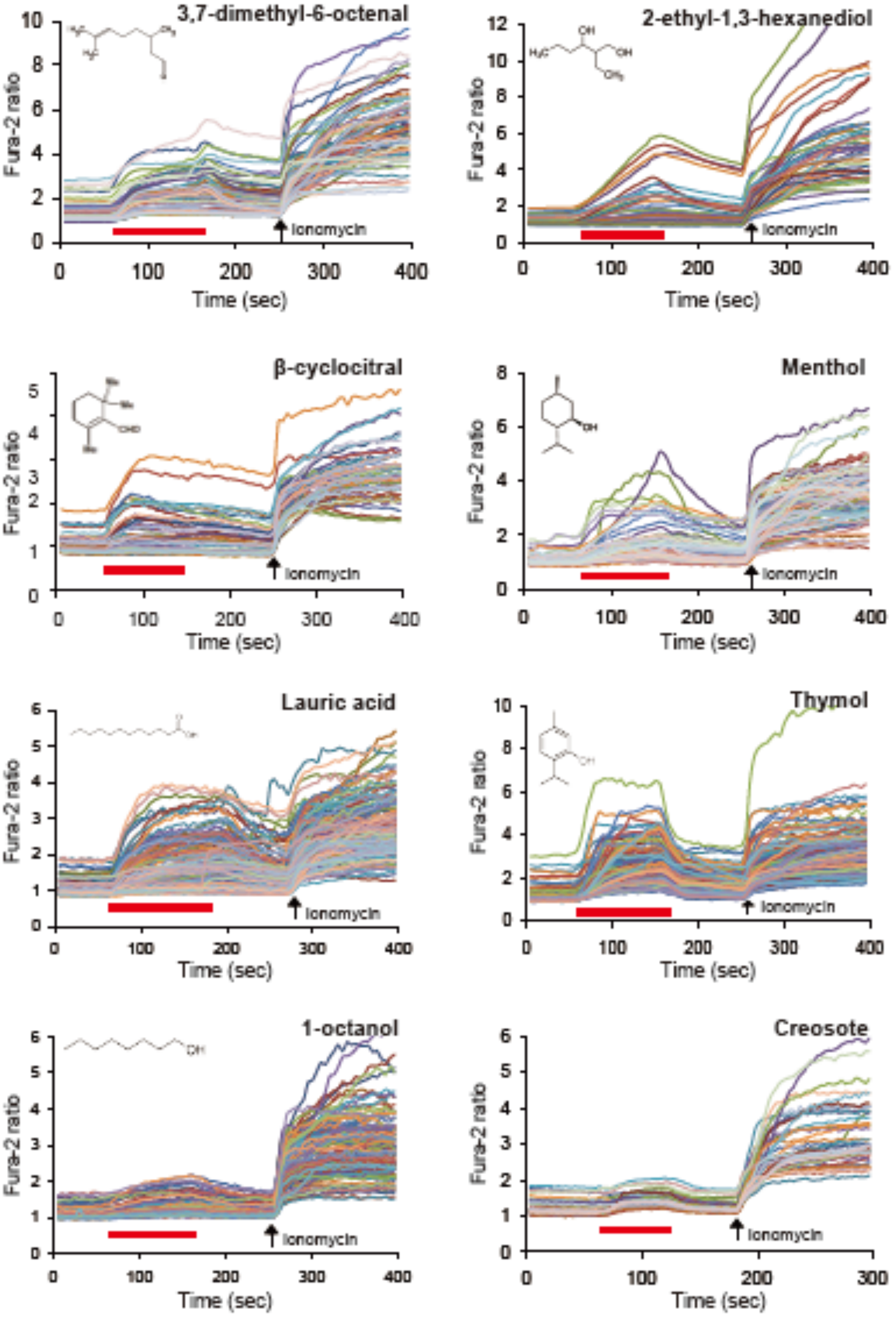
(A)



