Identification and Tissue Localization of Olfactory Proteins in the Antenna and Head of *Glossina* Species

Steven Reuben Ger Nyanjom

A thesis submitted in fulfilment for the Degree of Doctor of Philosophy in Biochemistry and Molecular Biology in the Jomo Kenyatta University of Agriculture and Technology

2011

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University

Signature ..... Date 23rd May 2011

Steven Reuben Ger Nyanjom

This thesis has been submitted for examination with our approval as Supervisors

1. Signature ..... Date 6th June 2011

Prof. Peter Onyimbo Lomo

JKUAT, Kenya

2. Signature ..... Date 27th May 2011

Dr. Daniel Khanani Masiga

International Center of Insect Physiology and Ecology (ICIPE), Kenya

3. Signature ..... Date 6th June 2011

Prof. Ahmed Hassanali

Kenyatta University, Kenya

DEDICATION

I would like to dedicate my thesis to my: Wife Christine Anyango and daughters Michelle and Misha; Parents the late Tom Nyanjom and Alice Marenya Nyanjom; Brothers: Robert, Jacob and Byron; Sisters: Jane, Florence and Joan. Your continued support and encouragements have enabled me reach this far and produce this thesis

ACKNOWLEDGEMENTS

I wish to convey my sincere gratitude to the following individuals for providing me with the inspiration and support while pursuing my PhD studies. I thank the almighty God for having kept me throughout the study period. My deepest thanks goes to the supervisors, Prof. Peter Onyimbo Lomo, Dr. Daniel Khanani Masiga and Prof. Ahmed Hassanali, for their constant motivation, valuable help, precious advice and patience throughout the research period.

I extend my gratitude to Prof. Christian Borgemeister, the Director General, for allowing me carry out my research at *Duduville* ICIPE. I reserve my thanks to both Dr. Ellie O. Osir and Dr. J. P. R. Ochieng-Odero, formerly of ICIPE, for the motivation to proceed with the research at the onset of the study.

Special thanks goes to all staff of Molecular Biology and Biotechnology Department. I am indebted to Dr. Paul Odhiambo Mireji and Dr. Martin Rono for detailed constructive comments about the thesis and manuscripts. Thanks to all my lab mates: J. Kabii, M. Kimondo, P. Seda, F. Khamisi, B. Aman, E. Kouam, A. Mebeasealassie, H. Kibogo, E. Machuka, E. Obura, J. Bargul, N. Ndungu, P. Kuria, E. Muchunga, P. Amwayi, C. Ngambi, and E. Waweru for their friendship, support and humour. It was fun sharing the laboratory together and I appreciate the memorable and valuable experience we had.

My gratitude also goes to my employer, Jomo Kenyatta University of Agriculture and Technology. I need to thank the Vice Chancellor, Prof. Mabel Imbuga for

granting me study leave to pursue my PhD studies. My sincere thanks to Prof. Gabriel Magoma (former department chairman), Dr. Daniel W. Kariuki and other members of the department for their support. Special thanks goes to the team at Wellcome Trust Sanger Institute, comprising of Dr. Batt Berriman, Dr. Michael A. Quail and Dr. Christiane Hertz-Fowler for the sequencing and generation of antennal and head ESTs. Without their involvement success of this project won't have been realised.

I acknowledge the pleasure of working with staff at ICIPE, capacity building (Lillian, Lisa and Margaret), colleagues in ARPPIS and DRIP programmes for their moral support and encouragement. Technical assistance of Joseph Saningo and all support staff while collecting tsetse flies at Nguruma are gratefully acknowledged.

I would extend my great appreciation to Netherlands SII Programme and WHO for granting me the scholarship and research grants respectively.

Finally, many thanks to my family members, wife Christine, daughters (Michelle and Misha), mother Alice M. Nyanjom, brothers (Robert, Jacob and Byron), sisters (Jane, Florence and Joan), uncles, aunties, cousins, nephews and nieces, for the patience, understanding and tremendous encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

AAT	Animal African Trypanosomiasis
CSPs	Chemosensory proteins
CAS	cDNA Annotation Software
cDNA	Complementary Deoxyribonucleic acid
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide Triphosphate
ESTs	Expressed Sequence Tags
GO	Gene Ontology
GOBP	General odorant binding proteins
GPCRs	G protein-coupled receptors
HAT	Human African Trypanosomiasis
IRD	Institut de Recherche pour le Dèveloppement
LD-PCR	Long distance Polymerase Chain Reaction
mRNA	Messenger Ribonucleic acid
NECT	Nifurtimox-Eflornithine Combination Therapy
OBPs	Odorant binding protein
ODEs	Odorant degrading enzymes
Ors	Odorant receptors
ORNs	Olfactory receptor neurons
PATTEC	The Pan African Tsetse and Trypanosomiasis Eradication Campaign
PBPs	Pheromone binding protein
RT-PCR	Reverse transcription Polymerase Chain Reaction

ABSTRACT

Tsetse flies use olfaction in search for food, mates and larviposition sites. Olfactory proteins [(odorant binding proteins (OBPs), pheromone binding proteins (PBPs), chemosensory protein (CSPs), odorant degrading enzymes (ODEs) and odorant receptors (Ors)], located within the antennae, play key role in this process. In this work, presence of olfactory proteins was investigated by constructing and sequencing cDNA libraries from *Glossina pallidipes* Austen, antennae; *Glossina palpalis gambiensis* Vanderplank, head and *Glossina tachinoides* Westwood, head. The Expressed sequence tags (ESTs) were clustered using cDNA Annotation™ Software (CAS) and annotated by blast searches. ESTs were generated from the antennal (1127) and head (906 for *G. p. gambiensis* and 830 for *G. tachinoides*) libraries, composed of 296 clusters (18 contigs and 278 singletons for *G. pallidipes* antennae), 305 clusters (36 contigs and 269 singletons for *G. p. gambiensis* head) and 232 clusters (54 contigs and 178 singletons for *G. tachinoides*). The analyses implicated ten (10) sequences in olfaction (2 OBPs from *G. pallidipes*, 5 OBPs from *G. p. gambiensis*, 2 OBPs and 1 CSP from *G. tachinoides*). Clustal alignment revealed a diverse multigene family while phylogenetic analysis supports the existence of different olfactory protein subfamilies.

To identify Dipteran orthologs, the three *Glossina* ESTs (*G. pallidipes* antennae, *G. p. gambiensis* head and *G. tachinoides* head) were clustered using StackPACK, then compared to *G. morsitans morsitans, Drosophila melanogaster, Anopheles gambiae, Aedes aegypti* and *Culex quinquefasciatus* proteomes. A total of 663

clusters for *G. pallidipes* (45 contigs and 618 singletons), 930 clusters for *G. p. gambiensis* (43 contigs and 387 singletons) and 444 clusters for *G. tachinoides* (40 contigs and 404 singletons) were generated. Nine putative OBPs (*G. pallidipes*: 2, *G. p. gambiensis*: 5, *G. tachinoides*: 2) and one putative CSP (*G. tachinoides*) were identified by BLAST search against the dipteran protein databases. Multiple sequence alignments revealed a diverse OBP gene family and a conserved CSP.

Phylogenetic analysis revealed a closely related multigene family that could have evolved separately along different evolutionary time. Reverse Transcription Polymerase Chain Reaction (RT-PCR) screening of male and female *G. pallidipes* tissues (antennae, head, thorax and abdomen) for presence of OBPs and CSPs homologs identified in *G. pallidipes*, *G. tachinoides* and *G. p. gambiensis*, and similar ones (putative OBPs from *G. m. morsitans*), revealed 7 none sex specific (tissue dependent), and 2 sex (male) specific *G. m morsitans* OBP homologues, one specific to the thorax tissue (Gmm_cn14014) and the other to both thorax and abdomen tissues (Gmm_GLAAS20TVB). Two putative OBPs identified in *G. pallidipes* and *G. p. gambiensis* (Gpacontig266 and Gphcontig184) were localised to the antennae tissue. Alignment of the sequenced amplicons revealed a diverse OBP and conserved CSP gene families. These results indicates that olfactory process in tsetse is a complex and interactive process involving established olfactory and non olfactory tissues, suggesting that tissues, other than antennae should also be targeted/considered in development of novel odor based technologies for tsetse control.

CHAPTER 1

1.0 INTRODUCTION AND LITERATURE REVIEW

1.1 Tsetse Fly Systematics and Habitats

Tsetse flies (Diptera: Glossinidae) belong to the genus *Glossina* which contains about 30 living taxa, 22 species and 8 subspecies (Krafsur, 2009). Three main groups occur i.e. *Morsitans* (*Glossina*), *Fusca* (*Austenina*) and *Palpalis* (*Nemorhina*) (FAO, 1982a). *Morsitans* (Savannah flies) include: *Glossina austeni*, *G. longipalpis*, *G. morsitans centralis*, *G. morsitans morsitans*, *G. morsitans submorsitans*, *G. swynnertoni* and *G. pallidipes*. *Fusca* are forest flies and species in this group include: *G. brevipalpis*, *G. fusca congolensis*, *G. fusca fusca*, *G. fuscipleuris*, *G. frezili*, *G. haningtoni*, *G. longipennis*, *G. medicorum*, *G. nashi*, *G. nigrofusca hopkinsi*, *G. nigrofusca nigrofusca*, *G. severini*, *G. schwetzi*, *G. tabaniformis* and *G. vanhoofi*. Riverine flies belongs to *Palpalis* group and examples of species in this group are *G. caliginea*, *G. fuscipes fuscipes*), *G. fuscipes martinii*, *G. fuscipes quanzensis*, *G. pallicera pallicera*, *G. pallicera newsteadi*, *G. palpalis palpalis*, *G. palpalis gambiensis* and *G. tachinoides* (Jordan, 1993).

The three *Glossina* groups are distributed discontinuosly with the *morsitans* group occupying much of the savannah (grassy woodland) of Africa (Figure 1.0). Their distribution appears to be limited by cold winter conditions in the south (Zimbabwe, Botswana), hot dry conditions north of West and Central Africa, scarcity of game animals on which to feed and lack of trees where they rest in the

shade and larviposit (Rogers and Robinson, 2006). The distribution patterns of *fusca* group takes three main types. *G. longipennis* is confined to dry parts of South East Sudan, southern boarder of Ethiopia and north-eastern parts of Somalia, Kenya and Tanzania. *G. brevipalpis* is widely scattered throughout eastern parts of Africa, from Ethiopia and Somalia in the north, to Mozambique and South Africa in the south. The remainder of the *fusca* group are limited to thickly forested areas of Africa with *G. tabaniformis, G. nashi* and *G. haningtoni* being restricted to rain forests; *G. fusca, G. medicorum, G. fuscipleuris* and *G. schwetzi* being forest edge species while *G. medicorummay* living in riverine habitats (FAO, 1982a). *Palpalis* group are found in similar forest habitats throughout very humid areas of Africa and extend into the mangrove swamps, riverine and lakeside forests. Members of this group do not move far away from free water rivers and lakes (Rogers and Robinson, 2006).

Distribution of tsetse flies is influenced mainly by climatic and ecological barriers, vegetation and availability of food. A combination of these factors work synergistically to limit distribution of tsetse in many parts of Africa. In the north of West Africa, characterized by very hot climates, low vegetation cover and dry conditions, tsetse flies are found near rivers with some bushes and trees which provide shade and cooler conditions (FAO, 1982b).

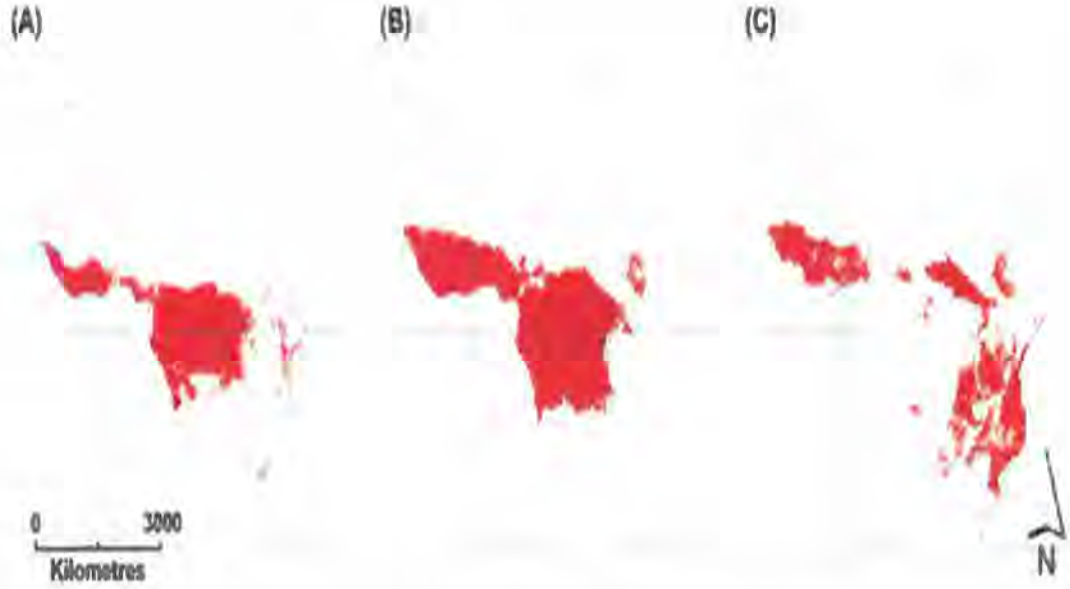

Figure 1.0. Distribution of the three tsetse groups in Africa, A-*Fusca* group; B–*Palpalis* group and C – *Morsitans* group (Cecchi *et al.,* 2008)

1.2 African Trypanosomiasis

Tsetse flies are vectors of trypanosomes (protozoan parasites of the genus *Trypanosoma*), which cause Human African Trypanosomiasis (HAT)/sleeping sickness in humans and Animal African trypanosomiasis (AAT)/nagana in animals. The parasites are cyclically transmitted through the bite of infected tsetse fly or mechanically by other blood sucking arthropods such as horse-flies (Steverding, 2008). The two main forms of HAT are caused by *Trypanosoma brucei gambiense* resulting in chronic infection in countries of western and central Africa and *T. b. rhodesiense* causing acute illness in countries of eastern and southern Africa. Trypanosomes cause pathology in different hosts e.g. *T. vivax* and *T. congolense* are major pathogens of cattle, *T. simiae* causes high mortality in domestic pigs whereas *T. b. brucei* affects all livestock (Mugasa *et al.,* 2008).

The tsetse fly ingests stumpy forms of trypanosomes while taking blood from an infected mammalian host. The ingested parasites move to the fly's midgut where it transform into procyclic forms and upon leaving the midgut develop into epimastigotes which then proceeds to the salivary glands where they develop to the mature metacyclic forms.

The metacyclic trypanosomes get into tissues and blood system of mammalian host along with the saliva from a bite of an infected fly as it takes its blood meal. The parasite proliferates at the site of infection and cause inflammation which later swells. It multiplies in the lymph node and blood stream where they transform into trypomastigotes and later invades the brain and spinal cord (Blum *et al.,* 2001). The trypanosome can then be passed from the blood stream of an infected host to tsetse fly as it feeds. It then undergoes another cycle of development lasting about three weeks before the infection is mature and the fly becomes infective to new hosts (Figure 1.1). After that period the fly remains effective for the rest of its life. In humans, the disease is characterized by general malaise, headache, fever, oedema and anemia (Checchi *et al.,* 2008) and when untreated results into death (Masiga *et al.,* 2002).

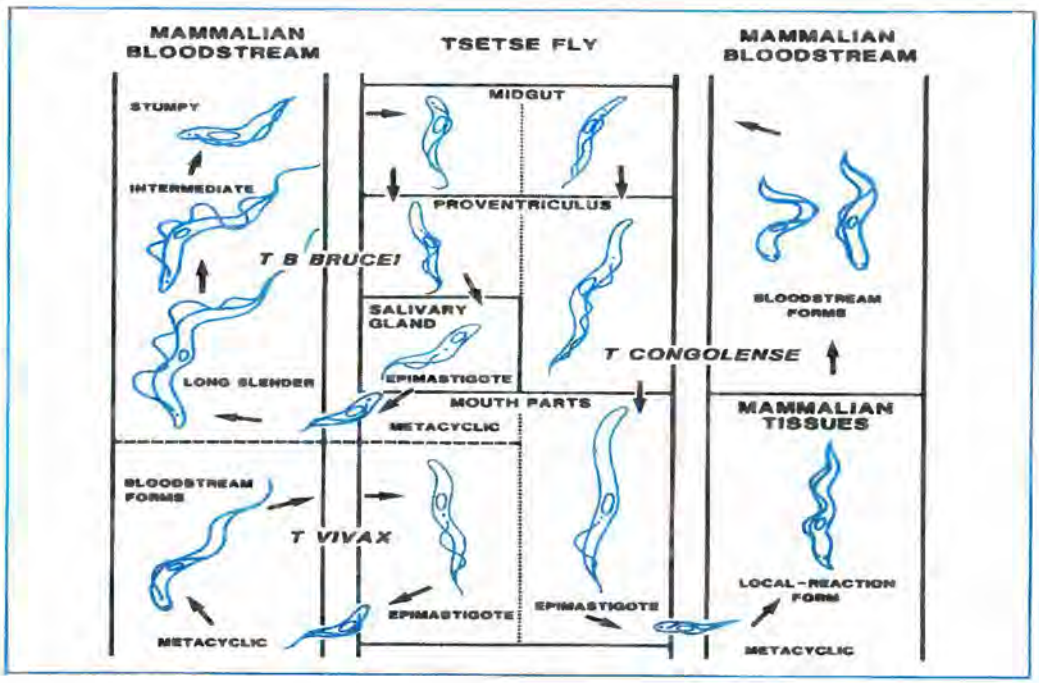

Figure 1.1 Life Cycle of *Trypanosome brucei brucei*, *Trypanosome congolense* and *Trypanosome vivax* in mammalian host and Tsetse vector (Annual Report of the International Laboratory for Research on Animal Diseases, ILRAD 1981).

Trypanosomiasis is endemic in vast region of Africa defined by the infestation range of tsetse vector. The fly inhabits over a third of the continent south of the Sahara where it exposes an estimated 66 million people to the risk of contracting HAT (Maudlin, 2006). It is estimated that 30% of about 150 million cattle in tsetse-affected areas are exposed to the risk of infection (Simarro *et al.,* 2008). Nagana causes about three (3) million cattle deaths every year and farmers are required to administer high doses of expensive trypanocidal drugs, many of which the parasites have developed resistance (Geerts *et al.,* 2001). AAT depresses every aspect of livestock production and limit availability of meat and milk products, leaving much of the human population malnourished.

The disease also reduces productivity, causes abortions and prevents use of draught animal power. The consequences are that people have to rely on manual tillage which lowers food production. Economic losses in agricultural and cattle production are estimated at US$ 4.75 billion and US$ 1.2 billion per year respectively (Simarro *et al.*, 2008). Sub-Saharan Africa is having a high population growth rate and environmental degradation with severe local disruptions in rainfall patterns, perhaps a reflection of climate change on a global scale. As a result poverty levels are high and there is scarcity of food. Effective control of trypanosomiasis could significantly increase livestock production in Africa, as well as greatly benefit arable farming where animals provide most of the traction power in agriculture.

1.3 Trypanosomiasis control

Management of African trypanosomiasis depends on active surveillance, treatment of infected hosts and on vector control. Tsetse control and chemotherapy are the most commonly used methods. The problem of drug resistance is common although new drug targets have been reported (Cross, 2010; Frearson *et al.*, 2010). Different methods have been used to control the tsetse fly. Most of these methods have proven to be unsustainable. Considerable effort have been made to control trypanosomes but the complex interactions of the pathogenic parasite in the vertebrate host and insect vector hinders development of better intervention methods. As a result, to control trypanosomiasis, an integrated approach that utilizes several tools has to be adopted (Simarro *et al.*, 2008).

1.3.1 Chemotherapy

Few drugs are currently available for treatment of HAT and include pentamidine, suramin, melarsoprol, eflornithine, nifurtimox and diminazene aceturate. Pentamidine and suramin are used to treat the first stage of HAT, melarsoprol and eflornithine are effective in treatment of second stage of HAT while nifurtimox and diminazene aceturate are alternative drugs (Priotto *et al.,* 2009). The actual mode of action for these drugs is unknown and proposed mechanisms could involve disruption of multiple cellular processes (e.g. binding to nucleic acid, disruption of kinetoplast DNA, inhibition of RNA-editing and mRNA trans-splicing of trypanosomes) and inhibition of numerous enzymes (Checchi *et al.,* 2007). AAT is primarily controlled in sub-Saharan Africa by three trypanocides: isometamidium chloride, homidium (bromide and chloride) and diminazene aceturate (Geerts *et al.,* 2001). Different treatment strategies are adopted for trypanocidal drug usage which could be routine block treatments, strategic block treatments, monitoring and treatment of infected animals or monitoring and treatment of clinical cases (Priotto *et al.,* 2007). The development of drug resistance (Geerts *et al.,* 2001), compounded by toxicity and high cost of available drugs necessitated the search for improved chemotherapeutics. Treatment of sleeping sickness can be improved by combination of nifurtimox and eflornithine (Priotto *et al.,* 2009) and it has been demonstrated that Nifurtimox-Eflornithine Combination Therapy (NECT) is as effective treatment of second stage of HAT patients (Chatelain and Ioset, 2009).

1.3.2 Vaccine development

Currently, there is no vaccine for trypanosomiasis and its development is hampered by the antigenic variation of trypanosomes (Mehlert *et al.,* 2002). Trypanosomes have complicated bloodstream forms that evade the host's immune response. The antigenic variation is linked to change in composition of a glycoprotein that covers the surface of the trypanosome parasite. The parasite's distinct antigenic glycoprotein enables it to avoid host antibodies raised against their variant surface glycoproteins (VSGs) (Mehlert *et al.,* 2002). Despite considerable research efforts, antigenic variation presents a formidable obstacle towards vaccine development. It is hoped that comparative trypanosome genome studies may lead to identification of potential vaccine targets (Berriman *et al.,* 2005).

1.3.3 Trypanotolerance

Trypanotolerant cattle have been exploited as a method of trypanosomiasis control because the breeds are able to survive and be productive in tsetse-infected areas (Murray *et al.,* 2006). Examples of such breeds include N'Dama and West African short horn (Murray *et al.,* 1982). They are also known to be tolerant to tick infestation, gastrointestinal nematode infection, dermatophilosis and other tropical diseases (Mattioli *et al.,* 2000). However, the animals may succumb to trypanosomiasis under high challenge and physiological stress. These animals have been used as an option for sustainable livestock production (Roelants *et al.,* 1987). The mitigating factors that limit their use are limited distribution of trypanotolerant breeds in West Africa, their relatively small size, inherent low calving rate, low

meat and milk productivity and unsuitability as draught animals (D'Ieteren *et al.,* 1998).

1.3.4 Tsetse vector control

Different methods have been used to control tsetse fly with varying degrees of success. These include discriminative bush clearing and destruction of wild hosts (Rogers *et al.,* 1994), ground and aerial spraying with insecticides (Jordan, 1986), application of sterile insect technique (SIT) (Bailey, 1998; Vreysen *et al.,* 2000), insecticide treated livestock (Barrett, 1997) and insecticide-treated targets (Hargrove, 1980). Control of tsetse populations with insecticides proved to be effective in many areas of Africa (Jordan, 1986). The success of this resulted from rapid treatment of tsetse infested bush by ground and aerial spraying though eradication was not achieved because of re-infestation of cleared areas, high costs of spraying technology and there were also environmental objections about widespread use of insecticides and spread of insecticide resistance (FAO, 1993; Rogers *et al.,* 1994). Sterile insect technique (SIT) was successfully used to eradicate *Glossina austeni* Newstead from Zanzibar (Vreysen *et al.,* 2000). A number of SIT control programs have been undertaken in Africa (Cuisance *et al.*, 1984; Brandl, 1988;) with considerable efforts initiated to implement the SIT program (Kabayo, 2002; Enserink, 2007). Though environmentally friendly, SIT is the most expensive tsetse control method and also requires complex logistics. It works best when preceded by initial suppression of tsetse populations by use of

targets, traps, pour-ons and chemotherapy as was the case in Zanzibar (Vreysen *et al.*, 2000).

Later, 'bait technology' was used to attract flies to visual 'targets' treated with insecticides (Laveissiere *et al.*, 1990; Dransfield *et al.*, 1990). The control method proved highly adaptable and was rapidly adopted by individuals or communities of farmers to control trypanosomiasis. Traps and targets were highly appropriate for small-scale, community-based participatory control operations and above all were environmentally friendly. Despite its relatively low cost and technical simplicity, the bait technology has proved difficult to sustain for poor farmers working without support through external funding. Also, field problems associated with bait control e.g. losses due to theft, damage from humans and wild animals, fire or loss from being washed away in the rainy season, were a draw back to its full use (Hargrove *et al.*, 2000).

The development of 'bait technology' led to the emergence and use of synthetic pyrethroids applied as dip or as a 'pour-on' (Thomson, 1987). Insecticide-treated livestock was regarded as a modification of trap and target technique as Insecticide-treated domestic animals are taken as "moving targets" (Leak *et al.*, 1995; 1996). The method is quite effective when cattle are present in large numbers and are treated on a regular basis. The insecticides also reduce the number of biting flies and ticks (Warnes *et al.*, 1999). Problems associated with this control method included: high cost, environmental concerns which tend to destroy invertebrate

fauna in dung (Vale *et al.,* 2004) and contamination of blood and milk (Bourn *et al.,* 2005). The insect vector remains at the center of hope for development of long term control of trypanosomiasis, as declared by African heads of states in 2000 (Kabayo, 2002). Member states like Ethiopia have made big steps by setting up insectaries to produce sterile males to be used in SIT programs (Enserink, 2007). Availability of genomic data allow studies e.g. host-parasite interactions, which can be used to develop refractory tsetse flies and genes related to olfaction can result in improvement of trapping technologies (Aksoy *et al,* 2005).

The focus of this work was on *G. pallidipes*, a savanna tsetse fly that is the main vector of AAT and whose odor composition of preferred and nonpreferred hosts has been determined (Gikonyo *et al.,* 2002). Synthetic repellants have also been developed and used along with traps to reduce tsetse fly populations (Saini and Hassanali, 2007). Riverine tsetse flies are also important vectors of HAT and recent studies have identified natural odors that can be used to improve performance of traps to control the fly population. Hence, comparative analysis of *G. pallidipes* with riverine species (*G. palpalis gambiensis* and *G. tachinoides*) may help determine the differences in vectorial capacity as well as habitat and host selection preferences.

1.4 *Glossina* Species

1.4.1 Glossina pallidipes

Glossina pallidipes are the principal vectors of nagana with causative trypanosomes being *Trypanosoma congolense*, *T. vivax* or *T. brucei brucei*. In eastern Africa *G. pallidipes* predominates and occurs in many countries including Ethiopia, Sudan, Somalia, Kenya, Tanzania, Uganda, DRC Congo, Zambia, Zimbabwe and Mozambique [Figure 1.2(a)]. It is abundant along the coastal regions and certain river valleys in Somalia (Rogers and Robinson, 2006). Population genetic studies of *G. pallidipes,* supported by allozyme, mitochondrial and microsatellite markers, have shown high levels of genetic differentiation with restricted gene flow (Krafsur and Wohlford 1999; Ouma *et al.,* 2003; 2005) with the southern *G. pallidipes* population having experienced a severe and prolonged bottleneck (Krafsur, 2002). Examples of *G. pallidipes* hosts are bushbuck, warthog, bushpig and buffalo while oribi, impala, waterbuck, hartebeest and reedbuck are rarely fed on (Turner, 1987; Grootenhuis and Olubayo, 1993). Baits, insecticide spraying and SIT have been used to control *G. pallidipes* (Dutoit, 1954).

1.4.2 *Glossina palpalis gambiensis* and *Glossina tachinoides*

Both *G. palpalis gambiensis* and *G. tachinoides* are riverine flies restricted to humid areas, mangrove swamps, rain forest, lake shores and gallery forests along rivers (FAO, 1982a). *G. p. gambiensis* are localized in humid areas of Senegal, Cameroon, along the coast of Angola and north of Mali [Figure 1.2(b)] while *G. tachinoides* are found distributed from Guinea in the west to Central African

Republic in the east, along Sudan-Ethiopia border, northern Nigeria and Chad [Figure 1.2(c)].

Population genetic studies have not been carried out for *G. tachinoides*. However, low allozyme diversities have been reported for *G. p. gambiensis* (Elsen and Roelants, 1999) while microsatellite studies have shown that there is significant genetic variation in X-linked loci, indicating existance of genetic differentiation on macro- and micro-geographic scales (Solano *et al.*, 1999). The main host for both species is man and cattle. Porcupines are preferred by *G. tachinoides* while *G. p. gambiensis* prefer feeding on reptiles and bushbuck. Generally, *G. tachinoides* and *G. p. gambiensis* do not feed on wild pigs like warthog, bushpig, red river hog and giant forest hog but they mainly fed on domestic pigs. Hence, they can also transmit AAT but the disease is less severe. Trypanosome parasites transmitted by the two species include *T. vivax*, *T. congolense* and *T. brucei*. However, *palpalis* group are the princple vectors of sleeping sickness caused by *T. b. gambiense* (Courtis *et al.*, 2005). Insecticide-impregnated targets have been used to control *G. tachinoides* and *G. p. gambiensis* in West Africa and prospects of developing better traps are high with the identification of odors from human, cattle and pigs (Rayaisse *et al.*, 2010).

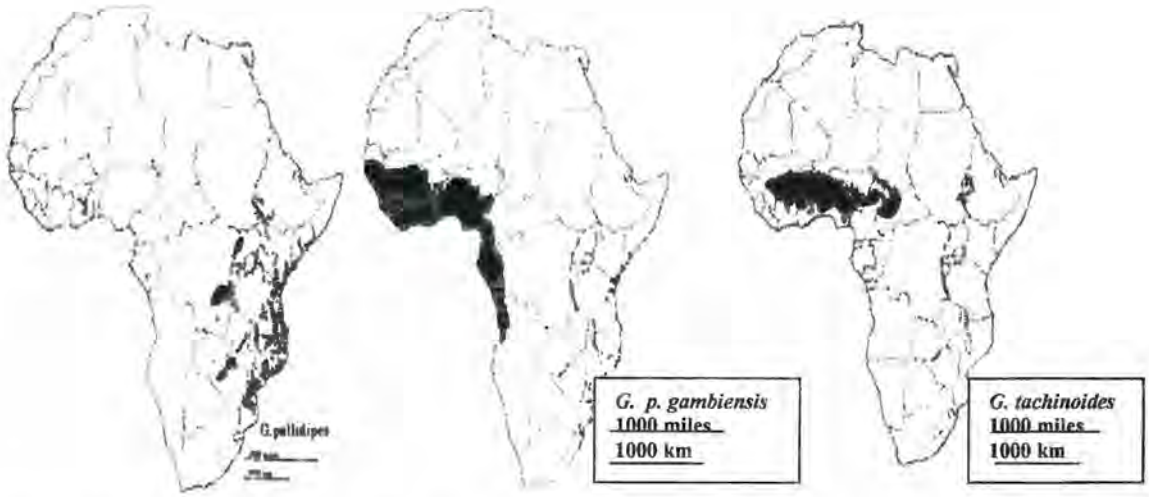

Figure 1.2 Distribution map of *G. pallidipes* (a), *G. palpalis gambiensis* (b), and *G. tachinoides* (c) (Pollock, 1982).

1.5 Tsetse fly Hosts Preference

Tsetse flies exhibit different feeding patterns on specific groups of vertebrate hosts. Animals like warthog, bushpig, cattle, bushbuck, elephant and buffalo are frequently fed on by tsetse flies while others like zebra, giraffe, black rhinoceros, waterbuck, hartebeest and impala are rarely fed on (Clausen *et al.,* 1998). Tsetse locates and recognize preferred host through olfactory and visual cues. The host produce odors which activate the fly beyond its visual range and at close vicinity to the host, the fly is stimulated by visual cues such as shape, size, colour and movement (Leak, 1998). Excretory products (urine, dung, breath and glandular secretions) emit attractive odors to tsetse which are detected by the antennae (Warnes, 1990).

Savannah tsetse flies are attracted by buffalo urine, reported to contain at least 7 phenolic compounds (Hassanali *et al.,* 1986; Madubunyi *et al.,* 1996; Owaga *et al.,* 1988). Body and breath odors are other sources of tsetse attractants (Vale and Hall, 1985) while body odors from Waterbuck repel tsetse flies (Gikonyo *et al.,* 2000). Of the known tsetse fly attractants 4-cresol (4-methylphenol) and 3-n-propylphenol are common to buffalo and Ox (Saini *et al.,* 1993). Guaicol (2-methoxyphenol), 3-isopropyl-6-methylphenol and a series of (C_8-C_{13}) methylketones in waterbuck are moderate repellents (Torr *et al.,* 1996), while 4-methylguaicol (2-methoxy-4-methylphenol) and δ-octalactone are strong repellents (Gikonyo *et al.,* 2002). Feeding response of *G. pallidipes, G. m. morsitans* and *G. fuscipes* to preferred hosts have been studied with a range of natural odors identified (Gikonyo *et al.,* 2002; Omolo *et al.,* 2009). The Identified volatile odors (attractants and repellents) from tsetse hosts find application in providing alternative means that can be exploited to control the fly by developing more potent odors through structural studies (Saini and Hassanali, 2007).

1.6 Insects Olfactory System

Living organisms detect chemicals from their external environment through chemosensory systems, broadly divided into physical (mechanical, sound, vision and temperature) and chemical (gustation and olfaction) systems. Both gustation (sense of taste) and olfaction (sense of smell) systems are similar as they transduce chemical signals into perception (Chandrashekar *et al.,* 2006). The olfactory system performs the complex task of detecting, differentiating the quality and assessing the

concentration of thousands of different odors e.g. aldehydes, esters, ketones, alcohols, alkenes, carboxylic acids, amines, imines, thiols, halides, nitriles, sulphides and ethers. Animals have developed sophisticated olfactory systems to decode chemical information from their environments with great precision (Hildebrand and Shepherd, 1997; Vogt, 2005). The functional organisation of the vertebrate and invertebrate olfactory system is highly conserved (Bargmann, 2006). The simplicity of Insects' olfactory system makes it an attractive and ideal model to study chemoreception (Rutzler and Zwiebel, 2005; de Bruyne and Baker, 2008).

The antennae, main olfactory organ in insects, is responsible for translating the chemical odorant messages into neuronal electrical activities to elicit behavioral-physiological responses. Its surface is innervated by sensilla which are also distributed in other parts of the insect's body and are often used for purposes unrelated to feeding including mechano-, hygro-, and thermo reception (Boeckh *et al.,* 1987; de Bruyne *et al.,* 2001). There occur three morphological and functionally different auxiliary cells per sensillum. These are thecogen, trichogen and tormogen cell. The thecogen encloses the neurons from the axons up to the outer dendritic segment, trichogen cell which is the largest secretes the cuticle of the hair shaft during development while tormogen cell secretes part of the hair base during development (Hansson, 1999). The cuticular forms of sensilla are very variable. For example, in the locust and *Drosophila* there are three distinct types, two of basiconic and one of coeloconic sensilla (Stocker, 1994).

The long hairs with thick walls are trichoid sensilla, short finger-like projections (basiconic pegs), flat plates level with the general surface of the cuticle (plate or placoid sensilla) and short pegs sunk in depressions of the cuticle and opening to the exterior via a relatively restricting opening (coeloconic sensilla) (Wegener *et al.,* 1997). The sensilla contain one or several bipolar olfactory receptor neurons (ORNs) which detect chemical stimuli at the dendritic end and transform it into distinct neuronal signalling at the axonal end. The ORNs arise from epithelial cells and during development, pre-sensillum cluster of cells are formed, some of which move basally to become neurons but three remain apical to form the accessory cells that surrounds the neurons (Endo *et al.,* 2007). The neurons determine the scope and accuracy that odors are identified and processed while accessory cells provide the extracellular environment that supports their function (Park *et al.,* 2002). These cells (neurons and accessory) form sensilla that contain sensillum lymph that bath neuronal dendrites within the lumen (Zacharuk, 1985). The cuticle of sensilla are perforated by numerous small pores which permit entry of chemicals or odors (Steinbrecht, 1997). An ORN is differentiated into three compartments namely: the outer dendrite (exposed to the sensilum lymph), the inner dendrite and the soma (both tightly wrapped by the thecogen cell) and the axon (covered with a glial sheath). ORNs are situated in different types of sensilla each bearing 2-4 ORNs that send their afferent projections to the glomeruli in the antennal lobe, the first olfactory processing center (Stocker, 1994). The axons from the neurons extend without synapses, to the antennal lobes, ending in arborizations that form the olfactory glomeruli. In general, each neuron ends in one glomerulus and all neurons

responding to the same or group of compounds, end in the same glomerulus as it is the case for sensilla on the palps and antennae (Vosshall *et al.*, 2000).

The perireceptor events involved in odor coding range from odorant capture to activation of ORNs with generation of electrical messages through signal transduction processes. Olfactory proteins are involved in odor perception, transport, clearance, signal transduction and include: Odorant binding Proteins (OBPs), Pheromone Binding Proteins (PBPs), Chemosensory proteins (CSPs), Odorant Degrading Enzymes (ODEs) and Odorant receptors (Ors) (Pelosi *et al.*, 2006; de Bruyne and Baker, 2008). The OBPs, PBPs, CSPs and ODEs are expressed in the support cells and secreted into the sensilla lumen (Steinbrecht *et al.*, 1992; Sandler *et al.*, 2000) while ORNs express only one or a few types of Ors (Clyne *et al.*, 1999; Benton, 2006).

1.6.1 Odorant and Pheromone Binding Proteins

Odorant binding Proteins (OBPs) and Pheromone Binding Proteins (PBPs) are a diverse multigene family of antennal specific proteins containing about 130 - 150 amino acids residues (Pelosi and Maida, 1995; Robertson *et al.*, 1999). They are small (14-20 kda) proteins with a signal peptide, a hydrophobic domain between residues 40 - 60 and six conserved cysteines (Prestwich, 1993; Biessmann *et al.*, 2002). A comparison between the amino acid sequences of PBP and two OBPs (general odorant binding proteins (GOBP1 and GOBP2), shows that they are conserved at certain amino acids hence they belong to the same class (Breer *et al.*,

1990; Vogt *et al.,* 1991a,b). GOBPs expressed in both male and female antennae (Breer *et al.,* 1990) and located in the sensilla basiconica (Laue *et al.,* 1994), are more similar to each other (Pelosi and Maida, 1995) while PBPs are mainly expressed in male antennae and located in the pheromone sensitive sensilla (sensilla trichodea) (Raming *et al.,* 1989). The distribution of OBPs in different types of sensilla is related to the odorants detected and associated with distinct classes of ORNs (Vogt *et al.,* 1991a). Both OBPs and PBPs are involved in perireceptor events but differ in the molecules they bind with OBPs binding more general odors while PBPs bind pheromones (Du and Prestwich, 1995). OBPs are synthesized by accessory cells and secreted into sensillum lymph surrounding dendrites projections of ORNs (Steinbrecht *et al.,* 1995). They are thought to bind volatile hydrophobic odorants entering this fluid and transport them to membrane bound odorant receptor (Or) located in the ORNs (Sandler *et al.,* 2000). Insects OBPs have been isolated from a variety of insect species including *Antheraea polyphemus* (Raming *et al.,* 1989), *Antheraea pernyi* (Breer *et al.,* 1990), *Heliothis virescens* (Krieger *et al.,* 1993), *Bombyx mori* (Gong *et al.,* 2009), *Drosophila melanogaster* (McKenna *et al.,* 1994; Pikielny *et al.,* 1994; Kim and Smith, 2001), *Apis mellifera* (Foret and Maleszka, 2006), *Anopheles gambiae* (Biessmann *et al.,* 2002; Xu *et al.,* 2003), *Culex pipens quinquefasciatus* (Pelletier and Leal, 2009), *Aedes aegypti* (Zhou *et al.,* 2008) and *G. m. morsitans* (Liu *et al.,* 2010).

The discovery of great diversity of OBPs in insects suggests that these proteins could ensure the molecular coding of odorants and be the first step in odorant

discrimination (Jacquin-Joly *et al.*, 2001; Biessmann *et al.*, 2002). Although the functional role ascribed to these soluble proteins in the perireceptor events is still unclear, it has been hypothesized that they may have the function to: recognize odors and pheromone, solubilize the hydrophobic odorant/pheromone components in the sensillum lymph (Maibeche-Cosine *et al.*, 1997), transport the odor through the sensillum lymph (Ziegelberger, 1995), mediate its delivery to the specific Ors located in the ORNs (Jacquin-Joly and Merlin, 2004;), degrade, clear (Prestwich, 1993) and act as scavenger protein for receptor deactivation (Rutzler and Zwiebel, 2005).

1.6.2 Chemosensory Proteins

Another soluble protein that is found within the insect's sensilium lymph is chemosensory protein (CSPs). It also contains α-helical domains just like the OBPs, and folded in two different patterns (Lartigue *et al.*, 2002). Unlike, OBPs, CSPs are better conserved with often greater than 50% identical residues even between members of phylogenetically distant species and share no sequence similarity with OBPs (Ozaki *et al.*, 2008). They consist of about 120 amino acid residues and are characterized by four conserved cysteines forming two disulphide links between adjacent residues, resulting in formation of two small loops of eight and four amino acids (Angeli *et al.*, 1999). Despite their different structures, OBPs and CSPs represent two different classes of proteins performing similar roles in perireceptor events by binding and transporting odors (Campanacci *et al.*, 2003; Zhou *et al.*, 2004). They are also expressed in non-chemosensory tissues implying

that they are involved in functions other than olfaction (Jacquin-Joly *et al.,* 2001; Sabatier *et al.,* 2003; Gong *et al.,* 2007). Chemosensory proteins may also be implicated in immune response, circadian cycles or development (Sabatier *at al.,* 2003).

The first identified CSPs were OS-D proteins in *Drosophila melanogaster* antennae (McKenna *et al.,* 1994; Pikielny *et al.,* 1994). Several CSPs have been identified in different insects orders which include Lepidoptera (Jacquin-Joly *et al.,* 2001; Picimbon *et al.,* 2001; Gong *et al.,* 2007), Orthoptera (Angeli *et al.,* 1999; Ban *et al.,* 2003), Hymenoptera (Briand *et al.,* 2002; Ishida *et al.,* 2002), Blattoidea (Kitabayashi *et al.,* 1998; Picimbon and Leal, 1999), and Hemiptera (Jacobs *et al.,* 2005).

1.6.3 Odorant Degrading Enzymes (ODEs)

Odors and pheromones are xenobiotic substances and once they have activated Ors they need to be degraded to prevent interference with the recognition system. Odorant degrading enzymes (ODEs) are involved in the deactivation and are found in high levels within the sensillum lymph (Pelosi, 1996). In Insects, these detoxifying enzymes specific for species' sex pheromones have been identified and include esterases, oxidases and glutathione transferases. (Vogt and Riddiford, 1981; Vogt *et al.,* 1985; Rybczynski *et al.,* 1990). In *Drosophila* a NADPH-cytochrome P450 oxidoreductase is thought to be involved in olfactory clearance (Hovemann *et al.,* 1997). In *M. sexta* the pheromone consists of two or more aldehydes

(Tumlinson *et al.*, 1989) and the aldehyde oxidase is presumably capable of metabolising the entire signal. This is in contrast to *A. polyphemus* pheromone which is a mixture of an acetate ester and an aldehyde. The esterase is known to degrade the acetate ester component of the pheromone (Vogt *et al.*, 1985) while the aldehyde component is broken down by an aldehyde oxidase (Rybczynski *et al.*, 1990). Similarly, the silk moth *B. mori* pheromone consists of an alcohol (bombykol), degraded by either an oxidase or dehydrogenase (Kasang *et al.*, 1989) and an aldehyde (bombykal), degraded by an aldehyde oxidase (Rybczynski *et al.*, 1990).

The detoxifying mechanism is varied and in several Lepidoptera species, the functional group of the pheromone is first modified into a more hydrophilic one, with consequent increase of the compound's solubility. The modified pheromone is not capable of further excitation of the ORNs resulting in termination of the signal (Prestwich *et al.*, 1989; Vogt *et al.*, 1990). In *A. polyphemus*, the pheromone interacts with the receptor in its reduced form and is then converted to the oxidized form thereby deactivating the pheromone bound to the PBP (Kaissling, 1998).

1.6. 4 Odorant Receptors (Ors)

Odorant receptors (Ors) are a multigene family of about 370 – 400 amino acids that belong to the seven transmembrane G protein-coupled receptors (GPCRs) superfamily (Mombaerts, 1999) and are expressed in ORNs (Benton, 2006). The GPCRs are involved in a variety of cellular processes including hormonal

regulation, neurotransmission and photoreception. It is a protein complex that has a membrane receptor for external signal reception, a heterotrimeric transducer (G protein) and one of several effector enzymes (e.g. phospholipase C or adenyl cyclase) (Breer, 2003). The G-proteins comprise of α, β and γ subunits. The α subunit is responsible for GTP binding and hydrolysis to GDP hence the G-proteins are generally referred to by this α subunit (Kalidas and Smith, 2002). The receptors have seven transmembrane regions with the C-terminus of the protein situated inside the cell cytoplasm and the N-terminus situated outside the cell, defining an external recognition domain (Buck and Axel, 1991).

In *Drosophila*, ORNs express one specific Or to the cell type and a ubiquitous one, Or83b, whose protein product dimerizes with the specific receptor and mediates its transport to olfactory cilia (Larsson *et al.,* 2004). The ORNs expressing the same Or converge on a common glomeruli in the antennal lobe. The number of Or genes reported varies with *Drosophila* having 62 Ors encoded by 60 genes (Clyne *et al.,* 1999) while the *Anopheles* mosquito has 85 Ors identified (Fox *et al.,* 2001). Generally, insects Ors are very divergent with little sequence conservation within and across insect orders and species (Hill *et al.,* 2002). However, *Heliothis* receptor HR2 shares a high degree of sequence identity (60–80%) with *Drosophila* (Or83b) and *Anopheles* (AgamGPROr7). Different orthologs of this unique Or subtype have been identified in several insect species which include moths (*Antheraea permyi, Bombyx mori*), honey bee (*Apis mellifera*), blowfly (*Calliphora erythrocephela*), yellow mealworm (*Tenebrio molitor*) (Krieger *et al.,* 2003), *M. brasicae* (Jacquin-

Joly *et al.*, 2001) and yellow fever mosquito, *Aedes aegypti* (Melo *et al.*, 2004). In many insect species, Or83b is highly conserved, suggesting it could be involved in unique and essential functions in insect olfaction (Jones *et al.*, 2005). It is thought to be involved in Or activation as a dimerization partner (Vosshall *et al.*, 2000; Krieger *et al.*, 2003).

1.7 Mechanism of Olfactory Signal Transduction

Olfactory signal transduction is widely conserved across many organisms, including mammals, fish, crustaceans, nematodes and insects (Hildebrand and Shepherd, 1997; Krieger and Breer, 1999). Although many olfactory systems are complex, there is a general agreement that the transduction processes are initiated by highly specific interaction between an odor molecule and Ors on the dendrites of ORNs. Ors play a dual role, first they discriminate different odors and secondly, they transfer chemical message from extracellular to intracellular face of the ORN upon binding the ligand (Buck, 1996; Hildebrand and Shepherd, 1997). In mammals and nematodes binding of odorant to Ors activates G protein signalling cascade. This stimulates adenylate and guanylate cyclase, resulting in increase intracellular concentrations of cyclic AMP (cAMP) and GMP (cGMP) in mammals and nematodes respectively. Binding of either cAMP or cGMP activates and opens the cyclic nucleotide-gated (CNG) channels, allowing cations to enter into the neurons, producing an action potential that travels down the axon to the brain (Pellegrino and Nakagawa, 2009).

Before insect Or genes were identified, it was proposed that GPCR-mediated second messenger pathway was involved based on identification of phospholipase C and inositol-1,4,5-trisphosphate (IP_3) (Krieger and Breer, 1999). Increase in IP_3 concentration caused opening of IP_3-dependent Ca^{2+} channels in the outer dendrite membrane, causing Ca^{2+} influx in the cell. These inward currents form a receptor potential that elicits a discharge of action potentials which travel down ORN axons to antennal lobe where olfactory message is decoded (Krieger *et al.,* 1997). Recent studies provide evidence that insect Ors are ligand-gated non-specific cation channels. Activation of G-protein by ligand-bound receptor stimulates adenyl cyclase. This generates cAMP which binds intracellular side of CNG channel causing an influx of Na^+ and Ca^{2+}, resulting in depolarizaton of ORNs (Wicher *et al.,* 2008). Divergent views also propose a model involving a ligand-gated ion channel, formed by Or83b/Or complex, that is directly opened by binding of odorants to cause influx of cations (Sato *et al.,* 2008). These proposals provide new insights on insects olfactory signal transduction mechanisms involving either ion or ligand-gated channels.

1.8 Justification

Effective control of tsetse flies depends on different complementary approaches that entails an integrated approach. Chemotherapy and tsetse vector control are the most widely used methods despite limitations linked to development of resistance, high costs and degradation to the environment. The use of *Bacillus thuringiensis* (Bt) toxin genes has been successfully applied in crops as an alternative insecticide

to control insect pests. However, Bt toxins mostly target lepidopteran pests and there is need to develop strategies that reduce current reliance on insecticides to control insect vectors. One potential approach is to interfere with the tsetse olfactory system because of the vital role it performs in foraging for food, mates and sites to deposit their larvae. Determining molecular and structural aspects of olfactory proteins is an active area of research in insects chemical ecology as it promises alternative approach to controlling insects by interfering with its ability to find suitable mates and hosts through volatile chemical odors (semiochemicals). Such olfactory based approaches have been applied successfully in 'push-pull' strategies with plants that produce repellents (push) and traps that are blended with attractants (pull) to control insect pests. An alarm pheromone synthase gene from aphid has also been cloned into *Arabidopsis thaliana* to produce plants that repelled aphids and attracted beneficial insects. Thus, there is also need to develop environmentally friendly control strategies for insect vectors of human and animal diseases, such as mosquitoes, sand flies, horn flies and tsetse flies.

Tsetse flies use olfaction to find mates for reproduction and locate hosts for nutrition by detecting volatile semiochemicals in their external environment. The antennae is the main olfactory organ that allows odors to enter through pores and are transported to the olfactory receptor neurons (ORNs) through the sensillum lymph. A lot of research has been done to study insect olfaction system at the molecular level to help identify olfactory proteins and functional components involved in signal recognition and transduction mechanism. The main olfactory

proteins include Odorant binding Proteins (OBPs), Pheromone Binding Proteins (PBPs), Chemosensory proteins (CSPs), Odorant Degrading Enzymes (ODEs) and Odorant receptors (Ors). It is important to undertake studies that will help identify olfactory genes in different tsetse species and determine their functions as it will contribute towards understanding the crucial role olfaction plays in tsetse chemical ecology as well as genomic divergence that may underline behavioural and physiological differences among the *Glossina* groups. Such information will be useful in developing novel and innovative ways of controlling tsetse flies to interfere with host interactions and mating behavior.

This study focussed on odorant binding proteins (OBPs) which are proposed to be the first molecules involved in odor perception during perireceptor events. *G. pallidipes* was selected because it is the principal vector of nagana and transmits *T. congolense*, *T. vivax* or *T. brucei brucei.* Moreover, its feeding patterns to preferred and nonpreferred hosts has been documented with a range of attractants and repellents already identified. The need to compare OBPs from riverine and savannah tsetse vectors necessitated the inclusion of *G. tachinoides* and *G. palpalis gambiensis* in this study. The latter two species are important riverine vectors of the disease trypanasomiasis. Thus, the three selected species make a good set for comparative studies because of the differences in their ecological habitats and host animal preference.

1.9 General Objective

To identify and characterise olfactory proteins in antennae of *G. pallidipes* and heads of *G. palpalis gambiensis* and *G. tachinoides*.

1.9.1 Specific Objectives

1. To profile the expression of genes in olfactory organs of *G. pallidipes, G. tachinoides* and *G. palpalis gambiensis.*

2. To identify putative olfactory protein homologs using bioinformatic approaches.

3. To assess tissue specificity of identified olfactory proteins in different body organs.

CHAPTER 2

2.0 IDENTIFICATION OF PUTATIVE OLFACTORY PROTEINS IN *GLOSSINA* SPECIES (DIPTERA: GLOSSINIDAE)

2.1 Introduction

This chapter reports analysis of OBPs and CSP obtained from sequencing of *G. pallidipes* antennal, *G. p. gambiensis* and *G. tachinoides* head libraries. The characteristic features (number of amino acids, molecular weight and isoelectric point) of the identified OBPs and CSP were predicted and then the OBPs and CSP compared with homologs identified in other insects. Phylogenetic relationship of the putative OBPs and CSP was also determined.

2.2 Materials and Methods

2.2.1 Study area and preparation of test insects

Adult *G. pallidipes* were sampled in Nguruman, Kajiado District, Kenya (1° 50'S; 36° 05'E) (between 14th and 24th July 2005), using NGU traps baited with acetone and cow urine (Brightwell *et al.*, 1991), (Figure 2.0). Acetone was dispensed from 200 ml bottles each with a 2 mm diameter hole in the lid while cow urine was dispensed from suitable tins with the top covered by polythene and a 2 X 4 cm slot cut in the tin just below the rim. Cages were emptied after 24 hours catch. The area consists mostly of Acacia woodland with thickets along stream beds and is infested mainly by *G. pallidipes* Austen and smaller populations of *G. longipennis* Corti. Identification of *G. pallidipes* from other *Glossina* species was based on morphological characters that distinguish the species from *G. longipennis* (FAO,

1982a). The *G. pallidipes* sampled (4500) were immediately preserved in RNA*later (*Ambion Inc, Austin, TX) and transferred to the laboratory for analysis. *Glossina p. gambiensis* and *G. tachinoides* adults were reared from 250 and 220 pupae respectively, and exposed to acetone and cow urine for ten (10) minutes, all under standard laboratory conditions (temperature 25 ± 1°C; relative humidity 75 ± 10%; fed on sterilized pig blood after every 24 hours using an artificial membrane) in the insectary at ICIPE (Moloo, 1971). The pupae were kindly donated by Dr. Philippe Solano of Institut de Recherche pour le Dèveloppement (IRD), Bukina Faso.

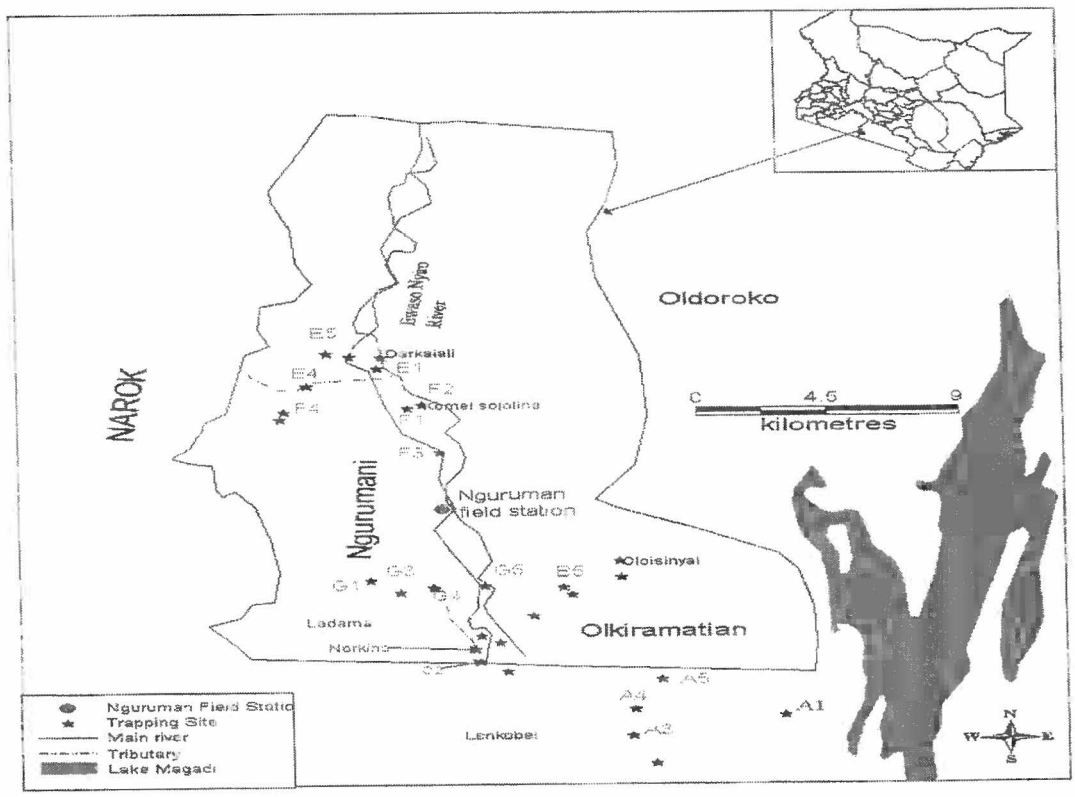

Figure 2.0 A map of Nguruman showing sampling sites (Dransfield *et al.,* 1990)

2.2.2 Construction of *G. pallidipes, G. p. gambiensis* and *G. tachinoides* cDNA libraries

Antennae were isolated from *G. pallidipes* (2000 male and female) (without sexing) using the standard methods of Pollock, (1982), and stored in RNA*later* (Ambion Inc, Austin, TX). Heads were similarly, isolated from *G. p. gambiensis* (100 males and 93 females) and *G. tachinoides* (93 males and 75 females) and stored at -80°C. Antennae were used for *G. pallidipes* as it is the main olfactory organ while head was used for *G. p. gambiensis* and *G. tachinoides* as it contains the antennae and other sensory appendages including mouthparts and sensilla. Poly (A) RNA were isolated from antenna (*G. pallidipes*) and heads (*G. p. gambiensis* and *G. tachinoides*) tissues using MicroPoly(A) Purist kit (Ambion Inc, Austin, TX). The process involved disruption and homogenization of the tissues followed by centrifugation at 4°C for 15 minutes. The clear lysate was transferred to a fresh tube and mixed with Oligo(dT) cellulose at room temperature. Centrifugation was done at 4,000g for 3 minutes to pellet the Oligo(dT) cellulose followed by two washes with lysate wash buffer.

The poly(A) RNA was then eluted with 200 μl of THE RNA Storage Solution and centrifuged at 5000g for 2 minutes. Final Ploy(A) RNA selection was carried out by re-suspending the Oligo(dT) cellulose then adding it to the eluted poly(A) RNA after which the mixture was heated for 5 minutes at 75°C. The tube was then rocked gently for 15 minutes at room temperature and Oligo(dT) cellulose

transferred back to the spin column followed by centrifugation at 5,000g for 20 seconds. After two washes, the poly(A) RNA was eluted into a fresh collection tube. To precipitate the eluted poly(A) RNA, 20 μl of 5M ammonium acetate, 1 μl glycogen and 550 ul absolute ethanol were added and the mixture left at -20°C overnight. RNA was recovered by centrifugation at 12,000g for 30 minutes at 4°C and pellet re-suspended in 5 μl THE RNA storage solution. The quality of the poly(A)RNA was checked on a 0.3% agarose/ethidium bromide (0.1 μg/ml) gel run at 60V for two (2) hours.

Complementary cDNA (cDNA) was constructed from respective RNAs using SMART cDNA library synthesis kit (BD Biosciences, Franklin Lakes, NJ) according to the manufacturer's instructions. Briefly, single strand cDNA was synthesized from 3 μl poly(A) RNA (129μg/ml) using PowerScript reverse transcriptase, 1 μl SMART IV oligonucleotide and 1 μl CDS III/3' PCR primer. The tube contents were mixed, spun briefly, then incubated at 42°C for 1 hour in a 9800 Fast Thermal cycler (Applied Biosystems, Austin, TX) and then placed on ice to terminate first strand synthesis. Second-strand synthesis was performed by long distance (LD) PCR protocol using 5' PCR primer and CDS III/3' primer as sense and anti-sense primers respectively. The two primers also create *SfiI A* and *B* restriction enzyme sites at the end of nascent cDNA. Two identical amplification reactions were set up using 3 ul of first cDNA synthesis product. To both tubes the following were added in order: 79 ul PCR grade water, 10 ul 10X advantage 2 PCR

buffer, 2 ul 50X dNTP mix, 2 ul 5' PCR primer, 2 ul CDS III/3' PCR primer and 2 ul 50X advantage 2 polymerase mix, to make a total volume of 100 ul.

The contents were placed in a 9800 Fast Thermal cycler (Applied Biosystems, Austin, TX) with the following PCR program: 95°C for 1 minute, 30 cycles – 95°C for 10 seconds; 68°C for 6 minutes.

Integrities of the cDNAs were verified on 1.2% agarose gel (stained with ethidium bromide (0.5μg/ml)), immediately treated with 4 ul of proteinase K (20 ug/ul) at 45°C for 20 minutes and subsequently concentrated to a final volume of 79 ul using Microcon YM-100 column (Millipore, Bedford, MA). The cDNA concentrates were digested with *SfiI* restriction enzyme at 50°C for 2 hours and the resultant fragments size-fractionated on a ChromaSpin-400 drip column (BD Biosciences, Franklin Lakes, NJ). Validated profiles on 1% agarose gel (stained with ethidium bromide (0.5μg/ml)) revealed large (> 3000bp), medium (500bp to 3000bp) or small (<500bp) fragment fractions, which were individually concentrated (to 10 ul) using Microcon YM-100 columns (Millipore, Bedford, MA). Sequences in each fraction were ligated into lambda TripIEx2 vector (3 ul cDNA, 0.5 ul lambda TripIEx2 vector (500ng/ul), 0.5 ul 10X ligation buffer, 0.5 ul ATP (10 mM) and 0.5 ul T4 DNA ligase) (BD Biosciences, Franklin Lakes, NJ) and the resultant ligation mixture (4 ul) packaged into lamda vectors using Gigapack III gold packaging extract (Stratagene, La Jolla, CA). The packed phage libraries were plated using *E. coli* XL1-Blue competent cells, with transformation efficiency determined by percentage recombinant clones. Resultant plaques formed (~200 per library) were

eluted and plasmids purified using Qiagen plasmid purification kit (Qiagen Madison, WI) and integrity (size and quality) of respective cDNA insert validated via PCR using vector specific primers flanking the insert followed by fractionation of the fragments on 1% agarose gel stained with ethidium bromide (0.5μg/ml).

2.2.3 Sequencing of *G. pallidipes, G. p. gambiensis* and *G. tachinoides* cDNA libraries

Expressed sequence tags (ESTs) were generated by Sanger (capillary) sequencing from a total of 2863 cDNA library clones using an ABI3730 Sequencer and are available for download from ftp://ftp.sanger.ac.uk/pub/pathogens/Glossina/. Briefly, during capillary electrophoresis, the extension products of the cycle sequencing reaction enter the capillary as a result of electrokinetic injection. A high voltage charge applied to the buffered sequencing reaction forces the negatively charged fragments into the capillaries. The extension products are separated by size based on their total charge. The electrophoretic mobility of the sample can be affected by the run conditions: the buffer type, concentration, and pH; the run temperature; the amount of voltage applied; and the type of polymer used. Shortly before reaching the positive electrode, the fluorescently labeled DNA fragments, separated by size, move across the path of a laser beam. The laser beam causes the dyes on the fragments to fluoresce. An optical detection device on Applied Biosystems genetic analyzers detects the fluorescence. The Data Collection Software converts the fluorescence signal to digital data, then records the data in a *.ab1 file. Because each dye emits light at a different wavelength when excited by

the laser, all four colors, and therefore all four bases, can be detected and distinguished in one capillary injection (Applied Biosystems Chemistry Guide).

2.2.4 Analyses of *G. pallidipes, G. p. gambiensis* and *G. tachinoides* EST sequence data

The ESTs (generated from the cDNA library sequencing) were processed and analysed using cDNA Annotation Software™ (CAS) (Guo *et al.,* 2009). The software integrates different modules into a pipeline system with the final output piped into a tab-delimited file imported into an Excel spreadsheet (Valenzuela *et al.,* 2003). Briefly, primer and vector sequences were striped and the ESTs clustered using CAP3 assembler (Huang and Madan, 1999) of the software generating singletons and consensus sequences. Singletons are sequences not included in any clusters and consensus sequences contain two or more sequences within a cluster.

Both singletons and consensus sequences, referred hereafter as clusters, were functionally annotated using blastx (Altschul *et al.,* 1997) against nonredundant (NR) protein database of GenBank (Wheeler *et al.,* 2010) and gene ontology (GO) database (Ashburner *et al.,* 2000). Rpsblast identified conserved protein domains (Schaffer *et al.,* 2001) in Pfam (Bateman *et al.,* 2000), SMART (Letunic *et al.,* 2002) and KOG (Tatusov *et al.,* 2003) databases. Finally, signal peptides in the clusters were predicted using the SignalP program (Nielsen *et al.,* 1997). Subsequent clusters with putative olfactory protein domains (identified from the analyses with CAS software) were characterised for their molecular weights (MW)

and isoelectric points (pIs) using ProtParam software (Gasteiger *et al.,* 2005) and aligned against putative OBPs sequences of *G. morsitans morsitans* downloaded from GeneDB (Hertz-Fowler *et al.,* 2004) and other insects OBPs (*Drosophila, Musca domestica, Aedes aegypti, Anopheles gambiae, Heliothis virescens, Bombyx mori, Antheraea pernyi, Phyllopertha diversa* and *Apis mellifera*) in GenBank (Wheeler *et al.,* 2010), using ClustalW2 software (Larkin *et al.,* 2007).

Genetic distance between the putative OBPs and those in the databases were calculated using PHYML (Guindon and Gascuel, 2003) software, based on neighbour joining algorithm, maximum likelihood model and WAG substitution matrix (Whelan and Goldman, 2001). The generated consensus tree, based on 100 bootstrap replicates, was viewed using TreeDyn software (Chevenet *et al.,* 2006).

2.3 Results

2.3.1 Poly(A) RNA and cDNA Quality

Antennae and head were used as tissues to construct cDNA libraries. Poly(A) RNA was reverse transcribed to make first strand which formed template for second strand cDNA synthesis. The poly(A) RNA was found to be intact and assessment of its quality from agarose gel revealed that it looked like a diffuse smear from about 250 bp to about 1kb for *G. pallidipes* (Figure 2.1a), and ranges from 250 bp to about 2 kb for *G. tachinoides* and *G. palpalis gambiensis* (Figure 2.2a and 2.3a). cDNA synthesis was done by long-distance PCR (LD PCR) as it allows construction of cDNA library using nanogram amounts of poly(A) RNA with a

high percentage of full-length clones (Chenchik *et al.,* 1998). A moderately strong smear of cDNA from 250 bp to 2 kb was observed for *G. pallidipes* (Figure 2.1b) and from 250 bp to 2 kb for *G. tachinoides* and *G. palpalis gambiensis* (Figure 2.2b and 2.3b).

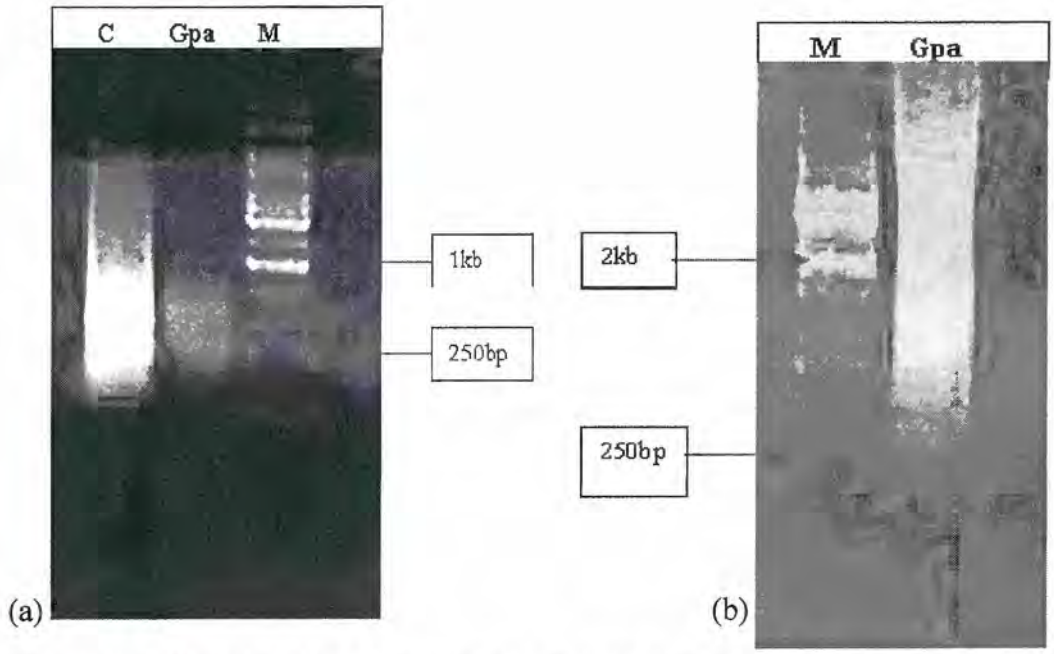

Figure 2.1 Poly(A) RNA (a) and cDNA (b) from *Glossina pallidipes* antenna (Gpa). 5 ul of poly(A) RNA and cDNA was resolved on 1% agarose gel stained with ethidium bromide. M-1 kb ladder, C - Control RNA.

