Characterization of Olfactory Responsive Genes in Selected Tsetse Species: Annotation and Comparative Analyses of Chemosensory Proteins in the Genus *Glossina*

UNIVERSITY *of the*
WESTERN CAPE

Rosaline Wanjiru Macharia
3378728

Advisor: Prof. Alan Christoffels
Co-supervisors: Dr. Daniel Masiga
Dr. Paul Mireji

Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the South African National Bioinformatics Institute, Faculty of Natural Sciences, University of Western Cape

Abstract

Tsetse fly (Diptera: Glossinidae) is the sole cyclical vector of African trypanosomes, the causative agents of the neglected tropical disease, African trypanosomiasis. Approximately 70 million people and 50 million livestock in sub-Saharan Africa are at risk of contracting the disease. Tsetse spread the disease to their vertebrate hosts during blood meal feeding. A lack of efficacious vaccines/drugs against the disease has made vector (tsetse) control a feasible option; vector control is mainly achieved through application of insecticides, sterile insect technique (SIT) and odor-baited traps. Whereas insecticides have undesirable environmental impacts, the high cost of SIT limits it application. Consequently, use of host-derived odors is an attractive option for tsetse control. However, this method is faced with a challenge of differential responses exhibited by *Glossina* species, which has hindered establishment of a universal control strategy. In this study it was hypothesized that the differential responses exhibited by various tsetse species are encoded by their chemosensory proteins. Availability of complete genome sequences of five tsetse species provided the opportunity to test the study hypothesis. For this, genome-wide annotation of chemosensory genes was carried out in four species: *Glossina austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes*. The chemosensory genes in these four *Glossina* species were compared to those reported in *G. m. morsitans* and other closely related dipterans (fruit fly, housefly and mosquito). Expression abundance for the genes identified in *G. m. morsitans* was determined in non-olfactory tissues and the binding dynamics of an olfactory specific binding protein; Obp83a1 compared among the five tsetse species. The data revealed a reduced but rather conserved chemosensory repertoire in all tsetse species in relation to other insects compared. High expression abundances of some odorant-binding proteins and chemosensory specific proteins in non-olfactory tissues suggest their involvement in reproduction and development. Molecular dynamics carried out on Obp83a1 homologs potentially supports its engagement in host seeking as suggested by an earlier study on *G. m. morsitans*. This study has provided a comprehensive chemosensory repertoire necessary for undertaking functional genomics on tsetse chemosensory genes to unravel novel molecules for development of improved control strategies. The results will enhance our understanding of tsetse's chemical sensing and provide insights into development of improved control approaches.

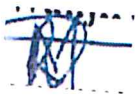
Key words

Olfaction, chemosensory genes, annotation, tsetse biology, RNA profiling, odorant binding, vector control, trypanosomiasis, *Glossina*, receptors

Declaration

I declare that "Characterization of Olfactory Responsive Genes in Selected Tsetse Species: Annotation and Comparative Analyses of Chemosensory Proteins in the Genus *Glossina*" is my own work, that it has not been submitted for any other degree or examination in any other university, and that all sources I have used have been indicated and acknowledged through complete referencing.

Name: Rosaline Wanjiru Macharia **Date**: 15th December, 2015

Signed: 

Acknowledgements

I give thanks to God for the good health and sufficient grace to complete this endeavor. May this achievement be for his glory!

Numerous people played key roles toward completion of this thesis. First, I would like to appreciate my advisor; Prof. Alan Christoffels and co-supervisors Dr. Daniel Masiga and Dr. Paul Mireji who provided immeasurable support and guidance throughout my PhD studies. I am grateful for your close follow-up on my progress and for allowing me to exercise intellectual freedom. Secondly, I would like to thank Prof. Serap Aksoy for her directed feedback throughout my research. I also am grateful to Dr. Grace Murilla and her team whom I had a chance to work with on various occasions.

I appreciate the support I received from the MBBU team at *icipe* where I was based. Your criticism and feedback not only provided insight to my work but also an opportunity to improve my soft skills. In a special way, I thank Collins Omogo for his assistance with graphics and Esther Waweru for always attending to me with a smile. Similarly, I would like to thank all the capacity building staff at *icipe* for ensuring my comfort while at work and facilitating my participation in various conferences and courses.

I acknowledge my donors; DAAD who financed my studies and NIH-FIC who funded my research. I would like to thank SANBI staff members; Maryam and Samantha for their prompt correspondence at all times. I also appreciate computational support that I received from International Livestock Research Institute through Dr. Wamalwa and Allan Oarth.

The mentorship of Dr. Henri Kariithi cannot go unmentioned. Kariithi not only inspired me but was keen to bring the best out me through "tough" and timely feedback whenever I consulted him. Similarly, am thankful for the help I got from Thomas Musyoka of Rhodes University while working on molecular dynamics.

All this would not have been possible without the support, prayers and sacrifice of my family members (Mom, Margaret, Sister, Carol, Brother, Raphael, nephew Carnell and my grandma, Leah) who had to keep up with fewer hours of me. Your support is what kept me focused on my goal. I pray that the lord may expand your territories and strengthen the cord of love among us. I also appreciate my friends: Lucy, Alice, Christine, Hellen, Grace, Maureiq, Nelly, Erick, Ngao, Ian, Caleb, Cyrus, George and my fellow scholars for being there to check my sanity as I walked this arduous journey.

Dedication

This thesis is dedicated to the Macharia's for their love and support and in loving memory of our late dad, Moses Macharia.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

1.0 Background Information & Literature Review

1.1 Epidemiology and Transmission of African Trypanosomiasis

African trypanosomiasis is a neglected vector-borne disease caused by protozoan parasites of order Kinetoplastida (Alsford *et al.*, 2013). The disease affects both humans and livestock; approximately 70 Million people and 50 Million of livestock in rural areas are at risk of contracting the disease. The number of reported cases have dropped significantly over the years with less than 10,000 cases having been reported in 2009 (Figure 1.1) and only 7139 cases in 2010 (Simarro *et al.*, 2011). Nevertheless, reported cases of *Trypanosoma brucei rhodesiense* infections are thought to be underestimated given that the disease mainly occur in remote areas with limited access to health facilities (Odiit *et al.*, 2006).

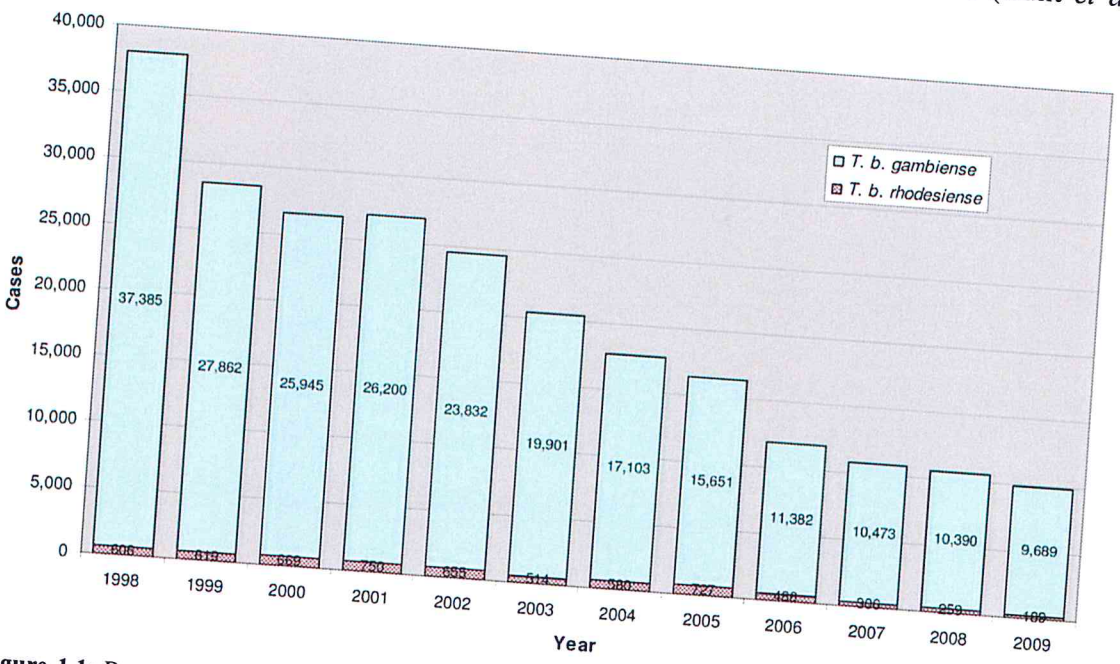


Figure 1.1: Reported cases of human African trypanosomiasis between 1998 and 2009 (Figure adapted from Simarro *et al.*, 2009).

Trypanasoma brucei rhodesiense is responsible for acute form of Human African Tryapanosomiasis (HAT) experienced in east and southern parts of Africa. This acute form of HAT is transmitted by *G. m. morsitans, G. pallidipes, G. swynnertoni and G. austeni* (Aksoy, Hao and Strickler, 2002; Gooding and Krafsur, 2005; Krafsur, 2009). On the other hand, the chronic form of HAT experienced in west and central Africa is caused by *T. b. gambiense*,

which is transmitted *by G. palpalis, G. f. fuscipes and G. tachinoides* (Gooding and Krafsur, 2005). The latter form of HAT accounts for over 95% of reported cases while the "Rhodesian" HAT accounts for less than 5% of the reported cases. Other trypanosomes such as *T. vivax, T. b. brucei*, and *T. congolense* are responsible for acute nagana/African Animal Trypanosomiasis (AAT) in cattle. Their relatives, *T. simiae* and *T. suis*, are effective pathogens of the chronic AAT and Suoma in domestic pigs (Mugasa *et al.*, 2008). Animals such as antelopes, cattle, dogs and pigs are reservoirs of *T. b. rhodesiense* but play an insignificant role in spreading *T. b. gambiense* (Kabasa, 2007).

Tsetse flies (Diptera: Glossinidae), the sole cyclical vectors of the African Trypanosoma parasites; inhabit about 40% (38 countries) of Africa. Tsetse flies are classified into three sub-groups: i.e. Morsitans (Savannah flies), Fusca (Forest flies) and the Palpalis (riverine flies) based on a combination of morphological, behavioral, geographical distribution and molecular differences (Gooding and Krafsur, 2005). The three sub-groups of tsetse are further sub divided into four sub-genera including fusca, morsitans, palpalis and austeni which have been implicated as the leading vectors of trypanosomiasis (Aksoy *et al.*, 2002).

Transmission of HAT mainly occur during social-economic activities (Alsford *et al.*, 2013) and the disease is transmitted either mechanically or biologically (Masiga *et al.*, 2002). Other modes of transmission include mother to child, blood transfusion and accidental infections that may occur in a laboratory setting.

1.2 Clinical Manifestation and Management of African Trypanosomiasis

The two forms of trypanosomiases manifests through two distinct stages (Brun et al., 2010). (i) Haemo-lymphatic stage, which involves the spread of trypanosomes throughout the blood and lymphatic system of the higher vertebrate hosts. This stage is characterized by common symptoms like headache, fever, inflammation and lymphadenopathy. (ii) Meningoencophalitic stage, which is characterized by brain-barrier crossing of trypanosomes where they cause infection in the central nervous system (CNS) (Brun *et al.*, 2010). At this stage, sleep, speech and psychiatric disorders are reported. If left untreated, the disease progresses to coma and later to death (Duggan and Hutchinson, 1966). The main control strategies for management of African trypanosomiasis include chemotherapy (for treatment) and vector control (for disruption of parasite transmission).

1.2.1 Chemotherapy

Chemotherapy of HAT uses old and expensive drugs such as pentamidine, suramin, melasoprol and eflornithine; all of which have shown various degrees of resistance and/or toxicity (Balasegaram *et al.*, 2009; Brun *et al.*, 2010; Anene *et al.*, 2001). Further, the treatment schedules of the available HAT drugs are overly prolonged, excruciatingly painful and require continuous hospitalization to monitor the patients (Matovu *et al.*, 2001). In addition, cross-resistance between melasoprol and petamidine has been observed *(Baker et al.*, 2013). Poor infrastructure in most sub-Saharan African countries hinders availability and delivery of drugs to the rural populations (Barrett *et al.*, 2007). Further, antigenic variation of trypanosomal surface glycoprotein which occurs inside the mammalian bloodstream has frustrated efforts towards development of effective vaccines. This has in-turn limited disease management to active surveillance, prophylaxis, chemotherapy and vector control (Aksoy *et al.*, 2001; Horn and McCulloch, 2010). Prophylaxis use is however not recommended due to drug toxicity and the low chances of infection (Brun *et al.*, 2010).

Similarly, treatment of AAT relies on relatively old trypanocidal drugs such as isometamidium chloride, homidium bromide and diminazene aceturate (Alsford *et al.*, 2013). It is estimated that about 35 million doses of these drugs are administered by farmers annually, a therapy that is relatively expensive to maintain. Similar to HAT, drug resistance in AAT has been reported in highly endemic regions (Geerts *et al.*, 2001). Based on the available tsetse and parasite genomic data, there are on-going efforts to understand resistance in anti-Trypanosoma drugs; it is hoped that novel and perhaps more efficient therapies can be developed through these efforts (Alsford *et al.*, 2013).

1.2.2 Vector Control and Disease Surveillance

Following the second major outbreak of HAT (1920s and 1940s), control measures such as trapping, clearing of bushes and disease surveillance were put in place in most endemic countries (Brun *et al.*, 2010). These strategies nearly eliminated the disease by 1960s, but deployed traps were vandalized by civil war after independence, leading to resurgence of the disease in West Africa in the 1990s. Nevertheless, vector control remains the corner stone of the disease suppression in Africa (Brun *et al.*, 2010).

Tsetse population control through life-cycle disruption has shown profound success in reducing the number of flies (Barrett *et al.*, 2007). Vector control is implemented either by use of bait-technology (Green, 1994) and/or Sterile Insect Technique (SIT) (Feldmann *et al.*,

2005). The SIT is achieved through introduction of sterile male insects (on a ratio of 10:1, sterile to wild) and is only effective within the concept of area-wide integrated pest management (AW-IPM). The SIT has been applied successfully in eradication of *G. austeni* from Zanzibar (Vreysen *et al.*, 2000). Nevertheless, SIT is costly and thus has not been implemented widely (Aksoy, 2003; Gooding and Krafsur, 2005). Alternatively, bait technology, which exploits host excreta to trap and kill, or repel insects from the hosts (Omolo *et al.*, 2009) has been adopted to control tsetse population. Use of baits is cheaper and easy to implement compared to SIT. However, implementation of bait technology is tedious and its application in rural areas has been disrupted by lack of funds and damage by humans and/or wild animals or by heavy storms (Solano *et al.*, 2010).

Development of attractants and repellents that are currently used for tsetse control took many years due to the tedious bio-assay techniques applied in their screening (Aksoy, unpublished). Their application has however gained popularity since the development of a successful attractant against the Savannah species based on cow and buffalo odor (Logan and Birkett, 2007). However, no effective attractants have been established for the riverine flies (Palpalis sub-group) such as *G. f. fuscipes*; the major vector for HAT in sub-Saharan Africa (Omolo *et al.*, 2009).

With genomic data becoming increasingly available, discovery of novel bioactive molecules is feasible (Liu *et al.*, 2010). Thus, it is important to take advantage of available genomic data to understand the mechanism of chemical sensing in disease vectors. These efforts coupled with the World Health Organization (WHO), HAT control and surveillance worldwide alliance re-established in 2012 will play a significant role in combating and perhaps eliminating this neglected disease.

1.3 Chemoreception in insects

Chemoreception (the organisms' ability to detect biologically relevant chemicals in their environment) is crucial for the organism's survival. It mediates seeking of food, selection of suitable mates and escape from predators. Due to its role in ecological processes, chemoreception has gained popularity among researchers whose interest is in pest control and/or animal and human health. Chemosensation process in insects share similarities with that of mammals in that their receptors function as hetero-dimers and that they have preference for sweet, bitter tastes and/or aversion to bitter substances (Zhao *et al.*, 2003; Larsson *et al.*, 2004; Benton *et al.*, 2006). On the other hand, the insect chemosensory systems significantly differ from that of mammals, .suggesting differences in their evolution.

Chemosensation is divided into two broad categories namely, olfaction and gustation. Olfaction refers to the detection of volatile chemicals present in the environment, while gustation refers to the detection of non-volatile chemicals in diet of an organism. Majority of organisms have evolved different sets of sensory organs to detect volatile and non-volatile chemicals. Taste cells project their axons to the nucleus of the solitary tract (Hamilton and Norgren, 1984) while olfactory sensory neurons (OSNs) project their axons towards olfactory bulb (Mori *et al.*, 1999).

1.3.1 Olfaction

Olfaction forms basis of host selection, ovipostioning and predator detection. Olfaction has been exploited as a tool control of disease vectors through manipulation of their behavior (Carey and Carlson, 2011). Olfaction takes place in the OSNs, which are mainly located on antennae and maxillary pulpy of adult insects. The sensory neurons extend their dendrites into hair-like structures, called sensilla. Each sensillum houses between one and 50 neurons, depending on the insect species (Ochieng *et al.*, 1998). The sensilla consists of pores through which odorants pass into the fluid lymph surrounding the OSNs (Riesgo-Escovar *et al.*, 1997). Effectiveness of sense of smell is dependent on the ability of peripheral proteins to selectively detect and rapidly inactivate stray odorants (Leal, 2011). Various proteins and receptors are involved in perireceptor events of odorant processing such as uptake, binding, transport and inactivation of odor molecules (Leal, 2011). These include chemosensory proteins (CSPs), ionotropic receptors (IRs), odorant binding proteins (OBPs), pheromone binding proteins (PBPs), odorant degrading enzymes (ODEs), odorant receptors (ORs) and sensory neuron membrane proteins (SNMPs).

1.3.1.1 Ionotropic Receptors (IRs) and Ionotropic Glutamate Receptors (iGluRs)

With exception of two co-receptor IRs (IR8a and IR25a), all other IRs are expressed in sensory dendrites of coeloconic sensilla, as well as gustatory neurons in the proboscis and mechanosensory neurons (Benton *et al.*, 2009). The IRs share structural domains with ionotropic glutamate receptors (iGluRs), but their sequences are highly divergent from kainate, alpha-amino-3-hydroxy-5-methyl-isoxazole-4-propionic acid (AMPA), or N-methyl-D-aspartate receptor (NMDA) classes of iGluRs (Benton *et al.*, 2009). Classic iGluRs are characterized by two glutamate-binding modules separated by an ion channel pore; iGluRs are known to function in the CNS by binding onto glutamate, a neurotransmitter (Mayer, 2011). On the other hand, IRs lack the critical residues coordinating the glutamate binding,

and have been shown to act as receptors for volatiles (Benton *et al.*, 2009). Further, identity across the IR family ranges between 10- 70% in *Drosophila* suggesting maintenance of their ability to act as ion channels upon ligand binding. Hence, unlike mammals, insects have evolved additional iGluR-like genes to function as chemosensory receptor family enhancing their olfaction sensitivity. Unlike ORs, IRs are considered more ancestral and have been reported to participate in non-olfactory roles in lower organisms (Groh-Lunow *et al.*, 2014).

1.3.1.2 Odorant Receptors (ORs)

Insect ORs, which are characterized by the presence of seven transmembrane domains, were first established in *D. melanogaster* (Clyne, Warr and Carlson, 2000). The ORs lack sequence homology with mammalian G-protein-coupled receptors (GPCRs) (Benton *et al.*, 2006; Marshall *et al.*, 2010) and exhibit a reversed transmembrane topology with their N-terminal resting in the cytoplasm (Benton *et al.*, 2006). Benton and colleagues (2006) also believed that ORs form heterodimers with non-canonical co-expressed receptor (Orco/Or83b) to create an ion-gated channel similar to that of GPRCs. Orco acts both as a chaperon protein and as a cognate co-receptor contributing to early tuning of the odorant receptors (Kaupp, 2010). Orco (Or83b) differs from other olfactory receptors in that it is the only OR that is expressed along with a neuron-specific conventional OR that interacts with odorant ligands. Other ORs are highly variable in sequence within and across various insect species (Clyne *et al.*, 2000; Sato *et al.*, 2008). Nevertheless, neurons expressing the same odorant target the same glomerular structures in the antennal lobe (Fishilevich and Vosshall, 2005).

Ubiquitous co-expression of a specific OR and Orco is thought to render cation channels permeable to sodium (Na^+) ions, potassium (K^+) ions, and calcium (Ca^{2+}) ions, leading to odor sensitivity in insects. The ttargeting of the glomerulus in the brain by OSNs suggests a similarity of insect olfaction to that of the vertebrates (Yao and Carlson, 2010).

1.3.1.3 Odorant Binding Proteins (OBPs)

Insect OBPs constitute most insect sensillum proteins (Hekmat-scafe *et al.*, 2002; Zwiebel and Takken, 2004). The OBPs were first described in 1981 in *Antheraea polyphemus* moth. They appear as small globular molecules made of approximately 150 amino acid residues with a conserved six cysteine residues that help to maintain their protein tertiary structure (Xu *et al.*, 2009). They are highly diverse showing 5-80 % sequence identity and are broadly classified into general OBPs (GOBPs) and pheromone OBPs (PBPs), which bind

odors with different sensitivity and specificity (Rogers, Steinbrecht and Vogt, 2001). Various roles have been described for GOBPs, including recognition and transportation of volatile odorants through sensilla lymph to the ORs (Liu *et al.*, 2012). Indeed, OBPs are considered the link between the external environment and odorant receptors; they bind onto hydrophobic odorants and transport them to the odorant receptors through the aqueous lymph (Hekmat-scafe *et al.*, 2002; Ishida and Leal, 2005). Comparably, PBPs are associated with pheromone binding, transportation and clearance of odors from the lymph (Hallem and Carlson, 2006). The role of PBPs in olfaction has been demonstrated using lush (OBP76a) found in fruit fly's sensilla which specifically binds to a pheromone component 11-cis vaccenyl acetate (cVA) and *An. gambiae* (Agor1) which acts a receptor to the 4-methylphenol component of human sweat (Hallem *et al.*, 2006).

Compared to other insects, few OBPs(n=20) were reported in the first genome of tsetse (*G. m. morsitans*) (Liu *et al.*, 2010). This number was latter revised to (n=32) through manual annotation(Obiero, 2014). Eight of these OBPs were reported to be highly transcribed in the antennae of female flies. Liu and colleagues (2010) suggested that three of the OBPs are likely to function as olfactory specific genes given that they were highly transcribed in female species. Females have a higher demand for blood meal as compared to their male counterparts.

1.3.1.4 Chemosensory Specific Proteins (CSPs)

Unlike OBPs, CSPs are more conserved across many insect species, but are slightly smaller in size (~130 aa long) compared to the OBPs. The CSPs harbor a signature of four conserved cysteine residues, which results in different 3D protein structures compared to the structures of the OBPs. However, to date, there is little available evidence of their involvement in olfaction. CSPs are expressed in various insect tissues, including the antennae, head, thorax, legs, wings, ovaries, wing disks and eyes (Gong *et al.*, 2007; Sato *et al.*, 2008). These proteins are thought to play a role in transporting chemical compounds to the receptors on the dendritic membranes (Gong *et al.*, 2009). Other proposed roles of CSPs include CO_2 detection, legs regeneration, pheromone transportation and larval development (Briand *et al.*, 2002; Forêt *et al.*, 2007). In the tsetse fly; *G. m. morsitans*, a total of five CSPs have been described (Liu *et al.*, 2012). Similar to the case of the OBPs, three of the five CSPs were shown to be highly expressed in the antennae, suggesting their involvement in olfaction. Like in other insects, the tsetse CSPs depict high sequence divergence and close phylogenetic

relationships, suggesting that dipteran CSPs may have arisen from three ancestral genes (Liu *et al.*, 2012).