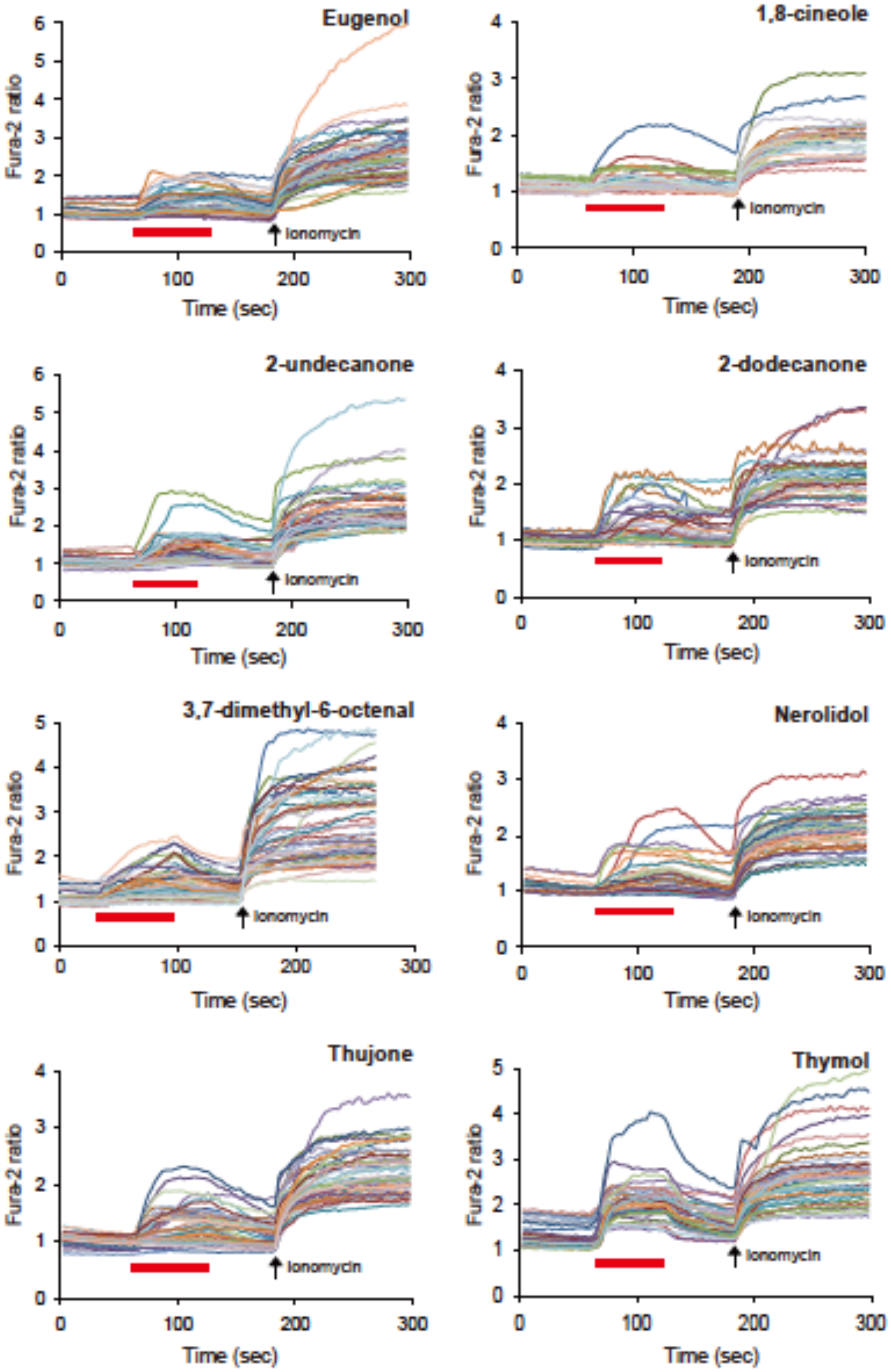
Postive compunds to activate TmTRPA1b	Postive compunds to activate TmTRPA1c	Postive compunds to activate TmTRPA1a
Eugenol	Eugenol	Eugenol
Cinnamaldehyde	Cinnamaldehyde	Carvacrol (0.5mM)
1,8-Cineole	1,8-Cineole	2-Undecanone
AITC	AITC	Geranylacetone (0.5 mM)
Carvacrol (0.5mM)	Carvacrol (0.5mM)	Terpinen-4-ol
Creosote (0.1%)	Creosote (0.1%)	Nerol (1 mM)
Methyl jasmonate	Methyl jasmonate	
2-Undecanone	2-Undecanone	
Nerolidol	Nerolidol	
Thujone	Thujone	
2-Dodecanone	2-Dodecanone	
Geranylacetone (0.5 mM)	Geranylacetone (0.5 mM)	
Myrtenal	Myrtenal	
Terpinen-4-ol	Terpinen-4-ol	
β-Citronellol	β-Citronellol	
α-Terpineol	α-Terpineol	
Nerol (0.5 mM)	Nerol (0.5 mM)	
Diallyl disulfide	Diallyl disulfide	
3,7-Dimethyl-6-octenal	3,7-Dimethyl-6-octenal	
2-Ethyl-1,3-hexanediol	2-Ethyl-1,3-hexanediol	
β-cyclocitral	β-cyclocitral	
Menthol (3 mM)	Menthol (3 mM)	
Geraniol	Geraniol	
Carveol	Carveol	
Thymol	Thymol	
Lauric acid	Lauric acid	
1-Octanol	1-Octanol	
Negative compounds to activate	Negative compounds to activate	Negative compounds to activate
Camphor	Camphor	Cinnamaldehyde
o-Methoxyphenol	o-Methoxyphenol	1,8-Cineole
2-Methoxy-4-methylphenol	2-Methoxy-4-methylphenol	AITC
Linoleic acid	Linoleic acid	Creosote (0.1%)
N,N-Diethyl-2-phenylacetamide	N,N-Diethyl-2-phenylacetamide	Methyl jasmonate
Octanoic acid	Octanoic acid	2-Dodecanone
Decanoic acid	Decanoic acid	α-Terpineol
1-Octanal	1-Octanal	Diallyl disulfide
Borneol	Borneol	3,7-Dimethyl-6-octenal
Coumarin	Coumarin	2-Ethyl-1,3-hexanediol
Methyl salicylate	Methyl salicylate	Geraniol
Verbenone	Verbenone	Carveol
		Lauric acid
		1-Octanol
		Nerolidol
		Thujone
		Myrtenal
		β-Citronellol
		β-cyclocitral
		Menthol (3 mM)
		Thymol
		Camphor
		o-Methoxyphenol
		2-Methoxy-4-methylphenol
		Linoleic acid
		N,N-Diethyl-2-phenylacetamide
		Octanoic acid
		Decanoic acid
		1-Octanal
		Borneol
		Coumarin
		Methyl salicylate
		Verbenone