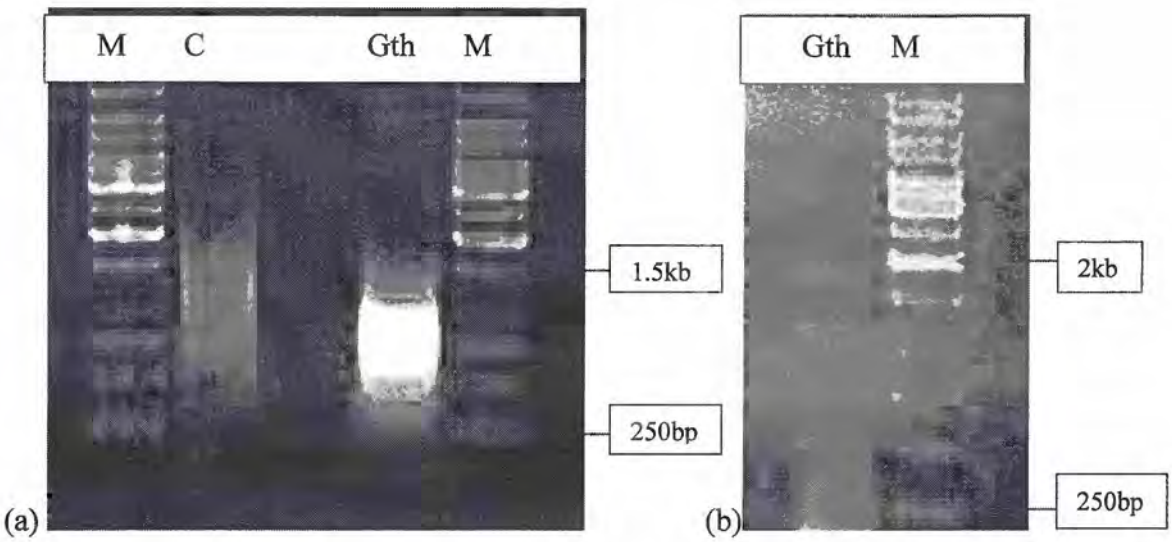

Figure 2.2 Poly(A) RNA (a) and cDNA (b) from *Glossina tachinoids* head (Gth). 5 ul of poly(A) RNA and cDNA was resolved on a 1% agarose gel stained with ethidium bromide. M-1 kb ladder, C - Control RNA.

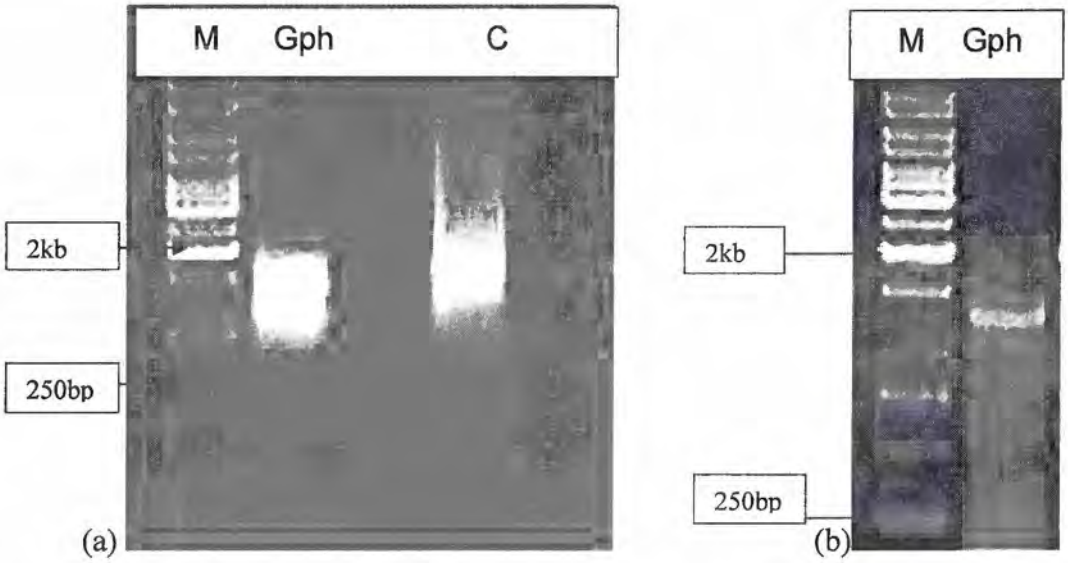

Figure 2.3 Poly(A) RNA (a) and cDNA (b) from *Glossina palpalis gambiensis* head (Gph). 5 ul of poly(A) RNA and cDNA was resolved on a 1% agarose gel stained with ethidium bromide. M-1 kb ladder, C - Control RNA.

2.3.2 Summary of Expressed Sequence Tags (ESTs)

Expressed sequenced tags (ESTs) obtained from *G. pallidipes* antennal, *G. palpalis gambiensis* and *G. tachinoides* head libraries were 1127, 906 and 830 respectively. These generated 18 consensus sequences and 278 singletons in *G. pallidipes* antennal library; 36 consensus sequences and 269 singletons in *G. p. gambiensis* head library. Similarly, 54 and 178 consensus sequences and singletons respectively were generated from *G. tachinoides* (Table 2.0). Clusters from *G. pallidipes* antennae, *G. p. gambiensis* head and *G. tachinoides* head were compared by the program blastx, rpsblast (Altschul *et al.,* 1997) to NR protein database of GenBank (Wheeler *et al.,* 2010), GO database (Ashburner *et al.,* 2000) and conserved domains database of GenBank (Marchler-Bauer *et al.,* 2002). Nine categories of transcripts were derived from manual annotation of the clusters. The clusters for *G. pallidipes*, *G. p. gambiensis* and *G. tachinoides* are as shown in Table 2.1. In *G. pallidipes* antennae, odorant/pheromone binding proteins were represented in 0.7% of the clusters and 0.9% of the sequences; energy metabolism had 3.7% of the clusters and 3.4% of the sequences, transcription factors with 4.7% of the clusters and 4.3% of the sequences and cytoskeletal having 7.8% and 7.4% of the clusters and sequences respectively.

Table 2.0 Number of ESTs sequenced and clusters generated from *Glossina pallidipes* antennal, *Glossina tachinoides* head and *Glossina palpalis gambiensis* head libraries

	ESTs sequenced	**Consensus**	**Singletons**	**Clusters**
G. pallidipes	1127	18	278	296
G. p. gambiensis	906	36	269	305
G. tachinoides	830	54	178	232
Total	**2863**	**108**	**725**	**833**

Table 2.1 Summary of clusters found in *Glossina pallidipes* antennae, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head libraries

G. pallidipes					***G. p. gambiensis***				***G. tachinoides***			
Category	**Clusters**	**Seqs**	**% Clusters**	**% Seqs**	**Clusters**	**Seqs**	**% Clusters**	**% Seqs**	**Clusters**	**Seqs**	**% Clusters**	**% Seqs**
Odorant/Pheromone Binding	2	3	0.7	0.9	5	6	1.6	0.9	3	4	1.4	0.7
Energy Metabolism	11	11	3.7	3.4	23	33	7.5	5.2	24	46	10.8	8.4
Transcription Factors	14	14	4.7	4.3	17	22	5.6	3.5	29	33	13.1	6.0
Cytoskeletal	23	24	7.8	7.4	11	14	3.6	2.2	6	15	2.7	2.7
Transporters	21	22	7.1	6.8	23	27	7.5	4.3	12	16	5.4	2.9
Signal transduction	27	36	9.1	11.1	15	26	4.9	4.1	7	34	3.2	6.2
Protein Function	56	60	18.9	18.6	90	100	29.5	15.8	56	261	25.2	47.5
Unknown	49	54	16.6	16.7	29	162	9.5	25.6	30	53	13.5	9.7
Hypothetical	93	99	31.4	30.7	92	242	30.2	38.3	65	82	29.3	14.9
Total	**296**	**323**			**305**	**632**			**222**	**549**		

2.3.2.1 *Glossina pallidipes* clusters with matches to nonredundant (NR), Conserved Domains (CDD) and Gene ontology (GO) databases

Two clusters (265 and 266) from *G. pallidipes* antennae are related to *D. erecta* Olfactory Specific-F and *An. gambiae* AgamOBP1 respectively (Table 2.2). Cluster 265 did not have a signal peptide and a total of eleven clusters may be involved in energy transduction processes with fourteen clusters implicated in transcriptional factors. Cytoskeletal, transporters and signal transduction all had an average of 20 clusters. Fifty six clusters were matching proteins of known function that may be involved in basic metabolism and majority of the clusters were either unknown (49 clusters) or hypothetical (93 clusters) (Appendix I).

Of the 296 clusters, only thirty one (31) are possibly secretory proteins as they had a signal peptide signature. The remaining clusters have no conclusive indication of either a leader signal peptide or complete ORF, probably due to the diminished sequence quality. Many of the clusters have high identity to insects (*Drosophila, Anopheles gambiae, Vasdavidius concursus, Chrysomya putoria, Hypoderma sinense, Lucilia cuprina* and *Aedes aegypti*), confirmation that the annotation was good. Other clusters had best matches to proteins known from mammals, *Arabidopsis thaliana, Caenorhabditis briggsae* or other organisms. Clusters with no significant matches to either NR protein database of GenBank (Wheeler *et al.*, 2010) or GO database (Ashburner *et al.*, 2000) accounted for 17.9%.

2.3.2.2 *Glossina palpalis gambiensis* clusters with matches to nonredundant (NR), Conserved Domains (CDD) and Gene ontology (GO) databases

Four clusters from *G. p. gambiensis* (184, 204, 206 and 255) had significant matches to olfactory proteins from *Drosophila* (Table 2.2). Cluster 195 had homolog to transcription factor IIIB from *Dictyostelium discoideum* (Chung *et al.*, 2007). Its involvement in odorant/pheromone binding is based on GO annotation. Altogether, twenty three (23) clusters were linked to energy metabolism while transcriptional factors, cytoskeletal and transporters had 17, 11 and 23 clusters respectively. Fifteen clusters were associated with signal transduction molecules and ninety clusters annotated as proteins of various functions. The unknown and hypothetical clusters had 29 and 92 clusters respectively (Appendix II).

Notably, 4 clusters were homologs to *G. m. morsitans* proteins with cluster 174 being related to *G. m. morsitans* imaginal disc growth factor 4 and encodes a gene related to chitinase. Cluster 207 is similar to MnFe superoxide, cluster 221 represents *G. m. morsitans* homolog (Gmfb8) with unknown function and cluster 259 is homolog to *G. m. morsitans* ferritin. Among the signal transduction clusters, two clusters are related to vision, one cluster to *Homo sapiens* G-Protein Coupled Receptor (GPCR 52) and one arrestin. The percentage of clusters having signal peptide and those without significant matches to NR protein database was 13.4% and 4.9% respectively. Many high identities of species were of insects, few *Ixodes scapularis* and other organisms which included mammals, plants, parasites and nematodes.

2.3.2.3 *Glossina tachinoides* clusters with matches to nonredundant (NR), Conserved Domains (CDD) and Gene ontology (GO) databases

Three clusters are probably involved in odorant binding with clusters 63 and 151 having significant matches to *D. melanogaster* PBP/OBP and cluster 216 similar to *Musca domestica* odorant binding protein 3 (Table 2.2). None of the *G. tachinoides* olfactory protein clusters had signal peptide. The number of clusters involved in energy metabolism, transcription factors and cytoskeletal were 24, 29 and 6 respectively. Transporters, signal trasduction and protein function had 12, 7 and 56 clusters respectively. Thirty clusters were classified as unknown and sixty five clusters as hypothetical (Appendix III). Cluster 125 was the only identified paralog from *G. m. morsitans*. As reported for *G. p. gambiensis,* signal transduction had clusters related to opsin, arrestin and olfactory receptor homologs from *Bos taurus*. The percentage of clusters with signal peptide stands at 29.3% while twelve clusters, all under unknown category, had no positive hits from NR protein database. Distribution of species varied among Insects with *Drosophila*, being the most dominant. Other hits included plants, ticks, nematodes and parasites.

Table 2.2 Functional annotation of odorant/pheromone cDNA clusters from *Glossina pallidipes* antennae library, *Glossina palpalis gambiensis* and *Glossina tachinoides* head libraries producing best hits to nonredundant (NR) protein database of GenBank, gene ontology (GO) and conserved domains (CDD) database

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
Gpacontig265	1	Olfactory Specific F	emb|CAE00444.1|	7.0E-18	*Drosophila erecta*	Pfam PBP/GOBP family	0.029	pheromone binding	3.0E-18	0005550	No
Gpacontig266	2	AgamOBP1	gb|AAO12081.1|	3.0E-30	*Anopheles gambiae*	Pfam PBP/GOBP family	4.0E-19	pheromone binding	1.0E-30	0005550	Yes
Gphcontig184	2	Olfactory Specific F	ref|NP_524241.1|	5E-36	*Drosophila melanogaster*	Smart Insect PBP/OBP domains	8E-18	pheromone binding	3.0E-37	0005550	Yes
Gphcontig195	1	transcription factor IIIB	ref|XP_640891.1|	5E-05	*Dictyostelium discoideum*	Smart Putative single-stranded NAs	0.077	pheromone binding	3.0E-04	0005550	No
Gphcontig204	1	CG11852-PA	gb|AAM50923.1|	2E-19	*Drosophila melanogaster*	Pfam Odorant binding protein	2E-29				No
Gphcontig206	1	hypothetical protein	emb|CAJ01441.1|	2E-09	*Drosophila pseudoobscura*	Pfam Insect PBP-binding family	4E-09	pheromone binding	1.0E-09	0005549	No
Gphcontig255	1	GA16322-PA	gb|EAL28074.1|	1E-45	*Drosophila pseudoobscura*	Pfam PBP/GOBP family	2E-15	odorant binding	2.0E-46	0005549	No
Gthcontig63	2	RH74005p	gb|AAM29645.1|	1E-44	*Drosophila melanogaster*	Pfam Insect pheromone-binding family	1E-42	odorant binding	6E-46	0005549	No
Gthcontig151	1	odorant binding protein 83g	gb| AAN13232.1 |	9E-19	*Drosophila melanogaster*	Smart Insect PBP/OBP domains	3E-10	odorant binding	4E-20	0005549	No
Gthcontig219	1	odorant binding protein 3	gb|AAV74624.1|	5E-07	*Musca domestica*	Smart Insect PBP/OBP domains	5E-08	pheromone binding	2E-08	0005550	No

2.3.3 Molecular weights and Isoelectric points

The numbers of amino acids for putative *Glossina* OBPs are in the range of 103 to 210. Predicted molecular weights (MWs) also vary considerably between 11.9 kDa to 25.1 kDa while pIs reported are mainly alkaline implying that OBPs in *Glossina* species are positively charged at physiological pH in the antennae (Table 2.3).

Table 2.3 Predicted Molecular weight and Isoelectric points for putative *Glossina* OBPs

Putative *Glossina* OBP	**Number of amino acids**	**Molecular Weight (kDa)**	**Isoelectric point (pI)**
Gpacontig265	136	15.4	9.77
Gpacontig266	188	21.9	8.49
Gphcontig184	192	22.8	9.49
Gphcontig195	195	22.4	8.51
Gphcontig204	178	20.5	9.1
Gphcontig206	111	13.1	10.31
Gphcontig255	210	25.1	9.14
Gthcontig63	198	22.8	9.46
Gthcontig151	126	15	9.27
Gthcontig219	103	11.9	9.43

kDa – Kilo Daltons; GpacontigXXX - *Glossina pallidipes* antennae contigs; Gphcontig XXX - *Glossina palpalis* head contigs; GthcontigXXX - *Glossina tachinoides* head contigs

2.3.4 Multiple Sequence Alignment

An alignment of six putative *Glossina* OBPs (Gpacontig265; Gpacontig266; Gphcontig184; Gthcontig63, Gthcontig151 and Gthcontig219) with *G. m. morsitans* OBPs downloaded from geneDB (Hertz-Fowler *et al.,* 2004) showed that they had characteristic features of Insect OBPs (Figure 2.4). Two putative *Glossina* OBPs (Gpacontig266 and Gphcontig184) had six conserved cysteine amino acids while Gpacontig265, Gthcontig151 and Gthcontig219 all had one cysteine residue. Gthcontig63 had the characteristic four cysteine residues for chemosensory proteins (CSPs). Five *G. m. morsitans* putative OBPs (cn15569, cn15565, cn15220, cn7403 and GLAER58TV) had 6 conserved cysteines; three *G. m. morsitans* putative OBPs (cn13968, cn14014 and cn15331) had 5 cysteines. The putative OBPs (cn13435 and GLAAS20TVB) had 4 cysteines with cn14707 and cn15567 each having 3 cysteines while GLAC953TV contained 1 cysteine. The putative *Glossina* OBPs are low in sequence similarity (Figure 2.4), demonstrating the highly divergent nature of this protein family even in insects of the same genus.

The five putative *Glossina* OBPs and CSP were also aligned with selected insects OBPs (Figure 2.5) and CSPs (Figure 2.6) extracted from GenBank (Wheeler *et al.,* 2010). The OBPs included olfactory Specific-F (*Drosophila erecta)*, olfactory Specific-E (*Drosophila simulans*), OBP3 (*Musca domestica*), OBP (*Drosophila willistoni*), OBP (*Aedes aegypti*), OBP (*Anopheles gambiae)* GOBP2 (*Heliothis virescens)*, GOBP1 (*Bombyx mori),* GOBP2 (*Bombyx mori)*, GOBP2 (*Antheraea*

pernyi), OBP2 (*Phyllopertha diversa)*, PBP (*Bombyx mori)*, PBP (*Heliothis virescens)*, OBPASP5 (*Apis mellifera)* and ABP8 (*Manduca sexta)*. The CSPs used in the alignment were from *Heliothis virescens*, *Bombyx mori*, *Apis mellifera*, *Microplitis mediator*, *Glossina morsitans morsitans*, *Tribolium castaneum*, *Aphis gossypii* putative, *Sclerodermus guani*, *Spodoptera exigua*, *Plutella xylostella*, *Drosophila melanogaster* and *Anopheles gambiae*.

The alignment revealed that OBP sequences are diverse unlike the CSPs which are more highly conserved at certain amino acid residues. They have four conserved cysteines, three highly conserved aromatic residues, a lysine residue between cysteines 3 and 4 and a glutamine residue at the fourth amino acid from cysteine 4. The *Glossina* CSP ortholog (Gthcontig63) have the first, second and third aromatic amino acids as phenylalanine, tryptophan and tyrosine respectively (Figure 2.6).

It was noted that Antennal-binding protein (ABP) from *Manduca sexta* also had four conserved cysteines though at different positions from CSPs. The four cysteine amino acids are in similar position with the other cysteines found in OBPs, suggesting that ABP belongs to the OBP family.

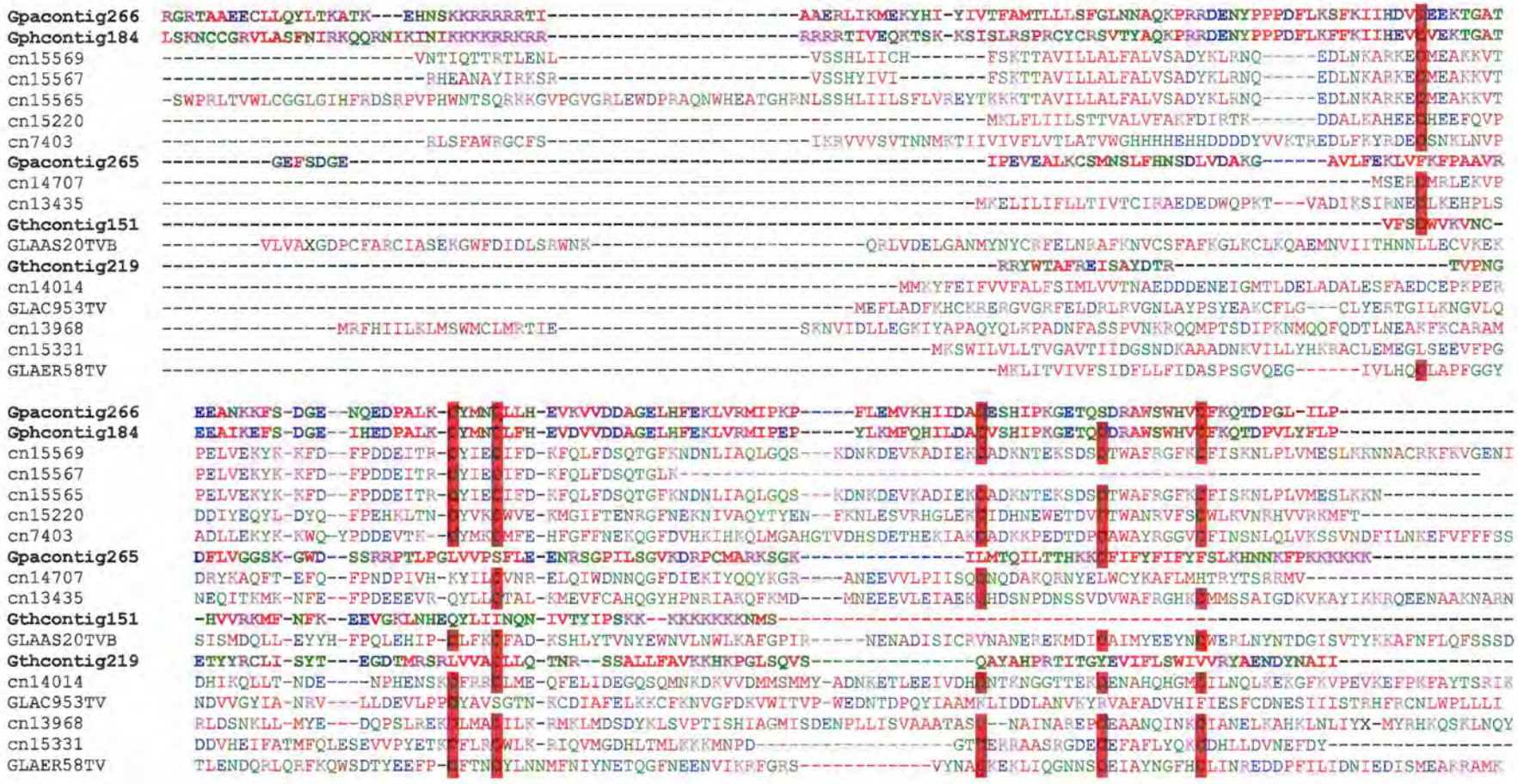

Figure 2.4. Alignment of putative *Glossina pallidipes* antennal, *Glossina palpalis gambiensis* and *Glossina tachinoides* head OBPs with *Glossina morsitans morsitans* OBPs. Multiple Sequence alignment was done using ClustalW2 software. The conserved cysteine residues are shaded red. GpacontigXXX - *Glossina pallidipes* antennae contigs; GphcontigXXX – *Glossina palpalis* head contigs; GthcontigXXX - *Glossina tachinoides* head contigs. *Glossina morsitans morsitans* contigs (cn15565, cn15567, cn15569, cn15220, cn7403, cn14707, cn13435, cn13435, GLAAS20TVB, cn14014, GLAC953TV, cn13968, cn15331 and GLAER58TV).

```
OSFDerecta        ------------------------------------MALNGFGRRVSASVLLIALSLLSGALILPPAAAQRDENYPPPGILKMAKPFH-DACVEKTGVSEAAIKEFSD-GEIHEDEKL
OBP3Mdomestica    ------------------------------------MALKGLGRGVSASVLLIVLSLLTGAPILPRATAQRDDNYPQPVKLKMAKPLH-DACVENTGVIEAAIKEFSD-GEIHEDENL
OSEDsimulans      ----------------------------------------------MVKYPLILFLIG-CAAAQ--EPRRDGEWPPPAILKLAKHFH-DICAPKTGVTDEAIKEFSD-GQIHEDEAL
GK14216PADwilli   ----------------------------------------------MIRYCFLLCLFGGYAIGQ--APRRDAEYPPPGLLKMGKHFH-DICVEKTGATDEAIKEFSD-GEIHEDEKL
OBPAaegypti       ------------------------------------MNGSVVF-----VLSALVS-----LSVGD-VTPRRDAEYPPPEFLEAMKPLR-EICIKKTGVTEEAIIEFSD-GKVHEDENL
OBPAnopheles      ------------------------------------MGHDSCWSSRWRVLAALVIFQCAILMVRS-DEPRRDANYPPPELLEKMKPMH-DACVAETGASEDAIKRFSD-QEIHEDDKL
Gpacontig266      RGRTAAEECLLQYLTKATK---EHNSKKRRRRRTIAAERLIKMEKYHIYIVTFAMTLLLSFGLNNAQKPRRDENYPPPDFLKSFKIIH-DVCEEKTGATEEANKKFSD-GENQEDPAL
Gphcontig184      LSKNCCGRVLASFNIRKQQRNIKINIKKKKRRKRRRRRRTIVEQKTSKKSISLRSPRCYCRSVTYAQKPRRDENYPPPDFLKFFKIIH-EVCVEKTGATEEAIKEFSD-GEIHEDPAL
GOBP2Heliothis    ----------------------------------------------MTSKSCLLLVAMVTLTTS-VMGTAEVMSHVTAHFGKALEECREESGLS-AEVLEEFQHFWRE-DFEVVHREL
GOBP2Antheraea     ----------------------------------------------MGYK--LLLMYIAIVIDS-VIGTAEVMSHVTAHFGKALEECRDESGLS-PEILNEFKHFWSE-DFDVVHREL
GOBP2Bombyx        ----------------------------------------------MFSF--LILVFVASVADS-VIGTAEVMSHVTAHFGKTLEECREESGLS-VDILDEFKHFWSD-DFDVVHREL
GOBP1Bombyx        ----------------------------------------------MWK---LVVVLTVNLLQG-ALTDVYVMKDVTLGFGQALEQCREESQLT-EEKMEEFFHFWND-DFKFEHFEL
PBPBombyx          ----------------------------------------------MSIQGQIALALMVNMAVGSVDASQEVMKNLSLNFGKALDECKREMTLT-DAINEDFYNFWKE-GYEIKNRET
PBPH.virescens     ----------------------------------------------MMSVR--LMLVVAVWLCLR-VDASQDVMKNLSMNFAKPLEDCKKEMDLP-DSVTTDFYNFWKE-GYEFTNRHT
ABP8Manducasexta-------------------------------------------------------------MKALLVLAACLVLAQALTDEQKEKLKKHKSECLSETKVEEQLVNKLKAGDYKAENDNL
OBP_ASP5Apis      -------------------------------------------------------MHVKSVLLLITIVTFVALKPVKSMSADQVEKLAKNMRKSCLQKIAITEELVDGMRR-GEFPDDHDL
OBP2Phyllopertha-------------------------------------------------------------------------KEHGQKVLEQIIDYATSKADSLGVSPEDMKLLMEKKFPT------SREG
Gpacontig265      ------------------------------------------------------------GEFSDGEIPEVEALKCSMNSLFHNSDLVDAKGAVLFEKLVFKFPAAVRD-FLVGGSKGW
Gthcontig151      ------------------------------------------------------------VFSCWVKVNCHVVRKMFN-FKEEVGKLNHEQYLIINQNIVTYIPSKKK-KKKKKKNMS
Gthcontig219      ------------------------------------------------------------------------------------------------RRYWTAFREISAYDT----RTVPNG

OSFDerecta        KCYMNCFFHEIEVVDDKGDVHLEKLFATVPL-----SLRDKLVEMSKGCVH--PEGDTL-CHKAWWFHQCWKKADPKHYFLP--------------------------
OBP3Mdomestica    KCYMNCFFHEIEVVDDHGDVHLEKLFATVPL-----PMRDNLMEMSKGCVH--PEGDTL-CHKVWWFHQCWKKADPKHYFLP--------------------------
OSEDsimulans      KCYMNCLFHEIEVVDDNGDVHMEKLFNAIPGE----KLRNILMEASKGCMH--PEGDTL-CHKAWWFHQCWKKADPVHYFLV--------------------------
GK14216PADwillistoni KCYMNCLFHEFDVVDDNGDVHLEKLFNAVPG-----KLRDILMKMSENCIH--PEGDTL-CHKAWWFHQCWKKADPVHYFLLMINDTPKLLMGVAREGGLMHQHSGNT
OBPAaegypti       KCYMNCLFHEAKVVDDTGHVHLEKLHDALPD-----SMHDIALHMGKRCLY--PEGENL-CEKAFWLHKCWKESDPKHYFLI--------------------------
OBPAnopheles      KCYMNCLFHQAGVVNDKGEFHYVKIQDFLPE-----SMHLITLNWFKRCLY--PEGENG-CEKAFWLNKCWKTRDPVHYFLP--------------------------
Gpacontig266      KCYMNCLLHEVKVVDDAGELHFEKLVRMIPK-----PFLEMVKHIIDACESHIPKGETQ-SDRAWSWHVCFKQTDPGL-ILP--------------------------
Gphcontig184      KCYMNCLFHEVDVVDDAGELHFEKLVRMIPE-----PYLKMFQHILDACVSHIPKGETQ-CDRAWSWHVCFKQTDPVLYFLP--------------------------
GOBP2Heliothis    GCAIICMSNKFSLLQDDSRMHHVNMHDYVKSFPNGHVLSEKLVELIHNCEKKYD-TMTDDCDRVVKVAACFKVDAKAAGIAP--EVTMIEAVMEKY------------
GOBP2Antheraea    GCAIICMSNKFSLLKDDTRIHHVNMHDYVKSFPNGEVLSAKMVNLIHNCEKQYD-DITDECDRVVKVAACFKVDAKKEGIAP--EVAMIEAVIEKY------------
GOBP2Bombyx       GCAIICMSNKFSLMDDDVRMHHVNMDEYIKGFPNGQVLAEKMVKLIHNCEKQFD-TETDDCTRVVKVAACFKKDSRKEGIAP--EVAMIEAVIEKY------------
GOBP1Bombyx       GCAIQCMSRHFNLLTDSSRMHHENTDKFIKSFPNGEILSQKMIDMIHTCEKTFD-SEPDHCWRILRVAECFKDACNKSGLAPSMELILAEFIMESEADK---------
PBPBombyx         GCAIMCLSTKLNMLDPEGNLHHGNAMEFAKKHGADETMAQQLIDIVHGCEKSTP-ANDDKCIWTLGVATCFKAEIHKLNWAPSMDVAVGEILAEV-------------
PBPHeliothisvirescens GCAILCLSSKLELLDQEMKLHHGKAQEFAKKHGADDAMAKQLVDMIHGCSQSTPDATDDPCMKALNVAKCFKAKIHELNWAPSMELVVGEVLAEV-------------
ABP8Manducasexta  KKYALCMMMKSELMTKEGKFKKDVALSKVPN----PADKPMVEKLIDTCLAN---KGNTPHQTAWNYVKCYHEKDPKHAIFL--------------------------
OBP_ASP5Apis      QCYTTCIMKLLRTFKNGN-FDFDMIVKQLEIT---MPPEEVVIGKEIACR-N-EEYTGDDCQKTYQYVQCHYKQNPEKFFFP--------------------------
OBP2Phyllopertha  QCMPSCVNKKFGLQKADGTLNKEYRYSEMENVKAIDEEIYNKMNSVWDCV-N-GADGTDECDTGMKVVTCMKEESEKLGLSKDAIGF---------------------
Gpacontig265      DSSRRPTLPGLVVPSFLEENRSGPILSGVKDR----PCMARKSGKILMTQILTTHKKCFIFYFIFYFSLKHNNKFPKKKKKK--------------------------
Gthcontig151      ------------------------------------------------------------------------------------------------------------
Gthcontig219      ETYYRCLISYTEGDTMRSRLVVACLLQTNRSSALLFAVKKHKPGLSQVSQAYAHPRTITGYEVIFLSWIVVRYAENDYNAII--------------------------
```

Figure 2.5. ClustalW2 alignment of putative *Glossina* OBPs with other insect OBPs downloaded from GenBank. The conserved cysteine residues are shaded red and conserved amino acids shaded grey. GpacontigXXX - *Glossina pallidipes* antennae contigs; GphcontigXXX - *Glossina palpalis* head contigs; GthcontigXXX - *Glossina tachinoides* head contigs.

```
Gthcontig63 -----YGRGLKNFLTIAVAVVLMTIAVI-AEEQYTTKFDNIDVDEILASDRLFDNYFKCL VDEG--KCT-PEGRELKKTLPDALETACTKCNDKQKATVDKVIRFLTEKKPEQWKALQAK 111
GmmorsitansCSP1------MKYLTIVAVIATLSAVVVMGAEEKYTTKYDDVDVDEVLKSDRLFKNYYNCL IDQG--KCT-PDARELKKSLPDALQTECSKCSEKQKKTSEKVIKHLMDHKPEEWKVLQTK 108
AgambiaeCSP1---------MKHLTMVAIFA---MVVVLASAQKYTDKFDNIDVDRVLSNDRILNNYLKCL LDKG--PCT-QEGRELK-TLPDALKTNCEKCSEKQRTSSRKVIAHLEERKPQEWKKLLDK 104
MmediatorCSP1-----------MKVAIIFLAIIAVALAATTKTYTSKFDDVDVDGILGSDRLLRNYVNCL LDRG--PCT-KEGVTLKEILPDALATSCESCTEKQKTKSEKVIRHLVNNKKELWDELAVK 105
AgossypiiCSP1 IAVVCVVLAAFAVDQTVGAPQKDAVAASGPAYTTKYDHIDVDQVLASKRLVNSYVQCL LDKK--PCT-PEGAELRKILPDALKTQCAKCNATQKNAALKVVDRLQKDYDAEWKQLLDK 117
HvirescensCSP1----------MALARPDGAAYTDKYDNVDLDEILSNRRLLVPYVKCI LDQG--KCA-PDAKELKEHIIEALENECGKCTEAQKKGTRRVIGHLINNEADYWNELTAK 94
SexiguaCSP1 ---------MKSFIVLCLFGLAAVAMARPDGSTYTDRYDNINLDEILGNRRLLTPYIKCI LEEG--KCT-PDGKELKSHIREALEQNCAKCTDAQRNGTRRVLGHLINNEEESWNRLKAK 108
BmoriCSP1 ---------MKCLTIAALLFVAGLSIAEK----YTDKYDNIDVDEILENRKLLVPYIKCV LDEG--RCT-PDGKELKAHIKDGMQTACAKCTDRQKVSARKIVKHIKQHEADYWEQMKAK 104
PxylostellaCSP1------MKSAAFIALFLIGKAVCEDK--PTYTTKYDNIDLDEILSSERLLTGYVNCL LDQG--PCT-PDGKELKHTLPDAIDNDCRKCTQKQKEGSDRVMGYIIEYRPNDWAKLEKK 106
AmelliferaCSP1---------MRHNYIVILILSLLTWTYAEELYSDKYDYVNIDEILANDRLRNQYYDCF IDAG--SCLTPDSVFFKSHITEAFQTQCKKCTEIQKQNLDKLAEWFTTNEPEKWNHFVEI 107
SguaniCSP1 ----------MKSIIVVFTVFAIVFAQESTPEYYTGKWNDLNTHDIVDNARLFKKYKQCI MAETNTGCP-QEVIELKKVLPEALETVCSKCSPVQVEKIRDTLKYVCEKRKTDFDDILKH 109
TcastaneumCSP1------MLILQIAHLCAQFCLLAAIFTCVKPQLTRISDEAIESTLNDRRYLLRQLKCA TGEA--PCD-PVGRRLKSLAPLVLRGSCPQCTPQEMKQIQKVLAFVQKNYPKEWNKILHQ 109
DmelanogasterCSP1----MLLLNKNRVISLVVNFIFLIILISSSVQADERNINKLLNNQVVVSRQIMCI LGKS--ECD-QLGLQLKAALPEVITRRCRNCSPQQAQKAQKLTTFLQTRYPDVWAMLLRK 108
                                                                                                                                  
Gthcontig63             YDPSGEYLKKYR---AEAEKRGIKVINLMLRVECYSPMCNMCK--- 151
GmmorsitansCSP1         YDPEGIYYSKYK---ARDAKA------------------------- 126
AgambiaeCSP1            YDPEGIYKSKFE---KINKRS------------------------- 122
MmediatorCSP1           YDPNNEYRKKYE---DQAKAKGINV--------------------- 127
AgossypiiCSP1           WDPKREHFQKFQQFLAEEKKKGFTKF-------------------- 143
HvirescensCSP1          FDPEKKYVQKYEKELKEVKA-------------------------- 114
SexiguaCSP1             YDPQSKYTVKYELELRKLKQ-------------------------- 128
BmoriCSP1               YDPKDEFKEIYEGFLAGQN--------------------------- 123
PxylostellaCSP1         YLSDGSYKKKYLEKKNASENNGDSKSTEAKNKDDEEKKSKGDGEEK 152
AmelliferaCSP1          MIKKKDEGA------------------------------------- 116
SguaniCSP1              IDPEGTHRPKFEEKFGTLGC-------------------------- 129
TcastaneumCSP1          YAG------------------------------------------- 112
DmelanogasterCSP1       YDSA------------------------------------------ 112
```

Figure 2.6. Multiple Sequence Alignment of *Glossina tachinoides* CSP (Gthcontig63) with other insect CSPs downloaded from GenBank. The alignment was done using ClustalW2 software. The conserved cysteine residues are shaded green (*) and conserved amino acids are shaded blue and marked by (: and .). Aromatic amino acid residues are shaded yellow. Accession numbers of other Insects CSPs are: *Heliothis virescens* CSP1 gi|21898556|; *Bombyx mori* CSP1 gi|112983050|; *Apis mellifera* CSP1 gi|118150502|; *Microplitis mediator* CSP1 gi|126508768|; *Glossina morsitans morsitans* CSP1 gi|281426841|; *Tribolium castaneum* CSP1 gi|113951685|; *Aphis gossypii* putative CSP1 gi|215254064|; *Sclerodermus guani* putative CSP1 gi|91983607|; *Spodoptera exigua* CSP1 gi|122894084|; *Plutella xylostella* CSP1 gi|122894080|; *Drosophila melanogaster* putative CSP1 gi|48994224| and *Anopheles gambiae* putative CSP1 gi|48994214|.

2.3.5 Phylogenetic analysis of Putative *Glossina* OBPs

There is high bootstrap support for *B. mori* and *H. virescens* PBPs (100%), *M. domestica* OBP3 and *D. erecta* OS-F (96%), *H. virescens* and *A. pernyi, B. mori* GOBP2 (96%), Gpacontig266 and Gphcontig184 (83%) (Figure 2.8). The weak bootstrap values for *M. sexta* ABP8, Gpacontig265 and *A. mellifera* OBPASP5 could imply they are not related to the classical olfactory proteins (OBPs, PBPs and GOBPs) while PBPs and GOBPs could be more similar than to GOBP1. The clustering is supported by spacing pattern of the first conserved cysteines where it is the same for PBP, GOBP1 and GOBP2; Gpacontig266, Gphcontig184, OBPs, OS-E and OS-F and different for ABP8, Gpacontig265, OBPASP5, Gthcontig151 and Gthcontig219 (Figure 2.6). The phylogenetic tree in Figure 2.8 could indicate the existence of different olfactory protein subfamilies and shows the relationship between putative *G. pallidipes* OBPs (Gpacontig265; Gpacontig266), *G. p. gambiensis* OBP (Gphcontig184) and *G. tachinoides* OBPs (Gthcontig151 and Gthcontig219) with other insect's OBPs downloaded from GenBank.

The sequence similarity tree in Figure 2.9 shows the relationships between *G. tachinoides* CSP (Gthcontig63) with CSP1 from *H. virescens*, *B. mori, A. mellifera, M. mediator*, *G. m. morsitans*, *T. castaneum*, *A. gossypii*, *S. guani*, *S. exigua*, *P. xylostella*, *D. melanogaster* and *A. gambiae*. It is noted that the *Glossina* CSPs are more similar than to other Insects CSPs. There is good bootstrap support for many

terminal relationships with CSP1 from Insects of different orders randomly distributed.

2.3.5 Phylogenetic analysis of Putative *Glossina* OBPs

There is high bootstrap support for *B. mori* and *H. virescens* PBPs (100%), *M. domestica* OBP3 and *D. erecta* OS-F (96%), *H. virescens* and *A. pernyi, B. mori* GOBP2 (96%), Gpacontig266 and Gphcontig184 (83%) (Figure 2.8). The weak bootstrap values for *M. sexta* ABP8, Gpacontig265 and *A. mellifera* OBPASP5 could imply they are not related to the classical olfactory proteins (OBPs, PBPs and GOBPs) while PBPs and GOBPs could be more similar than to GOBP1. The clustering is supported by spacing pattern of the first conserved cysteines where it is the same for PBP, GOBP1 and GOBP2; Gpacontig266, Gphcontig184, OBPs, OS-E and OS-F and different for ABP8, Gpacontig265, OBPASP5, Gthcontig151 and Gthcontig219 (Figure 2.6). The phylogenetic tree in Figure 2.8 could indicate the existence of different olfactory protein subfamilies and shows the relationship between putative *G. pallidipes* OBPs (Gpacontig265; Gpacontig266), *G. p. gambiensis* OBP (Gphcontig184) and *G. tachinoides* OBPs (Gthcontig151 and Gthcontig219) with other insect's OBPs downloaded from GenBank.

The sequence similarity tree in Figure 2.9 shows the relationships between *G. tachinoides* CSP (Gthcontig63) with CSP1 from *H. virescens*, *B. mori, A. mellifera, M. mediator*, *G. m. morsitans*, *T. castaneum*, *A. gossypii*, *S. guani*, *S. exigua*, *P.*

xylostella, *D. melanogaster* and *A. gambiae*. It is noted that the *Glossina* CSPs are more similar than to other Insects CSPs. There is good bootstrap support for many terminal relationships with CSP1 from Insects of different orders randomly distributed.

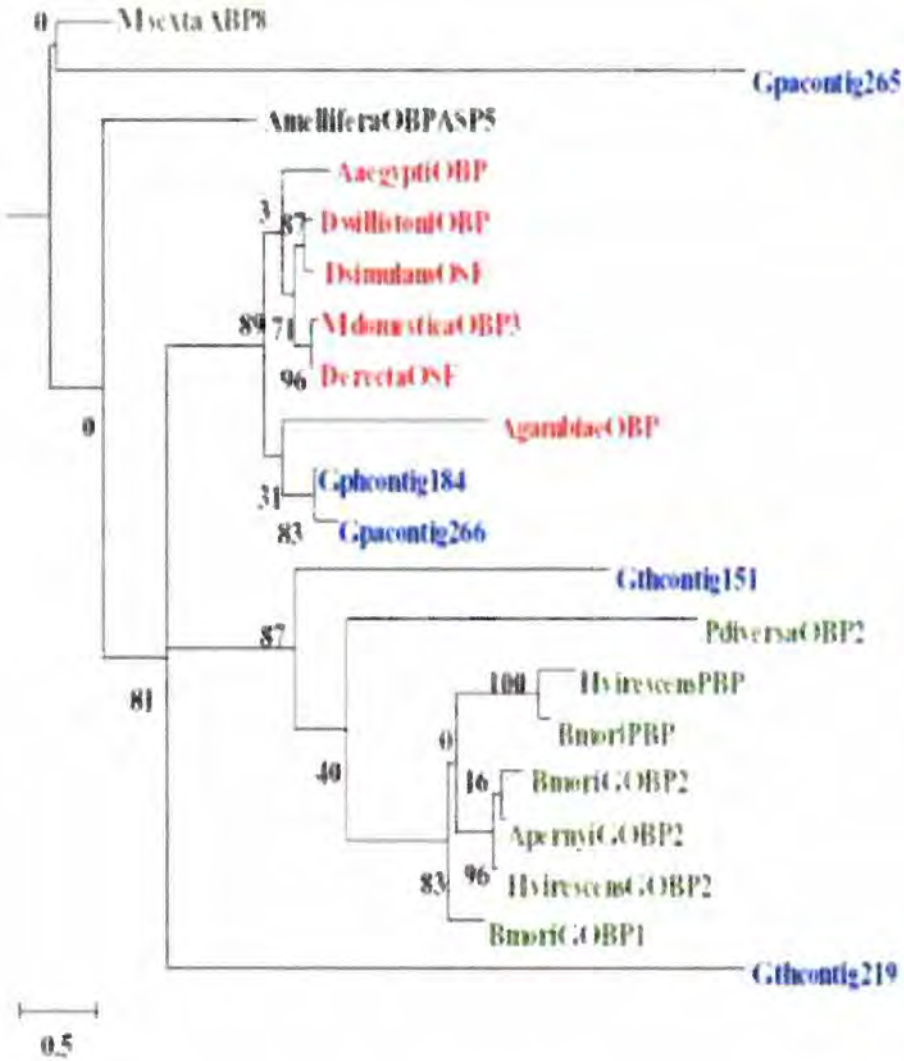

Figure 2.7Phylogenetic relationships of *Glossina* OBPs (blue) including [*Glossina pallidipes* OBPs - Gpacontig265, Gpacontig266; *Glossina palpalis gambiensis* OBP - Gphcontig184 and *Glossina tachinoides* OBPs - Gthcontig151 and Gthcontig219], with other OBPs downloaded from GenBank including Dipteran (red) *Drosophila erecta* olfactory Specific E, *Drosophila simulans* olfactory Specific E, *Musca domestica* OBP3, *Drosophila willistoni* OBP, *Aedes aegypti* OBP, *Anopheles gambiae* OBP; Lepidopteran (green) *Heliothis virescens* GOBP2, *Bombyx mori* GOBP1, *Bombyx mori* GOBP2, *Antheraea pernyi* GOBP2, *Phyllopertha diversa* OBP2, *Bombyx mori* PBP, *Heliothis virescens* PBP, *Manduca sexta* ABP8 and Hymenopteran (black) *Apis mellifera* OBPASP5. The tree was constructed from an alignment shown in Figure 2.6 using maximum likelihood model with PHYML program. Numbers on branches show values of 100 times replication bootstrap analysis.

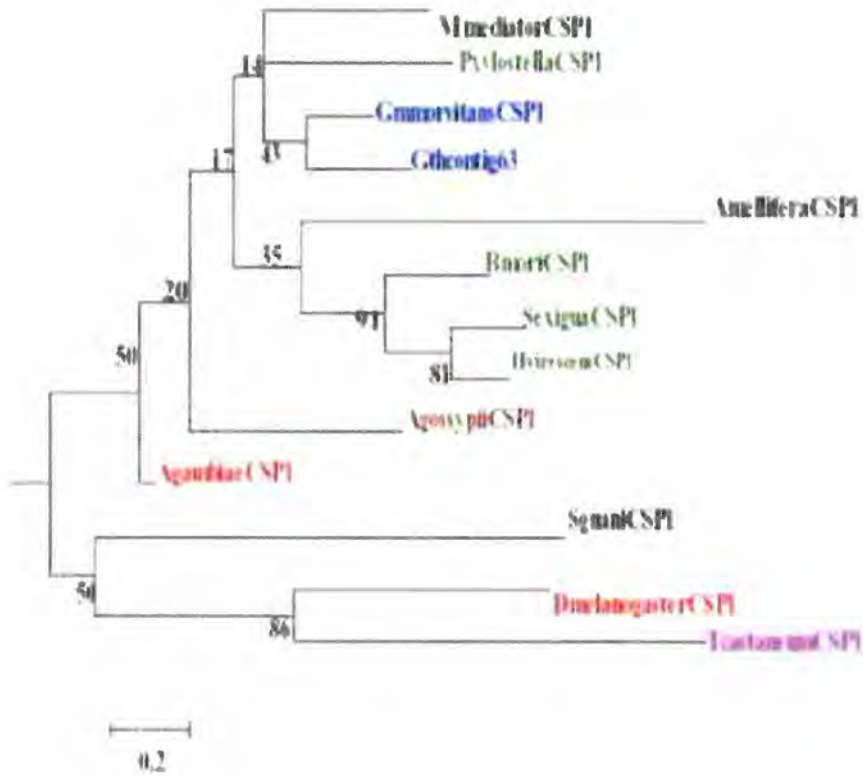

Figure 2.8 Phylogenetic relationships of *Glossina* CSPs (blue) including *Glossina tachinoides* CSP-Gthcontig63 and *Glossina morsitans morsitans* CSP1) with other insect CSPs downloaded from GenBank including CSP1 from Lepidopteran (green) *Heliothis virescens*, *Bombyx mori*, *Spodoptera exigua* and *Plutella xylostella*; Hymenopteran (black) *Apis mellifera*, *Microplitis mediator* and *Sclerodermus guani*; Dipteran (red) *Drosophila melanogaster* and *Anopheles gambiae*; Coleopteran (pink) *Tribolium castaneum* and Hemipteran (brown) *Aphis gossypii*. The tree was constructed from an alignment shown in Figure 2.7 using maximum likelihood model with PHYML program. Numbers on branches show values of 100 times replication bootstrap analysis.

2.4 Discussion

Clustering of the *G. pallidipes* antennal, *G. p. gambiensis* and *G. tachinoides* head ESTs generated large numbers of singletons in the three libraries, an indication of a highly diverse collection of mRNA species in antennae and head. A variety of molecules were identified from both the antennal and head libraries which could be implicated in either olfaction, energy metabolism, transcription, structural, transport activity, signal transduction or general metabolism. The different cellular and metabolic molecules identified may reflect complex processes that occur within the *Glossina* antennae and head thus contribute to the survival of tsetse flies in their habitats. Interestingly, similar types of genes were observed in the three libraries, implying that *G. pallidipes*, *G. p. gambiensis* and *G. tachinoides* genes may perform similar functions in these tissues. Proteins with transmembrane domains that were reported included opsins, retinin and arrestins. Apart from olfaction, they mediate sense of taste, vision and pain. Arrestins are involved in desensitization of GPCRs that mediate neurotransmission, olfactory and visual sensory reception (Spaethe and Briscoe, 2004). Identification of energy metabolism genes suggests that most processes occurring within the antennae and head require energy which could be provided by hydrolysis of ATP. In this work putative OBPs were identified from *G. pallidipes* (2), *G. p. gambiensis* (5) and *G. tachinoides* (2 OBPs and 1 CSP). This confirms occurrence of OBPs and CSPs in *Glossina*. Few OBPs were reported in this work, presumably because they are expressed at extremely low levels and proved hard to find using EST approaches (Robertson *et al.*, 1999). However, this will be made certain when the complete *Glossina* genome sequence

is available even though twenty two OBPs have also been identified in. *G. m. morsitans* (Liu *et al.,* 2010). Two putative *Glossina* OBPs (Gpacontig266 and Gphcontig184) had the hallmark features of OBP protein family which includes a signal peptide, a major hydrophobic domain and particularly the six conserved cysteine residues. The other eight putative OBPs had neither signal peptide nor six cysteines. Classification of these proteins as OBPs was based on significant similarity to other OBPs from NR protein and GO database. The predicted amino acids, MWs and pIs for the tsetse OBPs are similar to those of other insects (Vogt, 2003). The predicted pIs are mainly basic at physiological pH, unlike lepidopteran OBPs found to be acidic (Horst *et al.,* 2001; Lee *et al.,* 2002) while dipteran is between 4 and 10 (Kruse *et al.,* 2003; Wogulis *et al.,* 2006). The low sequence similarity of OBPs within and across insect species is consistent suggesting that insects OBPs are under strong adaptive selection (Vogt, 2005). Insect species are numerous and occupy diverse biological niches. Over time they have evolved to have different lifestyles and feeding ecologies where they encounter a wide range of chemicals. In order to locate mates, nutrients, resting sites, oviposition and larviposition sites, insects rely on their chemical senses, particularly gustatory and olfactory. Hence, the great diverse OBPs play an important role in insect's behavioral ecology. The large number of OBPs found in different insects could account for the wide range of chemicals each species detect in its environment (de Bruyne and Baker, 2008). Alignment of the selected insect and putative *Glossina* OBPs revealed a highly conserved region, LKCYMNC, around the second and third conserved cysteine.

Similar results have been reported for the yellow fever mosquito, *A. aegypti* with *D. melanogaster* and *An. gambiae* (Zhou *et al.*, 2008). The conserved amino acids in OBPs are an indication that these regions may perform an important function in the proteins.

Unlike OBPs, CSPs were more conserved at certain amino acids and had the characteristics four cysteines (Angeli *et al.*, 1999). They have been identified in both chemosensory and non-chemosensory tissues of different insects suggesting they are involved in functions other than olfaction (Sabatier *at al.*, 2003; Gong *et al.*, 2007). Despite identifying a CSP in *G. tachinoides*, it remains unclear if it could serve both chemosensory and non-chemosensory role. Phylogenetic analysis revealed occurrence of different olfactory proteins that could have evolved from a common ancestor. Further evidence is required to determine the rate of evolution for the *Glossina* OBPs which could pinpoint the extent of divergence. However, the good bootstrap values indicate that olfactory proteins from different insect orders are quite similar implying they could have evolved from a common ancestor and diversified to perform similar role in different insects.

The major finding of this EST project was identification of OBPs and CSP in *Glossina*. This open avenues for studying molecular basis of olfaction in *Glossina* with the possibility of developing models to account for different behavioural responses to hosts during feeding and better methods to control tsetse vectors based on olfactory driven behaviours.

CHAPTER 3

3.0 COMPARATIVE ANALYSIS OF *GLOSSINA* EXPRESSED SEQUENCE TAGS (ESTS): IDENTIFICATION OF ORTHOLOGS ODORANT-BINDING PROTEINS AND CHEMOSENSORY PROTEINS OF DIPTERAN SPECIES

3.1 Introduction

In this study, a comparative analysis of three *Glossina* ESTs libraries (*G. pallidipes* antennal, *G. tachinoides* and *G. p. gambiensis* head) with various Dipteran proteomes, including *G. m. morsitans*, *Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* was done. A total of nine putative OBPs and one CSP were identified with homologs reported for ubiquinol-cytochrome-*c* reductase and salivary gland OBPs. Comparative sequence analysis of the OBP and CSP homologs revealed a highly diverse and conserved multi gene protein family respectively, supported by good bootstrap values under phylogenetic analysis. These results should speed up determination of precise role the identified orthologs play in contributing to vectorial capacity of the analysed Dipteran insects.

3.2 Materials and Methods

3.2.1 Processing of *Glossina pallidipes* antennal, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head ESTs

The Expressed sequence tags (ESTs) libraries (*G. pallidipes* antennae (1127), *G. tachinoides* head (830) and *G. p. gambiensis* head (906) were downloaded from the Sanger Center ftp site (ftp://ftp.sanger.ac.uk/pub/pathogens/Glossina/).

The sequences were trimmed of primer and vector sequences, clustered using StackPACK™ v2.2 (Miller *et al.,* 1999), generating singletons (sequences not included in any clusters) and contigs (consensus sequences containing two or more sequences). For each organism, singletons and contigs were analysed identically, thus were combined in a single set of sequences, referred hereafter as clusters.

3.2.2 Bioinformatics analyses of *Glossina pallidipes* antennal, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head ESTs against *Glossina morsitans morsitans* Proteins

The resulting sequences were compared to the *G. m. morsitans* proteins available at GeneDB (Hertz-Fowler *et al.,* 2004) using wublastx (Gish and David, 1993). The dataset (*G. m. morsitans* proteins) contains sequences derived from different tissues and developmental stage specific libraries: head (2,700 ESTs), midgut (21,662 ESTs), reproductive organs (3,438 ESTs), salivary gland (27,426 ESTs), larvae (2,304 ESTs), pupae (2,304 ESTs), fat body (20,257 ESTs), male and female whole bodies (19,968 ESTs) (Hertz-Fowler *et al.,* 2004). StackPACK™ v2.2 was used in clustering the ESTs since the *G. m. morsitans* ESTs derived from different libraries had also been clustered using StackPACK software. *G. m. morsitans* sequences which gave the best hits (E-value <0.0) were collected as putative orthologs and grouped into eight different categories: odorant/pheromone binding, structural, transport, signal transduction, nucleic acid metabolism, energy metabolism, salivary gland and basic metabolism.

3.2.3 Bioinformatics analyses of *Glossina pallidipes* antennal, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head ESTs against Selected Dipteran Proteins

These sequences (from *G. pallidipes* antennae, *G. tachinoides* head and *G. p. gambiensis* head) were then used to search phylogenetically related sequences using blastx (Altschul *et al.,* 1997) against the *D. melanogaster* (FlyBase5.13) (Ensembl), *An. gambiae* (AgamP3.5), *Ae. aegypti* (AaegL1.2) and *C. quinquefasciatus* (CpipJ1.2) (VectorBase) protein databases (Lawson *et al.,* 2009). The sequences grouped as putative odorant/pheromone binding proteins from the three *Glossina* species (*G. pallidipes, G. tachinoides* and *G. p. gambiensis*) and the selected Dipteran species (*G. m. morsitans*, *D. melanogaster*, *An. gambiae, Ae. aegypti* and *C. quinquefasciatus*) were aligned by ClustalW2 (Larkin *et al.,* 2007). The genetic distance was determined by phylogenetic analysis using PHYML (Guindon and Gascuel, 2003) and the neighbour joining algorithm, maximum likelihood model and WAG substitution matrix (Whelan and Goldman, 2001). The consensus tree was generated based on 100 bootstrap replicates and viewed using TreeDyn (*Chevenet et al.,* 2006).

3.3 Results

3.3.1 Clustering of *Glossina pallidipes* antennal, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head ESTs

A total of 2863 ESTs were clustered by StackPACK^{TM} v2.2 and generated 618 singletons and 45 contigs in *G. pallidipes* antennal library.

Similarly, 387 singletons and 43 contigs were generated in *G. p. gambiensis* head library while 404 singletons and 40 contigs were generated in the *G. tachinoides* head library (Table 3.0). Singletons included sequences not found in any clusters and contigs are consensus sequences containing two or more sequences. Singletons and contigs are hereafter referred to as clusters, which could encode possible transcripts. Compared with the clusters generated from cDNA Annotation Software (CAS), StackPACK™ v2.2 software generated more singletons and contigs resulting in a high number of sequences within the clusters. The singletons and consensus/contigs generated by the two softwares (CAS and StackPACK™ v2.2) included EST clones that could represent the same gene when compared with other protein sequences in different databases.

Table 3.0. Summary of EST sequences and clusters generated from *Glossina pallidipes* antennae, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head libraries

	ESTs	**Singletons**	**Contigs**	**Total Sequences in Clusters**
G. pallidipes	1127	618	45	516
G. palpalis gambiensis	906	387	43	529
G. tachinoides	830	404	40	431
Total	**2863**	**1409**	**128**	**1476**

3.3.2 Categories of *Glossina pallidipes* antennal, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head clusters

In order to identify orthologs between the four *Glossina* species, the clusters (singletons and contigs) were compared by wublastx (Gish and David, 1993) to the

G. m. morsitans proteins from the GeneDB database (Hertz-Fowler *et al.*, 2004). This resulted in 120 (18.1%), 154 (35.8%) and 170 (38.3%) positive hits for the *G. pallidipes*, *G. p. gambiensis* and *G. tachinoides* clusters respectively, which were grouped into eight categories: odorant/pheromone binding, structural, transport, signal transduction, nucleic acid metabolism, energy metabolism, salivary gland and basic metabolism (Table 3.1). The majority of the clusters from *G. pallidipes*, *G. p. gambiensis* and *G. tachinoides* (199 sequences in total) fall into the basic metabolism category and included male-specific sperm, iron transport ferritin, reproduction-associated vitellogenin, kinases and trypsin inhibitor Kunitz domain. A total of eighty-one clusters were linked to salivary glands category. Other categories identified included odorant/pheromone binding, signal transduction, energy and nucleic acid metabolism with 8, 23, 54 and 44 clusters respectively. Structural and transport categories each had 16 clusters.

Table 3.1. Number of clusters found in *Glossina pallidipes* antennae, *Glossina palpalis gambiensis* head and *Glossina tachinoides* head

Category	**Number of *G. pallidipes* clusters**	**Number of *G. palpalis gambiensis* clusters**	**Number of *G. tachinoides* clusters**
Odorant/Pheromone Binding	2	5	1
Structural	5	6	5
Transport	5	6	5
Signal transduction	8	6	9
Nucleic Acid Metabolism	8	15	21
Energy Metabolism	6	23	25
Salivary gland	29	26	26
Metabolism	57	64	78
Total	**120**	**151**	**170**

3.3.2.1 *Glossina pallidipes* clusters with matches to *Glossina morsitans morsitans* protein database

The 120 clusters from *G. pallidipes* antennae yielded a diverse large number of genes with similarity ranging from 23% to 100% (Table 3.2). Two clusters (cn30 and cn31) had low matches to PBP-related protein (29% similarity) and general odorant-binding protein (23% similarity) respectively. Six clusters were linked to energy metabolism with clusters gpafl-1a01.p1k and gpafl-6d02.p1k having a high similarity to NADH-ubiquinone oxidoreductase (100% similarity) and cytochrome c oxidase (93% similarity) respectively. A total of eight clusters were implicated in nucleic acid metabolism while both structural and transport category each had five clusters. Eight clusters gave matches to signal transduction molecules and a greater proportion of the clusters were either related to salivary gland (29 clusters) or basic metabolism (57 clusters) (Appendix IV).

Table 3.2 Olfactory related clusters from *Glossina pallidipes* antennae, *Glossina palpalis gambiensis* and *Glossina tachinoides* head producing best matches to *Glossina morsitans morsitans* protein GeneDB

Cluster	Size (bp)	Best GeneDB Match	E-value
gpacn30	771	29% to LAR-001B13.b PBP-related protein 5 precursor	0.00067
gpacn31	529	23% to cn15569 General odorant-binding protein 99a	0.994
gphcn109	746	31% to LAR-001B13.b PBP-related protein 5 precursor	0.00064
gphfl-8a06.p1k	488	26% to cn7404 General odorant-binding protein 99b precursor	0.99994
gphfl-5g01.p1k	568	67% to cn14025 Odorant binding protein	1.1E-44
gphfl-8d10.p1k	658	35% to cn15569 General odorant-binding protein 99a precursor	8.5E-22
gphfl-9d09.p1k	58	40% to cn15569 General odorant-binding protein 99a precursor	0.997
gthfl-5d10.q1k	393	28% to cn15569 General odorant-binding protein 99a precursor	6.1E-07

gpacnxx – *Glossina pallidipes* antennae; gphcn/gphf1xxxx – *Glossina p. gambiensis* head; gthf1xxxx – *Glossina tachinoides* head.

3.3.2.2 *Glossina palpalis gambiensis* clusters with matches to *Glossina morsitans morsitans* protein database

Blast results for *G. p. gambiensis* (154 clusters) matches to *G. m. morsitans* protein database are compiled as shown in Appendix V. Clusters cn109 and gphfl-5g01.p1k had significant matches to PBP (31% similarity) and OBP (67% similarity) respectively while clusters gphfl-8a06.p1k, gphfl-8d10.p1k and gphfl-9d09.p1k were all related to general odorant-binding proteins (GOBPs) with percentage of identity ranging from 26% to 40% (Table 3.2). Twenty-three clusters have significant indications of being involved in energy transduction processes

while the nucleic acid metabolism category included fifteen clusters. Notably, structural proteins, transport and signal transduction categories each had six clusters. The structural proteins included actin, cuticle and histone while transport category had acyl carrier proteins, ADP/ATP translocase, cation and tricarboxylate transport protein. Most signal transduction clusters are involved in vision. Salivary proteins category had 26 clusters and 64 clusters were attributed to basic metabolism category.

3.3.2.3 *Glossina tachinoides* clusters with matches to *Glossina morsitans morsitans* protein database

Significant matches reported for the 170 clusters from *G. tachinoides* were similar to those of *G. p. gambiensis* (Appendix VI). Only one cluster (gthfl-5d10.q1k) was related to GOBP with 28% identity (Table 3.2), while energy and nucleic acid metabolism categories had 25 and 21 clusters respectively. Structural and transport category each had 5 clusters with majority of structural clusters being adult cuticle protein and one cluster (gthfl-4g01.q1k) had tubulin fragment which wasn't reported for *G. palpalis gambiensis*. The transport category had a GDP-fructose transporter and sodium: neurotransmitter symporter which were also not identified in *G. p. gambiensis*. Transcripts implicated in signal transduction were 9 while salivary gland proteins constituted 26 clusters. Majority of the clusters (78) represented various proteins involved in basic metabolic processes with cluster cn65 having ejaculatory bulb-specific protein-3 (95% identity) as its ortholog.

3.3.3 Comparison of *Glossina* Clusters with *Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* protein databases

In order to predict functions of reported orthologs between the dipteran insects, *Glossina* clusters within the eight categories were used to search *D. melanogaster* (Ensembl) (Appendix VII), *An. gambiae*, *Ae. aegypti* and *C. quinquefasciatus* (VectorBase) protein databases (Appendix VIII) because they are phylogenetically closely related and, importantly, have a well annotated genome (Lawson *et al.*, 2009). Nine clusters were identified through BLAST search to belong to odorant/pheromone binding category. The structural category was represented by 10 clusters, which included mainly cuticle protein, histones, prion, proline/alanine rich regions, armadillo and profilin. Twelve clusters were found to be in the transport category and could be involved in transport of nucleotides, cations and organic molecules. The cell or neuron signalling category had 10 clusters with a majority implicated in vision and included arrestin, opsin, phosrestin, rhodopsin, and dopamine receptor. Orthologs linked to nucleic acid metabolism were mainly ribosomal proteins, few transcription initiation factors and core binding factors. Homologs reported under the energy metabolism category were mainly enzymes involved in mitochondria transport chain. Different orthologs were reported for salivary gland category and included ribosomal protein, retinin like protein, thiolase, adenosine, fibrinogen, salivary growth factor, rhodopsin receptor, cuticular proteins. It could be that the *Glossina* salivary gland clusters are not well annotated hence the discrepancy in finding orthologs in the dipteran species.