1.3.1.5 Odorant Degrading Enzymes (ODEs)

For effective navigation of the environment, an organism must not only be able to detect the important semiochemicals, but must also quickly inactivate the chemical signal as soon as the message is conveyed (Leal, 2011). This is achieved by odorant degrading enzymes (ODEs) (e.g esterases) that reside in the sensilla lymph. Some of the ODEs have been identified. For instance, *A. polyphemus* pheromone degrading enzyme (ApolPDE) (Leal, 2011) has been shown to be present at the pheromone sensitive sensilla from pupal stage and reaching its peak at adult stage. Adult ApolPDE is involved in degradation of E6Z11-16OAc (a sex pheromone component) (Ishida and Leal, 2005). Other ODEs include: antennal aldehyde oxidases, aldehyde dehydrogenases, epoxide hydrolases, glutathione-S-transferases, and cytochrome P450s; all which have been described through *in vitro* experiments (Blomquist and Vogt, 2003).

1.3.1.6 Sensory Neuron Membrane Proteins (SNMPs)

The SNMPs are members of a gene family characterized by human CD36 domain. In humans, SNMPs are involved in varied roles including transportation of fatty acids and cell-cell recognition (Nichols and Ã, 2008). In insects, SNMPs are known to associate with chemosensory sensilla and are highly conserved throughout Diptera (Vogt *et al.*, 2009). SNMPs 1 and 2 were first discovered in *Manduca sexta* and found to share up to 40% sequence identity (Rogers et al., 2001). The SNMP1 ortholog has been shown to play an essential role in the detection of *D. melanogaster* male pheromone cVA (Jin, Ha and Smith, 2008) but no evidence of SNMP2's involvement in olfaction has been described to date. Genome annotation of *G. m. morsitans* (Obiero, 2014) reported presence of orthologs for both SNMP1 and SNMP2 in tsetse fly, but no functional studies have been conducted on them.

1.3.2 Gustation

Gustation (taste perception) is employed by insects for detection of sugars, bitter compounds and pheromones. Gustation is mediated by GRs, which are expressed on taste pegs located on various insect body parts including proboscis, legs, wings among others (Falk,1976).

1.3.2.1 Gustatory Receptors (GRs)

Similar to other insect chemoreceptors, GRs were first described in *Drosophila* shortly after discovery of ORs (Clyne *et al.*, 2000). Their phylogenetic relationship with ORs suggested a functional overlap between these two protein families (Robertson *et al.*,2003). Insect GRs are members of a large GPCR family, which are co expressed with other GRs in single receptor neurons (Isono *et al.*, 2010). Unlike *Drosophila* that has up to eight genes encoding sugar receptors, the tsetse fly; *G. m. morsitans* lacks all the sugar receptors (Obiero *et al.*, 2014). On the other hand, *G. m. morsitans* was found to have expanded the CO_2 responsive gene (Gr21a), implying a reliance on CO_2 by tsetse flies to locate their hosts (Torr *et al.*, 2006).

1.4 Genomics and Transcriptomics

The study of gene structure and expression profiles is crucial in determining their functions. These type of studies are broadly categorized into genomics (study of entire genomes) and trancriptomics (study of transcriptomes from given tissues and/or at given physiological state).Accurate gene identification coupled with the determination of their abundance in a given physiological state is critical in understanding biology of an organism (Wang *et al.*, 2009). Genomic studies aim at determining the gene structure, while transcriptomics focuses on mapping splice junctions and gene expression at RNA levels (Dong and Chen, 2013). Biologists have used various approaches including candidate gene approach, microarray technology and sequence based approach to correlate between genotypic and phenotypic characteristics of multicellular organisms (Morozova *et al.*, 2009). Each of these technologies has their advantages and disadvantages in resolving long outstanding biological questions.

1.4.1 Candidate Gene Approach

Candidate gene approach involves investigation of total RNA derived from cells, tissues or disease states for transcripts of interest based on low-throughput technologies such as Northern blot analysis (Alwine et al., 1977). This technique is simple, but requires large amounts of RNA as input. Alternatively, quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) (Becker-André and Hahlbrock, 1989) that requires less amounts of input mRNA has been employed in abundance estimation of mRNA (Morozova et al., 2009).

1.4.2 Micro-Array Technology

The advent of the micro-array technique that could achieve expression of thousands of transcripts at the same time was a great achievement in the field of transcriptomics. The technology slowly replaced a single-gene expression approach as it allowed detection of non-coding RNAs, single nucleotide polymorphisms (SNPs) and alternative splicing events (Becker-André and Hahlbrock, 1989). Despite its power in evaluating abundance of over a thousand transcripts, the technique does not detect novel transcripts adequately, suffers noise and is costly; factors that have made it less common (Morozova *et al.*, 2009).

1.4.3 RNA-Seq: A Transcriptome Profiling Tool

Cheaper sequencing techniques have advanced to directly identify and quantify novel transcripts thus replacing micro-array technology in transcriptome analysis (Morozova *et al.*, 2009). Among them, RNA-Seq has gained wide application in many laboratories. It relies on deep sequencing in which RNA is used to construct a library of complementary DNA (cDNA) (Wang *et al.*, 2009) segments with linked adaptors on both sides as illustrated in Figure 1.2. RNA-Seq experiments are executed in four steps: (1) Sample preparation and library construction; (2) Sequencing; (3) Assembly; and (4) Downstream analysis.

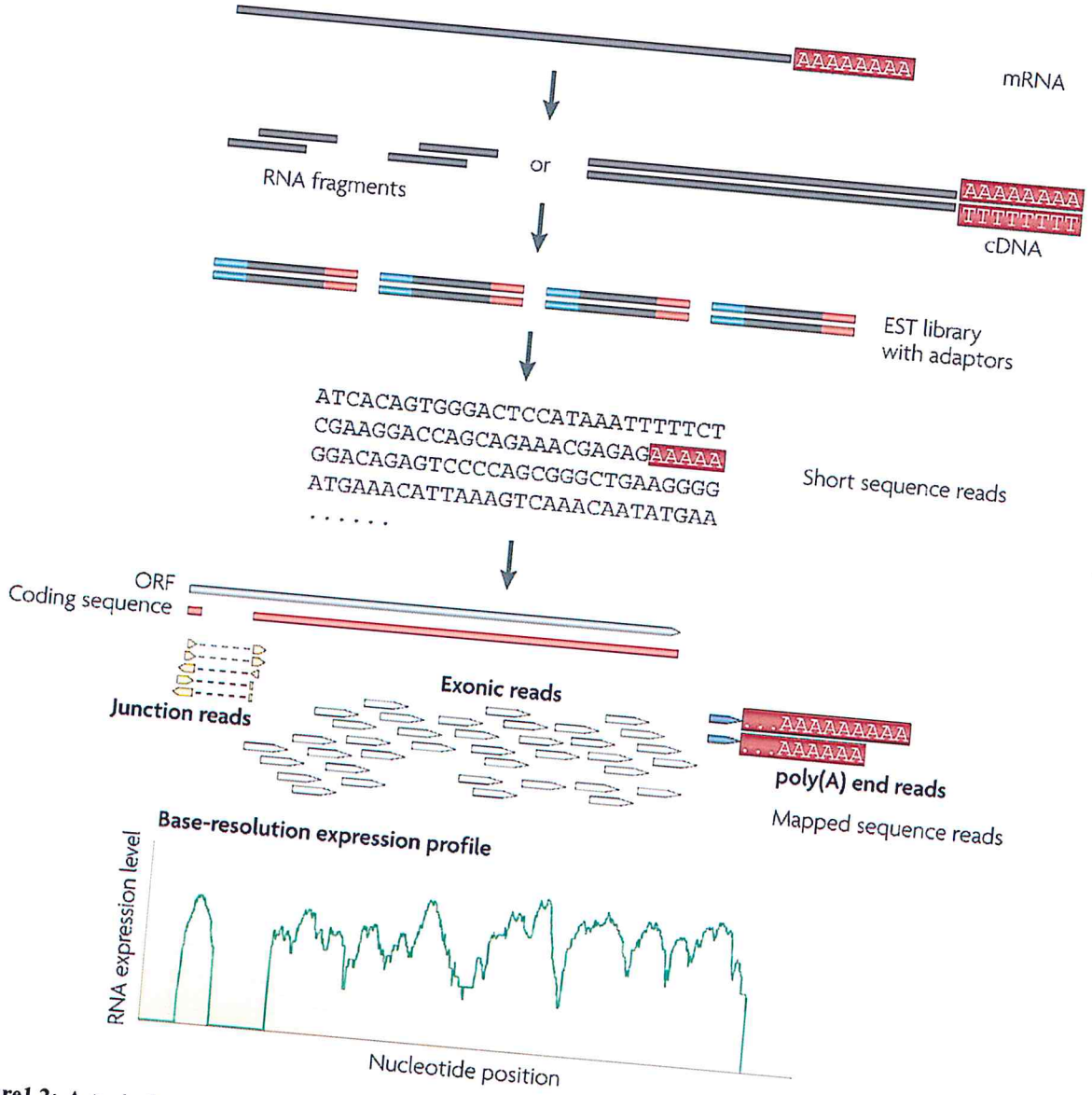


Figure1.2: A typical RNA-Seq experiment. Long RNAs stretches are first converted into a library of cDNA through fragmentation. Adaptors (blue) are then added to each cDNA fragment and high-throughput sequencing technology is applied to obtain short reads. The short reads are assembled and used in prediction of open reading frames (ORF) which are then mapped on the reference genes regions and identified into exonic reads, junction reads and the poly A end read based on the region they map onto. RNA expression is then estimated for each nucleotide in the read (Figure adapted from Wang *et al.*, 2009)

Unlike hybridization and expressed sequence tag (ESTs) sequencing, RNA–Seq is not limited to known transcripts, presents less background signal and reveals the location of transcription boundaries with higher precision (Costa *et al.*, 2010). More so, RNA-Seq data allows correct gene annotation and the technique shows high reproducibility (Costa *et al.*,

2010). RNA-Seq has gained popularity in studying non-model organisms due to its added advantage over micro-array (Vera *et al.*, 2008).

1.4.3.1 RNA Preparation and Library Construction

This step involves isolation of RNA samples followed by further processing of the extracted RNA molecules, determined by the purpose for which the transcriptome is needed. For complete transcriptome analysis, rRNA molecules are depleted via hybridization with rRNA sequence-specific 5 -biotin labelled oligonucleotide probes, followed by their probe removal using streptavidin-coated magnetic beads from total RNA (Costa *et al.*, 2010). Like in DNA sequencing, double stranded cDNA library is prepared using fragmented cDNAs or fragmented RNA with the latter being preferred for maintenance of strand information that is useful in informing gene orientation.

1.4.3.2 Sequencing

Different next generation technologies can be applied to sequencing of transcriptomes and genomes. An early method that is employed in transcriptome sequencing is pyrosequencing (Roche 454) which is based on bioluminescence. The pyrosequencing protocol correlates release of a pyrophosphate to the amount of light produced. It however suffers from insertion or deletion (indel) sequencing error type (Costa *et al.*, 2010). On the other hand, Illumina and Heliscope rely on colour dyes to monitor reversible terminators that are employed in a cyclic manner. The two techniques differ in the number of dyes used, but they both introduce substitution error and under representation of A-T pairs. In contrast, ABI Solid is based on sequencing by ligation (SBL) and the use of one or two nucleotide probes that are based on colour space (Costa *et al.*, 2010). The latter, has advantage over others in that it shows improved accuracy in detection of Single Nucleotide Variation (SNV) in the resultant sequence. Millions of reads ranging between 30-400 bp are yield depending on the sequencing technology used.

1.4.3.3 Assembly

High resolution sequencing techniques such as RNA-Seq and whole genome sequencing technique yield higher base pair resolution unlike low-throughput technologies that only detect abundant transcripts (Martin and Wang, 2011). However, they output short reads, which introduce complexity into their downstream analyses. This shortfall necessitates reconstruction of full-length transcripts through stacking together the short reads; often

referred to as transcriptome assembly. This process requires high performance computing to handle the numerous amounts of data and sequencing errors introduced by different sequencing techniques.

A number of tools including Velvet (Zerbino and Birney, 2008), ABYSS (Simpson *et al.*, 2009), Soap *de novo* (Xie *et al.*, 2012) among others, have been developed and applied in assembly of short genomic reads but suffer shortcomings that limit their application in transcriptome assembly. The shortcomings include: (i) reliance on sequencing depth to identify repeats, thus classifying abundant reads as possible repeats, (ii) lack of strand information to resolve overlapping transcripts since both strands are sequenced in genomic experiments unlike in RNA-Seq and (iii) their inability to resolve ambiguity resulting from gene variants that share exons (Wang *et al.*, 2009).

Successful assembly depends on proper designing of an RNA-Seq experiment, choice of sequencing technique with either paired or single-end reads and pre-processing of reads to remove artefacts. To date three assembly strategies have been documented and reviewed by Martin and Wang (2011). They include reference-based strategy, de novo assembly strategy and combined strategy.

1.4.3.3.1 Reference-Based Assembly Strategy

This strategy, also referred to as *ab initio* assembly technique, relies on the availability of a reference genome upon which assembly of the transcriptome is constructed (Grabherr *et al.*, 2011; Martin and Wang, 2011). The strategy is implemented in three steps, which include alignment of reads using splice-aware aligners such as Bowtie (Langmead, 2010), Tophat (Trapnell *et al.*, 2010) or GSNAP (Wu and Nacu, 2010). Clustering of overlapping reads into a graph that contains all possible isoforms is done, followed by resolution of the individual isoforms. Reference based strategy has been implemented using various tools including; Cufflinks (Trapnell *et al.*, 2010) and Scripture (Guttman *et al.*, 2010), and is known to be highly sensitive, thus can be used to assemble transcripts of low abundance (Wu and Nacu, 2010). It requires less random access memory (RAM) and can be executed through parallel computing. Its performance is not affected by sequencing artefacts (Trapnell *et al.*, 2010). However, its success is dependent on the quality of reference genome used and it suffers introduction of unmapped gaps in regions that do not match uniquely on to the reference genome (Martin and Wang, 2011).

1.4.3.3.2 De novo Assembly Strategy

For this strategy, no reference genome is required. This strategy can be applied in assembly of non-model organisms. It is based on the principle of building transcripts on a De Bruijn graph using repetitive short-reads (Martin and Wang, 2011). The transcripts are further merged into contigs to eliminate redundancy. Implementation of this strategy has been done in various software including, but not limited to Rnnotator (Martin *et al.*, 2010) NN, Multiple-K (Surget-Groba and Montoya-Burgos, 2010) and TransABYSS (Robertson *et al.*, 2010). Trinity (Grabherr *et al.*, 2011) is yet another software that is widely used in de novo transcriptome assembly. Unlike other programs, Trinity uses a greedy stepwise approach that first assembles the unique set of reads and then pools together the sets of unique sequences that overlap and uses them to create De Bruijn graph for each group of sequences (Martin and Wang, 2011). It has been shown to perform better than other de novo assembly software (Grabherr *et al.*, 2011).

Unlike a reference based strategy, success of *de novo* assembly does not depend on the correct alignment of known splice regions in the reference sequence. Again, it has the ability to unveil more novel transcripts as compared to reference based strategy despite the presence of longer introns (Martin and Wang, 2011). Nevertheless, *De novo* based approach requires high computing resources as compared to reference based strategies. In addition, *De novo* based approaches are more sensitive to sequencing artefacts and requires deep sequencing with higher coverage (Martin *et al.*, 2010).

1.4.3.3.3 Combined Assembly Strategy

Combined assembly strategy exploits high sensitivity of reference based assemblers, which complements the ability of de novo assemblers to detect previously unknown transcripts (Martin and Wang, 2011). Implementation of this strategy could either be done through alignment followed by assembly or vice versa. In cases where a reference genome is available, the align then assemble approach is used. In the case where the reference genome is either derived from a close species or its quality is in question, assemble then align approach is preferred (Martin and Wang, 2011). Increasing read coverage and the use of different read types can also be used to improve quality of transcriptome assembly (Costa *et al.*, 2010). Combined strategy is however not yet implemented in available software and thus its shortcomings remain unknown.

The choice of transcriptome assembly strategy ultimately depends on the existence of reference genome, the type of data to be assembled and the goal of assembly. Statistical measures are used to describe the quality of a genome assembly, with the most common being N50 which is defined in terms of the length of the shortest contig/scaffold in the assembly (Yandell and Ence, 2012).

1.4.3.4 Downstream Analysis

To date, there are no 'gold-standard' tools for performing down-stream analysis on transcriptome data (Costa *et al.*, 2010). The choice is rather influenced by the sequencing technique employed to generate data and the kind of analysis to be performed on the data. Nevertheless, it is important to filter out low quality reads before embarking on any type of analysis. This helps to save on computational resources needed to accomplish a given task (Costa *et al.*, 2010). Common downstream applications of transcriptome data involve gene identification and annotation, quantification of gene and isoform abundance and differential expression analysis.

Prior to genome annotation, a computational intensive process (known as repeat masking) is necessary to avoid false gene inclusion. This involves identification of regions of low-complexity such as homopolymeric nucleotides and transposons and coding them as 'N' or 'a' in case of Adenine 't' in case of Thiamine, 'g' in case of Guanine and 'c' in case of Cytosine (Stein, 2001; Yandell and Ence, 2012). Tools such as RepeatMasker (Smit *et al.*, 1996) have been used successfully for this task.

1.4.4 Gene Finding

Scientists rely on accurate gene prediction and annotation to deduce evolutionary relationships among species from the growing volumes of genomic data. Eukaryotic gene prediction is however made complex by low gene density in the corresponding genomes and thus requires sophisticated software (McElwan, Unpublished; 2007).

To achieve accuracy, any gene prediction software should define exact boundaries of protein-protein coding regions including the regulatory region. To the contrary, most software is designed to identify protein-coding regions of the genes leaving out the regulatory regions that confer specificity of gene expression. Available gene predictors are classified into: (i) similarity based software that use cDNAs, ESTs and homologs from databases and (ii) Motif-based '*ab initio*' software that use nucleotide content to predict structure of genes (McElwan, Unpublished; 2007). The latter are said to be more accurate than the similarity

based software. This is associated to incompleteness and sequencing errors that could be introduced by so called similar sequence data (Mathé *et al.*, 2002).

Commonly used gene predictors include Fgenesh (Salamov, 2000); based on hidden Markov models (HMM) and suitable for finding genes in human, *Drosophila*, plants, yeast and nematodes. A newer version of this program, Fgenesh+ incorporates similarity data and has been shown to yield better results (McElwan, Unpublished; 2007). Glimmer *(Salzberg et al.,* 1999) is another predictor that was first designed to predict genes in prokaryotes based on an interpolated Markov model (IMM), but was later modified to GlimmerM (Majoros *et al.,* 2003), that includes an algorithm for predicting alternative splice sites in eukaryotic genomes. Similar to Fgenesh, GENMark (Borodovsky and McIninch, 1993) predicts genes based on HMM and has been modified to work for eukaryotic and viral genomes (Besemer and Borodovsky, 2005). Combining two or more predictors could help improve on accuracy of gene predictions.

Other gene predictors include AUGUSTUS (Stanke and Morgenstern, 2005) (http://bioinf.uni-greifswald.de/augustus/) *ab initio* gene finder for eukaryotic genomic sequences. Unlike Fgenesh, AUGUSTUS has the ability to predict alternative splicing and untranslated regions of genes and exons. So far, AUGUSTUS has been trained with various species data and has been shown to outperform other *ab initio* gene predictors (Stanke and Morgenstern 2005).

1.4.5 Gene Annotation

Gene annotation is divided into two broad categories namely structural annotation (the process of defining coding regions in a genome) and functional annotation (assigning functions to identified genes) (Stein, 2001). Structural annotation (Yandell and Ence, 2012) entails construction of protein-encoding structures intron-exon junctions, lengths, coding and non-coding regions, and untranslated regions (Curwen et al., 2004) as shown in Figure 1.3.

Annotation is executed in two distinct phases that include (i) computational phase which overlaps with automatic gene predictions covered above and (ii) annotation phases (Yandell and Ence, 2012) which involve nucleotide annotation, protein annotation and process annotation (Stein, 2001).

1.4.5.1 Annotation Pipelines

Annotation of whole genomes can be achieved either using computational pipelines which rely on evidence from a combination of bioinformatics application or online annotation

tools else referred to as framework annotation, contributed to by the research communities. The latter approach is to date valued as the most reliable approach for accurate genome annotation (Curwen *et al.*, 2004). Nevertheless, manual annotation is slow and does not offer a practical approach to the rising need for genome annotation. Hence, automated annotation pipelines are commonly used (Curwen *et al.*, 2004). Some of the most widely used resources include Ensemble (http://www.ensembl.org/index.html) and the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) which provides access to annotation data. Others include University of California Santa Cruz (UCSC) genome browser (Kent *et al.*, 2002) and the Distributed Annotation System (DAS) *(Dowell et al.*, 2001) which use "framework annotation" approaches. Community databases such as Saccharomyces genome database (SGD) (Cherry *et al.*, 1998) for *Saccharomyces cerevisiae* and FlyBase (Drysdale, 2008) for *Drosophila* rely on manual annotation which is performed using genome browsers such as Apollo (Lewis *et al.*, 2002) and Artemis (Rutherford *et al.*, 2000).

Various annotation tools are developed with an aim to develop accurate gene models. However, most of them achieve up to 80% accuracy at exon level (Reese and Guigó, 2006). For example, Maker annotation pipeline (Cantarel *et al.*, 2008), a standalone pipeline, makes use of available data such as proteins, transcripts and ESTs to model structure of genes of a given genome. In addition, to improve on annotation quality, search for homologs in related species is carried out using algorithms such as Cluster of Orthologous genes (COG) (Tatusov et al., 2000), INPARANOID (O'Brien *et al.*, 2005) and/or OrthoMCL (Li *et al.*, 003). Alternatively, search of shared functional domains/motifs is carried out using databases such as InterProScan (Mulder and Apweiler, 2007) and Pfam (Sonnhammer *et al.*, 1997).

Successful annotation of gene models is preliminary to further analysis such as 3D-modelling that calumniate in defining the functions played by the encoded proteins.