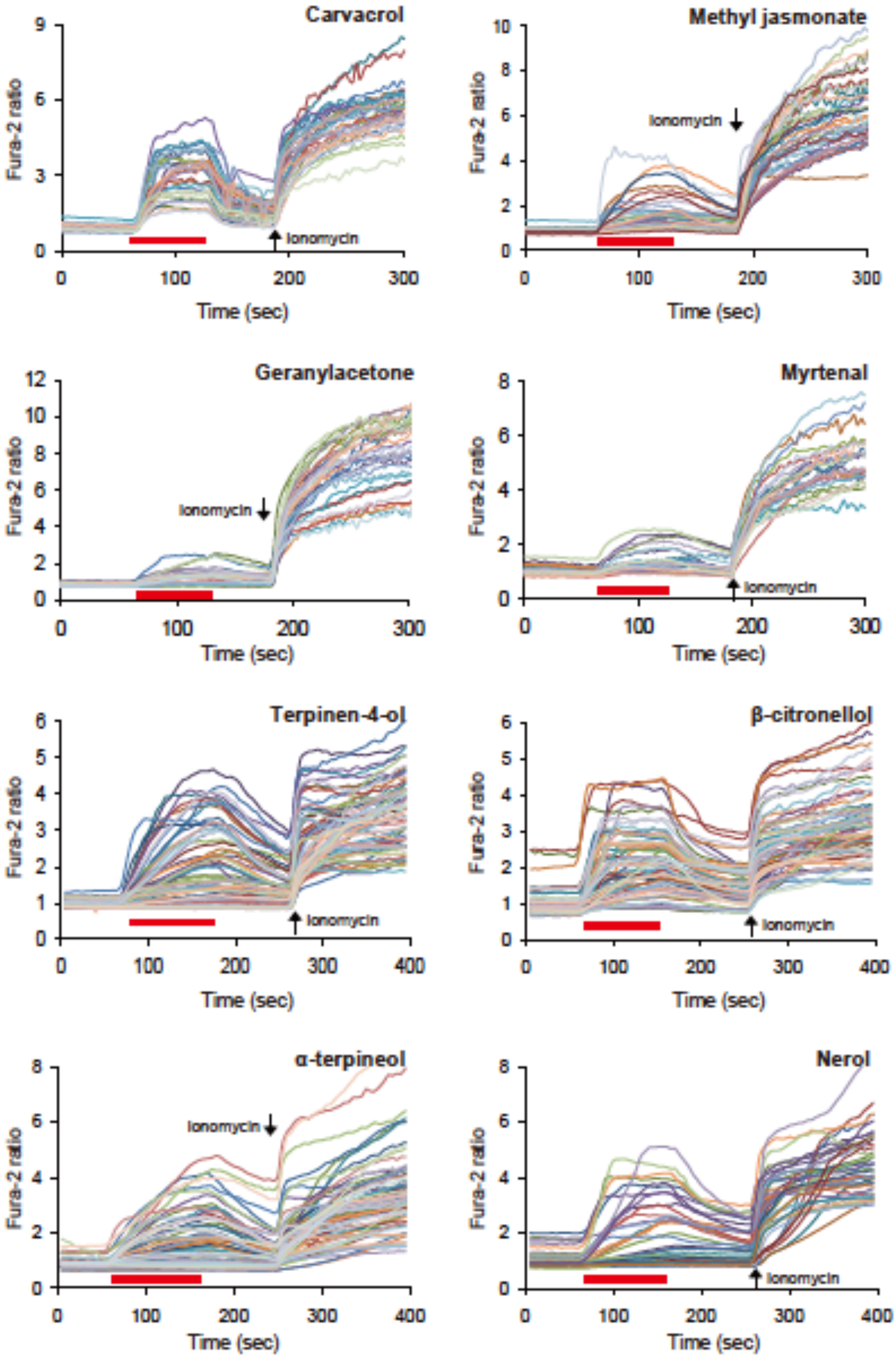
(B)