3.3.4 Identification of Putative OBP genes

A total of nine putative odorant-binding proteins (OBPs) and one chemosensory protein (CSP) were identified in *G. pallidipes* (2 OBPs), *G. p. gambiensis* (5 OBPs), *G. tachinoides* (2 OBPs and 1 CSP) clusters by searches against *G. m. morsitans, D. melanogaster*, *An. gambiae*, *Ae. aegypti* and *C. quinquefasciatus* protein databases (Table 3.3). The two *G. pallidipes* OBPs (*GpOBP1*, *GpOBP2*) and one of the *G. p. gambiensis* OBP (*GpgOBP1*) were most likely related to genes encoding pheromone binding proteins (PBPs) in *G. m. morsitans, D. melanogaster* and *C. quinquefasciatus* with percentage identity varying from 23% to 70.1%. The four other *G. p. gambiensis* OBPs had different homologs, with *GpgOBP2* having no homolog in *D. melanogaster* and *GpgOBP4* predicted to encode general odorant binding protein (GOBP) (31.1% sequence identity) in *C. quinquefasciatus*. Interestingly, *GpgOBP3* had a match against GOBP (26% identity) in *G. m. morsitans* and ubiquinol-cytochrome-*c* reductase (63.6% identity) in *D. melanogaster, An. gambiae* and *Ae. aegypti* while the homologous gene reported in *C. quinquefasciatus* was a mitochondrial containing fragment having 55.7% sequence identity. The other OBP gene (*GpgOBP5*) was similar to the *G. m. morsitans GmmOBP1* (40% identity), and had no match in *Drosophila* and mosquitoes. The only CSP gene reported in this study was in *G. tachinoides* (*GtCSP*); it had a high degree of similarity (95% identity) to the Ejaculatory Bulb Serum Protein-3 (EBSP-3) in *G. m. morsitans*. It also matched a Pherokine-3 protein (49.2% identity) in *D. melanogaster*, the CSP1 protein (49.6% identity) in *C. quinquefasciatus*, as well as the Olfactory Specific-D (OS-D)/PhBP protein

(61% identity) and the Antennal protein 10(A10)/PBP/OS-D (60% identity) in *Ae. aegypti* and *An. gambiae* respectively. The other two *G. tachinoides* genes (*GtOBP1* and *GtOBP2*) could be linked to GOBP, OBP or PBP. *GtOBP2* gene had a match against the TsalI salivary gland protein (45% identity) in *G. m. morsitans.*

Table 3.3 Putative Odorant Binding Proteins (OBPs) identified from *Glossina pallidipes* antennal, *Glosinna tachinoides* and *Glossina palpalis gambiensis* head libraries

Cluster	Size (aa)	Best *G. morsitans* Match (Identities, *E* value)	Best *D. melanogaster* Match (Identities, *E* value)	Best *An. gambiae* Match (Identities, *E* value)	Best *Ae. aegypti* Match (Identities, *E* value)	Best *C. quinquefasciatus* Match (Identities, *E* value)	Signal Peptide
Gpacn30 (OBP1)	143	PBPRP-5 (29%, 6.7e-4)	Pbprp-3 (70.1%, 8.e-41)	OBP17 (51.7%, 2e-037)	OBP (48.2%, 3e-36)	OBP (48.3%, 2e-35)	1 - 11aa
Gpacn31 (OBP2)	98	GmmOBP1 (23%, 0.994)	Os-E (66.7%, 4.6e-13)	OBP17 (49.5%, 2e-20)	OBP (43.8%, 3e-18)	OBP/ PhBP (45.5%, 5e-19)	1 - 16aa
Gphcn109 (OBP1)	123	PBPRP-5 (31%, 6.4e-4)	Os-E (63%, 5.1e-39)	OBP17 (53.7%, 1e-34)	OBP (52%, 1e-34)	OBP/ PhBP (52%, 3e-34)	1 - 14aa
Gphfl-5g01.p1k (OBP2)	99	OBP (67%, 1.1e-44)	No Hits	OBP (32.6%, 1e-8)	OBP (28.3%, 2e-7)	OBP (29.3%, 5e-9)	1 - 22aa
Gphfl-8a06.p1k (OBP3)	88	GOBP99b (26%, 0.994)	Ubq-cyt c red (72.9%, 1.4e-40)	Ubq-cyt.c red (53.9%, 1e-20)	Ubq-cyt.c red (64%, 4e-23)	Mitochondrial (55.7%, 1e-22)	No Signal P
Gphfl-8d10.p1k (OBP4)	136	GmmOBP1 (35%, 8.5e-22)	OBP83g (61.4%, 3.4e-65)	OBP9 (32.4%, 2e-17)	OBP 99c (31.6%, 4e-17)	GOBP (31.1%, 6e-17)	1 - 18aa
Gphfl-9d09.p1k (OBP5)	45	GmmOBP1 (40%, 0.997)	No Hits	No Hits	No Hits	No Hits	NA
Gthcn65 (CSP)	123	EBSP-3 (95%, 2.8e-62)	Pherokine-3 (49.2%, 2e-37)	PBP/A10/OS-D (60%, 5e-41)	OS-D/PhBP (61%, 1e-41)	CSP 1 (49.6%, 8e-35)	1 - 18aa
Gthfl-5d10.q1k (OBP1)	77	GmmOBP1 (28%, 6.1e-7)	OBP83g (63.6%, 2.1e-18)	OBP9 (28.6%, 0.002)	OBP 99c (29.5%, 8e-4)	GOBP 99a (27.3%, 0.002)	No Signal P
Gthfl-8b10.p1k (OBP2)	58	Tsal1 (45%, 0.74)	No Hits	OBP17 (42.6%, 9.0e-9)	OBP 56a (41.4%, 2e-8)	OBP/ PhBP (39.3%, 1e-7)	1 - 17aa

GpacnXX - *Glossina pallidipes* antennae cluster; GphcnXXX - *Glossina palpalis* head cluster; GthcnXX - *Glossina tachinoides* head cluster; Gphf1-XXXX - *Glossina palpalis* head singleton; Gthf1-XXXX - *Glossina tachinoides* head singleton.

The multiple alignments of the OBP gene sequences between *G. pallidipes, G. p. gambiensis*, *G. tachinoides* and *G. m. morsitans* revealed a highly diverse protein family (Figure 3.0). Most relationships have strong bootstrap support with the lowest identity being 13% between GmmOBP3 and GtOBP2, GmmOBP18, GmmOBP19 (Figure 3.4). Two OBPs (GpgOBP2 and GmmOBP17) are probably not related to the other *Glossina* OBPs as they didn't form any clusters. There is no bootstrap support for GmmOBP8 and GpOBP1, GmmOBP18 and GmmOBP19. The closely related sequences that clustered together with high identities of > 90% included: GtOBP1, GmmOBP11 and GpgOBP4 (97%); GmmOBP21 and GmmOBP22 (96%); GmmOBP and GmmOBP1 (98%); GmmOBP13 and GmmOBP14 (91%); GmmOBP5 and GmmOBP6 (94%); GmmOBP10 and GmmOBP12 (97%) and with GmmOBP9 (96%) (Figure 3.3).

Comparison of the putative *Glossina* OBPs with the OBPs from the four Dipteran (Figure 3.1) also showed a highly diverse gene family; a result also reported in other studies of several insect species (Zhou *et al.*, 2008; Liu *et al.*, 2010). Clustering of the clades indicates that the mosquitoes OBPs could have evolved separately from the *Glossina* and *Drosophila* OBPs (Figure 3.4). This is supported by evolutionary studies which indicate that Diptera split into two lineages about 260 million years ago, i.e. Brachycera (which includes both *Glossina* and *Drosophila*) and Nematocera (which includes mosquitoes). Clustering of *G. m. morsitans* and *D. melanogaster* OBPs separately from mosquitoes OBPs was also reported by Liu *et al.*, 2010. The OBPs identified from *G. pallidipes* (GpOBP1 and (GpOBP2) and *G. p. gambiensis* (GpgOBP1) are all related to GmmOBP8 with at least 69% bootstrapping support (Figure 3.4). GmmOBP8 was predicted to encode

classic OBPs and was found to be more expressed in the antennae of females than males *G. m. morsitans* (Liu *et al.,* 2010). Hence, the three OBPs (GpOBP1, GpOBP2 and GpgOBP1) could be involved in mediating olfaction process in *G. pallidipes* and *G. p. gambiensis.*

Both GpgOBP4 and GtOBP1 share a common ancestor with GmmOBP11 with the closest *D. melanogaster* homolog being OBP83g (84% bootstrap support). The two proteins (GmmOBP11 and DmelOBP83g) have been reported to play an important role in olfaction (Liu *et al.,* 2010; Graham and Davies, 2002; Hekmat-Scafe *et al.,* 2002) and this could implicate GpgOBP4 and GtOBP1 in odor binding suggesting occurrence of orthologs in the different *Glossina* species. Clustering of *D. melanogaster* OS-E and Pbprp3 to GmmOBP9 has also been reported and it was observed that GmmOBP9 is highly expressed in antennae (Liu *et al.,* 2010) while OS-E and Pbprp3 (OS-F) are normally co-expressed with the *D. melanogaster* pheromone-binding protein LUSH in one neuron. The clustering analysis shows that some of *G. m. morsitans* OBPs are more closely related: GmmOBP13 and GmmOBP14 (30% bootstrap), GmmOBP15 and GmmOBP4 (47% bootstrap), GmmOBP6 and GmmOBP5 (45% bootstrap), GmmOBP22 and GmmOBP21 (68% bootstrap), GmmOBP3 and GmmOBP7 (37% bootstrap) (Figure 3.4). Interestingly, GmmOBP1 has a strong bootstrap support with GmmOBP (100%) and was found to be the only *Glossina* OBP that clustered with the three mosquitoes OBPs (AaegOBP99c, AgamOBP9 and CquiOBP99a). GmmOBP1 has been implicated in blood feeding in insects (Liu *et al.,* 2010).

Multiple sequence alignment of *G. tachinoides* CSP (GtCSP) with the identified orthologs in *D. melanogaster*, *An. gambiae*, *Ae. aegypti*, *C. p. quinquefasciatus* and selected homologous CSPs (*Bombyx mori* Chemosensory protein 4; *Papilio xuthus* Chemosensory proteins; *Tribolium castaneum* Chemosensory protein 9; *Heliothis virescens* Chemosensory protein 1; *Mamestra brassicae* Chemosensory protein; *Schistocerca gregaria* Chemosensory protein; *Locusta migratoria* Chemosensory protein; *Apis mellifera* Chemosensory protein 1; *Anopheles gambiae* Chemosensory protein 5 and *Apis cerana cerana* Chemosensory protein), downloaded from GenBank revealed a highly conserved class of olfactory proteins (Zhou *et al.*, 2006).

The amino acid sequence of GtCSP was also consistent with the CSP family of other insects in many features i.e. predicted size of 123 amino acids, signal peptide of about 18 amino acids (Table 3.3), four conserved cysteines, three highly conserved aromatic residues, a lysine residue between cysteines 3 and 4 and a glutamine residue at the fourth amino acid from cysteine 4 (Figure 3.2). The *G. tachinoides* CSP (GtCSP) have the first, second and third aromatic amino acids as phenylalanine, tryptophan and tyrosine respectively (Figure 3.2). The 4 conserved cysteines and the 3 conserved aromatic residues are probably involved in formation of 2 disulfide bridges and in the binding sites respectively (Angeli *et al.*, 1999).

```
GpOBP1     ---------------------------------IVTFAMTLLLSFGLNNAQKPRRDENYPPPDFLKSFKIIHDVVVEKT-GATEEAIKEFSDG------------EIHEDPALKYYMNVL 74
GmmOBP8    --------------------------MKKYHIYIVTFAITLLMSFGLNNAQKPRRDENYPPPDFLKSFKIIHDVVVEKT-GATEEAIKEFSDG------------EIHEDPALKYYMNVL 81
GpgOBP1    ------------------------------------------------------PRRDENYPPPDFLKFFKIIHEVVVEKT-GATEEAITEFSDG------------EIHDDPALKYYMNVL 55
GpOBP2     ------------------------------------------------------------------------------HYFVEEKT-GVTGGANKKFSDG------------ENQEDPALKYYMNVL 36
GmmOBP10   ------------------------------------------------------------------------------------------EAIKEFSEG------------NIHEDEALKYYMNVF 24
GmmOBP12   ------------------------------------------------------------------------------------------EAIREFSDG------------EVHEDEALKYYMNVF 24
GmmOBP9    ------------------------MTLSGKYRFLTVYTMLIVLLSAWTRAQQPRRDDEYPPPAILKLAKPFHDIVVEQT-GVKEEAIKEFSDG------------EIHEDEALKYYMNVL 83
GmmOBP18   ------------------------------------------------------------------------------------H-MRYAEYLLAALYN------------NTEHTESFKYYLQII 29
GmmOBP19   ------------------------------------------------------------------------------------H-MRYAEYLLAALYN------------NTEHTESFKYYLQII 29
GpgOBP4    ------------------------------MKLFLIILSTTVALVN----------AKFDIRTKDDALKAHEEVHEEF-NVPDDIYEQYLDY------------QFPEHKLTNYYVKLW 66
GmmOBP11   ------------------------------------------------------------TKDDALKAHEEVHEEF-QVPDDIYEQYLDY------------QFPEHKLTNYYVKLW 44
GtOBP1     ------------------------------------------------------------------------------------------------------------EKVKLW 7
GmmOBP1    ------------------------------MKTTAVIL-LALFALVS----------ADYKLRNQEDLNKARKEVMEAK-KVTPELVEKYKKF------------DFPDDEITRYYIEII 66
GmmOBP     ------------------------------MKTTAVIL-LALFALVS----------ADYKLRNQEDLNKARKEVMEAK-KVTPELVEKYKKF------------DFPDDEITRYYIEII 66
GmmOBP2    ------------------------------MKTIIVIVFLVTLATVWGHHHHEHHDDDDYVVKTREDLFKYRDEVSNKL-NVPADLLEKYKKW------------QYPDDEVTKYYMKVM 77
GmmOBP21   ------------------------------------------------------AEDEDWQPKTVADIKSIRNEVLKEH-PLSNEQITKMKNF------------EFPDEEEVRQYLLVT 53
GmmOBP22   ------------------------------------------------------------------DDFFQMSERVMRLE-KVPDRYKAQFTEF------------QFPNDPIVHKYILVV 42
GmmOBP16   ------------------------------------------------------------------------------------ENFNAFQSI------------DMEPDRKFKYYAHVL 24
GmmOBP14   ------------------------------------------------------------ATEEQMRSAANLMRDVVLPKFPKVSKETADGIRNG------------NLSDNKDAKYYINVV 50
GmmOBP3    IPCFARCIASEKGWFDIDLSRWNKQRLVDELGANMYNYCRFELNRAFKNVCSFAFKGLKCLKQAEMNVIITHNNLLECVKEKSISMDQLLEYY------------HFPQLEHIPVLFKVF 108
GmmOBP7    FPCFTNCYLNNMFNIYNETQGFNEENVIKRFGRSVYNACKEKLIQGN-NSCEIAYNGFHCLINREDDPFILIDNIEDISMEAKRAMKECLHKFNTDEWQYLSDYVRFPVQEPIPVYTAVF 119
GmmOBP13   ------------------------------------------------------------TKDDFEKILQSVREDM-QINENDLRTLSAS------------PNDVSEGVKYYMKVV 44
GmmOBP15   ------------------------------------------------------------TRETLQNYVKTVVIEENISTKDLKLFMAWN------------FSNISNEGKVFFSVF 45
GmmOBP4    ----------------------------MKSKIIFDTRVRRKVMFRVTLILLAIVTSALFSENRYMEFLADFKHVKRERGVGRFELDRLRVGN------------LAYPSYEAVFLGVL 80
GmmOBP17   ------------------------------------------------------------------------NIPGRFNLPNYSLTDHFLGQIEYA--------------EIAPKNAKVFLRVW 38
GmmOBP5    ---MRFHIILKLMSWMCLMCAIESKNVIDLLEGKMYAPAQYQLKPADNFASSPVNKRQMPTSDIPKNMQQFQDTLNEAKFKCARAMRLDSNKLLMY---------EDQPSLREKVLMAII 108
GpgOBP3    ------------------------------------------------------------MPLRSFLKDFIALPVVKADDEELVDPKV------------TLREQCNNKAHISAL 43
GmmOBP6    ------------------------------------MFKLLLVTVLMLGILSVEAEIDVQEEIAKFILLANEVREEVGAKEADIQDLIHKH-------------PSAGQEGKVLRAVL 69
GmmOBP20   ------------------------------------------------------------------------------------MLFTLFLIVFIFS-------------SREASALNETSRFV 27
GpgOBP2    ------------------------------------------------------------------LHDAHFVVTINVELEKRQGKNYI------------KFKKLVAMVQPKNTT 38
GtOBP2     ------------------------------------------------------FFFFFFFFFVQLTIYYVIIIIFSISNNLYPRKKDY--------------LISLRNSAWMGVR 48
```

```
GpOBP1     FHEVNVVDDAGE-LHFEKLVRMIPEPF-----LEMVKHIIDAESH------IPKGETQMDRAWSWHVTFKQTDPVLYFLP------------------------------- 143
GmmOBP8    FHEVNVVDDAGE-LHFEKLVRMIPEPF-----LEMVKHIIDAESH------IPKGETQMDRAWSWHVTFKQTDPVLYFLP------------------------------- 150
GpgOBP1    FHEVDVVDDAGE-LHFEKLVRMIPEPY-----LKMFQHILDAVSH------IPKGETQMDRAWSWHVTFKQTDPVLYFL-------------------------------- 123
GpOBP2     LH-VKVFDDAGE-LHFEKLVGMIPKPF-----LEMVKHIIDAESH------IPKGETQMDRAWSWHVTFKQTDP------------------------------------- 98
GmmOBP10   FHELGLVDDKGD-VHLETLHQSMPGSF-----VDLILKPAQHTVH--------PEGDTLHHKAWWFHQWWKKADPVVSNLMEET---------------------------- 94
GmmOBP12   FHELGAVDDKGD-VHLETLNLIMPGSF-----VEAILKPAQHIIH--------PEGDTLHHKAWWFHQWWKKADPEVSNLAQESL--------------------------- 95
GmmOBP9    FHEFDVVDDNGD-VHLEKLFSKIPAAL-----RDLLMEASKGTVH--------PEGDTLHHKAWWFHQWWKKADPVHYFLV------------------------------- 150
GmmOBP18   FDSLGLVDSNNQ-VNLEKLINFAPTEI-----HEHILELHRAIDTQRKLVDVIPAGKDSTDIVYTTSQWYYELKPASREYIEYMMH-------------------------- 109
GmmOBP19   FDSLGLVDSNNQ-VNLEKLINFAPTEI-----HEHILELHRAIDTQRKLVDVIPAGKDSTDIVYTTSQWYYELKPASREYIEYMMH-------------------------- 109
GpgOBP4    IEKMGIFTENRG-FNEKNIIAQYTFEN--YKNLESVRHGLEKTIDH------NEWETDVTTWANRVFSTWLKVNRHVVR--------------------------------- 136
GmmOBP11   VEKMGIFTENRG-FNEKNIVAQYTYEN--FKNLESVRHGLEKTIDH------NEWETDVTTWANRVFSTWLKVNRHVVRKMFT------------------------------ 118
GtOBP1     LEKMGIITENCG-FNEKNVKVPYSLEN--FQKLESVRQGPEKTLDH------NEWETEVTTWGYRVFSTWVKVNCHVVR--------------------------------- 77
GmmOBP1    FDKFQLFDSQTG-FKNDNLIAQLGQSK---DNKDEVKADIEKTADK------NTEKSDSTTWAFRGFKTFISKNLPLVMESLKKN---------------------------- 141
GmmOBP     FDKFQLFDSQTG-FKNDNLIAQLGQSK---DNKDEVKADIEKTADK------NTEKSDSTTWAFRGFKTFISKNLPLVMESLKKN---------------------------- 141
GmmOBP2    FEHFGFFNEKQG-FDVHKIHKQLMGAHGTVDHSDETHEKIAKTADK------KPEDTDPTAWAYRGGVTFINSNLQLVKSSVN------------------------------ 153
GmmOBP21   ALKMEVFCAHQG-YHPNRIAKQFKMDMN----EEEVLEIAEKTHDS------NPDNSSVDVWAFRGHKTMMSSAIGDKVKAYIKKRQEENAAKNA----------------- 137
GmmOBP22   NRELQIWDNNQG-FDIEKIYQQYKGRAN----EEVVLPIISQTNQD-------AKQRNYELWCYKAFLTILDTQVGEWFKEDVRRQQTRTLTNGHQ----------------- 126
GmmOBP16   LSNLKYLNTFSGKFDIEDFKQQDGIED-------EDVAVIAKTKKL------YDNINDPTEYGFNILQTILMFEPTE------------------------------------ 88
GmmOBP14   MEMMRTMKKGKFLYEGALKQVDLLMPDS---YKEEYRPGLAKTKDS------ANGIKNNTDAAYAVLSTLRGEITQFVFP--------------------------------- 121
GmmOBP3    ADKSHLYTVNYEWNVLNWLKAFGPIRN------ENADISICRVNAN------EREKMDITAIMYEEYNTWERLNYNTDGISVTYKKALKKIFNF------------------- 190
GmmOBP7    VYKMQLYNHRLRSWNIAAMQRLLGVPA--------EHANIENTLSLS-----KRRNNNMTAWIYKEMTTFSLSQ---------------------------------------- 180
GmmOBP13   MEKQGHFK--NGALLEEAVIKSLESSPADHNDQNQMSAIVKETKKE--------IGSNETETAFKVSMTLREHKVDFEI---------------------------------- 113
GmmOBP15   HEKIGLTINGVLQKKIAFGHLKRIFDR------ETAEFVLGETVNL--------VGKDKTETAYQFEKTLFNIEYNRLAK--------------------------------- 111
GmmOBP4    YERTGILKNGVLQNDVLKKNVGYIANR------VLLDEVLPPTYAVS--------GTNKTDIAFELKKTFKNVGFDKVWITVPWEDNTDPQYIAAMRLIDDLANVKYRVAFA 178
GmmOBP17   YKKMGILKENLV-TSAGPIPELRQHMR-------ECNEVATEWAQN-------QSNGDETEFAWSFYTTMHESLVKCLT------------------------------- 102
GmmOBP5    LKRMKLMDSDYKLSVPTISHIAGMISDENPLLISVAAATASNTNNA-------INAREPTEAANQINKTIANELKAHKLNLIY------------------------------- 184
GpgOBP3    YNKYQECNDRVN--------------SR---------SKTLENTMEE------------LMDYVSELDHTVAHTLFSKLK--------------------------------- 88
GmmOBP6    MKKYEVLDANGKLVKSVALEHAKKFTNSDENKLKIAGTIIDMTSAMD-------TVGDTTEAAEQYSETFKKQADTYGITLEI------------------------------ 145
GmmOBP20   LKEPNVRFAQMR--CAEKYPDARPFPN-------YPDTPANHTYVY--------------LFYKLGLIDLRSRDLDVVYLIKI------------------------------ 88
GpgOBP2    FKFENLFDGNKE---LTDATNKVLTDS--------WQDVWNELGKD--------FNQGFADALYLILTPVADSISYDDYF--------------------------------- 99
GtOBP2     LRNLRKSGFMFFYGKK-STGSVCLKQT-------CHDQARSHTVSPFGI---DSQASIITFTISRNGSGIIRTNFSKCSSPVSS------------------------------ 121
```

Figure 3.0. Alignment of putative *Glossina pallidipes* (GpXXXX), *Glossina palpalis gambiensis* (GpgXXXX) and *Glossina tachinoides* (GtXXXX) OBPs with *Glossina morsitans morsitans* (GmmXXXX) OBPs. Multiple Sequence alignment was done using ClustalW2 software. The numbers on the right of the alignment indicate amino acid residues in the sequence. The conserved cysteine residues, characteristics of insects OBPs, are shaded red.

```
GpOBP1                  ----------------------------------------IVTFAMTLLLSFGLNNAQKPRRDENYPPPDFLKSFKIIHDVWVEKT----GATEEAIK 54
GmmOBP8                 ----------------------------------MKKYHIYIVTFAITLLMSFGLNNAQKPRRDENYPPPDFLKSFKIIHDVWVEKT----GATEEAIK 61
GpgOBP1                 ------------------------------------------------------------PRRDENYPPPDFLKFFKIIHEVWVEKT----GATEEAIT 35
GpOBP2                  ------------------------------------------------------------------------------------HYFWEEKT----GVTGGANK 16
FBpp0078304(Os-E)       -------------------------------------MVKYPLILLLIGCAAAQEP----RRDGEWPPPAILKLGKHFHDIWAPKT----GVTDEAIK 53
GmmOBP9                 ----------------------------MTLSGKYRFLTVYTMLIVLLSAWTRAQQP----RRDDEYPPPAILKLAKPFHDIWVEQT----GVKEEAIK 63
FBpp0078305(Pbprp3)     ----------------------------MALNGFGRRVSASVLLIALSLLSGALILPPAAAQRDENYPPPGILKMAKPFHDAWVEKT----GVTEAAIK 67
AAEL009449-PA(OBP)      -------------------------------------MNGSVVFVLSAL--VSLSVGDVT-PRRDAEYPPPEFLEAMKPLREIWIKKT----GVTEEAII 56
AAEL013018-PA(OBP)      -------------------------------------MIRFIVFVSSCL--VAVSIADVT-PRRDAEYPPPELLQALKPLRDIWQKKT----GVSDEAIL 56
AGAP003309-PA(OBP17)    ---------------------------------------MKLVTFVFAALLCCSMTLGDTT-PRRDAEYPPPELLEALKPLHDIWLGKT----GVTEEAIK 57
CPIJ007604(OBP/PhBP)    ----------------------------------MAARCAKTLVLFSAVLGIAVVVLADVT-PRRDAEYPPPELLEALKPLHDIWAKKT----GVTDEAII 62
CPIJ007617(OBP)         ---------------------------------------MFTTTVLGLGLLLLLQVGLISCEEPRRDAEYPPPEFLVKMKPMHDEWVAET----GASEDAIK 59
AAEL015499-PA(OBP56a)   -------------------------------------MILLNMAVVLLEVMLTLAADKPIPRRDAEYPPPFVLEISKKPHKMWVAST----GVSEAAIK 58
GpgOBP4                 -------------------------------------MKLFLIILSTTVALVN------------AKFDIRTKDDALK-AHEEWHEEF----NVPDDIYE 46
GmmOBP11                ------------------------------------------------------------------TKDDALK-AHEEWHEEF----QVPDDIYE 24
GtOBP1                  ------------------------------------------------------------------------------------------------
FBpp0078266(OBP83g)     ------------------------------MQSQSLLLIVAAVATFLVAQTT------------AKFLLKDHADAEK-AFEEWREDY----YVPDDIYE 52
AAEL005772-PA(OBP99c)   ---------------------------------------MKVFIAVFALIAVAA------------AEFTVSTTEDLQR-YRTEWVSSL----NIPADYVE 45
CPIJ017326(GOBP99a)     ---------------------------------------MKLFIAIFALIAVAT------------ADFTVKTTDDLQT-YRSEWVSSL----SISDELVA 45
AGAP000278-PA(OBP9)     -------------------------------------MLKFVVALLAFTAVVS------------AEFVVQTREDLLA-YRAEWVKSL----GVSDELVE 46
GmmOBP1                 -------------------------------------MKTTAVILLALFALVS------------ADYKLRNQEDLNK-ARKEWMEAK----KVTPELVE 46
GmmOBP                  -------------------------------------MKTTAVILLALFALVS------------ADYKLRNQEDLNK-ARKEWMEAK----KVTPELVE 46
GmmOBP2                 ---------------------------------------MKTIIVIVFLVTLATVWGHHHHEHHDDDDYVVKTREDLFK-YRDEWSNKL----NVPADLLE 57
GmmOBP21                ------------------------------------------------------------AEDEDWQPKTVADIKS-IRNEWLKEH----PLSNEQIT 33
GmmOBP13                ------------------------------------------------------------------------------TKDDFEKILQSWREDM----QINENDLR 24
GmmOBP15                ------------------------------------------------------------------------------TRETLQNYVKTWVIEE----NISTKDLK 24
GmmOBP4                 -------------------------------------MKSKIIFDTRVRRKVMFRVTLILLAIVTSALFSENRYMEFLADFKHWKRER----GVGRFELD 59
GmmOBP5                 ----------------------------MRFHIILKLMSWMCLMCAIESKNVIDLLEGKMYAPAQYQLKPADNFASSPVNKRQMPT----SDIPKNMQ 66
GmmOBP6                 -------------------------------------MFKLLLVTVLMLGILS----------VEAEIDVQEEIAKFILLANEWREEV----GAKEADIQ 49
GmmOBP22                ------------------------------------------------------------------------DDFFQMSERWMRLE----KVPDRYKA 22
GmmOBP14                ------------------------------------------------------------------ATEEQMRSAANLMRDVWLPKFP---KVSKETAD 30
GmmOBP3                 -------IDQSQYNSSEVLLEKFKNYAYWSHEEIPCFARCIASEKGWFDIDLSRWNKQRLVDELGANMYNYCRFELNRAFKNVWSFAFKGLKCLKQAEMN 93
GmmOBP7                 FKQWSDTYEEFPCFTNCYLNNMFNIYNETQGFNEENVIKRFGRSVYNACKEKLIQGNNSCEIAYNGFHCLINREDDPFILIDNIEDISMEAKRAMKECLH 100
```

```
GpOBP1                EFSDGE-------------------IHED---PALKYYMNLLFHEVNVVDDAGELHFEKLVRMIP-----EPFLEMVKHIIDAESHI-PKGETQLDRAWSWHVFFKQTDPVL----YFLP 126
GmmOBP8               EFSDGE-------------------IHED---PALKYYMNLLFHEVNVVDDAGELHFEKLVRMIP-----EPFLEMVKHIIDAESHI-PKGETQLDRAWSWHVFFKQTDPVL----YFLP 150
GpgOBP1               EFSDGE-------------------IHDD---PALKYYMNLLFHEVDVVDDAGELHFEKLVRMIP-----EPYLKMFQHILDAVSHI-PKGETQLDRAWSWHVFFKQTDPVL----YFL- 123
GpOBP2                KFSDGE-------------------NQED---PALKYYMNLLLH-VKVFDDAGELHFEKLVGMIP-----KPFLEMVKHIIDAESHI-PKGETQLDRAWSWHVFFKQTDP---------- 98
FBpp0078304(Os-E)     EFSDGQ-------------------IHED---EALKYYMNLLFHEFEVVDDNGDVHMEKVLNAIPG----EKLRNIMMEASKGLIH---PEGDTLLHKAWWFHQFWKKADPVH----YFLV 141
GmmOBP9               EFSDGE-------------------IHED---EALKYYMNLLFHEFDVVDDNGDVHLEKLFSRIP-----AALRDLLMEASKGLVH---PEGDTLLHKAWWFHQFWKKADPVH----YFLV 150
FBpp0078305(Pbprp3)   EFSDGE-------------------IHED---EKLKYYMNLFFHEIEVVDDNGDVHLEKLFATVP-----LSMRDKLMEMSKGLVH---PEGDTLLHKAWWFHQFWKKADPKH----YFLP 154
AAEL009449-PA(OBP)    EFSDGK-------------------VHED---ENLKYYMNLLFHEAKVVDDTGHVHLEKLHDALP-----DSMHDIALHMGKRLLY---PEGENLLEKAFWLHKFWKESDPKH----YFLI 143
AAEL013018-PA(OBP)    EFSDGK-------------------VHED---EKLKYYMNLLFHEAKVVDDTGHVHLEKLHDALP-----DSMRDIAMHMGKRLLY---PEGENLLEKAFWLHKFWKESDPKH----YFLI 143
AGAP003309-PA(OBP17)  KFSDEE-------------------IHED---EKLKYYMNLLFHEAKVVDDNGDVHLEKLHDSLP-----SSMHDIAMHMGKRLLY---PEGETLLDKAFWLHKFWKQSDPKH----YFLV 144
CPIJ007604(OBP/PhBP)  EFSDGK-------------------IHED---EKLKYYMNLLFHEAKVVDDNGDVHLEKLHDSLP-----NSMHDIAMHMGKRLLY---PEGENLLEKAFWLHKFWKQADPKH----YFLV 149
CPIJ007617(OBP)       RFSDQE-------------------IHED---DNLKYYMNLLFHKAGVVNDKGEFHYVKIQDFLP-----ESMHLITLNWFKRLLY---PQGENLLEKAFWLNKFWKERDPVH----YFLP 146
AAEL015499-PA(OBP56a)RFSDED-------------------IFEDD--EKLKYYMQLLFEKLRYTDDKGELHLGKVMDSVP-----EEYEDIALKMGSKLLK---PKGKTQLERAFWYHKFWKTSDPVVSICDYVFL 150
GpgOBP4               QYLDYQ-------------------FPEH---KLTNYYVKLWIEKMGIFTENRGFNEKNIIAQYT--FENYKNLESVRHGLEKLIDHN-EWETDVLTWANRVFSFWLKVNRHVVR------ 136
GmmOBP11              QYLDYQ-------------------FPEH---KLTNYYVKLWVEKMGIFTENRGFNEKNIVAQYT--YENFKNLESVRHGLEKLIDHN-EWETDVLTWANRVFSFWLKVNRHVVRKMFT-- 118
GtOBP1                ------------------------------------EKLKVKLWLEKMGIITENCGFNEKNVKVPYS--LENFQKLESVRQGPEKLLDHN-EWETEVLTWGYRVFSFWVKVNCHVVR------ 77
FBpp0078266(OBP83g)   KYLNYE-------------------FPAH---RRTSFVKLFLEKLELFSEKKGFDERAMIAQFT--SKSSKDLSTVQHGLEKLIDHN-EAESDVLTWANRVFSFWLPINRHVVRKVFA-- 146
AAEL005772-PA(OBP99c)KFKKWE-------------------FPED---DTTMYYIKLVFNKMQLFDDTEGPLVDNLVHQLA--H-GRD-AEEVRTEVLKLVDKN-TD-NNAHWAFRGFKFFQKNNLSLIKASIKKD 138
CPIJ017326(GOBP99a)   KYRKWD-------------------FPED---DTTQYYIKLIFNKMELFDDNNGPIVDNLVLQLA--H-GRD-ADEVRAEILKLVDKN-TD-DNSHWAFRGFKFFQTNNLQLIKASIKKD 138
AGAP000278-PA(OBP9)   KYKSWN-------------------FPED---DTTQYYIKLIFNKMQLFDDTNGPIVDNLVVQLA--H-GRD-ANEVREEIVKLAGSN-TD-GNVHWAFRGFQFFQKNNLSLIKASVKKD 139
GmmOBP1               KYKKFD-------------------FPDD---EITRYYIELIFDKFQLFDSQTGFKNDNLIAQLG--Q-SKDNKDEVKADIEKLADKN-TEKSDSLTWAFRGFKFFISKNLPLVMESLKKN 141
GmmOBP                KYKKFD-------------------FPDD---EITRYYIELIFDKFQLFDSQTGFKNDNLIAQLG--Q-SKDNKDEVKADIEKLADKN-TEKSDSLTWAFRGFKFFISKNLPLVMESLKKN 141
GmmOBP2               KYKKWQ-------------------YPDD---EVTKYYMKLMFEHFGFFNEKQGFDVHKIHKQLMGAHGTVDHSDETHEKIAKLADKK-PEDTDPLAWAYRGGVFFINSNLQLVKSSVN-- 153
GmmOBP21              KMKNFE-------------------FPDE---EEVRQYLLLTALKMEVFCAHQGYHPNRIAKQFK----MDMNEEEVLEIAEKLHDSN-PDNSSVDVWAFRGHKFMMSSAIGDKVKAYIKK 127
GmmOBP13              TLSASP-------------------NDVS---EGVKYYMKLVMEKQGHFKN-GALLEEAVIKSLESSPADHNDQNQMSAIVKEFKK---EIGSNEFETAFKVSMFLREHKVDFEI------ 113
GmmOBP15              LFMAWN-------------------FSNIS--NEGKLFFSLFHEKIGLTIN-GVLQKKIAFGHLKR----IFDRETAEFVLGEFVN---LVGKDKFETAYQFEKLFNIEYNRLAK----- 111
GmmOBP4               RLRVGN-------------------LAYPS--YEAKLFLGLYERTGILKN-GVLQNDVLKKNVG----YIANRVLLDEVLPPLYA---VSGTNKLDIAFELKKFKNVGFDKVWITVPWE 151
GmmOBP5               QFQDTLNEAKFKCARAMRLDSNKLLMYEDQPSLREKLLMAILKRMKLMDSDYKLSVPTISHIAGMISDENPLLISVAAATASNCNNA-INAREPLEAANQINKLIANELKAHKLNLIY-- 184
GmmOBP6               DLIHKH-------------------PSAG--QEGKLLRALMKKYEVLDANGKLVKSVALEHAKKFTNSDENKLKIAGTIIDMCSAM-DTVGDTLEAAEQYSEFKKQADTYGITLEI-- 145
GmmOBP22              QFTEFQ-------------------FPND---PIVHKYILLVNRELQIWDNNQGFDIEKIYQQYKG----RANEEVVLPIISQLNQD--AKQRNYELWCYKAFLILDTQVGEWFKEDVRR 115
GmmOBP14              GIRNGN-------------------LSDN---KDAKLYINLVMEMMRTMKKGKFLYEGALKQVDLLMP--DSYKEEYRPGLAKLKDSA-NGIKNNLDAAYAVLSFLRGEITQFVFP----- 121
GmmOBP3               VIITHNN-LLECVKEKSISMDQLLEYYHFPQLEHIPLLFKLFADKSHLYTVNYEWNVLNWLKAFG-----PIRNENADISICRVNANE-REKMDILAIMYEEYNFWERLNYNTDGISVTYK 207
GmmOBP7               KFNTDEWQYLS--------------DYVRFPVQEPIPLYTRLFVYKMQLYNHRLRSWNIAAMQRLLG-------VPAEHANIENLLSLSKRRNNNMLAWIYKEMTLFSLSQ---------- 190
```

Figure 3.1. Multiple Sequence Alignment of putative *Glossina* OBPs with Dipteran insect (*Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*) OBPs downloaded from Ensembl and Vector Base. The alignment was done using ClustalW2 software. The conserved cysteine residues are shaded red and the numbers on the right of the alignment indicate amino acid residues in the sequence. GpXXXX - *Glossina pallidipes* OBPs; GpgXXXX - *Glossina palpalis gambiensis* OBPs; GtXXXX - *Glossina tachinoides* OBPs; GmmXXXX -*Glossina morsitans morsitans* OBPs; FBppXXXX - *Drosophila melanogaster* OBPs; AGAPXXXX - *Anopheles gambiae* OBPs; AAELXXXX - *Aedes aegypti* OBPs; CPIJXXXX - *Culex pipens quinquefasciatus* OBPs.

```
GmmCSP4                -MKSTF--CCLALLVVY----------LSVIVTAQKSYTNKFDGVDVDSVLSNERILTNYIKCLME--KGPCT-PEGRELKELLPDALISDCTKCTDVQKKNSQKVINYLRANRPGEWKLLLNKYDPSG 113
FBpp0072311(Pherokine)-MKASL--ALVFCVCVG----------LAAAAP-EKTYTNKYDSVNVDEVLGNNRVLGNYLKCLMD--KGPCT-AEGRELKRLLPDALHSDCSKCTEVQRKNSQKVINYLRANKAGEWKLLLNKYDPQG 112
GmmCSP2                -MLRFFGICIVTAIIAWDSVRSLPHPPATTAAPFKQSYDNKFDNVDLDEILGQERLLKNYVKCLEG--TGPCT-PDGKMLKETIPDAMATDCAKCTPKQKYGSEKVTHFLIDNRPEDWERLEKIYDPAG 125
GtCSP                  ----LTIAVAVVLMTIA----------VIA--EEQYTTKFDNIDVDEILASDRLFDNYFKCLVD--EGKCT-PEGRELKKTLPDALETACTKCNDKQKATVDKVIRFLTEKKPEQWKALQAKYDPSG 108
GmmCSP3(EBSP-3)        -MKFLTIVVAVVLMTVA----------VIA--EEQYTTKFDNIDVDEILASDRLFDNYFKCLVD--EGKCT-PEGRELKKSLPDALETACAKCNDKQKATVDKVIRFLTEKKPDQWKALQAKYDPAG 111
GmmCSP1                -MKYLTIVAVIATLSAV----------VVMGAEEKYTTKYDDVDVDEVLKSDRLFKNYYNCLID--QGKCT-PDARELKKSLPDALQTECSKCSEKQKKTSEKVIKHLMDHKPEEWKVLQTKYDPEG 113
AAEL002022-PA(OS-D)    ----MKIMIVIAFALLA----------VASAQEEQYTTKYDNIDVEEILKSDRLFNNYFKCLMD--EGPCT-PDGNELKRILPEALQTNCAKCSESQRAGAIKVINYMIENRAEQWKALQEKYDPEN 110
AGAP008052-PA(A10)     ----MKLFVAIAFALLA----------LAAAQ-EQYTTKYDGIDLDEILKSDRLFNNYFKCLMD--EGRCT-PDGNELKKILPEALQTNCEKCSEKQRSGAIKVINYVIENRKEQWDALQKKYDPEN 109
CPIJ002618(CSP1)       ----MKLLIV--FALVA----------LVAAQDSTYTNKYDNIDVDEILKSDRLFKNYFNCLID--QGPCT-PDATELKQSLPDALENNCSKCTPKQKEVGYKVVGWLINNRPEEWNVLRAKYDPEN 108
BmoriCSP4              ------MKVLIVLSCVL----------VAVLADDKYTDKYDKINLQEILENKRLLESYMDCVLG--KGKCT-PEGKELKDHLQEALETGCEKCTEAQEKGAETSIDYLIKNELEIWKELTAHFDPDG 108
PxuthusCSP8b           ----MASKSLIILLCVV----------ASVWCT--YTNKYDNINYEEVVSNEKLLNKYADCILD--RGRCT-AEAKELKEHLKEAIENGCKECTEKQKEGTNYVIRYLIENKRNLWEELCAKFDPDG 108
TcastaneumCSP9         ----MTAIVFLLALACL----------KTYVSSQEYLVP-QNIDVDEILKNDRLTRNYLDCVLG--KGKCT-PEGEELKKDIPEALQNGCAKCNEKHKEGVRKVIHHLIENKPNWWQELESKFDPQG 109
HvirescensCSP1         ---------------MA----------LARPDGAAYTDKYDNVDLDEILSNRRLLVPYVKCILD--QGKCA-PDAKELKEHIIEALENECGKCTEAQKKGTRRVIGHLINNEADYWNELTAKFDPEK 99
MbrassicaeCSP          -MKSCIVLCVLSVAVMA----------LARPEEAHYTDRYDSVDLDEILGNRRLMVPYIKCILD--QGKCA-PDAKELKEHIREALENECGKCTETQKNGTRRVIGHLINHEDAYWKELTAKYDPQS 113
PxuthusCSP4a           -MKSVVLLCLFVVLATA----------VAGP-AGQYTDRYDSINIDEILGNRRLLIPYVHCLLD--QGKCT-PEGKELKSHIKEALENNCAKCTPPQPTRSRQVIGHLINHEPTFWNELCQKYDPTS 112
PxuthusCSP4b           -----------MAIATA----------MPSP-SETYTDRFDNIDLDEIIGNRRLLVPYIHCVLD--KGKCT-PDGKELKSHIAEAIENDCAKCTEVQRKGTRKVLGHLINNEQEFWDELTARYDPEH 102
PxuthusCSP4c           -MK--VLVCFLLALVII----------AQARSNRKYTSRYDNIDLDDVLANRRLLVSYIKCLLD--LGKCS-SEGKELKSHINDALEDNCAQCTDVQKSGFRRVIGHLINHEKDYWGKLTAKYDPER 111
SgregariaCSP-sg5       ------------------------------------EEKYTTKYDNVNLDEILANDRLLNKYVQCLLEDDESNCT-ADGKELKKDIPDALSNECAKCNEKQKEGTKKVLKHLINHKPDVWAQLKAKYDPDG 94
SgregariaCSP-sg4       ------------------------------------EEKYTTKYDNVNLDEILANDRLLNKYVQCLLEDDESNCT-ADGKELKSVIPDALSNECAKCNEKQKEGTKKVLKHLINHKPDVWAQLKAKYDPDG 94
LmigratoriaCSP         ------MQKCTLALLLA----------CLVAAAAAYTTKYDNIDLDDVLHNDRLLKKYHECLLSDSDASCT-PDGKELKAAIPDALTNECAQCNEKQKAGAEKVIRFLIKEKPDLWTPLEKKYDPTG 110
AmelliferaCSP1         -MRHNYIVILILSLLTW----------TYAEELYSDKYDYVNIDEILANDRLRNQYYDCFID--AGSCLTPDSVFFKSHITEAFQTQCKKCTEIQKQNLDKLAEWFTTNEPEKWNHFVEIMIK-- 110
AngambiaeCSP5          -MRKVWLLASVVLAFLDFVK---------SQEVARTLYSTRYDNLDIDTILASNRLVTNYVDCLLS--RKPCP-PEGKDLKRILPEALRTKCARCSPIQKENALKIITRLYYDYPDQYRALRERWDPSG 116
GmmCSP5            MKRIMISRVCVTCVIYLLMA----------AVVVEC-------DEKNINKLLNNQVIVSRQIMCVLE--KSPCD-QLGRQLKAALPEVIVRNCRNCSPQQAQNAQKLTTFLQTKYPDVWAMLLRKYKT-- 108
AngambiaeCSP4      MSSKALPNLFMLSAAVIVVMA----------ALVIVGPQPAAANDSQNINRLLNNQVIVSRQIMCVLE--KSPCD-QLGRQLKAALPEVIQRNCRNCSPQQAQNAQKLTNFLQTRYPEVWAMLIRKYGAV- 117
AceranaCSP          MASAIKALLIVCALLVYTVT----------AETEEGQSGRSRVSDEQLNMALSDQRYLRRQLKCALG--EAPCD-PVGRRLKSLAPLVLRGACPQCSPEETRQIKKVLSHIQRTYPKEWSMIVQQYAGVS 117
```

Figure 3.2. Multiple Sequence Alignment of *Glossina tachinoides* chemosensory protein (GtCSP) with selected homologous sequences download from GenBank. Th numbers on the right of the alignment indicate amino acid residues in the sequence. The conserved cysteine residues are shaded green and conserved amino acids ar shaded blue. Aromatic amino acid residues are shaded yellow. GmmXXXX - *Glossina morsitans morsitans* CSPs; FBpp0072311 - *Drosophila melanogaster* Pherokine AGAP008052 - *Anopheles gambiae* antennal protein 10/Olfactory specific-D; AAEL002022 - *Aedes aegypti* Olfactory specific-D/Pheromone binding protei CPIJ002618 - *Culex pipens quinquefasciatus* Chemosensory protein 1; BmoriCSP4 – *Bombyx mori* Chemosensory protein 4; PxuthusXXXXX - *Papilio xuthu* Chemosensory proteins; TcastaneumCSP9 – *Tribolium castaneum* Chemosensory protein 9; HvirescensCSP1 - *Heliothis virescens* Chemosensory protein 1 MbrassicaeCSP - *Mamestra brassicae* - Chemosensory protein; SgregariaCSP-sg5 - *Schistocerca gregaria* Chemosensory protein; LmigratoriaCSP - *Locusta migratori* Chemosensory protein; AmelliferaCSP1 - Apis mellifera Chemosensory protein 1; AngambiaeCSP5 - *Anopheles gambiae* Chemosensory protein 5; AceranaCSP - *Api cerana cerana* Chemosensory protein; EBSP-3 - Ejaculatory bulb Serum protein 3

There is good bootstrap support of 97% between GtCSP and GmmCSP3, identified to be an ejaculatory bulb Serum protein 3 (EBSP-3) (Figure 3.5). High bootstrap values also support other othologs which includes *An. gambiae* A10/OS-D and *Ae. aegypti* OS-D/PhBP (92%), *S. gregaria* CSP-sg4 and 5 (93%), *M. brassicae* CSP and *H. virescens* CSP1 (91%), *An. gambiae* CSP4 and GmmCSP5 (97%) with *Apis cerana cerana* CSP (92%). However, there was moderate support for the other possible orthologs e.g. *C. quinquefasciatus* CSP1 and GmmCSP1 (50%), *L. migratoria* CSP and *S. gregaria* CSP-sg4/5 (58%), *P. xuthus* CSP8b and *B. mori* CSP4 (52%). Clustering of the CSPs indicates that Dipteran (*Culex*, *Glossina*, *Anopheles* and *Aedes*), Orthoptera (*L. migratoria* and *S. gregaria*), Lepidoptera (*P. xuthus, M. brassicae, H. virescens* and *B. mori*) and Coleoptera (*T. castaneum*) could have diverged separately from *Glossina* (GmmCSP2, GmmCSP4 and GmmCSP5), *A. mellifera* CSP1, *An. gambiae* CSP4 and CSP5, *A. c. cerana* CSP and *D. melanogaster* Pherokine.

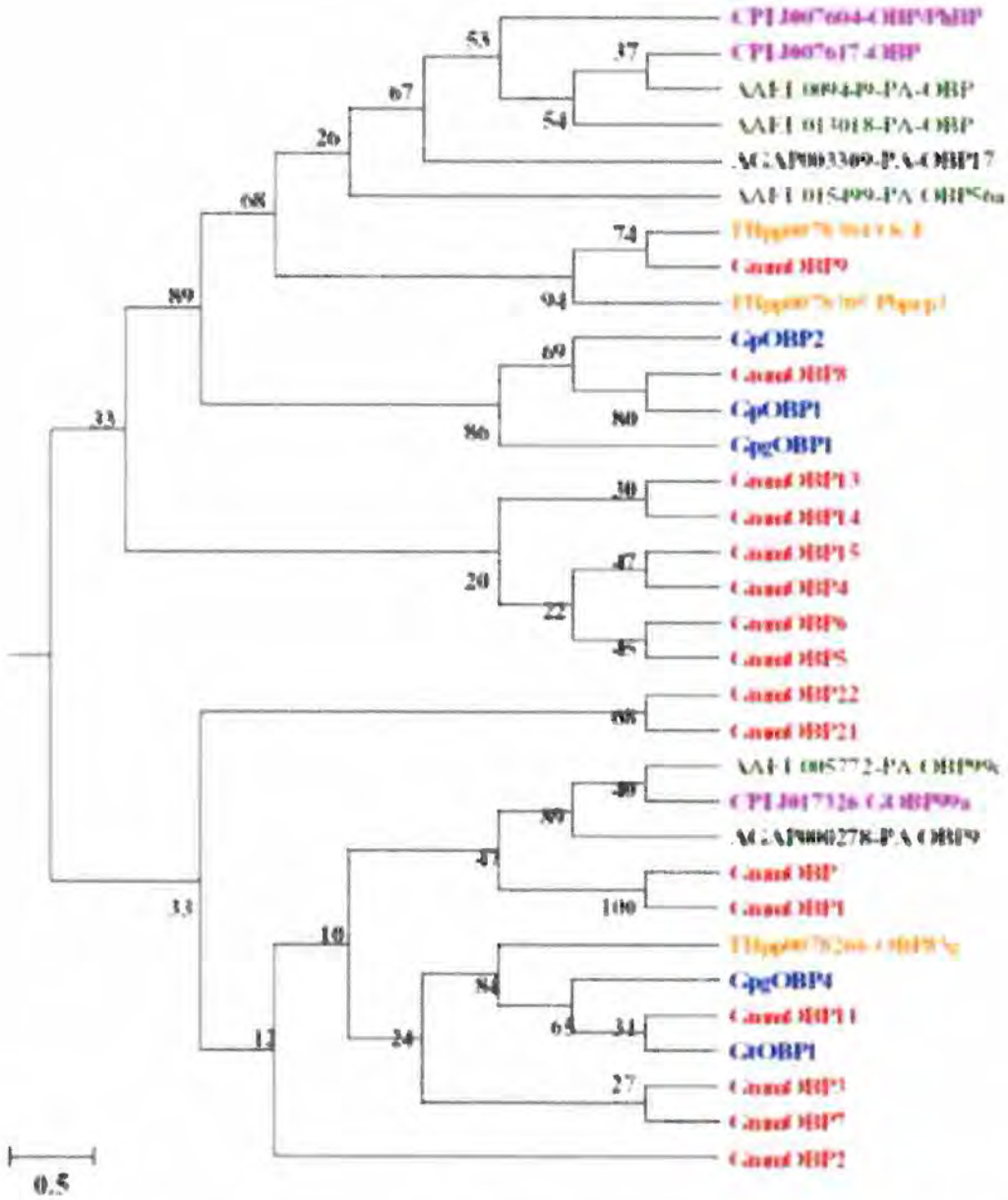

Figure 3.4 Phylogenetic comparison of the OBP family members in Diptera including *Glossina morsitans morsitans* (GmmXXXX) OBPs (red); *Glossina pallidipes* (GpXXXX), *Glossina palpalis gambiensis* (GpgXXXX) and *Glossina tachinoides* (GtXXXX) OBPs (blue); *Drosophila melanogaster* (FBppXXXX) OBPs (orange); *Anopheles gambiae* (AGAPXXXX) OBPs (black); *Aedes aegypti* (AAELXXXX) OBPs (green) and *Culex quinquefasciatus* (CPIJXXXX) OBPs (pink). The tree was constructed from an alignment shown in Figure 3.2 using maximum likelihood model with PHYML program. Bootstrap support (100 replications) is indicated on the branches.

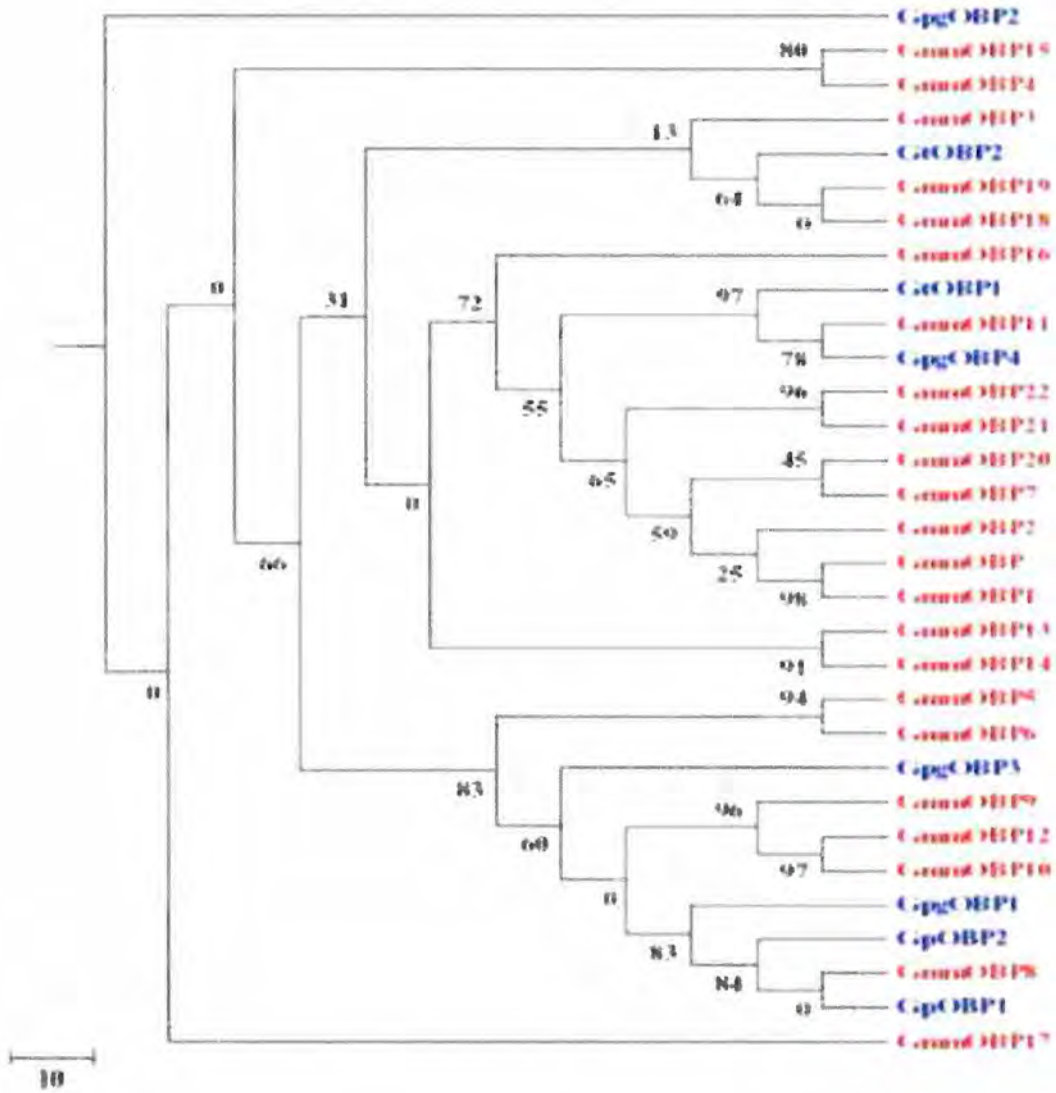

Figure 3.3 Phylogenetic relationships of identified putative OBPs (blue) in *Glossina pallidipes* - GpXXXX, *Glossina palpalis gambiensis* - GpgXXXX and *Glossina tachinoides* - GtXXXX with *Glossina morsitans morsitans* (GmmXXXX) OBPs (red). The tree was constructed from an alignment shown in Figure 3.1 using maximum likelihood model with PHYML program. Numbers on branches show values of 100 times replication bootstrap analysis.

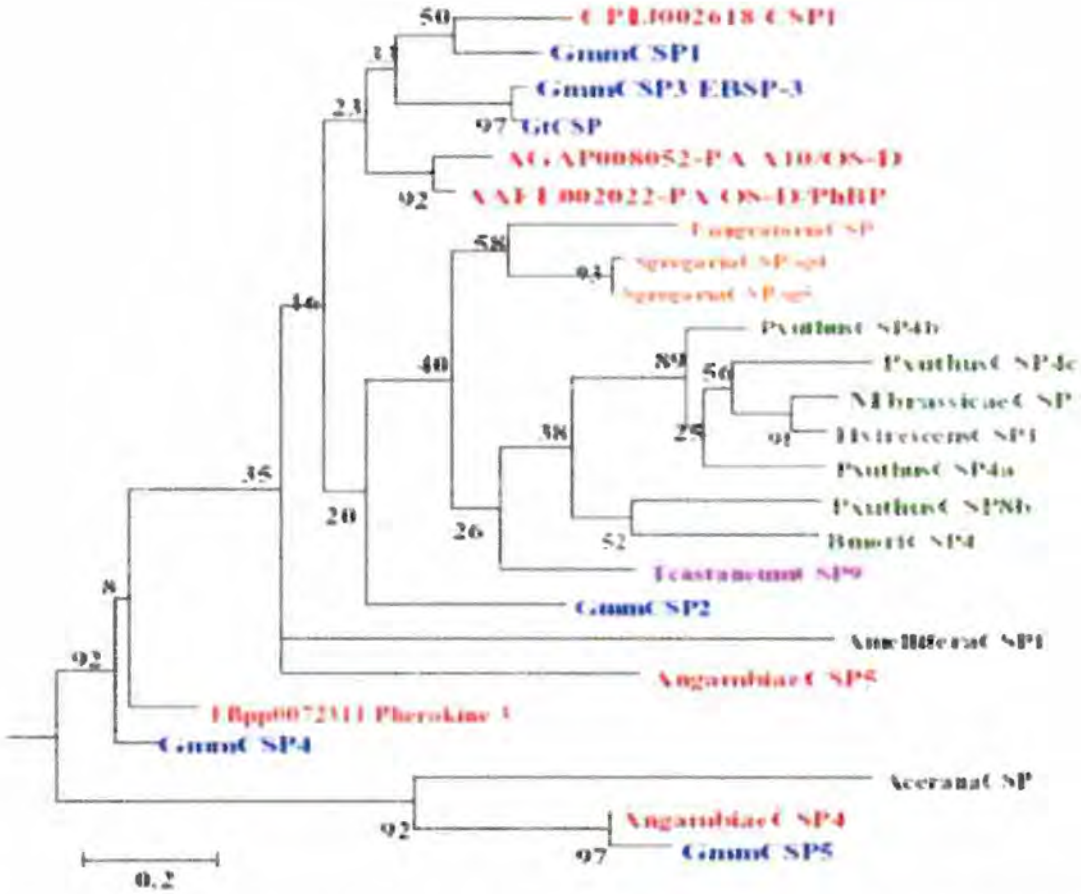

Figure 3.5. Phylogenetic relationships of *Glossina tachinoides* (GtCSP) and *Glossina morsitans morsitans* (GmmXXXX) chemosensory proteins (blue) with other insect CSPs downloaded from GenBank including *Drosophila melanogaster* Pherokine (FBpp0072311) (red), *Anopheles gambiae* antennal protein 10/Olfactory specific-D (AGAP008052) (red), *Anopheles gambiae* Chemosensory protein 5 (AngambiaeCSP5) (red), *Aedes aegypti* Olfactory specific-D/Pheromone binding protein (AAEL002022) (red), *Culex quinquefasciatus* Chemosensory protein 1 (CPIJ002618) (red), Lepidopteran (green) *Bombyx mori* Chemosensory protein 4 (BmoriCSP4), *Papilio xuthus* Chemosensory proteins (PxuthusXXXXX); *Heliothis virescens* Chemosensory protein 1 (HvirescensCSP1) and *Mamestra brassicae* - Chemosensory protein (MbrassicaeCSP); Orthopteran (orange) *Schistocerca gregaria* Chemosensory protein (SgregariaCSP- sg4 and sg5), *Locusta migratoria* Chemosensory protein (LmigratoriaCSP); Hymenopteran (black) *Apis mellifera* Chemosensory protein 1 (AmelliferaCSP1) and *Apis cerana cerana* Chemosensory protein (AceranaCSP); Coleopteran (pink) *Tribolium castaneum* Chemosensory protein 9 (TcastaneumCSP9). The tree was constructed from an alignment shown in Figure 3 using maximum likelihood model with PHYML program. Numbers on branches show values of 100 times replication bootstrap analysis. EBSP-3 - Ejaculatory bulb Serum protein 3

3.4 Discussion

A total of ten putative olfactory proteins (9 OBPs and 1 CSP) were identified from *Glossina* species, *G. pallidipes*, *G. p. gambiensis* and *G. tachinoides,* by comparison of head and antennae EST libraries to *G. m. morsitans*, *D. melanogaster*, *An. gambiae, Ae. aegypti* and *C. quinquefasciatus* protein databases. Analysis of the homologs revealed a diverse set of genes implicated in different processes indicating the complexity of the *Glossina* antennae and head. The putative OBPs identified in this work is small comparable to the number identified in *G. m. morsitans* (Liu *et al.,* 2010) and *Apis mellifera* (Foret and Maleszka, 2006). Most studies have a repertoire of more than twice as many OBP genes (Krieger *et al.,* 1993; Kim and Smith, 2001; Xu *et al.,* 2003;). This may be due to the short truncated sequences but it is most likely due to the small scale of the EST project, compared to the various genomic studies on which some of the gene discovery efforts are based (Ishida *et al.,* 2004; Zhou *et al.,* 2008; Gong *et al.,* 2009; Pelletier and Leal, 2009). It is very probable that more OBP and CSP genes are present in the unsequenced fraction of the ESTs. Completion of the *G. m. morsitans* genome and other species of tsetse flies may reveal more olfactory genes.

The characteristic features of the predicted OBPs (small size of about 120 amino acids, conserved 6 cysteines, signal peptide) identified in *G. pallidipes*, *G. p. gambiensis* and *G. tachinoides*, suggest that the OBPs are conserved across different Insect orders. Presence of a signal peptide confirms that the OBPs are secretory, while the 6 cysteines ensures formation of 3 disulfide bridges that enables the protein to fold into its 3-dimenstional tertiary structure necessary for binding of odors (Sandler *et al.,* 2000; Kruse *et al.,* 2003;

Wogulis *et al.,* 2006). The Insect OBP gene family is very large, quite diverse in sequence and the different homologs identified from Dipteran species show that tsetse flies could also be expressing different olfactory proteins which may play significant roles to adjust to specific life cycle stages, adaptation to different habitats and response to various odors emitted by hosts.

The PBPs, predominantly expressed in males (Raming *et al.,* 1989) may be important in response to female tsetse fly's surface cuticular paraffins or synthetic 15,19-dimethyltritriacontane that has been reported to evoke response in male *G. m. morsitans* to elicit a mating response (Huyton *et al.,* 2008a,b). The OBPs/or GOBPs, expressed in both male and female Insects (Breer *et al.,* 1990; Vogt *et al.,* 1991a) may participate in odor induced host location in tsetse flies (Gikonyo *et al.,* 2003; Saini and Hassanali, 2007; Omolo *et al.,* 2009). Both PBPs and OBPs are most probably involved in binding and transport of hydrophobic pheromone/odors and delivery to odorant receptors (Ors) located in the dendritic membrane of olfactory receptor neurons (ORNs) as proposed in Insects olfaction models (Jacquin-Joly and Merlin, 2004; Rutzler and Zwiebel, 2005). Although there are no experimental studies done to determine functions of *Glossina* OBPs, the non-olfactory homologous genes reported in this work (ubiquinol-cytochrome-*c* reductase, mitochondrial fragment and TsaII) may pinpoint that some *Glossina* OBPs could be involved in functions not related to olfaction. Salivary odorant binding proteins have been identified in salivary transcriptome of some insects (Ribeiro *et al.,* 2004; Kalume *et al.,* 2005) and their tertiary structure predicted (Paramasivan *et al.,* 2007). They belong to the D7 family of proteins and are thought to play a role either in endogenous or housekeeping

function (Calvo *et al.*, 2010). It will be interesting to determine the role of salivary OBP reported in *Glossina*.

Phylogenetic analysis of the *Glossina* OBPs suggests a diverse gene family that has undergone rapid evolution by duplication. This could be as a result of various odors produced by different hosts located in different ecological zones (Gikonyo *et al.*, 2002; Omolo *et al.*, 2009) hence the existence of many OBPs. Sequence similarity tree of the diptera Insects show that *Glossina* and *Drosophila* OBP genes evolved from a common ancestor independent from mosquitoes OBPs. The evolution could have taken place over 260 million years ago by speciation event (Liu *et al.*, 2010).

The high sequence identities of GtCSP reported in this study to its Dipteran species indicate presence of chemosensory protein (CSP) in tsetse flies. Chemosensory proteins are small soluble proteins of about 120 amino acids that are unrelated to OBPs. They are expressed widely in various tissues and are postulated to be involved in both olfactory and non-olfactory functions (Campanacci *et al.*, 2003). The strong sequence identity (95%) to *G. m. morsitans* homolog (EBSP-3) suggests that GtCSP could be involved in functions related to mating. Interestingly, the *Drosophila* Pherokine-3 homolog is strongly expressed during metamorphosis (Sabatier *et al.*, 2003) and may implicate GtCSP in tissue remodeling. A similar molecule, pherokine-2 (Phk-2), is related to olfactory specific-D (OS-D)/Antennal protein 10 (A10) (Phk-1). Olfactory specific-D (OS-D) is the first CSP gene identified in antennae of *D. melanogaster* by substractive hybridization experiments (McKenna *et al.*, 1994; Pikielny *et al.*, 1994). Other homologs of this protein family

include CSP1 from *C. p. quinquefasciatus*, OS-D/PhBP from *Ae. aegypti* and A10/PBP/OS-D from *An. gambiae*. The different CSP family members reported suggest broad functions GtCSP could be implicated to perform.

Unlike OBPs, CSPs are highly conserved multigene family with about 50% identity residues even in Insects of different orders (Gong *et al.*, 2007). This is supported by strong bootstrap support and phylogenetic analysis indicates that tsetse CSPs are more closely related to other insects CSPs and could have evolved rapidly by speciation events. The CSP reported in this work from *G. tachinoides* (1), *G. m. morsitans* (5) (CSP1-gi|281426841|; CSP2-gi|281426843|; CSP3-gi|281426845|; CSP4-gi|281426847| and CSP5-gi|281426849|) and other Diptera and Hymenoptera genomes (4 in *D. melanogaster* (Zhou *et al.*, 2006), (7 in *An. gambiae* (Zhou *et al.*, 2006) and (6 in *Apis mellifera* Foret *et al.*, 2007), are small in number relative to those identified in Lepidopteran and Orthopteran species (17 in *B. mori* (Gong *et al.*, 2007), (22 in *Locusta migratria* Foret *et al.*, 2007), (11 in *Papilio xuthus* Ozaki *et al.*, 2008). This could mean that CSPs are involved in many functions where they occur in large numbers and the small numbers could indicate that their expression levels are low in *Drosophila*, *Anopheles*, *Apis* and *Glossina*.

CHAPTER 4

4.0 TISSUE LOCALISATION OF ODORANT-BINDING PROTEINS (OBPS) AND CHEMOSENSORY PROTEIN (CSP) IN *GLOSSINA PALLIDIPES* (DIPTERA: GLOSSINIDAE)

4.1 Introduction

Tsetse flies use olfactory and visual cues to locate preferred hosts, attract potential mates, find larviposition sites and explore the environment. Olfactory sensillia, located in different olfactory and gustatory tissues, detect volatile hydrophobic odorants (semiochemicals) that enter the sensillium lymph through small pores that perforate its surface. Soluble proteins, chemosensory proteins (CSPs) and odorant binding proteins (OBPs), are found in high concentration within the lymph (Pelosi *et al.*, 2006), and are thought to bind hydrophobic odorants entering the sensillium fluid, transport them to Odorant receptors (Ors) located in the olfactory receptor neurons (ORNs) to evoke a response through either a ligand-gated or ion channel G-protein Coupled Receptor (GPCR) pathways (Wicher *et al.*, 2008; Pellegrino and Nakagawa, 2009).