1.5 Protein Structure Prediction

Tertiary structure of a protein is important in predicting its function as it is more conserved compared to its DNA counterpart (Chothia and Lesk, 1986). Thus, protein sequences are gaining greater significance in annotation systems that aim to integrate structural data into the annotation process (Reeves *et al.*, 2009). The structure of a protein could either be resolved experimentally through Mass spectrometry (MS) (Aebersold and Mann, 2003), X-ray crystallography (Smyth and Martin, 2000) or Nuclear Magnetic Resonance spectroscopy (NMR) (Wüthrich, 1990). The protein structure can also be predicted computationally. The latter approach has gained more popularity due to its reduced

cost. Computational structure prediction is classified into two categories, which include homology modelling and *de novo (ab initio)* approach. Homology modelling involves threading of the sequence relying on similarity of a known structure (Law and Sansom, 2004). It entails finding known structures (template) that are related to the sequence to be modelled based on similarity search methods such as PSI-BLAST (Altschul *et al.,* 1997) and profile HMMs (Eddy, 1996) against the protein Data Bank (PDB). The identified templates are aligned with the sequence and evaluation of the generated models is done to select the best structure (Marti-Renom *et al.,* 2003). On the other hand, *ab initio* protein modelling predicts the structure without relying on similarity between the modelled sequence and any of the known structures (Baker and Sali, 2001). This approach generates structural models based on the principle of free energy state. They assume minimum global free energy of the protein at its native state and conducts large-scale search of conformational space for structures in low free energy, which reduces its efficiency (Ginalski, 2006; Zhang, 2008). This approach is thus only preferred in case no template is available for use in homology modelling.

Once the structure of the protein has been determined, its interaction with other proteins and/or with small compounds such as drug molecules can be determined computationally. The relative binding energies can also be determined and compared with those of similar proteins. A number of docking software has been developed to perform this type of analysis.

1.6 Computational Protein-Ligand Docking and Molecular Interactions

Molecular docking refers to the computational prediction of a protein-protein or protein-ligand interaction necessary for a protein to perform its function (Sousa *et al.,* 2006). Computational docking has gained popularity among scientists due to its reduced cost as compared to experimental methods such as X-ray crystallography and NMR (Hernández-santoyo, *et al.* 2013).

Docking requires two components including the ligand and the target protein to generate a complex. Its success is dependent on correct binding site prediction and the target protein flexibility. The procedure lampoons the natural molecular interactions exhibited by a protein upon binding of a ligand assuming lowest energy trajectory (Sousa *et al.,* 2006). Various algorithms and programs such as AutoDock Vina (Trott and Olson, 2010), GLIDE (Halgren *et al.,* 2004), FlexX and DARWIN (Taylor and Burnett, 2000) have been developed for rigid and/or flexible docking. Molecular dynamics simulation is gaining popularity as an evaluation process for docking performance. The simulation experiments generate a root-

mean square deviation (RMSD) value; which is a measure of change in protein structure over time. Successful docking is considered to have an RMSD less than 2A° (Hernández-santoyo *et al.,* 2013). Similar to docking, a variety of molecular dynamic tools are available. Most commonly used include GROMACS (Spoel *et al.*, 2005) which uses either Amber or optimized potential for liquid simulations (OPLS) force field and CHARMM (Brooks *et al.,* 1983) that uses empirical energy for modeling protein motions.

1.6.1 Structure Modeling and Ligand-Binding of Insect Chemosensory Proteins

Despite the heightened research in characterization of chemosensory proteins in insects, identification of their potential ligands and determination of their docking ability has not been well studied (Venthur *et al.,* 2014). This has limited our knowledge on potential targets for control of disease vectors and pests. For instance, expression profiling of OBPs in *G. m. morsitans* identified three proteins that putatively bind to host odors (Liu *et al.,* 2010). Homology modeling of these proteins, virtual screening of their potential ligands and docking analysis may provide further insight on their suitability as targets for improvement of vector control.

CHAPTER 2

2. 0 Research Aims and Rationale

2.1 Theme

The theme of this thesis was to identify, annotate and compare putative olfactory responsive gene loci in five tsetse fly species. The five *Glossina* species were selected to represent the three sub-groups of the insect. The selected species included representatives of the Savannah sub-group (*G. austeni, G. m. morsitans* and *G. pallidipes*), forest sub-group species (*G. brevipalpis*), and riverine sub-group (*G. f. fuscipes*). These species are of economic importance as they are among the major vectors of African trypanosomes. For instance, *G. f. fuscipes* is the major vector of sleeping sickness causing pathogens in East Africa while *G. pallidipes* which carries the nagana causing trypanosomes is the most abundant species in the region.

Six classes of olfactory associated genes that play a role in host seeking mate selection and identification of suitable larvipositing sites in the tsetse flies were identified and annotated. The identified genes were compared to those reported in the recently fully sequenced genome of *G. m. morsitans* (Liu *et al.*, 2010, 2012; Obiero *et al.*, 2014) and to closely related Dipterans (*D. melanogaster* and *An. gambiae*). Transcriptomic abundance of genes reported in *G. m. morsitans* was determined in non-olfactory tissues in order to examine their presumed involvement in non-olfactory functions. In addition, *Glossina* homologs of Obp83a1 that is implicated in host seeking (Liu *et al.*, 2010) were modeled and their binding properties to known tsetse baits/repellents evaluated.

2.2 Study Objectives

2.2.1 Aim

To characterize olfactory responsive genes that encodes the major chemosensory proteins (CSPs, GRs, IRs, OBPs, ORs and SNMPs) in five *Glossina* species.

2.2.2 Specific Objectives

(i) To annotate and determine evolutionary relationships of putative olfactory responsive genes in the genomes of *G. austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes*.

(ii) To determine the expression abundance of annotated chemosensory genes in selected non-olfactory tissues of *G. m. morsitans*.

(iii) To determine molecular docking properties of olfactory-specific odorant-binding protein, Obp83a in *G. austeni, G. brevipalpis, G. f. fuscipes, G. m. morsitans* and *G. pallidipes*

2.3 Research Rationale

Eradication of African trypanosomiasis in sub-Saharan Africa is hindered by a lack of effective vaccine and treatment (Welburn, Maudlin and Simarro, 2009). The WHO estimates that about 40% of African population, which translates to 37 sub-Saharan African countries, is at risk of contracting Human African Trypanosomiasis (HAT or sleeping sickness) (Kioy, *et al.,* 2004) caused by *Trypanasoma brucei*. A major outbreak of HAT reported between 1999 and 2003 in Western Kenya suggested that the country is at the risk of increased HAT infection due to human immigration, increased human population and deforestation (Bossche *et al.,* 2010). On the other hand, African Animal Trypanosomiasis (AAT or Nagana), which affects domestic animals, has an estimated global economic cost of US$ 4.75 billion per annum (FAO, 2013). The increased risk of HAT and economic loss due to AAT necessitates search for novel or improved vector (tsetse fly) control programs to suppress disease transmission.

Olfaction plays a crucial role in survival of organisms. For instance, disease vectors including tsetse rely on their olfactory organs to locate their specific hosts, food, mates, breeding sites and predators (Fuss and Ray, 2009). Various *in vitro* studies have reported expression of olfaction responsive genes on the maxillary palpi and antennae of the insect (de Bruyne and Baker, 2008; Fuss and Ray, 2009). Recently, vector research paradigm has shifted towards whole genome sequencing and scientists are exploiting olfactory related knowledge to design repellents for livestock and human protection (Aksoy, 2003, 2010). Consequently, efforts to understand the insect olfactory system have been initiated including analysis of complete genome sequences. For instance, preliminary annotation of OBPs in *G. m. morsitans* genome revealed at least 20 differentially expressed OBPs *(Liu et al.,* 2010), suggesting a feasibility in deciphering molecular processes that mediate host finding in tsetse. Despite the efforts by scientists to control tsetse populations, a universal bait to target all species has not been found to date, mainly due to differential responses exhibited by tsetse species towards host odors. This necessitated the study of other economically important tsetse

species such as, *G. austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes* in order to establish specific molecular characteristics that could be responsible for differential responses exhibited by tsetse species.

2.4 Study Hypothesis and Research Strategy

This study hypothesized that differential host seeking behaviors exhibited by tsetse species reside in the properties of the proteins encoded by their chemosensory genes. To test this hypothesis, genome-wide structural and functional annotation of the major chemosensory proteins was carried out on four tsetse species (*G. austeni, G. brevipalpis. G. f. fuscipes* and *G. pallidipes*) and compared with those of *G. m. morsitans* and closely related dipterans. To further test their functionality, the binding properties of an olfactory specific protein, Obp83a1, earlier named as OBP8 in *G. m. morsitans* (Liu *et al.*, 2010) were compared among the five *Glossina* species. In addition, the potential non-olfactory roles of identified proteins were tested by determining their expression profiles in non-olfactory tissues. The expression profiling was only determined in *G. m. morsitans* due to availability of transcriptomes and because related studies had been carried out in the olfactory organs of this *Glossina* species (Liu *et al.*, 2010, 2012).

2.5 Study Outcome

The data generated in this study provide a basis for undertaking functional studies to unravel tsetse molecular olfactory processes at species level. These data also contribute to scientific knowledge on tsetse biology and its evolutionary relatedness with other dipteran insects. The chemosensory proteins identified in this study potentially play important pleitropic function and could act as molecular targets for development of new control strategies for vector control.

CHAPTER 3

3. 0 Annotation and Comparative Analysis of Chemosensory Gene Families in *Glossina*

3.1 Introduction

Tsetse flies (*Glossina spp.)* are the sole cyclical vectors of African trypanosomes that cause devastating human African trypanosomiasis (HAT, sleeping sickness) and animal African trypanosomiasis (AAT, nagana) across sub-Saharan Africa (Aksoy, 2003). It is estimated that approximately 70 million people and 50 million cattle inhabiting tsetse fly infested areas are at risk of contracting trypanosomiasis (Simarro *et al.* 2011; FAO 2014), and that nagana accounts for up to $ 4.75 billion annual losses (FAO 2013). Currently, there are no prophylactic drugs or vaccines against HAT. Moreover, the available chemotherapeutic remedies are not ideal due to their toxicity, difficulty in administration and growing resistance (Brun *et al.* 2010; FAO 2013).

Sustainable control of African trypanosomiases requires a vector control component (Hocking *et al.*, 1963). Vector control efforts (to suppress tsetse populations) have included trapping using baited traps and targets (Dransfield *et al.*, 1990). The baits comprise of phenolic components present in animal urine and/or breath, 1-octen-3-ol, CO_2 and acetone among other chemical blends that mimic host odors (Hall *et al.*, 1984a). In addition, chemicals such as guaiacol (methyl phenols), d-octalactone and methyl ketones *(Gikonyo et al.* 2002; 2003) have been used as repellents to protect animals from tsetse bites. Differences in response to available baits have been reported among tsetse species and/or between males and female flies (Gikonyo *et al.*, 2003; Mireji *et al.*, 2003). Particularly, *G. f. fuscipes*, a palpalis/riverine species and a major vector of HAT, does not respond to any known attractants but is thought to respond to kairomones released by monitor lizards (Omolo *et al.*, 2009). This differentiation of responses to odors is shown by the varied host preference in the different *Glossina* sub-groups (Späth, 2000; Muturi *et al.*, 2011). Recognition of odor else known as chemoreception is crucial to disease transmission by insect vectors.

Chemoreception in tsetse and other insects is mediated by a group of peri-receptor and surface proteins/receptors encoded by different gene families (Vieira, Sánchez-Gracia and Rozas, 2007). The chemosensory proteins include the odorant binding proteins (OBPs), chemosensory-specific proteins (CSPs) and sensory neuron membrane proteins (SNMPs). On

the other hand, ionotropic receptors (IRs), odorant receptors (ORs) and gustatory receptors (GRs) constitute chemoreceptor families involved in olfaction. Genes encoding various chemosensory proteins are expressed at different olfactory receptor neurons (ORNs) located mainly on the surface of antennae and in fewer numbers on the maxillary palpi *(Andersson et al.,* 2013; Mamidala *et al., 2013*).

The OBPs and CSPs are small soluble proteins present in sensillum lymph of insect where they recognize and solubilize hydrophobic odor molecules, thus shuttling them to the dendritic membrane where they bind to cognate receptors (Leal, 2011; Kulmuni and Havukainen, 2013). The two protein families are characterized by the presence of a signal peptide and α-helices joined by disulphide bonds (Ozaki *et al.,* 2008). OBPs, are ~150 aa long, and are characterized by presence of six conserved cysteine residues. The CSPs are slightly shorter (~ 130 aa) and have four conserved cysteines (Leal, 2011). However, unlike OBPs, CSPs are implicated in non-olfactory functions such as leg regeneration and larval development in other insects such as *Drosophila* (Mameli *et al.,* 1996). Expression of OBPs and CSPs has been linked to host seeking by adult female in *G. m. morsitans* (Liu *et al.* 2010a; 2012).

A third class of proteins that play a role in olfaction is the SNMPs whose domain (CD36) is homologous to that of human scavenger proteins that participates in lipid uptake (Ronderos and Smith, 2009). An earlier study by Jin and colleagues (2008) demonstrated involvement of SNMP1 in chemoreception as a requirement for pheromone detection by *Drosophila* (Jin *et al.,* 2008)

Evolution of ORs has been linked to the ability of insects to distinguish odors under terrestrial conditions given that their aquatic crustacean ancestors do not harbor any ORs (Robertson *et al.,* 2003; Krang *et al.,* 2012). The insect ORs are highly diverse and are characterized by a reversed N-terminal topology and presence of a seven trans-membrane domain (Benton et al., 2006). Specific ORs combine with non-conventional OR; Orco (Or83b), to form functional ion channels that confer specificity to a variety of semiochemicals (Benton *et al.,* 2006; Hallem *et al.,* 2006). Fewer and diverse ORs were identified in *G. m. morsitans* relative to *D. melanogaster* genome, but with an expansion of cis-vaccenyl acetate (cVA) receptor OR67d; a gene that plays a critical role in recognition of the male pheromone (Obiero *et al.,* 2014). On the other hand,insect GRs are responsible for distinguishing between odor tastes and contact pheromones (Vieira *et al.,* 2007; Montell, 2009). Fewer GRs were also identified in tsetse than in *D. melanogaster* and other Diptera

(Krang *et al.*, 2012). Notably, no GRs for sugar were identified in *G. m. morsitans* (Obiero *et al.*, 2014).

The IRs, like ORs, function in complexes formed by up to three subunits and one or two of co-receptors (Ir25a and Ir8a) (Benton *et al.*, 2009; Abuin *et al.*, 2011). However, unlike ORs, IRs are expressed by coeloconic olfactory neurons (Benton *et al.*, 2009), and show responses to a variety of odors including acids, aldehydes, amines and humidity (Yao *et al.*, 2005). Between two and three heterodimers in IRs, similar to those observed in ORs, are required to form functional complexes involved in distinct odor perception (Benton *et al.*, 2009; Rytz, Croset and Benton, 2013). Antennal IRs are not similar to ionotropic glutamate receptors (iGluRs), but are thought to have higher specificity to volatiles than ORs (Benton *et al.*, 2009). Characterization of IRs has not been reported among *Glossina* species to date.

Overall, the chemosensory genes of insects are divergent and evolve through duplication, pseudogenisation and/or deletion incidences (Niimura and Nei, 2006). Functional olfactory genes have been reported to be under natural selection in humans (Voight *et al.*, 2006) and in insects such as *Drosophila* (Gardiner *et al.*, 2008). Positive selection confers fitness advantage to a given species relative to the rest of the population and/or increases its genetic diversity. On the other hand, negative (purifying) selection is known to remove deleterious alleles (Delport *et al.*, 2010).

Recent characterization of major chemosensory protein gene families (OBPs and CSPs) (Liu *et al.*, 2010, 2012; Attardo *et al.*, 2014; Obiero *et al.*, 2014) and identification of genes encoding GRs and ORs in *G. m. morsitans* (Obiero *et al.*, 2014) provided a platform for comparative genomics among tsetse species. This study hypothesized that differential host seeking behavior observed among tsetse species is dictated by diversity in their chemosensory genes. This hypothesis was tested by annotation, comparative phylogenetic analyses and evaluation of the signatures of selection pressures acting on six chemosensory gene families (CSPs, GRs, IRs and iGluRs, OBPs, ORs and SNMPs) from five *Glossina* species. The choice of insects used in comparative analysis was informed by their evolutionary grouping in the tree of life (Wiegmann *et al.*, 2011). Results obtained from this study provide a baseline for undertaking functional studies to enhance the understanding of tsetse speciation and differential host-selection.

3.2 Materials and Methods

3.2.1 Genome and Transcriptome Sequences

Complete genome sequences of *G. austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes* were retrieved from the VectorBase database (Release VB-2014-12) *(Lawson et al.,* 2009). Paired end RNA-Seq reads sequenced on Illumina HiSeq platform from different tsetse fly tissues (whole body, heads, female reproductive organs, guts and salivary glands) (Table S2) were obtained from Aksoy's lab, Yale School of Public Health.

3.2.2 *Ab initio* Gene Model Prediction

To assess the quality of the transcriptomes, the RNA-Seq reads were analyzed using FASTQC (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were then mapped onto their respective reference genomes using Bowtie2 aligner (Langmead, 2010). The mapped reads were then assembled using the Cufflinks suits of programs (Trapnell *et al.,* 2010) to yield a set of transcripts for use in gene model prediction. Automatic gene model prediction was performed using Maker v2.28 (Holt and Yandell, 2011). Within Maker, the *ab initio* gene predictor, Semi-HMM-based Nucleic Acid Parser (SNAP) (Korf, 2004), was trained for *Glossina spp.* starting with a *Drosophila* hidden Markov model (HMM) as the training seed. All *Glossina* expressed sequence tags (ESTs) available in GenBank at the time of this study were downloaded. They included: *G. pallidipes* (n=1127), *G. brevipalpis* (n=407), and *G. f. fuscipes* (n=2). To provide additional evidence for the genes modeled by Maker, a total of 945,752 insecta proteins available in UniProt database were used. Modeled protein sequences were subjected to domain/motif searches against InterProScan v5 (Quevillon *et al.,* 2005) database and the results exported into Blast2Go v 2.8 (Conesa *et al.,* 2005) for Gene Ontology (GO) mapping. Putative chemosensory related genes were selected based on their GO annotation. Bam alignment files generated from the RNA-Seq reads mapping were used to provide further evidence for the intron-exon junctions. In addition, the annotated proteins were probed for definitive domains including OS-D like domain for CSPs, PBP/GOBP domain in OBPs, 7tm-6 for ORs, 7tm-7 for GRs and Lig-Chan, ANF, and NMDA domains for IRs using Delta Blast (Boratyn *et al.,* 2012) against the Conserved Domain Database (CDD) at NCBI.

In order to validate the predicted gene models a comparison was done to those available in VectorBase (Release VB-2014-12) (Lawson *et al.,* 2009). The *ab initio*

annotation was not done for *G. austeni* due to lack of transcriptome data at the time of this study. VectorBase models were adopted for further analysis for consistency in naming. Complete proteomes and gene loci feature files for the four newly sequence tsetse species; *G. austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes* were retrieved from VectorBase database (Release VB-2014-12).

3.2.3 Identification and Annotation of Chemosensory Genes

To identify amino acid sequences for homologs annotated in *G. m. morsitans (Liu et al.*, 2010, 2012; Obiero *et al.*, 2014) and/or those of *D. melanogaster* (Vieira *et al.*, 2007), BLASTp searches were conducted with an e-value cutoff of $\leq 1.0e^{-5}$. Presence of definitive domain(s) expected in each gene family including chemosensory specific OS-D like domain for CSPs, PBP/GOBP domain for OBPs, 7tm-6 for ORs, 7tm-7 for GRs and Lig-Chan, ANF, and NMDA domains for IRs was confirmed through domain searches using Delta Blast (Boratyn *et al.*, 2012) against the CDD (Marchler-Bauer *et al.*, 2005). Gene loci that showed incomplete domains and/or having incomplete sequences were curated using Artemis genome viewer (Rutherford *et al.*, 2000) where possible. Flanking regions of the gene loci (in respective scaffolds) were interrogated for Open Reading Frames (ORF) using NCBI's ORF-Finder (Rombel *et al.*, 2002). Results of ORF-Finder were used to manually curate the gene models observing rules of intron-exon junction and the subsequent sequences re-blasted against NCBI's non-redundant database to confirm homology before inclusion into the genes list. Genes with incomplete or no conserved functional domains were considered as putative pseudogenes.

3.2.4 Comparative Phylogenetic Analysis

For easier comparison, the identified *Glossina* genes were renamed after their closest *Drosophila* homologs. Among them, *Glossina* OBPs with no homologs in *D. melanogaster* retained their names assigned by Liu and colleagues (Liu *et al.*, 2010) for *G. m. morsitans*. Chemosensory gene sequences from *D. melanogaster, An. gambiae*, and *M. domestica* were sourced from Flybase (Gelbart *et al.*, 1996), UniProt (Apweiler *et al.*, 2004), and Scott *et al.*, (2014) (through Hugh Robertson, University of Illinois), respectively. The OBP sequences for *C. capitata* were obtained from GenBank (Benson *et al.*, 2011) using the published Accession numbers (Fly *et al.*, 2014). Among the selected relatives, both *D. melanogaster*

and *An. gambiae* were considered out-group species based on their relationship to other species in the tree of life (Wiegmann *et al.*, 2011). To compare the sequences, multiple sequence alignments for each class of genes were generated using MUltiple Sequence Comparison by Log-Expectation (MUSCLE v3.6) (Edgar *et al.*, 2004) using default settings. The resulting alignments were manually edited using standalone Jalview v2 (Waterhouse *et al.*, 2009) (See supplementary figures 3 1.1-3, 3.4.1-3), then converted into Phylip format using ClustalX v2. 1. The best substitution model for the alignment was determined using ProtTest server v3.2.1 (Abascal *et al.*, 2005). Phylogeny inference for the aligned sequences were deduced using a Maximum-likelihood approach as implemented in RAxML v 8 (Stamatakis, 2014) with 1000 bootstrap iterations. The obtained phylogenetic trees were viewed and rendered using FigTree viewer (Abascal *et al.*, 2005).

3.2.5 Selection Analysis

To assess the influence of natural selection in evolution on the identified genes, codon alignment of *Glossina* orthologs was done using Prank (Löytynoja, 2014) and their corresponding phylogenetic trees constructed using RAxML v 8.2.0 (Stamatakis, 2014). Signatures of natural selection on the orthologs were evaluated by calculating ratios of nonsynonymous to synonymous substitutions (d_N/d_S) in codeml in PAML v 4 (Yang, 2007). Three site models including M1a (Nearly neutral), M2a (Positive Selection) and M8 (beta & w) were evaluated against their null models to test for selection using log-likelihood ratio (LRT). Corresponding p-value was calculated to test for significance of selection. A threshold of <= 0.05 was used to consider a gene to be under significant positive selection. Similarly, selection analysis was carried out using the HyPhy package (Pond *et al.*, 2005) hosted on Datamonkey web server (Delport *et al.*, 2010). In this case, neighbor joining trees were constructed within the package and an appropriate model of nucleotide evolution was determined for each alignment, prior to analysis. Two algorithms; Mixed effects model of Evolution (MEME) (Murrell *et al.*, 2012) and PARRIS (Scheffler *et al.*, 2006) were used to identify sites under episodic selection taking recombination events into account. A p-value of <=0.05 was implemented to estimate the rate of false positives (type I error) in which neutrally evolving sites may be erroneously reported to be under selection.