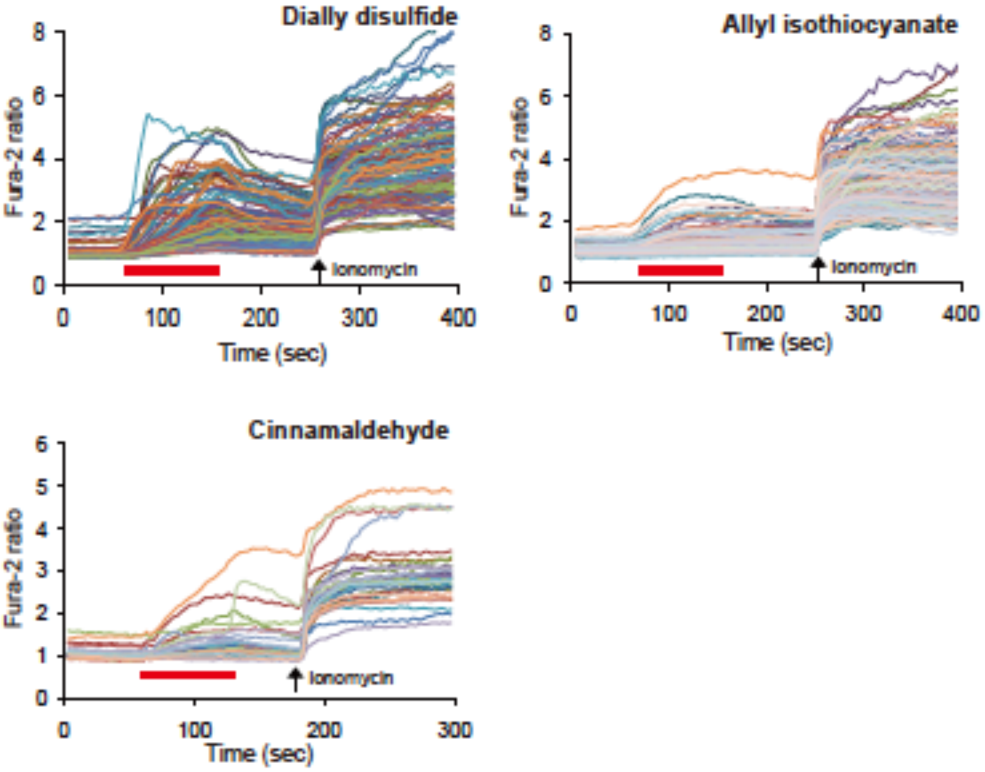

Carveol
Fura-2 ratio
Time (sec)
Ionomycin
2-dodecanone
Fura-2 ratio
Time (sec)
Ionomycin
Geraniol
Fura-2 ratio
Time (sec)
Ionomycin
Eugenol
Fura-2 ratio
Time (sec)
Ionomycin
Nerolidol
Fura-2 ratio
Time (sec)
Ionomycin
Thujone
Fura-2 ratio
Time (sec)
Ionomycin
Myrtenal
Fura-2 ratio
Time (sec)
Ionomycin
CHO
Me
Terpinen-4-ol
Fura-2 ratio
Time (sec)
Ionomycin
OH
Pr-i
Me