Odorant binding proteins have been identified in many insects olfactory and non-olfactory tissues including: *Antheraea polyphemus* (Raming *et al.*, 1989), *Antheraea pernyi* (Breer *et al.*, 1990), *Heliothis virescens* (Krieger *et al.*, 1993), *Bombyx mori* (Gong *et al.*, 2009), *Drosophila melanogaster* (McKenna *et al.*, 1994; Pikielny *et al.*, 1994; Kim and Smith, 2001), *Apis mellifera* (Foret and Maleszka, 2006), *Anopheles gambiae* (Biessmann *et al.*, 2002; Xu *et al.*, 2003), *Culex pipens quinquefasciatus* (Pelletier and Leal, 2009), *Aedes aegypti* (Zhou *et al.*, 2008) and *G. m. morsitans* (Liu *et al.*, 2010). Examples of insect

orders where CSPs have been identified are Lepidoptera (Jacquin-Joly *et al.*, 2001; Picimbon *et al.*, 2001; Gong *et al.*, 2007), Orthoptera (Angeli *et al.*, 1999; Ban *et al.*, 2003), Hymenoptera (Briand *et al.*, 2002; Ishida *et al.*, 2002), Blattoidea (Kitabayashi *et al.*, 1998; Picimbon and Leal, 1999), and Hemiptera (Jacobs *et al.*, 2005).

In *Glossina morsitans morsitans*, OBPs have been identified and reported to be highly expressed in female than the male antennae (Liu *et al.*, 2010). Identification of OBPs and CSPs, coupled with the highly conserved nature of these proteins across different insect orders, suggests that these proteins could play an important role in recognizing and delivering hydrophobic odorants to Ors (Jacquin-Joly *et al.*, 2001; Biessmann *et al.*, 2002; Pelosi *et al.*, 2006). Furthermore, these proteins may be encapsulins as they have also been reported in both chemosensory as well as non-chemosensory tissues (Pelletier and Leal, 2009). This chapter reports detection of functional homologs of putative OBPs identified from *G. pallidipes*, *G. tachinoides*, *G. p. gambiensis* and *G. morsitans morsitans* in male and female *G. pallidipes* tissues (Antennae, head, thorax and abdomen).

4.2 Materials and Methods

4.2.1 Tsetse flies

Adult *G. pallidipes* were collected in Nguruman, Kajiado District, Kenya (1° 50'S; 36° 05'E) (between 6th and 11th February 2008), using NGU traps baited with acetone and cow urine as described earlier (Chapter 2, section 2.3). *G. pallidipes* males and females were sorted, immediately preserved in RNA*later* (Ambion Inc, Austin, TX) and transferred to the laboratory for analysis. Antennae, heads (without antennae), thorax and abdomen were isolated from *G. pallidipes* (100 males and 100 females), using the standard methods of

Pollock (1982), stored at -80°C and frozen in liquid nitrogen prior to use.

4.2.2 Screening for the putative OBP transcripts in *G. pallidipes* tissues

Total RNA was extracted from antennae, head (minus antennae), thorax and abdomen of *G. pallidipes* (100 males and 100 females flies), using RNAgents Total RNA Isolation System (Promega, Madison, WI), and treated with RNase-free DnaseI (Fermentas, Glen Burnie, MD). Integrity of extracted RNA was validated by electrophoresis in 1.0% agarose (Sigma – Aldrich Chemie, Gmbh) RNA denaturing gel in 0.3% agarose/ethidium bromide (0.1 ug/ml) gel run at 60V for two hours. The yield and quality of RNA was determined spectrophotometrically (Sambrook and Russell, 2001). The genomic DNA was extracted from the legs of the tsetse fly by conventional phenol-chloroform DNA extraction method and digested with RNAase A (Sambrook and Russell, 2001). The RNA was used for cDNA synthesis and DNA extracted was used for detection of gDNA contamination in each transcription product. Reverse transcriptions were conducted using RevertAid^{TM} H M-MuLV reverse transcriptase (Fermentas, Glen Burnie, MD). Oligod$(T)_{18}$ was used as primer in the first strand cDNA synthesis in the reaction mix that consisted of 5X reaction buffer, 5μg total RNA, 2.5 μM oligo-dT, 20U/μl RNAse inhibitor, 500 μM dNTPs, 200 U/μl M-MuLV reverse transcriptase and H_20 in a total volume of 20 μl. This was incubated at 65°C for 5 minutes to maximize primer-RNA template binding, reverse-transcribed at 42°C for 60 minutes and heat inactivation at 70°C for 5 minutes. Integrity of each cDNA was validated through PCR amplification of *G. m. morsitans* GAPDH specific primers; forward (5'-TAAAATGGGTGGATGGTGAGAGTC-3') and reverse (5'-CTACGATGAAATTAAGGCAAAAGT-3') (product size 377bp) (Attardo *et al.,* 2006).

Standard PCR reactions were conducted with OBP specific primers as described by Marone *et al.* (2001). Briefly, 1 ul cDNA products were amplified with 0.2 ul Phusion DNA polymerase (0.02u/ul) (New England Biolabs, Ipswich, MA), 4 ul 5X Phusion GC buffer, 0.5 ul DMSO, 0.5 ul dNTPs (10mM) and 13 ul PCR grade water, in the presence of 0.5 ul of each primer (10pmol). Reactions were carried out in 9800 Fast Thermal Cycler (Applied Biosystems, Foster City, CA). Cycle conditions were as follows: initial denaturation at 98°C for 1 minute; 33 cycles of 98°C for 1 minute, 55°C for 1 minutes, 72°C for 1 minutes; final extension at 72°C for 8 minutes. The primers (Table 4.0) were manually designed for *G. m. morsitans* OBPs (Liu *et al.,* 2010), *G. pallidipes, G. tachinoides* and *G. p. gambiensis* OBPs, to span a predicted intron in the gene to exclude contamination of cDNA with genomic DNA.

4.2.3 Sequencing and analysis of putative OBP transcripts in *G. pallidipes* tissues

The amplicons (5 ul) were run and visualized in 1% agarose gel (stained with ethidium bromide (0.5μg/ml)), gel purified using QIAquick gel extraction kit (Qiagen Madison, WI) and directly sequenced using ABI 3730 DNA sequencing system (Applied Biosystems, Foster City, CA). Consensus sequences were generated by pairwise alignment of the forward and reverse sequences using BioEdit software, (Hall, 1999) and function predicted by searches against non-redundant GenBank protein (Wheeler *et al.,* 2010) and GO (Ashburner *et al.,* 2000) databases using blast (Altschul *et al.,* 1997). Similarities between the protein sequences and selected homologs were estimated by multiple sequence alignments using ClustalW2 software (Larkin *et al.,* 2007). Where necessary, manual adjustment was done to align the conserved cysteines in the alignments.

Table 4.0 Primers of putative *Glossina* OBPs used in RT-PCR

Primer No.	Putative OBP	Forward primer (5' ⟶ 3')	Reverse primer (5' ⟶ 3')
1	Gpacontig265	GGGGAATTCAGTGATGGAGAA	CCCCCCCCCCAAGGGGGGGTT
2	Gpacontig266	GCCTCGAGATGGAAAAGTATCATATT	GCTTCGAACTATTACGGCAAAAAGTAAAG
3	Gthcontig63	GCCTCGAGATGACCATCGCGGTTATA	GCTTCGAATTAAACCTTAATGCCACG
4	Gphcontig184	GCCTCGAGACATACGCACAGAAGCCA	GCTTCGAACTATTACGGCAAAAAGTAAAG
5	Gthcontig151	GCCTCGAGATGGGAATTATAACCGAA	GCTTCGAACTAATTGAACATTTTACG
6	Gmm_cn7403	GCCTCGAGATGAAAACAATTATCGTG	GCTTCGAACTATTAATTGACACTGGACTT
7	Gmm_cn15569	GCCTCGAGATGGAAGCGAAGAAAGTC	GCTTCGAACTATTAGTTCTTTTTCAAACT
8	Gmm_GLAAS20TVB	CCTCGAGATGTATAACTATTGTCGC	GCGAATTCCTAATCAGAAGATGAAAATTG
9	Gmm_cn15567	AAGACTACCGCCGTTATATTG	GAAAGCTAAGATTGGGAAGGT
10	Gmm_cn14707	GCCTCGAGATGTCTGAACGTTGCATG	GCTTCGAACTATTAAACCATTCGCCTACT
11	Gmm_cn15220	GCCTCGAGATGAAACTATTTTTAATA	GCTTCGAACTACTAAGTAAACATTTTACG
12	Gmm_cn13435	GCCTCGAGATGAAGGAATTAATTTTG	GCTTCGAACTATTAAGCATTTTTCGCTGC
13	Gmm_cn15565	GCGAGCTCGGGGAGAGTCCGAACCCT	GCTTCGAACTATTAGTTCTTTTTCAAACT
14	Gmm_cn14014	GCCTCGAGATGATGAAATATTTCGAA	GCAAGCTTCTATCATTCTTTGACTTCGGG
15	Gmm_GLAC953TV	GCCTCGAGATGGAATTTCTTGCGGAT	GCAAGCTTCTATTAAGCAAAAGCAACTCT
16	Gmm_GLAER58TV	GCCTCGAGATGAAATTAATTACCGTT	GCAAGCTTCTATTAAGCAGTTTCGATATT
17	Gmm_cn15331	GCCTCGAGATGAAATCTTGGATTCTC	GCAAGTAACTTCTAATAATCAAACTCGTT
18	Gmm_cn13968	GCCTCGAGATGAGATTTCATATCATA	GCAAGCTTCTATTAATAAATCAAATTCAA
19	Gthcontig219	GCCTCGAGATGCGATCGCGCTTGGTC	GCTTCGAACTACGTAATAGTGCGTGG

Glossina pallidipes antennae (Gpa); *Glossina tachinoides* head (Gth); *G. palpalis* head (Gph); *G. morsitans morsitans* (Gmm). cn stands for consensus which represents the cluster sequence of each OBP. Primers designed from putative OBP ESTs have been assigned arbitrary primer numbers 1 to 19.

4.3 Results

4.3.1 Determination of Total RNA Quality

Assessment of total RNA extracted from antennae, head (without antennae), thorax and abdomen on 0.3% agarose gel showed a band smear between 250bp to 2kb for both *G. pallidipes* male and female (Figure 4.0).

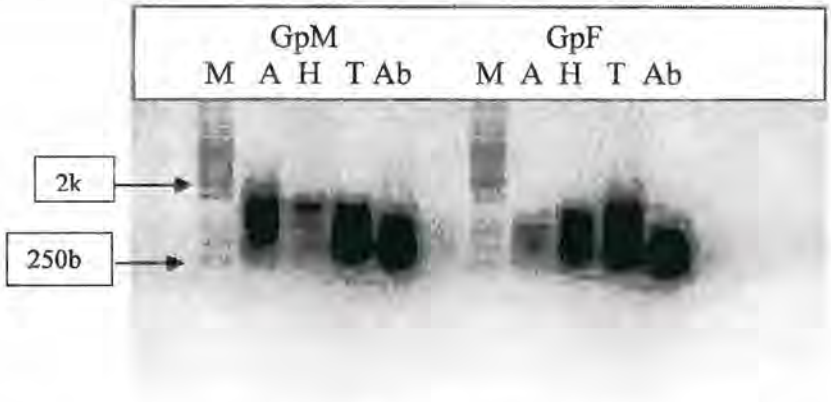

Figure 4.0 Analysis of total RNA isolated from *G. pallidipes* male (GpM) and female (GpF) antennae (A), head (H), thorax (T) and abdomen (Ab). 5ul of total RNA was resolved on a 0.3% agarose gel stained with ethidium bromide. M-1kb ladder.

4.3.2 Determination of First Strand cDNA Integrity

All cDNA fractions from antennae, head (without antennae), thorax and abdomen of *G. pallidipes* Male (GpM) and Female (GpF), amplified with internal control primer, GAPDH, produced a band of comparable intensity and expected product size of 377bp (Figure 4.1).

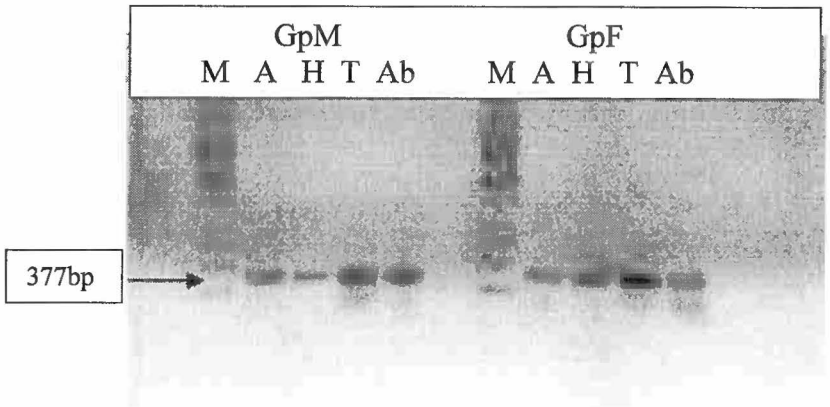

Figure 4.1 Amplification of first strand cDNA synthesized from *G. pallidipes* male (GpM) and female (GpF) antennae (A), head (H), thorax (T) and abdomen (Ab) with GAPDH internal control primer of band size 377bp. 5ul of the PCR product was resolved on a 1% agarose gel stained with ethidium bromide. M-1 kb ladder.

4.3.3 Amplification of First Strand cDNA with designed putative OBP primers

A set of nineteen primers (Table 4.0) were used to amplify *G. pallidipes* Male (GpM) and Female (GpF) antennae (A), head (H), thorax (T) and abdomen (Ab). No amplifications of either GpM or GpF tissues were reported for primers number 1 (Gpacontig265), 5 (Gthcontig151), 9 (Gmm_cn15567), 11 (Gmm_cn15220), 13 (Gmm_cn15565), 15 (Gmm_GLAC953TV), 16 (Gmm_GLAER58TV) and 18 (Gmm_cn13968). The primers, 8 (Gmm_GLAAS20TVB) and 14 (Gmm_cn14014), amplified fragments localised to GpM thorax and abdomen while primers 2 (Gpacontig266) and 4 (Gphcontig184) amplified antennae of both GpF and GpM. In general, the proportion of amplified fragments in antennae, head, thorax and abdomen was similar in both GpF and GpM (Table 4.1).

Table 4.1 Summary of amplified *Glossina pallidipes* body parts

	GpF						GpM				
Primer No.	**Putative OBP**	**A**	**H**	**T**	**Ab**	**Primer No.**	**Putative OBP**	**A**	**H**	**T**	**Ab**
1	Gpacontig265	X	X	X	X	1	Gpacontig265	X	X	X	X
2	Gpacontig266	√	X	X	X	2	Gpacontig266	√	X	X	X
3	Gthcontig63	√	X	√	X	3	Gthcontig63	√	X	√	√
4	Gphcontig184	√	X	X	X	4	Gphcontig184	√	X	X	X
5	Gthcontig151	X	X	X	X	5	Gthcontig151	X	X	X	X
6	Gmmcn7403	√	√	√	√	6	Gmm_cn7403	√	X	√	√
7	Gmmcn15569	√	√	√	X	7	Gmm_cn15569	√	X	√	√
8	GmmGLAAS20TVB	X	X	X	X	8	Gmm_GLAAS20TVB	X	X	√	√
9	Gmmcn15567	X	X	X	X	9	Gmm_cn15567	X	X	X	X
10	Gmmcn14707	√	√	√	√	10	Gmm_cn14707	√	X	√	√
11	Gmmcn15220	X	X	X	X	11	Gmm_cn15220	X	X	X	X
12	Gmmcn13435	X	√	√	X	12	Gmm_cn13435	√	X	√	√
13	Gmmcn15565	X	X	X	X	13	Gmm_cn15565	X	X	X	X
14	Gmmcn14014	X	X	X	X	14	Gmm_cn14014	X	X	√	X
15	GmmGLAC953TV	X	X	X	X	15	Gmm_GLAC953TV	X	X	X	X
16	GmmGLAER58TV	X	X	X	X	16	Gmm_GLAER58TV	X	X	X	X
17	Gmmcn15331	√	√	X	X	17	Gmm_cn15331	√	X	√	√
18	Gmmcn13968	X	X	X	X	18	Gmm_cn13968	X	X	X	X
19	Gthcontig219	√	√	√	√	19	Gthcontig219	√	√	√	√

Glossina pallidipes Male (GpM) and Female (GpF) antennae (A), head (H), thorax (T) and abdomen (Ab) amplified fragments with primers number 1 to 19. √- represents amplification and X-represents No amplification.

Fifty (50) amplified fragments were resolved on a 1% agarose/ethidium bromide (0.1ug/ml) gel along with 1 Kb+ ladder. Most of the amplified fragments were of the size range between 250bp to 500bp (Figures 4.2 a, b and c).

Figure 4.2a. PCR amplification of GpF tissues with designed OBP primers. M–1 kb ladder, C–Control DNA (100ng/ul), DNA Samples – 1-GpF_2A; 2-GpF_4A; 3-GpF_3A; 4-GpF_3T; 5-$GpF_{19}A$; 6-$GpF_{19}H$; 7-$GpF_{19}T$; 8-$GpF_{19}Ab$; 9-GpF_6A; 10-GpF_6H; 11-GpF_6T; 12-GpF_6Ab; 13-GpF_7A; 14-GpF_7H; 15-GpF_7T; 16-$GpF_{10}A$; 17-$GpF_{10}H$. [1-GpF_2A - Sample1 GpF Antennae amplified with primer number 2]

Figure 4.2b. PCR amplification of GpF and GpM tissues with designed OBP primers. M–1 kb ladder, C–Control DNA (100ng/ul), DNA Samples – 18-$GpF_{10}T$; 19-$GpF_{10}Ab$; 20-$GpF_{12}H$; 21-$GpF_{12}T$; 22-$GpF_{17}A$; 23-$GpF_{17}T$; 24-GpM_2A; 25-GpM_4A; 26-GpM_6A; 27-GpM_6T; 28-GpM_6Ab; 29-GpM_3A; 30-GpM_3T; 31-GpM_3Ab; 32-$GpM_{19}A$; 33-$GpM_{19}H$; 34-$GpM_{19}T$. [34-$GpM_{19}T$ – Sample 34 GpM Thorax amplified with primer number 34]

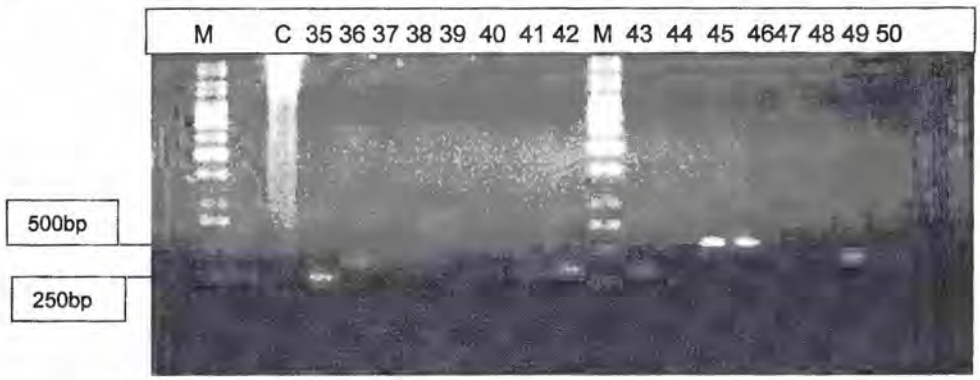

Figure 4.2c. PCR amplification of GpM tissues with designed OBP primers. M–1 kb ladder, C–Control DNA (100ng/ul), DNA Samples – 35-$GpM_{19}Ab$; 36-GpM_7A; 37-GpM_7T; 38-GpM_7Ab; 39-GpM_8T; 40-GpM_8Ab; 41-$GpM_{10}A$; 42-$GpM_{10}T$; 43-$GpM_{10}Ab$; 44-$GpM_{12}A$; 45-$GpM_{12}T$; 46-$GpM_{12}Ab$; 47-$GpM_{14}T$; 48-$GpM_{17}A$; 49-$GpM_{17}T$; 50-$GpM_{17}Ab$ [36-GpM_7A- Sample36 GpM Antennae amplified with primer no. 7]

4.3.4 Annotation of amplified *Glossina* sequences

A total of eighty seven (87) amplified fragments were sequenced bi-directionally and functional annotation carried out using blastx program against NR protein database of GenBank (Wheeler *et al.*, 2010) and GO database (Ashburner *et al.*, 2000). All consensus sequences had an estimated size range of between 268bp to 522bp (comparable with the amplified fragments in Figures 4.2a,b,c) and are possibly related to olfaction based on the high E-values and similar homologs to other insect's olfactory proteins (Table 4.2). Similar results were reported for both *G. pallidipes* male and female with twenty-eight of the consensus sequences having OBP homologs to *Drosophila* and one consensus sequence to *A. gambiae* OBP homolog. Five of the consensus sequences had a signal peptide, an indication that they are secretory proteins.

Alignment of *G. pallidipes* male and female sequences with selected OBP homologs revealed that they had between three to six conserved cysteine residues with the exception of consensus sequence, con37_P7 and con46_P12, which had only one cysteine (Figure 4.3a & b). The OBPs are low in sequence similarity, demonstrating the highly divergent nature of this protein family in insects unlike the deduced CSPs homologs that had four remarkably conserved cysteine amino acids, a lysine residue between cysteines 3 and 4 and a glutamine residue at the fourth amino acid from cysteine 4 (Figure 4.4). In addition, three highly conserved

aromatic residues was revealed from the alignment with the first, second and third aromatic amino acids being phenylalanine, tryptophan and tyrosine respectively. A similar conserved pattern has been reported for lepidopteran CSPs at positions 26, 81 and 94 and is probably involved in ligand binding in all insect CSPs (Briand *et al.,* 2002; Gong *et al.,* 2007).

Table 4.2 Functional annotation of Consensus sequences against nonredundant and Gene Ontology database

Consensus No.	Size (bp)	Blastx match to NR database	Accession No	Species	E-value	SignalP	Best Match to GO database	E-Value	GO No	Description
con1_P2	418	Os-E	ref|NP_524242.2|	*Drosophila melanogaster*	6.0E-36	Yes	Os-E	5.4E-36	0005549/ 0005550	MF odorant/pheromone binding
con2_P4	352	PBP- protein 3	ref|NP_524241.1|	*Drosophila melanogaster*	4.0E-31	No	Pbprp3	1.9E-31	0005549	MF odorant/pheromone binding
con3_P3	269	EBP III	ref|NP_524966.1|	*Drosophila melanogaster*	6.0E-38	No	EBP III	3.0E-37	0007552/0009615	BP metamorphosis/ response to virus
con4_P3	301	EBP III	ref|NP_524966.1	*Drosophila melanogaster*	6.0E-38	No	EBP III	3.0E-37	0007552/0009615	BP metamorphosis/ response to virus
con9_P6	416	Obp 99b	gb|ABW78183.1|	*Drosophila melanogaster*	2.0E-29	No	Obp99b	3.2E-33	0005549	MF odorant binding
con10_P6	418	Obp 99b	gb|ABW78183.1|	*Drosophila melanogaster*	4.0E-31	No	Obp99b	1.1E-34	0005549	MF odorant binding
con11_P6	487	Obp 99b	gb|ABW78399.1|	*Drosophila melanogaster*	5.0E-31	Yes	Obp99b	1.1E-33	0005549	MF odorant binding
con12_P6	414	Obp 99b	gb|ABW78399.1|	*Drosophila melanogaster*	3.0E-35	No	Obp99b	1.1E-34	0005549	MF odorant binding
con13_P7	283	Obp44a	ref|NP_610358.1|	*Drosophila melanogaster*	9.0E-24	No	Obp99b	3.4E-27	0005549	MF odorant binding
con14_P7	282	Obp44a	ref|NP_610358.1|	*Drosophila melanogaster*	9.0E-03	No	Obp44a	1.1E-30	0005549	MF odorant binding
con15_P7	291	Obp44a	ref|NP_610358.1|	*Drosophila melanogaster*	7.0E-31	No	Obp44a	8.3E-31	0005549	MF odorant binding
con16_P10	273	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	3.0E-06	No	Obp8a	2.0E-09	0005549	MF odorant binding
con17_P10	274	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	3.0E-06	No	Obp8a	2.0E-08	0005549	MF odorant binding
con18_P10	269	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	3.0E-06	No	Obp8a	2.0E-08	0005549	MF odorant binding
con19_P10	270	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	3.0E-06	No	Obp8a	2.0E-08	0005549	MF odorant binding
con20_P12	424	Obp 99c	gb|ABW78717.1|	*Drosophila melanogaster*	2.0E-43	No	Obp99c	1.3E-41	0005549	MF odorant binding
con21_P12	522	Obp 99c	gb|ABW78717.1|	*Drosophila melanogaster*	1.0E-43	Yes	Obp99c	1.0E-41	0005549	MF odorant binding
con24_P2	417	Os-E	ref|NP_524242.2|	*Drosophila melanogaster*	3.0E-35	Yes	OS-E	2.3E-35	0005549	MF odorant/pheromone binding
con25_P4	364	AF437884_1 OBP	gb|AAL84179.1|	*Anopheles gambiae*	8.0E-32	No	Pbprp3	9.0E-33	0005549	MF odorant/pheromone binding
con26_P6	439	Obp 99b	gb|ABW78399.1|	*Drosophila melanogaster*	1.0E-36	No	Obp99b	5.7E-36	0005549	MF odorant binding
con27_P6	489	Obp 99b	gb|ABW78399.1|	*Drosophila melanogaster*	3.0E-30	Yes	Obp99b	3.6E-30	0005549	MF odorant binding
con28_P6	488	Obp 99b	gb|ABW78399.1|	*Drosophila melanogaster*	5.0E-03	No	Obp99b	1.1E-33	0005549	MF odorant binding

Table 4.2 Cont.

Consensus No.	Size (bp)	Blastx match to NR database	Accession No	Species	E-value	SignalP	Best Match to GO database	E-Value	GO No	Description
con29_P3	298	EBP III	ref|NP_524966.1|	*Drosophila melanogaster*	6.0E-38	No	EBP III	3.0E-37	0007552/0009615	BP metamorphosis/ response to virus
con30_P3	298	EBP III	ref|NP_524966.1|	*Drosophila melanogaster*	6.0E-38	No	EBP III	3.0E-37	0007552/0009615	BP metamorphosis/ response to virus
con31_P3	302	EBP III	ref|NP_524966.1|	*Drosophila melanogaster*	3.0E-38	No	EBP III	3.0E-37	0007552/0009615	BP metamorphosis/ response to virus
con36_P7	283	Obp44a	ref|NP_610358.1|	*Drosophila melanogaster*	2.0E-30	No	Obp44a	1.7E-30	0005549	MF odorant binding
con37_P7	288	Obp44a	ref|NP_610358.1|	*Drosophila melanogaster*	2.0E-24	No	Obp44a	8.2E-28	0005549	MF odorant binding
con38_P7	296	Obp44a	ref|NP_610358.1|	*Drosophila melanogaster*	7.0E-31	No	Obp44a	8.4E-31	0005549	MF odorant binding
con39_P8	519	Obp 83ef	ref|NP_731042.1|	*Drosophila melanogaster*	1.0E-16	No	Obp83ef	9.2E-24	0005549	MF odorant binding
con40_P8	511	Obp83ef	ref|XP_001358932.2|	*Drosophila pseudoobscura*	5.0E-21	No	Obp83ef	8.8E-25	0005549	MF odorant binding
con41_P10	268	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	3.0E-06	No	Obp8a	2.0E-08	0005549	MF odorant binding
con42_P10	277	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	3.0E-06	No	Obp8a	2.0E-08	0005549	MF odorant binding
con43_P10	277	Obp 8a	ref|NP_727322.1|	*Drosophila melanogaster*	7.0E-07	No	Obp8a	5.9E-09	0005549	MF odorant binding
con 47_P14	437	Obp 19b	ref|NP_608391.2|	*Drosophila melanogaster*	5.0E-26	No	Obp19b	2.1E-26	0005549	MF odorant binding

con20_P12 – consensus sequence number 20 (derived from pair wise alignment of 20_P12F and 20_P12R) and amplified by primer number 12. Os-E - Olfactory specific E; PBP/Pbprp3 - Pheromone binding protein; Obp - Odorant binding protein; EBP III - Ejaculatory Bulb Protein III; MF – Molecular Function; BP – Biological Process

```
con1_P2              ---TFAMTLLLSFGLNNAQ---------------KPRRDENYPPPDFLKSFKIIHDVTVEKTGATEEAIKEFSDGE----IHEDPALKYYMNFLFHEVN
con24_P2             ---TFAMTLLLSFGLNNAQ---------------KPRRDENYPPPDFLKSFKIIHDVTVEKTGATEEAIKEFSDGE----IHEDPALKYYMNFLFHEVN
con25_P4             -------------------------------------RRDENYPPPDFLKSFKIIHDVTVEKTGATEEAIKEFSDGE----IHEDPALKYYMNFLFHEVN
con2_P4              -------------------------------------DAMKIYPPPDFLKSFKIIHDVTVEKTGATEEAIKEFSDGE----IHEDPALKYYMNFLFHEVN
OsEDro_melanogaster  --MVKYPLILLLIGCAAAQ---------------EPRRDGEWPPPAILKLGKHFHDIGAPKTGVTDEAIKEFSDGQ----IHEDEALKYYMNFLFHEFE
OBPAnopheles_gambiae-MLVQDSPLLLLVLLLVTQCLDGADCSTTTTQRPAPRRDGQYPPPETLAFLRPLGKLALEETGVSPEAVKRFSDADP---FDDNRALKYYMDFMFRVTN
con10_P6             ----------LVTLATVWG-----------HHHHEHHDDDDY-VVKTREDLFKYRDEISNKLNVPADLLEKYKKWQ----YPDDEVTKYYMKFMFEHFG
con26_P6             --------LFVVTLATVWG-----------HHHHEHHDDDDY-VVKTREDLFKYRDEISNKLNVPADLLEKYKKWQ----YPDDEVTKYYMKFMFEHFG
con9_P6              ----------IWSHWQQFG-----------VITIMNIMMMMITCLKTREDLFKYRDEISNKLNVPADLLEKYKKWQ----YPDDEVTKYYMKFMFEHFG
con14_P7             ------------------------------------------------------------------TPELVEKYKKFD----FPDDEITRYYIEFIFDKFQ
con36_P7             -----------------------------------------------------------------SDELVEKYKKFD----FPDDEITRYYIEFIFDKFQ
con13_P7             ------------------------------------------------------------------TAMLVEKYKKFD----FPDDEITRYYIEFIFDKFQ
con16_P10            ------------------------------------------------------------LEKVPDRYKAQFTEFQ----FPNDPIVHKYILFVNRELQ
con41_P10            -----------------------------------------------------------------KVPDRYKAQFTEFQ----FPNDPIVHKYILFVNRELQ
con17_P10            ------------------------------------------------------------SKKYLHRYKAQFTEFQ----FPNDPIVHKYILFVNRELQ
GOBP1Heli_virescens  ----MPGVLRALLLLAAAA----------------PLLADVNVMKDVTLGFGQALDKGREESQLTEEKMEEFFHFWRDDFKFEHRELGHAIQYMSRHFN
GOBP2Bombyx_mori     ----MFSFLILVFVASVAD----------------SVIGTAEVMSHVTAHFGKTLEEGREESGLSVDILDEFKHFWSDDFDVVHRELGHAIIYMSNKFS
PBPBombyx            -MSIQGQIALALMVNMAVG----------------SVDASQEVMKNLSLNFGKALDEGKKEMTLTDAINEDFYNFWKEGYEIKNRETGHAIMYLSTKLN
con20_P12            ----------LLTIVTCIR------------------AEDEDWQPKTVADIKSIRNEILKEHPLSNEQITKMKNFE----FPDEEEVRQYLLYTALKME
MSSPG1C.capitata     -----MKYFIVILAAVVLA------------------QAQDDWKIKSANEVNDIRREIHKEHLFNEELQKHEEVLP----LPDEDVVRNYEVFVFTKWE
ABP3Manduca_sexta    -MITATLHVVFALLGFVYG------------------AKNKPVFSEEIKEIIQTVHDEIVGKTGVSEEDIANCENGI----FKEDVKLKYYMFLLEVAG
SalivaryOBPC.tarsali--MVSVTSIAVVLAFAVIS------------------ASSDDAVADSKQQYLKNKELQGKLLQVPAEAMEQYQRFE----YPENHETFYYIRIIGILQG
OBPASP5A_mellifera   --MHVKSVLLLITIVTFVA----------------LKPVKSMSADQVEKLAKNMRKSHLQKIAITEELVDGMRRGE----FPDDHDLQYYTFIMKLLR
PBP2D.melanogaster   MSHLVHLTVLLLVGILCLG-----------------ATSAKPHEEINRDHAAELANEIKAETGATDEDVEQLMSHD----LPERHEAKYLRAFVMKKLQ
OBP56eAedes_aegypti  KFSSALSFCSLQLAWRFVT-------------ELQCANSDEEKKAQAKEMMRGMAEEIKKEGATDEDVEALLEDK----TPETEVQKYFLSIFQHQFQ
OBP2Phy_diversa      ----------------------------------------KEHGQKVLEQIIDYATSIADSLGVSPEDMKLLMEKK----FPTSREGQYMPSFVNKKFG
```

```
con1_P2                  VVDDAGELHFEKLVRMIPEPF------LEMVKHIIDAESHIPKGETWDRAWSWHVYFKQTDPVLYF--------------------- 139
con24_P2                 VVDDAGELHFEKLVRMIPEPF------LEMVKHIIDAESHIPKGETWDRAWSWHVYFKQTDPVLY---------------------- 138
con25_P4                 VVDDAGELHFEKLVRMIPEPF------LEMVKHIIDAESHIPKGETWDRAWSWHVYFKQTDLMCF---------------------- 120
con2_P4                  VVDDAGELHFEKLVRMIPEPF------LEMVKHIIDAESHIPKGETWDRAWSWHVYFKQTDP-------------------------- 117
OsEDro_melanogaster      VVDDNGDVHMEKVLNAIPG------ELRNIMMEASKGIHPEGDTL-WHKAWWFHQWWKKADPVHYFLV------------------- 141
OBPAnopheles_gambiae     VTDDRGELHMGKLLEHVPT------EFEDIALRMGVRTRPKGKDV-WERAFWFHKWWKTSDPVHYYLV------------------- 157
con10_P6                 FFNEKQGFDVHKIHKQLMGAHGT-VDHSDETHEKIAKCADKKPE-DPW--AWAYGVYFINSNLQC---------------------- 138
con26_P6                 FFNEKQGFDVHKIHKQLMGAHGT-VDHSDETHEKIAKCADKKPE-DPW--AWAYGVYFINSNLQLVKSS------------------ 144
con9_P6                  FFNEKQGFDVHKIHKQLMGAHGT-VDHSDETHEKIAKCADKKPE-DPW--AWAYGVYFINSNLQ----------------------- 138
con14_P7                 LFDSQTGFKNDNLIAQLGQSK----DNKDEVKADIEKCADKNTEKS-WTWAFRGFKYFISKNLPL----------------------- 93
con36_P7                 LFDSQTGFKNDNLIAQLGQSK----DNKDEVKADIEKCANKNTEKS-WTWAFRGFKYFISKNLPLV---------------------- 94
con13_P7                 LFGSQTGFKNDNLIAQLGQSK----DNKDEVKTDIEKCAXKNIKNQ-IHVLGRSAASNVSLAR-IYLW------------------- 93
con16_P10                IWDNNQGFDIEKIYQQYKGRA-----NEEVVLPIISQWNQDAKQREWWYK---AFLYILD---------------------------- 90
con41_P10                IWDNNQGFDIEKIYQQYKGRA-----NEEVVLPIISQWNQDAKQREWWYK---AFLYILD---------------------------- 88
con17_P10                IWDNNQGFDIEKIYQQYKGRA-----NEEVVLPIISQWNQDAKQREWWYK---AFLYILD---------------------------- 90
GOBP1Heliothis_virescens LLTDSSRMHHDNTEKFIQSFPNG-EVLARQMVELIHSWEKQFDHEEHWWRISHLADWFKSSCVQRGIAPSMELMMTEFIMEAEAR-- 164
GOBP2Bombyx_mori         LMDDDVRMHHVNMDEYIKGFPNG-QVLAEKMVKLIHNWEKQFDTETDWTRVVKVAAWFKKDSRKEGIAP--EVAMIEAVIEKY---- 160
PBPBombyx                MLDPEGNLHHGNAMEFAKKHGAD-ETMAQQLIDIVHGWEKSTPANDKWIWTLGVATYFKAEIHKLNWAPSMDVAVGEILAEV----- 164
con20_P12                VFCAHQGYHPNRIAKQFKMDMN-----EEEVLEIAEKWHDSNPDNSSVDVWAFRGHYMMSSAIGDKVKAYIKKRQVN---------- 140
MSSPG1C.capitata         FSNARNRFKKDRLVRQFEPVLK-----REEIDEIIGRWADKNEQGSPVDVWVRFQQYVSRSEIPPNFLKIIGKL------------- 142
ABP3Manduca_sexta        LADEDGTVDYDMLVSLIPEEY------SERASKMIFAWNHLDTPEK-DKCQRSFH-YTYEKDPEFYFLF------------------ 141
SalivaryOBPCulex_tarsalisHFEDEQGLQVDKLFALSNMGK-----SKEEFTELVNGCQAQVGEDVSHYCHKAYILWFWKQFREWKKTAGSGEEAA----------- 148
OBPASP5_Apis_mellifera   TFKN-GNFDFDMIVKQLEITMP--PEEVVIGKEIVAVYRNEEYTGDDKQKTYQYVQHHYKQNPEKFFFP------------------ 143
PBP2Dro_melanogaster     IMDESGKLNKEHAIELVKVMSKHDAEKEDAPAEVVAKIEAIETPEDHLDAAFAYEEKIYEQMKEHGLELEEH--------------- 150
OBP56eAedes_aegypti      ISDG-KRFNKDGFMQLSAMMFGEDQEKMATAEEIAEEQSSVENADR-QLSVDIKEVEKAMDKRGIKMEK---------------- 151
OBP2Phyllopertha_diversa LQKADGTLNKEYRYSEMENVKAIDEEIYNKMNSVWDKWVINGADGT-WDTGMKVVTWMKEESEKLGLSKDAIGF------------- 130
```

Figure 4.3a Multiple Sequence Alignment of GpM and GpF antennae and head sequences (con1_P2; con2_P4; con9_P6; con10_P6; con13_P7; con14_P7; con16_P10; con17_P10; con20_P12; con24_P2; con25_P4; con26_P6; con36_P7; con41_P10) with selected homologs. Accession numbers are: OsE *Drosophila melanogaster* gi|24644477|; OBP *Anopheles gambiae* gi|19071272|; GOBP1 *Heliothis virescens* gi|1255931|; GOBP2 *Bombyx mori* gi|112984436|; PBP *Bombyx mori* gi|112984442|; EBPIII *Drosophila melanogaster* gi|24762502|; MSSPG1 *Ceratits capitata* gi|6682279|; ABP3 *Manduca sexta* gi|18140731|; Salivary OBP *Culex tarsalis* gi|215259537|; OBPASP5 *Apis mellifera* gi|18140749|; PBP2 *Drosophila melanogaster* gi|24643509|; OBP56e *Aedes aegypti* gi|157103281|; OBP2 *Phyllopertha diversa* gi|6521353|. Con1_P2 – consensus sequence number 1 (derived from pairwise alignment of 1_P2F and 1_P2R) and amplified by primer number 2. OsE-Olfactory-specific E; OBP-Odorant-binding protein; GOBP1-general odorant-binding protein; PBP-Pheromone-binding protein; MSSPG1-male specific serum polypeptide gamma 1; ABP3-antennal binding protein 3; OBPASP5-odorant binding protein ASP5. The conserved cysteine residues are shaded red and conserved amino acids shaded grey.

```
GOBP1Heliothis_virescens ----------------------MPGVLRALLLLAAAAPLLADVNVMKDVTLGFGQALDKVREESQLTEEKMEEFFH-FWRDDFKFEHRELGVA
GOBP2Bombyx_mori         ----------------------MFSFLILVFVASVADSVIGTAEVMSHVTAHFGKTLEEVREESGLSVDILDEFKH-FWSDDFDVVHRELGVA
PBPBombyx                -------------------MSIQGQIALALMVNMAVGSVDASQEVMKNLSLNFGKALDEVKKEMTLTDAINEDFYN-FWKEGYEIKNRETGKA
con11_P6                 -GHIGKRRRDEFYLIFFFKNFSLLSLAYATVWGHHHHEHHDDDDXVVKTREDLFKYRDEVSNKLNVPA-DLLEKYK-KWQ---YPDDEVTKEY
con28_P6                 IGHIGKRGRDEFYLIFFFKNFSLLSLAYATVWGHHHHEHHDDDDYVVKTREDLFKYRDEVSNKLNVPA-DLLEKYK-KWQ---YPDDEVTKEY
con12_P6                 ----------------------LVTLA--TVWGHHHHEHHDDDDYVVKTREDLFKYRDEVSNKLNVPA-DLLEKYK-KWQ---YPDDEVTKEY
con27_P6                 FGHIGKRRRDEFYLIFFFKNFSLLSLAYATVWGHHHHEHHDDDDYVVKTREDLFKYRDEVSNKLNVPA-DLLEKYK-XWQ---YPDDEVTKEY
con15_P7                 ------------------------------------------------------------------S-ELVEKYK-KFD---FPDDEITREY
con38_P7                 ------------------------------------------------------------------RHR-ELVEKYK-KFD---FPDDEITREY
con37_P7                 ------------------------------------------------------------------VTP-ELVEKYK-KFD---FPDDEITREY
con18_P10                ------------------------------------------------------------------EKVPD-RYKAQFT-EFQ---FPNDPIVHKY
con43_P10                -----------------------------------------------------------------CLEKVPD-RYKAQFT-EFQ---FPNDPIVHKY
con19_P10                ------------------------------------------------------------------EKVPD-RYKAQFT-EFQ---FPNDPIVHKY
con42_P10                ------------------------------------------------------------------EKVPEYRYKAQFT-EFQ---FPNDPIVHKY
OsEDro_melanogaster      ----MVKYP----LILLLIG-------CAAAQ----EPRRDGEWPPPAILKLGKHFHDIVAPKTGVTD-EAIKEFS-DGQ---IHEDEALKEY
OBPAnopheles_gambiae     ---MLVQDSPLLLLVLLLVTQCLDGADCSTTTTQRPAPRRDGQYPPPETLAFLRPLGKLVLEETGVSP-EAVKRFS-DADP--FDDNRALKEY
con39_P8                 ----------IDRSRMSAHSLSRANALNRYYFLRIYFKNNYNKNLKYFLFKEMNINNLEVVKEKSISMDQ----LLEYYH---FPQLEHIPEL
con40_P8                 ----------NERSRMSAHSLSRANALNRYYFLRIYFKNNYNKNFIYFLFAEMTHNNLEVVKEKSISMDQ----LLEYYH---FPQLEHIPEL
ABP3Manduca_sexta        --------------------MITATLHVVFALLGFVYGAKNKPVFSEEIKEIIQTVHDEVVGKTGVSEEDIANCENGIFK-----EDVKLKEY
con21_P12                -------YRSLVAPGVRGGLLLYLVGLLFYSRSFISQIRAEDEDWQPKTVADIKSIRNEVLKEHPLSNEQITKMKNFEFP-----DEEEVRQY
MSSPG1C.capitata         -------------------------MKYFIVILAAVVLAQAQDDWKIKSANEVNDIRPEVHKEHLFNEELQKHEEVLPLP-----DEDVVRNY
SalivaryOBPCulex_tarsalis----------------------MVSVTSIAVVLAFAVISASSDDAVADSKQQYLKNKELVGKLLQVPAEAMEQYQRFEYP-----ENHETFEY
OBPASP5_Apis_mellifera   --------------------MHVKSVLLLITIVTFVALKPVKSMSADQVEKLAKNMRKSVLQKIAITEELVDGMRRGEFP-----DDHDLQEY
PBP2Dro_melanogaster     --------------------MSHLVHLTVLLLVGILCLGATSAKPHEEINRDHAAELANEVKAETGATDEDVEQLMSHDLP-----ERHEAKEL
OBP2Phyllopertha_diversa ------------------------------------------KEHGQKVLEQIIDYATSVADSLGVSPEDMRLLMEKKFP-----TSREGQEM
OBP56eAedes_aegypti      ---------------KFSSALSFCSLQLAWRFVTELQCANSDEEKKAQAKEMMRGMAEEVKKKEGATDEDVEALLEDKTP-----ETEVQKEF
con47_P14                ----------------------MLYSLRFGIMLVVTNAEDDDENEIGMTLDELADALEEDVEPKPERDHIKQLLTNDE-------NPHENSKEF
con46_P12                ------------------------------------------------------------------AYLTNNNIPPVNH-EKP-----CKRVSGTK
```

```
GOBP1Heliothis_virescens   IQCMSRHFNLLTD-SSRMHHDNTEKFIQSFPNGEV-LARQMVELIHSCEKQFDHEEDH-CWRISHLADCFKSSCVQRGIAPSMELMMT 155
GOBP2Bombyx_mori           IICMSNKFSLMDD-DVRMHHVNMDEYIKGFPNGQV-LAEKMVKLIHNCEKQFDTETDD-YTRVVKVAACFKKDSRKEGIAP-EVAMI  154
PBPBombyx                  IMCLSTKLNMLDP-EGNLHHGNAMEFAKKHGADET-MAQQLIDIVHGCEKSTPANDDK-CIWTLGVATCFKAEIHKLNWAPSMDVAVG 154
con11_P6                   MKCMFEHFGFFNE-KQGFDVHKIHKQLMGAHGTVD-HSDETHEKIAKCADNETRRHSLRLGLSWR---CVLHQFFT------------ 158
con28_P6                   MKCMFEHFGFFNE-KQGFDVHKIHKQLMGAHGTVD-HSDETHEKIAKCADFYTRRHSLRLGLSWR---CVLHQFFT------------ 159
con12_P6                   MKCMFEHFGFFNE-KQGFDVHKIHKQLMGAHGTVD-HSDETHEKIAKCADKKPEDTDP-CAWAYRGGVCFINSNLQ------------ 137
con27_P6                   MKCMFEHFGFFNE-KQGFDVHKIHKQLMGAHGTVD-HSDETHEKIAKCADXKPEDTDP-CAWAYRGGVRSPQFFT------------ 160
con15_P7                   IECIFDKFQLFDS-QTGFKNDNLIAQLGQSK---D-NKDEVKADIEKCADKNTEKSDS-CTWAFRGFKCFISKNLPLVDGKF------ 97
con38_P7                   IECIFDKFQLFDS-QTGFKNDNLIAQLGQSK---D-NKDEVKADIEKCADKNTEKSDS-CTWAFRGFKCFISKNLPLVMLK------- 98
con37_P7                   IE-IFDKFQLFDS-QTGFKNDNLIAQLGQSK---D-NKDEVKADIEILRRQEYRKIR----FMYLGVPRLQMFHQESTFG-------- 93
con18_P10                  ILCVNRELQIWDN-NQGFDIEKIYQQYKGRA-----NEEVVLPIISQCNQDAKQRNYELCYKA---FLCILD---------------- 89
con43_P10                  ILCVNRELQIWDN-NQGFDIEKIYQQYKGRA-----NEEVVLPIISQCNQDAKQRNYELCYKA---FLCILDT--------------- 92
con19_P10                  ILCVNRELQIWDN-NQGFDIEKIYQQYKGRA-----NEEVVLPIISQCNQDAKQRNYELCYKA---FLCILD---------------- 89
con42_P10                  ILCVNRELQIWDN-NQGFDIEKIYQQYKGRA-----NEEVVLPIISQCNQDAKQRNYELCYKA---FLCILD---------------- 90
OsEDro_melanogaster        MNCLFHEFEVVDD-NGDVHMEKVLNAIPGEK-----LRNIMMEASKGCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV---------- 141
OBPAnopheles_gambiae       MDCMFRVTNVTDD-RGELHMGKLLEHVP-TE-----FEDIALRMGVRCTRPKGKDVCERAFWFHKCWKTSDPVHYYLV---------- 157
con39_P8                   FKCFADKSHLYTV-DYEWNVLNWLKAFGPIR--------NENADISICRVNANEREKMICAIFVRVLLGTSKLRYPP--FPPTYKKAL 160
con40_P8                   FKCFADKSHLYTV-DYEWNVLNWLKAFGPIR--------NENADISICRVNANEREKMICAIMYEEYNCWERLNYNTDGISVTYKKAL 162
ABP3Manduca_sexta          MFCLLEVAGLADE-DGTVDYDMLVSLIPEEY------SERASKMIFACNHLDTPEKDK-CQRSFDVHKCTYEKDPEFYFLF------- 141
con21_P12                  LLCTALKMEVFCA-HQGYHPNRIAKQFKMDMN-----EEEVLEIAEKCHDSN-PDNSSVDVWAFRGHKCMMSSAIGDKVNAYIKKRQE 162
MSSPG1C.capitata           EVCVFTKWEFSNA-RNRFKKDRLVRQFEPVLK-----REEIDEIIGRCADKN-EQGSPVDVWVYRFQQCVSRSEIPPNFLKIIGKL-- 142
SalivaryOBPCulex_tarsalis  IRCIGILQGHFED-EQGLQVDKLFALSNMGK-----SKEEFTELVNGCQAQVGEDVSCYCHKAYIPLMCFWKQFREWKKTAGSGEEAA 148
OBPASP5_Apis_mellifera     TTCIMKLLRTFK--NGNFDFDMIVKQLEITMP----PEEVVIGKEIACRNEE--YTGDDCQKTYQYVQCHYKQNPEKFFFP------- 143
PBP2Dro_melanogaster       RACVMKKLQIMDE-SGKLNKEHAIELVKVMSKHDAEKEDAPAEVVAKCEAIETPEDHCDAAFAYEECIYEQMKEHGLELEEH------ 150
OBP2Phyllopertha_diversa   PSCVNKKFGLQKA-DGTLNKEYRYSEMENVKAIDEEIYNKMNSVWDKC-VINGADGTDECDTGMKVVTCMKEESEKLGLSKDAIGF-- 130
OBP56eAedes_aegypti        LSCFQHQFQISD--GKRFNKDGFMQLSAMMFGE---DQEKMATAAEECSSVE---NADRCQLSVDIKECVEKAMDKRGIKMEK----- 151
con47_P14                  RRCLMEQFELIDEGQSQMNKDKVVDMMSMMYAD---NKETLEEIVDHCNTKNGATTEK-CENAHQHGMCILNQLKENGL--------- 143
con46_P12                  TGRFSPPFGYEKS-TGKTQGNRFLILPIGGG------GLAVSKEVAPCPAAN---------FPFLGEE-------------------- 76
```

Figure 4.3b Multiple Sequence Alignment of GpM and GpF thorax and abdomen sequences (con11_P6; con12_P6; con15_P7; con18_P10; con19_P10; con21_P12; con27_P6; con28_P6; con37_P7; con39_P8; con38_P7; con40_P8; con42_P10; con43_P10; con46_P12; con47_P14) with selected homologous sequences. Accession numbers are: OsE *Drosophila melanogaster* gi|24644477|; OBP *Anopheles gambiae* gi|19071272|; GOBP1 *Heliothis virescens* gi|1255931|; GOBP2 *Bombyx mori* gi|112984436|; PBP *Bombyx mori* gi|112984442|; EBPIII *Drosophila melanogaster* gi|24762502|; MSSPG1 *Ceratits capitata* gi|6682279|; ABP3 *Manduca sexta* gi|18140731|; Salivary OBP *Culex tarsalis* gi|215259537|; OBPASP5 *Apis mellifera* gi|18140749|; PBP2 *Drosophila melanogaster* gi|24643509|; OBP56e *Aedes aegypti* gi|157103281|; OBP2 *Phyllopertha diversa* gi|6521353|. Con12_P6 – consensus sequence number 12 (derived from pairwise alignment of 12_P6F and 12_P6R) and amplified by primer number 6. OsE-Olfactory-specific E; OBP-Odorant-binding protein; GOBP1-general odorant-binding protein; PBP-Pheromone-binding protein; MSSPG1-male specific serum polypeptide gamma 1; ABP3-antennal binding protein 3; OBPASP5-odorant binding protein ASP5. The conserved cysteine residues are shaded red and conserved amino acids shaded grey

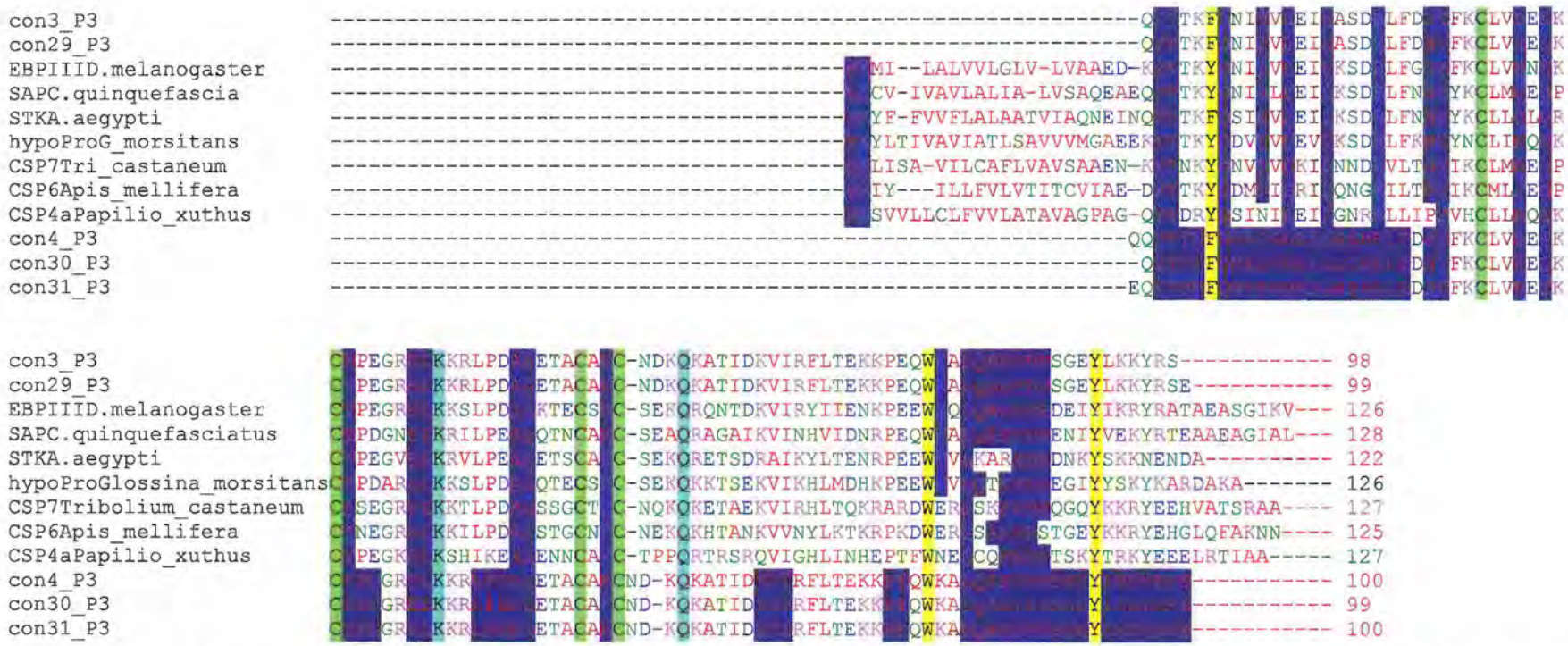

Figure 4.4 Multiple Sequence Alignment of GpM and GpF antennae, head, thorax and abdomen sequences (con3_P3; con29_P3; con4_P3; con30_P3 and con31_P3) with selected homologous sequences. Accession numbers are: EBPIII *Drosophila melanogaster* gi|24762502|; SAP *Culex quinquefasciatus* gi|170033663|; STK *Aedes aegypti* gi|157125736|; hypothetical protein *Glossina morsitans* gi|77415652|; CSP7 *Tribolium castaneum* gi|113951719|; CSP6 *Apis mellifera* gi|118150500|; CSP4a *Papilio xuthus* gi|207107814|. con29_P3 – consensus sequence number 29 (derived from pairwise alignment of 29_P3F and 29_P3R) and amplified by primer number 3. EBPIII-ejaculatory bulb protein III; SAP-sensory appendage protein; STK-serine/threonine kinase; hypoPro-hypothetical protein; CSP7-chemosensory protein 7; The conserved cysteine residues are shaded green and conserved amino acids shaded blue. Aromatic amino acid residues are shaded yellow.

4.5. Discussion

This study established that odorant binding proteins (OBPs) are located not only in *G. pallidipes* antennae but also in other body parts namely head, thorax and abdomen. Using reverse transcriptase polymerase chain reaction (RT-PCR), OBPs were amplified from both *G. pallidipes* male and female tissues using primers designed from *G. pallidipes, G. p. gambiensis, G. tachinoides* and *G. morsitans morsitans*, hence the possibility of occurrence of similar OBPs within the same and different tsetse species. Most of the amplified fragments from *G. pallidipes* male and female were similar and this could indicate that OBPs perform the same functions in both male and female tsetse flies. The antennae, being the main olfactory organ, was not amplified by some of the OBP gene primers as expected. This could be due to OBP gene primers from other *Glossina* species that are either not specific or localised within the male and female *Glossina pallidipes* antennae. The two OBP genes (Gmm_GLAAS20TVB and Gmm_cn14014) that were localised in male tissues could be sex specific OBPs whose role needs to be investigated since their identified homologs in *D. melanogaster* (Obp83ef and Obp19b), are involved in perception of chemical stimuli and binding of odorants. The two novel *Glossina* OBPs (Gpacn266 and Gphcn184) were only localized in both *G. pallidipes* male and female antennae, could represent OBPs that play an important role in tsetse olfaction. Most of the identified OBPs homologs were from *Drosophila* an indication of close relationship between these groups of sequences.

The other olfactory proteins reported in *D. melanogaster* were pheromone binding protein-related protein 3 and chemosensory protein homolog, ejaculatory bulb protein III, reported to function in reproduction and humoral response (Ludwig *et al.,* 1991; Sabatier *et al.,* 2003). It is similar to OBP in *Mamestra brassicae* in that they can detect vaccenly acetate as a ligand and could represent orthologs that have conserved function (Bohbot *et al.,* 1998). Chemosensory proteins are small, highly conserved soluble proteins expressed in the sensillium lymph (Leal *et al.,* 1999; Scaloni *et al.,* 1999) and function as carriers binding a range of aliphatic compounds, esters and pheromone blends (Nagnan-Le Meillour *et al.,* 2000; Briand *et al.,* 2002). They have been identified in both chemosensory and non-chemosensory tissues of different insects (McKenna *et al.,* 1994; Maleszka and Stange, 1997; Ozaki *et al.,* 2008), implying that they could also be involved in functions other than olfaction. It is notable that only one CSP gene was identified in this work. Small numbers of CSPs have also been identified in complete genomes of *D. melanogaster* (4 CSPs), *A. gambiae* (7 CSPs), *A. mellifera* (6 CSPs) while a large number were identified in Lepidoptera, Orthoptera and phytophagous insects (Angeli *et al.,* 1999; Robertson *et al.,* 1999; Zhou *et al.,* 2006). Identification of OBPs and CSP in antennae, head, thorax and abdomen could imply other functions of olfactory proteins in tsetse flies. Localisation of the olfactory proteins in different body parts of male and female *G. pallidipes* need to be investigated further to quantify their expression levels and determining their possible functions.

CHAPTER 5

5.0 GENERAL DISCUSSION AND FUTURE DIRECTIONS

This study reports the identification, characterization and tissue localisation of nine (9) putative odorant binding proteins (OBPs) and one (1) chemosensory protein (CSP) in the tsetse (*Glossina pallidipes*, *Glossina palpalis gambiensis* and *Glossina tachinoides*). The different genes identified in the tsetse antennae and head suggests occurance of many cellular and metabolic processes. Characterization of OBP and CSP genes open avenues for further analysis to elucidate their functional role and molecular basis of olfaction in tsetse flies that could be vital in determining the vectorial capacity of different *Glossina* species in parasite transmission. The discovery of orthologs in phylogenetically related dipteran species *G. morsitans morsitans* (Liu *et al.*, 2010), *Drosophila melanogaster* (Pikielny *et al.*, 1994), *Anopheles gambiae* (Xu *et al.*, 2003), *Aedes aegypti* (Zhou *et al.*, 2008) and *Culex quinquefasciatus* (Pelletier and Leal, 2009) suggests that the OBPs and CSP identified in *G. pallidipes*, *G. p, gambiensis* and *G. tachinoides* may be involved in olfactory process. These results are consistent with other Expressed sequence tag (EST) and genomic studies implicating OBPs and CSPs in olfaction. However, no mutant tsetse flies defective for any OBP or CSP gene have been described, thus the *in vivo* function of these proteins is unknown. Tsetse with mulfunctional OBP or CSP genes can, however be produced and examined for behavioural response to known attractants and repellents. Such studies have been done with *Drosophila*

LUSH mutants, reported to be defective for pheromone evoked behavior (Kim and Smith, 2001; Xu *et al.,* 2005). This study also established that both OBPs and CSPs are found in chemosensory (antennae) and non-chemosensory tissues (head, thorax and abdomen). This could imply that they are also involved in functions not related to olfaction. Identification of male specific *G. m. morsitans* OBP homologs in *G. pallidipes* males calls for further investigation to ascertain their precise role as previously identified OBPs in *G. m. morsitans* reported genes that were highly transcribed in the female antennae. Localisation of OBPs and CSPs in *G. pallidipes* males and females will be interesting to investigate further to determine their expression profile within olfactory as well as other chemosensory organs. In addition, ligand binding studies with already identified attractants and repellents will generate more information on mechanism involved in odor-OBP/or CSP interactions. Such information will accelerate discovery of new potent odors that can be used to improve existing tsetse control methods (Omolo *et al.,* 2009; Rayaisse *et al.,* 2010). These results can be used to study molecular basis of olfaction in *Glossina* and help develop environmentally friendly control methods based on odor behavior as have been used in 'push-pull' pest control strategies with recombinant aphid's alarm pheromone synthase gene in *Arabidopsis thaliana* (Beale *et al.,* 2006). Thus, identification of *Glossina* OBPs provides alternative targets that can be investigated to control tsetse as it may interfere with insect host location and mating behavior (Zhou *et al.,* 2010).

REFERENCES

Aksoy, S., Maudlin, I., Dale, C., Robinson, A. S. and O'Neill, S. L. (2001). Prospects for control of African trypanosomiasis by tsetse vector manipulation. *Trends in Parasitology* **17:** 29–35.

Aksoy, S., Berriman, M., Hall, N., Hattori, M., Hide, W. and Lehane, M. (2005). A case for a *Glossina* genome project. *Trends in Parasitology* **21:** 107–111.

Altschul, S. F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.

Annual Report of the International Laboratory for Research on Animal Diseases, ILRAD, 1981.

Angeli, S., Ceron, F., Scaloni, A., Monti, M., Monteforti, G., Minnocci, A., Petacchi, R. and Pelosi, P. (1999). Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria. European Journal of Biochemistry* **262**: 745–754.

Applied Biosystems Chemistry Guide (Second Edition). DNA Sequencing by Cappilary Electrophoresis.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25, 25 – 29.

Attardo, G. M., Strickler-Dinglasan, P., Perkin, S. A. H., Caler, E., Bonaldo, M. F., Soares, M. B., El-Sayeed, N. and Aksoy, S. (2006). Analysis of fat body transcriptome from the adult tsetse fly, *Glossina morsitans morsitans. Insect Molecular Biology* **15:** 411–424.

Bailey, S. (1998). Tsetse fly eliminated on Zanzibar. Nuclear News January, 56 – 61.

Ban, L., Scaloni, A., Brandazza, A., Angeli, S., Zhang, L., Yan, Y. and Pelosi, P. (2003). Chemosensory proteins of Lo*custa migratoria. Insect Molecular Biology* **12**: 125–134.

Bargmann, C. I. (2006). Comparative chemosensation from receptors to ecology. *Nature* **444:** 295–301.

Barrett, J. C. (1997). *Economic issues in trypanosomiasis control.* NRI Bulletin 75. Chatham, UK. National Resources Institute.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Research* **28**: 263–266.

Beale, M. H., Birkett, M. A., Bruce, T.J., Chamberlain, K., Field, L.M., Huttly, A.K., Martin, J.L., Parker, R., Phillips, A.L., Pickett, J.A., Prosser, I.M., Shewry, P.R., Smart, L.E., Wadhams, L.J., Woodcock, C.M. and Zhang, Y. (2006). Aphid alarm pheromone produced by transgenic plants affects aphid and parasitoid behavior. *Proceedings of the National Academy of Science, U S A.* **103**: 10509-10513.

Benton, R. (2006). On the ORigin of smell: odorant receptors in insects. *Cellular and molecular life sciences* **63:** 1579–1585.

Berriman, M. *et al.,* (2005). The Genome of the African Trypanosome *Trypanosoma brucei. Science* **309**: 416–422.

Biessmann, H., Walter, M. F., Dimitratos, S. and Woods, D. (2002). Isolation of cDNA clones encoding putative odourant binding proteins from the antennae of the malaria transmitting mosquito, *Anopheles gambiae. Insect Molecular Biology* **11:** 123–132.

Blum, J., Nkunku, S. and Burri, C. (2001). Clinical description of encephalopathic syndromes and risk factors for their occurrence and outcome during melarsoprol treatment of human African trypanosomiasis. *Tropical Medicine & International Health* **6**: 390–400.

Boeckh, J., Ernst, K. D. and Selsam, P. (1987). *Neurophysiology and neuroanatomy of the olfactory pathway in the cockroach.* In: Olfaction and taste, Vol IX (Roper, S.D, Atema, J, eds), pp. 39 – 43. New York: New York Academy of Sciences.

Bohbot, J., Sorbian, F., Lucas, P. and Meillour, P. N. (1998). Functional characterization of a new class of odorant binding proteins in the Moth *Mamestra brassicae. Biochemical and Biophysical Research Communication* **253**: 489–494.

Bourn, D., Maitima, J., Motsamai, B., Blake, R., Nicholson, C. and Sundstol, F. (2005). *Livestock and environment*. Chapter 9 In: Livestock and wealth creation: improving the husbandry of animals kept by resource-poor people in developing countries. Nottingham University Press.

Brandl, F. E. (1988). *Economics of trypanosomiasis control in cattle*. University of Hohenheim, Kiel, Germany, Wissenschaftsverlag Vaulk.

Breer, H., Boekhoff, I. and Tareilus, E. (1990). Rapid kinetics of second messenger formation in olfactory transduction. *Nature* **345**: 65–68.

Breer, H. (2003). Sense of smell: recognition and transduction of olfactory signal. *Biochemical Society transactions* **31**: 113–116.

Briand, L., Swasdipan, C., Nespoulous, V., Bezirard, F., Blon, J. C., Huet, P., Ebert, P. and Pernollet, J. C. (2002). Characterization of a chemosensory protein (ASP3c) from honeybee (*Apis mellifera* L) as a brood pheromone carrier. *European Journal of Biochemistry* **269:** 4586–4596.

Brightwell, R., Dransfield, R. D. and Kyorku, C. (1991). Development of a low-cost trap and odour baits for *Glossina pallidipes* and *Glossina longipennis* in Kenya. *Medical Veterinary Entomology* **5**: 153–164.

Buck, L. and Axel, R. (1991). A novel multigene family may encode odorant receptors: A molecular basis for odour recognition. *Cell* **65:** 175–187.

Buck, L. (1996). Information coding in the vertebrate olfactory system. *Annual Review of Neuroscience* **19**: 517–544.