3.3 Results

3.3.1 Annotation and genomic arrangement of chemosensory genes across tsetse species

Numbers of the annotated chemosensory genes are summarized in Table 3.1. The tsetse species were found to have fewer chemosensory genes than other insects evaluated in this study. Compared to other dipterans used in this study *Glossina* species have relatively conserved gene copies in all the chemosensory protein classes. The IRs, OBPs and ORs were observed to have a few incomplete sequences and/or missing functional domains and were considered putative pseudogenes. Those proteins found to be missing the functional domains include Obp73a in all tsetse species, Obp56h in *G. austeni*, Obp20, Or85e and Gr33a in *G. brevipalpis*, SNMP1 and Or56a in *G. f. fuscipes*, and one copy of Or67d in *G. pallidipes*. No pseudogenes were found among the CSPs, GRs and SNMPs. The rest of the identified genes were found to have definitive domain signatures (7tm_7 superfamily in GRs, 7tm_6 in ORs, PBP, ANF- receptor and Lig_Chan in IRs). The OBPs and CSPs had PBP-GOBP and OS-D domains respectively, while the SNMPs had CD36 family signatures. In terms of sequence length, the GRs and ORs identified in *G. austeni, G. brevipalpis, G. f. fuscipes,* and *G. pallidipes* were 269-480 aa and 295-508 aa long, respectively. Similarly, CSPs and OBPs were 108-178 aa and 108-257 aa long, respectively. The SNMPs and IRs had longer sequences than other gene families, being 384-540 aa and 407-1070 aa long, respectively (See Table S3.1.1-3.1.6 for more annotation details).

Table 3.1: Summary of putative chemosensory genes annotated in *Glossina* species: *G. austeni, G. brevipalpis, G. f. fuscipes, G. m. morsitans* and *G. pallidipes* against selected dipterans

Species	CSPs±	GRs	IRs/iGluRs	OBPs	IRs	SNMPs	References
G. austeni	5	14	28	29	40 (5)	2	Macharia *et al.*, 2016
G. brevipalpis	4	11	28	28	42 (5)	2	,,
G. f. fuscipes	5	14	31 (2)	30 (3)	42 (6)	2	,,
G. pallidipes	5	14	30 (1)	30 (2)	42 (3)	2	,,
G. m. morsitans	5	14	30 (2)	30 (3)	46 (3)	2	Liu *et al.*, 2010; Liu *et al.*, 2012; Obiero *et al.*, 2014
D. melanogaster	4	60 (13)	66(9)	52	62 (2)	2	Clyne *et al.*, 2000; Robertson *et al.*, 2003; Vieira *et al.*, 2007; Benton

Species	CSPs±	GRs	IRs/iGluRs	OBPs	IRs	SNMPs	References
							et al., 2009
An. gambiae	8	76	48	82	79	2	Fox *et al.*, 2001; Hill *et al.*, 2002, Croset *et al.*, 2010
M. domestica	5	103	110	>87	86	2	Scott *et al.*, 2014

±CSPs – chemosensory specific proteins, GRs – gustatory receptors, IRs/IGluRs- ionotropic receptors/ionotropic glutamate receptors, OBPs- odorant binding proteins, ORs- odorant receptors, SNMPs- sensory neuron membrane proteins.
Number of genes showed in parenthesis represents putative pseudogenes i.e. either incomplete genes or genes missing functional domain.

Generally, the chemosensory genes in the four tsetse species analyzed depicted a sparse distribution across their respective genomes (Tables S3.1.1-3.1.6). This study revealed a general genome-wide dispersion of the chemosensory genes in all the tsetse species analyzed. Fourteen loci were duplicated. The loci included one CSP (Ejbp3; that have two copies namely Ejbp3A and Ejbp3B), three GRs (Gr21a; with three copies namely Gr2a1, Gr2a2 and Gr21a3, Gr28b; two to three copies per genome Gr28bB, Gr28bC, and/or Gr28bD and Gr59f; with two copies). Two OBPs (Obp83a which has four copies; Obp83a1-4, Obp56e; with two copies Obp56e1 and Obp56e2 and eight ORs (Or7a with three copies: Or7a1-3, Or45a with three copies: Or45a1-3, Or67d with five copies: Or67d1-5 and Or56a with two copies: Or56a1 and Or56a2, Or43a, Or46a, Or63a, and Or67c with two copies, each). All the four copies of Obp83a homolog were in tandem across the five tsetse genomes, and represented evidence of structural gene variation and rearrangement (Figure S3.1). One of the Obp83a copies was located on the reverse strand. In contrast, duplicated ORs including three copies of Or45a, two copies of Or7a and four to six copies of Or67d homologs were located in disparate scaffolds.

3.3.2 Comparative Analyses of Putative Chemosensory Genes

3.3.2.1 Chemosensory protein families

Sequence alignment of Obp56i and Obp19 from selected dipterans showed variation of amino acids at the third and fourth conserved cysteine residues (labeled c3 and c4 in Figure S3.2).The sequences of Obp56i and Obp19 showed sequence deletions in the positions 225-233 and 241-245 respectively. In contrast, their homologs from *D. melanogaster* and *M. domestica* showed amino acid conservation around the same regions.

Phylogenetic relationships among the OBPs identified in *Glossina* species against those in *C. capitata, M. domestica* and *D. melanogaster* are shown in Figures 3.1.1 - 3.1.3. About 68.9% (n = 29) of the *Glossina* OBPs were grouped into the Classic subfamily (Hekmat-scafe *et al.*, 2002) (with six conserved cysteines) (Figure 3.1.1) while six OBPs in each of the tsetse species were grouped into the Minus-C subfamily (with less than the conventional six cysteines) (Figure 3.1.2). No Plus-C/Atypical subfamily members were identified in the *Glossina* species (Figure 3.1.3). Expansions of Obp56e (two copies) and Obp83a (four copies) classic subfamily were observed in all tsetse species (Figure 3.1.1), while *M. domestica* and *C. capitata* had three and two copies of gene encoding Obp83a, respectively. The Obp56a, Obp56h, Obp19b, and Obp19d, Obp99c were among OBPs expanded in *M. domestica.*

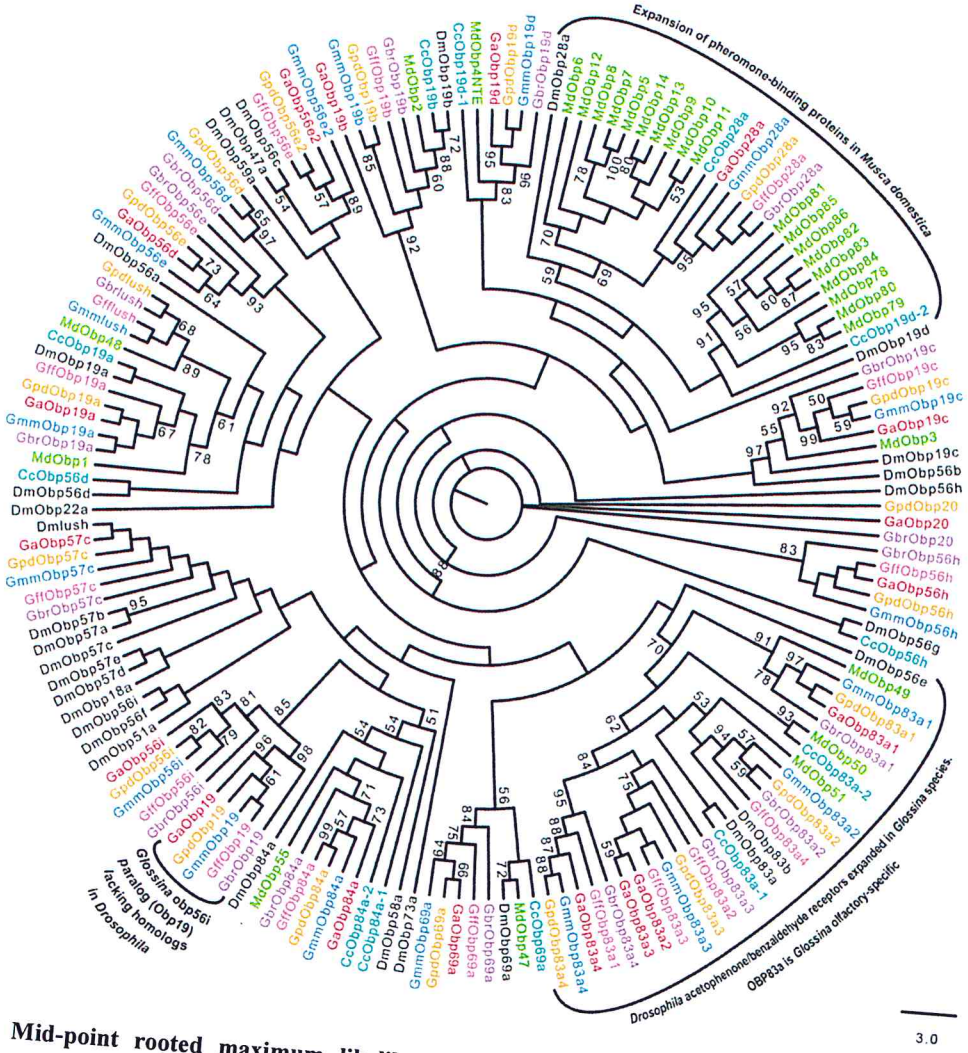


Figure 3.1.1: Mid-point rooted maximum likelihood phylogenetic tree of Classic odorant binding proteins. Insect classic OBPs are characterized by six conserved cysteine residues. Different symbols and

colours depict OBPs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* **(Gpd*),** *Drosophila melanogaster* **(Dm*),** *Ceratitis caiptata* **(Cc*)** and *Musca domestica* **(Md*).** The symbol * represents the name of the specific OBP. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations.

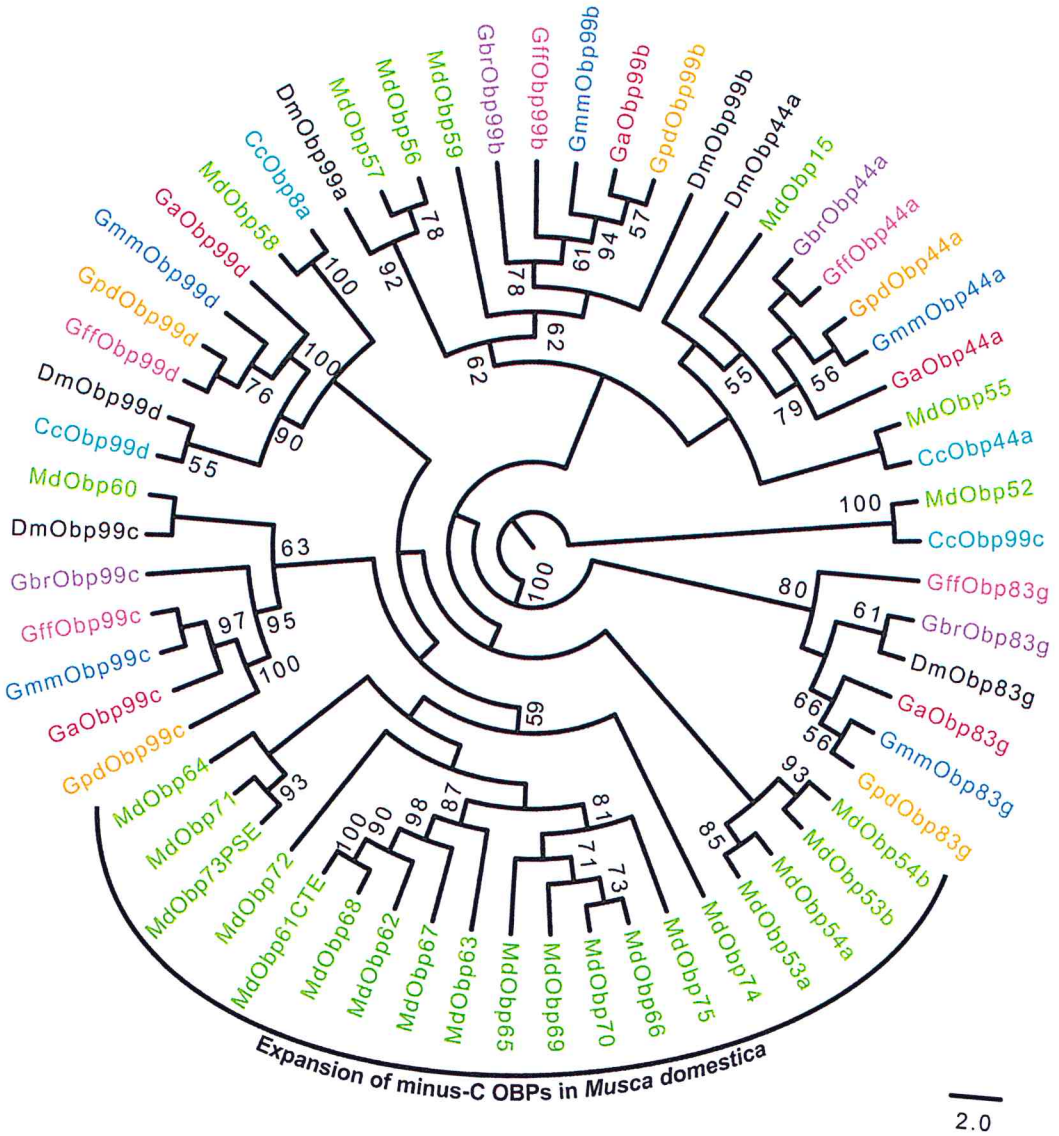


Figure 3.1.2: Mid-point rooted maximum likelihood phylogenetic tree of Minuc-C odorant binding proteins. The minus-C OBPs have less than six conserved cysteine residues (Missing C1or C2 and/or C5). Different symbols and colours depict OBPs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* (Gpd*), *Drosophila melanogaster* **(Dm*),** *Ceratitis capitata* (Cc*) and *Musca domestica* (Md*). The symbol * represents the name of the specific OBP. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations.

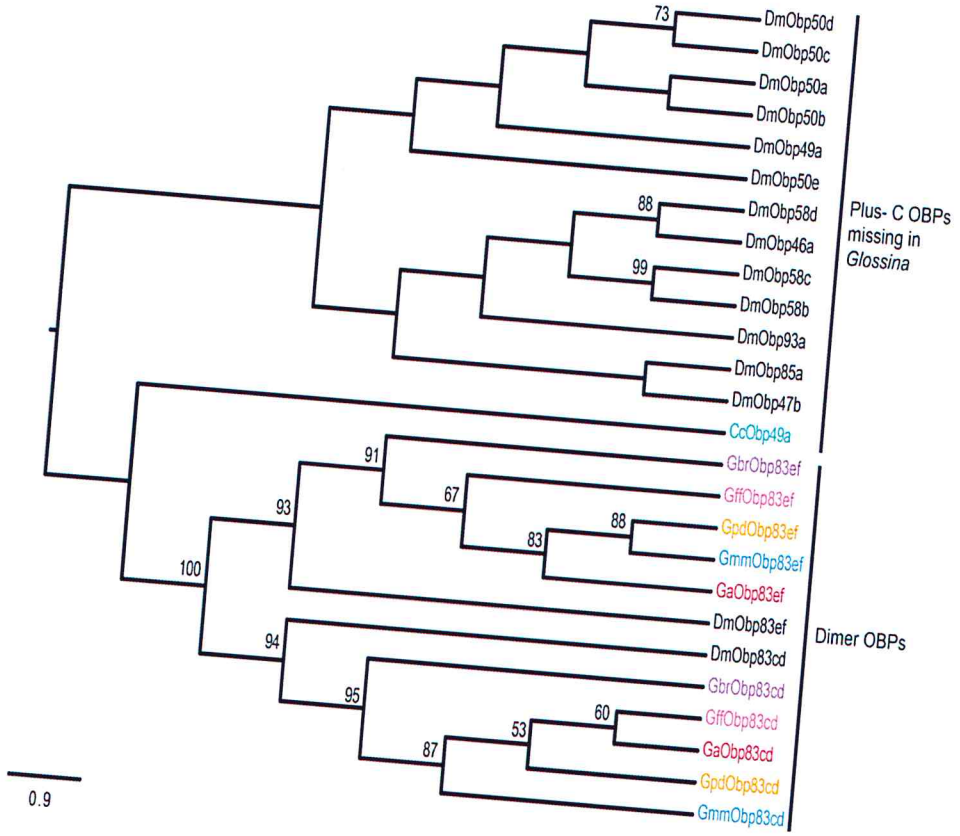


Figure 3.1.3: Mid-point rooted maximum likelihood phylogenetic tree of Plus-C and Classic-Dimer odorant binding proteins. The Plus-C OBPs are characterized by having more than six cysteines and a conserved proline residue. The Classic-dimers have two conserved domains of classic sub-family. Different symbols and colours depict OBPs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* (Gpd*), *Drosophila melanogaster* **(Dm*),** *Ceratitis capitata* **(Cc*).** The symbol * represents the name of the specific OBP. Sequence alignment was performed using MuSCLE v3.8.31 and

phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations.

Phylogenetic analysis of the CSPs revealed four distinct clades (A – D) (Figure 3.2). *M. domestica* and all tsetse species (except *G. brevipalpis*) had two copies of ejaculatory–bulb specific protein (Ejbp3).

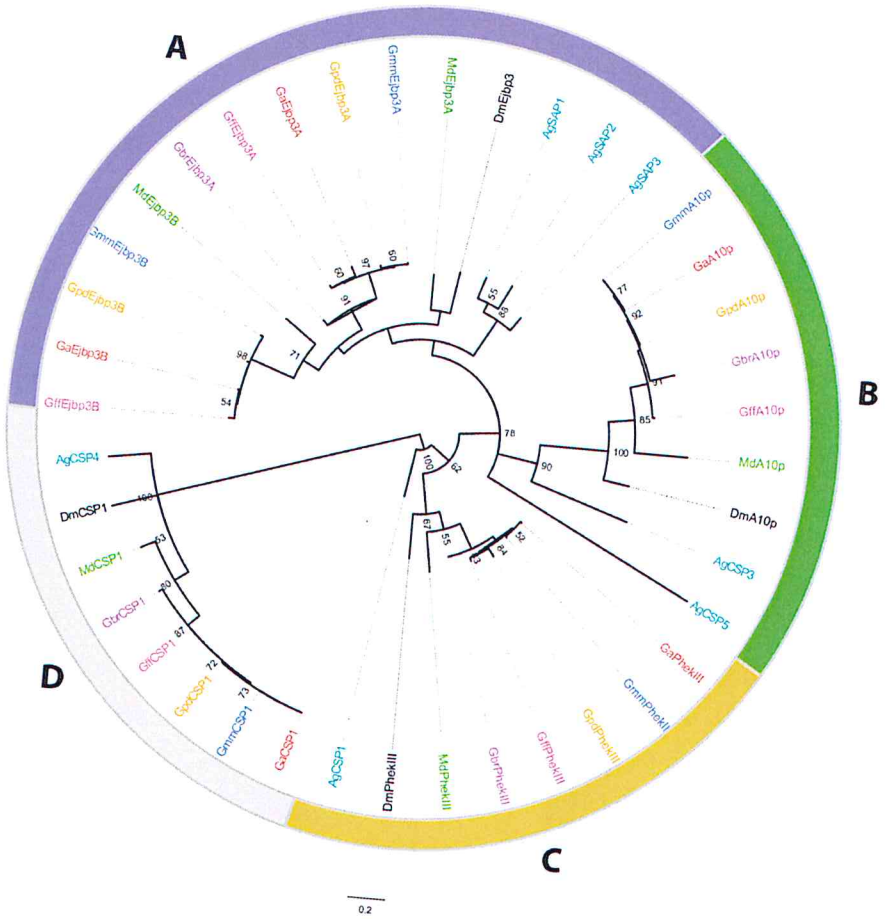


Figure 3.2: Mid-point rooted maximum likelihood phylogenetic tree of chemosensory proteins (CSPs). Clade A shows duplication of ejaculatory bulb protein 3 (Ejbp3 in four tsetse species). Different symbols and colours depict CSPs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* **(Gpd*),** *Drosophila melanogaster* **(Dm*),** *Anopheles gambiae* **(Ag*)** and *Musca domestica* (Md*). The symbol * represents the name of the specific CSP. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations.

The two orthologs of SNMP1 and SNMP2 in *Drosophila* were present in all tsetse species that were included in this study. Two SNMP sub-clades with one-to-one orthology across all insects were identified (Figure 3.3).

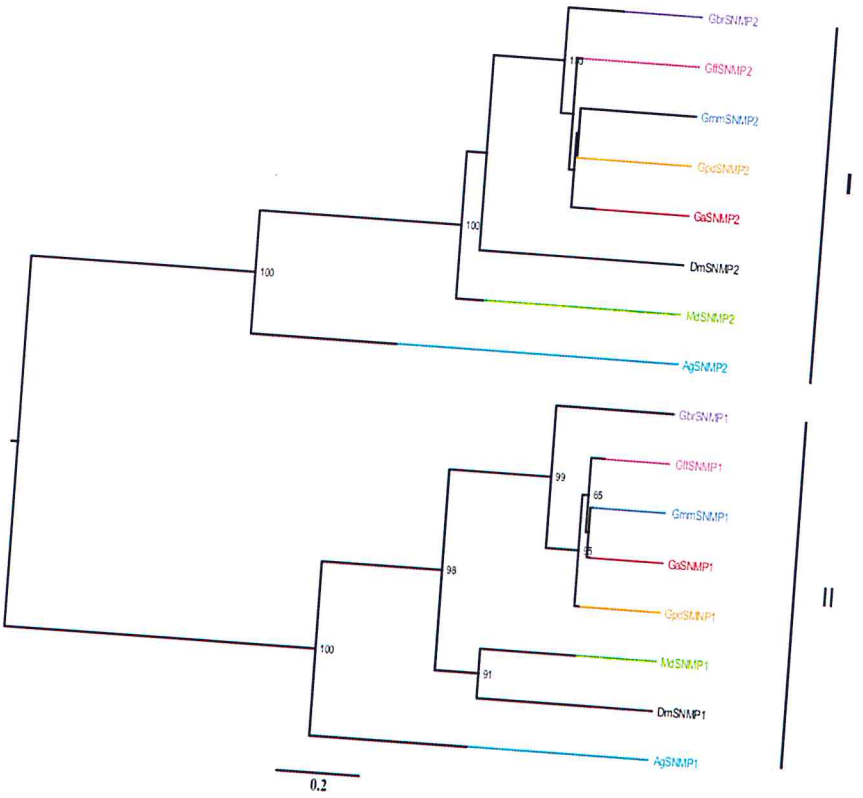


Figure 3.3: Mid-point rooted maximum likelihood phylogenetic tree of sensory neuron membrane proteins (SNMPs). Phylogeny reconstruction yielded two clades I and II each showing one to one orthology of the specific SNMP from different insect species. Different symbols and colours depict SNMPs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* **(Gpd*),** *Drosophila melanogaster* **(Dm*),** *Anopheles gambiae* **(Ag*)** and *Musca domestica* **(Md*).** The symbol * represents the name of the specific SNMP. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations.

3.3.2.2 Chemoreceptor Gene Families

Phylogenetic relationships of GRs identified in *Glossina* genes and their homologs in *An. gambiae, M. domestica* and *D. melanogaster* are shown in Figure 3.4. In all the tsetse species, there was expansion of Gr21a, associated with CO_2 detection (Jones et *al.*, 2007; Kwon *et al.*, 2007). *An. gambiae,* on the other hand has expanded Gr63a, a protein co-

expressed with Gr21a and involved in CO_2 detection (Jones *et al.*, 2007). No homologs to sugar receptors in *D. melanogaster* (Isono *et al.*, 2010) were identified in any of the five *Glossina* species (Figure 3.4). *Drosophila melanogaster* Gr43a, implicated in internal fructose sensing (Miyamoto *et al.*, 2012) was found to be absent in all five tsetse species analyzed in this study.