3,7-dimethyl-6-octenal
Fura-2 ratio
0
2
4
6
8
10
Ionomycin
Time (sec)
0
100
200
300
400
2-ethyl-1,3-hexanediol
Fura-2 ratio
0
2
4
6
8
10
12
Ionomycin
Time (sec)
0
100
200
300
400
β-cyclocitral
Fura-2 ratio
0
1
2
3
4
5
Ionomycin
Time (sec)
0
100
200
300
400
Menthol
Fura-2 ratio
0
2
4
6
8
Ionomycin
Time (sec)
0
100
200
300
400
Lauric acid
Fura-2 ratio
0
1
2
3
4
5
6
Ionomycin
Time (sec)
0
100
200
300
400
Thymol
Fura-2 ratio
0
2
4
6
8
10
Ionomycin
Time (sec)
0
100
200
300
400
1-octanol
Fura-2 ratio
0
1
2
3
4
5
6
Ionomycin
Time (sec)
0
100
200
300
400
Creosote
Fura-2 ratio
0
1
2
3
4
5
6
Ionomycin
Time (sec)
0
100
200
300

(C)

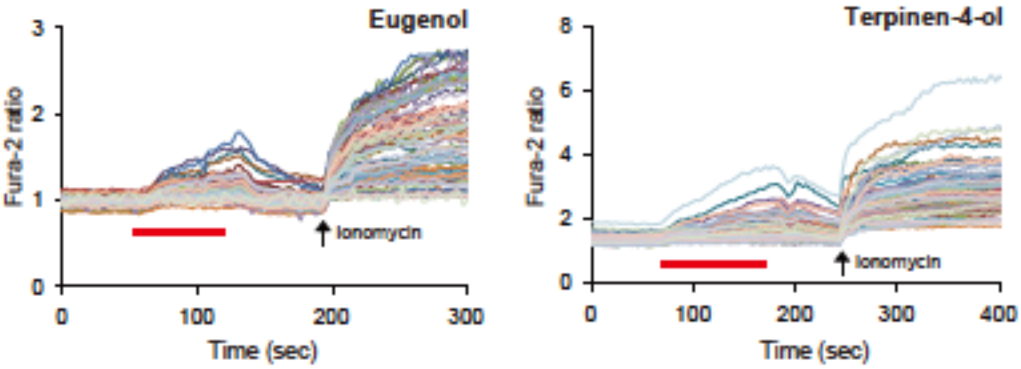

Eugenol
1,8-cineole
2-undecanone
2-dodecanone
3,7-dimethyl-6-octenal
Nerolidol
Thujone
Thymol
Fura-2 ratio
Time (sec)
Ionomycin
0
100
200
300
0
1
2
3
4
5
6


Carvacrol
Fura-2 ratio
0
3
6
9
Ionomycin
0
100
200
300
Time (sec)
Methyl jasmonate
Fura-2 ratio
0
2
4
6
8
10
Ionomycin
0
100
200
300
Time (sec)
Geranylacetone
Fura-2 ratio
0
2
4
6
8
10
12
Ionomycin
0
100
200
300
Time (sec)
Myrtenal
Fura-2 ratio
0
2
4
6
8
Ionomycin
0
100
200
300
Time (sec)
Terpinen-4-ol
Fura-2 ratio
0
1
2
3
4
5
6
Ionomycin
0
100
200
300
400
Time (sec)
β-citronellol
Fura-2 ratio
0
1
2
3
4
5
6
Ionomycin
0
100
200
300
400
Time (sec)
α-terpineol
Fura-2 ratio
0
2
4
6
8
Ionomycin
0
100
200
300
400
Time (sec)
Nerol
Fura-2 ratio
0
2
4
6
8
Ionomycin
0
100
200
300
400
Time (sec)



Dially disulfide
Fura-2 ratio
8
6
4
2
0
Ionomycin
0
100
200
300
400
Time (sec)
Allyl isothiocyanate
Fura-2 ratio
8
6
4
2
0
Ionomycin
0
100
200
300
400
Time (sec)
Cinnamaldehyde
Fura-2 ratio
6
5
4
3
2
1
0
Ionomycin
0
100
200
300
Time (sec)

(D)



Eugenol
Fura-2 ratio
3
2
1
0
Ionomycin
0
100
200
300
Time (sec)
Terpinen-4-ol
Fura-2 ratio
8
6
4
2
0
Ionomycin
0
100
200
300
400
Time (sec)