Calvo, E., Sanchez-Vargas, I., Favreau, A. J., Barbian, K. D., Pham, V. M., Olson, K. E. and Ribeiro, J. M. C. (2010). An insight into the sialotranscriptome of the West Nile mosquito vector, *Culex tarsalis*. *BMC Genomics* **11**: 51. doi:10.1186/1471-2164-11-51

Campanacci, V., Lartigue, A., Hallberg, B. M., Jones, T. A., Giudici-Orticoni, M. T., Tegoni, M. and Cambillau, C. (2003). Moth chemosensory protein exhibits drastic conformational changes and cooperativity on ligand binding. *Proceedings of the National Academy of Sciences* U.S.A. **100**: 5069–5074.

Cecchi, G., Mattioli, R. C., Slingenbergh, J. and Rocque, D, L. (2008). Land cover and tsetse fly distributions in sub–Saharan Africa. *Medical and Veterinary Entomology* doi: 10.1111/j.1365-2915.2008.00747.x

Chandrashekar, J., Hoon, M. A., Ryba, N. J. P. and Zuker, C. S. (2006). The receptors and cells for mammalian taste. *Nature* **444**: 288–294.

Chatelain, E. and Ioset, J-R. (2009). Drug discovery and development for neglected diseases: the DNDi model. In Preclinical Drug Development, *Genesis* 52 – 61.

Checchi, F., Piola, P., Ayikoru, H., Thomas, F., Legros, D. and Priotto, G. (2007). Nifurtimox plus Eflornithine for Late-Stage Sleeping Sickness in Uganda: A Case Series. *PloS Neglected Tropical Disease* **1**(2):e64.

Chenchik, A., Zhu, Y. Y., Diatchenko, L., Li, R., Hill, J. and Siebert, P. D. (1998). *Generation and use of high-quality cDNA from small amounts of total RNA by SMART PCR*. In Gene Cloning and Analysis by RT-PCR (Bio Techniques Books, MA), pp. 305 – 319.

Chevenet, F., Brun, C., Banuls, A.L., Jacq, B. and Chisten, R. (2006). TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**: 439.

Chung, T., Siol, O., Dingermann, T. and Winckler, T. (2007). Protein Interactions Involved in tRNA Gene-Specific Integration of *Dictyostelium discoideum* Non-Long Terminal Repeat Retrotransposon TRE5-A. *Molecular and Cellular Biology* **27**: 8492–8501.

Clausen, P. H., Adeyemi, I., Bauer, B., Breloeer, M., Salchow, F. and Staak, C. (1998). Host preferences of tsetse flies (Diptera: Glossinidae) based on bloodmeal identifications. *Medical Veterinary of Entomology* **12**: 169–180.

Clyne, P. J., Warr, C. G., Freeman, M. R., Lessing, D., Kim, J. and Carlson, J. R. (1999). A novel family of divergent seven-transmembrane proteins: Candidate odorant receptors in *Drosophila*. *Neuron* **22**: 327-338.

Courtis, F., Dupont, S., Zeze, D. G., Jamonneau, V., Sane, B., Coulibaly, B., Cuny, G. and Solano, P. (2005). Human African Trypanosomiasis: Urban transmission in the focus of Bonon (Cote d'Ivoire). *Tropical Medicine in International Health* **10**(4): 340-346.

Cross, G. A. M. (2010). Drug discovery: Fat-free proteins kill parasites. *Nature* **464**: 689–690.

Cuisance, D., Politzar, H., Merot, P. and Tamboura, I. (1984). Les lachers de males irradies dans la campagne de lutte integree contre les glossines dans la zone pastorale de Sideradougou, Bukina Faso. *Revne d'Elevage et de medicine Veterinaire des Pays Tropicaux* **37**: 449–467.

de Bruyne, M., Foster, K. and Carlson, J. (2001). Odour coding in the *Drosophila* antenna. *Neuron* **30:** 537–552.

de Bruyne, M. and Baker, T. C. (2008). Odor detection in Insects: Volatile codes. *Journal of Chemical Ecology* **34**: 882–897.

D'Ieteren, G. D. M., Authie, E., Wisocq, N. and Murray, M. (1998). Trypanotolerance, an option for sustainable livestock production in areas at risk from trypanosomiasis. *Revue Scientifique et Technique* **17**: 154–175.

Dransfield, R. D., Brightwell, R., Kyorku, C. and Williams, B. (1990). Control of tsetse (Glossinidae) populations using traps at Nguruman, southwest Kenya. *Bulletin of Entomological Research* **80**: 265–276.

Du, G. and Prestwich, G. D. (1995). Protein structure encodes the ligand binding specificity in pheromone binding proteins. *Biochemistry* **34**: 8726–8732.

Dutoit, R. (1954). Trypanosomiasis in Zululand and the control of tsetse by chemical means. *Onderstepoort Journal of Veterinary Research* **26**: 317–387.

Elsen, P. and Roelants, P. (1999). Isozymic comparison of five laboratory lines of tsetse flies belonging to the two subspecies of *Glossina palpalis* (Diptera: Glossinidae). *Annals of Tropical Medicine and Parasitology* **93**: 97–104.

Endo, K., Aoki, T., Yoda, Y., Kimura, K. and Hama, C. (2007). Notch signal organizes the *Drosophila* olfactory circuitry by diversifying the sensory neuronal lineages. *Nature Neuroscience* **10**: 153–160.

Enserink, M. (2007). Welcome to Ethiopia's Fly Factory. *Science* **317**: 310–313.

FAO (1982a). Training manual for tsetse control personnel. Vol. 1: *Tsetse biology, systematics and distribution; techniques*. Pollock, J. N. (Ed). Pp. 1 – 40.

FAO (1982b). Training manual for tsetse control personnel. Vol. 2: *Ecology and behaviour of tsetse*. Pollock, J. N. (Ed). Pp. 50 – 70.

FAO (1993). Training manual for tsetse control personnel. Vol. 5: *Insecticides for tsetse and trypanosomiasis control using attractive bait techniques*, Rome, FAO, 88.

Forêt, S. and Maleszka, R. (2006). Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*). *Genome Research* **16**: 1404–1413.

Forêt, S., Wanner, K. W. and Maleszka, R. (2007). Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling. *Insect Biochemistry and Molecular Biology* **37**: 19–28.

Fox, A. N., Pitts, R. J., Robertson, H. M., Carlson, J. R. and Zwiebel, L. J. (2001). Candidate odorant receptors from the malaria vector mosquito *Anopheles gambiae* and evidence of down-regulation in response to blood feeding. *Proceedings of the National Academy of Sciences* **98**: 14693-14697.

Frearson, J. A., Brand, S., McElroy, S. P., Cleghorn, L. A. T., Smid, O., Stojanovski, L., Price, H. P., Guther, M. L. S., Torrie, L. S., Robinson, D. A., Hallyburton, I., Mpamhanga, C. P., Brannigan, J. A., Wilkinson, A. J., Hodgkinson, M., Hui, R., Qiu, W., Raimi, O. G., van Aalten, D. M. F., Brenk, R., Gilbert, I. H., Read, K. D., Fairlamb, A. H., Ferguson, M. A. J., Smith, D. F. and Wyatt, P. G. (2010). *N*-myristoyltransferase inhibitors as new lead to treat sleeping sickness. *Nature* **464**: 728–732.

Gasteiger, E., Hoogland, C., Gattike, A., Duvaud, S., Wilkins, M. R., Appel, R. D. and Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server*; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. Pages 571 – 607.

Geerts, S., Holmes, P. H., Eisler, M. C. and Diall, O. (2001). African bovine trypanosomiasis: the problem of drug resistance. *Trends in Parasitology* **17**: 25–28.

Gikonyo, N. K., Hassanali, A., Njagi, P. G. N. and Saini, R. K. (2000). Behaviour of *Glossina morsitans morsitans* Westwood (Diptera: Glossinidae) on Waterbuck *Kobus defassa* Ruppel and feeding membranes smeared with Waterbuck serum indicates the presence of allomones. *Acta Tropica* **77**: 295-303.

Gikonyo, N. K., Hassanali, A., Njagi, P. G. N., Gitu, P. M. and Midiwo, J. O. (2002). Odour composition of preferred (Buffalo and Ox) and nonpreferred (Waterbuck) hosts of some Savannah tsetse flies. *Journal of Chemical Ecology* **28:** 961-973.

Gikonyo, N. K., Hassanali, A., Njagi, P. G. N. and Saini, R. K. (2003). Responses of *Glossina morsitans morsitans* to blends of electroantennographically active compounds in the odours of its preferred (Buffalo and Ox) and none preferred (waterbuck) hosts. *Journal of Chemical Ecology* **29**: 2331-2346.

Gish, W. and David, J. S. (1993). Identification of protein coding regions by database similarity search. Nature Genetics **3**: 266-272.

Gong, D., Zhang, H., Zhao, P., Lin, Y., Xia, Q. and Xiang, Z. (2007). Identification and expression pattern of chemosensory protein gene family in the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology* **37**: 266–277.

Gong, D., Zhang, H., Zhao, P., Lin, Y., Xia, Q. and Xiang, Z. (2009). The Odorant binding protein gene family from the genome of Silkworm, *Bombyx mori*. *BMC Genomics* **10**: 332 doi:10.1186/1471-2164-10-332.

Graham, L. A. and Davies, P. L. (2002). The odorant-binding proteins of *Drosophila melanogaster*: annotation and characterization of a divergent gene family. *Gene* **292**: 43–55.

Grootenhuis, J. G. and Olubayo, R. O. (1993). Disease research in the wildlife livestock interface in Kenya. Vet. Q. **15**: 55–59.

Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematics Biology* **52**: 696–704.

Guo, Y., Ribeiro, J. M., Anderson, J. M. and Bour, S. (2009). dCAS: a desktop application for cDNA sequence annotation. *Bioinformatics* **25**: 1195–1196.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic acids symposium series* **41**: 95–98.

Hansson, B. S. (1999). Insect Olfaction. Springer-Verlag Berlin Heidelberg.

Hargrove, J. W. (1980). The effect of model size and Ox odour on the alighting responses of *Glossina morsitans* Westwood and *G. pallidipes* Austen (Diptera: Glossinidae). *Bulletin of Entomology Research* **70**: 229–234.

Hargrove, J. W., Omolo, S., Msalilwa, J. S. I. and Fox, B. (2000). Insecticide treated cattle for tsetse control: the power and the problems. *Medical Veterinary Entomology* **14**: 123–130.

Hassanali, A., McDowell, P. G., Owaga, M. L. A. and Saini, R. K. (1986). Identification of tsetse attractants from excretory products of a wild host animal, *Syncerus caffer*. *Insect Science Application* 7: 5-9.

Hekmat-Scafe, D. S., Scafe, C. R., McKinney, A. J. and Tanouye, M. A. (2002). Genome-Wide analysis of the Odorant-Binding Protein gene family in *Drosophila melanogaster*. *Genome Research* **12**: 1357–1369.

Hertz-Fowler, C., Peacock, C. S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A. C., Rajandream, M-A. and Barrell, B. (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Research* **32**: Database issue D339-D343.

Hildebrand, J. G. and Shepherd, G. M. (1997). Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. *Annual Review of Neuroscience* **20**: 595–631.

Hill, C. A., Fox, A. N., Pitts, R. J., Kent, L. B., Tan, P. L., Chrystal, M. A., Crunchy, A., Collins, F. H., Robertson, H. M. and Zwiebel, L. J. (2002). G protein-coupled receptors in *Anopheles gambiae*. *Science* **298**: 176-178.

Horst, R., Damberger, F., Luginbuhl, P., Guntert, P., Peng, G. and Nikonova, L. (2001). NMR structure reveals intermolecular regulation mechanism for pheromone binding and release. *Proceedings of National Academy of Science USA* **98**: 14374–14379.

Hovemann, B. T., Sehimeyer, F and Malz, J. (1997). *Drosophila* melanogaster NADPH-cytochrome P450 oxidoreductase: pronounced expression in antennae may be related to odorant clearance. *Gene* **189**: 213–219.

Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research* **9**: 868-877

Huyton, P. M., Langley, P. A., Carlson, D. A., Coates, T. W. (2008a). The role of sex pheromones in initiation of copulatory behaviour by male tsetse flies, *Glossina morsitans morsitans*. *Physiological Entomology* **5**: 243–252.

Huyton, P. M., Langley, P. A., Carlson, D. A., Schwarz, M. (2008b). Specificity of contact sex pheromones in tsetse flies, *Glossina* spp. *Physiological Entomology* **5**: 253–264.

Ishida, Y., Chiang, V. and Leal, W. S. (2002). Protein that makes sense in the Argentine ant. *Naturwissenschaften* **89**: 505–507.

Ishida, Y., Chen, A. M., Tsuruda, J. M., Cornel, A. J., Debboun, M. and Leal, W. L. (2004). Intriguing olfactory proteins from the yellow fever mosquito, *Aedes aegypti*. *Naturwissenschaften* **91**: 426–431.

Jacobs, S. P., Liggins, A. P., Zhou, J. J., Pickett, J. A., Jin, X. and Field, L. M. (2005). OS-D-like genes and their expression in aphids (Hemiptera: Aphididae). *Insect Molecular Biology* **14**: 423–432.

Jacquin-Joly, E., Vogt, R. G., Francois, M. and Meillour, P. N. (2001). Functional and expression pattern of chemosensory proteins expressed in antennae and pheromonal gland of *Mamestra brassicae*. *Chemical Senses* **26**: 833-844.

Jacquin-Joly, E. and Merlin, C. (2004). Insect olfactory receptors: Contribution of molecular biology to chemical ecology. *Journal of Chemical Ecology* **30**: 2359–2397.

Jones, W. D., Nguyen, T. A., Kloss, B., Lee, K. J. and Vosshall, L. B. (2005). Functional conservation of an insect odorant receptor gene across 250 million years of evolution. *Current Biology* **15**: R119–R121.

Jordan, A. M. (1986). Trypanosomiasis Control and African Rural Development. Longman, London.

Jordan, A. M. (1993). *Tsetse flies (Glossinidae)*. In Medical Insects and Arachnids. R. P. Lane and R. W. Crosskey (Eds). Chapman and Hall, London.

Kabayo, J. P. (2002). Aiming to eliminate tsetse from Africa. *Trends in Parasitology* **18**: 473–475.

Kaissling, K. E. (1986). Chemo-electrical transduction in insect olfactory receptors. *Annual Review of Neuroscience* **9**: 121–145.

Kaissling, K. E. (1998). Pheromone deactivation catalyzed by receptor molecules: a quantitative kinetic model. *Chemical Senses* **23**: 385–395.

Kalume, D. E., Okulate, M., Zhong, K., Reddy, R., Suresh, S., Deshpande, N., Kumar, N. and Pandey, A. (2005). A proteomic analysis of salivary glands of female *Anopheles gambiae* mosquito. *Proteomics* **5**: 3765–3777.

Kalidas, S. and Smith, D. P. (2002). Novel genomic cDNA hybrids produce effective RNA interference in adult *Drosophila*. *Neuron* **33**: 177–184.

Kasang, G., Nicholls, M. and van Proff, L. (1989). Sex pheromone conversion and degradation in antennae of the silkworm moth *Bombyx mori* L. *Experentia* **45**: 81–87.

Kim, M. S. and Smith, D. P. (2001). The invertebrate odorant binding protein LUSH is required for normal olfactory behaviour in *Drosophila*. *Chemical Senses* **26**: 195–199.

Kitabayashi, A. N., Arai, T., Kubo, T. and 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.

Krafsur, E. S and Wohlford, D. L. (1999). Breeding structure of *Glossina pallidipes* populations evaluated by mitochondrial variation. *The Journal of Heredity* **90**: 635-642.

Krafsur, E. S (2002). Population structure of the tsetse fly *Glossina pallidipes* estimated by allozyme, microsatellite and mitochondrial gene diversities. *Insect Molecular Biology* **11**: 37–45.

Krafsur, E. S. (2009). Tsetse flies: Genetics, evolution, and role as vectors. *Infection, Genetics and Evolution* **9**: 124–141.

Krieger, J., Ganble, H., Raming, K. and Breer, H. (1993). Odorant binding proteins of *Heliothis virescens*. *Insect Biochemical Molecular Biology* **23**: 449–456.

Krieger, J., Mameli, M. and Breer, H. (1997). Elements of the olfactory signaling pathways in insect antennae. *Invertebrate Neuroscience* **3**: 137–144.

Krieger, J. and Breer, H. (1999). Olfactory reception in invertebrates. *Science* **286**: 720–723.

Krieger, J., Klink, O., Mohl, C., Raming, K. and Breer, H. (2003). A candidate olfactory receptor subtype highly conserved across different insect orders. *Journal of Comparative Physiology* **189**: 519-526.

Kruse, S. W., Zhao, R., Smith, D. P. and Jones, D. N. (2003). Structure of a specific alcohol-binding site defined by the odorant binding protein LUSH from *Drosophila melanogaster*. *Nature Structural Biology* **10:** 694–700.

Larkin, M. A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007). ClustalW and ClustalX version 2. *Bioinformatics* **23**: 2947-2948.

Larsson, M. C., Domingos, A. I., Jones, W. D., Chiappe, M. E., Amrein, H. and Vosshall, L. B. (2004). Or83b Encodes a Broadly Expressed Odorant Receptor Essential for *Drosophila* Olfaction. *Neuron* **43**: 703–714.

Lartigue, A., Campanacci, V., Roussel, A., Larsson, A.M., Jones, T.A., Tegoni, M. and Cambillau C. (2002). X-ray structure and ligand binding study of a moth chemosensory protein. *Jorunal of Biological Chemistry* **277**: 32094-32098

Laue, M., Steinbrecht, R. A. and Ziegelberger, G. (1994). Immunocytochemical localization of general odorant binding proteins in olfactory sensilla of the silkmoth *Antheraea polyphemus*. *Naturwissenschaften* **81**: 178–181.

Laveissiere, C., Vale, G. A., Gouteux, J. P. (1990). *Bait methods of tsetse control*. In: Curtis, C. F. (Ed), Appropriate Technology in Vector Control. CRC Press, Boca Raton, FL, pp. 47 – 74.

Lawson, D. et al. (2009). VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Research* **37**: D583-587; PMID: 19028744

Leak, S. G. A., Rowlands, G. J. and Mulatu, W. (1995). A trial of cypermethrin pour-on insecticide to control *Glossina pallidipes* and *G. Morsitans submorsitans* in southwest Ethiopia. *Bulletin of Entomology Research* **85**: 241–251.

Leak, S. G. A., Peregrine, A. S., Rowlands, G. J. and Mulatu, W. (1996). Use of impregnated targets for the control of tsetse flies (*Glossina* spp) and trypanosomiasis occurring in cattle in an area of southwest Africa with high prevalence of drug resistant trypanosomes. *Tropical Medicine of International Health* **1**: 599–609.

Leak, S. (1998). *Tsetse Biology and Ecology*. Their role in the Epidemiology and Control of Trypanosomiasis. New York, NY, USA: CABI Publishing.

Leal, W. S., Nikonova, L. and Peng, G. (1999). Disulfide structure of the pheromone binding protein from the silkworm moth, *Bombyx mori*. *FEBS Letters* **464**: 85–90.

Lee, D., Damberger, F. F., Peng, G., Horst, R., Guntert, P. and Nikonova, L. (2002). NMR structure of the unliganded *Bombyx mori* pheromone-binding protein at physiological pH. *FEBS Letters* **531**: 314-318.

Letunic, I., Goodstadt, L., Dickens, N. J., Tobias Doerks, T., Joerg Schultz, J., Richard Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. and Bork, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Research* **30**: 242-244.

Liu, R., Lehane, S., He, X., Lehane, M., Hertz-Fowler, C., Berriman, M., Pickett, J. A., Field, L. M. and Zhou, J. J. (2010). Characterisations of odorant-binding proteins in the tsetse fly *Glossina morsitans morsitans*. *Cellular Molecular Life Science* **67**: 919–929.

Ludwig, M. Z., Uspensky, I. I., Ivanov, A. I., Kopantseva, M. R., Dianov, C. M. Tamarina, N. A. and Korochkin, L. I. (1991). Genetic control and expression of the major ejaculatory bulb protein (PEB-me) in *Drosophila melanogaster*. *Biochemical Genetics* **29**: 215–239.

Madubunyi, L. C., Hassanali, A., Ouma, W., Nyarango, D. and Kabii, J. (1996). Chemoecological role of mammalian urine in host location by tsetse, *Glossina* spp. (Diptera: Glossinidae). *Journal of Chemical Ecology* **22**: 1187-1199.

Maibeche-Coisne, M., Sobrio, F., Delaunay, T., Lettere, M., Dubroca, J., Jacquin-Joly, E. and Nagnan-Le Meillour, P. (1997). Pheromone binding proteins of the moth *Mamestra brassicae*: specificity of ligand binding. *Insect Biochemistry and Molecular Biology* **27**: 213–221.

Maleszka, R. and Stange, G. (1997). Molecular cloning, by a novel approach, of a cDNA encoding a putative olfactory protein in the labial palps of the moth *Cactoblastis cactorum*. *Gene* **202**: 39–43.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. and Bryant, S. H. (2002). CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic acid Research* **30**: 281–

283.

Marone, M., Mozzetti, S., De Ritis, D., Pierelli, L. and Scambia, G. (2001). Semiquantitative RT-PCR analysis to assess the expression levels of multiple transcripts from the same sample. *Biological Procedures Online* **3**: 19–25.

Masiga, D. K., Okech, G., Irungu, P., Ouma, J., Wekesa, S., Ouma, B., Guya, S. O. and Ndung'u, J. M. (2002). Growth and mortality in sheep and goats under high tsetse challenge in Kenya. *Tropical Animal Health Production* **34:** 489–501.

Mattioli, R. C., Pandey, V. S., Murray, M. and Fitzpatrick, J. L. (2000). Immunogenetic influences on tick resistance in African cattle with particular reference to trypanotolerant N'Dama (*Bos Taurus*) and trypanosusceptible Gobra zebu (*Bos indicus*) cattle. *Acta Tropica* **75**: 263–277.

Maudlin, I. (2006). African trypanosomiasis. *Annals of tropical medicine and parasitology* **100**: 679–701.

McKenna, M. P., Hekmat-Scafe, D. S., Gaines, P. and Carlson, J. R. (1994). Putative *Drosophila* pheromone-binding proteins expressed in a subregion of the olfactory system. *Journal of Biological Chemistry* **169**: 16340–16347.

Mehlert, A., Bond, C. S and Ferguson, M. A. J. (2002). The glycoforms of a *Trypanosoma brucei* variant surface glycoprotein and molecular modeling of a glycosylated surface coat. *Glycobiology* **12:** 607–612.

Melo, A. C. A., Rutzler, M., Pitts, R. J. and Zwiebel, L. J. (2004). Identification of a chemosensory receptor from the yellow fever mosquito, *Aedes aegypti*, that is highly conserved and expressed in olfactory and gustatory organs. *Chemical Senses* **29**: 403–410.

Miller, R. T., Alan, G. C., Chella, G., Burke, J., Ptitsyn, A. A., Broveak, T. R. and Hide, W. A. (1999). A Comprehensive Approach to Clustering of Expressed Human Gene Sequence: The Sequence Tag Alignment and Consensus Knowledge Base. *Genome Research* **9**: 1143-1155.

Moloo, S. K. (1971). An artificial feeding technique for *Glossina*. *Parasitology* **63:** 507–512.

Mombaerts, P. (1999). Seven-transmembrane proteins as odorant and chemosensory receptors. *Science* **286**: 707-711.

Mugasa, C. M., Schoone, G. J., Ekangu, R. A., Lubega, G. W., Kager, P. A. and Schallig, H. (2008). Detection of *Trypanosoma brucei* parasites in blood samples using real-time nucleic acid sequence-based amplification. *Diagnostic Microbiology of Infectious Diseases* **61**: 440–445.

Murray, M., Morrison, W. I. And Whitelaw, D. D. (1982). Host susceptibility to African trypanosomiasis: trypanotolerance. *Advanced Parasitology* **21**: 1–68.

Murray, M., D'leteren, G. D. M. and Teale, A. J. (2006). *Trypanotolerance*. The Trypanosomiasis (Ed. By I. Maudlin, Holmes, P.H. and Miles, M. A), pp. 461 – 477.

Nagnan-Le Meillour, P., Cain, A. H., Jacquin-Joly, E., Francois, M. C., Ramachandran, S., Maida, R. and Steinbrecht, R. A. (2000). Chemosensory proteins from the proboscis of *Mamestra brassicae*. *Chemical Sense* **25**: 541–553.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1–6.

Omolo, M. O., Hassanali, A., Mpiana, S., Esterhuizen, J., Lindh, J., Lehane, M. J., Solano, P., Rayaisse, J. B., Vale, G. A., Torr, S. J. and Tirados, I. (2009). Prospects for Developing Odour Baits To Control *Glossina fuscipes* spp., the Major Vector of Human African Trypanosomiasis. *PLoS Neglected Tropical Disease* **3**(5): e435.

Ouma, J. O., Cummings, M. A., Jones, K. C. and Krafsur, E. S. (2003). Characterization of microsatellite markers in the tsetse fly, *Glossina pallidipes* (Diptera: Glossinidae). *Molecular Ecology Notes* **3**: 450–453. doi: 10.1046/j.1471-8286.2003.00480.x.

Ouma, J. O., Marquez, J.G. and Krafsur, E. S. (2005). Macrogeographic population structure of the tsetse fly, *Glossina pallidipes* (Diptera: Glossinidae). *Bulletin of Entomological Research* **95**: 437–447. doi:10.1079/BER2005376

Owaga, M. L. A., Hassanali, A. and McDowell, P. G. (1988). The role of 4-cresol and 3-n-propylphenol in the attraction of tsetse flies to buffalo urine. *Insect Science Application* **9**: 95-100.

Ozaki, K., Utoguchi, A., Yamada, A. and Yoshikawa, H. (2008). Identification and genomic structure of chemosensory proteins (CSP) and odorant binding proteins (OBP) genes expressed in foreleg tarsi of the swallowtail butterfly *Papilio xuthus*. *Insect Biochemistry and Molecular Biology* **38**: 969–976.

Paramasivana, R., Sivaperumalb, R., Dhananjeyana, K. J., Thenmozhia, V. and Brij Kishore Tyagib, B. K. (2007). Prediction of 3-Dimensional Structure of Salivary Odorant-Binding Protein-2 of the Mosquito *Culex Quinquefasciatus*, the Vector of Human Lymphatic Filariasis. *In Silico Biology* **7**: 1–6.

Park, S. K., Shanbhag, S. R., Dubin, A. E., De Bruyne, M., Wang, Q., Yu, P., Shimoni, N., D'Mello, S., Carlson, J. R., Harris, G. L., Steinbrecht, R. A. and Pikielny, C. W. (2002). Inactivation of olfactory sensilla of a single morphological type differentially affects the response of *Drosophila* to odors. *Journal of neurobiology* **51**: 248–260.

Pelosi, P. and Maida, R. (1995). Odorant binding proteins in insects. *Comparative Biochemistry and Physiology* **111B**: 503–514.

Pelosi, P. (1996). Perireceptor events in olfaction. *Journal of Neurobiology* **30**: 3–19.

Pelosi, P., Zhou, J.-J., Ban, L. P. and Calvello, M. (2006). Soluble proteins in Insect chemical communication. *Cellular Molecular Life Sciences* **63:** 1658–1676.

Pellegrino, M. and Nakagawa, T. (2009). Smelling the difference: controversial ideas in Insect Olfaction. *Journal of Experimental Biology* **212:** 1973–1979.

Pelletier, J. and Leal, W. S. (2009) Genome Analysis and Expression Patterns of Odorant-binding Proteins from the Southern House Mosquito *Culex pipiens quinquefasciatus*. *PLoS ONE* **4**(7): e6237.

Picimbon, J. F. and Leal, W. S. (1999). Olfactory soluble proteins of cockroaches. *Insect Biochemistry and Molecular Biology* **29:** 973–978.

Picimbon, J. F., Dietrich, K., Krieger, J. and Breer, H. (2001). Identity and expression pattern of chemosensory proteins in *Heliothis virescens* (Lepidoptera, Noctuidae). *Insect Biochemistry and Molecular Biology* **31**: 1173–1181.

Pikielny, C. W., Hasan, G., Rouyer, F. and Rosbash, M. (1994). Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. *Neuron* **12**: 35–49.

Pollock, J. N. (1982). Training manual for tsetse control personnel. Food and Agriculture Organization of the United Nations.

Prestwich, G. D., Graham, S., Handley, M., Latli, B., Streinz, L. and Tasayco, M. J.

(1989). Enzymatic processing of pheromones and pheromone analogs. *Experientia* **45**: 263–270.

Prestwich, G. D. (1993). Bacterial expression and photoaffinity labeling of a pheromone binding protein. *Protein Science* **2**: 420–428.

Prestwich, G. D., Du, G. and LaForest, S. (1995). How is pheromone specificity encoded in proteins? *Chemical Senses* **20**: 461–469.

Priotto, G., Kasparian, S., Ngouama, D., Ghorashian, S., Arnold, U., Ghabri, S. and Karunakara, U. (2007). Nifurtimox-Eflornithine Combination Therapy for Second-Stage *Trypanosoma brucei gambiense* Sleeping Sickness: A Randomized Clinical Trial in Congo. *Clinical Infectious Diseases* **45**: 1435–1442.

Priotto, G., Kasparian, S., Mutombo, W., Ngouama, D., Ghorashian, S., Arnold, U., Ghabri, S., Baudin, E., Buard, V. and Kazadi-Kyanza, S. (2009). Nifurtimox-eflornithine combination therapy for second-stage African *Trypanosoma brucei gambiense* trypanosomiasis: a multicentre, randomised, phase III, non-inferiority trial. *The Lancet* **374**: 56–64.

Raming, K., Krieger, J. and Breer, H. (1989). Molecular cloning of an insect pheromone binding protein. *FEBS Letters* **356**: 215–218.

Rayaisse, J, B., Tirados, I., Kaba, D., Dewhirst, S,Y., Logan, J.G., Diarrassouba, A., Saloul, E., Omolo, M. O., Solano, P., Lehane, M. J., Pickett, J. A., Vale, G. A., Torr, S. J. and Esterhuizen, J. (2010). Prospects for the Development of Odour Baits to Control the Tsetse Flies *Glossina tachinoides* and *G. palpalis s.l.*. *PloS Neglected Tropical Disease* **4**(3): e632.

Ribeiro, J. M. C., Charlab, R., Pham, V. M., Garfield, M. and Valenzuela, J. (2004). An insight into the salivary transcriptome and proteome of the adult female mosquito *Culex pipiens quinquefasciatus*. *Insect Biochemistry and Molecular Biology* **34**: 543–563.

Robertson, H. M., Martos, R., Sears, C. R., Todres, E. Z., Walden, K. K. O. and Nardi, J. B. (1999). Diversity of odourant binding proteins revealed by an expressed sequence tag project on male *Manduca sexta* moth antennae. *Insect Molecular Biology* **8**: 501–518.

Roelants, G. E., Fumoux, F., Pinder, M., Queval, R., Bassinga A. and Authie, E. (1987). Identification and selection of cattle naturally resistant to African trypanosomiasis. *Acta Tropica* **44**: 55–66.

Rogers, D. J., Hendrickx, G. and Slingenbergh, J. (1994). Tsetse flies and their control. *Rev. Sci. Tech. Off. Int. Epiz* **13**(4): 1075-1124.

Rogers, D. J. and Robinson, T. P. (2006). *Tsetse distribution.* In Maudlin, I., Holmes, P. H. and Miles, M. A. (Eds), The Trypanosomiases, CABI Publishing, UK. Pp. 139 – 179.

Rutzler, M. and Zwiebel, L.J. (2005). Molecular biology of insect olfaction: recent progress and conceptual models. *Journal of Comparative Physiology A.*

Rybczynski, R., Vogt, R. G. and Lerner, M. R. (1990). Antennal-specific Pheromone degrading Aldehyde oxidases from the moths *Antheraea polyphemus* and *Bombyx mori.* The Journal of Biological Chemistry **265**: 19712–19715.

Sabatier, L., Jouanguy, E., Dostert, C., Zachary, D., Dimareq, J. L., Bulet, P. and Imler, J. L. (2003). Pherokine-2 and -3. *European Journal of Biocmistry* **270**: 3398–3407.

Saini, R. K., Hassanali, A., Ahuya, P., Andoke, J. and Nyandat, E. (1993). Close range responses of tsetse flies *Glossina morsitans morsitans* Westwood (Diptera: Glossinidae) to host body kairomones. *Discovery and innovation* **5**: 149–153.

Saini, R. K. and Hassanali A. (2007). 4-Alkyl-substituted Analogue Of Guaiacol Shows Greater Repellency To Savannah Tsetse (Glossina spp.). *Journal of Chemical Ecology* **33**: 985–995.

Sambrook, J. P. and Russell, D. (2001). Molecular Cloning: A Laboratory Manual (Third Edition). Cold Spring Habour Laboratory (CSHL) Press.

Sandler, B. H., Nikonova, L., Leal, W. S. and Clardy, J. (2000). Sexual attraction in the silkworm moth: structure of the pheromone binding protein, bombykol complex. *Chemical Biology* **7**: 143–151.

Sato, K., Pellegrino, M., Nakagawa, T., Vosshall, L. B. and Touhara, K. (2008). Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature* **452**: 1002–1006.

Scaloni, A., Monti, M., Angeli, S. and Pelosi, P. (1999). Structural analysis and disulfide-bridge pairing of two odorant-binding proteins from *Bombyx mori. Biochemical Biophysical Research Communication* **266**: 386–391.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V. and AltschulS. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29:** 2994-3005.

Simarro, P. P., Jannin, J. and Cattand, P. (2008). Eliminating human African trypanosomiasis: Where do we stand and what comes next. *PloS Medicine* **5**(2): e55.

Solano, P., de la Rocque, S., Cuisance, D., Geoffroy, B., de Meeus, T., Cuny, G. and Duvalle, G. (1999). Intraspecific variability in natural populations of *Glossina palpalis gambiensis* from West Africa, revealed by genetic and morphometric analyses. *Medical and Veterinary Entomology* **13**: 401–407.

Spaethe, J. and Briscoe, A. D. (2004). Early Duplication and Functional Diversification of the Opsin Gene Family in Insects. *Molecular Biology and Evolution* **21**: 1583-1594.

Steinbrecht, R. A., Ozaki, M. and Ziegelberger, G. (1992). Immunocytochemical localization of pheromone binding protein in moth antennae. *Cell Tissue Research* **270**: 287–302.

Steinbrecht, R. A., Lave, M. and Ziegelberger, G. (1995). Immunolocalization of pheromone binding protein and general odorant binding protein in olfactory sensilla of the Silk Moths *Antheraea* and *Bombyx*. *Cell and Tissue Research* **282**: 203–217.

Steinbrecht, R. A. (1997). Pore structures in insect olfactory sensilla: A review of data and concepts. *International Journal of Insect Morphology and Embryology* **26**: 229–245.

Steverding, D. (2008). The History of African Trypanosomiasis. *Parasite Vectors* **1**:3.

Stocker, R. F. (1994). The organization of the chemosensory system in *Drosophila melanogaster*. *Cell and Tissue Research* **275**: 3–26.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L,, Nikolskaya, A.N,, Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I,, Yin, J.J. and Natale, D.A. (2003). The COG database: an updated version includes eukaryotes. *BioMed Central Bioinformatics* **4**: 41.

Thomson, M. C. (1987). The effect on tsetse flies (*Glossina* spp.) of deltamethrin applied to cattle. *Tropical Pest Management* **33**: 329–335.

Torr, S. J., Mangwiro, T. N. C. and Hall, D. R. (1996). Responses of *Glossina pallidipes* (Diptera: Glossinidae) to synthetic repellents in the field. *Bulletin of Entomology Research* **86**: 609-616.

Tumlinson, J. H., Brennan, M, M., Doolittle, R. E., Mitchell, E. R., Brabham, A., Mazomenos, B. E., Baumhover, A. H. and Jackson, D. M. (1989). *Archives of Insect Biochemistry and Physiology* **10**: 255–272.

Turner, D. A. (1987). The population ecology of *Glossina pallidipes* Austen (Diptera: Glossinidae) in the Lambwe Valley, Kenya. I. Feeding behaviour and activity patterns. *Bulletin of Entomology Research* **77**: 317–333.

Vale, G. A. and Hall, D. R. (1985). The use of 1-octen-3-ol, acetone and carbon dioxide to improve baits for tsetse flies, *Glossina* spp. (Diptera: Glossinidae). *Bulletin of Entomology Research* **75:** 219–231.

Vale, G. A., Grant, I. F., Dewhurst, C. F. and Aigreau, D. (2004). Biological and chemical assays of pyrethroids in cattle dung to cattle. *Bulletin of Entomology Research* **94**: 273–282.

Valenzuela, J. G., Francischetti, I. M., Pham, V. M., Garfield, M. K. and Ribeiro, J. M. C. (2003). Exploring the salivary gland transcriptome and proteome of the *Anopheles stephensi* mosquito. *Insect Biochemistry and Molecular Biology* **33**: 717–732.

Vogt, R. G. and Riddiford, L. M. (1981). Pheromone binding and inactivation by moth antennae. *Nature* **293:** 161–163.

Vogt, R. G., Riddiford, L. M. and Prestwich, G. D. (1985). Kinetic properties of a sex pheromone-degrading enzyme: the sensillar esterase of *Antheraea polyphemus*. *Proceedings of National Academy of Science USA* **82**: 8827–8831.

Vogt, R. G., Rybczynski, R. and Lerner, M. R. (1990). *The biochemistry of odorant reception and transduction*. In: Chemosensory Information Processing, NATO ASI Series H, Vol. 39. D. Schild, Ed. Springer-Verlag, Berlin, pp. 33 – 76.

Vogt, R. G., Prestwich, G. D and Lerner, M. R. (1991a). Odorant binding protein subfamilies associate with distinct classes of olfactory receptor neurons in insects. *Journal of Neurobiology* **22**: 74–84.

Vogt, R. G., Rybezynski, R. and Lerner, M. R. (1991b). Molecular cloning and sequencing of general odourant binding proteins GOBP1 and GOBP2 from the tobacco hawk moth *Manduca Sexta*; comparison with other insect OBPs and their signal peptides. *Journal of Neuroscience* **11**: 2972–2984.

Vogt, R. G. (2003). *Biochemical diversity of odor detection: OBPs, ODEs and SNMPs*. In G. J. Blomquist and R. G. Vogt. Insect Pheromone Biochemistry and Molecular Biology. Elsevier, Oxford. pp. 391– 445.

Vogt, R. G. (2005). *Molecular basis of pheromone detection in insects.* In L. I. Gilbert, K. Iatro. And S. Gill (eds). Comprehensive insect physiology, biochemistry, pharmacology and molecular biology. Volume 3 Endocrinology. Elsevier, London. pp. 753–804.

Vosshall, L. B., Wong, A. M. and Axel, R. (2000). An olfactory sensory map in the fly brain. *Cell* **102**: 147-159.

Vreysen, M. J. B., Saleh, K. M., Ali, M. Y., Abdulla, A. M., Zhu, Z. K., Juma, k. G., Dyck, V. A., Msangi, A. R., Mkonyi, P. A. and Feldmann, H. V. (2000). *Glossina austeni* (Diptera: Glossinidae) Eradicated on the island of Unguja, Zanzibar, using the sterile insect techniques. *Journal of Economic Entomology* **93**: 123–135.

Warnes, M. I. (1990). The effect of host odour and carbon dioxide on the flight of tsetse flies (*Glossina* spp) in the laboratory. *Journal of Insect Physiology* **36**: 607–611.

Warnes, M. L., Van den Bossche, P., Mudenge, D., Robinson, T. P., Shereni, W. and Chadenga, V. (1999). Evaluation of insecticide-treated cattle as a barrier to re-invasion of tsetse to cleared areas in northeastern Zimbabwe. *Medical Veterinary Entomology* **13**: 177–184.

Wegener, J. W., Hanke, W. and Breer, H. (1997). Second messenger controlled membrane conductance in locust (*Locust migratoria*) olfactory neurons. *Journal of Insect Physiology* **43**: 595–605.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schrimi, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A. and Wagner, L. (2010). Database resources of the National Centre for Biotechnology Information: *Nucleic Acids Research* **38**: D5-D16; doi:10.1093/nar/gkp967.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology Evolution* **18**: 691–699.

Wicher, D., Schafer, R., Bauernfeind, R., Stensmyr, M. C., Heller, R., Heinemann, S. H. and Hansson, B. S. (2008). *Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. *Nature* **452**: 1007–1011.

Wogulis, M., Morgan, T., Ishida, Y., Leal, W. S. and Wilson, D. K. (2006). The crystal structure of an odorant binding protein from *Anopheles gambiae*: evidence for a common ligand release mechanism. *Biochemical Biophysics Research Communication* **339**: 157–164.

Xu, P. X., Zwiebel, L. J. and Smith, D. P. (2003). Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, An*opheles gambiae*. *Insect Molecular Biology* **12**: 549–560.

Xu, P. X., Atkinson, R., Jones, D. N. M. and Smith, D. P. (2005). *Drosophila* OBP LUSH is Required for Activity of Pheromone-Sensitive Neurons. *Neuron* **45**, 193-200.

Zacharuk, R. Y. (1985). *Antennae and sensilla.* In G. A. Kerkut and L. I. Gilbert (Eds). Comprehensive insect physiology biochemistry and pharmacology, Vol. 6, Nervous system: sensory. Pergamon press, Oxford. pp. 1 – 69.

Zhou, J. J., Huang, W., Zhang, G. A., Pickett, J. A. and Field, L. M. (2004). "Plus-C" odorant binding proteins from the southern house mosquito *Culex pipiens quinquefasciatus*. *PloS One* **4**:e6237.

Zhou, J. J., Kan, Y., Antoniw, J., Pickett, J. A. and Field, L. M. (2006). Genome and EST analyses and expression of a gene family with putative functions in insect chemoreception. *Chemical Senses* **31**: 453–465.

Zhou, J.-J., He, X.-L., Pickett, J. A. and Field, L. M. (2008). Identification of odorant-binding proteins of the yellow fever mosquito *Aedes aegypti*: genome annotation and comparative analyses. *Insect Molecular Biology* **17**: 147–163.

Zhou, J.-J., Field, L. M. and He, X. L. (2010). Insect Odorant-Binding Proteins: Do They Offer an Alternative Pest Control Strategy? *Outlooks on Pest Management* **21**, 31-34.

Ziegelberger, G. (1995). Redox-shift of the pheromone binding protein in the silkmoth *Antheraea polyphemus. European Journal of Biochemistry* **232**: 706–711.

APPENDICES

Appendix I - Functional annotation of cDNA clusters from *Glossina pallidipes* antennae library producing best hits to nonredundant (NR) protein database of GenBank, gene ontology (GO) and conserved domains database (CDD) database

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
Odorant/ Pheromone Binding											
265	1	Olfactory Specific F	emb|CAE00444.1|	7.0E-18	*Drosophila erecta*	Pfam PBP/GOBP family	0.029	pheromone binding	3.0E-18	0005550	No
266	2	AgamOBP1	gb|AAO12081.1|	3.0E-30	*Anopheles gambiae*	Pfam PBP/GOBP family	4.0E-19	pheromone binding	1.0E-30	0005550	Yes
Energy Metabolism											
1	1	NADH DH subunit 4	ref|YP_133766.1	0.11		Kog G protein-coupled receptors	0.68				No
20	1	NADH DH subunit 2	ref|NP_694903.1	0.027		Kog Predicted membrane protein	0.071	kinase activity	0.029		No
60	1	cytochrome b	gb|AAT99389.2|	0.074	*Vasdavidius concursus*	Smart Cytochrome b-561	0.059				Yes
107	1	NADH DH subunit 2	ref|YP_271927.1	0.4		Kog UDP-galactose transporter	0.044				No
109	1	GA10478-PA	gb|EAL33178.1|	9.0E-76	*Drosophila pseudoobscura*	Kog Cytochrome c oxidase	8.0E-61	cytoc-c oxidase activity	9.0E-76		No
148	1	NADH DH subunit 5	emb|CAI38858.1|	6.3		Kog Long chain fatty acid elongase	0.095				No
207	1	NADH DH subunit 5	gb|AAK21325.1|	5.0E-06	*Chrysomya putoria*	Pfam NADH DH subunit 5	9.0E-08				No
234	1	NADH DH subunit 6	dbj|BAD90827.1	2.5		Kog Predicted DHHC-type Zn-finger	1.5				Yes
237	1	cytochrome oxidase	gb|AAR11468.1|	0.027	*Hypoderma sinense*	Kog Voltage-gated Ca2+ channels	0.014	sleep	0.002		No
254	1	ENSANGP00000014099	gb|EAA11886.2	4.0E-04		Kog Ca2+/Mg2+-permeable channels	0.002	NADH DH activity	3.0E-05		No
271	1	NADH DH subunit 2	ref|NP_694903.1	0.076		Pfam Protein of unknown function	0.034				No
Transcription Factors											
11	1					Smart DNA Topoisomerase IV	0.093				No
40	1	transcription initiation F	ref|XP_629530.1	0.025		Smart Ribonuclease III family	0.28				Yes
44	1	receptor-like pro kinase	ref|NP_914396.1|	0.034	*Oryza sativa*	Kog Translation initiation factor 3	0.11				Yes
48	1	SJCHGC01957 protein	gb|AAX27763.1|	0.003	*Schistosoma japonicum*	Smart Ribosomal protein L11/L12	0.024				Yes
55	1	hypothetical protein	ref|XP_642344.1	0.027		Kog Alpha-1,2 glucosyltransferase	0.015	mRNA processing	0.052		Yes
90	1	Friedlin nascent p/peptide	emb|CAC22620.1	0.001		Kog Uncharacterized conserved pro	0.072	nucleic acid binding	5.0E-04	0003677	No
97	1	unknown	gb|AAX27955.1|	1.0E-07	*Schistosoma japonicum*	Kog Ribosome biogenesis protein	2.0E-05	snRNA binding	5.0E-05	0017069	No
104	1	TonB-dependent receptor	gb|AAZ44891.1	4.4		Kog RNA polymerase II	0.075				No
136	1	hybrid proline-rich protein	gb|AAA33132.1|	6.0E-05		Kog 5'-3' exonuclease HKE1	1.0E-04	DNA binding	3.0E-04		Yes

Appendix I
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
167	1					Smart tRNA-adenosine deaminase	0.091				No
289	1					Smart Histone 2A	0.11				No
290	1	unnamed protein	dbj|BAA97098.1|	5.0E-06	*Arabidopsis thaliana*	Kog Translation initiation factor	9.0E-04	embryonic/larval devmnt	2.0E-04	0003723	No
296	1	TFIID subunit	ref|XP_629530.1	3.4		Pfam Domain of unknown function	0.04				No
Cytoskeletal											
8	1	TonB, C-terminal	gb|AAZ48780.1	0.004		Kog Histone H1	0.025	locomotory behavior	0.023		No
54	1	unnamed protein product	ref|XP_503326.1	9.5		Smart bacterial histone like domain	0.054				No
69	1	guanylyl cyclase	emb|CAB52252.1|	9	*Tetrahymena pyriformis*	Smart Ras-like GTPases	0.005				No
70	1	Serpentine Receptor	gb|AAC25815.1	2		Smart LITAF	0.16				No
71	1					Smart Four-disulfide core domains	0.4				No
83	1	Nucleoporin Nup205	ref|XP_416176.1|	10	*Gallus gallus*	Pfam BT1 family	0.15				No
108	1	Cleavage factor 6	gb|AAQ97763.1	0.061		Smart Protein phosphatase 2A	0.078				No
116	1	Serpentine receptor	ref|NP_504379.1	2.9		Smart DNA polymerase type-B family	0.4				No
118	1	sperm autoantigenic protein	gb|AAH75539.1	4.4		Kog RNA polymerase II complex	0.68				No
127	1	hypothetical protein	emb|CAB96200.1|	4. 0E-07	*Capsella rubella*	Pfam CAP protein	4.0E-04	eggshell formation	1.0E-06	0003779	No
130	1					Smart Histone 2A	1.8				No
131	1	Huntington disease gene	ref|XP_573634.1|	4.0E-07	*Rattus norvegicus*	Kog RhoA GTPase effector	0.003	spliceosome assembly	2.0E-08	0003676	No
168	1	expressed protein	gb|AAT85077.1|	6.0E-06	*Oryza sativa*	Smart Cytochrome b-561	0.12	contractile ring formation	5.0E-06	0003779	No
178	1	SJCHGC09076 protein	gb|AAW26562.1|	0.027	*Schistosoma japonicum*	Kog Uncharacterized conserved protein	0.024	Maintenance of chromatin	0.088		No
179	2	SJCHGC09076 protein	gb|AAW26562.1|	8.0E-15	*Schistosoma japonicum*	Kog RNA polymerase II C-terminal domain	0.005	cartilage condensation	7.0E-04	0005201	No
187	1	hypothetical protein	gb|EAA75452.1	2.0E-06		Kog WASP-interacting protein	4.0E-04	contractile ring formation	1.0E-06	0003779	No
202	1	unnamed protein	emb|CAG02796.1|	9.0E-07	*Tetraodon nigroviridis*	Pfam CAP protein	2.0E-04	axonogenesis	2.0E-05	0003779	No
210	1					Kog DHHC-type Zn-finger proteins	1.8				No
213	1	formin homology protein B	dbj|BAC16797.1|	7.0E-08	*Dictyostelium discoideum*	Kog Adenylate cyclase-protein	0.01	cytoskeletal protein binding	1.0E-06	0008092	No
215	1	Hypothetical protein	emb|CAE58429.1|	1.0E-06	*Caenorhabditis briggsae*	Kog WASP-interacting protein	7.0E-04	structural cytoskeleton compt	3.0E-07	0005200	No
235	1	hypothetical protein	gb|AAU28478.1	0.51		Pfam Domain of unknown function	0.006	Microtubule complex	0.04		No
258	1					Smart Acidic/basic fibroblast factor	0.16				No

Appendix I
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
Transporters											
3	1					Kog mannosyltransferase family	1.8				No
5	1	cytochrome b	gb|AAB91353.1|	0.15	*Dennyus d. distinctus*	Kog Na+:Ca2+ antiporter	0.29				No
15	1	hypothetical protein	ref|XP_646009.1	0.018		Pfam 7TM chemoreceptor	0.22				No
22	1					Kog Ferric reductase	0.098				No
42	1	ADP/ATP translocase	gb|AAF32322.1|	1.0E-58	*Lucilia cuprina*	Kog Mith ADP/ATP carrier proteins	1.0E-47	ATP:ADP antiporter	3.0E-59	0015207	No
43	2	aspartic acid-rich protein	emb|CAA07355.1|	0.003	*Plasmodium falciparum*	Kog Secretory carrier membrane protein	1.1				No
49	1	SAM-methyltransferase	emb|CAG21321.1	6.7		Kog Nuclear transport receptor	0.24				No
52	1	CWC22	gb|ABA27413.1|	0.37	*Bigelowiella natans*	Smart Ras-like GTPases	0.22				No
56	1	unnamed protein product	dbj|BAC87585.1|	6.0E-07	*Homo sapiens*	Pfam ATP synthase A chain	0.021				No
75	1	SJCHGC01957 protein	gb|AAX27763.1|	4.0E-07	*Schistosoma japonicum*	Kog Ferric reductase-like proteins	0.001	odontogenesis	0.039		No
91	1					Smart rRNA adenine dimethylases	1.6				No
92	1	phosphatase	ref|XP_728980.1|	0.016		Kog Na+/K+ ATPase	0.06				No
98	1					Kog Cl- channel CLC-7	0.61				No
139	1					Kog LDL B-like protein	0.5				No
146	1	CG11739-PC, isoform C	gb|AAL90157.1	8.0E-05		Pfam Tricarboxylate carrier	4.0E-07	tricarboxylate carrier	3.0E-06		No
152	1	variola B22R-like protein	ref|NP_955149.1	7.3		Kog Cytochrome P450	0.3				No
166	1	unnamed protein	dbj|BAC86300.1|	0.007	*Homo sapiens*	Kog Ferric reductase-like proteins	0.43				No
208	1	ear1 protein-like	dbj|BAD28171.1	2.7		Kog Ca2+/Mg2+-permeable channels	0.63				No
221	1	Zea mays permease 1	gb|AAB60909.1|	0.51	*Arabidopsis thaliana*	Pfam Ribosomal prokaryotic protein	0.17				No
224	1	succinyl-CoA synthetase	ref|NP_700961.1	0.16		Kog Permease-major facilitator	0.048				No
260	1	hypothetical protein	ref|NP_705155.1	0.002		Kog Protein transporter-TRAM	7.0E-04	K^+ transporter activity	0.018		No
Signal transduction											
6	1	Unknown protein	gb|AAH96212.1|	6.0E-04	*Homo sapiens*	Kog RhoA GTPase effector	1.0E-04	nucleic acid binding	2.0E-04	0003723	Yes
9	3	Salivary proline protein	sp|P04280|PRP1	1.0E-16		Kog RhoA GTPase effector	8.0E-11	structural molecule activity	5.0E-18	0005201	No
10	1	unnamed protein product	emb|CAG02796.1|	9.0E-06	*Tetraodon nigroviridis*	Kog RhoA GTPase effector	0.11	nucleic acid binding	6.0E-04	0003723	No
16	4	receptor type 1	gb|AAR12889.1|	8.2		Kog Signaling protein	1.7				No
24	1	CG3983-PB, isoform B	ref|NP_732199.2	9.0E-35		Kog GTPase	2.0E-24	hydrolase activity	5.0E-15	0003924	No

Appendix I Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
36	1	similar to Nebulin	ref|XP_851603.1|	4.3	*Canis familiaris*	Pfam Diaphanous GTPase-binding	0.17				No
41	1	Adenylate cyclase 8	ref|XP_418437.1|	5	*Gallus gallus*	Smart G protein alpha subunit	0.12				No
80	1	unknown	gb|AAX27955.1|	1.0E-12	*Schistosoma japonicum*	Kog RhoA GTPase effector	1.0E-06	ISG15 carrier activity	2.0E-05	0005522	No
85	3	pherophorin-dz1 protein	emb|CAD22154.1|	2.0E-62	*Volvox nagariensis*	Kog RhoA GTPase effector	7.0E-39	dense fibrillar component	2.0E-37	0003723	No
93	1	Wiskott-Aldrich syndrome	ref|NP_001001206.1	0.45		Kog Rac1 GTPase effector	0.063				No
138	1	Protein kinase	ref|ZP_00769315.1	2.0E-06		Kog RhoA GTPase effector	2.0E-07	transcription factor activity	7.0E-06	0003779	No
142	1	unnamed protein	emb|CAG08643.1|	1.0E-11	*Tetraodon nigroviridis*	Kog RhoA GTPase effector	8.0E-09	ribonucleoprotein complex	1.0E-07	0003723	No
143	1	Hypothetical protein	emb|CAE70686.1|	1.0E-08	*Caenorhabditis briggsae*	Kog RhoA GTPase effector	2.0E-09	collagen type V	1.0E-08	0003723	No
165	1	formin homology protein	dbj|BAC16796.1|	2.0E-06	*Dictyostelium discoideum*	Kog RhoA GTPase effector DIA	1.0E-07		4.0E-06	0003924	No
176	1	GA20725-PA	gb|EAL31318.1|	6.0E-10	*Drosophila pseudoobscura*	Kog CDP-diacylglycerol synthase	9.0E-08	rhodopsin mediated signal	2.0E-11		No
177	1	NAD(+) synthase	gb|ABA04487.1	0.4		Kog guanine/nucleotide exchange	0.93				No
182	2	mannose-6-phosphate	gb|AAO34909.1	2.1		Kog Protein tyrosine phosphatase	0.059				No
189	1	formin	gb|EAN92551.1	2.0E-07		Kog Wiskott Aldrich syndrome	0.005	GTPase regulator activity	6.0E-07	0003676	No
195	1	Hypothetical protein	emb|CAE69130.1|	3.0E-05	*Caenorhabditis briggsae*	Kog Adenylate cyclase protein	0.002	GTPase regulator activity	5.0E-05	0008321	No
196	2	capsid associated protein	ref|NP_848457.1	1.0E-06		Pfam CAP protein	8.0E-05	GTPase regulator activity	3.0E-07	0008321	No
199	1	regulatory protein	gb|AAA33306.1|	0.002	*Emericella nidulans*	Kog RhoA GTPase effector	1.0E-04	actin filament bundle	0.002	0003676	No
200	1	diaphanous-related formin	ref|XP_636106.1	3.0E-06		Kog RhoA GTPase effector	0.001	ISG15 carrier activity	6.0E-05	0005522	No
275	1	putative membrane protein	ref|YP_294122.1	1.0E-24		Kog RhoA GTPase effector	7.0E-19	dense fibrillar component	1.0E-16	0003723	No
276	1	putative membrane protein	ref|YP_293961.1	7.0E-16		Kog RhoA GTPase effector	1.0E-17	Rac protein signal	3.0E-15	0003723	No
278	1	ferlin OHproline/rich	emb|CAH97750.1	5.6		Smart G protein alpha subunit	0.027				No
287	1	glycoprotein	emb|CAB62280.1|	1.0E-20	*Volvox c. f. nagariensis*	Kog RhoA GTPase effector DIA	2.0E-20	N/A binding	4.0E-17	0003723	No
294	1	PELP1	gb|AAC17708.2|	2.0E-07	*Homo sapiens*	Kog RhoA GTPase effector	3.0E-08	embryonic/ larval develop	2.0E-07	0003723	No
Protein Function											
2	2					Smart Protein phosphatase	0.058				No
12	1					Smart c4 zinc finger	0.2				No
14	1					Smart Interleukin-7	0.92				No
37	1					Smart G protein alpha subunit	0.52				No

Appendix I Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
57	1	hypothetical protein	ref|NP_700707.1	2.0E-07		Pfam 7TM chemoreceptor	2.0E-05	ubiquitin-protein ligase	1.0E-04	0003677	No
61	1					Smart LITAF	0.24				No
62	2	SJCHGC01957 protein	gb|AAX27763.1|	0.028	*Schistosoma japonicum*	Pfam Geminivirus AL3 protein	0.087				No
68	1	SJCHGC01957 protein	gb|AAX27763.1|	2.0E-04	*Schistosoma japonicum*	Kog Predicted RNA methylase	0.025				Yes
76	3	unknown	gb|AAX27955.1|	3.0E-21	*Schistosoma japonicum*	Kog Ribosome biogenesis protein	4.0E-10	Ca^+/K^+ channel activity	7.0E-10	0005267	No
87	1	proenkephalin	gb|AAU95755.1|	0.1	*Bombina orientalis*	Pfam Vert endogenous opioids	0.086	perception of pain	0.008		No
89	1	unknown	gb|AAX27955.1|	5.0E-10	*Schistosoma japonicum*	Kog Telomerase elongation inhibitor	6.0E-06	Embryonic morphogenesis	3.0E-04	0003702	No
95	1	GH01724p	ref|NP_610284.1	4.0E-35		Kog Protein tyrosine phosphatase	1.0E-35	Golgi biogenesis	7.0E-24		No
121	1	formin homology protein A	dbj|BAC16796.1|	2.0E-07	*Dictyostelium discoideum*	Pfam CAP protein	7.0E-04	Lamellipodium / actin	2.0E-06	0003677	No
125	1	SJCHGC01964 protein	gb|AAX24673.2|	0.008	*Schistosoma japonicum*	Pfam Acyltransferase family	0.16				No
134	1	SPAPB15E9.02c	ref|NP_001018275.1	0.015		Kog Molecular chaperone	0.008				No
141	1	Glycosyl transferase	gb|ABB32237.1	9.7		Kog spliceosome subunit	0.44				No
144	1	TLL2 protein	gb|AAH13871.1|	0.007	*Homo sapiens*	Kog CBF1-interacting corepressor	0.098				No
145	1					Smart Glycosyl hydrolase family 10	0.69				No
155	1	super cysteine rich protein	gb|AAB05810.1|	1.0E-04	*Homo sapiens*	Pfam Metallothionein	0.005				No
157	1	Sec14d containing protein	gb|EAK90545.1	9.5		Smart endonuclease III	0.056				No
158	1	SJCHGC01957 protein	gb|AAX27763.1|	0.22	*Schistosoma japonicum*	Pfam Arginine-tRNA-protein transferase	0.031				No
162	1					Smart Glycosyl hydrolases family 32	2.1				No
163	1	unknown	gb|AAX27911.1|	2.0E-06	*Schistosoma japonicum*	Pfam DUP family	0.74				No
169	1	unknown	gb|AAX27911.1|	2.0E-06	*Schistosoma japonicum*	Kog CBF1-interacting corepressor	0.004	odontogenesis	0.03		No
170	1					Pfam Sre, C	0.29				No
175	1					Smart E3 ubiquitin-protein ligases	1.4				No
181	1	hypothetical protein	gb|AAH49534.1	3.0E-04		Smart LITAF	0.003	neurotransmitter secretion	0.008		No
183	1	ENSANGP00000025094	gb|EAA43936.2	3.0E-05		Kog finger-containing proteins	0.027				No
185	1					Smart Presenilin	0.04				No
191	1	SJCHGC01957 protein	gb|AAX27763.1|	0.016	*Schistosoma japonicum*	Pfam Acetyl co-enzyme A carboxylase	0.65				Yes
192	1					Kog Permease of the major facilitator	0.25				No

Appendix I
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
203	1					Smart DNA binding domain-helix-turn	0.36				No
204	1	KIAA1657 protein	dbj|BAB33327.1|	0.009	*Homo sapiens*	Kog Chitin synthase/hyaluronan synthase	0.055				No
206	1	reverse transcriptase	emb|CAI96711.1|	3.2	*Aedes aegypti*	Pfam Tryptophan/tyrosine permease	0.46				No
209	1	TcC31.7	gb|AAC14069.1|	0.23	*Trypanosoma cruzi*	Smart prohibitin homologues	0.14				No
212	1					Pfam Surfeit locus protein 2	0.18				No
222	1	ENSANGP00000016593	gb|EAA00286.3	0.002		Smart Tissue inhibitor-metalloproteinase	0.44	sulfate permease	1.0E-.04		No
223	1					Smart RIO-like kinase	0.48				No
225	1	phosphoribosyltransferase	gb|AAO35570.1	5.1		Smart GTPase-activator protein	0.15				No
228	1					Pfam Poxvirus P4B major core protein	0.16				No
231	1	similar to Furin precursor	ref|XP_850069.1|	1.5	*Canis familiaris*	Smart Ly-6 antigen / uPA receptor	0.18				No
233	1					Pfam Envelope glycoprotein	0.33				No
239	1	Zgc: 55396 protein	gb|AAH71500.1|	0.003	*Danio rerio*	Smart LITAF	0.042				Yes
241	1	WRKY11	gb|AAQ20911.1|	0.009	*Oryza sativa*	Pfam Cytochrome C oxidase subunit II	0.12				No
242	1	putative chaperonin	ref|XP_730128.1	0.012		Smart basic region leucin zipper	0.019				No
244	1	Uba1 gene product-related	ref|XP_729708.1	0.006		Kog Putative receptor CCR1	0.024				No
250	1	putative chaperonin	ref|XP_730128.1	0.007		Pfam Borrelia ORF-A	4.0E-04				No
267	2	condensin component cnd2	pir||T49494	0.88	*Neurospora crassa*	Pfam Protein of unknown function	0.003				No
273	1	major surface glycoprotein	emb|CAC43461.1|	1.2	*Pneumocystis carinii*	Pfam 4-OHphenylacetate 3-OHlase	0.17				No
277	1	similar to Torsin family 1	ref|NP_001034286.1	1.5		Pfam Rotavirus VP3 protein	0.055				No
279	1					Smart Phosphoinositide 3-kinase	0.79				No
280	1	KIAA1657 protein	dbj|BAB33327.1|	1.1	*Homo sapiens*	Pfam Papillomavirus E5	0.025				No
288	1					Pfam SCAMP family	1.8				Yes
292	1	ClpC/MecB	emb|CAD74986.1	3.3		Smart Low MW phosphatase family	0.35				No
295	1					Smart bacterial periplasmic substrate	0.27				No
Unknown											
7	1	unknown	gb|AAX27955.1|	0.003	*Schistosoma japonicum*	Kog Actin filament-binding protein	0.17				Yes
13	1					Pfam Poxvirus serine/threonine	0.024				No
18	1	unnamed protein	dbj|BAC86300.1|	0.025	*Homo sapiens*	Kog Endocytosis/signaling protein	1.8				No

Appendix I
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
23	1	unknown	gb|AAX27911.1|	0.52	*Schistosoma japonicum*	Kog Cytochrome oxidase subunit III	0.3				No
27	2	putative chaperonin	ref|XP_730128.1	0.07		Kog Uncharacterized conserved protein	0.097				No
29	2	putative chaperonin	ref|XP_730128.1	0.048		Smart hypothetical protein domain	0.19				No
31	1	putative chaperonin	ref|XP_730128.1	0.004		Pfam Domain of unknown function	0.006				No
39	1					Pfam Protein of unknown function	0.38				No
46	1	unknown	gb|AAX27911.1|	0.005	*Schistosoma japonicum*	Pfam LacY proton/sugar symporter	0.068				Yes
65	1					Kog Predicted membrane protein	0.59				No
66	1	unnamed protein	ref|XP_454192.1	0.095		Smart Ras-Guanine nucleotide factor	0.1				No
67	1	unnamed protein	ref|XP_502130.1	2.0E-06		Kog Ferric reductase	5.0E-04				No
72	1	unknown	gb|AAX27911.1|	0.083	*Schistosoma japonicum*	Kog Cytochrome P450	0.027				No
81	1	unknown	gb|AAX27955.1|	0.008	*Schistosoma japonicum*	Kog Secretory carrier membrane protein	0.56				Yes
99	1	unknown	ref|NP_704103.1	0.012		Kog Ca2+/Mg2+-permeable channels	0.048				No
101	2	unnamed protein product	dbj|BAC87052.1|	0.002	*Homo sapiens*	Smart Glycosyl hydrolases family 32	0.3				No
106	1	unnamed protein	dbj|BAC87430.1|	0.89	*Homo sapiens*	Smart Syntaxin N-terminal domain	0.14				No
114	1	unnamed protein	dbj|BAC87575.1|	0.004	*Homo sapiens*	Smart Putative GTP-ase	0.013				No
115	1	unnamed protein	ref|NP_059419.1	1.1		Pfam CHD5-like protein	0.094				No
123	1	unnamed protein	emb|CAG10585.1|	3.3	*Tetraodon nigroviridis*	Pfam recA bacterial DNA recom protein	0.13				No
124	1					Kog WD40 repeat-containing protein	0.35				No
126	1	unnamed protein	dbj|BAC86958.1|	0.11	*Homo sapiens*	Kog Uncharacterized conserved protein	0.043				No
150	1					Kog Predicted membrane protein	1.9				No
151	1	predicted protein	ref|XP_366618.1|	1.3	*Magnaporthe grisea*	Pfam Protein of unknown function	0.19				No
159	1	unnamed protein	dbj|BAC87575.1|	5.0E-15	*Homo sapiens*	Pfam Sre, C	0.007				No
160	1	unknown	gb|AAX27955.1|	0.053	*Schistosoma japonicum*	Smart PRE_C2HC, PRE_C2H2 domain	0.28				No
171	1	SJCHGC09731 protein	gb|AAX31025.1|	3.3	*Schistosoma japonicum*	Pfam Uncharacterised protein family	0.14				No
172	1					Pfam Protein of unknown function	1.5				No
173	1	unknown	gb|AAX27911.1|	0.1	*Schistosoma japonicum*	Kog Endosomal membrane proteins	0.042				No
180	1					Pfam Uncharacterised protein family	0.28				Yes