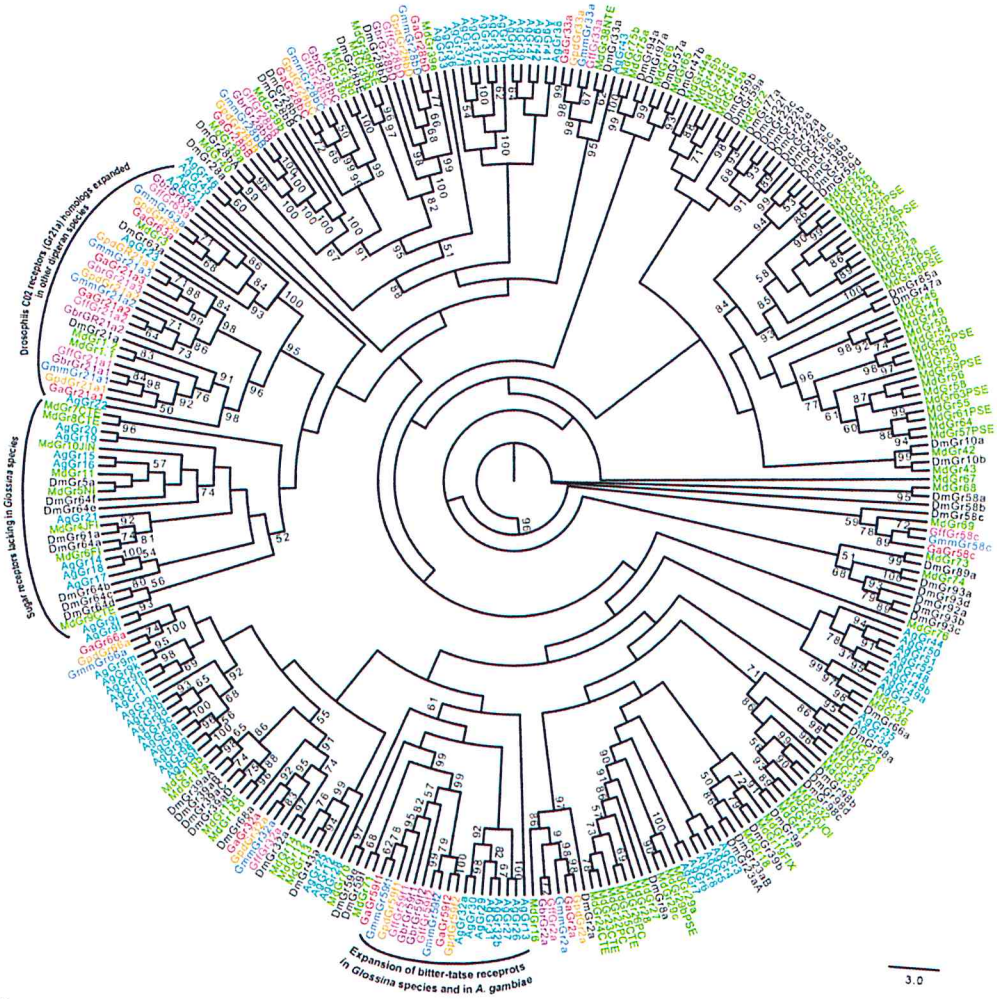


Figure3.4: Phylogenetic tree of gustatory receptors (GRs). The resulting clades were identified with respect to function of the GRs in *Drosophila* Different symbols and colours depict GRs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* (Gpd*), *Drosophila melanogaster* **(Dm*),** *Anopheles gambiae* (Ag*) and *Musca domestica* (Md*). The symbol * represents the name of the specific GR. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations

A single copy of the non conventional co-receptor (Orco) was identified in all five tsetse species (Figure3.5). Up to 75 – 85 % amino acid identity was calculated among Orco in all tsetse and those of its homologs in *M. domestica, D. melanogaster* and *An. gambiae*. Phylogenetic analysis revealed 16 distinct clades among *Glossina* species, *D. melanogaster, M. domestica* and *An. gambiae* ORs (Figure 3.5). Three paralogs of Or45a which is responds to stress in *Drosophila* larvae (Vosshall and Stocker, 2007), were identified in all tsetse species. Expansion and genomic dispersion was also noted in Or7a. Phylogenetic analysis showed clustering of three *M. domestica* ORs with the *Glossina* Or7a homologs (Clade D, Figure 3.5). Clade A containing *Drosophila* Or67d homolog was also expanded in tsetse flies. Four *Glossina* species (*G. austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes*) had a total of five Or67d paralogs compared to six copies reported in *G. m. morsitans* (Obiero *et al.*, 2014). Other genes showing expansion in *Glossina* species include Or67c K) and Or43a (Figure 3.5).

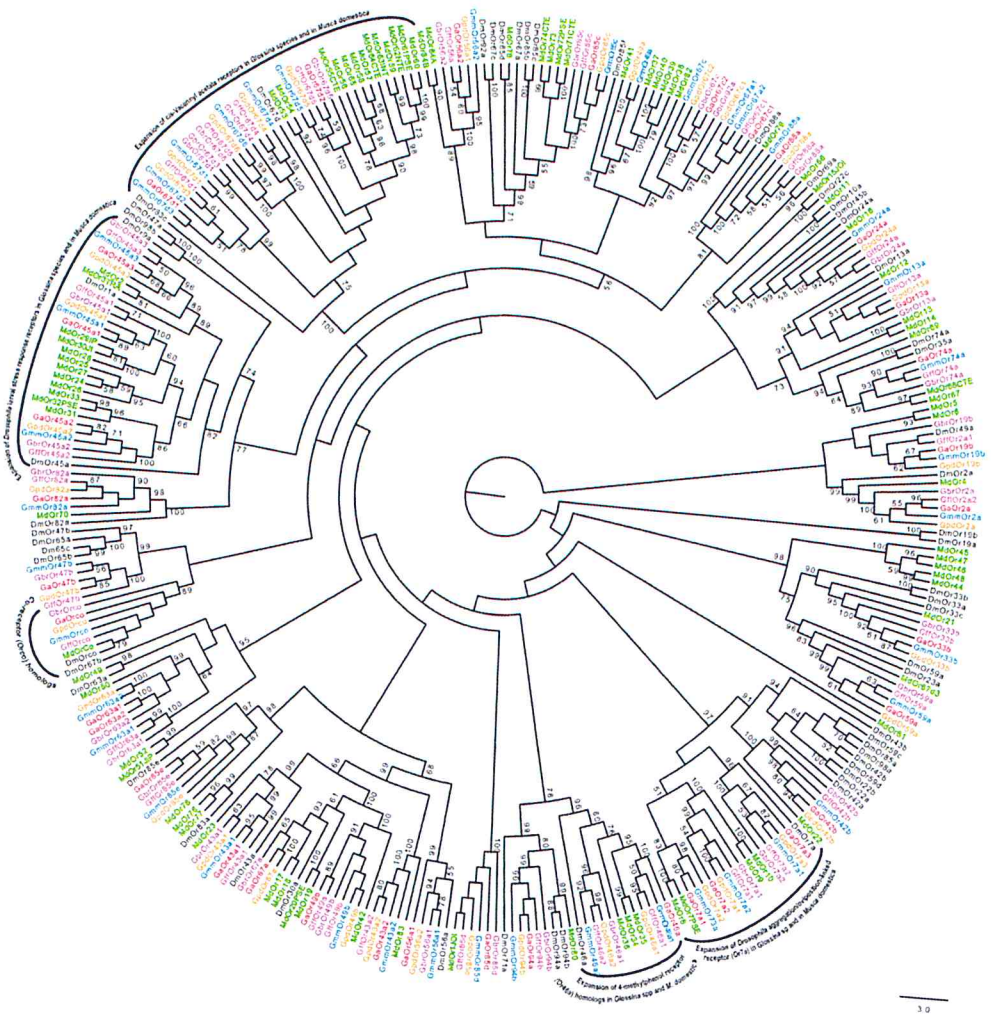


Figure 3.5: Phylogenetic tree of odorant receptors (ORs). Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship was inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations. Clades harboring characterized ORs were labeled based on the gene function Different symbols and colours were used to depict ORs from the different species at the terminal nodes: *Glossina austeni* **(Ga*)**, *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*)**, *Glossina pallidipes* (Gpd*), *Drosophila melanogaster* **(Dm*)** and *Musca domestica* (Md*). The symbol * represents the name of the specific OR.

Similar numbers of IRs/iGluRs were identified in all tsetse five species (Table 3.1). The homolog of a *Drosophila* Ir93a was not found in *G. austeni*. Phylogeny reconstruction of IRs and iGluRs yielded distinct clades with IRs (Figure 3.6.1-2) showing divergence from iGluRs receptors (Figure 3.6.3). A total of 13 *Glossina* IR homologs clustered with their antennal *Drosophila* orthologs (Ir40a, Ir25a, Ir8aa, Ir93a, Ir21a, Ir76a, Ir76b, Ir31a, Ir75c, Ir75a, Ir75d, Ir64a and Ir84a) (Figure 3.6.1). All the five *Glossina* species were seen to have less divergent IRs compared to other diptera (Figure 3.6.2). Further, *Drosophila*-specific antennal Ir84a and was found to have homologs in all five *Glossina* species studied here.

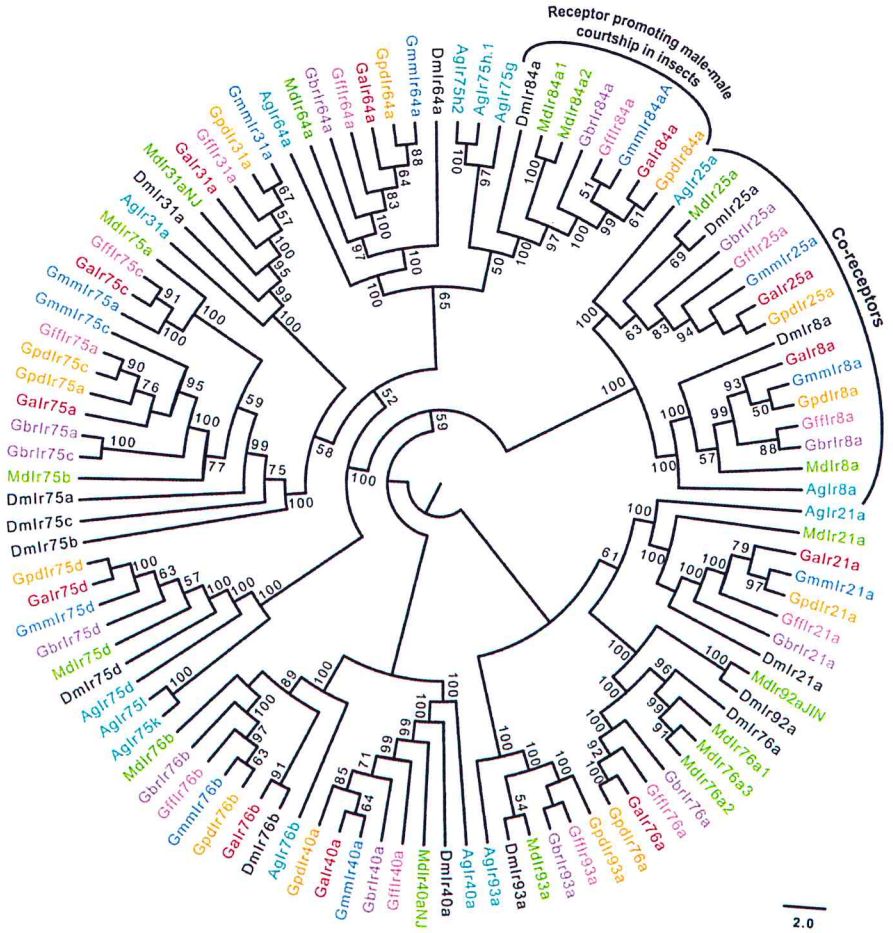


Figure 3.6.1: Mid-point rooted maximum likelihood phylogenetic tree of Antennal ionotropic receptors. Antennal IRs are primarily expressed at the antenna of the insect. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations. Different symbols and colours were used to depict IRs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* **(Gbr*),** *Glossina fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* (Gpd*), *Drosophila melanogaster* (Dm*), *Musca domestica* (Md*) and *Anopheles gambiae* (Ag*). The symbol * represents the name of the specific IR.

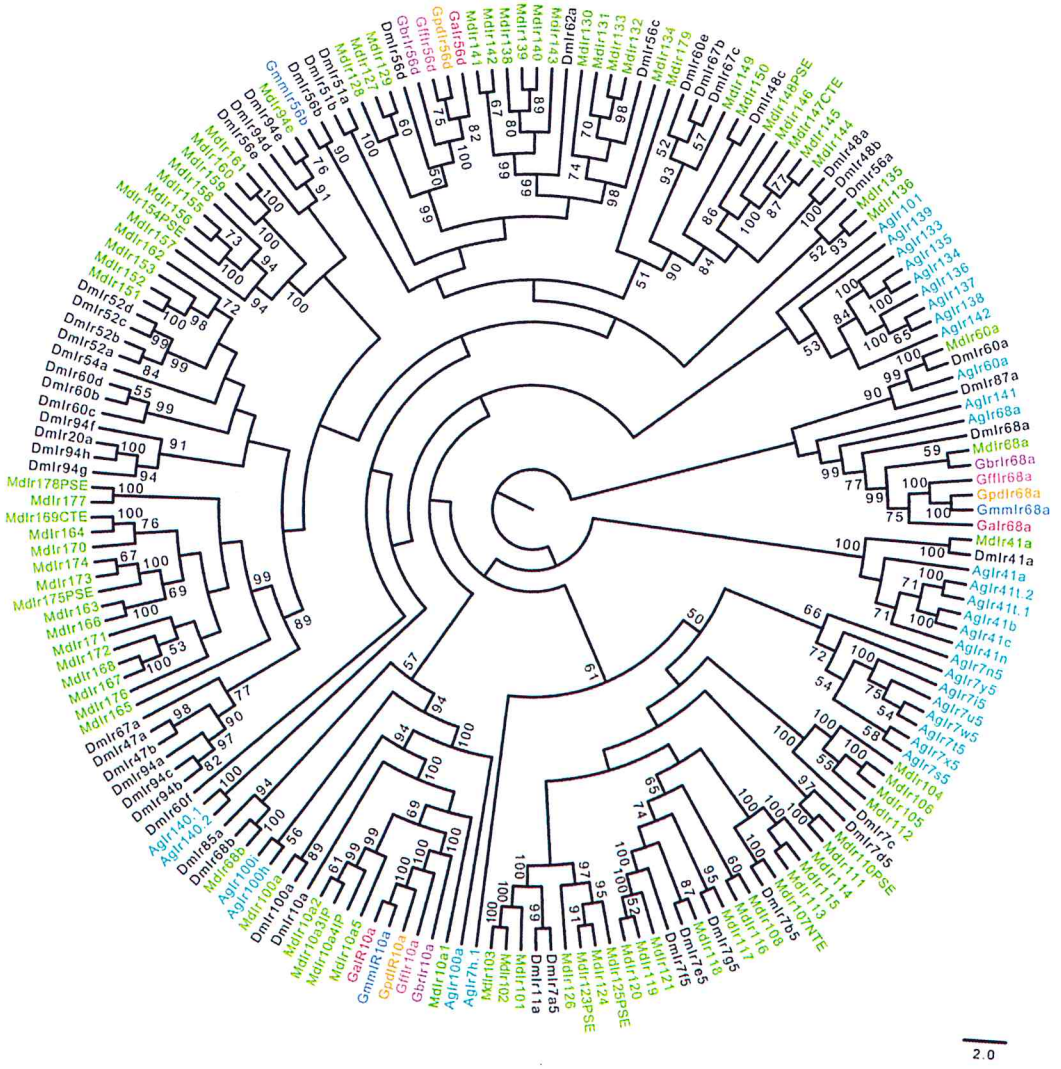


Figure 3.6.2: Mid-point rooted maximum likelihood phylogenetic tree of Divergent ionotropic receptors. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations. Different symbols and colours were used to depict IRs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* **(Gpd*),** *Drosophila melanogaster* **(Dm*)** and *Musca domestica* (Md*). The symbol * represents the name of the specific IR.

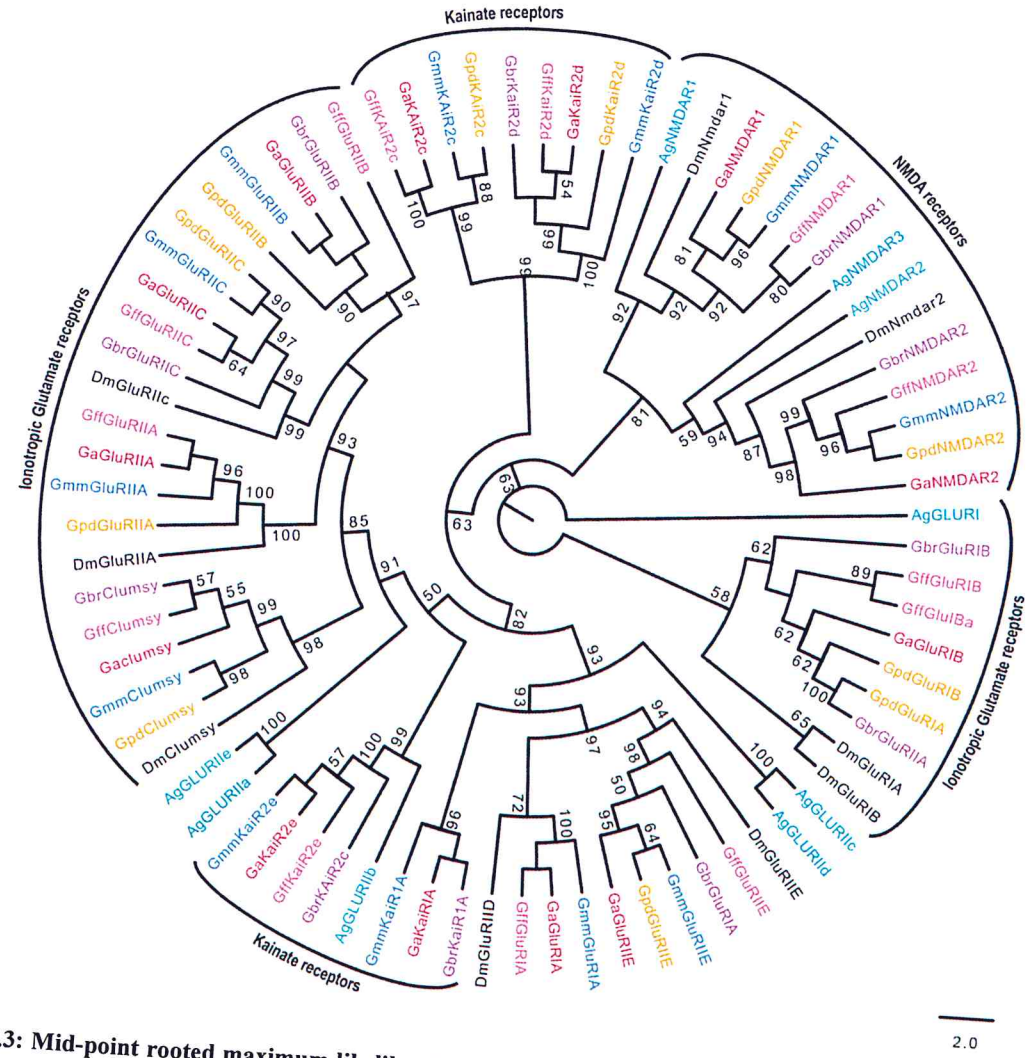


Figure 3.6.3: Mid-point rooted maximum likelihood phylogenetic tree of ionotropic Glutamate receptors. Sequence alignment was performed using MuSCLE v3.8.31 and phylogeny relationship inferred using RAxML v8 with best fitting Wheelan and Goldman (WAG) model and 1000 bootstrap iterations. Different symbols and colours were used to depict IRs from the different species at the terminal nodes: *Glossina austeni* **(Ga*),** *Glossina brevipalpis* (Gbr*), *Glossina fuscipes fuscipes* (Gff*), *Glossina morsitans morsitans* **(Gmm*),** *Glossina pallidipes* **(Gpd*),** *Drosophila melanogaster* (Dm*) and *Anopheles gambiae* **(Ag*).** The symbol * represents the name of the specific IGluR.

3.3.3 Selection analysis

The M8 (beta & w) codeml model was found to represent the data better than M1a and M2a models hence adopted in calculation of p-values. Nevertheless, some of the d_N/d_S (w1M8) values were too high to be considered reliable (Table S3.3). Such values result from low count of synonymous substitutions compared nonsynonymous substitutions. In addition,

majority (67.02%) (n=88) of the alignments were seen to have a significant p-value under the M8 model. Contrary, only a small subset (13.64%) of gene loci was significantly identified to be under selection in the HyPhy package (Table S3.4). (Gr21a, Gr28b, Obp83a and GluRIIA) (Table 3.2). Various factors such as the low number of sequences per gene loci and lack of divergence within sequences have been indicated to introduce false positives (type I error) and lack meaningful inference (Yang *et al.*, 2000; Poon *et al.*, 2009).

Table 3.2: Summary of four Glossina chemosensory gene loci identified to have signatures of positive selection: Selection analysis was performed using HyPhy package using MEME and PARRIS and compared with PAML –codeml using the M8-M8a model.

Gene id	lnL M8	lnL M8a	LRT	p –value	w1M8	Sites by MEME	Singleton (S) /Duplicate (D)	Number of codons analyzed	ΔLRT MEME
Obp83a	-529.17	-533.174	7.943	0.0048	1.075	29	D	498	57.927
Gr21a	-1642.19	-1656.36	28.346	1.014E-7	1.1865	39	D	621	21.87
GluRIIA	-1431.52	-1387.04	.34	0.00387	1.4264	2	S	1807	12.58
Gr28b	-1557.62	-1566.29	7.34	4.85E-5	1.75	44	D	569	6.98

lnL M8 is the likelihood of the experimental model (M8), lnL M8a is the likelihood of the null model (M8a), ΔLRT is the Likelihood Ratio Test = 2*(lnL M8- lnL M8a), w1M8 is the ratio of non-synonymous to synonymous mutations (d_N/d_S) predicted under M8 model/and p-value is the statistical measure of significance.