Appendix I
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
194	1	unnamed protein	dbj|BAC86300.1|	0.005	*Homo sapiens*	Kog Predicted lipoprotein	0.005				No
197	1	unnamed protein	dbj|BAC32804.1|	0.036	*Mus musculus*	Kog Sodium/hydrogen exchanger protein	0.15				No
201	1	unnamed protein	dbj|BAC04995.1|	0.021	*Homo sapiens*	Pfam Sec-indep protein translocase	0.016				No
211	1					Pfam Poxvirus of unknown function	0.28				No
216	1	WRKY11	gb|AAQ20911.1|	0.53	*Oryza sativa*	Kog Uncharacterized PilT protein	0.035				No
217	1					Smart VPS10 domain	0.48				No
229	1	unnamed protein	dbj|BAC86276.1|	5.8	*Homo sapiens*	Smart CLUSTERIN Beta chain	0.61				No
230	1	unnamed protein	dbj|BAC86178.1|	0.22	*Homo sapiens*	Smart Presenilin	0.005				No
245	1	unnamed protein	dbj|BAC86300.1|	0.006		Kog Uncharacterized conserved protein	0.13				Yes
259	1	SJCHGC01957 protein	gb|AAX27763.1|	0.046	*Schistosoma japonicum*	Pfam hypo plant mitochondrial protein	0.071				Yes
263	1	unnamed protein	dbj|BAC86958.1|	0.002	*Homo sapiens*	Pfam Sre, C	0.08				No
268	1					Pfam Protein of unknown function	1.3				No
274	1					Smart Ribosomal protein	0.79				No
281	1	SJCHGC03128 protein	gb|AAX26662.2|	8.0E-10	*Schistosoma japonicum*	Kog RhoA GTPase effector	4.0E-08				No
283	1	unnamed protein	dbj|BAC86958.1|	0.004	*Homo sapiens*	Smart lipopolysaccharide enzyme	0.11				No
Hypothetical											
4	1	hypothetical protein	ref|XP_675398.1	0.67		Kog Golgi-associated protein	0.13				No
17	4	hypothetical protein	ref|NP_473073.2	0.004		Pfam YMF19 plant mitochondrial protein	0.056				No
19	1	hypothetical protein	ref|XP_727001.1	0.03		Pfam Cyt c oxidase subunit III	0.085				Yes
21	1	hypothetical protein	ref|XP_644272.1	0.1		Kog P-type ATPase	0.19				No
25	1	hypothetical protein	ref|XP_646781.1	0.009		Smart Ubiquitin-conjugating enzyme E2	0.056				No
26	1	hypothetical protein	ref|XP_679980.1	5.7		Kog Uncharacterized conserved protein	0.047				No
28	1	hypothetical protein	ref|NP_703298.1	1.5		Kog ubiquitin ligase	0.41				No
30	1	hypothetical protein	emb|CAC13617.1	2.3		Pfam Protein of unknown function	0.11				No
32	1	hypothetical protein	ref|NP_701505.1	7.4		Pfam Major intrinsic protein	0.33				No
33	1	hypothetical protein	ref|XP_642253.1	0.013		Pfam LacY proton/sugar symporter	0.042				No
34	1	hypothetical protein	ref|XP_638906.1	9.0E-04		Smart Integrin beta subunits	0.063				No
35	1	hypothetical protein	ref|XP_645450.1	3.3		Pfam LacY proton/sugar symporter	0.77				No

Appendix I
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
38	2	hypothetical protein	ref|XP_643556.1	7.0E-04		Kog Secretory carrier membrane protein	0.078				No
45	1	hypothetical protein	emb|CAH86855.1	0.3		Pfam Domain of unknown function	0.089				No
47	1	hypothetical protein	ref|NP_701637.1	5.7		Smart Glycosyl hydrolase family 10	1.6				No
50	1	hypothetical protein	ref|XP_643102.1	0.02		Kog Alpha-1,2 glucosyltransferase	0.015				No
51	1	hypothetical protein	gb|EAN32317.1	7.4		Smart G protein alpha subunit	0.28				No
53	1	hypothetical protein	ref|XP_670256.1	0.51		Pfam 7TM chemoreceptor	0.053				No
58	1	hypothetical protein	ref|XP_640040.1	0.016		Smart Presenilin	0.2				No
59	1	hypothetical protein	gb|AAY61752.1	2.1		Pfam Domain of unknown function	2.0E-05				No
63	1	hypothetical protein	ref|XP_638906.1	0.018		Pfam Rhabdovirus matrix protein	0.16				No
64	1	hypothetical protein	ref|XP_643681.1	0.013		Smart Presenilin	0.036				Yes
73	2	hypothetical protein	pir||T31613	9.0E-06	*Caenorhabditis elegans*	Kog Lipid exporter ABCA1	4.0E-06				No
74	1	hypothetical protein	gb|EAN94557.1	5.8		Pfam LacY proton/sugar symporter	0.29				No
78	1	hypothetical protein	ref|NP_702127.1	0.092		Kog RNA-binding protein	0.045				Yes
79	1	hypothetical protein	ref|XP_646296.1	0.059		Kog 60S ribosomal protein	0.17				No
82	1	hypothetical protein	ref|XP_644272.1	0.13		Smart Homeodomain	0.099				Yes
84	1	hypothetical protein	gb|EAN93140.1	5.8		Smart NEAr Transporter domain	0.044				No
86	1	hypothetical protein	ref|NP_700707.1	0.001		Pfam Baculovirus of unknown function	0.015	nucleic acid binding	0.086		No
88	1	hypothetical protein	ref|XP_646201.1	0.003		Kog Uncharacterized conserved protein	0.023				No
94	1	hypothetical protein	pir||E22845	4.0E-05	*Trypanosoma brucei*	Kog Transporter, ABC superfamily	0.004				No
96	1	hypothetical protein	ref|NP_700669.1	0.048		Pfam YMF19 plant mitochondrial protein	0.035				No
100	1	hypothetical protein	ref|XP_642344.1	0.26		Pfam 7 transmembrane receptor	0.29				No
102	1	hypothetical protein	ref|XP_645969.1	9.9		Smart hypothetical Domains - Drosophila	0.46				No
103	1	hypothetical protein	ref|XP_643102.1	0.022		Pfam Fibronectin-binding protein	0.45				No
105	1	hypothetical protein	ref|NP_700715.1	3.3		Smart DNA polymerase type-B family	0.14				No
110	1	hypothetical protein	ref|ZP_00738855.1	1.5		Smart Broad-Complex	0.42				No
111	1	hypothetical protein	ref|XP_742832.1	5.7		Kog mannosyltransferase family	0.2				No
112	1	hypothetical protein	ref|NP_473342.1	1.1		Smart STE like transcription factors	0.63				No
113	1	hypothetical protein	gb|EAN87037.1	2		Kog C-4 sterol methyl oxidase	0.27				No

Appendix 1 Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
117	1	hypothetical protein	emb|CAI76869.1	7		Smart Presenilin	0.075				No
119	1	hypothetical protein	ref|NP_703238.1	3.7		Smart Ras-Guanine nucleotide factor	0.39				No
120	1	hypothetical protein	ref|XP_719067.1	0.29		Kog Uncharacterized conserved protein	0.001				No
122	1	hypothetical protein	ref|XP_637070.1	0.12		Kog Uncharacterized conserved protein	0.029				No
128	1	hypothetical protein	ref|NP_703246.1	0.87		Kog Predicted mitochondrial protein	1.4				No
129	1	Hypothetical protein	gb|AAO52249.1|	0.39	*Dictyostelium discoideum*	Kog Polytopic membrane protein	0.43				No
132	1	hypothetical protein	ref|XP_640208.1	0.001		Kog Histone H1	0.077				No
133	1	hypothetical protein	ref|XP_643102.1	0.88		Pfam Protein of unknown function	0.18				No
137	1	hypothetical protein	gb|AAA96601.1	0.011		Kog Ceramide glucosyltransferase	0.013				No
140	2	hypothetical protein	ref|XP_674063.1	0.61		Smart TLC/TRAM/LAG1 domains	0.034				No
147	1	hypothetical protein	ref|XP_415764.1|	5.7	*Gallus gallus*	Pfam Ribosomal protein S8e	1.5				No
153	1	hypothetical protein	ref|NP_703576.1	3.3		Smart Ras-Guanine nucleotide factor f	0.09				No
154	1	hypothetical protein	ref|NP_702771.1	4.4		Smart Presenilin	0.064				No
156	1	hypothetical protein	ref|XP_737285.1	2.3		Pfam Chordopoxvirus G3 protein	0.23				No
161	1	hypothetical protein	ref|XP_641754.1	0.32		Kog Acetylcholine receptor	0.28				No
164	1	hypothetical protein	gb|AAM67677.1	0.13		Kog Ribosomal protein	0.048				No
174	1	hypothetical protein	ref|XP_646911.1	0.18		Smart LITAF	0.12				No
184	1	hypothetical protein	ref|NP_703819.1	0.4		Pfam Yip1 domain	0.11				No
188	1	hypothetical protein	ref|XP_644272.1	2.2		Smart Alpha-amylase domain	0.61				No
190	1	Hypothetical protein	emb|CAE74867.1|	3.6	*Caenorhabditis briggsae*	Pfam Sec-independent protein translocase	0.058				No
193	1	hypothetical protein	pir||T31613	9.0E-06	*Caenorhabditis elegans*	Kog Uncharacterized conserved protein	0.011				Yes
198	1	hypothetical protein	ref|XP_642369.1	0.51		Kog Predicted transporter	0.37				No
186	1	hypothetical protein	ref|ZP_00367506.1	1.9		Smart Presenilin	0.13				No
205	1	hypothetical protein	ref|NP_701331.1	9.8		Smart VPS10 domain	0.33				No
214	1	hypothetical protein	ref|XP_640208.1	0.027		Smart TLC/TRAM/LAG domains	0.13				No
218	1	hypothetical protein	ref|XP_636844.1	0.005		Smart Presenilin	0.009				No
220	1	hypothetical protein	ref|XP_638906.1	0.45		Pfam 7TM chemoreceptor	0.002				No
232	1	hypothetical protein	ref|YP_456730.1	3.3		Kog Uncharacterized conserved protein	0.15				No

Appendix I Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
236	1	hypothetical protein	gb|EAN85835.1	0.028		Smart Presenilin	0.077				Yes
238	1	hypothetical protein	ref|XP_644272.1	0.048		Kog RNA-directed RNA polymerase	0.47				Yes
240	1	hypothetical protein	ref|XP_643556.1	0.009		Pfam 7TM chemoreceptor	0.06				No
243	1	hypothetical protein	ref|XP_638498.1	0.035		Kog CBF1-interacting corepressor	0.021				No
246	1	hypothetical protein	ref|XP_640040.1	0.046		Kog Predicted alpha/beta hydrolase	0.035				No
247	1	hypothetical protein	ref|XP_646009.1	0.016		Smart Formin Homology 2 Domain	0.014				Yes
248	1	hypothetical protein	ref|XP_635821.1	0.02		Pfam Rifin/stevor family	0.072				No
249	1	hypothetical protein	gb|EAN93140.1	0.067		Pfam Vps52 / Sac2 family	0.026				No
251	1	hypothetical protein	ref|XP_644272.1	2.0E-04	*Homo sapiens*	Smart N-terminal to some SET domains	0.26				Yes
253	1	hypothetical protein	ref|NP_113368.1	3.3		Kog Uncharacterized conserved protein	0.27				No
255	1	hypothetical protein	ref|XP_645982.1	0.009		Kog Uncharacterized conserved protein	0.007				Yes
256	1	hypothetical protein	gb|EAN86251.1	0.078		Kog Predicted RNA methylase	0.044				No
261	1	hypothetical protein	gb|EAL87103.1	5.4		Smart TopoisomeraseII	0.13				No
264	1	hypothetical protein	emb|CAH86969.1	2.0E-10		Kog finger-containing proteins	0.003				No
269	1	hypothetical protein	emb|CAH87320.1	4.4		Smart Galectin - galactose-binding lectin	0.22				No
270	1	hypothetical protein	ref|XP_666801.1	7.4		Kog Sodium/hydrogen exchanger protein	0.56				No
282	1	hypothetical protein	ref|XP_644272.1	0.13		Pfam Cytidylyltransferase family	0.015				No
284	1	hypothetical protein	ref|XP_728229.1	0.3		Pfam Borrelia ORF-A	0.001				No
285	1	hypothetical protein	ref|NP_701842.1	4.3		Pfam Domain of unknown function	0.071				No
286	1	hypothetical protein	gb|EAN76693.1	1.5		Smart DSRM/ZnF_C2H2 domains	0.21				No
291	1	hypothetical protein	ref|XP_629925.1	5.6		Pfam Sre, C	0.029				No

Appendix II - Functional annotation of cDNA clusters from *Glossina papalis gambiensis* head library producing best hits to nonredundant (NR) protein database of GenBank, gene ontology (GO) and conserved domains database (CDD) database

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
Odorant/Pheromone Binding											
184	2	Olfactory Specific F	ref|NP_524241.1|	5E-36	*Drosophila melanogaster*	Smart Insect PBP/OBP domains	8E-18	pheromone binding	3.0E-37	0005550	Yes
195	1	transcription factor IIIB	ref|XP_640891.1|	5E-05	*Dictyostelium discoideum*	Smart Putative single-stranded NAs	0.077	pheromone binding	3.0E-04	0005550	No
204	1	CG11852-PA	gb|AAM50923.1|	2E-19	*Drosophila melanogaster*	Pfam Odorant binding protein	2E-29				No
206	1	hypothetical protein	emb|CAJ01441.1|	2E-09	*Drosophila pseudoobscura*	Pfam Insect PBP-binding family	4E-09	pheromone binding	1.0E-09	0005549	No
255	1	GA16322-PA	gb|EAL28074.1|	1E-45	*Drosophila pseudoobscura*	Pfam PBP/GOBP family	2E-15	odorant binding	2.0E-46	0005549	No
Energy Metabolism											
15	2	mtAcyl carrier subunit 2	emb|CAA70290.1|	6E-18	*Drosophila melanogaster*	Kog NADH-ubiqui oxidoreductase	1E-13	NADH DH activity	3.0E-19		Yes
21	1	mtAcyl carrier subunit 2	emb|CAA70290.1|	8E-50	*Drosophila melanogaster*	Kog NADH-ubiqui oxidoreductase	2E-39	NADH DH activity	4.0E-51		No
32	2	NADH DH subunit 1	ref|YP_245509.1|	2E-58	*Haematobia irritans irritans*	Kog NADH DH subunit 1	1E-46	NADH DH activity	8.0E-51		Yes
33	2	NADH DH subunit 1	ref|YP_245509.1|	3E-64	*Haematobia irritans irritans*	Kog NADH DH subunit 1	3E-52	NADH DH activity	8.0E-57		No
34	4	cyt-c oxidase subunit III	ref|NP_075453.1|	2E-88	*Cochliomyia hominivorax*	Pfam Cyt-c oxidase subunit III	5E-94	cyt-c oxidase activity	4.0E-87		No
35	2	cyt-c oxidase subunit III	ref|NP_075453.1|	2E-91	*Cochliomyia hominivorax*	Pfam Cyt-c oxidase subunit III	1E-96	cyt-c oxidase activity	8.0E-92		Yes
46	1	cyt oxidase subunit II	gb|AAQ13680.1|	5E-66	*Drosophila wassermani*	Pfam Cyt-C oxidase subunit II	4E-64	cyt-c oxidase activity	1.0E-63		No
47	2	Cyt oxidase subunit 2	gb|AAX47684.1|	3E-69	*Calliphora sternalis sternalis*	Pfam Cyt-C oxidase subunit II	5E-64	cyt-c oxidase activity	7.0E-67		No
53	1	NADH subunit 2	ref|NP_075448.1|	2E-43	*Cochliomyia hominivorax*	Kog NADH DH subunits	3E-16	NADH DH activity	2.0E-36		No
54	1	NADH subunit 2	ref|NP_075448.1|	2E-45	*Cochliomyia hominivorax*	Kog NADH DH subunits	3E-17	NADH DH activity	2.0E-39		No
70	2	cyt c oxidase p/peptide	gb|AAP88317.1|	4E-23	*Drosophila simulans*	Pfam Cyt oxidase c subunit VIII	0.014	cyt-c oxidase activity	2.0E-24		No
72	1	NADH DH subunit	gb|AAK21325.1|	2E-16	*Chrysomya putoria*	Pfam NADH DH subunit 5 C-terminus	3E-17	NADH DH activity	8.0E-11		No
73	1	NADH DH subunit	gb|AAL88695.1|	2E-04	*Trichophthalma sp.*	Pfam NADH DH subunit 5 C-terminus	0.0004	Actin depolymerization	1.0E-05	0005522	No
113	1	NADH DH subunit 5	ref|YP_025917.1|	9.8	*Xiphinema americanum*	Smart delta serrate ligand	0.47				No
166	2	cytc oxidase subunit I	gb|AAX35708.1|	2E-91	*Drosophila ficusphila*	Pfam Cyt-C, Quinol oxidase p/peptide	6E-56	sleep	6.0E-92		No
177	1	NADH subunit 5	ref|NP_075455.1|	2E-57	*Cochliomyia hominivorax*	Pfam NADH DH subunit 5 C-terminus	1E-45	NADH DH activity	9.0E-48		No
196	1	Cyt-c oxidase p/peptide Va	gb|AAR30197.1|	3E-55	*Drosophila melanogaster*	Pfam Cyt c oxidase subunit Va	4E-45	cyt-c oxidase activity	2.0E-56	0005489	No
212	1	GA18962-PA	gb|EAL32669.1|	7E-42	*Drosophila pseudoobscura*	Pfam NADH-ubiqui oxidoreductase	5.0E-42	NADH DH activity	1.0E-42		No
228	1	IP05690p	ref|NP_724697.1|	3E-27	*Drosophila melanogaster*	Kog Ubiquinol-Cyt c reductase	9E-17	ubiquinol-cyt-c reductase	2.0E-28		No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
238	1	Dro. melanogaster CG9306	gb|AAR10178.1|	4E-56	*Drosophila yakuba*	Kog NADH:ubiqui oxidoreductase	1E-45	NADH DH activity	2.0E-57		No
248	1	CG3446-PA	gb|AAL48700.1|	2E-56	*Drosophila melanogaster*	Kog NADH:ubiqui oxidoreductase	2E-45	mit electron transport	9.0E-58	0005524	No
260	1	LD31474p	ref|NP_727921.1|	6E-95	*Drosophila melanogaster*	Kog NADH-ubiqui oxidoreductase	6E-90	NADH DH activity	3.0E-96		No
297	1	ENSANGP00000011766	gb|EAA00203.3|	6E-24	*Anopheles gambiae*	Kog NADH:ubiqui oxidoreductase	9E-25	NADH DH activity	2.0E-23		No
Transcription Factors											
17	6	SJCHGC01957 protein	gb|AAX27763.1|	2E-27	*Schistosoma japonicum*	Kog Predicted RNA binding protein	2E-10	small ribonucleoprotein	7.0E-10	0005267	No
25	1	similar to D. melano RpL15	gb|AAR10086.1|	1E-99	*Drosophila yakuba*	Pfam Ribosomal L15	1E-78	struct constituent of ribosome	1.0E-101	0003723	No
26	1	similar to D. melano RpL15	gb|AAR10086.1|	6E-89	*Drosophila yakuba*	Pfam Ribosomal L15	4E-77	struct constituent of ribosome	3.0E-90	0003723	Yes
31	1	diaphanous homologue	ref|XP_478998.1|	4E-06	*Oryza sativa*	Pfam CAP protein	0.002	ribosomal DNA binding	7.0E-05	0000182	No
97	1	RNA-binding protein	gb|EAN97599.1|	1.2	*Trypanosoma cruzi*	Smart Protein phosphatase 2A	0.21				No
116	1	T-box transcription factor	ref|NP_001027587|	0.51	*Ciona intestinalis*	Kog Flavin-containing monooxygenase	0.087				No
172	1	CG17489-PA.3	gb|AAL48927.1|	1E-100	*Drosophila melanogaster*	Kog 60S ribosomal protein	4E-61	struct constituent of ribosome	1.0E-101	0008097	No
179	1	D. melanogaster CG2099	gb|AAR10024.1|	1E-67	*Drosophila yakuba*	Kog 60S ribosomal protein	2E-43	struct constituent of ribosome	1.0E-66	0003723	No
185	1	CG8636-PA	gb|AAM50793.1|	2E-90	*Drosophila melanogaster*	Smart RNA recognition motif	8E-13	eukaryotic TIF 3 complex	1.0E-91	0003723	No
229	1	GA20486-PA	gb|EAL32194.1|	3E-56	*Drosophila pseudoobscura*	Pfam Ribosomal protein L36e	1E-40	struct constituent of ribosome	1.0E-55	0003723	No
239	1	SJCHGC01957 protein	gb|AAX27763.1|	2E-02	*Schistosoma japonicum*	Pfam plant mitochondrial protein	0.005	ribonuclease MRP complex	0.056		No
245	1	transcriptional regulator	ref|ZP_01000137.1|	1.2	*Oceanicola batsensis*	Smart Ras-like small GTPases	0.096				No
271	1	CG7726-PA	gb|AAM11143.|	5E-95	*Drosophila melanogaster*	Pfam ribosomal L5P family C-terminus	2E-41	RNA binding	2.0E-96	0003723	No
273	1	similar to D.melanogaster	gb|AAR10020.1|	2E-30	*Drosophila yakuba*	Kog 60S ribosomal protein	4E-32	struct constituent of ribosome	7.0E-32	0003723	No
275	1	GA19229-PA	gb|EAL33406.1|	1E-122	*Drosophila pseudoobscura*	Kog 40S ribosomal protein	1E-106	Struct constituent of ribosome	1.0E-123	0003723	No
278	1	60S ribosomal protein L23	gb|AAY66949.1|	2E-12	*Ixodes scapularis*	Kog 60S ribosomal protein	3E-11	Struct constituent of ribosome	6.0E-14		No
281	1	GA20722-PA	gb|EAL30266.1|	5E-90	*Drosophila pseudoobscura*	Pfam Core binding factor beta subunit	4E-75	transcription coactivator	6.0E-91	0046982	No
80	1	similar to Splicing factor	ref|XP_213658.3|	2E-02	*Rattus norvegicus*	Smart Phosphoinositide 3-kinase	8E-04	cytoskeletal protein binding	0.002	0003779	No
107	1	unnamed protein product	dbj|BAA97098.1|	8E-18	*Arabidopsis thaliana*	Kog Histone H1	2E-08	Struct constituent - cytoskeleton	6.0E-08	0005200	Yes
168	1	GA21525-PA	gb|EAL25823.1|	2E-48	*Drosophila pseudoobscura*	Pfam Insect cuticle protein	1E-25	structural constituent of cuticle	6.0E-49	0005214	No
183	1	CG13041-PA	gb|AAL49186.1|	2E-13	*Drosophila melanogaster*	Pfam Drosophila Retinin like protein	3E-08				Yes
197	1	GA21874-PA	gb|EAL34274.1|	3E-58	*Drosophila pseudoobscura*	Smart Profilin/Binds actin monomers	3E-38	actin filament organization	2.0E-59	0003779	No
226	2	CG13043-PA	ref|NP_648868.2|	2E-21	*Drosophila melanogaster*	Pfam Drosophila Retinin like protein	7E-17	structural constituent of cuticle	0.005	0005214	Yes

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
234	2	CG7216-PA	gb|AAM51018.1|	2E-22	*Drosophila melanogaster*	Kog Ultrahigh sulfur keratin	0.0002	structural constituent of cuticle	8.0E-24	0008012	Yes
243	2	RH70420p	gb|AAM29629.1|	1E-23	*Drosophila melanogaster*	Pfam Male specific sperm protein	0.0001	structural constituent of cuticle	3.0E-15	0008012	Yes
249	1	CG13042-PA	gb|AAY85051.1|	2E-26	*Drosophila melanogaster*	Pfam Drosophila Retinin like protein	1E-18	structural constituent of cuticle	2.0E-05	0005214	Yes
266	1	CG7178-PE, isoform E	gb|AAM52657.1|	2E-44	*Drosophila melanogaster*	Pfam Troponin	4E-13	troponin complex	2.0E-40	0003779	No
291	1	H3f3b protein	gb|AAH88835.1|	4E-68	*Mus musculus*	Kog Histones H3 and H4	1E-66	chromosome org & biogenesis	7.0E-69	0003677	No
Transporters											
4	1	hypothetical protein	emb|CAH87320.1|	2E-04	*Plasmodium chabaudi*	Pfam Sugar transporter	0.035	kinase activity	0.086		No
38	1	hypothetical protein	ref|XP_678650.1|	2E-12	*Plasmodium berghei*	Pfam Domain of unknown function	8E-06	nucleocytoplasmic transport	1.0E-06	0003702	No
57	2	succinyl-CoA synthetase	ref|NP_700961.1|	0.12	*Plasmodium falciparum 3D7*	Kog Voltage-gated Ca2+ channels	0.049				Yes
91	2	LD45288p	emb|CAA20899.2|	1E-105	*Drosophila melanogaster*	Pfam Synaptobrevin	3E-17	vesicle-mediated transport	1.0E-65	0005485	No
92	1	LD45288p	emb|CAA20899.2|	6E-58	*Drosophila melanogaster*	Pfam Synaptobrevin	9E-17	vesicle-mediated transport	5.0E-33	0005485	No
93	1					Pfam LacY proton/sugar symporter	0.17				No
102	1	Substrate transporter	ref|ZP_00845774.1|	1.5	*Rhodopseudomonas palustris*	Smart Guanine nucleotide factor	0.071				Yes
119	1	putative chaperonin	ref|XP_730128.1|	0.004	*Plasmodium yoelii yoelii*	Pfam HCO3- transporter family	0.006				Yes
126	1	SJCHGC01957 protein	gb|AAX27763.1|	0.005	*Schistosoma japonicum*	Kog Predicted transporter	0.008				No
130	1					KogNa+:iodide/myo-inositol symporters	0.21				No
154	1	K^+ voltage-gated channel	ref|XP_692589.1|	1.5	*Danio rerio*	Kog Voltage-gated K+ channel	1.2				No
155	1	K^+ voltage-gated channel	ref|XP_692589.1|	9.8	*Danio rerio*	Kog Putative Rab5-interacting protein	0.13				No
173	3	ATP synthase F0 subunit 6	ref|YP_133762.1|	8E-96	*Dermatobia hominis*	Pfam ATP synthase A chain	3E-48	proton-transporting ATP synthase	9.0E-95	0046933	No
176	1	CG6782-PA, isoform A	gb|AAL29051.1|	3E-70	*Drosophila melanogaster*	Kog Mit tricarboxylate carrier proteins	1E-53	mitochondrial citrate transport	7.0E-59	0015137	No
181	1	ADP/ATP translocase	gb|AAF32322.1|	6E-66	*Lucilia cuprina*	Pfam Mitochondrial carrier protein	5E-15	ATP:ADP antiporter activity	3.0E-66	0015207	No
191	1	ENSANGP00000011079	gb|EAA05846.2|	1E-34	*Anopheles gambiae str. PEST*	Pfam Mitochondrial ATP synthase	2E-22	proton transport	2.0E-34		No
192	1	CG8931-PA	gb|AAK92928.1|	7E-32	*Drosophila melanogaster*	Pfam Mitochondrial carrier protein	3E-06				No
201	1	ABC cobalt transport system	dbj|BAD04265.1|	0.28	*Onion yellows phytoplasma*	Pfam Srg, C	0.13				No
211	1	SD10334p	gb|AAK93568.1|	9E-62	*Drosophila melanogaster*	Kog Myosin assembly protein	6E-33	Mit outer membrane translocase	5.0E-63		No
214	1	CG18624-PA, isoform A	ref|NP_727209.1|	6E-18	*Drosophila melanogaster*	Kog Na+/K+ transporter	0.029	NADH dehydrogenase activity	3.0E-19		No
290	1	GA20881-PA	gb|EAL29617.1|	8E-94	*Drosophila pseudoobscura*	Kog Mitochondrial F1F0-ATP synthase	3E-67	proton-transporting ATP synthase	3.0E-94		Yes
293	1	NodD	emb|CAA88827.1|	0.23	*Azorhizobium caulinodans*	Kog Na+/H+ antiporter	0.037				No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
296	1	RE66761p	gb|AAM48442.1|	6E-10	*Drosophila melanogaster*	Kog Translocase outer mit membrane	6E-06	mit outer membrane translocase	3.0E-11		No
Signal transduction											
6	1	SJCHGC08168 protein	gb|AAW24991.1|	3E-06	*Schistosoma japonicum*	Kog Adenylate cyclase	0.002	G-protein/camp 2nd messenger	1.0E-05	0003779	No
13	2	Y22D7AR.1	gb|AAK29874.1|	3E-08	*Caenorhabditis elegans*	Kog RhoA GTPase effector	9E-08	spliceosome complex	2.0E-07	0008307	No
18	1					Pfam taste receptor protein	0.18				No
42	1	Cyclic nucleotide-binding	ref|ZP_00600408.1|	4.6	*Rubrobacter xylanophilus*	Smart Ras-like GTPases	0.082				No
61	1	Hypothetical protein	gb|AAC00575.1|	0.0001	*Arabidopsis thaliana*	Kog RhoA GTPase effector	0.008	nucleic acid binding	1.0E-04	0003723	No
66	1	GPCR 52	ref|NP_005675.2|	6.5	*Homo sapiens*	Smart Amb_all domain	0.47				No
74	7	opsin Rh1	gb|AAS88872.2|	1E-138	*Bactrocera dorsalis*	Pfam 7 transmembrane (rhodopsin)	7E-45	G-protein coupled photoreceptor	1.0E-135		No
75	2	Opsin Rh1	gb|AAA62725.1|	7E-66	*Unknown*	Pfam 7 transmembrane (rhodopsin)	1E-09	G-protein coupled photoreceptor	2.0E-66		Yes
98	1					Pfam 7tm Odorant receptor	0.25				No
109	1	similar to Or 10T2	ref|XP_597712.2|	5.7	*Bos taurus*	Smart Thrombospondin N-terminal	0.23				No
148	1	SJCHGC01957 protein	gb|AAX27763.1|	2E-05	*Schistosoma japonicum*	Smart Calpain-like thiol protease	0.005	small GTPase signal transduction	0.068	0005522	No
194	1	ENSANGP00000013043	gb|EAA09123.2|	3E-14	*Anopheles gambiae*	Smart calmodulin-binding motif	9E-06	myosin/ ATPase activity	0.003	0042623	No
244	4	arrestin1	emb|CAA55672.1|	5E-77	*Calliphora vicina*	Pfam Arrestin (or S-antigen)	5E-26	metarhodopsin inactivation	2.0E-74	0016030	No
288	1	CG9224-PA	ref|NP_476736.1|	5E-81	*Drosophila melanogaster*	Pfam von Willebrand factor type C	6E-15	torso signaling pathway	3.0E-82	0008083	No
299	1	ENSANGP00000012700	gb|EAA05425.2|	3E-79	*Anopheles gambiae*	Smart calcium binding motif	9E-09	rhodopsin mediated signaling	2.0E-79		No
Protein Function											
1	1	CG31075-PA	ref|NP_733183.1|	1E-67	*Drosophila melanogaster*	Pfam Aldehyde DH family	9E-60	aldehyde DH (NAD) activity	7.0E-69	0005489	No
2	1	SJCHGC09076 protein	gb|AAW26562.1|	4E-13	*Schistosoma japonicum*	Kog Predicted hydrolase	0.03				No
5	1	tRNA-dihydrouridine synthase	ref|XP_710328.1|	0.12	*Candida albicans SC5314*	Pfam hypothetical mitochondrial protein	0.006				No
12	4	formin	gb|EAN92551.1|	1E-05	*Trypanosoma cruzi*	Pfam CAP protein	0.004	cell migration-locomotory	3.0E-06	0008017	No
20	4	polyprotein	ref|YP_145791.1|	3E-20	*Varroa destructor virus 1*	Pfam RNA dependent RNA polymerase	2E-20				No
30	1	CG10217-PA, isoform A	ref|NP_732903.2|	0.005	*Drosophila melanogaster*	Pfam eubacterial secY protein	0.37				No
37	1	putative chaperonin	ref|XP_730128.1|	0.004	*Plasmodium yoelii yoelii*	Smart Intercrine alpha family	0.3				No
39	1	unnamed protein product	ref|XP_501207.1|	3E-07	*Yarrowia lipolytica*	Pfam Atrophin-1 family	9E-06	pancreatic ribonuclease activity	2.0E-05	0003676	No
40	1	GH17891p	gb|AAK52961.1|	2E-81	*Drosophila melanogaster*	Kog Glutamine PRPP amidotransferase	1E-55	de novo' IMP biosynthesis	1.0E-82		No
41	1	CG10078-PA, isoform A	ref|NP_729191.1|	5E-49	*Drosophila melanogaster*	Kog Glutamine PRPP amidotransferase	1E-41	de novo' IMP biosynthesis	3.0E-50		No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
52	1	hypothetical protein	ref|XP_640250.1|	0.001	*Dictyostelium discoideum*	Kog Cadherin EGF/LAG G-type receptor	0.084	spliceosome complex	0.022		No
55	1	SJCHGC01974 protein	gb|AAW26857.1|	5E-10	*Schistosoma japonicum*	Pfam Ferric reductase like transmembrane	0.034				No
58	1	putative Kazal-type serpin	gb|AAL09362.1|	0.18	*Trypanosoma cruzi*	Pfam Herpesvirus BMRF2 protein	0.36				No
59	1	diaphanous homologue	ref|XP_478998.1|	9E-06	*Oryza sativa*	Pfam CAP protein	0.012	mitochondrion	2.0E-05	0008017	No
62	1	Similar- dispatched homolog	ref|XP_785823.1|	5E-06	*Strongylocentrotus purpuratus*	Smart LITAF	0.005	neurotransmitter secretion	5.0E-04	0003676	No
63	1	putative chaperonin	ref|XP_730128.1|	0.023	*Plasmodium yoelii yoelii*	Pfam Envelope glycoprotein	0.028				No
64	1	mFLJ00419 protein	dbj|BAD21439.1|	0.018	*Mus musculus*	Pfam Atrophin-1 family	0.002	transcription corepressor activity	0.039		No
67	1	SJCHGC01957 protein	gb|AAX27763.1|	0.026	*Schistosoma japonicum*	Pfam Synapsin, N-terminal domain	0.11				Yes
76	1	GA22021-PA	gb|EAL28728.1|	1E-32	*Drosophila pseudoobscura*	Kog Translation initiation factor 3	5E-30	salivary gland cell death	3.0E-33		No
85	1	SJCHGC03138 protein	gb|AAX25791.2|	0.19	*Schistosoma japonicum*	Smart Presenilin	0.007	Ca^+ -activated K^+ channel activity	0.03	0005267	No
89	1	SJCHGC01957 protein	gb|AAX27763.1|	3E-16	*Schistosoma japonicum*	Kog Telomerase elongation inhibitor	2E-08	embryonic development	2.0E-06	0003723	Yes
90	1	SJCHGC01957 protein	gb|AAX27763.1|	0.51	*Schistosoma japonicum*	Smart LITAF	0.15				No
94	1					Kog Predicted nucleolar protein	1.9				No
96	1	P0518C01.25	ref|NP_914359.1|	0.06	*Oryza sativa*	Kog P-type ATPase	0.064				No
100	1	Phosphoribosyltransferase	gb|AAO35570.1|	1.1	*Clostridium tetani E88*	Smart SprT homologues	0.27				No
106	1	methionine sulfoxide reductase	gb|AAV62883.1|	0.87	*Streptococcus thermophilus*	Pfam Rft protein	0.12				No
117	1	SJCHGC01957 protein	gb|AAX27763.1|	1E-07	*Schistosoma japonicum*	Pfam Sec-independent protein translocase	3E-06	Ca^+ -activated K^+ channel activity	4.0E-04	0005267	Yes
121	1	SJCHGC01957 protein	gb|AAX27763.1|	0.002	*Schistosoma japonicum*	Pfam TB2/DP1/HVA22 family	0.008				No
123	1	hybrid proline-rich protein	gb|AAA33132.1|	0.019	*Unknown*	Smart STE like transcription factors	0.02				Yes
127	1	similar to FIP1 like 1	ref|XP_690394.1|	2.1	*Danio rerio*	Kog Cytochrome P450	0.19				No
128	1	similar to Nebulin	ref|XP_851603.1|	4.3	*Canis familiaris*	Pfam Diaphanous GTPase-binding Domain	0.17				No
129	1	SJCHGC01957 protein	gb|AAX27763.1|	0.013	*Schistosoma japonicum*	Pfam FtsH Extracellular	0.11				Yes
131	1	Polyadenylate protein	gb|EAL88129.|	5.6	*Aspergillus fumigatus Af293*	Pfam Prolipoprotein DAG transferase	0.38				No
134	1	CG9717-PA	gb|AAL48537.1|	0.06	*Drosophila melanogaster*	Smart Galanin	0.38	sulfate permease activity	0.002		No
135	1	SJCHGC01957 protein	gb|AAX27763.1|	9E-05	*Schistosoma japonicum*	Pfam YMF19 plant mitochondrial protein	0.13				No
139	1	LD31383p	ref|NP_572704.1|	0.006	*Drosophila melanogaster*	Pfam Atrophin-1 family	0.0004	Rab guanyl-nucleotide factor	2.0E-04	0017112	No
143	1	3-hydroxyacyl-CoA DH	gb|EAM76027.1|	7.6	*Kineococcus radiotolerans*	Pfam Cyt c oxidase assembly protein	0.13				No
144	1	putative chaperonin	ref|XP_730128.1|	0.018	*Plasmodium yoelii yoelii*	Pfam Domain of unknown function	0.26				Yes

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
145	1	Ubiquitin	ref|XP_889225.1|	7.5	*Mus musculus*	Kog (Down-regulated in metastasis)	0.14				No
146	1	3-oxoacyl-synthase	ref|ZP_00980631.1|	0.16	*Burkholderia cenocepacia*	Smart Villin headpiece domain	0.061				Yes
147	1	SJCHGC01957 protein	gb|AAX27763.1|	3E-04	*Schistosoma japonicum*	Kog CBF1-interacting corepressor	0.006				No
153	1	protein kinase	gb|EAL44480.1|	9.8	*Entamoeba histolytica*	Smart Glycosyl hydrolase family 10	0.33				No
160	1	Ab1-181	gb|AAP92554.1|	9.6	*Rattus norvegicus*	Smart K homology RNA-binding domain	0.05				No
167	1	GA12109-PA	gb|EAL26389.1|	0.58	*Drosophila pseudoobscura*	Pfam Protein of unknown function	0.071				No
174	1	imaginal disc growth factor	gb|ABC25096.1|	4E-58	*Glossina morsitans morsitans*	Kog Chitinase	2E-15	imaginal disc growth factor	5.0E-49	0008084	No
186	1	receptor-like protein kinase	ref|NP_181196.1|	1.1	*Arabidopsis thaliana*	Pfam Probable DNA packing protein	0.25	kinase activity	0.039		No
187	1	envelope glycoprotein	emb|CAB95107.1|	9.9	*Human immunodeficiency virus*	Pfam Photosystem II reaction centre I	0.17				No
189	1	unnamed protein product	dbj|BAB11454.1|	6E-08	*Arabidopsis thaliana*	Pfam CAP protein	0.002	protein binding	2.0E-06	0003676	No
190	1	C2 domain, putative	ref|XP_723898.1|	1.5	*Plasmodium yoelii yoelii*	Smart Presenilin	0.079				No
198	1	GA12593-PA	gb|EAL29998.1|	5E-12	*Drosophila pseudoobscura*	Smart Serine/Threonine protein kinases	0.021				No
207	1	MnFe superoxide dismutase	gb|AAT85826.1|	1E-113	*Glossina morsitans morsitans*	Pfam manganese superoxide dismutases	5E-40	antioxidant activity	8.0E-84		No
210	1	putative chaperonin	ref|XP_730128.1|	0.0007	*Plasmodium yoelii yoelii*	Pfam Prominin	0.26				Yes
213	1	RE70703p	gb|AAL68358.1|	3E-56	*Drosophila melanogaster*	Pfam Tim17/Tim22/Tim23 family	2E-05	mitochondrion	1.0E-15		No
215	1	GA11203-PA	gb|EAL30481.1|	1E-21	*Drosophila pseudoobscura*	Kog 4-OHphenylpyruvate dioxygenase	3E-24	4-OHphenylpyruvate dioxygenase	2.0E-22		No
216	1	adenosine deaminase	gb|AAQ23537.1|	6E-07	*Drosophila melanogaster*	Pfam 7TM chemoreceptor	0.25	adenosine deaminase factor	2.0E-08	0008083	No
217	1	Sodium channel protein type	ref|XP_600250.2|	4.6	*Bos taurus*	Kog Cyclic nucleotide phosphodiesterase	0.57				No
220	1	nerve growth factor	gb|AAH63835.1|	1.1	*Homo sapiens*	Kog Adaptor protein	0.02				No
221	1	Gmfb8	gb|AAL83358.1|	8E-05	*Glossina morsitans morsitans*	Pfam Domain of unknown function	6E-04				No
223	2	similar to putative protein	ref|XP_344805.1|	0.012	*Rattus norvegicus*	Pfam ALG6/ALG8 glycosyltransferase	0.25	DNA exonuclease activity	0.088	0009381	No
225	2	hypothetical protein	ref|XP_642201.1|	4E-04	*Dictyostelium discoideum*	Pfam Translocation protein Sec62	1E-04	pseudohyphal growth	0.036	0003677	No
227	1	serine proteinase	gb|AAW57295.1|	2E-07	*Delia antiqua*	Smart Trypsin-like serine protease	1E-06	trypsin activity/proteolysis	2.0E-04	0017080	No
230	1	putative enzyme	gb|AAN42522.2|	5E-23	*Shigella flexneri 2a str. 301*	Smart Domain associated with PX	0.095				No
232	1	CG1420-PA	ref|NP_651659.2|	1E-08	*Drosophila melanogaster*	Kog Pyruvate dehydrogenase E1	6E-07	pre-mRNA splicing factor y	5.0E-09	0008248	No
233	1	chromosome proteins	gb|AAC74010.1|f	1E-109	*Escherichia coli K12*	Pfam PspA/IM30 family	4E-07	cytokinesis	6.0E-05	0042623	No
237	1	hypothetical protein	pir||T31613	1E-05	*Unknown*	Pfam Borrelia ORF-A	3E-05	anaphase-promoting complex	0.03	0003704	No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
246	1	htpE	gb|AAT65047.1|	0.53	*Bacteriophage phiMFV1*	Smart LITAF	0.075				Yes
251	1	GA16794-PA	gb|EAL32306.1|	8E-33	*Drosophila pseudoobscura*	Kog Calcyclin-binding protein	1E-29	nuclear membrane lumen	4.0E-20		No
252	1	COG1230	ref|ZP_00109846.2|	7.1	*Nostoc punctiforme*	Smart Intron encoded nuclease repeat motif	0.098				No
254	1	Wiskott-Aldrich syndrome	emb|CAI21159.1|	3E-08	*Danio rerio*	Pfam CAP protein	0.002	actin nucleation	2.0E-06	0003677	No
259	1	ferritin heavy chain-like	gb|ABC48949.1|	0.0008	*Glossina morsitans*	Kog Ferritin	0.04				No
261	1	hypothetical protein	ref|XP_635201.1|	6E-05	*Dictyostelium discoideum*	Kog FERM domain protein	0.002	histone methylation	0.001	0003677	No
264	1	similar to Myol protein	ref|XP_623750.1|	9E-40	*Apis mellifera*	Kog Myotrophin and similar proteins	2E-40	DNA binding	1.0E-40	0003677	No
265	1	unknown	gb|AAX27911.1|	5E-05	*Schistosoma japonicum*	Pfam 7TM chemoreceptor	0.003	double-stranded RNA binding	0.018	0003725	No
267	1	hypothetical protein	ref|XP_638834.1|	1E-05	*Dictyostelium discoideum*	Kog Aminopeptidases of the M20 family	5E-06	response to aluminum ion	2.0E-05	0019789	Yes
268	1	GA17562-PA	gb|EAL30728.1|	9E-38	*Drosophila pseudoobscura*	Smart Protein tyrosine phosphatase	0.04	regulation of cell shape	2.0E-37		No
276	1	hypothetical protein	ref|ZP_00779277.1|	3E-05	*Thermoanaerobacter ethanolicus*	Pfam T-cell surface antigen CD2 protein	0.004	purine nucleotide binding	0.019	0005524	No
277	1	SJCHGC01957 protein	gb|AAX27763.1|	1E-06	*Schistosoma japonicum*	Kog CBF1-interacting corepressor CIR	1E-04	locomotory behavior	0.008		No
279	1					Pfam Human herpesvirus U26 protein	0.34				Yes
282	1	receptor-like protein kinase	ref|NP_914396.1|	0.001	*Oryza sativa*	Kog Alpha-1,2 glucosyltransferase	0.013				Yes
286	1	GH10454p	ref|NP_651385.3|	5E-36	*Drosophila melanogaster*	Smart HLH domain containing kinases	4E-12				No
287	1	polyprotein	gb|AAP49283.1|	2E-17	*Deformed wing virus*	Pfam RNA dependent RNA polymerase	5E-06				No
292	1	chromosome protein	ref|NP_703436.1|	1.5	*Plasmodium falciparum 3D7*	Smart TLC/TRAM/LAG1/CLN8 domains	0.042				No
295	1	Membrane glycosyltransferase	gb|AAN42667.2|	4E-49	*Shigella flexneri 2a str. 301*	Smart LITAF	0.004	Oligosaccharidebiosynthesis	1.00E-10		Yes
298	1	tRNA pseudouridine synthase	ref|YP_115618.1|	1.5	*Mycoplasma hyopneumoniae*	Smart Domain A in dwarfin family proteins	0.13				No
300	3	ATP-dependent RNA helicase	ref|YP_309759.1|	1E-64	*Shigella sonnei Ss046*	Kog ATP-dependent RNA helicase	5E-37	ATP-dependent helicase activity	4.0E-39	0003723	Yes
303	1	LD35854p	gb|AAK93286.1|	3E-06	*Drosophila melanogaster*	Pfam Protein of unknown function	1E-07				No
304	1	LD35669p	ref|NP_651209.1|	1E-101	*Drosophila melanogaster*	Pfam Autophagy protein Apg6	3E-79	salivary gland cell death	1.0E-102		No
305	1	GA14438-PA	gb|EAL28349.1|	4E-63	*Drosophila pseudoobscura*	Kog Typepeptidyl-prolylcis-isomerase	4E-06	FK506 binding	0.012	0005528	No
Unknown											
3	1	unknown	gb|AAX27911.1|	2E-07	*Schistosoma japonicum*	Kog Secretory carrier membrane protein	0.005	pseudouridine synthesis	0.008	0030519	No
8	1	unknown	gb|AAX27955.1|	2E-18	*Schistosoma japonicum*	Kog Ribosome biogenesis protein	4E-08	DNA topoisomerase type I	2.0E-08	0000217	No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
7	1	unknown	gb|AAX27911.1|	0.002	*Schistosoma japonicum*	Kog Ferric reductase-like proteins	0.0004				No
14	8	unknown	gb|AAX27955.1|	2E-17	*Schistosoma japonicum*	Kog CBF1-interacting corepressor	1E-06	zinc D-Ala-carboxypeptidase	3.0E-11	0004180	No
16	127	unknown	gb|AAX27955.1|	2E-07	*Schistosoma japonicum*	Pfam DIE2/ALG10 family	5E-06	nucleic acid binding	0.001	0003723	No
22	1					Smart semaphorin domain	1				No
27	1	unnamed protein product	dbj|BAC86300.1|	0.01	*Homo sapiens*	Pfam Domain of unknown function	0.022				No
29	1	unnamed protein product	dbj|BAC86300.1|	0.005	*Homo sapiens*	Pfam Domain of unknown function	0.061				No
81	1					Kog Mannosyltransferase	0.23				No
83	1					Pfam Herpesvirus virion protein	0.11				No
87	1	Unknown protein	ref|NP_919611.1|	1E-08	*Oryza sativa*	Pfam CAP protein	0.001	transcription mediator activity	6.0E-06	0003702	No
95	1					Smart meprin/TRAF homology	0.26				No
103	1	unnamed protein product	dbj|BAC86300.1|	0.1	*Homo sapiens*	Pfam TB2/DP1/HVA22 family	0.088				No
108	1	unknown	ref|NP_704103.1|	0.021	*Plasmodium falciparum*	Pfam Herpes virus tegument protein	0.08				No
114	1	unnamed protein product	dbj|BAC86300.1|	0.036	*Homo sapiens*	Pfam Protein of unknown function	0.31				No
118	1	unknown	gb|AAX27955.1|	0.047	*Schistosoma japonicum*	Pfam WzyE protein	0.005	RNA polymerase II transcription	0.088	0003704	No
133	1	unknown	gb|AAX27955.1|	1.9	*Schistosoma japonicum*	Smart Cytoplasmic phospholipase A2	0.36				No
138	1	unknown	gb|AAX27911.1|	0.0003	*Schistosoma japonicum*	Kog Predicted G-protein coupled receptor	0.08				No
140	1	unnamed protein product	emb|CAA91606.1|	0.004	*Saccharomyces cerevisiae*	Pfam Protein of unknown function	0.009				No
142	1					Kog (Half-A-TPR) containing protein	0.096				No
149	1					Pfam protein of unknown function	0.24				No
164	1					Pfam Conserved carboxylase domain	0.18				No
178	1	unnamed protein product	emb|CAA33190.1|	0.01	*Crithidia fasciculata*	Smart Presenilin	0.007				No
199	1	unnamed protein product	ref|NP_059418.1|	0.036	*Paramecium aurelia*	Kog Lipid exporter ABCA1	0.03				Yes
256	1					Smart Domain in homologues of a S	1.2				No
269	1	unknown	gb|AAX27955.1|	0.0008	*Schistosoma japonicum*	Kog Transporter, ABC superfamily	0.006				No
289	1	unnamed protein product	emb|CAF91062.1|	0.027	*Tetraodon nigroviridis*	Pfam Adenovirus E3B protein	0.018				No
294	1					Pfam RTA1 like protein	0.047				No
301	1	unknown	gb|AAX27955.1|	7E-06	*Schistosoma japonicum*	Smart Presenilin	0.0007	transcription regulator activity	0.023	003702	No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
Hypothetical											
9	1	hypothetical protein	emb|CAE75230.1|	1E-24	*Caenorhabditis briggsae*	Pfam Sec-independent protein translocase	2E-10	N/O-linked glycosylation	4.0E-13	0017069	Yes
10	1	hypothetical protein	emb|CAE75230.1|	9E-25	*Caenorhabditis briggsae*	Pfam Sec-independent protein translocase	2E-11	N/O-linked glycosylation	2.0E-12	0017069	No
19	25	hypothetical protein	ref|XP_642369.1|	0.018	*Dictyostelium discoideum*	Pfam Borrelia ORF-A	0.001				No
23	1	hypothetical protein	ref|XP_645055.1|	2.2	*Dictyostelium discoideum*	Pfam Protein of unknown function	0.24				No
24	1	hypothetical protein	ref|XP_646912.1|	0.071	*Dictyostelium discoideum*	Pfam Protein of unknown function	0.27				No
28	1	hypothetical protein	ref|XP_729957.1|	7.4	*Plasmodium yoelii yoelii*	Smart Presenilin	1				No
36	1	hypothetical protein	gb|AAZ68053.1|	1.7	*Ehrlichia canis str. Jake*	Pfam Mammalian taste receptor protein	0.01				No
43	1	hypothetical protein	ref|XP_644193.1|	0.05	*Dictyostelium discoideum*	Pfam Domain of unknown function	0.011	kinase activity	0.037		No
44	1	hypothetical protein	ref|XP_644272.1|	0.57	*Dictyostelium discoideum*	Pfam Uncharacterised protein family	0.14				No
45	1	hypothetical protein	ref|XP_642163.1|	0.078	*Dictyostelium discoideum*	Pfam Coronavirus nonstructural protein 4	0.002				No
48	1	hypothetical protein	ref|XP_720764.1|	4E-06	*Candida albicans SC5314*	Smart TLC/TRAM/LAG1/CLN8 domains	8E-06				No
49	1	hypothetical protein	ref|XP_638906.1|	0.001	*Dictyostelium discoideum*	Pfam Sec-independent protein translocase	2E-06	G-protein coupled receptor	0.005		No
50	1	hypothetical protein 4	pir||E22845	2E-06	*Trypanosoma brucei*	Pfam Domain of unknown function	5E-06				No
51	1	hypothetical protein 4	pir||E22845	0.052	*Trypanosoma brucei*	Smart Presenilin	0.007				No
56	4	hypothetical protein	ref|XP_642527.1|	0.52	*Dictyostelium discoideum*	Pfam Domain - MnhB of Na+/H+ antiporter	0.17				No
60	1	hypothetical protein	ref|NP_701006.1|	0.061	*Plasmodium falciparum 3D7*	Kog Uncharacterized membrane protein	0.054				No
65	1	hypothetical protein	ref|NP_197106.1|	0.051	*Arabidopsis thaliana*	Pfam Atrophin-1 family	0.002				No
68	1	hypothetical protein	ref|XP_638495.1|	0.18	*Dictyostelium discoideum*	Pfam SCAMP family	0.027				No
69	1	hypothetical protein	ref|XP_643220.1|	0.018	*Dictyostelium discoideum*	Pfam YMF19 plant mitochondrial protein	0.035				No
71	1	hypothetical protein	ref|XP_638906.1|	0.031	*Dictyostelium discoideum*	Kog Alpha-1,2 glucosyltransferase	0.042				No
77	1	hypothetical protein	ref|NP_700666.1|	0.027	*Plasmodium falciparum 3D7*	Pfam Domain of unknown function	0.006				No
78	1	hypothetical protein	emb|CAC09395.1|	0.036	*Neurospora crassa*	Smart Horizontally Transferred Domain	0.036				No
79	1	hypothetical protein	gb|AAN57898.1|	0.52	*Streptococcus mutans*	Smart Phosphoinositide 3-kinase	0.006				No
82	1	hypothetical protein	ref|XP_846856.1|	3.4	*Trypanosoma brucei*	Kog Mannosyltransferase	0.19				No
88	1	hypothetical protein	emb|CAJ04956.1|	0.0002	*Leishmania major*	Pfam Domain of unknown function	0.001				No
84	1	hypothetical protein	ref|NP_704585.1|	0.51	*Plasmodium falciparum 3D7*	Pfam Domain of unknown function	0.044				No
86	1	hypothetical protein	ref|XP_641074.1|	0.0001	*Dictyostelium discoideum*	Pfam Sec-independent protein translocase	0.0008	peptidyl-proline hydroxylation	0.023		No
99	1	hypothetical protein	gb|EAL87031.1|	0.002	*Aspergillus fumigatus Af293*	Smart Formin Homology	0.26				No
101	1	hypothetical protein	ref|XP_645982.1|	9.1	*Dictyostelium discoideum*	Kog Secretory carrier membrane protein	0.001				No
104	1	hypothetical protein	ref|XP_646781.1|	0.042	*Dictyostelium discoideum*	Kog Predicted UDP-galactose transporter	0.32				No
105	1	hypothetical protein	ref|XP_725254.1|	7.4	*Plasmodium yoelii yoelii*	Pfam Domain of unknown function	0.008				No
110	1	hypothetical protein	ref|XP_644272.1|	0.0002	*Dictyostelium discoideum*	Smart Eukaryotic DNA topoisomerase II	0.04				No
111	1	hypothetical protein	ref|XP_640207.1|	7.5	*Dictyostelium discoideum*	Pfam Pox virus E6 protein	0.077				No
112	1	hypothetical protein	ref|XP_644272.1|	0.053	*Dictyostelium discoideum*	Kog Oxoprolinase	0.076				No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
115	1	hypothetical protein	ref|XP_676757.1|	0.3	*Plasmodium berghei strain*	Pfam Mammalian taste receptor protein	0.017				No
120	1	hypothetical protein	ref|XP_643799.1|	0.005	*Dictyostelium discoideum*	Smart Presenilin	0.002	odontogenesis	0.01		No
122	1	hypothetical protein	ref|XP_635915.1|	0.0002	*Dictyostelium discoideum*	Kog Uncharacterized conserved protein	0.21	snoRNP binding	0.053	0030519	No
124	1	hypothetical protein	ref|NP_701505.1|	9.6	*Plasmodium falciparum 3D7*	Smart RIO-like kinase	0.23				No
125	1	hypothetical protein	ref|ZP_00236498.1|	0.019	*Bacillus cereus G9241*	Smart human TAF130/Drosophila TAF110	0.19				No
132	1	hypothetical protein	ref|XP_673325.1|	7.6	*Plasmodium berghei strain*	Smart TopoisomeraseII	0.88				No
136	1	hypothetical protein	ref|NP_702215.1|	0.3	*Plasmodium falciparum 3D7*	Pfam Borrelia ORF-A	0.0008				No
137	1	hypothetical protein	ref|XP_644809.1|	0.0005	*Dictyostelium discoideum*	Pfam Borrelia ORF-A	0.002				No
150	1	hypothetical protein	ref|XP_644809.1|	0.14	*Dictyostelium discoideum*	Pfam Sec-independent translocase protein	0.076				No
151	1	hypothetical protein	emb|CAH81582.1|	0.013	*Plasmodium chabaudi*	Pfam GPI transamidase subunit	0.005				No
152	1	hypothetical protein	gb|EAN76494.1|	0.1	*Trypanosoma brucei*	Pfam Protein of unknown function	0.13				No
157	1	hypothetical protein	ref|XP_644272.1|	0.006	*Dictyostelium discoideum*	Kog Predicted small molecule transporter	0.077				No
158	1	hypothetical protein	ref|XP_636844.1|	0.049	*Dictyostelium discoideum*	Pfam Sugar (and other) transporter	0.073				No
159	1	hypothetical protein	pir||T31613	0.0005	*Unknown*	Pfam NADH-ubiquinone oxidoreductase	0.0003	odontogenesis	0.03		No
161	2	hypothetical protein	gb|AAF18309.1|	0.062	*SVTS2 plectrovirus*	Kog Pleiotropic drug resistance proteins	0.12				No
162	2	hypothetical protein	gb|AAF18309.1|	0.036	*SVTS2 plectrovirus*	Pfam SCAMP family	0.011				Yes
163	5	hypothetical protein	gb|EAN85835.1|	0.22	*Trypanosoma cruzi*	Pfam YMF19 plant mitochondrial protein	0.18				Yes
165	2	hypothetical protein	gb|EAN86251.1|	0.21	*Trypanosoma cruzi*	Pfam Chordopoxvirus G3 protein	0.24				Yes
169	1	hypothetical protein	ref|XP_737285.1|	0.13	*Plasmodium chabaudi*	Smart Pentraxin/C-reactive protein	0.04				No
170	1	hypothetical protein	ref|NP_702053.1|	0.069	*Plasmodium falciparum 3D7*	Smart CLUSTERIN alpha chain	0.007				No
171	1	hypothetical protein	gb|EAN85835.1|	0.002	*Trypanosoma cruzi*	Pfam Cytochrome c oxidase	0.014	Mitochondria transporter activity	0.03		Yes
175	1	hypothetical protein	gb|EAN86251.1|	0.022	*Trypanosoma cruzi*	Smart minichromosome proteins	0.007				No
180	1	hypothetical protein	ref|XP_637981.1|	0.011	*Dictyostelium discoideum*	Pfam CHD5-like protein	0.1				No
182	1	hypothetical protein	gb|AAC74862.1|	7E-55	*Escherichia coli K12*	Pfam Protein of unknown function	3E-27				No
193	1	hypothetical protein	ref|NP_703646.1|	7.5	*Plasmodium falciparum 3D7*	Pfam Guanylate-binding protein	0.26				No
200	1	hypothetical protein	gb|AAX45893.1|	0.061	*Zygnema circumcarinatum*	Pfam Uncharacterised protein family	0.012				No
205	1	hypothetical protein	gb|EAN92483.1|	0.31	*Trypanosoma cruzi*	Smart LITAF	0.14				No
208	1	hypothetical protein	ref|NP_703519.1|	0.035	*Plasmodium falciparum 3D7*	Smart Presenilin	0.049				No
209	1	hypothetical protein	ref|NP_702447.1|	9E-05	*Plasmodium falciparum 3D7*	Smart Presenilin	0.0002	transcription factor activity	0.001	0003702	No
219	1	hypothetical protein	ref|XP_641763.1|	0.014	*Dictyostelium discoideum*	Pfam Protein of unknown function	0.017				No
222	3	hypothetical protein	ref|XP_636996.1|	8.8	*Dictyostelium discoideum*	Pfam Uncharacterised protein family	0.12				No
224	1	hypothetical protein	ref|XP_721601.1|	1.9	*Candida albicans SC5314*	Smart DNA polymerase type-B family	0.066				No
231	1	hypothetical protein	ref|XP_725833.1|	0.39	*Plasmodium yoelii yoelii*	Smart LITAF	0.004				No
235	1	hypothetical protein	pir||T31613	1E-05		Pfam Borrelia ORF-A	0.0001				No
236	1	hypothetical protein	gb|EAL87031.1|	0.046	*Aspergillus fumigatus Af293*	Pfam DIE2/ALG10 family	0.012				Yes
240	1	hypothetical protein	ref|NP_703519.1|	5.7	*Plasmodium falciparum 3D7*	Pfam Ion transport protein	0.087				No
241	1	hypothetical protein	ref|XP_644272.1|	0.012	*Dictyostelium discoideum*	Pfam 7TM chemoreceptor	0.002				No
250	1	hypothetical protein	emb|CAC13940.1|	2.5	*Mycoplasma pulmonis*	Pfam Apical membrane antigen 1	0.076				No

Appendix II
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
257	1	hypothetical protein	ref|XP_745683.1|	0.09	*Plasmodium chabaudi*	Pfam Protein of unknown function	0.003				No
258	1	hypothetical protein	ref|NP_700669.1|	0.043	*Plasmodium falciparum 3D7*	Pfam V-type ATPase	0.2				No
263	1	hypothetical protein	ref|NP_700669.1|	2.7	*Plasmodium falciparum 3D7*	Pfam Trehalose receptor	0.22				No
270	1	hypothetical protein	ref|XP_963466.1|	1.5		Smart Presenilin	0.12				No
272	1	hypothetical protein	ref|XP_643799.1|	3E-06	*Dictyostelium discoideum*	Smart Presenilin	0.0002	odontogenesis	0.066		No
274	1	hypothetical protein	gb|EAN85835.1|	0.006	*Trypanosoma cruzi*	Smart Presenilin	0.14				No
280	1	hypothetical protein	emb|CAE76809.1|	1.6	*Mycoplasma mycoides*	Pfam Protein of unknown function	0.004				No
283	1	hypothetical protein	gb|EAN31130.1|	1.2	*Theileria parva*	Pfam Disulfide bond formation protein	0.22				No
284	1	hypothetical protein	ref|XP_638906.1|	0.0001	*Dictyostelium discoideum*	Kog Secretory carrier membrane protein	0.008				No
285	1	hypothetical protein	ref|XP_646009.1|	1.1	*Dictyostelium discoideum*	Pfam Fijivirus P9-2 protein	0.029				No
302	1	hypothetical protein	ref|XP_670767.1|	0.007	*Plasmodium berghei strain*	Pfam 7TM chemoreceptor	0.008				No

Appendix III - Functional annotation of cDNA clusters from *Glossina tachinoides* head library producing best hits to nonredundant (NR) protein database of GenBank, gene ontology (GO) and conserved domains database (CDD) database

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	GO No	SignalP
Odorant/Pheromone Binding											
63	2	RH74005p	gb|AAM29645.1|	1E-44	*Drosophila melanogaster*	Pfam Insect pheromone-binding family	1E-42	odorant binding	6E-46	0005549	No
151	1	odorant binding protein 83g	gb| AAN13232.1 |	9E-19	*Drosophila melanogaster*	Smart Insect PBP/OBP domains	3E-10	odorant binding	4E-20	0005549	No
219	1	odorant binding protein 3	gb|AAV74624.1|	5E-07	*Musca domestica*	Smart Insect PBP/OBP domains	5E-08	pheromone binding	2E-08	0005550	No
Energy Metabolism											
9	3	CG4412-PA	ref|NP_477194.1|	2E-40	*Drosophila melanogaster*	Kog Mitochondrial F1F0-ATP synthase	3E-33	proton-transporting ATP synthase	1E-41		No
11	2	cyt c oxidase polypeptide IV	gb|AAP88303.1|	1E-71	*Drosophila simulans*	Kog Cytochrome c oxidase	3E-59	cyt-c oxidase activity	4E-72		Yes
32	2	LD14731p	gb|EAL25276.1|	7E-19	*Drosophila pseudoobscura*	Pfam Cyt c oxidase subunit VIIc	7E-07	cyt-c oxidase activity	4E-20		No
33	2	ENSANGP00000010167	ref|XP_625078.1|	4E-23	*Apis mellifera*	Kog F0F1-type ATP synthase	3E-27	proton-transporting ATP synthase	9E-24	0046933	No
35	5	cytochrome oxidase subunit 3	ref|YP_133763.1|	1E-108	*Dermatobia hominis*	Pfam Cytochrome c oxidase subunit III	1E-107	cytochrome-c oxidase activity	1E-109		No
36	6	cytochrome b	gb|AAG08980.1|	6E-93	*Scathophaga obscura*	Pfam Cytochrome b(C-terminal)/b6/petD	6E-26	ubiquinol-cyt-c reductase activity	2E-90	0005489	No
42	2	GA14437-PA	gb|EAL25670.1|	2E-41	*Drosophila pseudoobscura*	Pfam Cytochrome c oxidase subunit VIa	8E-26	cytochrome-c oxidase activity	2E-42		Yes
44	3	cytochrome oxidase subunit 2	gb|AAX47696.1|	1E-100	*Hemipyrellia fergusoni*	Kog Cytochrome c oxidase	7E-93	cytochrome-c oxidase activity	1E-95		No
47	2	CG9603-PA	ref|NP_652184.2|	5E-29	*Drosophila melanogaster*	Pfam Cytochrome c oxidase subunit VIIa	3E-09	cytochrome-c oxidase activity	3E-30		Yes
65	1	ENSANGP00000017699	gb|EAA13279.2|	2E-13	*Anopheles gambiae str.*	Kog Ubiquinol cyt c oxidoreductase	2E-13	spermatid cell development	6E-14		No
66	1	SJCHGC01957 protein	gb|AAX27763.1|	4E-10	*Schistosoma japonicum*	Kog Ubiquinol cyt c oxidoreductase	0.001				No
72	3	NADH dehydrogenase subunit	gb|AAK21326.1|	6E-61	*Chrysomya putoria*	Kog NADH dehydrogenase	6E-43	NADH dehydrogenase activity	2E-56		No
76	2	GA14517-PA	gb|EAL26698.1|	2E-65	*Drosophila pseudoobscura*	Pfam ATP synthase subunit C	3E-21	hydrogen-exporting ATPase	4E-66	0008289	No
89	2	cytochrome oxidase subunit I	gb|AAF28303.1|	3E-36	*Scathophaga tropicalis*	Pfam Cytochrome C	1E-05				No
96	1	ATP synthase A chain	gb|AAV41426.1|	3E-12	*Periplaneta americana*	Kog ATP synthase F0 subunit 6	1E-04	hydrogen-exporting ATPase	2E-08		Yes
97	1	ATP synthase F0 subunit 6	ref|YP_133762.1|	1E-43	*Dermatobia hominis*	Kog ATP synthase F0 subunit 6	5E-30	hydrogen-exporting ATPase	5E-43	0046933	No
101	1	NADH subunit 5	ref|NP_075455.1|	1E-104	*Cochliomyia hominivorax*	Pfam NADH DH subunit 5	2E-55	NADH dehydrogenase activity	4E-93		No
116	1	NADH dehydrogenase subunit	ref|YP_054477.1|	7.4	*Podura aquatica*	Pfam RofA transcriptional regulator	0.17				No
131	1	CG18624-PA, isoform A	ref|NP_727209.1|	9E-17	*Drosophila melanogaster*	Pfam Phosphatidylinositolglycan class N	0.009	NADH dehydrogenase activity	3E-18		No
135	1	IP05690p	ref|NP_724697.1|	2E-17	*Drosophila melanogaster*	Pfam Ubiquinol-cytochrome C reductase	6E-12	ubiquinol-cytochrome-c reductase	9E-19		No

Appendix III Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
144	1	CG9306-PA	gb|AAL48406.1|	3E-29	*Drosophila melanogaster*	Kog NADH:ubiquinone oxidoreductase	5E-21	NADH dehydrogenase activity	1E-30	0008017	No
155	1	LD31474p	ref|NP_727921.1|	1E-27	*Drosophila melanogaster*	Pfam NADH ubiquinone oxidoreductase	5E-15	NADH dehydrogenase activity	5E-29		Yes
181	1	CG8189-PA, isoform A	gb|AAR99111.1|	4E-20	*Drosophila melanogaster*	Pfam Mitochondrial ATP synthase B chain	8E-14	hydrogen-exporting ATPase	2E-21		Yes
207	1	NADH subunit 4	ref|NP_075456.1|	6E-85	*Cochliomyia hominivorax*	Pfam NADH-Ubiquinone/plastoquinone	1E-38	NADH DH (ubiquinone) activity	2E-81		No
Transcription Factors											
3	1	AT10293p	gb|AAO42670.1|	7E-27	*Drosophila melanogaster*	Kog Troponin	3E-07	double-stranded RNA binding	3E-28	0003725	No
4	1	CG5163-PA	emb|CAA58244.1|	3E-48	*Drosophila melanogaster*	Pfam Transcription initiation factor IIA	3E-15	transcription factor TFIIA complex	2E-49		No
8	2	CG15442-PA	gb|AAL48766.1|	6E-64	*Drosophila melanogaster*	Kog 60s ribosomal protein L15/L27	9E-55	structural constituent of ribosome	3E-65	0003723	Yes
10	2	CG15603-PA	ref|NP_996473.1|	5E-39	*Drosophila melanogaster*	Pfam BAF1/ABF1 chromatin factor	2E-05	chromatin/transcriptional repressor	0.0002	0016564	Yes
19	1	RNAse (dsRNA binding motif)	ref|YP_142697.1|	2.5	*Acanthamoeba polyphaga*	Pfam emp24/gp25L/p24 family	0.035				No
30	2	ribosomal protein S3	emb|CAD12886.1|	1E-89	*Drosophila virilis*	Kog 40S ribosomal protein S3	9E-84	structural constituent of ribosome	8E-90	0003676	Yes
31	1	hypothetical protein	ref|XP_723186.1|	0.18	*Candida albicans SC5314*	Smart Protein phosphatase 2A	0.06	RNA polymerase II activity	0.099	0003704	No
34	2	receptor-like protein kinase	ref|NP_914396.1|	0.035	*Oryza sativa*	Pfam Sec-independent protein translocase	3E-04	mRNA processing	0.037		No
48	1	GA20693-PA	gb|EAL32444.1|	2E-52	*Drosophila pseudoobscura*	Kog Mit/chloroplast ribosomal protein	2E-45	perception of sound	2E-53		No
50	1	CG11835-PA	ref|NP_608532.1|	2E-52	*Drosophila melanogaster*	Kog RNA polymerase II	9E-37	transcription factor activity	1E-53	0003702	No
64	1	DNA-directed RNA polymerase	emb|CAA41631.1|	7E-65	*Drosophila melanogaster*	Kog RNA polymerase III	3E-70	DNA-directed RNA polymerase	4E-66	0003677	No
91	1	Integrin alpha-8 precursor	ref|XP_796395.1|	0.006	*Strongylocentrotus purpuratus*	Kog Predicted DHHC/Zn-finger protein	0.008	transcription of nuclear rRNA	0.03		No
103	1	Drosophila melanogaster	gb|AAR10024.1|	7E-18	*Drosophila yakuba*	Kog 60S ribosomal protein	8E-07	structural constituent of ribosome	2E-19	0000049	Yes
115	1	condensin complex component	pir||T49494	0.056		Smart Ribosomal protein L11/L12	0.008				No
137	1	GA20486-PA	gb|EAL32194.1|	4E-48	*Drosophila pseudoobscura*	Pfam Ribosomal protein L36e	2E-36	structural constituent of ribosome	1E-48	0003723	No
139	1	CG1420-PA	ref|NP_651659.2|	3E-08	*Drosophila melanogaster*	Kog RNA splicing factor-Slu7p	2E-04	pre-mRNA splicing factor activity	2E-08	0008248	No
148	1	hypothetical protein	ref|XP_644272.1|	3E-05	*Dictyostelium discoideum*	Pfam Der1-like family	0.038	transcription of nuclear rRNA	0.051	0030519	Yes
143	1	transcription initiation factor	ref|XP_629530.1|	0.061	*Dictyostelium discoideum*	Smart Doublesex DNA-binding motif	0.036				No
154	1	Wiskott-Aldrich syndrome-like	ref|NP_082735.1|	2E-05	*Mus musculus*	Pfam Wilm's tumour protein	0.051	actin nucleation	9E-07	0003677	No
162	1	SJCHGC01957 protein	gb|AAX27763.1|	8E-05	*Schistosoma japonicum*	Pfam Ribonucleases P/MRP protein	0.004	transcription factor activity	0.039	0003723	No
164	1	ribosomal protein L11	gb|AAV34822.1|	7E-11	*Bombyx mori*	Kog 60S ribosomal protein L11	9E-11	protein binding	3E-11	0003723	Yes

Appendix III Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
167	1	GA19229-PA	gb|EAL33406.1|	2E-41	*Drosophila pseudoobscura*	Pfam Ribosomal protein S5	3E-19	structural constituent of ribosome	4E-42	0003723	No
168	1	60S ribosomal protein	gb|AAY66949.1|	4E-14	*Ixodes scapularis*	Pfam Ribosomal protein L14p/L23e	3E-08	structural constituent of ribosome	1E-15		No
173	1	CG17489-PC.3	gb|AAS93729.1|	2E-25	*Drosophila melanogaster*	Kog 60S ribosomal protein L5	8E-17	structural constituent of ribosome	1E-26	0008097	No
174	1	GA20722-PA	gb|EAL30266.1|	2E-39	*Drosophila pseudoobscura*	Kog Transcription factor CBF	2E-26	transcription coactivator activity	2E-40	0046982	No
200	1	GA10244-PA	gb|EAL28204.1|	3E-56	*Drosophila pseudoobscura*	Smart glycine rich nucleic binding domain	3E-04	nuclear mRNA splicing	0.0002		No
203	1	GM14667p	gb|AAM49965.1|	9E-67	*Drosophila melanogaster*	Kog E3 ubiquitin ligase	3E-26	nucleic acid binding	5E-68	0003676	No
212	1	Eukaryotic ribosomal protein	ref|XP_729142.1|	0.009	*Plasmodium yoelii yoelii*	Pfam Borrelia ORF-A	0.02				No
213	1	GA20722-PA	gb|EAL30266.1|	6E-34	*Drosophila pseudoobscura*	Pfam Core binding factor	5E-24	transcription coactivator activity	3E-35	0046982	No
Cytoskeletal											
12	6	GA20185-PA	gb|EAL33529.1|	4E-31	*Drosophila pseudoobscura*	Pfam MAGE family	3E-04	structural constituent-adult cuticle	6E-16	0008012	Yes
13	2	RH70420p	gb|AAM29629.1|	4E-24	*Drosophila melanogaster*	Kog Ultrahigh sulfur keratin protein	2E-04	structural constituent-adult cuticle	2E-15	0008012	Yes
60	3	CG13043-PA	ref|NP_648868.2|	2E-21	*Drosophila melanogaster*	Pfam Drosophila Retinin like protein	7E-17	structural constituent of cuticle	0.006	0005214	Yes
117	1	structural constituent of cell wall	ref|NP_683431.1|	9E-09	*Arabidopsis thaliana*	Kog Rac1 GTPase effector	2E-06	Ca-dependent phospholipid binding	9E-07	0005522	No
161	2	GA17562-PA	gb|EAL30728.1|	2E-22	*Drosophila pseudoobscura*	Kog Mitogen-activated protein kinase	0.098	actin filament organization	3E-22		No
191	1	MAP1B protein	gb|AAH17240.1|	0.002	*Homo sapiens*	Kog Pleiotropic drug resistance proteins	0.007	potassium channel activity	0.006	0005267	No
Transporters											
6	2	CG3725-PA, isoform A	ref|NP_476832.1|	1E-103	*Drosophila melanogaster*	Kog Ca2+ transporting ATPase	2E-72	calcium-transporting ATPase	1E-101		No
14	2	unnamed protein product	ref|XP_454192.1|	0.06	*Kluyveromyces lactis*	Kog Nuclear transport receptor LGL2	0.085				No
25	2	CG2968-PA	gb|AAL48825.1|	2E-62	*Drosophila melanogaster*	Kog Mitochondrial F1F0-ATP synthase	2E-51	proton-transporting ATP synthase	1E-63	0046933	No
43	2	GA20584-PA	gb|EAL33947.1|	8E-12	*Drosophila pseudoobscura*	Pfam Sec-independent protein translocase	0.008				
92	1	ABC transporter	ref|XP_670113.1|	1.2	*Plasmodium berghei strain*	Smart LITAF	0.081				No
100	1	ENSANGP00000018102	gb|EAA12925.2|	2E-61	*Anopheles gambiae*	Kog Mitochondrial tricarboxylate/carrier	7E-49	mitochondrial citrate transport	2E-51	0015137	No
111	1	SJCHGC01957 protein	gb|AAX27763.1|	0.002	*Schistosoma japonicum*	Pfam LacY proton/sugar symporter	0.027				No
129	1	GA15439-PA	gb|EAL28731.1|	1E-17	*Drosophila pseudoobscura*	Kog Myosin assembly protein	0.001	mit outer membrane translocase	6E-19		Yes
179	1	transposon protein	gb|ABA96402.1|	0.0002	*Oryza sativa*	Kog Synaptic vesicle protein Synapsin	7E-04	ISG15 carrier activity	0.0007	0005522	No
184	1	putative chaperonin	ref|XP_730128.1|	0.018	*Plasmodium yoelii yoelii*	Kog Na+/H+ antiporter	0.086				Yes
228	1	phosphatidylinositol 4-kinase	ref|NP_113372.1|	0.036	*Guillardia theta*	Pfam Acyltransferase family	0.05				No
231	1	putative protein kinase	ref|XP_549827.1|	0.86	*Oryza sativa*	Kog Na+-nucleoside cotransporter	5E-04				No

Appendix III Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
184	1	putative chaperonin	ref|XP_730128.1|	0.018	*Plasmodium yoelii yoelii*	Kog Na+/H+ antiporter	0.086				Yes
228	1	phosphatidylinositol 4-kinase	ref|NP_113372.1|	0.036	*Guillardia theta*	Pfam Acyltransferase family	0.05				No
231	1	putative protein kinase	ref|XP_549827.1|	0.86	*Oryza sativa*	Kog Na+-nucleoside cotransporter	5E-04				No
Signal transduction											
5	1	RH28686p	gb|AAX33382.1|	4E-62	*Drosophila melanogaster*	Kog Arrestin	2E-38	rhodopsin mediated signaling	2E-63	0016030	No
39	21	opsin Rh1	gb|AAS88872.2|	0	*Bactrocera dorsalis*	Kog G protein-coupled receptor	6E-25	phototransduction	0		No
46	8	arrestin1	emb|CAA55672.1|	1E-66	*Calliphora vicina*	Kog Arrestin	3E-52	metarhodopsin inactivation	8E-65	0016030	No
93	1	formin binding protein 30	ref|XP_230291.3|	0.021	*Rattus norvegicus*	Kog RhoA GTPase effector	0.009	ligand-dependent nuclear receptor	0.003	0003704	No
118	1	hydroxyproline-rich glycoprotein	emb|CAB62280.1|	2E-22	*Volvox carteri f. nagariensis*	Pfam Formin Homology Region 1	2E-16	small GTPase regulator activity	2E-17	0003677	No
197	1	arrestin1	emb|CAA55672.1|	9E-80	*Calliphora vicina*	Pfam Arrestin (or S-antigen)	4E-15	metarhodopsin inactivation	5E-79	0016030	No
198	1	similar to Olfactory receptor	ref|XP_870536.1|	0.86	*Bos taurus*	Smart TLC/TRAM/LAG1/CLN8 domains	0.009				No
Protein Function											
7	3	SJCHGC01957 protein	gb|AAX27763.1|	0.004	*Schistosoma japonicum*	Smart LITAF	0.078				No
16	1	conserved protein	ref|XP_743082.1|	0.41	*Plasmodium c. chabaudi*	Smart SERine Proteinase Inhibitors	0.37				No
17	1	putative chaperonin	ref|XP_730128.1|	0.079	*Plasmodium yoelii yoelii*	Smart Olfactomedin-like domains	0.3				No
20	1	aspartic acid-rich protein	emb|CAA07355.1|	0.004	*Plasmodium falciparum*	Smart LITAF	0.039				No
21	177	galactosidase α-polypeptide	gb|AAA84370.1|	0.0005		Pfam YMF19 plant mitochondrial protein	0.12				No
22	8	putative chaperonin	ref|XP_730128.1|	0.011	*Plasmodium yoelii yoelii*	Pfam YMF19 plant mitochondrial protein	0.14				Yes
24	2	maturase-like protein	emb|CAA10854.1|	0.44	*Euglena spirogyra*	Pfam Riboflavin kinase (Flavokinase)	0.25				No
26	2	CG11079-PA, isoform A	ref|NP_611785.1|	7E-18	*Drosophila melanogaster*	Pfam Mitochondrial ATPase inhibitor	4E-16	5-formyltetrahydrofolate cyclo-ligase	4E-19	0004857	Yes
27	2	LD46083p	gb|AAL90297.1|	4E-38	*Drosophila melanogaster*	Pfam Lyase	6E-20	fumarate hydratase activity	1E-39		Yes
28	1	SJCHGC01957 protein	gb|AAX27763.1|	6E-07	*Schistosoma japonicum*	Pfam Cytidylyltransferase family	4E-04	embryonic development	0.023	0003723	No
29	8	polyprotein	emb|CAD34006.2|	3E-38	*Deformed wing virus*	Pfam RNA dependent RNA polymerase	1E-27				No
37	1	CG7875-PA	ref|NP_476768.1|	2E-19	*Drosophila melanogaster*	Kog Calreticulin	3E-04	calmodulin binding	1E-20	0005516	Yes
52	2	SJCHGC01957 protein	gb|AAX27763.1|	0.017	*Schistosoma japonicum*	Pfam Protein of unknown function	0.11	calcium channel activity	0.09	0005262	Yes
56	2	putative chaperonin	ref|XP_730128.1|	0.014	*Plasmodium yoelii yoelii*	Smart Domain in homologues of a S	0.024				No
74	1	CG3590-PA	ref|NP_650586.2|	1E-67	*Drosophila melanogaster*	Kog Adenylosuccinate lyase	4E-55	adenylosuccinate lyase activity	4E-68		No

Appendix III Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
75	1	GH08719p	gb|AAL28209.1|	5E-10	*Drosophila melanogaster*	Kog Adenylosuccinate lyase	4E-06	adenylosuccinate lyase activity	2E-11		No
77	3	ENSANGP00000012700	gb|EAA05425.2|	1E-58	*Anopheles gambiae*	Smart calcium binding motif	9E-09	calcium ion sensing	7E-60		Yes
80	2	cement precursor protein 3B	gb|AAY29121.1|	6E-10	*Phragmatopoma californica*	Smart DNA-binding domain - MUTS family	2E-05	cell adhesion molecule binding	4E-08		Yes
82	2	CG7985-PA	ref|NP_650689.1|	7E-60	*Drosophila melanogaster*	Smart Calpain-like thiol protease family	0.045				Yes
85	2	receptor-like protein kinase	ref|NP_914396.1|	0.017	*Oryza sativa*	Smart Phospholipase A2	0.011				Yes
88	1	beta-galactosidase α-polypeptide	gb|AAA84370.1|	0.0002		Kog Tyrosine kinase negative regulator	0.002				No
95	1	CG5670-PD, isoform D	ref|NP_732575.1|	2E-98	*Drosophila melanogaster*	Pfam haloacid dehalogenase-like hydrolase	3E-06	regulation of cell shape	3E-99	0005524	No
99	1	expressed protein	ref|NP_565801.1|	4.3	*Arabidopsis thaliana*	Smart DNA/RNA-binding repeats	0.16				Yes
102	1	troponin T	gb|AAD33604.1|	2E-07	*Libellula pulchella*	Kog Troponin	0.005	mesoderm development	2E-08	0005523	No
106	1	Cyanate permease	ref|ZP_00697373.1|	1E-16	*Shigella boydii BS512*	Kog Protein of unknown function	3E-04				Yes
110	1	syntaxin	gb|EAN81394.1|	0.019	*Trypanosoma cruzi*	Pfam SCAMP family	0.56				Yes
119	1	CG13877-PA	ref|NP_612014.1|	3E-08	*Drosophila melanogaster*	Kog Multitransmembrane protein	0.025				No
121	1	GA18355-PA	gb|EAL28490.1|	4E-51	*Drosophila pseudoobscura*	Kog Aldehyde dehydrogenase	5E-46	succinate-semialdehyde DH activity	4E-50	0005489	No
123	1	P0518C01.25	ref|NP_914359.1|	0.003	*Oryza sativa*	Pfam TB2/DP1, HVA22 family	0.049				Yes
124	1	succinyl-CoA synthetase	ref|NP_700961.1|	0.065	*Plasmodium falciparum 3D7*	Smart TopoisomeraseII	0.084				No
125	1	MnFe superoxide dismutase	gb|AAT85826.1|	5E-04	*Glossina morsitans morsitans*	Kog Manganese superoxide dismutase	0.028				Yes
127	1	oligosacharyl transferase subunit	ref|NP_701033.1|	0.013	*Plasmodium falciparum 3D7*	Smart Cysteine-rich domain- Drosophila	0.094				Yes
128	1	rpoD	emb|CAA64574.1|	0.68	*Plasmodium falciparum*	Kog Tryptophan-rich basic nuclear protein	0.24				No
132	1	adenosine deaminase factor	gb|AAQ23537.1|	0.51	*Drosophila melanogaster*	Kog Ethanolamine kinase	0.46	adenosine deaminase activity	0.018	0008083	No
133	1	RH52407p	gb|AAX33371.1|	1E-59	*Drosophila melanogaster*	Kog Molecular chaperone	5E-23	response to heat	5E-61	0051082	Yes
136	1	putative enzyme	gb|AAN42522.2|	7E-14	*Shigella flexneri 2a str. 301*	Smart Domain associated with PX domains	0.095				No
140	1	putative diaphanous homologue	ref|XP_478998.1|	2E-06	*Oryza sativa*	Pfam Survival motor neuron protein	0.002	mRNA processing/RNA splicing	2E-05	0004722	No
146	1	GA14282-PA	gb|EAL32706.1|	6E-12	*Drosophila pseudoobscura*	Kog eIF2-interacting protein	1E-10	physiological process	1E-09	0017151	No
152	1	SJCHGC01957 protein	gb|AAX27763.1|	4E-06	*Schistosoma japonicum*	Kog CBF1-interacting corepressor proteins	6E-04			0005267	No
153	1	aspartic acid-rich protein	emb|CAA07355.1|	0.001	*Plasmodium falciparum*	Kog Alpha-1,2 glucosyltransferase	0.009				Yes
159	1	CG31715-PA	gb|AAM29484.1|	2E-05	*Drosophila melanogaster*	Myotrophin and similar proteins	0.009				No
160	2	hypothetical protein	ref|XP_638834.1|	1E-05	*Dictyostelium discoideum*	Kog Aminopeptidases of the M20 family	3E-05	regulation of nitrogen utilization	2E-05	0019789	Yes

Appendix III Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
172	1	cysteine repeat modular protein	ref|XP_726874.1|	0.24	*Plasmodium yoelii yoelii*	Pfam Ferric reductase like transmembrane	0.024				No
183	1	e receptor-like protein kinase	ref|NP_914396.1|	0.008	*Oryza sativa*	Smart Horizontally TransMembrane Domain	0.23				Yes
186	1	Membrane glycosyltransferase	gb|AAN42667.2|	9E-57	*Shigella flexneri 2a str. 301*	Pfam Tymovirus 45/70Kd protein	0.012	oligosaccharide biosynthesis	7E-22		No
187	1	SJCHGC01957 protein	gb|AAX27763.1|	0.079	*Schistosoma japonicum*	Kog Adenylate cyclase- calcitonin receptor	0.047				No
189	1	putative chaperonin	ref|XP_730128.1|	0.003	*Plasmodium yoelii yoelii*	Kog Alpha amylase	0.08				No
190	1	ATP-dependent RNA helicase	dbj|BAB34298.1|	4E-38	*Escherichia coli O157:H7*	Pfam DEAD/DEAH box helicase	6E-16	ATP-dependent helicase activity	1E-24		No
194	1	LD35669p	ref|NP_651209.1|	4E-70	*Drosophila melanogaster*	Pfam Autophagy protein	2E-57	autophagic cell death	2E-71		Yes
195	1	CG17282-PA	gb|AAR96163.1|	3E-11	*Drosophila melanogaster*	Kog Alpha-1,2 glucosyltransferase	0.014				No
196	3	RE19540p	gb|AAL90338.1|	2E-24	*Drosophila melanogaster*	Kog WD40 repeat-containing protein	4E-24	eggshell pattern formation	8E-26	0045502	No
202	1	CG8839-PD, isoform D	ref|NP_725139.1|	4E-88	*Drosophila melanogaster*	Pfam Amidase	4E-12				Yes
205	1	GA18185-PA	gb|EAL31565.1|	5E-72	*Drosophila pseudoobscura*	Smart Ubiquitin-conjugating enzyme	5E-37	ubiquitin conjugating enzyme activity	2E-71		No
209	1	similar to D. mela EG:22E5.9	gb|AAR10251.1|	8E-23	*Drosophila yakuba*	Kog Uncharacterized conserved protein	6E-16	integral to membrane	0.0001		No
217	1	ZNF285 protein	gb|AAH74824.1|	2.6	*Homo sapiens*	Pfam Cytochrome C oxidase subunit II	0.06				No
230	1	GA14438-PA	gb|EAL28349.1|	2E-39	*Drosophila pseudoobscura*	Kog Dolichol kinase	0.04				No
Unknown											
2	20	unknown	gb|AAX27955.1|	0.0002	*Schistosoma japonicum*	Smart Presenilin	0.03				No
38	4	unknown	gb|AAX27955.1|	4E-25	*Schistosoma japonicum*	Kog Telomerase elongation inhibitor	3E-12	N-linked glycosylation	6E-11	0005267	No
45	1	unknown	gb|AAX27911.1|	0.027	*Schistosoma japonicum*	Smart Presenilin	0.089				No
54	1					Pfam Uncharacterised protein family	0.057				No
55	1	CG6579-PA	ref|NP_609558.2|	7E-48	*Drosophila melanogaster*	Pfam Baculovirus protein of unknown function	0.007				No
61	1	unknown	gb|AAX27911.1|	0.0007	*Schistosoma japonicum*	Pfam YMF19 plant mitochondrial protein	0.019				No
62	1	integral membrane protein	pir||A71606	0.016	*Unknown*	Pfam Zona-pellucida-binding protein	0.015				No
69	2	GA11916-PA	gb|EAL26149.1|	1E-27	*Drosophila pseudoobscura*	Kog Unnamed protein	1E-28				No
73	1	integral membrane protein	pir||A71606	3.4	*Unknown*	Pfam Sec-independent protein translocase	0.13				No
83	1					Smart Presenilin	2.5				No
94	1	putative chaperonin	ref|XP_730128.1|	0.024	*Plasmodium yoelii yoelii*	Smart CLUSTERIN alpha chain	0.006				Yes
98	1					Smart Lipoxygenase homology 2	0.4				No

Appendix
III Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
104	1	coding region HP1516	ref|NP_208307.1|	3.4	*Helicobacter pylori 26695*	SmartTopoisomeraseII	0.12				No
108	1					Smart erythroblast transformation domain	0.17				No
114	1	transesterase	gb|ABA02244.1|	0.034	*Monascus pilosus*	Pfam YMF19 plant mitochondrial protein	0.064				No
120	1	unknown	ref|NP_704103.1|	0.003	*Plasmodium falciparum 3D7*	Pfam 7TM chemoreceptor	0.011				No
150	1	unknown	gb|AAX27911.1|	5.7	*Schistosoma japonicum*	Pfam Srg, C	0.048				No
166	1	unnamed protein product	emb|CAA97071.1|	0.021	*Saccharomyces cerevisiae*	Pfam Protein of unknown function	0.002				No
177	1	putative chaperonin	ref|XP_730128.1|	0.001	*Plasmodium yoelii yoelii*	Kog Uncharacterized conserved protein	0.063				Yes
178	1	unnamed protein product	dbj|BAC86300.1|	0.007	*Homo sapiens*	Smart Presenilin	0.008				No
193	1	LD35854p	gb|AAK93286.1|	1E-06	*Drosophila melanogaster*	Pfam Protein of unknown function	2E-06				No
199	1					Smart G protein alpha subunit	0.06				No
206	1					Smart Domain A in dwarfin family proteins	0.17				No
210	1					Pfam Herpesvirus egress protein UL20	0.28				No
211	1					Kog Beta-transducin family protein	0.099				No
216	1					Pfam Domain of unknown function	0.29				No
220	1					Pfam Eukaryotic protein of unknown function	0.21				No
221	1					Pfam Trehalose receptor	0.39				No
224	1	unnamed protein product	emb|CAF95619.1|	2.7	*Tetraodon nigroviridis*	Smart Glycoprotein hormone alpha chain	0.37				No
227	1					Smart Domain of Unknown Function	0.41				No
Hypothetical											
1	1	hypothetical protein	ref|XP_638538.1|	3E-10	*Dictyostelium discoideum*	Smart Presenilin	7E-05	nucleic acid binding	1E-07	0003676	No
15	1	hypothetical protein	gb|EAN85835.1|	0.13	*Trypanosoma cruzi*	Pfam Rft protein	0.075				Yes
18	1	hypothetical protein	ref|NP_704173.1|	3.3	*Plasmodium falciparum 3D7*	Pfam DNA polymerase family B	0.11				No
23	2	hypothetical protein	gb|EAN86251.1|	0.027	*Trypanosoma cruzi*	Smart DNA polymerase type-B family	0.017				No
40	2	hypothetical protein	ref|XP_636844.1|	0.019	*Dictyostelium discoideum*	Smart Ribosomal protein L11/L12	0.033				No
41	10	hypothetical protein	gb|EAL89832.1|	0.035	*Aspergillus fumigatus Af293*	Pfam Uncharacterised protein family	0.23				Yes
49	1	hypothetical protein	ref|XP_638906.1|	0.003	*Dictyostelium discoideum*	Pfam Sec-independent protein translocase	0.006				Yes
51	1	hypothetical protein	ref|XP_672587.1|	0.016	*Plasmodium berghei strain*	Pfam Sec-independent protein translocase	0.002				Yes

Appendix III
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
53	2	hypothetical protein	ref|NP_701936.1|	0.038	*Plasmodium falciparum 3D7*	Pfam Domain of unknown function	0.014				Yes
57	2	hypothetical protein	ref|XP_644272.1|	0.0002	*Dictyostelium discoideum*	Pfam Domain of unknown function	0.01				Yes
58	1	hypothetical protein	ref|XP_637301.1|	0.022	*Dictyostelium discoideum*	Kog Ribosome biogenesis protein	0.48				No
59	2	hypothetical protein	ref|XP_643220.1|	0.017	*Dictyostelium discoideum*	Smart domain in FBox and BRCT	0.41				Yes
67	1	hypothetical protein	ref|NP_703519.1|	0.87	*Plasmodium falciparum 3D7*	Pfam Sre, C	0.026				No
68	2	hypothetical protein	emb|CAD70304.1|	0.67	*Neurospora crassa*	Pfam Glycine rich protein family	0.077				Yes
70	1	hypothetical protein	emb|CAE06957.1|	2.1	*Synechococcus sp. WH 8102*	Smart Ubiquitin-conjugating enzyme E2	0.039				No
71	1	hypothetical protein	ref|XP_638495.1|	0.001	*Dictyostelium discoideum*	Pfam GtrA-like protein	0.019				No
78	1	hypothetical protein	ref|NP_704964.1|	1.3	*Plasmodium falciparum 3D7*	Pfam Myelin proteolipid protein	0.16				Yes
79	1	hypothetical protein	ref|XP_726011.1|	0.031	*Plasmodium yoelii yoelii*	Pfam Protein of unknown function	0.012				No
81	1	hypothetical protein	gb|EAA71490.1|	9.6	*Gibberella zeae PH-1*	Smart Protein phosphatase 2A homologues	0.3				No
84	1	hypothetical protein	gb|EAN85835.1|	0.007	*Trypanosoma cruzi*	Kog CAATT-binding transcription factor	0.34				No
86	1	hypothetical protein	gb|EAN77150.1|	0.31	*Trypanosoma brucei*	Smart LITAF	0.047				No
87	1	hypothetical protein	ref|XP_730830.1|	5.7	*Plasmodium yoelii yoelii*	Kog Serine/threonine protein kinase	0.16				No
90	1	hypothetical protein	gb|EAN85835.1|	0.011	*Trypanosoma cruzi*	Pfam Protein of unknown function	0.15				Yes
105	1	hypothetical protein	ref|NP_473073.2|	0.037	*Plasmodium falciparum 3D7*	Kog Multitransmembrane protein	0.1				Yes
107	1	hypothetical protein	ref|XP_644272.1|	0.006	*Dictyostelium discoideum*	Pfam Fibronectin-binding protein A	0.12				No
109	1	hypothetical protein	ref|XP_642253.1|	0.81	*Dictyostelium discoideum*	Pfam Protein of unknown function	0.44				No
112	1	hypothetical protein	ref|NP_700669.1|	0.093	*Plasmodium falciparum 3D7*	Pfam YMF19 plant mitochondrial protein	0.026				Yes
113	1	hypothetical protein	ref|XP_636980.1|	0.046	*Dictyostelium discoideum*	Pfam Accessory gene regulator B	0.014				No
122	1	hypothetical protein	ref|NP_703849.1|	5.8	*Plasmodium falciparum 3D7*	Smart Histone 2A	0.4				No
126	1	hypothetical protein	ref|XP_646781.1|	2E-05	*Dictyostelium discoideum*	Smart Presenilin	4E-05				No
130	1	hypothetical protein	ref|XP_643220.1|	0.009	*Dictyostelium discoideum*	Smart Cysteine-rich domain - Drosophila	0.097	integral to membrane	0.053		Yes
134	1	hypothetical protein	ref|XP_756946.1|	3.3	*Ustilago maydis 521*	Smart RIO-like kinase	0.09				No
138	1	hypothetical protein	ref|XP_640658.1|	4.4	*Dictyostelium discoideum*	Smart LITAF	0.009				No
141	1	hypothetical protein	ref|XP_642086.1|	0.041	*Dictyostelium discoideum*	Smart Presenilin	0.002	odontogenesis	0.077		Yes
142	1	hypothetical protein	ref|XP_644272.1|	0.001	*Dictyostelium discoideum*	Smart TLC/TRAM/LAG1/CLN8 domains	0.014				Yes

Appendix III
Cont.

Cluster No.	No of Seqs	Match to NR DB	Accession No	E-Value	Species Match	Best Rpsblast to CDD DB	E-value	Match to GO DB	E-Value	Go No	SignalP
145	1	hypothetical protein	ref|NP_909739.1|	0.001	*Oryza sativa*	Pfam Sec-independent protein translocase	0.037				No
147	1	hypothetical protein	ref|XP_643102.1|	0.001	*Dictyostelium discoideum*	Pfam Bacterial low temperature requirement	0.083				Yes
149	1	hypothetical protein	gb|EAN85835.1|	0.007	*Trypanosoma cruzi*	Pfam LacY proton/sugar symporter	0.04				Yes
157	1	hypothetical protein	gb|EAN85835.1|	0.012	*Trypanosoma cruzi*	Pfam Ion transport protein	0.028				No
163	1	hypothetical protein	gb|EAL87031.1|	0.16	*Aspergillus fumigatus Af293*	Smart MADS domain	0.18				No
165	1	hypothetical protein	ref|XP_644272.1|	0.013	*Dictyostelium discoideum*	Kog 60S ribosomal protein L38	0.01				Yes
169	1	hypothetical protein	ref|XP_644272.1|	0.005	*Dictyostelium discoideum*	Pfam Domain of unknown function	0.015				Yes
170	1	hypothetical protein	gb|EAN85835.1|	0.004	*Trypanosoma cruzi*	Pfam hypothetical plant mitochondrial	0.027				No
175	1	hypothetical protein	ref|XP_640860.1|	0.009	*Dictyostelium discoideum*	Pfam Baculoviridae late expression factor 5	0.018				No
176	1	hypothetical protein	ref|XP_643556.1|	0.021	*Dictyostelium discoideum*	Pfam Fijivirus P9-2 protein	0.016				Yes
180	1	hypothetical protein	gb|EAL87031.1|	0.033	*Aspergillus fumigatus Af293*	Pfam Rifin/stevor family	0.2				Yes
182	1	hypothetical protein	ref|XP_642253.1|	0.059	*Dictyostelium discoideum*	Pfam Cytochrome c oxidase subunit III	0.077				Yes
185	1	hypothetical protein	ref|XP_643102.1|	0.003	*Dictyostelium discoideum*	Kog Uncharacterized conserved protein	0.003				No
188	1	hypothetical protein	ref|XP_658760.1|	0.66	*Aspergillus nidulans FGSC*	Kog Uncharacterized membrane protein	0.14				No
192	1	hypothetical protein	emb|CAH87333.1|	0.078	*Plasmodium chabaudi*	Pfam Replication initiator protein A	0.31				No
201	2	hypothetical protein	ref|NP_701827.1|	5E-17	*Plasmodium falciparum 3D7*	Pfam Merozoite surface protein	4E-12	mesoderm development	3E-18	0005523	Yes
204	1	hypothetical protein	ref|XP_643799.1|	2.6	*Dictyostelium discoideum*	Pfam Borrelia ORF-A	0.073				Yes
208	1	hypothetical protein	emb|CAJ16304.1|	0.13	*Trypanosoma brucei*	Pfam Protein of unknown function	1.1				No
214	1	hypothetical protein	gb|EAN31143.1|	0.4	*Theileria parva*	Smart Pulmonary surfactant proteins	0.058	sensory perception-chemical stimulus	0.087		No
215	1	hypothetical protein	emb|CAB16287.1|	9.6	*Schizosaccharomyces pombe*	Pfam Retinoblastoma-associated protein	0.17				No
218	1	hypothetical protein	ref|XP_643556.1|	0.4	*Dictyostelium discoideum*	Smart semaphorin domain	0.18				No
222	1	hypothetical protein	ref|NP_704043.1|	0.78	*Plasmodium falciparum 3D7*	Smart TLC/TRAM/LAG1/CLN8 domains	0.009				No
223	1	hypothetical protein	ref|XP_638775.1|	9.7	*Dictyostelium discoideum*	Pfam S-adenosyl-L-homocysteine hydrolase	0.31				No
225	1	hypothetical protein	ref|NP_704010.1|	0.39	*Plasmodium falciparum 3D7*	Pfam Domain of unknown function	0.014				No
226	1	hypothetical protein	pir||T31613	0.001		Pfam Sec-independent protein translocase	0.004				No
229	1	hypothetical protein	gb|AAP95837.1|	5.7	*Haemophilus ducreyi*	Smart Zinc/CREB-binding domain	0.65				Yes

Appendix IV - *Glossina pallidipes* clusters producing best matches to *Glossina morsitans morsitans* protein GeneDB

Cluster	Size (bp)	Best GeneDB Match	E-value
Odorant/ Pheromone Binding			
cn30	771	29% to LAR-001B13.b PBP-related protein 5 precursor	0.00067
cn31	529	23% to cn15569 General odorant-binding protein 99a	0.994
Energy Metabolism			
cn33	402	34% to GMsg73e09.p1k Cytochrome b (Fragment)	0.9993
gpafl-1a01.p1k	128	100% to cn11620 NADH-ubiquinone oxidoreductase chain 4	0.79
gpafl-2e11.q1k	105	69% to Gmm-11491 putative Cytochrome P450 4d2	0.99
gpafl-2h05.q1k	136	46% to Gmm-8608 putative Cytochrome P450 18a1	0.18
gpafl-6c10.p1k	444	37% to Gmm-1831 putative Vacuolar ATP synthase	0.81
gpafl-6d02.p1k	581	93% to cn1999 Cytochrome c oxidase polypeptide IV	5.0E-88
Nucleic Acid Metabolism			
cn77	363	47% to cn16179 40S ribosomal protein S5a	0.999
cn95	234	39% to cn7969 putative Ribosomal protein S2 (Fragment)	0.16
gpafl-1e10.p1k	333	41% to LAR-003I21.g Ribosomal protein L6e	0.62
gpafl-1d08.p1k	228	35% to cn3371 RNA-binding protein squid	0.2
gpafl-2c08.p1k	734	28% to cn12317 Reverse transcriptase (Fragment)	0.034
gpafl-6c10.q1k	291	48% to LAR-005N05.g Ribosomal protein L15 (Fragment)	0.13
gpafl-7b08.p1k	507	44% to GMsg-6130 putative 60S ribosomal protein L13	0.79
gpafl-9b05.q1k	157	53% to cn16006 chromatin transcription complex subunit	0.94
Structural			
cn1	505	36% to GMre-18h04.q1k Collagen alpha-1 (II) chain precursor	0.016
cn101	177	47% to LAR-005F17.g Insect cuticle protein family protein	0.036
gpafl-8f02.p1k	105	58% to Gmm-2730 putative Troponin T, skeletal muscle	0.64
gpafl-9c09.p1k	194	36% to Gmm-2730 putative Troponin T, skeletal muscle	0.051
gpafl-9e01.q1k	242	63% to cn13510 Cuticle protein 1 (Bc-NCP1)	0.00057
Metabolism			
cn22	251	36% to cn3410 steamer duck	0.48
cn25	510	58% to cn6749 GS1-like protein	0.15
cn58	457	28% to LAR-009C04.g Mitochondrial cysteine	0.9999
cn61	441	36% to cn14513 Male-specific sperm protein	0.95
cn76	364	24% to cn3382 Dystrophin, isoform B	0.97
cn104	96	52% to cn9439 tRNA-dihydrouridine synthase 2-like	0.99993
cn119	173	54% to GLAFJ50TV Carboxylesterase	0.9
cn120	793	45% to GLAF815TV Arginyl-tRNA synthetase	0.99993
cn126	111	50% to Gmm-11822 Probable ATP-dependent helicase	0.89
cn124	299	37% to cn16612 PFTAIRE-interacting factor 2 (CG31483-PA)	0.21
cn132	210	86% to cn16564 BSD domain containing protein	4.40E-21
cn139	389	40% to Gmm-2796 putative G-rich selenoprotein	0.92
gpafl-1b12.p1k	407	82% to cn3698 Ferritin heavy chain-like	5.30E-32
gpafl-2a12.p1k	272	41% to cn9380 Ornithine decarboxylase antizyme	0.048
gpafl-2c03.p1k	226	40% to Gmm-3295 putative Polyadenylate-binding protein	0.64
gpafl-2c07.q1k	130	72% to cn369 Protein flightless-1 (Flightless-I)	0.994
gpafl-2c08.q1k	302	37% to cn7046 Methionyl-tRNA synthetase, mitochondrial precursor	0.44
gpafl-2f06.p1k	621	52% to Gmm-9567 putative Septin interacting protein	0.73
gpafl-2h05.p1k	112	40% to cn9428 Dynein light chain 2, cytoplasmic	0.998
gpafl-3b07.q1k	134	41% to Gmm-0994 putative Clathrin-associated adaptor complex	0.93
gpafl-3c12.q1k	208	41% to cn1143 Acyl-coa dehydrogenase	0.77
gpafl-4g07.p1k	97	41% to cn15926 Failed axon connections protein (Fragment)	0.45
gpafl-5e09.p1k	570	92% to cn482 NSFL1 cofactor p47 (p97 cofactor p47)	4.8E-72
gpafl-6a02.q1k	113	83% to cn15926 Failed axon connections protein (Fragment)	0.55

Appendix IV cont.

Cluster	Size (bp)	Best GeneDB Match	E-value
Transport			
cn125	98	54% to Mitchondrial carnitine/acylcarnitine carrier protein	0.72
cn130	112	31% to PUM-112D03.g Mit. inner membrane translocase	0.94
gpafl-1c09.q1k	160	50% to GMsg77d10.q1k Organic cation transporter protein	0.31
gpafl-1h05.p1k	374	93% to GMsg-8198 putative ADP, ATP carrier protein	8.9E-59
gpafl-6g09.q1k	244	44% to cn1167 carrier protein (ADP/ATP translocase)	0.055
Signal transduction			
cn81	352	38% to cn14990 G protein-coupled receptor kinase 1	0.999
cn138	165	63% to cn3603 Phosrestin-2 (Phosrestin II) (Arrestin A)	0.78
cn156	228	31% to cn5273 Opsin Rh1 (Outer R1-R6 photoreceptor cells opsin)	0.11
gpafl-1d06.q1k	188	37% to TUM013_P18.b Gustatory receptor trehalose 1	0.43
gpafl-1h08.q1k	133	32% to GMsg25c01.q1k Gustatory receptor candidate 39	0.998
gpafl-6a07.q1k	92	56% to cn10928 Transmembrane GTPase Marf	0.53
gpafl-7e07.q1k	224	33% to cn7444 Retinoblastoma-family protein	0.0086
gpafl-7f12.p1k	304	37% to Gmm-10446 putative Probable GPCR Mth-like 9 precursor	0.996
Salivary Gland			
cn29	100	85% to GMsg-4370 Heterogeneous nuclear ribonucleoprotein	0.87
cn45	1219	56% to GMsg-2943 putative Diaphanous protein	2.3E-5
cn52	591	72% to GMsg-88d06.q1k Tsal1 protein precursor	0.999
cn82	347	35% to GMsg-43c12.q1k CG12008-PA, isoform A	0.98
cn97	223	50% to GMsg-9481 putative Histone H2B	0.61
cn107	78	66% to GMsg-98d10.p1k Alanine aminotransferase 2	0.98
cn140	137	50% to GMsg-172c11.p1k BTB/POZ domain-containing protein 9	0.89
gpafl-1c01.p1k	96	60% to GMsg-5352 Raf homolog serine/threonine-protein kinase	0.13
gpafl-1f03.p1k	195	30% to GMsg-0826 Selenophosphate synthetase 2-like protein	0.74
gpafl-5b06.p1k	191	47% to GMsg-02f08.p1k Medea-B	0.85
gpafl-6a05.p1k	628	36% to Gmm-1932 putative 14-3-3 protein epsilon	0.052
gpafl-6a07.p1k	169	46% to cn3824 Peptidase S1 & S6, chymotrypsin/Hap domain	0.6
gpafl-6b05.p1k	235	38% to cn3380 Largest subunit of the RNA polymerase II complex	0.14
gpafl-6d02.q1k	177	38% to Gmm-6852 putative mRNA capping enzyme	0.2
gpafl-6d10.p1k	387	52% to Gmm-1691 putative NipSnap protein	0.77
gpafl-6e06.p1k	65	100% to cn12514 Probable sulfate permease C3H7.02	0.995
gpafl-6e08.p1k	113	36% to cn6008 Tropomyosin-1, isoforms 9A/A/B (Tropomyosin II)	0.998
gpafl-6g05.p1k	224	32% to cn12876 Ser/Thr protein kinase PAR-1alpha	0.81
gpafl-7a12.p1k	356	32% to Gmm-2574 putative Cathepsin L precursor	0.38
gpafl-7b08.q1k	124	39% to cn4833 Arginine kinase	0.92
gpafl-7b12.q1k	199	46% to Gmm-7627 putative Heparin sulfate O-sulfotransferase	0.992
gpafl-7c06.p1k	179	31% to GMsg-6407 putative Ubiquitin-conjugating enzyme	0.8
gpafl-7c02.p1k	678	65% to cn6218 Y-box binding protein	0.0032
gpafl-7c07.p1k	687	25% to Gmm-10891 putative CORTACTIN	0.7
gpafl-7c11.p1k	236	41% to cn48 Imaginal disc growth factor 4	0.65
gpafl-7d01.p1k	355	80% to PUP-003M08.g Bee antimicrobial peptide family protein	0.23
gpafl-7d07.q1k	382	55% to cn7452 Trypsin alpha precursor	0.96
gpafl-7e01.p1k	136	41% to cn3294 Acyl-coenzyme A oxidase 1, peroxisomal	0.96
gpafl-7b08.q1k	124	39% to cn4833 Arginine kinase	0.92
gpafl-7b12.q1k	199	46% to Gmm-7627 putative Heparin sulfate O-sulfotransferase	0.992
gpafl-7c06.p1k	179	31% to GMsg-6407 putative Ubiquitin-conjugating enzyme	0.8
gpafl-7c02.p1k	678	65% to cn6218 Y-box binding protein	0.0032
gpafl-7c07.p1k	687	25% to Gmm-10891 putative CORTACTIN	0.7
gpafl-7c11.p1k	236	41% to cn48 Imaginal disc growth factor 4	0.65
gpafl-7d01.p1k	355	80% to PUP-003M08.g Bee antimicrobial peptide family protein	0.23
gpafl-7d07.q1k	382	55% to cn7452 Trypsin alpha precursor	0.96

Appendix IV
cont.

Cluster	Size (bp)	Best GeneDB Match	E-value	Cluster	Size (bp)	Best GeneDB Match	E-value
gpafl-5g10.p1k	348	52% to GMsg-7068 putative TSC1 protein	0.22	gpafl-7e01.p1k	136	41% to cn3294 Acyl-coenzyme A oxidase 1, peroxisomal	0.96
gpafl-6a04.p1k	114	50% to GMsg-13d02.q1k Tsal1 protein precursor	0.997	gpafl-7e04.p1k	528	6% to cn6762 Calpain B, Contains: Calpain B catalytic subunit 1	3
gpafl-6f10.p1k	56	50% to GMsg-2943 putative Diaphanous protein	0.29	gpafl-7e10.p1k	105	53% to cn14858 Proteinase inhibitor I1, Kazal domain protein	0.997
gpafl-7a02.q1k	335	50% to GMsg-87c12.q1k Pleckstrin homology domain	0.97	gpafl-7f01.q1k	566	32% to cn6781 Metallothionein-1 (MT-1)	0.43
gpafl-7b07.q1k	85	42% to GMsg-36a05.q1k Tsal1 protein precursor	0.97	gpafl-8b01.q1k	135	41% to cn7946 Matrix metalloproteinase 1	0.93
gpafl-7d05.p1k	439	42% to GMsg-9481 putative Histone H2B	0.23	gpafl-8c02.p1k	125	36% to cn12771 Mod (Mdg4)-h55.7	0.98
gpafl-7e12.q1k	295	48% to GMsg-96h03.q1k Dosage compensation regulator	0.59	gpafl-8g10.p1k	393	58% to cn2171 5'-nucleotidase	0.0043
gpafl-7f02.q1k	61	100% to GMsg-1858 putative Diaphanous protein	0.84	gpafl-8e05.p1k	477	33% to Gmm-10891 putative CORTACTIN	0.53
gpafl-8e05.q1k	308	36% to GMsg-5145 putative DnaJ-like protein 60	0.00093	gpafl-9a03.q1k	143	38% to cn4721 Tetratricopeptide region domain containing protein	0.996
gpafl-8e10.p1k	269	43% to GMsg-15f09.p1k Mitotic checkpoint/RNA export protein	0.26	gpafl-9b01.q1k	279	42% to cn797 Ferritin light-chain	0.61
gpafl-9b12.q1k	146	23% to GMsg-6889 putative XL6 protein	0.994	gpafl-9c02.q1k	156	33% to PUF-105D09.b Alpha esterase	0.4
gpafl-9c02.p1k	153	63% to GMsg-08a08.p1k Gamma-tubulin complex component 3	0.8	gpafl-9e05.p1k	622	29% to Gmm-10891 putative CORTACTIN	0.49
gpafl-9g03.q1k	112	57% to GMsg-95g09.p1k Tsal1 protein precursor	0.991	gpafl-9e11.q1k	110	66% to cn4317 Protein expanded	0.12
gpafl-2c05.p1k	90	2% to Tse64g05.p1c S-adenosylmethionine synthetase 4	0.75				
gpafl-3b07.p1k	108	62% to Tse91c02.q1c Retrotransposon hot spot protein, RHS3	0.996				
gpafl-5e08.p1k	243	41% to Tse80e02.p1c RINT1-like protein (DmRINT-1)	0.27				
gpafl-7c07.q1k	276	37% to Tse129h11.q1c CG5427-PA	0.47				
gpafl-7e02.q1k	233	38% to Tse114g11.q1c MADF domain containing protein	0.6				
gpafl-7f07.q1k	126	46% to Tse59c01.p1c Leucine-rich repeat	0.72				

cnxxx - *Glossina pallidipes* consensus sequences; gpafl-xxxx.p1k – *Glossina pallidipes* singleton

Appendix V - *Glossina palpalis gambiensis* clusters producing best matches to *Glossina morsitans morsitans* protein GeneDB

Cluster	Size (bp)	Best GeneDB Match	E-value
Odorant Binding			
cn109	746	31% to LAR-001B13.b PBP-related protein 5 precursor (PBPRP-5)	0.00064
gphfl-8a06.p1k	488	26% to cn7404 General odorant-binding protein 99b precursor	0.99994
gphfl-5g01.p1k	568	67% to cn14025 Odorant binding protein	1.1E-44
gphfl-8d10.p1k	658	35% to cn15569 General odorant-binding protein 99a precursor	8.5E-22
gphfl-9d09.p1k	58	40% to cn15569 General odorant-binding protein 99a precursor	0.997
Energy Metabolism			
cn2	553	77% to cn9829 mitochondrial NADH-ubiquinone oxidoreductase	1.9E-27
cn4	634	68% to PUM-101L21.g Cytochrome oxidase subunit I	0.2
cn5	488	80% to Gmsg53b04.p1k Cytochrome oxidase subunit II	2.5E-18
cn6	676	97% to cn10291 Cytochrome c oxidase subunit 3	5.5E-47
cn15	684	100% to cn3694 Mitochondrial-processing peptidase subunit beta	5.0E-33
cn79	665	36% to PUM-113B01.g Cytochrome c oxidase polypeptide VIII	4.1E-01
cn88	420	97% to cn13470 Cytochrome c oxidase polypeptide VIII	1.6E-31
cn98	318	47% to cn5050 NADH-ubiquinone oxidoreductase chain 2	0.34
cn114	701	87% to cn5432 ATPase subunit 6	4.4E-41
cn115	142	91% to cn10291 Cytochrome c oxidase subunit 3	5.7E-06
cn116	508	91% to cn10291 Cytochrome c oxidase subunit 3	2.5E-37
gphfl-2a08.p1k	115	50% to cn10293 Cytochrome c oxidase subunit 3	1.0
gphfl-2g11.p1k	267	53% to cn3977 Cytochrome P450	6.9E-01
gphfl-3g03.q1k	103	66% to Tse65g07.p1c Cytochrome oxidase subunit II	9.97E-01
gphfl-5e01.p1k	593	94% to cn9606 ATP synthase subunit g, mitochondrial	8.9E-46
gphfl-5e07.p1k	607	80% to cn14982 Cytochrome c oxidase subunit 5A	1.1E-81
gphfl-5h02.p1k	567	97% to PUF-101C07.b NADH dehydrogenase	1.1E-49
gphfl-8b10.p1k	615	96% to cn13916 NADH dehydrogenase 1 beta subcomplex	3.2E-77
gphfl-8c10.p1k	578	97% to cn13743 NADH dehydrogenase 1 alpha subcomplex	1.1E-76

Cluster	Size (bp)	Best GeneDB Match	E-value
Metabolism			
cn3	463	100% to cn15353 Paramyosin, long form	3.1E-16
cn16	622	40% to cn12243 WD repeat protein 89	0.81
cn20	300	50% to cn4870 Annexin-B11	0.994
cn67	169	87% to cn8537 Derlin-1 (DER1-like protein 1)	0.68
cn68	168	37% to GLAAX73TH Lipase 1 precursor, putative	0.8
cn81	547	39% to cn16418 Larval serum protein 1 gamma	0.34
cn86	840	20% to Gmm-10891 putative CORTACTIN	0.89
cn89	709	99% to cn3089 Amidophosphoribosyltransferase	2.7E-88
cn91	396	52% to Gmm-10891 putative CORTACTIN	9.8E-01
cn118	402	41% to cn4283 DEAD box ATP-dependent RNA helicase	9.2E-22
gphfl-1a05.q1k	113	72% to GLAF430TV Serine protease inhibitor 4	9.9E-01
gphfl-1b02.p1k	80	92% to cn3507 Alanine-glyoxylate transaminase 1	7.8E-01
gphfl-1b02.q1k	101	87% to cn3507 Alanine-glyoxylate transaminase 1	2.3E-01
gphfl-1c12.p1k	601	31% to PUM-113B11.b Vacuolar ATP synthase proteolipid	8.0E-01
gphfl-1f12.q1k	255	40% to PUM-105P23.b Trypsin precursor	4.5E-01
gphfl-2c01.p1k	437	43% to Gmm-3119 putative Rexin L3	9.5E-01
gphfl-2c08.q1k	305	37% to cn10472 Smad7 (Fragment),	9.9E-01
gphfl-2c09.q1k	111	36% to cn5485 GPI inositol-deacylase	9.99E-01
gphfl-2c12.p1k	293	31% to cn10138 Carbonic anhydrase, eukaryotic family protein	8.6E-01
gphfl-2d11.p1k	355	27% to cn14999 Proteinase inhibitor I2, Kunitz metazoa domain	6.4E-01
gphfl-2e05.q1k	107	87% to cn5753 Transient-receptor-potential-like protein	1.0
gphfl-2f06.q1k	249	50% to LAR-007N03.b Mitochondrial associated cysteine-rich	6.8E-01
gphfl-2f10.p1k	199	36% to cn5197 Protein ariadne-1 (Ari-1)	9.8E-01
gphfl-2f11.p1k	132	45% to cn10094 Saposin-like type B, 1 domain containing protein	2.6E-01

Appendix V
cont.

Cluster	Size (bp)	Best GeneDB Match	E-value
gphfl-8e04.q1k	116	75% to cn5236 NADH-ubiquinone oxidoreductase chain 5	9.9E-01
gphfl-8e08.p1k	697	90% to cn9962 NADH dehydrogenase [ubiquinone] iron-sulfur protein	2.0E-100
gphfl-9c09.p1k	669	97% to cn1940 ATP synthase B chain, mitochondrial precursor	4.9E-104
gphfl-9e09.q1k	516	78% to cn14468 NADH dehydrogenase	2.8E-35
Nucleic Acid Metabolism			
cn1	685	83% to cn840 60S ribosomal protein L15	1.4E-81
cn49	841	30% to cn3379 RNA polymerase II largest subunit	4.7E-03
cn53	504	31% to cn3379 RNA polymerase II largest subunit	9.9E-01
cn111	896	97% to cn11956 RNA-directed RNA polymerase, P3D protein	1.2E-114
cn119	793	30% cn1504 Translation elongation factor 2 (EF-2)	0.93
gphfl-2a10.q1k	180	37% to GMsg-4148 putative RNA-binding protein cabeza	0.16
gphfl-2c01.q1k	146	44% to Gmsg-2447 putative 60S ribosomal protein L7	0.48
gphfl-2d08.p1k	219	35% to cn3371 RNA-binding protein squid	0.25
gphfl-2h11.q1k	198	50% to GMre-04d03.q1k Transcription factor Adf-1	0.998
gphfl-7d07.p1k	412	100% to Gmm-2661 Heterogeneous nuclear ribonucleoprotein	0.74
gphfl-8a07.p1k	399	100% to cn14492 60S ribosomal protein L36 (Protein minute(1)	1.2E-54
gphfl-8g08.p1k	410	100% to cn13635 60S ribosomal protein L38	3.7E-33
gphfl-8g10.p1k	702	92% to GMsg-6767 40S ribosomal protein S2	9.8E-115
gphfl-8h01.p1k	190	100% to cn13583 60S ribosomal protein L23 (L17A)	2.1E-14
gphfl-8h07.p1k	715	98% to cn2789 Ribosomal protein L5 (Fragment)	7.9E-111
Structural			
cn82	248	29% to cn4456 Actin-87E	0.7
cn110	547	88% to cn13615 Cuticle protein	8.0E-25
cn113	588	87% to cn5217 Adult cuticle protein 1 precursor (dACP-1)	3.2E-53
gphfl-5g03.p1k	548	92% to cn3635 Endocuticle structural glycoprotein ABD-4	5.9E-58
gphfl-5g11.p1k	646	97% to GLAAL91TH UNC45 (Smooth muscle cell protein1)	5.1E-70
gphfl-9d01.p1k	741	84% to cn718 Histone H3.3	1.1E-71

Cluster	Size (bp)	Best GeneDB Match	E-value
gphfl-2g01.p1k	390	37% to cn950 ATP-dependent RNA helicase p62	9.93E-01
gphfl-2g02.q1k	178	50% to cn604 Vitellogenin-2 precursor (Vitellogenin II)	7.6E-01
gphfl-2g06.q1k	112	71% to cn5758 Transient-receptor-potential-like protein	9.7E-01
gphfl-2g11.q1k	256	60% to cn50 Imaginal disc growth factor 4	0.7
gphfl-2h01.p1k	74	47% to cn12665 CG7769-PA (Damage-specific DNA protein)	0.9999
gphfl-3b08.p1k	175	29% to cn6 selenium-binding protein	0.87
gphfl-5b03.p1k	429	81% to cn49 Imaginal disc growth factor 4	3.6E-58
gphfl-5c02.p1k	474	20% to Gmm-10891 putative CORTACTIN	0.71
gphfl-5d04.p1k	615	32% to cn14311 SR family splicing factor SC35 (LD32469p)	0.000001
gphfl-5d08.p1k	89	100% to cn11741 PFTAIRE-interacting factor 1A (CG33719-PA,	0.0061
gphfl-5e02.p1k	492	53% to Gmm-2344 putative Congested-like trachea protein	0.0014
gphfl-5e06.p1k	601	34% to LAR-004G19.g Sperm mit-associated cysteine-rich protein	0.99992
gphfl-5e08.p1k	653	88% to cn9015 Profilin (Protein chickadee)	6.20E-90
gphfl-5f08.p1k	486	34% to cn4416 Gamma actin-like protein	1.0E-100
gphfl-5f10.q1k	505	37% to cn4930 1-pyrroline-5-carboxylate dehydrogenase 1	6.1E-15
gphfl-5g04.p1k	367	37% to PUP-005O21.g Vacuolar ATP synthase subunit D 1	0.95
gphfl-5h10.p1k	649	57% to Glycoprotein-N-acetylgalactosamine-3-ß-galactosyltransferase	3.6E-21
gphfl-7f03.p1k	181	37% to LAR-005I13.b Nascent polypeptide complex subunit α	2.4E-01
gphfl-8a05.p1k	487	44% to cn14907 Serine protease persephone precursor	5.1E-08
gphfl-8a10.q1k	209	36% to cn5649 Dihydroorotate dehydrogenase, mitochondrial	0.63
gphfl-8c06.p1k	486	35% to cn3341 Sn1-specific diacylglycerol lipase alpha	0.999
gphfl-8c08.q1k	334	46% to Gmm-8019 putative Diaphanous protein	0.97
gphfl-8d04.p1k	190	54% to cn12125 Protein spitz precursor	0.83
gphfl-8d07.p1k	598	93% to cn2376 Calcyclin-binding protein (CacyBP)	9.3E-58
gphfl-8e06.p1k	555	84% to cn3698 Ferritin heavy chain-like	1.0E-44
gphfl-8f01.p1k	574	70% to cn15801 Two-Kunitz protease inhibitor	3.3E-39
gphfl-8f08.p1k	553	39% to cn14188 Ankyrin	8.0E-10

Appendix V cont.

Cluster	Size (bp)	Best GeneDB Match	E-value
Transport			
gphfl-1a02.p1k	641	91% to cn3215 Acyl carrier protein, mitochondrial precursor (ACP)	3.0E-67
gphfl-1a02.q1k	564	52% to cn15856 Acyl carrier protein, mitochondrial precursor (ACP)	3.2E-20
gphfl-2d10.p1k	248	47% to cn7468 Sodium/potassium-transporting ATPase subunit	0.99
gphfl-5b10.p1k	572	84% to cn12247 Tricarboxylate transport protein	1.5E-70
gphfl-5c11.p1k	726	81% to cn1167 ADP, ATP carrier protein (ADP/ATP translocase)	7.5E-77
gphfl-8a09.p1k	147	50% to cn6319 Translocation-associated membrane protein 1	0.92
Signal transduction			
cn13	147	91% to cn5249 Rhodopsin 1 Fragment	5.2E-06
cn84	1216	93% to cn5253 opsin Rh1 (Outer R1-R6 photoreceptor cells opsin)	6.1E-155
cn112	765	86% to cn3533 Phosrestin-2 (Phosrestin II)	4.5E-96
gphfl-1e04.q1k	140	95% to cn5249 Rhodopsin 1 (Fragment)	8.6E-06
gphfl-1f09.q1k	141	53% to cn11426 Transmembrane and TPR repeat-containing protein	0.71
gphfl-5d02.p1k	540	29% to cn16801 Retinin-like protein family protein	0.00021
Salivary Gland			
cn8	227	40% to Tse104g08.q1c LD18186p (CG12489-PA)	0.96
cn14	2019	58% to GMsg68f01.p1ka Nucleosome assembly protein NAP-1	0.2
cn37	411	37% to Tse110g12.q1c Rabconnectin	0.92
cn56	364	47% to Tse104g08.q1c LD18186p (CG12489-PA)	0.17
cn58	315	34% to GMsg41h07.q1k Tsal1 protein precursor	0.54
cn66	175	50% to GMsg25g01.q1k Tsal1 protein precursor	0.96
cn95	155	37% to GMsg-0682 putative Zinc finger motif protein	9.9E-01
cn106	99	66% Gmsg07c04.p1k Tsal1 protein precursor	0.9992
cn108	529	65% to cn13931 Putative secreted salivary protein	5.4E-26
gphfl-1c12.q1k	468	43% to GMsg-9054 putative Pygopus protein	4.6E-01
gphfl-2a09.q1k	89	52% to Gmsg107c10.p1k crinkled	9.98E-01
gphfl-2f12.p1k	121	33% to Tse5e01.p1c Iron regulatory protein 1B (SD12606p)	7.8E-01
gphfl-8f09.p1k	523	79% to cn2901 Troponin I-b2	7.2E-37
gphfl-8h10.q1k	172	30% to Gmm-9756 putative Serine/threonine protein phosphatase	0.31
gphfl-9a01.p1k	681	98% to cn16483 Protein big brother	5.5E-89
gphfl-9c03.p1k	689	85% to cn11956 RNA-directed RNA polymerase, P3D protein family	2.3E-98
gphfl-9c05.p1k	653	88% to cn12618 Dorsal-ventral patterning protein	1.3E-82
gphfl-9c06.p1k	341	23% to cn14710 Major facilitator superfamily MFS_1	6.9E-01
gphfl-9d11.p1k	425	27% to cn9034 Lingerer protein type1	4.9E-01
gphfl-9f01.p1k	653	98% to cn460 Calmodulin-A (CaM A)	4.7E-88
gphfl-9f10.p1k	688	26% to cn6382 Prefoldin domain containing protein	9.93E-01
gphfl-10h01.p1k	1025	88% to cn4930 1-pyrroline-5-carboxylate dehydrogenase 1	1.4E-74

Cluster	Size (bp)	Best GeneDB Match	E-value
gphfl-2h09.q1k	120	55% to Gmsg-10331 putative DNA-directed RNA polymerase I	9.4E-01
gphfl-2g05.q1k	183	46% to GMsg74d10.q1k Protein Tube	9.7E-01
gphfl-2f06.p1k	228	42% to GMsg50g08.p1k Tsal1 protein precursor	6.1E-01
gphfl-2f10.q1k	220	30% toGMsg91g10.p1k Tsal1 protein precursor	0.08
gphfl-5c07.q1k	247	44% to GMsg-6886 putative Caspase-8 precursor	0.13
gphfl-5c09.q1k	136	42% to Tse115h06.q1c Nucleoporin 98-96 (Fragment)	0.73
gphfl-5d05.p1k	362	33% to Tse20e12.q1c Projectin (Fragment)	0.14
gphfl-5g09.p1k	760	37% to Tse31e04.q1c Predicted membrane protein	0.997
gphfl-5g10.q1k	206	47% to Tse68e10.q1c Kinesin-associated family protein	0.41
gphfl-5h09.p1k	157	33% to cn8884 Salivary gland growth factor-2	0.17
gphfl-7d12.q1k	115	42% to GMsg110h11.p1k Nuclear hormone receptor HR96 (dHR96)	0.9995

cnxxx - *Glossina palpalis gambiensis* consensus sequences; gphfl-xxxx.p1k - *Glossina palpalis gambiensis* head singleton.

Appendix VI - *Glossina tachinoides* clusters producing best matches to *Glossina morsitans morsitans* protein GeneDB

Cluster	Size (bp)	Best GeneDB Match	E-value	Cluster	Size (bp)	Best GeneDB Match	E-value
Odorant Binding				**Metabolism**			
gthfl-5d10.q1k	393	28% to cn15569 General odorant-binding protein 99a precursor	6.1E-07	cn3	610	90% to cn1796 ATP synthase coupling factor 6, mitoch precursor	4.8E-65
Energy Metabolism				cn9	660	95% to cn460 Calmodulin-A (CaM A)	1.0E-69
cn6	471	92% to cn9925 NADH-ubiquinone oxidoreductase chain 4	1.2E-23	cn17	1276	33% to Gmm-10891 putative CORTACTIN	9.6E-01
cn19	930	97% to cn10291 Cytochrome c oxidase subunit 3	5.4E-47	cn21	895	35% to Gmm-10148 putative Nuclear hormone receptor HR96	0.36
cn23	521	76% to cn7406 Ubiquinol-cytochrome c reductase complex	1.2E-30	cn25	505	35% to cn139 Carnitine palmitoyltransferase I	0.56
cn33	359	84% to cn14468 NADH dehydrogenase	9.2E-07	cn35	337	47% to cn50 Imaginal disc growth factor 4	0.063
cn57	555	93% to cn13205 Cytochrome c oxidase polypeptide Via	5.7E-53	cn36	336	33% to cn8930 Coproporphyrinogen III oxidase	0.94
cn58	886	47% to Gmsg82h07.p1k Cytochrome b (Fragment)	8.0E-31	cn38	311	45% to cn9299 Elongation factor 1-alpha (EF-1-alpha)	0.34
cn60	669	97% to cn3108 Probable cytochrome c oxidase polypeptide VIIa	7.0E-32	cn42	248	31% to cn12648 Sticky ch1 (Fragment)	0.69
cn69	754	96% to cn10291 Cytochrome c oxidase subunit 3	4.0E-45	cn43	247	41% to cn1251 Gelsolin (Fragment)	0.99
gthfl-1a11.p1k	563	95% to cn1999 Cytochrome c oxidase polypeptide IV	9.4E-87	cn47	212	32% to cn15475 T-cell lymphoma breakpoint associated target protein	0.64
gthfl-1a11.q1k	470	88% to cn1999 Cytochrome c oxidase polypeptide IV	1.1E-21	cn55	823	82% to cn4165 ATP synthase delta chain, mitochondrial	5.1E-61
gthfl-1c07.p1k	442	97% to cn13690 Cytochrome c oxidase subunit VIIc	5.1E-31	cn62	602	88% to cn3887 Laminin subunit gamma-1 precursor	1.3E-06
gthfl-1c07.q1k	381	87% to cn13690 Cytochrome c oxidase subunit VIIc	8.1E-26	cn65	613	95% to cn13725 Ejaculatory bulb-specific protein 3 precursor	2.8E-62
gthfl-2b11.q1k	570	69% to cn9925 NADH-ubiquinone oxidoreductase chain 4	6.1E-19	cn67	518	61% to cn2301 N(2),N(2)-dimethylguanosine tRNA methyltransferase	7.7E-06
gthfl-2c08.p1k	417	73% to Gmm-8685 putative Probable cytochrome P450 311a1	2.6E-01	cn68	424	96% to cn221 ATP synthase gamma chain, mitochondrial precursor	6.2E-41
gthfl-4a05.p1k	246	66% to GMsg53b04.p1k Cytochrome oxidase subunit II	9.6E-01	cn75	771	95% to cn5310 Calcium-tran ATPase sarcoplasmic/endo reticulum	4.1E-105
gthfl-4b01.q1k	298	66% to cn5432 ATPase subunit 6	2.6E-12	cn78	531	56% to cn3779 ATPase inhibitor homolog, mitochondrial precursor	2.2E-18
gthfl-4b11.q1k	157	91% to cn10291 Cytochrome c oxidase subunit 3	1.8E-08	cn83	685	83% to cn7418 ATP synthase lipid-binding protein	1.5E-45
gthfl-4c01.p1k	790	71% to GMsg144h12.p1k NADH dehydrogenase subunit 5	3.5E-24	gthfl-1d08.p1k	595	100% to cn13644 Trp protein	7E-9
gthfl-4h09.q1k	127	50% to GLABL12TV AAA+ ATPase	8.4E-01	gthfl-1e06.q1k	205	40% to cn6566 Protein YIPF5 (YIP1 family member 5)	0.49
gthfl-5b10.q1k	388	82% to cn13916 NADH DH [ubiquinone] 1 beta subcomplex	9 8E-37	gthfl-1e08.q1k	281	68% to cn16618 Beta-NAC-like protein	2.9E-17
gthfl-5c10.q1k	208	55% to cn13743 NADH DH [ubiquinone] 1 α-subcomplex	1.0E-01	gthfl-1g03.q1k	447	53% to Gmm-9433 putative Ubiquitin-conjugating enzyme E2-17 kDa	2.8E-01

Appendix VI
cont.