3.4 Discussion

In addition to chemosensory genes identified in *G. m. morsitans* (Liu et al., 2010, 2012; Obiero et al., 2014), annotation of these gene families in four additional tsetse genomes (*G. austeni, G. brevipalpis, G. f. fuscipes* and *G. pallidipes*) (Macharia *et al.*, 2016) has provided a comprehensive gene repertoire necessary for undertaking functional studies. These studies will enhance understanding of tsetse biology and facilitate selection of suitable novel molecules that could be used as targets for vector control. Overall, the five tsetse species compared depict a general gene conservation in terms of sequence length, structures, and copy numbers of the chemosensory genes across. Specifically, high conservation was

observed in OBPs and CSPs, which aid in trafficking of hydrophobic molecules across the sensillum lymph of insects (Dyer *et al.*, 2008). The two protein families are characterized by six and four conserved cysteine residues, respectively, with the latter being more conserved (Angeli *et al.*, 1999). This observation supports earlier observations by Sanchez-Gracia *et al.*, (2009) that the CSPs family is more conserved compared to the OBPs family (Vieira and Rozas, 2009). Majority (n=29) of OBPs identified across *Glossina* fall under the Classic subfamily, which is consistent with what has been reported in genomes of related insect species such as *Drosophila* and the Mediterranean fly. This result suggests that classic OBPs have conserved function across all insects as compared to other classes of OBPs. Expansion of Obp83a (previously named Obp8-10,12 in *G. m. morsitans* (Liu *et al.*, 2010)) was noted in all tsetse species. Liu and colleagues (2010) suggested that Obp83a could be olfactory-specific as it is expressed highly in starved females. Its expansion across all tsetse species studied here supports their argument for its probable participation in host seeking. Co-localization of the four copies within same scaffold suggests that they are recent paralogs that could be co-regulated. On the other hand, the presence of two *Glossina* odorant receptor paralogs (copies of Or45a and Or7a), in distantly located scaffolds, suggests a role of transposition in their emergence. Gene transposition has earlier been reported in three *Drosophila* species (*D. melanogaster, D. yakuba* and *D. simulans*) (Ponting *et al.*, 2001; Inohara and Nuez, 2002; Heger and Ponting, 2007). Therefore, findings of this study appear to support transposition as a valid mode of gene emergence in insects.

The complete loss of genes and/or distortion in their gene structure observed in *G. brevipalpis* could be attributed to evolutionary events given that it is the most ancient among the *Glossina* species studied. This confirms an assumption made by Gooding and colleagues (2005) who proposed that the oldest subgenus would exhibit more genetic differences if all tsetse species assumed a constant rate of evolution. Among the GRs, Gr32a and not Gr68a (pheromone receptors), were found present in all five *Glossina* species. Both Gr32a and Gr68a respond to pheromones in insects *(Isono et al.*, 2010). Additionally, Gr68a participates in sound reception (Isono et al., 2010). Absence of Gr68a in tsetse could imply that tsetse flies rely on a different receptor other than Gr68a for sound reception, or that the insects rely entirely on their tympanal organ for this function (Tuck, Windmill and Robert, 2009). Further, absence of Gr68a has been reported to reduce male-male courtship in *Drosophila* and perhaps may play the same role in tsetse flies (Montell, 2009). *Glossina* IRS shows conservation of copy numbers. Notably, the Ir84a have homologs in all tsetse species studied here. Ir84a is a candidate for phenyacetyaldehyde reception (Grosjean *et al.*, 2011) and has

been reported to promote male courtship (Grosjean *et al.*, 2011) in *Drosophila*. Presence of Ir84a in *Glossina* supports that male courtship is a conserved feature across tsetse species. On the other hand, the absence of Ir93a in *G. austeni* whose ligand is unknown (Silbering *et al.*, 2011) could potentially encode a defective response to either aldehydes, amines or carboxylic acids which are primarily recognized by IRs (Rytz *et al.*, 2013). Based on the number of chemosensory genes identified across *Glossina*, it is apparent that all tsetse fly species have a reduced chemosensory repertoire relative to *D. melanogaster* and *M. domestica*. This confirms findings reported in *G. m morsitans* (Attardo *et al.*, 2014; Obiero *et al.*, 2014). Noteworthy is the absence of all sugar receptors (Gr64a-f and Gr5a) in all tsetse species studied here. Presumably, this is due to the hematophagous nature of tsetse flies. This observation is contrasted with the conservation of sugar receptors in *M. domestica, D. melanogaster* and *An. gambiae*, which feed on nectar as primary or secondary source of nutrients. Also, tsetse species lack homologs to Gr43a, which has been attributed to internal fructose sensing in *Drosophila* (Miyamoto *et al.*, 2012). The Gr43a mutants shows an abolished preference of fructose but no difference in response to other sugars (Mishra *et al.*, 2013). In *An. gambiae*, Gr43a is expanded for reasons that are not clearly understood but could serve a similar function as in *Drosophila*. All tsetse species showed expansion of Gr21a homologs that mediates CO_2 recognition confirming that tsetse flies are attracted to their vertebrate hosts through this volatile gas (Torr *et al.*, 2006). Similar to *M. domestica* (Scott *et al.*, 2014), expansions of Or45a and Or67d that mediate stress response (Bellmann et al., 2010) and cVA reception (Wang and Anderson, 2010), respectively, in *Drosophila*, were noted in all tsetse species. Or45a in *Drosophila* is expressed only in larvae (Vosshall and Stocker, 2007) where it serves as a receptor for octyl acetate that trigger a repellency effect (Bellmann *et al.*, 2010). Though the significance of its expansion in tsetse is yet to be understood, it probably plays a role in recognizing any undesirable cues present in tsetse's uterus to be determined and Or43a, linked to benzaldehyde perception in *Drosophila* (Rollmann *et al.*, 2010).

Among the annotated *Glossina* OBPs, Obp19 (a gene that lacks homologs in *Drosophila*) was seen to have homologs in hemipterans, *Lygus lineoralis* and *Microplitis demolitor* and not in any of the close dipterans such as *M. domestica* or *Stomoxys calcitrans*. Moreover, it showed close phylogenetic relationship with Obp56i from all the *Glossina* species. This could imply that Obp19 is a recent paralog of Obp56i that assumes similar role to that of its homologs in hemipterans. Close phylogenetic relationship observed among *Glossina* OBPs and genes related to pheromone binding protein receptor proteins (PBPRPs)

from other insects including: Obp19d, Obp28a, Obp69a, Obp83a and Obp84a. This observation is similar to what was reported in *C. capitata* (Siciliano *et al.*, 2014) implying that the role of PBRPs is well conserved across insects.

Three of four gene loci (Table 3.2) showing strongest indication of positive selection are evolving under duplication suggesting rapid rate in their evolution as earlier reported in ants (Kulmuni and Havukainen, 2013) and in *Drosophila* (Almeida *et al.*, 2014). The three genes are potentially involved in host seeking and/or taste discrimination in tsetse species. The Gr21a has three copies in all the five *Glossina* species and is believed to play a role in detection of CO_2; a tsetse volatile cue from vertebrate hosts (Torr *et al.*, 2006). Its co-expressed counterpart, Gr63a had signatures of selection under PAML analysis but not under the HyPhy package and thus could not be conclusively interpreted. Similar to Gr21a, Obp83a has four copies in each of the five *Glossina* species characterized so far and has been reported to be highly expressed in adult females 48 hours post feeding (Liu *et al.*, 2010) suggesting its role in host seeking. The only singleton found to be under significant selection is GluRIIA. Its role in tsetse is not known to date. Rather, its homolog in *Drosophila* has been implicated in postsynaptic signaling at the neuromuscular junction (Morimoto *et al.*, 2010).

Though few genes were found to harbor signatures of natural selection, it is evident that those identified are inclined towards host seeking and perhaps are responsible for diverse host preference observed across different species. The discrepancy in the number of gene loci identified to be under positive selection by PAML and HyPhy package could be due to few sequences available for the analysis. This could mean that more genes may be identified as under natural selection as more genomes become available. In other insects, evolution of ORs has been linked to their feeding habits as reported by recent study on plant-feeding Drosophilidae; *Scaptomyza flava.* The study linked the herbivorous nature of *S. flava* to the loss of conserved ORs such as OR42b and OR85d in (Goldman-Huertas *et al.*, 2015). Similar evolution could be suggested in tsetse as some of the genes identified to be under pressure are linked to host recognition .Furthermore, the non-neutral evolution of tsetse chemoreceptor genes is consistent with what reported among five *Drosophila* genomes (McBride and Arguello, 2007). In their study, McBride and Arguello found that *Drosophila* GRs are prone to gene loss while ORs are prone to selection pressure. Comparably, this study found more GRs to be under positive selection in tsetse compared to ORs.

In addition to forces of natural selection, the observed behavioral differences exhibited by tsetse species could be as a result unraveled diversity in their signal transduction machinery and/or post translational modification in their respective chemosensory proteins. Two different odor transduction mechanisms have been proposed (Dionne and Dubin, 1994) in insects namely (i) receptor-mediated (ion-channel) mechanism, which does not rely on the G-protein signaling pathway (Sato *et al.*, 2008), and (ii) G-protein cascade approach in which binding of semiochemicals to ORs is thought to activate the cyclic-nucleotide pathway (Gomez-Diaz, Martin and Alcorta, 2004; Wicher *et al.*, 2008). To date, little is known about the interaction between the tsetse specific ORs and their corresponding ligands and their downstream processing in the fly's central nervous system (CNS). Receptor-ligand interaction marks the beginning of odor processing that leads to a behavioral response. Post-translational modification is known to permit change of the amino acid properties as a reaction towards physiological need of an organism (Prabakaran *et al.*, 2012). For instance, phosphorylation has been attributed to the elasticity of ion channels involved in signaling (Levitan, 1994). Thus, it is important to study the downstream processes involved in odor processing across tsetse species to identify any underlying differences responsible for their differential behavior. Additionally, tsetse species may have developed an adaptation to specific odors based on learning. This type of learning has been reported to influence host selection in tsetse (Bouyer *et al.*, 2007). It is therefore possible that learning could play a role in differentially recognizing odors observed across different tsetse species.

3.5 Conclusions

All the five tsetse species compared in this study have a conserved chemosensory gene repertoire which assumes a sparse distribution across their genomes. A few of the chemosensory genes are rapidly evolving through duplication and a few are under natural selection pressure, perhaps to confer adaptive behavioral responses to host odors. In addition, all the *Glossina* species appear to have a reduced chemosensory gene repertoire relative to other insect species such as mosquitoes and fruit fly.

Comparative analysis of the chemosensory genes across the three subgenera did not reveal obvious differences in the chemosensory genes that could explain differential host responses exhibited by these species in the field. It is possible that the chemosensory proteins undergo post-translational modification that may alter the signaling mechanism of odors in the central nervous system. Therefore, there is need to undertake functional studies on identified genes and to further study the odor signaling pathway across tsetse species. This

will enhance our understanding on possible factors that influence differential host responses across tsetse species.

Supplementary Data and Figures

Table S3.1.1: Metadata for annotated *Glossina* chemosensory specific proteins (CSPs)

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
Glossina austeni						
A10p	GAUT014421-PA	18	2	+	166	1568172-1568807
GaCSP1	GAUT038415-PA	515	2	-	108	65638-66091
GaEjbp3A	GAUT027332-PA	326	2	-	135	175155-176575
GaEjbp3B	GAUT027343-PA	326	2	-	130	184996-188773
GaPhekIII	GAUT046063-PA	711	2	-	123	29038-31271
Glossina pallidipes						
GpdA10p	GPAI012674-PA	173	2	+	178	320391-320853
GpdCSP1	GPAI011776-PA	167	2	-	108	320391-320853
GpdEjbp3A	GPAI029774-PA	377	2	-	182	162350-164278
GpdEjbp3B	GPAI029784-PA	377	2	-	168	170273-173411
GpdPhekIII	GPAI031814-PA	409	3	-	123	90936-92933
Glossina fuscipes						
GffA10p	GFUI014924-PA	1	2	-	178	2392190-2392874
GffCSP1	GFUI040903-PA	59	2	-	108	908419-908851
GffEjbp3A	GFUI003186-PA	186	4	-	177	173922-176773
GffEjbp3B	GFUI003196-PA	186	2	-	145	179497-184154
GffPhekIII	GFUI039843-PA	573	3	-	158	75439-85017
Glossina brevipalpis						
GbrA10p	GBRI045129-PA	9	2	+	157	3972663-3973296
GbrCSP1	GBRI011414-PA	16	2	+	108	1186389-1186789
GbrEjbp3A	GBRI020682-PA	26	3	+	141	607033-609776
GbrPhekIII	GBRI020713-PA	26	3	-	123	872937-875314

Table S3.1.2: Metadata for annotated Odorant binding proteins (OBPs)

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
Glossina austeni						
Galush	GAUT003576-PA	10	3	+	161	836822-839595
GaObp19a	GAUT045923-PA	70	8	+	199	406478-414262
GaObp19b	GAUT045912-PA	70	5	+	157	414995-417289
GaObp19c	GAUT045925-PA	70	4	+	181	417849-420327
GaObp19d	GAUT045144-PA	69	5	-	144	918179-922013
GaObp28a	GAUT048147-PA	7	6	+	170	1849265-1853703
GaObp44a	GAUT018078-PA	218	2	+	141	21530-22059
GaObp56e	GAUT030435-PA	375	4	+	170	21242-23016
GaObp56h	GAUT041055-PA	57		+	134	742801-743269
GaObp69a	-	196	2	+	123	277985-278364

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
GaObp73a	GAUT039149-PA	52	2	+	272	922243-928634
GaObp19	GAUT028974-PA	34	2	-	139	1076535-1079002
GaObp20	GAUT051622-PA	99	3	-	261	132022-133021
GaObp56d	GAUT040992-PA	57	3	+	212	243969-250483
GaObp56e2	GAUT029308-PA	357	2	+	134	75389-75859
GaObp56i	GAUT028968-PA	34	1	+	148	1032106-1032552
GaObp57c	GAUT026721-PA	316	5	-	155	299578-293311
GaObp83a1	GAUT019500-PA	232	4	-	147	105165-107441
GaObp83a2	GAUT029664-PA	232	2	+	92	141010-141267
GaObp83a3	GAUT019501-PA	232	3	+	198	131226-141845
GaObp83a4	GAUT019501-PA	232	3	+	150	131226-141845
GaObp83cd	GAUT030010-PA	368	3	-	240	18505-20900
GaObp83ef	GAUT030009-PA	368	3	+	257	27234-28161
GaObp83g	GAUT030008-PA	368	2	+	140	33879-34401
GaObp84a	GAUT044447-PA	675	3	+	108	27863-29439
GaObp8a	GAUT043978-PA	65	2	-	150	51053-53056
GaObp99b1	GAUT051640-PA	99	2	+	149	216183-216701
GaObp99b2	GAUT051645-PA	99	2	-	153	216778-217324
GaObp99c	GAUT051620-PA	99	2	+	140	134606-136059
Glossina pallidipes						
Gpdlush	GPAI017685-PA	20	4	-	125	1425471-1431914
GpdObp19	GPAI006440-PA	122	1	-	138	4800-5216
GpdObp19a	GPAI032191-PA	417	4	-	150	75020-77205
GpdObp19b	GPAI032193-PA	417	5	-	157	66478-68754
GpdObp19c	GPAI032197-PA	417	3	-	184	63397-64127
GpdObp19d	GPAI018668-PA	21	5	+	144	2135728-2140116
GpdObp20	GPAI045033-PA	81	5	-	328	2135728-2140116
GpdObp28a	GPAI017770-PA	210	3	-	85	14967-1695
GpdObp44a	GPAI004501-PA	10	2	-	141	2187795218832
GpdObp56d	GPAI008752-PA	13	4	-	192	1172942-1177154
GpdObp56e	GPAI008777-PA	13	3	+	196	1336406-1341746
GpdObp56e2	GPAI018009-PA	213	6	-	134	385474-393230
GpdObp56h	GPAI008860-PA	13	2	+	134	1810580181104
GpdObp56i	-	45	4	+	126	155318-155695
GpdObp57c	GPAI009631-PA	148	5	+	126	264217-267946
GpdObp69a	-	17	2	+	118	
GpdObp83a1	GPAI013560-PA	180	4	-	156	135662140784
GpdObp83a2	GPAI013557-PA	180	5	+	150	174444-175177
GpdObp83a3	GPAI013558-PA	180	4	+	174	168581-171435
GpdObp83a4	GPAI013555-PA	180	4	+	147	165720-166848
GpdObp83cd	GPAI031702-PA	405	3	+	240	42191-44525
GpdObp83ef	GPAI031704-PA	405	3	-	257	36016-36925
GpdObp83g	GPAI031703-PA	405	2	-	140	30052-30571
GpdObp84a	GPAI005408-PA	116	5	-	147	863930-866624
GpdObp8a	GPAI041909-PA	68	2	-	158	210470-211189

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
GpdObp99b1	GPAI045017-PA	81	2	+	144	376046-376547
GpdObp99b2	GPAI045022-PA	81	2	-	153	37661537715
GpdObp99c	GPAI045024-PA	81	2	+	154	30183130237
GpdObp73a	-	74	2	-	118	325289-32505
Glossina fuscipes						
Gfflush	GFUI025618-PA	327	4	+	122	141534-145323
GffObp19	GFUI007906-PA	13	1	-	132	1332928-1333326
GffObp19a	GFUI000760-PA	JFJR 01012814	4	-	132	62038-64042
GffObp19b	GFUI000759-PA	JFJR 01012815	5	-	157	53651-55725
GffObp19c	GFUI000757-PA	JFJR 01012816	3	-	189	50704-51407
GffObp19d					114	
GffObp28a	GFUI048313-PA	7	3	-		1817622-1819513
GffObp44a	GFUI004675-PA	117	2	+	150	181227-181754
GffObp56d	GFUI008988-PA	14	1	+	141	263291-263782
GffObp56e	GFUI008564-PA	145	2	-	163	474544-475026
GffObp56h	GFUI009068-PA	14	2	+	138	776017-776483
GffObp56i	GFUI007894-PA	13	6	+		1288885-1298117
GffObp57c	GFUI026749-PA	341	2	+	134	64570-66051
GffObp69a	GFUI040667-PA	595	6	-	254	36454-45097
GffObp83a1	GFUI048612-PA	80	4	-	128	887783-888883
GffObp83a2	GFUI048613-PA	80	4	-	122	883137-886169
GffObp83a3	GFUI017944-PA	80	4	-	148	885305..885119
GffObp83a4		80	3	-	159	879899..879642
GffObp83cd	GFUI049167-PA	832	3	+	107	2175-450
GffObp83ef	GFUI004156-PA	112	3	-	145	731842-732784
GffObp83g	GFUI004155-PA	112	2	-	242	725365-725880
GffObp84a	GFUI027466-PA	352	3	+	258	297252-298513
GffObp8a	GFUI045274-PA	707	3	+	140	84705-88483
GffObp99b1	GFUI035804-PA	48	3	-	>85	1028188-1030308
GffObp99b2	GFUI035776-PA	48	3	+	184	1027533-1028107
GffObp99c	GFUI035783-PA	48	2	-	163	1107114-1108488
GffObp20	-	48			153	
GffOBp73a	-				146	
Glossina brevipalpis					>88	
Gbrlush	GBRI030526-PA	43	7	-	222	600222-606859
GbrObp19	GBRI036202-PA	5	2	+	181	2274177-2274862
GbrObp19a	GBRI035551-PA	58	4	+	149	1163684-1166420
GbrObp19b	GBRI035552-PA	58	5	+	156	1171185-1173340
GbrObp19c	GBRI035549-PA	58	2	+	129	1175394-1175905
GbrObp19d	GBRI010734-PA	163	5	-	148	106459-111816
GbrObp20	GBRI012886-PA	181	3	-	263	253648-254657
GbrObp28a	GBRI045128-PA	9	3	+	100	4065050-4066455
GbrObp44a	GBRI026688-PA	368	2	+	142	50617-51173

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
GbrObp56d	GBRI016471-PA	211	4	+	175	188284-193310
GbrObp56e	GBRI016436-PA	211	2	-	152	69280-71060
GbrObp56e2	GBRI010929-PA	165	2	-	253	295002-295826
GbrObp56h	GBRI040269-PA	77	2	+	134	261505-261991
GbrObp56i	GBRI036199-PA	5	4	+	363	2276830-2280407
GbrObp57c	GBRI041963-PA	83	4	-	146	874548-882752
GbrObp69a	GBRI013864-PA	191	4	+	109	66042-68381
GbrObp83a1	GBRI031755-PA	47	4	-	158	1006248-1008915
GbrObp83a2	GBRI031753-PA	47	4	+	151	1046297-1046962
GbrObp83a3	GBRI031754-PA	47	4	-	154	1038529-1040123
GbrObp83a4	GBRI031756-PA	47	4	+	179	1024308-1036548
GbrObp83cd	GBRI031703-PA	47	3	+	239	684542-68782
GbrObp83ef	GBRI031705-PA	47	3	-	254	678967-679889
GbrObp83g	GBRI031704-PA	47	2	-	140	74746-7550
GbrObp84a	GBRI023685-PA	304	4	-	176	74746-75503
GbrObp8a	GBRI009351-PA	151	4	-	167	382122-385965
GbrObp99b1						
GbrObp99b2	GBRI012898-PA	181	3	-	149	361796-362320
GbrObp99c	GBRI012882-PA	181	2	+	164	258700-25925

Table S3.1.3: Metadata for annotated *Glossina* Sensory membrane proteins (SNMPs)

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
Glossina austeni						
GaSNMP1	GAUT049266-PA	85	7	+	540	412-4831
GaSNMP2	GAUT008732-PA	142	7	-	411	12468-15563
Glossina pallidipes						
GpdSNMP	GPAI010405-PA	153	7	-	540	74092-79270
GpdSNM2	GPAI029269-PA	369	5	-	377	11351-20597
Glossina fuscipes						
GffSNMP1	GFUI000887-PA	JFJR01012825	7	+	540	15099-19239
GffSNMP2	GFUI009502-PA	152	6	+	423	567526-576060
Glossina brevipalpis						
GbrSNMP1	GBRI029848-PA	14	10	+	391	2947412-2960911
GbrSNMP2	GBRI009197-PA	41	6	+	384	585363-593851

Table S3.1.4: Metadata for annotated Glossina Gustatory Receptors (GRs)

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
Glossina austeni						
GaGr21a1	GAUT050702-PA	91	5	-	460	825820-827456
GaGr21a2	GAUT041339-PA	58	13	-	460	403004-406548
GaGr21a3	-	58	3	-		401294-402776
GaGr28bD	GAUT018371-PA	21	5	-	384	1288618-1290861
GaGr28bC	GAUT037007-PA	498	4	+	463	181007-183576
GaGr2a	GAUT018378-PA	21	5	-	410	1408991-1410484
GaGr33a	GAUT030746-PA	37	3	-	398	822030..823331
GaGr58c	GAUT018082-PA	218		+	337	51995-53596
GaGr59f1	GAUT016799-PA	202	6	-	480	418171-425976
GaGr59f2	GAUT032734-PA	402	6	+	404	13568-19746
GaGr63a	GAUT042077-PA	5	6	+	425	2235841-2244048
GaGr66a	GAUT025297-PA	2	3	-	397	2031438-2032923
GaGr32a	GAUT018813-PA	225	1	+	342	297724-298752
Glossina pallidies						
GpdGr21a	GPAI014620-PA	18	4	-	437	2126014-2127571
GpdGr212	GPAI045887-PA	86	6	-	425	409943-411461
GpdGr2a3	-	86	4	-	433	410297-409943
GpdGr8bD	GPAI035388-PA	48	4	-	443	460073-462639
GpdGr2a	GPAI037163-PA	523	5	-	408	12416-13888
GpdGr32a	GPAI019874-PA	237	7	+	360	47096-53066
GpdGr33a	GPAI039461-PA	59	3	-	405	844645-86637
GpdGr58c	GPAI004494-PA	10	4	-	366	216084-2216345
GpdGr59f1	GPAI040289P	1	7	+	41	121972-12944
GpdGr59f2	GPAI040385-PA	61	9	-	27	714203-72482
GpdGr63a	GPAI007341-PA	12	7	+	474	636615-6445
GpdGr66a	GPAI024994-PA	2	4	+	343	885714-8809
Gpd28bC	GPAI043562-PA	75	4	+	463	111282-11231
Glossina fuscipes						
GffGr21a1	GFUI005702-PA	123	4	+	453	323330-324871
GffGr21a2	GFUI034303-PA	462	1	-	373	130506-134124
GffGr21a3	GFUI041369-PA	604	4	+	408	127218-127726
GffGr28bB	GFUI018032-PA	233	4	-	446	425543-430888
GffGr28bC	GFUI027606-PA	355	4	-	462	149208-151105
GffGr2a	GFUI026404-PA	339	5	-	418	10488-11985
GffGr33a	GFUI051944-PA	934	3	+	442	8609-1967
GffGr59f1	GFUI022205-PA	284	5	+	436	290713-294736
GffGr59f2	GFUI025370-PA	321	4	+	445	210603-21322
GffGr63a	GFUI036605-PA	4	8	+	434	2263346-227282
GffGr66a	GFUI041074-PA	5	7	-	368	106507-1071392
GffGr58c	-	117	4	+	374	214562..215622
GffGr32a	-	417	7	-	375	57967..56940