Cluster	Size (bp)	Best GeneDB Match	E-value
gthfl-5e08.q1k	371	45% to cn9962 NADH DH [ubiquinone] iron-sulfur protein 7	5.4E-11
gthfl-5h04.q1k	176	23% to PUM-113B01.g Cytochrome c oxidase polypeptide VIII	9.7E-01
gthfl-6b01.q1k	398	88% to cn5432 ATPase subunit 6	1.6E-40
gthfl-7h05.p1k	662	85% to cn11626 NADH dehydrogenase subunit 4	8.2E-20
Nucleic Acid Metabolism			
cn8	901	81% to cn2826 60S ribosomal protein L5	1.3E-30
cn10	1160	91% to cn11956 RNA-directed RNA polymerase	4.3E-164
cn34	353	52% to cn3189 Activating transcription factor	6.6E-01
gthfl-2h05.p1k	199	50% to cn8660 28S ribosomal protein S5	7.6E-01
gthfl-4c07.q1k	330	78% to cn30 Ribosomal protein L35Ae	3.0E-27
gthfl-4e10.p1k	460	37% to LAR-005F13.g 60S ribosomal protein L36	1.7E-01
gthfl-4f03.q1k	241	63% to cn1435 39 kDa FK506-binding nuclear protein	3.1E-01
gthfl-5a07.q1k	379	93% to cn14492 60S ribosomal protein L36	3.3E-47
gthfl-5f03.q1k	154	58% to GMre-05b12.q1k 40S ribosomal protein S3a	5.9E-01
gthfl-5g08.q1k	205	89% to cn13635 60S ribosomal protein L38	1.7E-03
gthfl-5h01.q1k	209	100% to cn13583 60S ribosomal protein L23 (L17A)	6.5E-16
gthfl-5h07.q1k	409	92% to cn2795 Ribosomal protein L5	3.6E-35
gthfl-5g10.q1k	458	63% to cn7965 40S ribosomal protein S2	2.7E-41
gthfl-6c06.q1k	130	42% to cn6031 60S acidic ribosomal protein P1 (RP21C)	9.7E-01
gthfl-6d11.q1k	265	43% to cn12819 Eukaryotic TIF 5A (eIF-5A)	9.9E-01
gthfl-8b08.p1k	281	42% to cn3189 Activating transcription factor	5.5E-01
gthfl-8c02.q1k	457	47% to cn11956 RNA-directed RNA polymerase	1.7E-03
gthfl-16e08.p1k	976	99% to LAR-007D18.b Transcription initiation factor IIA	2.1E-53

Cluster	Size (bp)	Best GeneDB Match	E-value
gthfl-1h09.q1k	213	40% to cn11882 Ser/threonine-protein phosphatase 4 catalytic subunit	9.92E-01
gthfl-2d09.p1k	489	85% to cn1876 Adenylosuccinate lyase	2.2E-71
gthfl-2g07.q1k	144	39% to Gmm-9432 putative Protein-cysteine N-palmitoyltransferase	9.3E-01
gthfl-3b12.q1k	121	38% to cn2313 Microsomal glutathione S-transferase-like protein	1.0E+00
gthfl-3h06.q1k	540	41% to cn8813 Brain protein 44	8.1E-01
gthfl-4a09.p1k	228	52% to cn884 Dolichyl-diphosphooligosaccharide glycosyltransferase	2.8E-01
gthfl-4b05.p1k	347	42% to Gmm-10891 putative CORTACTIN	9.7E-01
gthfl-4c04.q1k	148	53% to cn15796 Aldose 1-epimerase family protein	9.6E-01
gthfl-4c07.p1k	348	54% to cn15902 Troponin T-4 (Fragment)	2.7E-09
gthfl-4c11.p1k	259	57% to cn12991 RabGAP/TBC	8.0E-01
gthfl-4d04.q1k	128	48% to Gmm-8207 putative Myosin IA	9.4E-01
gthfl-4d07.p1k	157	75% to cn4101 Dynamin associated protein isoform Dap160-1	9.95E-01
gthfl-4d11.q1k	256	33% to GLAC352TV Superoxide dismutase, copper/zinc binding	9.5E-01
gthfl-4e09.p1k	202	82% to cn482 NSFL1 cofactor p47 (p97 cofactor p47)	4.6E-11
gthfl-4f10.p1k	663	31% to cn6792 Aldehyde dehydrogenase, mitochondrial	5.1E-14
gthfl-4g08.q1k	230	81% to cn3244 Superoxide dismutase [Mn], mitochondrial precursor	4.1E-07
gthfl-4g11.q1k	404	87% to GLAAL91TH UNC45 homolog A-Smooth muscle cell protein 1	2.1E-27
gthfl-4h06.p1k	156	41% to cn2955 TXBP181-like protein (CG2072-PA)	5.1E-01
gthfl-4h12.p1k	696	64% to cn10818 Heat shock protein Hsp40	2.1E-41
gthfl-5a03.q1k	103	43% to cn16764 Calreticulin precursor	9.94E-01
gthfl-5b04.q1k	248	39% to LAR-002O19.b Fatty acyl-CoA elongase	6.8E-01
gthfl-5e06.q1k	393	45% to cn3698 Ferritin heavy chain-like	2.3E-12
gthfl-5f01.q1k	368	47% to cn15801 Two-Kunitz protease inhibitor	3.7E-07

Appendix VI
cont.

Cluster	Size (bp)	Best GeneDB Match	E-value
Structural			
cn52	911	57% to cn13836 Adult cuticle protein 1 precursor (dACP-1)	2.6E-62
cn53	624	60% to cn8742 Adult cuticle protein 1 precursor (dACP-1)	2.1E-54
gthfl-4g01.q1k	190	66% to cn5187 Tubulin alpha-1 chain	3.5E-01
gthfl-5b03.q1k	211	58% to cn5223 Adult cuticle protein 1 precursor (dACP-1)	7.7E-10
gthfl-6d01.q1k	435	48% to Gmm-02e05.p1k Histone H3.3 type 1	9.6E-08
Transport			
gthfl-1h10.p1k	156	37% to Gmm-7520 putative Probable GDP-fructose transporter	3.1E-01
gthfl-4a12.p1k	682	99% to Gmm-3336 Sodium/potassium-transporting ATPase	1.1E-92
gthfl-4b10.q1k	535	81% to cn12247 Tricarboxylate transport protein	1.9E-61
gthfl-4c11.q1k	117	91% to cn1169 ADP, ATP carrier protein)	7.8E-06
gthfl-4h02.q1k	331	44% to cn7983 Sodium:neurotransmitter symporter	9.1E-01
Signal transduction			
cn1	97	37% to Gmm-7284 putative Probable GPCR Mth-like precursor	0.55
cn15	229	56% to cn5258 Opsin Rh1 (Outer R1-R6 photoreceptor cells)	0.016
cn59	732	83% to cn3533 Phosrestin-2 (Phosrestin II) (Arrestin A)	2.4E-84
cn87	237	75% to Gmm-10447 putative Probable GPCR Mth	1.9E-01
gthfl-1e10.q1k	146	84% to cn5249 Rhodopsin 1	2.6E-02
gthfl-4d09.q1k	129	41% to PUP-005M12.g G protein-coupled receptor kinase 1	0.97
gthfl-7b01.p1k	692	80% to cn3528 Phosrestin-2 (Phosrestin II) (Arrestin A)	2.0E-89
gthfl-8b07.q1k	246	76% to cn3533 Phosrestin-2 (Phosrestin II) (Arrestin A)	2.8E-12
gthfl-18b01.p1k	1357	85% to cn5034 Phosrestin-1 (Phosrestin I) (Arrestin B)	9.6E-110
Salivary Gland			
cn2	571	51% to Tse128b04.q1c 60S ribosomal protein L28-like protein	5.6E-09
cn11	1784	98% to Gmsg-0907 putative Opsin Rh1	1.2E-205
gthfl-5f02.q1k	328	37% to cn16368 NEB-CGP=CAMP generating peptide	1.0
gthfl-5f09.q1k	80	72% to Gmm-11567 putative Tunen	1.0
gthfl-5f11.q1k	159	35% to Gmm-10891 putative CORTACTIN	9.7E-01
gthfl-5g03.q1k	117	62% to cn3876 Phosphoinositide 3-kinase	9.9E-01
gthfl-6a01.q1k	495	91% to cn8844 Protein big brother	9.1E-46
gthfl-6c06.p1k	144	53% to cn7284 Peptidase S1 and S6, chymotrypsin/Hap	5.2E-01
gthfl-6c09.q1k	478	73% to cn1934 ATP synthase B chain, mitochondrial precursor	1.8E-26
gthfl-6c10.q1k	107	42% to cn261Protein Mo25 (dMo25)	9.9E-01
gthfl-6d12.q1k	265	34% to GLAAQ06TV Moesin/ezrin/radixin homolog 2	2.3E-01
gthfl-6e04.p1k	653	30% to Gmm-1607 putative Hook protein	9.2E-01
gthfl-6e08.q1k	315	32% to cn14304 Palmitoyltransferase ZDHHC3	1.0
gthfl-6f01.p1k	307	45% to cn9227 ATP-dependent RNA helicase vasa	5.3E-15
gthfl-6g09.q1k	329	70% to cn11468 Protein Peter pan	9.0E-05
gthfl-7g04.p1k	718	94% to cn2130 Ubiquitin-conjugating enzyme E2 G2	5.9E-91
gthfl-8a07.p1k	208	30% to cn10843 Zinc finger, CCCH-type domain containing protein	1.9E-01
gthfl-8a10.q1k	194	40% to cn8570 Chaperonin	7.4E-01
gthfl-8b01.p1k	313	91% to cn16483 Protein big brother	2.0E-36
gthfl-8b02.p1k	217	58% to cn797 Ferritin light-chain	3.7E-01
gthfl-8b07.p1k	312	53% to cn12215 Zinc finger/leucine zipper protein DALF isoform C3	9.97E-01
gthfl-8c01.p1k	290	27% to Gmre-05a07.p1k DNA primase small subunit	1.0
gthfl-8c05.q1k	127	60% to cn4278 Protein jagunal homolog 1	9.92E-01
gthfl-8c07.p1k	579	32% to Gmm-2730 putative Troponin T, skeletal muscle	0.9999
gthfl-8c10.p1k	153	46% to cn5744 Transient-receptor-potential-like protein	0.83
gthfl-8c12.p1k	425	50% to Gmm-2574 putative Cathepsin L precursor	0.998
gthfl-14h01.p1k	1041	73% to cn12539 Nesprin	1.4E-36

Appendix VI
cont.

Cluster	Size (bp)	Best GeneDB Match	E-value
cn74	563	65% to cn13931 Putative secreted salivary protein	5.8E-26
cn81	533	29% to Gmsg124b06.q1k Tsal1 protein precursor	0.994
gthfl-1c05.p1k	431	65% to Gmsg-6716 putative A kinase anchor protein 200	6.7E-23
gthfl-1e09.q1k	98	66% to Gmsg-5145 putative DnaJ-like protein 60	0.999
gthfl-1g09.p1k	161	43% to GMsg30g11.q1k Tsal1 protein precursor	0.79
gthfl-1h03.p1k	506	42% to Tse91c02.q1c Retrotransposon hot spot protein, RHS3	0.45
gthfl-1h08.q1k	566	80% to Gmsg13a02.p1k DNA-directed RNA polymerase III	1.2E-68
gthfl-3b01.q1k	134	87% to GMsg15f09.p1k Mitotic checkpoint protein & RNA protein	4.3E-02
gthfl-4d11.p1k	94	32% to Gmsg-7692 putative Arginine kinase	9.5E-01
gthfl-5b09.p1k	487	66% to GMsg75b08.p1k Glycylpeptide N-tetradecanoyltransferase	9.4E-01
gthfl-5b10.p1k	286	42% to Tse28c06.q1c Protein roadkill	4.8E-01
gthfl-5d08.q1k	237	48% to GMsg-5449 Selenophosphate synthetase 2-like protein	9.95E-01
gthfl-5e06.p1k	125	62% to GMsg53b04.p1k Cytochrome oxidase subunit II	0.00064
gthfl-5g02.q1k	244	52% to Gmsg-9068 Serine/threonine protein phosphatase PP-V	0.32
gthfl-6c05.q1k	186	42% to GMsg55h10.p1k Mucolipin	0.52
gthfl-8a02.p1k	85	44% to GMsg32e12.q1k Peptidase,trypsin-like serine & cysteine	0.74
gthfl-8b09.p1k	184	42% to GMsg161a05.q1k Myb protein	2.0E-01
gthfl-8b10.p1k	371	45% to Gmsg19h08.p1k Tsal1 protein precursor	0.74
gthfl-8b11.p1k	198	53% to Gmsg107f12.p1k phospholipid-transporting ATPase VA	9.6E-01
gthfl-9b01.p1k	335	53% to Tse14c06.p1c Pro3 precursor	9.95E-01

cnxxx - *Glossina tachinoides* consensus sequences; gthfl-xxxx.q1k - *Glossina tachinoides* head singletons

Appendix VII - *Glossina pallidipes*, *Glossina palpalis gambiensis* and *Glossina tachinoides* clusters and their best matches to *Drosophila melanogaster* protein databases

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Odorant Binding				
Gpa-cn30	Pbprp3	FBpp0078305	8.0E-41	3R:1798209-1798945
Gpa-cn31	Os-E	FBpp0078304	4.6E-13	3R:1802188-1802355
Gpgh-cn109	Os-E	FBpp0078304	5.1E-39	3R:1802125-1802480
Gth-cn65	Pherokine 3	FBpp0072311	8.2E-37	2R:20858474-20858779
Gthfl-5d10.q1k	OBP83g	FBpp0078266	2.1E-18	3R:1937471-1937602
Gpghfl-8a06.p1k	CG30354	FBpp0087782	1.4E-40	2R: 4564888-4565097
Gpghfl-8d10.p1k	OBP83g	FBpp0078266	3.4E-65	3R:1937474-1937917
Gpghfl-5g01.p1k	No Hits			
Gthfl-8b10.p1k	No Hits			
Structural				
Gpgh-cn110	Cuticle protein	FBpp0079123	8.2E-12	2L: 7740959-7741057
Gpgh-cn113	CG31904	FBpp0079125	2.5E-05	2L: 7734853-7734906
Gpghfl-9d01.p1k	Histone H3.3A	FBpp0078650	3.7E-82	2L:5055259-5055727
Gpghfl-5g03.p1k	Cuticular protein	FBpp0077993	0.62	3L:21288980-21289039
Gpghfl-5e08.p1k	chickadee	FBpp0078864	1.2E-87	2L:5973601-5979522
Gpghfl-5g11.p1k	Tom 34	FBpp0081284	2.5E-87	3R:3861765-3862187
Gth-cn52	CG32113	FBpp0271898	4.5	3L:12752356-12752406
Gth-cn53	CG7214	FBpp0079122	4.1E-12	2L:7744217-7744309
Gpa-cn1	No Hits			
Gpafl-8f02.p1k	CG7224	FBpp0099502	0.0069	2L: 7999170-7999202
Transport				
Gpafl-1h05.p1k	CG8026	FBpp0087691	0.71	2R:5076267-5076332
Gpafl-7d12.q1k	CG11739	FBpp0078553	1.9E-11	3R:205465-205654

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Salivary Gland				
Gpgh-cn108	CG13043	FBpp0075192	0.012	3L:16298643-16298687
Gpghfl-5c07.p1k	Ribosomal protein	FBpp0078416	2.5E-89	3R:1291433-1291979
Gpghfl-2f12.p1k	CG9717	FBpp0085041	1.4E-06	3R:26412320-26412385
Gth-cn74	CG13044	FBpp0075225	0.11	3L:16295906-16295956
Gth-cn81	CG12929	FBpp0111973	1.3E-33	2R:5473796-5473957
Gthfl-1h08.q1k	RNA polymerase I subunit	FBpp0077714	2.0E-05	2L:407574-407651
Gthfl-1e06.p1k	technical knockout	FBpp0070443	3.7E-76	X:2336416-2336742
Gpa-cn29	CG18557	FBpp0077373	9.8	2L:2844799-2844819
Gpafl-7f02.q1k	Cuticular protein 65Az	FBpp0076704	8.6	3L:6154157-6154177
Gpafl-9b12.q1k	CG15311	FBpp0071371	7.6	X:10082268-10082291
Gpafl-5b06.p1k	stranded at second	FBpp0081089	2.5	3R:3005897-3005932
Gpgh-cn37	Ribosomal protein L27A	FBpp0077142	1.5	2L:4457414-4457458
Gpghfl-2f06.p1k	No Hits			
Gpghfl-2h09.q1k	adenosine 2	FBpp0078851	3.7	2L:6044056-6044094
Gpghfl-5h09.p1k	CG33174	FBpp0073656	3.7	X:13702517-13702540
Gth-cn2	Ribosomal protein L27A	FBpp0077142	8.2E-78	2L:4457384-4457964
Gth-cn11	Neuropeptide F-like Receptor	FBpp0074640	0.56	3L:20068662-20068778
Gthfl-1e09.q1k	CG15390	FBpp0077444	0.64	2L:2375406-2375441
Gthfl-8a02.p1k	CG5674	FBpp0080575	4	2L:17972088-17972108
Gthfl-5e06.p1k	mitochondrial Cyt. c oxidase	FBpp0100177	6.9E-13	dmel_mitochondrion_genome: 3638-3706
Gthfl-5g02.q1k	Thiolase	FBpp0072135	9	2R:19754402-19754452

Appendix VII cont.

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Gpghfl-5e02.p1k	CG18327	FBpp0086651	0.0031	2R:10055529-10055648
Gth-cn75	Calcium ATPase	FBpp0072125	3.5E-141	2R:19816838-19817618
Gthfl-4a12.p1k	ATP7	FBpp0271765	2.3E-07	X:11756985-11757122
Gthfl-4b10.q1k	CG9582	FBpp0079379	5.8	2L:9025241-9025270
Gthfl-1h10.p1k	CG14778	FBpp0070244	3.2	X:1364274-1364300
Gpa-cn130	Ankyrin	FBpp0088239	6.9	4:139375-139474
Gpghfl-1a02.q1k	No Hits			
Signal transduction				
Gpgh-cn84	Dopamine receptor	FBpp0288722	0.14	3R:10003804-10006023
Gpgh-cn112	Arrestin 2	FBpp0076326	3.5E-28	3L:8640809-8641310
Gth-cn59	Arrestin 2	FBpp0076326	3.5E-22	3L:8640809-8641268
Gthfl-7b01.p1k	CG32683	FBpp0071397	8.8E-22	X:10315511-10321501
Gthfl-18b01.p1k	CG10625	FBpp0271739	9.4	3L:5533568-5533801
Gpghfl-5d02.p1k	CG13040	FBpp0075189	0.029	3L:16318549-16318868
Gpa-cn138	CG17834	FBpp0079298	1.5	2L:8536069-8536228
Gpafl-7e07.q1k	Ptx1	FBpp0271894	4.8	3R:26750265-26754895
Gpafl-7f12.p1k	No Hits			
Nucleic Acid Metabolism				
Gpgh-cn1	Ribosomal protein L15	FBpp0112467	4.1E-124	3LHet:2398667-2398930
Gpgh-cn89	Phosphoribosylamidotransferase	FBpp0081261	2.5E-108	3R:3735796-3736107
Gpgh-cn118	CG10077	FBpp0076649	9.8	3L:6947540-6947581
Gpghfl-8g10.p1k	string of pearls	FBpp0079500	6.0E-159	2L:9896474-9897028
Gpghfl-8h01.p1k	Ribosomal protein L23	FBpp0071808	2.1E-23	2R:18743058-18743162
Gpghfl-8g08.p1k	Ribosomal protein L38	FBpp0110412	1.3E-49	2R:403615-403824
Gpghfl-8h07.p1k	Ribosomal protein L5	FBpp0110420	1.1E-61	2L:22428526-22428855

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Metabolism				
Gpa-cn132	CG18869	FBpp0073060	7.2	3L:4060537-4060572
Gpafl-5e09.p1k	p47	FBpp0088069	2.2E-47	2R:3352651-3353002
Gpafl-8c02.p1k	CDP diglyceride synthetase	FBpp0076411	3.0E-18	3L:8122553-8122633
Gpgh-cn89	Phosphoribosylamidotransferase	FBpp0081261	2.5E-108	3R:3735796-3736107
Gpgh-cn118	CG10077	FBpp0076649	9.8	3L:6947540-6947581
Gpghfl-8d07.p1k	CG3226	FBpp0070953	7.7E-40	X:6547313-6547495
Gpghfl-8f09.p1k	wings up A	FBpp0074298	1.8E-39	X:18000765-18003514
Gpghfl-8f08.p1k	Ank2	FBpp0111609	8.2	3L:7703123-7703218
Gpghfl-5d04.p1k	cyclophilin-33	FBpp0086101	0.74	2R:13432932-13433061
Gpghfl-5e08.p1k	chickadee	FBpp0078864	1.2E-87	2L:5973601-5979522
Gpghfl-5e02.p1k	CG18327	FBpp0086651	0.0031	2R:10055529-10055648
Gpghfl-9c05.p1k	short gastrulation	FBpp0073879	1.2E-62	X:15500008-15500256
Gpghfl-9f01.p1k	Dynamin associated protein 160	FBpp0081031	8.7	2L:21140789-21140842
Gpghfl-9f10.p1k	Autophagy-specific gene 6	FBpp0083975	2.5E-109	3R:19875410-19875694
Gpghfl-9a01.p1k	Brother	FBpp0072643	5.2E-95	3L:1647959-1648426
Gpghfl-5f10.q1k	CG31075	FBpp0084452	1	3R:22811511-22811549
Gpghfl-10h01.p1k	CG31075	FBpp0084452	2.7E-96	3R:22811454-22811885
Gth-cn3	CG12027	FBpp0073228	5.4E-16	3L:5011883-5012020
Gth-cn9	Dynamin associated protein 160	FBpp0081031	6.4	2L:21140789-21140842
Gth-cn55	lethal (1) G0230	FBpp0071373	1.5E-78	X:10107326-10107774
Gth-cn68	ATP synthase-gamma chain	FBpp0084907	5.0E-36	3R:25573582-25573749
Gth-cn75	Calcium ATPase at 60A	FBpp0072125	3.5E-141	2R:19816838-19817618
Gth-cn78	CG13551	FBpp0071943	1.4E-09	2R:19265671-19270185
Gth-cn83	CG1746	FBpp0085135	8.5E-68	3R:27043502-27044712

Appendix VII cont.

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Gpghfl-9a01.p1k	Brother	FBpp0072643	5.2E-95	3L:1647959-1648426
Gpghfl-5d04.p1k	cyclophilin-33	FBpp0086101	0.74	2R:13432932-13433061
Gpghfl-9d01.p1k	Histone H3.3A	FBpp0078650	3.7E-82	2L:5055259-5055727
Gth-cn56	Ribosomal protein S3	FBpp0083802	9.3E-128	3R:19185044-19185535
Gthfl-8b01.p1k	Brother	FBpp0072643	3.0E-38	3L:1647959-1648177
Gthfl-5g08.q1k	lilliputian	FBpp0077318	3.7	2L:2950300-2950335
Gthfl-5h01.q1k	Ribosomal protein L23	FBpp0071808	7.8E-26	2R:18743049-18743162
Gthfl-5h07.q1k	Ribosomal protein L5	FBpp0110423	2.2E-35	2L:22427727-22428399
Gthfl-6f01.p1k	Dead box protein 80	FBpp0112438	0.22	3LHet:2423709-2423774
Gthfl-6a01.q1k	Brother	FBpp0072643	3.5E-22	3L:1648259-1648408
Gthfl-4c07.q1k	Ribosomal protein	FBpp0078416	9.0E-34	3R:1291582-1291704
Gthfl-16e08.p1k	TfIIA-S-2	FBpp0070177	1.9E-14	X:905246-905392
Gthfl-5g10.q1k	string of pearls	FBpp0079500	9.3E-33	2L:9896330-9896626
Gthfl-1h08.q1k	RNA polymerase I	FBpp0077714	2.5E-18	2L:407385-407705
Gpa-cn95	CG7702	FBpp0083080	2.4	3R:14498679-14498708
Gpghfl-8a07.p1k	Ribosomal protein L36	FBpp0070150	1.9E-81	X:522498-522891
Gpghfl-5c07.q1k	CG31814	FBpp0080157	3.1	2L:13678375-13678407
Gpghfl-7d07.p1k	CG32529	FBpp0074570	9.1	X:19796431-19796448
Gthfl-8c02.q1k	CG14642	FBpp0088629	0.94	3R:68094-68123
Gthfl-5a07.q1k	Ribosomal protein L36	FBpp0070150	9.6E-72	X:522498-522873
Gpafl-2c08.p1k	No Hits			
Gpgh-cn49	No Hits			
Gpgh-cn53	No Hits			
Gth-cn8	No Hits			
Gthfl-8b08.p1k	No Hits			

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Gthfl-2d09.p1k	CG3590	FBpp0082816	2.6E-83	3R:12846069-12846308
Gthfl-8b01.p1k	Brother	FBpp0072643	3.0E-38	3L:1647959-1648177
Gthfl-6f01.p1k	Dead box protein 80	FBpp0112438	0.22	3LHet:2423709-2423774
Gthfl-4f10.p1k	CG31075	FBpp0084452	1.7	3R:22811511-22811549
Metabolism				
Gthfl-4h12.p1k	CG2887	FBpp0071393	2.5E-16	X:10371602-10371763
Gthfl-4c07.p1k	upheld	FBpp0100015	2.5E-13	X:13489076-13489254
Gthfl-6c09.q1k	ATP synthase-2C subunit b	FBpp0076134	9.4E-25	3L:9723956-9724066
Gthfl-6a01.q1k	Brother	FBpp0072643	3.5E-22	3L:1648259-1648408
Gthfl-7g04.p1k	Bruce	FBpp0081717	0.34	3R:6139271-6139357
Gthfl-14h01.p1k	Muscle-specific protein 300	FBpp0091017	8.3E-18	2L:5204509-5204868
Gpafl-4g07.p1k	CG10317	FBpp0082769	1.2	3R:12195606-12195738
Gpafl-6a02.q1k	CG12360	FBpp0082186	7.9	3R:8853086-8853115
Gpafl-6b05.p1k	No Hits			
Gpafl-9e11.q1k	CG3884	FBpp0086913	8	2R:8836858-8836878
Gpafl-7a12.p1k	CG7304	FBpp0075329	5	3L:15673312-15673341
Gpafl-7d01.p1k	shuttle craft	FBpp0080266	7.1	2L:15114145-15114180
Gpafl-7e04.p1k	No Hits			
Gpgh-cn3	CG42319	FBpp0289014	4.9	2R:9151311-9151343
Gpgh-cn20	No Hits			
Gpgh-fl-1b02.q1k	CG3546	FBpp0070647	4.9	X:4453460-4453486
Gpghfl-2d11.p1k	No Hits			
Gpghfl-8e06.p1k	Ferritin 1 heavy chain homologue	FBpp0084995	0.00014	3R:26211709-26211771
Gthfl-4e09.p1k	p47	FBpp0088069	1.2	2R:3352651-3352683
Gthfl-5g03.q1k	CG40113	FBpp0112518	0.54	2RHet:2239984-2240007
Gthfl-4c04.q1k	mitochondrial Cyt. c oxidase	FBpp0100176	1.3E-11	dmel_mitochondrion_genome:2908-3009

Appendix VII cont.

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Energy Metabolism				
Gpafl-6d02.p1k	CG10396	FBpp0085425	7.0E-19	2R:1241966-1242157
Gpgh-cn15	CG3731	FBpp0082459	4.0E-20	3R:10717814-10718014
Gpghfl-8b10.p1k	CG9306	FBpp0080051	6.9E-79	2L:13370374-13370517
Gpghfl-8c10.p1k	CG3446	FBpp0070885	2.5E-78	X:6243768-6244230
Gpghfl-8e08.p1k	CG2014	FBpp0084806	4.3E-108	3R:25341696-25342028
Gpghfl-9c09.p1k	CG17300	FBpp0075626	1.1E-10	3L:13079667-13079936
Gpghfl-9e09.q1k	CG10320	FBpp0112014	1.1E-29	2R:17545870-17546120
Gpghfl-5e01.p1k	CG7211	FBpp0079104	2.0E-27	2L:7747647-7747754
Gpghfl-5e07.p1k	Cyt. c oxid subunit Va	FBpp0081947	1.8E-73	3R:7645521-7645877
Gpghfl-5h02.p1k	CG5548	FBpp0073777	4.1E-46	X:14958200-14958364
Gth-cn3	CG12027	FBpp0073228	5.4E-16	3L:5011883-5012020
Gth-cn57	CG30093	FBpp0086341	1.5E-22	2R:11831745-11831879
Gth-cn60	CG34172	FBpp0111278	0.04	2L:2192626-2192700
Gthfl-1a11.q1k	CG13397	FBpp0079316	1.7	2L:8410206-8410241
Gthfl-1a11.p1k	CG10396	FBpp0085425	2.2E-19	2R:1241966-1242157
Gthfl-1c07.q1k	CG2249	FBpp0087485	1.1E-33	2R:5937084-5937371
Gthfl-1c07.p1k	CG12907	FBpp0087440	8.2	2R:6241122-6241154
Gthfl-5e08.q1k	CG2014	FBpp0084806	6.1E-10	3R:25342074-25342142
Gthfl-4h06.q1k	CG18624	FBpp0071092	2.4E-27	X:7787181-7787330
Gpgh-cn88	CG7181	FBpp0078037	1.1E-38	3L:21532175-21532454
Gpafl-1a01.p1k	Mit. NADH-ubiq. oxidoreductase	FBpp0100183	3.4	dmel_mitochondrion_genome:8210-8254
Gpgh-cn2	Mit. NADH-ubiq. oxidoreductase	FBpp0100187	2.7E-44	dmel_mitochondrion_genome:11918-12085
Gpgh-cn4	mitochondrial Cyt. c oxidase subunit I	FBpp0100176	1.5E-107	dmel_mitochondrion_genome:2650-2859
Gpgh-cn5	mitochondrial Cyt c oxidase subunit II	FBpp0100177	3.5E-78	dmel_mitochondrion_genome:3545-3724
Gthfl-4g11.q1k	Translocase of outer membrane	FBpp0081284	6.3E-29	3R:3862014-3862214
Gthfl-1d08.p1k	transient receptor potential	FBpp0084879	1.3E-12	3R:25740321-25740398
Gpghfl-5h10.p1k	No Hits			
Gpghfl-9c03.p1k	No Hits			
Gpghfl-8a05.p1k	No Hits			
Gpghfl-9c06.p1k	No Hits			
Gthfl-8c01.p1k	No Hits			
Gthfl-4g08.q1k	No Hits			
Gthfl-6g09.q1k	No Hits			

Appendix VII cont.

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Gpgh-cn6	Mit. Cyt c oxidase nit III	FBpp0100180	2.2E-107	dmel_mitochondrion_genome:5108-5509
Gpgh-cn7	Mit. NADH-ubiq. oxidoreductase	FBpp0100182	3.3E-06	dmel_mitochondrion_genome:6508-6573
Gpgh-cn114	mitochondrial ATPase subunit 6	FBpp0100179	6.9E-73	dmel_mitochondrion_genome:4398-4568
Gpgh-cn115	mitochondrial Cyt c oxidase subunit III	FBpp0100180	0.69	dmel_mitochondrion_genome:5357-5383
Gpgh-cn116	mitochondrial Cyt c oxidase subunit III	FBpp0100180	1.0E-69	dmel_mitochondrion_genome:5087-5398
Gpghfl-3g03.q1k	CG15456	FBpp0077015	3.8	X:20297904-20297921
Gth-cn6	mitochondrial NADH-ubiquinone oxidoreductase chain 4	FBpp0100183	1.4E-45	dmel_mitochondrion_genome:8510-8677
Gth-cn19	mitochondrial Cytochrome c oxidase subunit III	FBpp0100180	2.5E-97	dmel_mitochondrion_genome:5108-5509
Gth-cn55	lethal (1) G0230	FBpp0071373	1.5E-78	X:10107326-10107774
Gth-cn58	mitochondrial NADH-ubiquinone oxidoreductase chain 1	FBpp0100187	1.3E-09	dmel_mitochondrion_genome:11780-11851
Gth-cn69	mitochondrial Cytochrome c oxidase subunit III	FBpp0100180	2.8E-121	dmel_mitochondrion_genome:5087-5425
Gth-cn70	mitochondrial Cytochrome c oxidase subunit II	FBpp0100177	5.7E-102	dmel_mitochondrion_genome:3320-3769
Gthfl-2b11.q1k	No Hits			
Gthfl-5b10.q1k	CG9306	FBpp0080051	1.8E-30	2L:13370374-13370517
Gthfl-7h05.p1k	mitochondrial NADH-ubiquinone oxidoreductase chain 4	FBpp0100183	8.6E-72	dmel_mitochondrion_genome:8771-8986
Gthfl-5h04.q1k	CG6723	FBpp0081881	1.5	3R:7242355-7242378
Gthfl-6b01.q1k	mitochondrial ATPase subunit 6	FBpp0100179	1.3E-55	dmel_mitochondrion_genome:4392-4598

Appendix VII cont.

Cluster	Best Ensembl Match *D. melanogaster* Gene	Protein ID	E-value	Chr. Location
Gthfl-4b01.q1k	mitochondrial ATPase subunit 6	FBpp0100179	9.4E-09	dmel_mitochondrion_genome:4632-4727
Gthfl-4b11.q1k	mitochondrial Cyt c oxidase subunit III	FBpp0100180	1.1E-05	dmel_mitochondrion_genome:5351-5398
Gthfl-4c01.p1k	Mit. NADH-ubiquinone oxidoreductase	FBpp0100182	6.0E-46	dmel_mitochondrion_genome:7003-7128
Gthfl-4h09.q1k	CG33174	FBpp0073656	9.9	X:13702517-13702540

Gpa-cnxxx - *Glossina pallidipes* consensus sequences; Gpaf-xxxx.q1k - *Glossina pallidipes* antennae singletons; Gpgh-cnxxx - *Glossina palpalis gambiensis* consensus sequences; Gpghfl-xxxx.q1k - *Glossina palpalis gambiensis* head singletons; Gth-cnxxx - *Glossina tachinoides* consensus sequences; Gthfl-xxxx.q1k - *Glossina tachinoides* head singletons

Appendix VIII - *Glossina pallidipes*, *Glossina palpalis gambiensis* and *Glossina tachinoides* clusters and their best matches to *Anopheles gambiae*, *Aedes aegypti* and *Culex quinguefasciatus* protein databases

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Odorant Binding												
Gpa-cn30	OBP17	AGAP003309-PA	2.0E-37	2R: 35,643,124-35,644,111	OBP	AAEL009449-PA	3.0E-36	1.397:1,059,826-1,064,022	OBP	CPIJ007617	2.0E-35	3.150:672930-673546
Gpgh-cn109	OBP17	AGAP003309-PA	1.0E-34	2R: 35,643,124-35,644,111	OBP	AAEL009449-PA	1.0E-34	1.397:1,059,826-1,064,022	Insect OBP/ PhBP	CPIJ007604	3.0E-34	3.150:170718 - 174721
Gth-cn65	PBP/A10/OS-D	AGAP008052-PA	5.0E-41	3R: 4,868,812-4,869,195	OS-D/PhBP	AAEL002022-PA	1.0E-41	1.47:2,418,811-2,419,194	CSP 1	CPIJ002618	8.0E-35	3.42:1,053,230-1,053,607
Gthfl-5d10.q1k	OBP9	AGAP000278-PA	0.002	X: 5,035,500-5,035,999	OBP 99c	AAEL005772-PA	8.0E-04	1.174:1,533,074-1,533,545	GOBP 99a	CPIJ017326	0.002	3.984:154,116-154,609
Gpghfl-8a06.p1k	Ubiquinol-cyt.c redu hinge	AGAP002245-PA	1.0E-20	2R: 18,083,401-18,083,756	Ubiquinol-cyt.c redu hinge	AAEL010801-PB	4.0E-23	1.509:141,153 - 141,587	Mitochondrial	CPIJ008356	1.0E-22	
Gpghfl-8d10.p1k	OBP9	AGAP000278-PA	2.0E-017	X: 5,035,500-5,035,999	OBP 99c	AAEL005772-PA	4.0E-17	1.174: 1,533,074-1,533,545	GOBP	CPIJ017326	6.0E-17	
Gpghfl-5g01.p1k	OBP	AGAP004263-PA	1.0E-08	2R: 53,468,012-53,469,656	OBP	AAEL011966-PA	2.0E-07	1.638:389,885-404,032	OBP	CPIJ015482	5.0E-09	3.639:116,905-125,388
Gthfl-8b10.p1k	OBP17	AGAP003309-PA	9.0E-09	2R: 35,643,124-35,644,111	OBP 56a	AAEL015499-PA	2.0E-08	1.2733:6,977-7,429	Insect OBP/ PhBP	CPIJ007604	1.0E-07	3.150:170,719-174,721
Structural												
Gpgh-cn110	Cuticular protein 15	AGAP008459-PA	2.0E-07	3R: 10,897,827-10,898,198	Cuticle protein	AAEL002181-PA	7.0E-07	1.51:1,124,480-1,124,944	cuticle protein	CPIJ003482	2.0E-06	3.48:207,283-207,796
Gpgh-cn113	Alanine-rich region profile	AGAP008454-PA	3.0E-05	3R: 10,888,884-10,889,372	Alanine-rich region	AAEL017402-PA	1.0E-04	1.51:1,134,955-1,135,681	cuticle protein	CPIJ003453	9.0E-06	3.48:38,673-39,135
Gpghfl-9d01.p1k	Histone H3	AGAP001813-PA	4.0E-70	2R: 10,860,617-10,861,148	histone H3.3	AAEL006158-PA	4.0E-70	1.192:1,123,692-1,124,168	histone H3.3 type 2	CPIJ017187	5.0E-70	3.876:94,629-95,108
Gpghfl-5g03.p1k	cuticular protein 75	AGAP009871-PA	5.0E-37	3R: 44,607,075-44,608,893	pupal cuticle protein	AAEL003259-PA	2.0E-38	1.82:349,372-357,541	pupal cuticle protein	CPIJ009331	1.0E-35	3.242:229,831-231,878
Gpghfl-5e08.p1k	Profilin/allergen	AGAP009861-PA	1.0E-53	3R: 44,535,146-44,555,606	profilin	AAEL013353-PD	1.0E-53	1.827:268,874-304,152	profilin	CPIJ001546	8.0E-54	3.17:367,863-385,968
Gpghfl-5g11.p1k	Tetratricopeptide TPR-1	AGAP003727-PA	6.0E-41	2R: 42,626,025-42,628,951	Armadillo	AAEL009168-PA	2.0E-38	1.375:491,598-493,244	translocase of outer membrane	CPIJ001395	1.0E-37	3.20:866,214-869,072
Gth-cn52	Cuticular protein 16	AGAP008460-PA	3.0E-06	3R: 10,900,320-10,900,829	Alanine-rich region profile	AAEL017402-PA	7.0E-07	1.51:1,134,955-1,135,681	cuticle protein	CPIJ003474	1.0E-06	3.48:188,147-188,597
Gth-cn53	Cuticular protein 1	AGAP008444-PA	1.0E-05	3R: 10,866,569-10,867,012	cuticle protein	AAEL002181-PA	2.0E-05	1.51:1,124,480-1,124,944	cuticle protein	CPIJ003473	7.0E-05	3.48:186,440-186,946

Appendix VIII cont.

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Gpa-cn1	Prion	AGAP008439-PA	0.005	3R: 10,805,155-10,807,128	Proline-rich region	AAEL005017-PA	0.021	1.139:780,568-782,310	Proline-rich region	CPIJ000931	0.036	3.9:1,953,966-1,959,660
Gpafl-8f02.p1k	No Hits				No Hits				No Hits			
Transport												
Gpafl-1h05.p1k	ADP,ATP carrier protein 2	AGAP002358-PC	1.0E-54	2R:20,598,090-20,599,178	ADP,ATP carrier protein	AAEL004855-PA	2.0E-46	1.132:1,668,937-1,682,495	ADP,ATP carrier	CPIJ005941	3.0E-54	3.104:143,246-144,296
Gpafl-7d12.q1k	Tricarboxylate/iron carrier	AGAP004519-PA	6.0E-05	2R: 57,327,659-57,329,455	No Hits				sideroflexin	CPIJ008314	7.0E-05	3.167:673,160-677,888
Gpgh-cn101	Longin	AGAP012717-PA	1.0E-101	UNKN: 24,915,268-24,916,389	snare protein sec22	AAEL012875-PA	1.0E-105	1.753:388,957-389,921	transport protein SEC22	CPIJ017814	1.0E-105	3.1019:93,459-94,353
Gpghfl-5b10.p1k	Adenine nucleotide translocator 1	AGAP007653-PA	1.0E-11	2L: 48,762,165-48,765,808	tricarboxylate transport protein	AAEL005991-PA	1.0E-63	1.184:106,061-115,898	tricarboxylate transport protein	CPIJ013697	2.0E-64	3.492:260,522-261,566
Gpghfl-5c11.p1k	ADP,ATP carrier protein 2	AGAP002358-PC	1.0E-62	2R:20,598,090-20,599,178	ADP,ATP carrier protein	AAEL004855-PA	1.0E-54	1.132:1,668,937-1,682,495	ADP,ATP carrier protein	CPIJ005941	4.0E-63	3.104:143,246-144,296
Gpghfl-5e02.p1k	Mit. substrate/solute carrier	AGAP002704-PA	2.0E-26	2R: 25,826,373-25,828,086	Mitochondrial substrate carrier	AAEL008718-PA	1.0E-27	1.344:870,632-889,122	Mitochondrial substrate carrier	CPIJ002886	9.0E-27	3.37:1,054,405-1,056,054
Gth-cn75	Calcium-transporting ATPase	AGAP006186-PB	7.0E-97	2L: 27,903,046-27,928,073	calcium-transporting ATPase	AAEL006582-PA	3.0E-98	1.211:1,168,009-1,213,899	calcium-transporting atpase sarcoplasmic	CPIJ018021	5.0E-96	3.1063:57,960-81,174
Gthfl-4a12.p1k	ATPase, P-type cation exchange	AGAP002858-PE	4.0E-99	2R:28,371,857-28,387,114	Na+/K+ ATPase alpha subunit	AAEL012062-PC	2.0E-98	1.650:512,971-547,252	sodium/potassium-transporting ATPase alpha chain	CPIJ005966	9.0E-57	3.104:497,581-503,775
Gthfl-4b10.q1k	Mitochondrial carrier protein	AGAP007653-PA	3.0E-09	2L: 48,762,165-48,765,808	tricarboxylate transport protein	AAEL005991-PA	1.0E-54	1.184:106,061-115,898	tricarboxylate transport protein	CPIJ013697	4.0E-55	3.492:260,522-261,566
Gthfl-1h10.p1k	No Hits				No Hits				No Hits			
Gpa-cn130	No Hits				No Hits				No Hits			
Gpghfl-1a02.q1k	Acyl carrier protein (ACP)	AGAP010464-PA	8.0E-14	3L: 3,944,950-3,945,956	acyl carrier protein precursor	AAEL011689-PA	1.0E-11	1.603:180,673-194,411	Acyl carrier protein (ACP)	CPIJ006168	7.0E-08	3.109:639,109-645,442
Signal transduction												
Gpgh-cn84	putative rhodopsin receptor 3	AGAP012985-PA	1.0E-90	2R:671,584-672,696	GPCR (Rhod)opsin Family Source	AAEL006498-PA	7.0E-88	1.208:1,715,269-1,716,390	opsin-1	CPIJ012052	1.0E-89	3.358:153,562-154,668
Gpgh-cn112	Arrestin	AGAP010134-PA	2.0E-51	3R: 49,303,097-49,304,807	phosrestin ii (arrestin a) (arrestin 1)	AAEL013535-PA	2.0E-52	1.864:398,197-406,521	phosrestin ii	CPIJ003101	3.0E-50	3.39:314,782-318,945
Gth-cn59	Arrestin	AGAP010134-PA	1.0E-42	3R: 49,303,097-49,304,807	phosrestin ii (arrestin a)	AAEL013535-PA	1.0E-43	1.864:398,197-406,521	phosrestin ii	CPIJ003101	2.0E-42	3.39:314,782-318,945

Appendix VIII cont.

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Gthfl-7b01.p1k	Arrestin	AGAP010134-PA	4.0E-57	3R: 49,303,097-49,304,807	phosrestin ii (arrestin a)	AAEL013535-PA	4.0E-55	1.864:398,197-406,521	phosrestin ii	CPIJ003101	2.0E-56	3.39:314,782-318,945
Gthfl-18b01.p1k	arrestin, Arr2-like	AGAP006263-PA	1.0E-44	2L: 28,771,909-28,773,907	phosrestin i (arrestin b) (arrestin 2)	AAEL003116-PA	5.0E-45	1.78:2,394,436-2,395,931	phosrestin-1	CPIJ014777	4.0E-45	3.588:39,146-40,948
Gpghfl-5d02.p1k	No Hits				No Hits				No Hits			
Gpa-cn138	No Hits				No Hits				No Hits			
Gpafl-7f12.p1k	No Hits				No Hits				ucrose transport protein	CPIJ012070	0.055	3.377:161,231-163,559
Nucleic Acid Metabolism												
Gpgh-cn1	60S ribosomal protein L15	AGAP004919-PA	5.0E-73	2L: 6,015,542-6,016,396	ribosomal protein L15	AAEL001488-PA	4.0E-51	1.34:3,448,327-3,448,923	60S ribosomal protein L15	CPIJ008103	4.0E-73	3.184:249,409-250,210
Gpgh-cn89	Amidophosphoribosyl transferase	AGAP000179-PA	1.0E-76	X: 2,993,095-2,995,080	amidophosphoribosyltransferase	AAEL003581-PB	2.0E-80	1.92:1,708,067-1,709,917	amidophosphoribosyltransferase	CPIJ009001	4.0E-78	3.210:183,267-185,112
Gpgh-cn118	RNA helicase, DEAD-box type, Q motif	AGAP010656-PA	2.0E-27	3L: 8,078,239-8,100,371	DEAD box ATP-dependent RNA helicase	AAEL013985-PA	7.0E-27	1.976:331,090-339,354	ATP-dependent RNA helicase DDX54	CPIJ008599	4.0E-23	3.202:434,116-436,557
Gpghfl-8g10.p1k	Ribosomal protein S5,	AGAP003768-PA	1.0E-106	2R: 43,124,620-43,125,724	40S ribosomal protein S2	AAEL010168-PB	1.0E-106	1.458:871,944-872,978	40S ribosomal protein S2	CPIJ010326	1.0E-103	3.295:280,049-282,560
Gpghfl-8h01.p1k	Ribosomal protein L14b	AGAP010252-PA	3.0E-14	3R: 51,662,651-51,663,450	60S ribosomal protein L23	AAEL013097-PA	3.0E-14	1.789:366,699-367,486	RL23	CPIJ011325	4.0E-14	3.310:461,499-462,272
Gpghfl-8g08.p1k	60S ribosomal protein L38	AGAP010163-PA	5.0E-29	3R: 49,777,851-49,778,063	Ribosomal protein L38e	AAEL005451-PA	1.0E-30	1.157:1,727,679-1,727,891	60S ribosomal protein L38	CPIJ005613	3.0E-30	3.95:734,463-734,675
Gpghfl-8h07.p1k	60S ribosomal protein L5	AGAP009031-PA	4.0E-75	3R: 24,355,978-24,357,240	ribosomal protein L5	AAEL004325-PA	2.0E-90	1.114:1,986,958-1,990,971	60S ribosomal protein L5	CPIJ010112	4.0E-89	3.276:498,078-499,646
Gpghfl-9a01.p1k	Core binding factor, beta	AGAP000317-PA	1.0E-71	X: 5,610,066-5,616,769	Core binding factor, beta	AAEL005829-PA	1.0E-70	1.176:643,909-677,173	Core binding factor, beta subunit	CPIJ019164	7.0E-70	3.1457:33,910-44,867
Gpghfl-5d04.p1k	Eukaryotic TIF-3 subunit G	AGAP007668-PA	9.0E-65	2L: 48,887,512-48,888,419	eukaryotic TIF-3	AAEL012661-PA	6.0E-69	1.722:40,930-42,245	eukaryotic TIF-3	CPIJ000190	2.0E-67	3.3:1,729,170-1,730,055
Gpghfl-9d01.p1k	Histone H3	AGAP001813-PA	4.0E-70	2R: 10,860,617-10,861,148	histone H3.3	AAEL006158-PA	4.0E-70	1.192:1,123,692-1,124,168	histone H3.3 type 2	CPIJ017187	5.0E-70	3.876:94,629-95,108
Gth-cn56	K Homology, prokaryotic type	AGAP001910-PA	3.0E-78	2R: 12,009,718-12,011,048	40S ribosomal protein S3	AAEL008192-PB	1.0E-79	1.305:1,304,669-1,321,925	40S ribosomal protein	CPIJ013941	4.0E-79	3.525:165,880-175,665
Gthfl-8b01.p1k	Core binding factor, beta subunit	AGAP000317-PA	2.0E-29	X: 5,610,066-5,616,769	Core binding factor, beta subunit	AAEL005829-PA	7.0E-29	1.176:643,909-677,173	Core binding factor, beta subunit	CPIJ019164	9.0E-29	3.1457:33,910-44,867
Gthfl-5g08.q1k	60S ribosomal protein L38	AGAP010163-PA	0.01	3R: 49,777,851-49,778,063	Ribosomal protein L38e	AAEL005451-PA	0.01	1.157:1,727,679-1,727,891	60S ribosomal protein	CPIJ005613	0.024	3.95:734,463-734,675

Appendix VIII cont.

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Gthfl-5h01.q1k	Ribosomal protein L14b/L23e	AGAP010252-PA	8.0E-16	3R: 51,662,651-51,663,450	60S ribosomal protein L23	AAEL013097-PA	8.0E-16	1.789:366,699-367,486	RL23	CPIJ011325	9.0E-16	3.310:461,499-462,272
Gthfl-5h07.q1k	Ribosomal protein L5	AGAP009031-PA	2.0E-06	3R: 24,355,978-24,357,240	ribosomal protein L5	AAEL004325-PA	1.0E-18	1.114:1,986,958-1,990,971	60S ribosomal protein	CPIJ010112	2.0E-18	3.276:498,078-499,646
Gthfl-6f01.p1k	RNA helicase, DEAD-box type, Q motif	AGAP008578-PA	1.0E-16	3R: 13,262,401-13,264,364	DEAD box ATP-dependent RNA helicase	AAEL009285-PA	3.0E-15	1.387:804,914-823,594	ATP-dependent RNA helicase vasa	CPIJ009286	3.0E-15	3.218:487,325-489,515
Gthfl-6a01.q1k	Core binding factor, beta	AGAP000317-PA	1.0E-27	X: 5,610,066-5,616,769	Core binding factor, beta subunit	AAEL005829-PA	5.0E-28	1.176:643,909-677,173	Core binding factor, beta subunit	CPIJ019164	2.0E-27	3.1457:33,910-44,867
Gthfl-4c07.q1k	Histone H5	AGAP002754-PA	2.0E-12	2R: 26,675,074-26,675,740	ribosomal protein L35A	AAEL000823-PB	2.0E-12	1.17:998,091-998,675	60S ribosomal protein	CPIJ008184	5.0E-14	3.183:728,929-729,634
Gthfl-16e08.p1k	TIF factor IIA	AGAP004370-PA	5.0E-46	2R: 55,365,431-55,365,931	(TFIIA)	AAEL002140-PA	5.0E-48	1.50:2,297,312-2,297,791	TIF IIA gamma chain	CPIJ007398	3.0E-47	3.147:699,798-700,264
Gthfl-5g10.q1k	Ribosomal protein S5	AGAP003768-PA	2.0E-29	2R: 43,124,620-43,125,724	40S ribosomal protein S2	AAEL010168-PB	5.0E-29	1.458:871,944-872,978	40S ribosomal protein	CPIJ010326	1.0E-26	3.295:280,049-282,560
Gthfl-1h08.q1k	RNA polymerase Rpb2,	AGAP008151-PA	3.0E-64	3R: 6,212,809-6,238,165	RNA polymerase Rpb2,	AAEL017156-PA	7.0E-64	1.84:2,256,737-2,268,271	DNA-directed RNA polymerase II	CPIJ007655	1.0E-33	3.159:169,512-173,328
Gpa-cn95	No Hits				No Hits				No Hits			
Gpghfl-8a07.p1k	Ribosomal protein L36e	AGAP002921-PB	9.0E-48	2R:29,615,444-29,616,107	ribosomal protein L36	AAEL000010-PB	1.0E-46	1.1:4,236,762-4,237,585	No Hits			
Gpghfl-5c07.q1k	No Hits				No Hits				No Hits			
Gpghfl-7d07.p1k	No Hits				No Hits				No Hits			
Gthfl-8c02.q1k	No Hits				No Hits				No Hits			
Gthfl-5a07.q1k	Ribosomal protein L36e	AGAP002921-PB	3.0E-41	2R:29,615,444-29,616,107	ribosomal protein L36	AAEL000010-PB	2.0E-40	1.1:4,236,762-4,237,585	No Hits			
Gpafl-2c08.p1k	No Hits				No Hits				hypothetical protein	CPIJ009839	0.031	3.253:560,664-561,263
Gpgh-cn49	Proline-rich region	AGAP005529-PB	0.001	2L: 16,654,663-16,656,903	Sec24B protein	AAEL004977-PA	0.047	1.137:1,147,263-1,174,647	Proline-rich region	CPIJ017787	0.008	3.1119:98-27,917
Gpgh-cn53	Chitin binding protein, peritrophin-A	AGAP001005-PA	0.02	X: 19,092,108-19,095,128	No Hits				No Hits			
Gth-cn8	Cysteine-rich region	AGAP013541-PA	0.01	X:20,029,435-20,030,978	No Hits				No Hits			
Gthfl-8b08.p1k	No Hits				No Hits				Zinc finger, C2H2-like	CPIJ006352	0.093	3.116:592,218-594,934

Appendix VIII cont.

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Energy Metabolism												
Gpafl-6d02.p1k	Cyt. c oxidase subunit IV	AGAP008727-PA	2.0E-55	3R: 16,158,434-16,159,082	cytochrome c oxidase subunit iv	AAEL013009-PA	5.0E-57	1.775:455,628-461,779	cytochrome c oxidase subunit IV	CPIJ004823	2.0E-54	3.76:758,796-762,083
Gpgh-cn15	Peptidase M16, zinc-binding site	AGAP000935-PA	9.0E-29	X: 17,669,389-17,671,168	mitochondrial processing peptidase	AAEL005435-PB	2.0E-28	1.156:1,726,989-1,729,306	mitochondrial processing peptidase beta subunit	CPIJ013398	1.0E-28	3.456:95,436-98,497
Gpghfl-8b10.p1k	Complex 1 LYR protein	AGAP009865-PA	8.0E-50	3R: 44,586,368-44,587,032	NADH:ubiquinone dehydrogenase	AAEL010230-PA	2.0E-50	1.464:19,653-31,462	NADH:ubiquinone dehydrogenase	CPIJ002649	3.0E-21	3.42:1,327,695-1,328,454
Gpghfl-8c10.p1k	GRIM-19	AGAP009652-PA	3.0E-53	3R: 37,625,238-37,625,878	mitochondrial NADH:ubiquinone oxidoreductase B16.6 subunit	AAEL005693-PB	6.0E-50	1.170:43,032-43,570	cell death-regulatory protein GRIM19	CPIJ000908	1.0E-52	3.9:1,540,609-1,541,061
Gpghfl-8e08.p1k	NADH_UbQ_OxRdtase-like_20kDa	AGAP000170-PA	2.0E-79	X: 2,905,106-2,905,729	NADH-plastoquinone oxidoreductase	AAEL008072-PA	3.0E-78	1.299:1,166,704-1,167,351	NADH dehydrogenase iron-sulfur protein 7	CPIJ018869	7.0E-78	3.1453:49,857-50,495
Gpghfl-9c09.p1k	ATPase, F0 complex, subunit B, mitochondrial	AGAP001138-PA	3.0E-76	2R: 569,086-570,174	mitochondrial ATP synthase b chain	AAEL005610-PA	2.0E-77	1.164:50,521-58,740	ATP synthase B chain	CPIJ006067	1.0E-77	3.106:527,151-528,237
Gpghfl-9e09.q1k	NADH_UbQ_OxRdtase_B12	AGAP012374-PA	3.0E-20	3L: 41,232,911-41,233,246	NADH dehydrogenase	AAEL007054-PA	3.0E-21	1.233:1,318,060-1,318,383	NADH-ubiquinone oxidoreductase B12 subunit	CPIJ014084	3.0E-19	3.496:66,701-67,045
Gpghfl-5e01.p1k	ATPase, F0 complex, subunit G, mitochondrial	AGAP009491-PA	2.0E-36	3R: 34,710,853-34,711,314	ATPase, F0 complex, subunit G	AAEL006509-PA	9.0E-33	1.209:719,920-720,340	hydrogen-transporting ATP synthase, G-subunit	CPIJ002431	1.0E-33	3.32:671,140-673,562
Gpghfl-5e07.p1k	Cytochrome c oxidase, subunit Va	AGAP011159-PA	2.0E-48	3L: 18,111,843-18,112,304	cytochrome c oxidase polypeptide	AAEL014944-PA	6.0E-47	1.1350:13,581-14,039	cytochrome c oxidase subunit 5A	CPIJ008200	3.0E-45	3.168:362,317-362,775
Gpghfl-5h02.p1k	NADH_UbQ_OxRdtase_B18_su	AGAP007574-PA	6.0E-25	2L: 47,791,258-47,792,012	NADH dehydrogenase, coupling factor	AAEL008490-PA	4.0E-24	1.328:823,062-832,150	Zinc finger, MYND-type	CPIJ000198	1.0E-24	3.3:1,833,672-1,841,944
Gth-cn3	ATPase, F0 complex, subunit F6, mitochondrial subgroup	AGAP004616-PA	1.0E-26	2R: 58,515,049-58,515,445		AAEL002813-PA	2.0E-26	1.68:1,541,837-1,563,004	mitochondrial ATP synthase coupling factor 6	CPIJ008665	8.0E-27	3.198:433,410-434,623
Gth-cn57	Cytochrome c oxidase, subunit VIa	AGAP000851-PA	1.0E-36	X: 15,724,507-15,725,070	cytochrome c oxidase	AAEL003234-PA	3.0E-33	1.82:1,176,603-1,177,001	cytochrome c oxidase, subunit VIA	CPIJ004219	4.0E-23	3.63:241,673-242,099
Gth-cn60	Cytochrome c oxidase, subunit VIIa	AGAP000109-PA	6.0E-17	X: 1,792,841-1,793,381	cytochrome c oxidase, subunit VIIA	AAEL007752-PA	1.0E-14	1.278:1,401,045-1,401,670	cytochrome c oxidase subunit VIIa	CPIJ015566	2.0E-15	3.680:230,253-230,725

Appendix VIII cont.

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Gthfl-1a11.q1k	Cytochrome c oxidase subunit IV	AGAP008727-PA	7.0E-14	3R: 16,158,434-16,159,082	cytochrome c oxidase subunit iv	AAEL005170-PA	1.0E-14	1.144:82,069-101,730	cytochrome c oxidase subunit IV	CPIJ004823	6.0E-14	3.76:758,796-762,083
Gthfl-1a11.p1k	Cytochrome c oxidase subunit IV	AGAP008727-PA	6.0E-55	3R: 16,158,434-16,159,082	cytochrome c oxidase subunit iv	AAEL013009-PA	2.0E-56	1.775:455,628-461,779	cytochrome c oxidase subunit IV	CPIJ004823	3.0E-54	3.76:758,796-762,083
Gthfl-1c07.q1k	Cytochrome c oxidase subunit VIIc	AGAP007621-PB	1.0E-12	2L: 48,354,815-48,355,245	cytochrome c oxidase, subunit VIIC	AAEL011514-PA	3.0E-12	1.586:394,316-394,624	zinc finger protein	CPIJ008759	1.0E-11	3.205:97,140-100,516
Gthfl-1c07.p1k	Cytochrome c oxidase subunit VIIc	AGAP007621-PB	5.0E-15	2L: 48,354,815-48,355,245	cytochrome c oxidase, subunit VIIC	AAEL011514-PA	1.0E-14	1.586:394,316-394,624	zinc finger protein	CPIJ008759	4.0E-14	3.205:97,140-100,516
Gthfl-5e08.q1k	NADH_UbQ_OxRdtase-20kDa	AGAP000170-PA	2.0E-10	X: 2,905,106-2,905,729	NADH-plastoquinone oxidoreductase	AAEL008072-PA	2.0E-10	1.299:1,166,704-1,167,351	NADH dehydrogenase iron-sulfur protein 7	CPIJ018869	2.0E-10	3.1453:49,857-50,495
Gthfl-4h06.q1k	NADH_UbQ_OxRdtase_MNLL	AGAP000849-PA	4.0E-16	X: 15,717,168-15,717,335	NADH dehydrogenase	AAEL003423-PB	6.0E-13	1.86:474,964-475,131	NADH dehydrogenase	CPIJ015857	2.0E-13	3.679:40,646-40,816
Gpgh-cn88	No Domain	AGAP004400-PA	6.0E-05	2R: 55,672,938-55,673,356	conserved hypothetical protein	AAEL007490-PB	0.018	1.258:944,532-944,911	No Hits			
Gpafl-1a01.p1k	No Hits				No Hits				No Hits			
Gpghfl-3g03.q1k	No Hits				No Hits				No Hits			
Gth-cn6	No Hits				NADH/Ubiquinone/ plastoquinone	AAEL009076-PA	5.0E-18	1.368:182,965-183,204	No Hits			
Gth-cn23	No Hits				Ubiquinol-cytochrome C reductase	AAEL000182-PA	7.0E-12	1.3:2,042,509-2,042,761	Ubiquinol-cytochrome C reductase	CPIJ000728	9.0E-12	3.6:1,764,455-1,764,704
Gth-cn55	No Hits				ATP synthase delta chain	AAEL002504-PB	7.0E-40	1.59:964,612-972,428	ATP synthase delta chain	CPIJ012335	6.0E-40	3.375:305,600-306,213
Gthfl-2b11.q1k	No Hits				NADH/Ubiquinone/ plastoquinone (complex I)	AAEL009076-PA	2.0E-15	1.368:182,965-183,204	No Hits			
Gthfl-5b10.q1k	Complex 1 LYR protein	AGAP009865-PA	4.0E-23	3R: 44,586,368-44,587,032	NADH:ubiquinone dehydrogenase	AAEL010230-PA	2.0E-24	1.464:19,653-31,462	No Hits			
Gthfl-7h05.p1k	NADH-quinone oxidoreductase	AGAP012727-PA	4.0E-14	UNKN: 25,591,108-25,592,481	No Hits				No Hits			

Appendix VIII cont.

Cluster	*An. gambiae* Description	Protein ID	E-value	Chr. Location	*Ae. aegypti* Description	Protein ID	E-value	Genomic Location Supercont	*Culex quinquefasciatus* Description	Protein ID	E-value	Genomic Location Supercont
Salivary Gland												
Gpgh-cn108	Retinin-like protein	AGAP006148-PA	2.0E-06	2L: 27,158,621-27,159,093	Retinin-like protein	AAEL001704-PA	1.0E-04	1.39:666,749-667,183	Retinin-like protein	CPIJ009100	3.0E-06	3.241:27,182-27,621
Gpghfl-5c07.p1k	Histone H5	AGAP002754-PA	7.0E-44	2R: 26,675,074-26,675,740	ribosomal protein L35A	AAEL000823-PB	2.0E-44	1.17:998,091-998,675	60S ribosomal protein L35a	CPIJ008184	2.0E-43	3.183:728,929-729,634
Gpghfl-2f12.p1k	Sulphate transporter	AGAP002331-PA	0.01	2R: 19,954,166-19,956,783	sulphate transporter	AAEL006372-PA	0.002	1.202:885,946-968,456	sulfate transporter	CPIJ012147	6.0E-04	3.394:339,528-347,182
Gth-cn74	Retinin-like protein	AGAP006148-PA	2.0E-06	2L: 27,158,621-27,159,093	Retinin-like protein	AAEL001704-PA	2.0E-04	1.39:666,749-667,183	Retinin-like protein	CPIJ009100	3.0E-06	3.241:27,182-27,621
Gth-cn81	UPF0546 membrane protein	AGAP012180-PA	7.0E-20	3L: 38,406,794-38,407,382	conserved hypothetical protein	AAEL005466-PA	9.0E-17	1.158:1,966,829-1,967,281	conserved hypothetical protein	CPIJ017962	1.0E-15	3.1053:54,127-54,615
Gthfl-1h08.q1k	RNA polymerase Rpb2 domain 7	AGAP008151-PA	3.0E-64	3R: 6,212,809-6,238,165	RNA polymerase Rpb2, domain 3	AAEL017156-PA	7.0E-64	1.84:2,256,737-2,268,271	DNA-directed RNA polymerase II 140 kDa polypeptide	CPIJ007655	1.0E-33	3.159:169,512-173,328
Gthfl-1e06.p1k	Ribosomal protein S12/S23	AGAP013102-PA	8.0E-53	2R:11,822,409-11,823,412	30S ribosomal protein S12	AAEL008169-PA	2.0E-52	1.304:67,556-67,888	30S ribosomal protein S12	CPIJ008389	2.0E-51	3.185:297,859-298,677
Gpghfl-2f06.p1k	Fibrinogen, alpha/beta chain	AGAP002005-PA	0.082	2R: 13,533,900-13,535,451	No Hits				No Hits			
Gpghfl-2h09.q1k	No Hits				No Hits				No Hits			
Gpghfl-5h09.p1k	No Hits				salivary gland growth factor	AAEL003214-PA	0.039	1.82:1,757,685-1,769,965	No Hits			
Gth-cn2	Ribosomal protein L15	AGAP011706-PA	3.0E-45	3L: 31,717,517-31,718,791	No Hits				60S ribosomal protein L15	CPIJ015042	3.0E-45	3.626:186,334-187,743
Gth-cn11	putative rhodopsin receptor 1	AGAP013149-PA	1.0E-126	2R:736,038-737,150	No Hits				opsin-1	CPIJ011571	1.0E-122	3.361:253,314-254,429
Gthfl-1e09.q1k	No Hits				No Hits				No Hits			
Gthfl-8a02.p1k	No Hits				No Hits				No Hits			
Gthfl-5e06.p1k	No Hits				No Hits				No Hits			
Gthfl-5g02.q1k	No Hits				No Hits				No Hits			

Gpa-cnxxx - *Glossina pallidipes* consensus sequences; Gpaf-xxxx.q1k - *Glossina pallidipes* antennae singletons; Gpgh-cnxxx - *Glossina palpalis gambiensis* consensus sequences; Gpghfl-xxxx.q1k - *Glossina palpalis gambiensis* head singletons; Gth-cnxxx - *Glossina tachinoides* consensus sequences; Gthfl-xxxx.q1k - *Glossina tachinoides* head singleton