GffGr28bE	-	235	4	-	414	430888..430235
Glossina brevipalpis						
GbrGr21a1	GBRI008315-PA	144	3	-	429	228455-2298
GbrGr21a2	GBRI004163-PA	145	1	+	444	320168-32288
GbrGr21a3	-	114		+	483	322215-32274
GbrGr28bD	GBRI016968-PA	21	4	-	440	828774-80444
GbrGr2a	GBRI016977-PA	21	6	-	346	905147-906777
GbrGr28E	GBRI039848-PA	74	4	+	321	221852-226499
GbrGr59f1	GBRI043822-PA	93	6	+	426	481031-488570
GbrGr59f2	GBRI043906-PA	93	7	+	395	949716-954740
GbrGr63a	GBRI014933-PA	1	9	+	454	1371407-137713
GbrGr33a	-	6	4	+	269	1379126..137950

Table S3.1.5: Metadata for annotated *Glossina* Odorant Receptors (ORs)

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
Glossina austeni						
GaOr13a	GAUT014395-PA	18	6	-	466	1294296-1298135
GaOr19b	GAUT050371-PA	8	7	-	445	3312163-3317121
GaOr24a	GAUT004311-PA	113	7	-	458	128786-132433
GaOr2a	GAUT045920-PA	70	3	+	394	480470-482494
GaOr33b	GAUT028888-PA	34	8	+	508	470445-480096
GaOr43a1	GAUT021583-PA	258	7	+	354	254902974
GaOr43a2	GAUT000836-PA	JMRR01017845	7	-	342	26406-34495
GaOr7a1	GAUT050213-PA	8	4	-	342	1602836-1602105
GaOr7a2	GAUT050213-PA	8	3	-	442	1597269-1597090
GaOr42b	GAUT022268-PA	266	7	+	379	328913-33254
GaOr45a1	GAUT044021-PA	65	5	-	405	38719238960
GaOr45a2	GAUT022034-PA	261	6	-	397	376409-379734
GaOr45a3	GAUT028238-PA	33	7	-	335	356645-362680
GaOr46a	GAUT011101-PA	15	2	+	351	1981339-1982450
GaOr47b	GAUT016620-PA	200	7	+	429	51228-53865
GaOr49b	GAUT005608-PA	121	6	-	485	207215-215611
GaOr56a1	GAUT042364-PA	602	5	+	251	81198-86639
GaOr56a2	GAUT042360-PA	602	4	+	393	75448-78119
GaOr59a	GAUT018044-PA	217	2	-	384	2492672052
GaOr63a1	-	10	5	-	352	1103692-1103504
GaOr63a	GAUT003629-PA	10	6	-	>164	1101200-1106895
GaOr67c1	GAUT038273-PA	50	3	-	269	1292778-1298372
GaOr67a	GAUT018383-PA	21	6	-	420	1437626-1437174
GaOr67c2	GAUT032244-PA	3	4	+	295	22667162679
GaOr67d1	GAUT021320-PA	253	10	-	344	19169-2267
GaOr67d2	GAUT051820-PA	9	5	+	392	84231-38936
GaOr67d5	GAUT021321-PA	253	4	-	407	24222-34589

GaOr74a	GAUT035779-PA	468	5	-	404	55131-56628
GaOr7a3	GAUT050214-PA	8	3	-	394	1604220-1607212
GaOr82a	GAUT003281-PA	108	5	-	299	540697-542872
GaOr85b	GAUT005460-PA	120	4	-	438	46304-51854
GaOr85d	GAUT006649-PA	12	6	+	401	80362-98058
GaOr85e	GAUT040462-PA	560	5	-	451	18651-12045
GaOr88a	GAUT036655-PA	487	3	-	420	163257-16621
GaOr94a	GAUT005363-PA	11	6	-	246	260914-262156
GaOrco	GAUT034813-PA	445	10	-	497	15129-46682
Glossina pallidipes						
GpdOr13a	GPAI034871-PA	479	4	-	378	7687-11530
GpdOr19b	GPAI027642-PA	338	3	-	393	85099-87347
GpdOr24a	GPAI015219-PA	197	4	-	362	131063-133020
GpdOr2a	GPAI004010-PA	107	3	-	394	943507-946652
GpdOr33b	GPAI034198-PA	45	2	+	380	1044239-1045446
GpdOr43a1	GPAI039623-PA	5	7	+	388	574081-57829
GpdOr43a2	GPAI039631-PA	5	2	+	287	519406- 519700
GpdOr7a1	GPAI031316-PA	3	4	+	362	2526168-2528893
GpdOr7a2	GPAI031326-PA	3	3	+	394	2527583.252831
GpdOr42b	GPAI029610-PA	371	4	-	269	17560-21114
GpdOr45a1	GPAI041951-PA	68	5	-	405	508520-510986
GpdOr45a2	GPAI026906-PA	323	5	-	395	16085-19449
GpdOr45a3	GPAI014680-PA	18	5	-	410	2507760-2510760
GpdOr46a1	GPAI009882-PA	14	3	-	388	700661-701953
GpdOr46a2	GPAI009200-PA	143	5	+	369	49237..50041
GpdOr47b	GPAI039539-PA	5	5	+	331	93062-94564
GpdOr49	GPAI001497-PA	JMRQ01006307	3	-		2536-3830
GpdOr49b	GPAI004557-PA	10	5	-	298	26544262662902
GpdOr56a1	GPAI045424-PA	83	6	+	365	689099-693647
GpdOr56a2	GPAI045426-PA	83	4	+	382	683333-686066
GpdOr59a	GPAI039747-PA	5	2	+	384	1426752-1428008
GpdOr63a	GPAI017649-PA	20	8	-	371	1165231-1171743
GpdOr67c1	GPAI041241-PA	657	2	+	329	32493-33542
GpdOr67a	GPAI037164-PA	523	10	-	376	55119-65818
GpdOr67c2	GPAI033169-PA	43	5	+	406	18938519379
GpdOr67d1	GPAI012943-PA	105	3	-	455	778996-782573
GpdOr67d3	GPAI012945-PA	105	4	-	313	771993-778475
GpdOr67d4	GPAI046202-PA	88	7	+	332	956516-961886
GpdOr67d5	GPAI002749-PA	0	5	-	380	5037832-5039272
GpdOr67d6	GPAI042230-PA	69	9	+	373	1111489-1121348
GpdOr7a3	GPAI031315-PA	3	3	+	394	2514726-2517401
GpdOr82a	GPAI024118-PA	28	7	-	428	1605921-1610562
GpdOr85b	GPAI001626-PA	JMRQ01006330	4	+	332	12958-18883
GpdOr85c	GPAI040919-PA	63	4	+	405	677245-679224
GpdOr85d	GPAI002024-PA	0	6	-	415	978630-980945

GpdOr85e	GPAI004056-PA	108	6	+	465	471781-473530
GpdOr88a	GPAI027550-PA	335	5	+	764	81179-10785
GpdOr94b	GPAI009882-PA	145	3	-	342	31320-31078
GpdOrco	GPAI035133-PA	481	8	-	477	20629-28693
Glossina fuscipes						
GffOr13a	GFUI014938-PA	1	7	+	477	2644857-2650323
GffOr2a1	GFUI043297-PA	65	3	-	393	113446-112907
GffOr24a	GFUI032492-PA	42	4	-	365	190458-192424
GffOr2a2	GFUI028755-PA	371	4	+	441	1134-5774
GffOr33b	GFUI007794-PA	13	4	+	387	820037-820798
GffOr43a1	GFUI003104-PA	107	6	+	327	562923-567104
GffOr43a2	GFUI003105-PA	107	5	+	290	508357-512244
GffOr7a1	GFUI003499-PA	10	4	-	376	122853-12564
GffOr42b	GFUI028213-PA	363	3	-	379	298922-302539
GffOr45a1	GFUI008162-PA	141	6	-	405	149517-152073
GffOr45a2	GFUI032116-PA	421	4	-	362	119677-11920
GffOr45a3	GFUI005658-PA	123	7	-	503	32427-37272
GffOr46a1	GFUI037305-PA	514	2	-	379	29704-3090
GffOr46a2	GFUI034469-PA	467	8	-	478	24341-13025
GffOr47b	GFUI045476-PA	713	6	+	406	28879-3223
GffOr49b	GFUI009257-PA	14	7	+	264	813504-181692
GffOr56a1	GFUI038138-PA	532	4	-	138	147381-15196
GffOr56a2	GFUI038147-PA	532	3	-	387	156749-159310
GffOr59a	GFUI042981-PA	64	2	+	336	439177-440282
GffOr63a	GFUI027054-PA	347	7	-	410	553756258
GffOr67c1	GFUI051694-PA	922	5	-	346	39892-5066
GffOr67d1	GFUI007388-PA	137	2	-	217	07939-10865
GffOr67d4	GFUI043789-PA	672	6	+	364	30656-36090
GffOr67d5	GFUI036188-PA	49	4	+	35	46737746868
GffOr67d6	GFUI022534-PA	28	10	+	368	43112-5885
GffOr74a	GFUI022472-PA	289	5	-	404	778-1028
GffOr7a2	GFUI003500-PA	10	3	-	394	132922-135699
GffOr82a	GFUI053522-PA	9	6	+	405	2146338-2149094s
GffOr85b	GFUI022126-PA	283	4	+	337	75605-7679
GffOr85c	GFUI047908-PA	79	4	-	345	2453-14660
GffOr85d	GFUI049134-PA	82	12	-	496	779971-787789
GffOr85e	GFUI037003-PA	50	5	-	465	165317-167064
GffOr88a	GFUI024278-PA	309	7	+	450	20570-26379
GffOr94b	GFUI012941-PA	184	4	-	304	482702-482103
GffOrco	GFUI035140-PA	47	9	+	463	95255-10431
Glossina brevipalpis						
GbrOr13a	GBRI045111-PA	9	2	-	416	3617747361726
GbrOr19b	GBRI018062-PA	231	4	+	414	85079-8859
GbrOr24a	GBRI036522-PA	5	4	-	375	3803138-3804854
GbrOr2a	GBRI035583-PA	58	3	+	393	1251539125492
GbrOr33b	GBRI036342-PA	5	4	-	374	2787260-279346
GbrOr43a1	GBRI002464-PA	102	6	+	350	9217-9561

GbrOr30a	GBRI016989-PA	21	8	-	357	925700-93083
GbrOr7a1	GBRI044639-PA	99	5	-	454	59914-60046
GbrOr42b	GBRI034666-PA	54	3	+	355	13935571-1394170
GbrOr45a1	GBRI009897-PA	158	5	-	405	6520-67771
GbrOr45a2	GBRI026647-PA	366	4	-	309	10230-9775
GbrOr45a3	GBRI008361-PA	144	7	-	562	478678-485648
GbrOr46a	GBRI028428-PA	3	4	-	411	1299053-1300843
GbrOr47b	GBRI026891-PA	36	7	-	382	1247663-1251882
GbrOr49b	GBRI015995-PA	209	5	+	326	183313-185626
GbrOr56a1	GBRI011898-PA	172	4	+	256	350661-351631
GbrOr56a2	GBRI011904-PA	172	4	+	389	345764-347359
GbrOr59a	GBRI011358-PA	16	2	-	384	724514-725775
GbrOr63a1	GBRI031244-PA	45	8	+	405	1372025137823
GbrOr63a2	GBRI031534-PA	46	5	-	357	1743896-174361
GbrOr92a	GBRI002179-PA	0	3	+	343	963120-963358
GbrOr67c	GBRI02158-PA	356	5	+	386	34418-3721
GbrOr67d1	GBRI017432-PA	224	7	+	280	237077-237607
GbrOr67d3	-	224	6	+	375	238586..238724
GbrOr67d4	-	387	5	-	387	6101..5852
GbrOr67d5	GBRI017598-PA	228	4	-	385	86282-187675
GbrOr67d6	GBRI040021-PA	75	4	+	391	421880-422953
GbrOr7a2	GBRI044640-PA	99	3	-	394	604341-606663
GbrOr82a	GBRI018811-PA	154	7	-	663	93037-92507
GbrOr85b	GBRI027004-PA	372	4	-	427	105938-107449
GbrOr85c	GBRI041284-PA	7	5	+	418	3478923-3483783
GbrOr85d	GBRI030235-PA	42	10	+	439	1220757-1240238
GbrOr85e	GBRI005734-PA	126	3	-	376	369996-369418
GbrOr88a	GBRI013056-PA	183	3	+	295	464204-463800
GbrOr94b	GBRI012762-PA	17	4	-	340	2487344-2488500
GbrOrco	GBRI030714-PA	440	5	-	230	65492-57864
GbrOr74a	-	55	6	+	401	63828-65251

Table S3.1.6: Metadata for annotated *Glossina* Ionotropic/Ionotropic Glutamate Receptors (IRs/iGluRs)

Gene Name	Gene ID	Scaffold	Exons	Strand	Length (aa)	Co-ordinates
Glossina austeni						
GaGluRIA	GAUT036857-PA	491	15	-	923	12641-26444
GaGluRIB	GAUT010844-PA	15	7	-	490	271878-289187
GaGluRIIA	GAUT032862-PA	406	9	+	610	227062-235839
GaGluRIIB	GAUT032862-PA	406	8	+	838	227062-235839

GaGluRIIC	GAUT018821-PA	225	9	-	818	320480-326531
GaGluRIIE	GAUT036856-PA	491	12	-	851	7165-11545
GaIR10a	GAUT051652-PA	99	7	+	580	264729-270327
GaIr21a	GAUT029664-PA	360	8	+	893	151916-160965
GaIr25a	GAUT011688-PA	165	9	+	915	419483-427269
GaIr31a	GAUT019628-PA	234	6	+	604	253145-256269
GaIr40a	GAUT028361-PA	33	12	+	799	1168341-1176017
GaIr56b	GAUT017831-PA	214	1	+	613	20365-22206
GaIr64a	GAUT035430-PA	45	11	-	589	490121-495546
GaIr68a	GAUT051179-PA	95	3	+	613	198684-202160
GaIr75a	GAUT013397-PA	17	6	-	503	1846089-1855137
GaIr75a2	GAUT013397-PA	17	5	-	547	1846910-1846101
GaIr75d	GAUT003875-PA	10	5	+	582	2831325-2834095
GaIr76a	GAUT051343-PA	96	11	-	726	662805-684595
GaIr76b	GAUT037856-PA	4	11	+	530	4002993-4003496
GaIr84a	GAUT038749-PA	520	8	+	745	111057-129675
GaIr8a	GAUT002274-PA	0	4	-	537	5757150-5758962
GaKaiR1A	GAUT026102-PA	308	16	-	295	71221-75175
GaKAiR2c	GAUT023024-PA	276	5	-	439	319882-321448
GaKaiR2d	GAUT005991-PA	125	13	-	1121	167157-176447
GaKaiR2e	GAUT026111-PA	308	12	+	848	103397-112434
GaNMDAR1	GAUT031582-PA	397	13	-	993	52340-61072
GaNMDAR2	GAUT008471-PA	13	6	-	470	2514984-2524581
GaClumsy	-	100	8	+	363	109611-215915
Glossina pallidipes						
GpdClumsy	GPAI011564-PA	162	15	-	1205	347557-357125
GpdGluRIA	GPAI006854-PA	125	13	+	670	651212-667233

GpdGluRIB	GPAI010111-PA	14		+	440	2333491-2347296
GpdGluRIIA	GPAI011561-PA	162	11	+	934	358155-366744
GpdGluRIIB	GPAI011561-PA	162	12	+	916	2354879..2355256
GpdGluRIIC	GPAI019869-PA	237		-	865	73886-78958
GpdGluRIIE	GPAI006854-PA	125	12	+	934	656655..657506
GpdIR10a	GPAI045043-PA	81	11	+	422	427515-428839
GpdIr21a	GPAI016226-PA	1	8	+	897	2318938-2327793
GpdIr25a	GPAI011331-PA	15	9	-	915	2393818-2401767
GpdIr31a	GPAI007758-PA	131	5	+	560	35981236238
GpdIr40a	GPAI004624-PA	110	11	-	834	272452-28262
GpdIr56b	GPAI022505-PA	26	1	+	613	43351-432192
GpdIr64a	GPAI032358-PA	41	10	+	550	839108-844584
GpdIr68a	GPAI017485-PA	209	4	-	574	287965-291718
GpdIr75a	GPAI036018-PA	4	6	+	1002	404760-415937
GpdIr75c	GPAI036018-PA	4	5	+	571	414465..45274
GpdIr75d	GPAI025294-PA	2	5	-	556	3454055-3456953
GpdIr76a	GPAI027894-PA	33	9	-	436	1763148-1774106
GpdIr76b	GPAI044391-PA	7	11	-	622	147031-154107
GpdIr84a	GPAI022870-PA	273	6	+	635	224716-231135
GpdIr8a	GPAI042411-PA	6	10	-	920	1453560-1459816
GpdIr93a	GPAI006139-PA	11	12	+	870	2423153243279
GpdKaiR1A	GPAI006142-PA	11	16	-	958	2386410-240035
GpdKAiR2c	GPAI029067-PA	362	6	-	463	12129-21576
GpdKaiR2d	GPAI010422-PA	153	12	-	997	231756-245490
GpdKaiR2e	GPAI006139-PA	11	12	+	870	2423153-2432799
GpdNMDAR1	GPAI006944-PA	126	14	+	1000	484327-497091
GpdNMDAR2	GPAI030510-PA	38	8	-	898	1382581-1393263

Glossina fuscipes

GffClumsy	GFUI019198-PA	247	13	+	1022	406109-413460
GffGluRIA	GFUI016186-PA	214	12	+	731	480630-486116
GffGluRIB	GFUI018591-PA	23	13	+	970	926941-981018
GffGluRIIA	GFUI019200-PA	247	12	+	1797	38785-5402060
GffGluRIIC	GFUI031610-PA	413	11	+	787	27607-3321
GffGluRIIE	GFUI041857-PA	618	13	-	851	29499-134101
GffIR10a	GFUI035802-PA	48	5	-	579	978940-984632
GffIr21a	GFUI017944-PA	232	6	-	890	15894-15684
GffIr25a	GFUI008852-PA	148	9	+	915	92690-499168
GffIr31a	GFUI031962-PA	41	6	+	591	58898-592064
GffIr40a	GFUI025996-PA	331	4	-	783	72875-80318
GffIr56b	GFUI041337-PA	602	1	-	613	150593346
GffIr64a	GFUI028023-PA	35	8	+	420	1286620-1289121
GffIr68a	GFUI019558-PA	250	4	-	592	256288-260192
GffIr75a	GFUI029180-PA	37	2	-	475	106926-107812
GffIr75c	GFUI029178-PA	37	3	-	639	98981-99779
GffIr75d	GFUI031962-PA	41	6	+	591	582898-592064
GffIr76a	GFUI043801-PA	673	9	+	664	33058-37788
GffIr76b	GFUI005590-PA	122	15	+	615	396974-400240
GffIr84a	GFUI004860-PA	118	6	+	612	48159848767
GffIr8a	GFUI020203-PA	25	8	+	876	743611-749393
GffIr93a	GFUI000063-PA	13	13	-	848	97023-106414
GffKAiR2c	GFUI009601-PA	154	6	-	428	63328-65417
GffKaiR2d	GFUI000460-PA	JFJR01011458	11	+	1071	152550-161731
GffKaiR2e	GFUI000063-PA	JFJR01008464	13	-	848	97023-106414
GffNMDAR1	GFUI045184-PA	702	10	-	949	114635-120420
GffNMDAR2	GFUI050910-PA	8	9	-	933	331214-340774

Glossina brevipalpis

GbrClumsy	GBRI004368-PA	116	15	-	1014	341723-347401
GbrGluRIA	GBRI037007-PA	61	14	-	1766	771920-783580
GbrGluRIB	GBRI006509-PA	12	13	-	931	2534017-2580729
GbrGluRIIA	GBRI004366-PA	116	12	+	604	350718-363113
GbrGluRIIB	GBRI004366-PA	116	10	+	363	
GbrGluRIIC	GBRI013356-PA	188	12	+	870	69252-376922
GbrGluRIIE	GBRI037007-PA	61	15	-	1766	771920..783580
GbrIR10a	GBRI012928-PA	181	9	+	761	452936-459660
GbrIr21a	GBRI001929-PA	0	3	-	406	8179738-8181089
GbrIr25a	GBRI023337-PA	2	10	+	894	4826984-4833513
GbrIr31a	GBRI000712-PA	0	5	+	634	2050653-2053202
GbrIr40a	GBRI039411-PA	71	10	-	793	494018-497181
GbrIr56b	GBRI033584-PA	50	3	+	613	920674-925264
GbrIr64a	GBRI012051-PA	174	9	+	648	408946-417138
GbrIr68a	GBRI033291-PA	4	6	+	664	4545805-4549340
GbrIr75a	GBRI016181-PA	20	7	-	1067	1115179-1128075
GbrIr75c	GBRI016181-PA	20	6	+	463	1122899.. 1122282
GbrIr75d	GBRI012020-PA	174	3	-	343	166465-167616
GbrIr76a	GBRI018928-PA	244	13	-	616	22656-28696
GbrIr76b	GBRI009997-PA	159	17	+	555	333681-343624
GbrIr84a	GBRI002787-PA	105	6	+	643	104851-110341
GbrIr8a	GBRI010267-PA	15	4	+	523	1360620-1362318
GbrIr93a	GBRI006799-PA	132	5	+	736	44007-54028
GbrKaiR1A	GBRI006802-PA	132	15	-	825	5046-16988
GbrKAiR2	GBRI006799-PA	132	12	+	808	44007-54028
GbrKaiR2d	GBRI029815-PA	41	14	+	1070	418769-429647
GbrNMDAR1	GBRI013857-PA	191	13	+	407	47354-48899
GbrNMDAR2	GBRI040612-PA	79	8	-	913	110001-12266

GbrGluRIID	-	61	-	687	775629-775024

Table S 3.2: *Glossina* transcriptomes used as training sets for gene prediction in the Maker2 annotation pipeline

Species	Tissue	Sex
G. pallidipes	Heads	Female
G. pallidipes	Heads	Male
G. pallidipes	Gut	Female
G. pallidipes	Lactating	Female
G. pallidipes	Non-lactating	Female
G. pallidipes	Whole body	Male
G. pallidipes	Salivary glands	Mixed
G. f. fuscipes	Heads	Mixed
G. f. fuscipes	Lactating	Female
G f. fuscipes	Non-lactating	Female
G. f. fuscipes	Whole body	Male
G. f. fuscipes	Reproductive organs	Female
G. f. fuscipes	Salivary glands	Mixed
G f fuscipes		
G. brevipalpis	Whole body	Mixed
G. brevipalpis	Larvae	1st and 2nd instar mixed
G. brevipalpis	Pupae	Mixed age

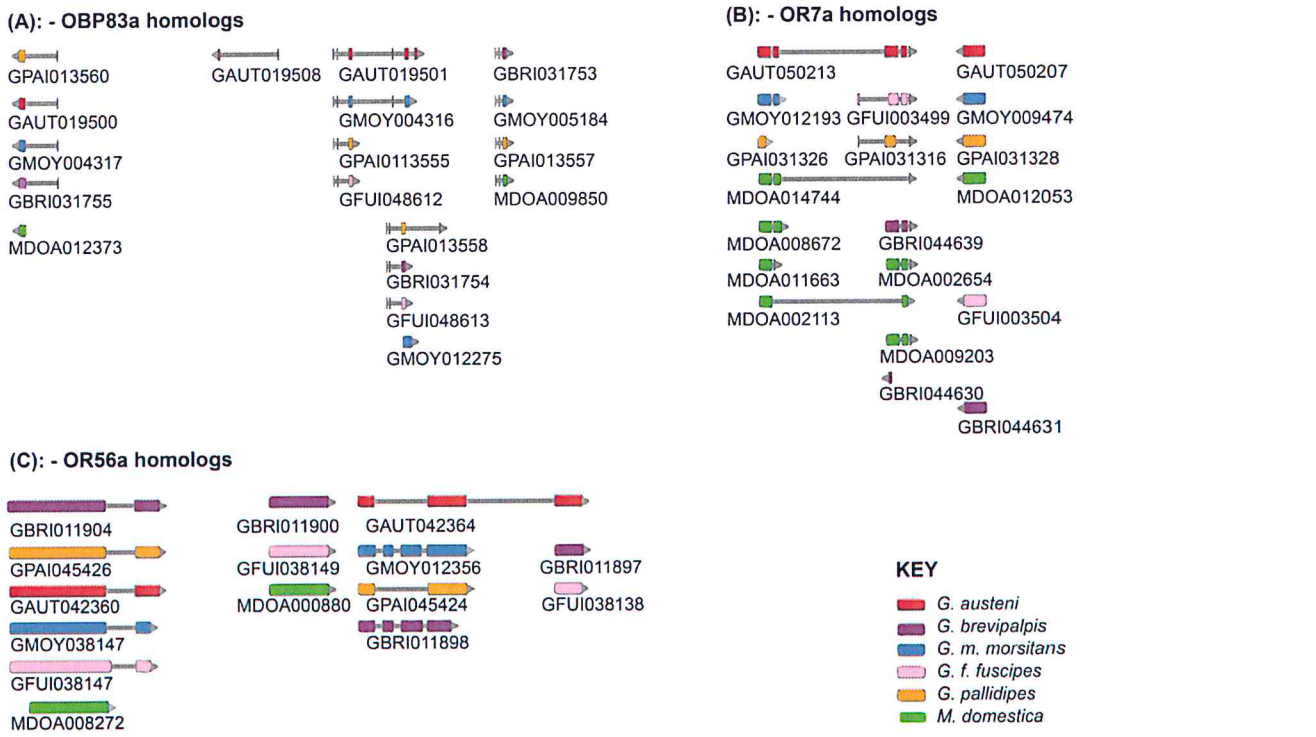


Figure S3. 1: VectorBase; Web Apollo screenshots illustrating gene structure and arrangement of selected duplicated (OBP83A, OR7A and OR56a) chemosensory genes across five *Glossina* genomes including *G. austeni* (GAUI*), *G. brevipalpis* (GBRI*), *G. f. fuscipes* (GFUI*), *G. m. morsitans* (GMOY*) and *G. pallidipes* (GFU*), and in *M. domestica* (MDOA*).

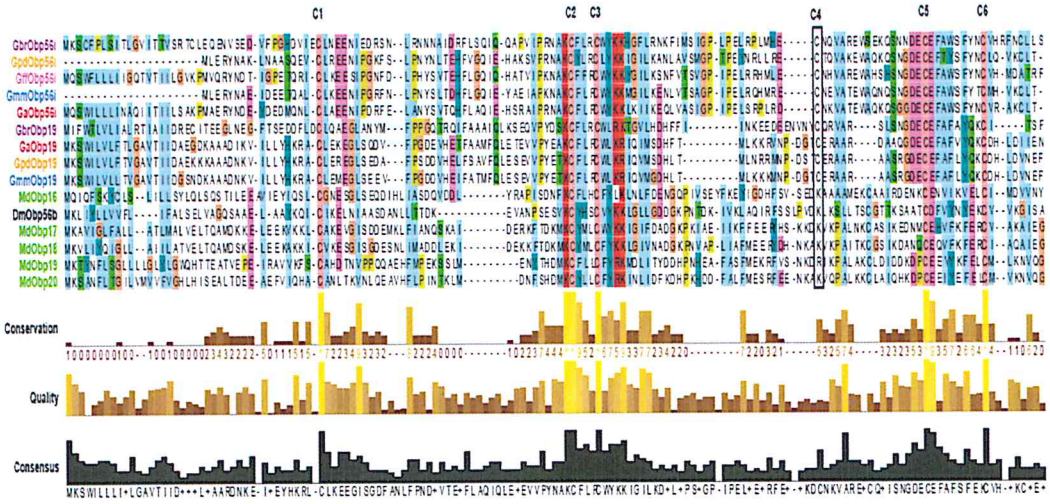


Figure S 3.2: Alignment of Obp19 and Obp56i from *Glossina*, Obp16-20 from *M. domestica* and Obp56i from *D. melanogaster* species. Variation of amino acids between conserved cysteine(s) C3 and C4 show deletion in Obp56i and Obp19 from *Glossina.*

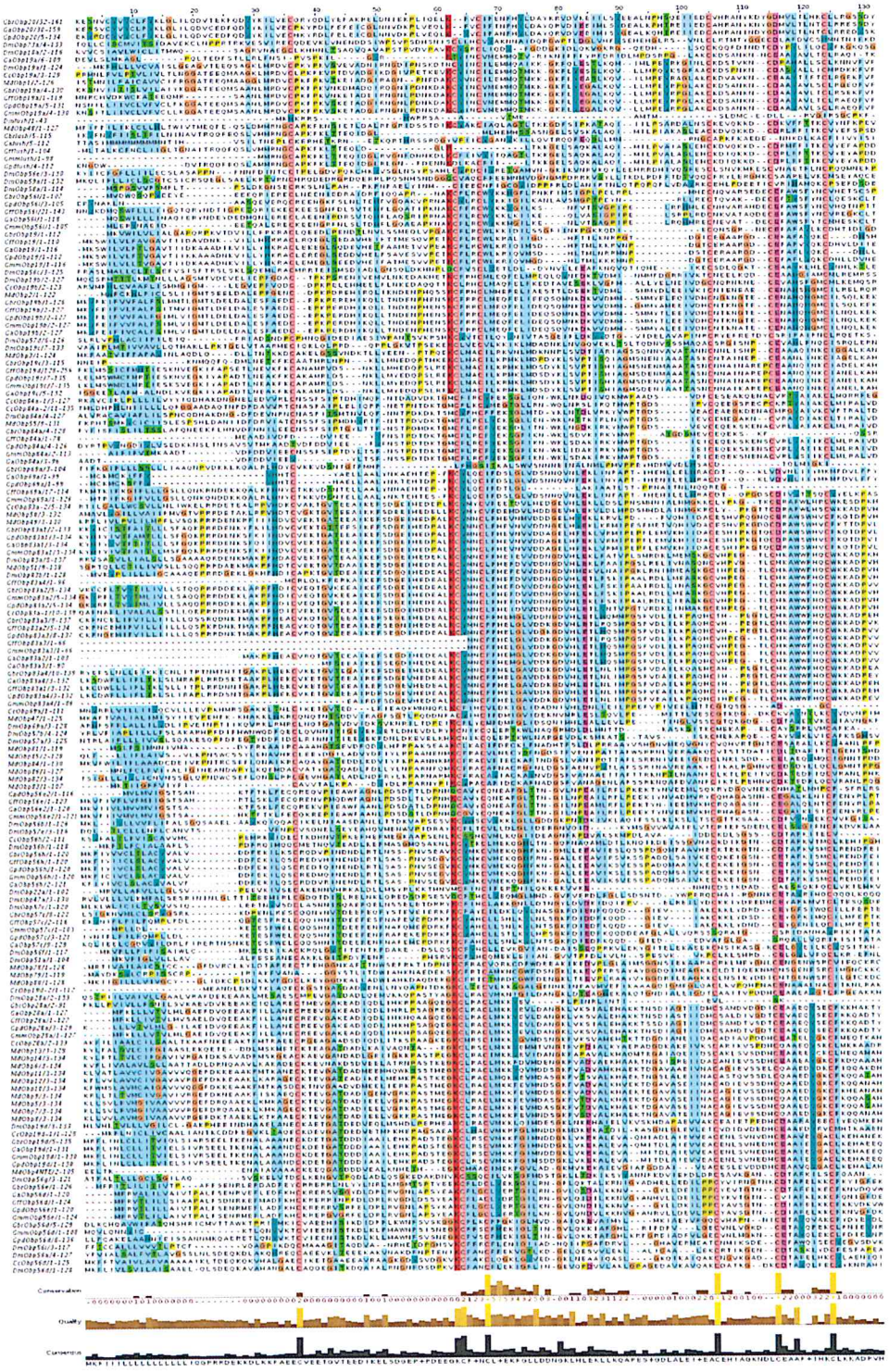


Figure S 3.3 1: Alignment of amino acid sequences of Classic OBPs identified in five *Glossina* species against those of closely related Diptera The six conserved cysteine residues are highlighted in orange.

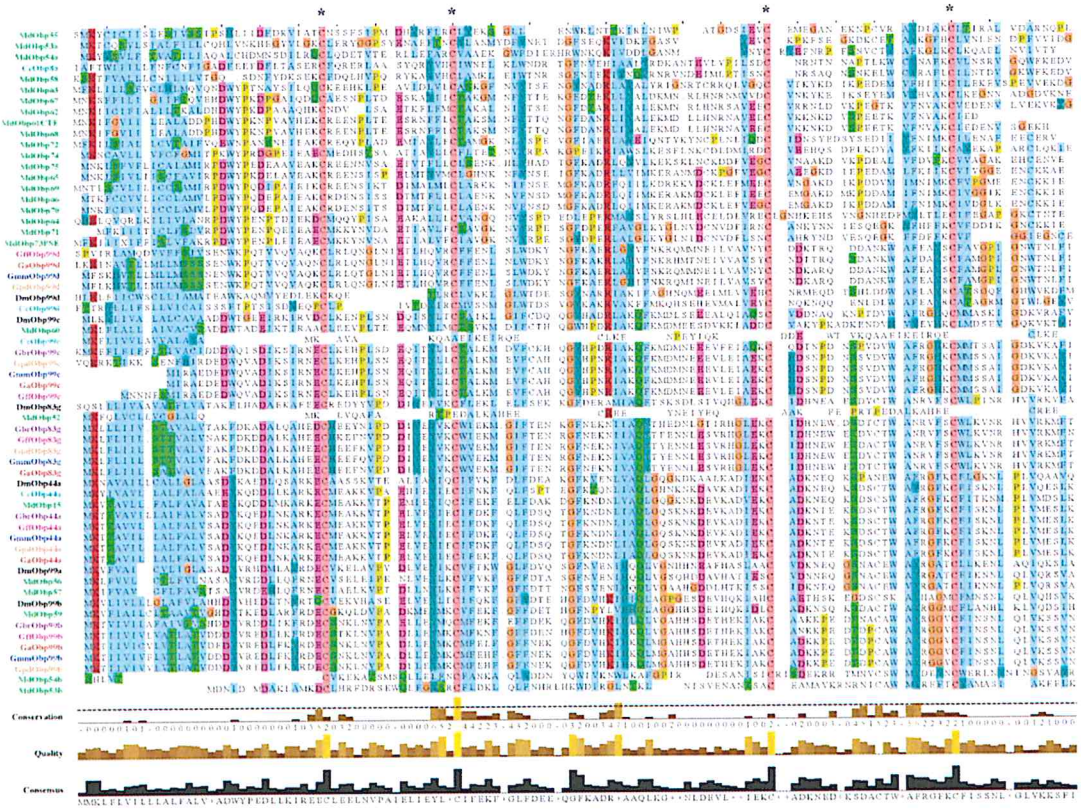


Figure S 3.3 2: Alignment of amino acid sequences of Minus-C OBPs identified in five *Glossina* species against those of closely related Diptera The four conserved cysteine residues are highlighted in orange and marked with an asterisk (*).

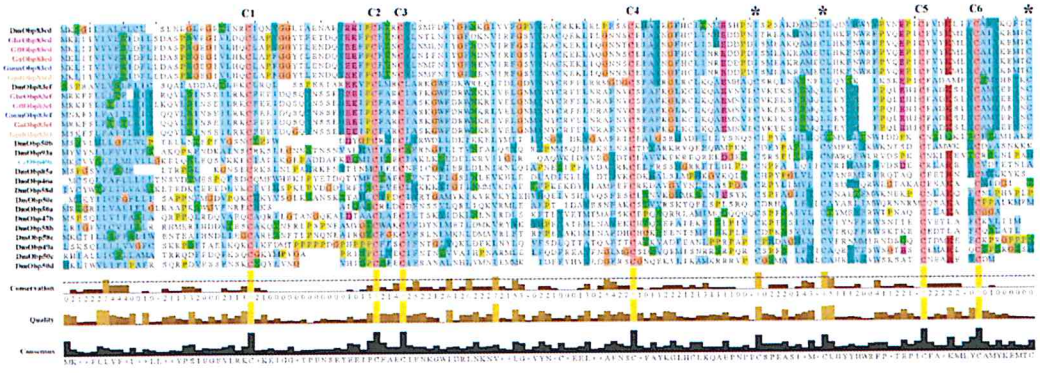


Figure S 3.3 2: Alignment of amino acid sequences of Plus-C and Classic-dimer OBPs identified in five *Glossina* species against those of closely related Diptera The six conserved cysteine residues are highlighted in orange and marked with an asterisk (*). The less conserved cysteines are boxed.

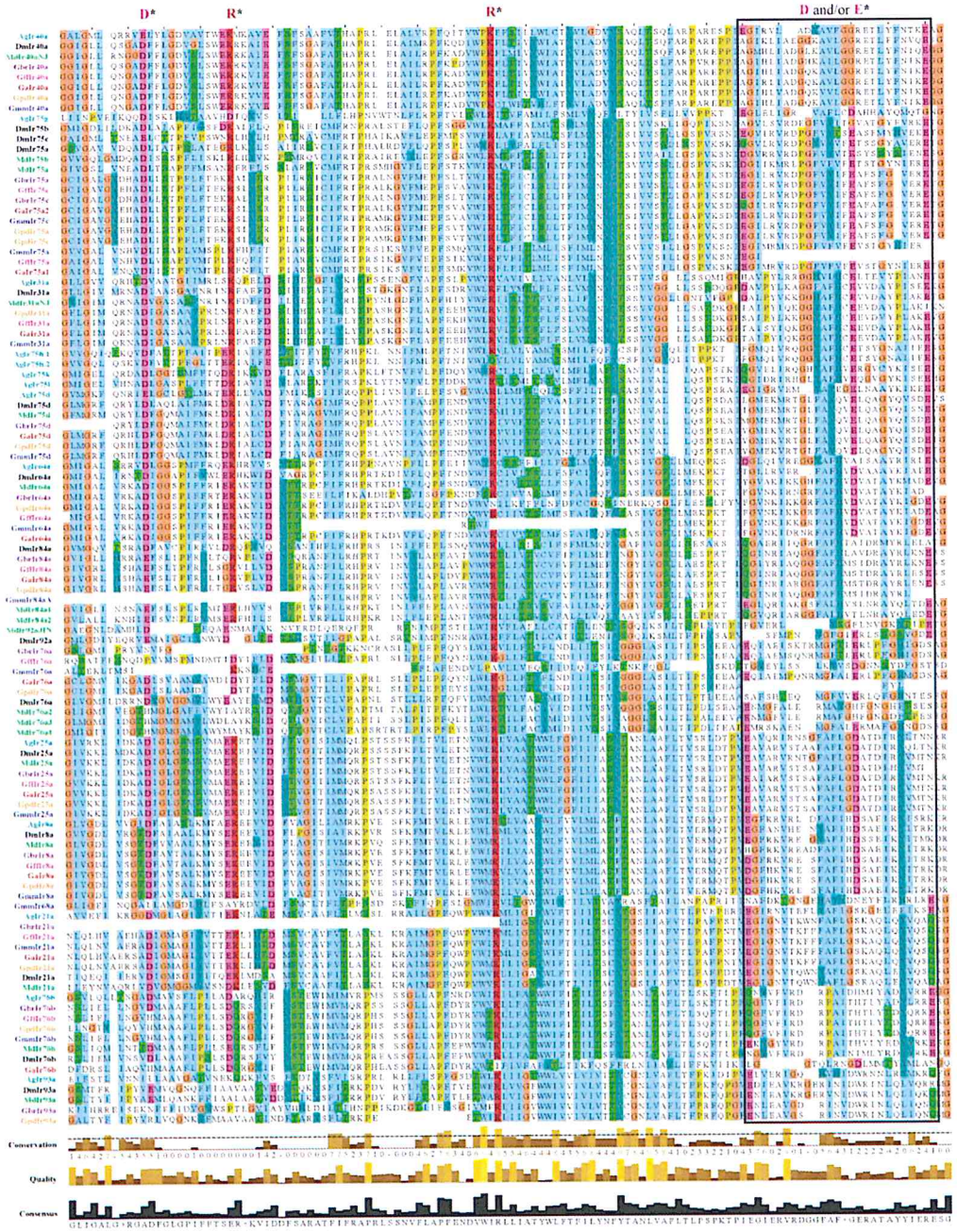


Figure S 3.4.1: Alignment of amino acid sequences of Ionotropic receptors identified in five *Glossina* species against those of closely related Diptera. Conserved residues that form ligand binding domain are highlighted by an asterisk (*)

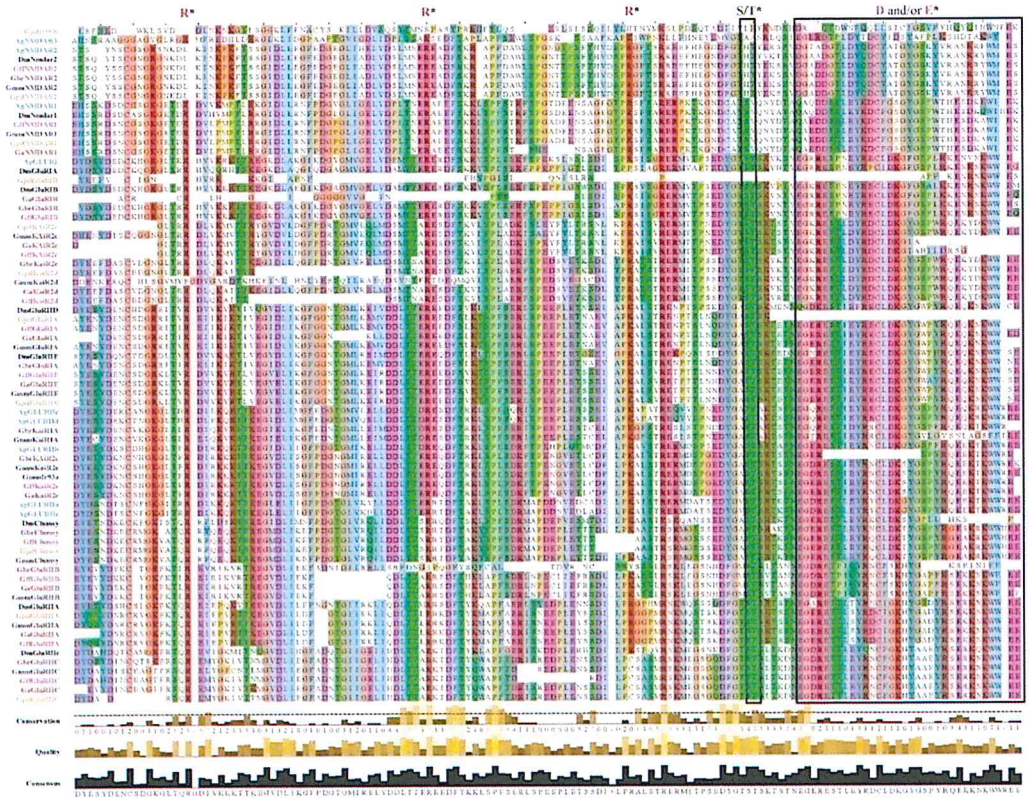


Figure S 3.4.2: Alignment of amino acid sequences of Ionotropic Glutamate Receptors (iGluRs) identified in five *Glossina* species against those of closely related Diptera. Conserved residues that form ligand binding domain are highlighted by an asterisk (*).

CHAPTER 4

4.0 Profiling of the Tsetse Fly, *Glossina morsitans morsitans* (Diptera; Glossinidae) Chemosensory Genes in Non-Olfactory Tissues

4. 1 Introduction

Glossina morsitans morsitans (Diptera; Glossinidae) is an important vector of African trypanosomes that mainly cause nagana in domestic animals across sub-Saharan Africa (Liu *et al.*, 2012). Like other insects, tsetse flies rely on visual and/or chemosensory cues to locate hosts for a blood meal, suitable mates and habitable larvipositioning sites (Vosshall, 2003).

Tsetse's chemosensation is mediated by multiprotein families. Among them, are two classes of globular proteins reffered to as odorant binding proteins (OBPs) and chemosensory proteins (CSPs) that are abundantly expressed in the insect sensilla (Shanbhag *et al,* 1995; Shanbhag *et al.,* 2001). The OBPs and CSPs transport hydrophobic semiochemicals to the sensory neurons (Gong *et al.,* 2009). These proteins have been reported in non-olfactory tissues of other insects. For instance, CSPs have been implicated in functions such as leg-regeneration, wing and larval development (Jacquin-joly and Merlin, 2004; Gong *et al.,* 2007). Similarly, expression of OBPs has been reported in tissues other than the anntenae. An example include expression of *Aedes aegypti*'s Obp22 in male reproductive organs (Li *et al.,* 2008) and a dual role of Obp10 in *Helicoverpa* species (Sun et al., 2012). Unlike OBPs and CSPs, specific neuron membrane protein (SNMPs) play a role in pheromone reception and their expression has been reported in sensory neurons of the silk moth, *Antheraea polyphemus* (Rogers *et al.*, 1997).

In addition to chemosensory proteins, insect chemical sensing also involves three types of receptors including odorant receptors (ORs), ionotropic receptors (IRs) and gustatory receptors (GRs). The GRs play a key role is taste recognition (Clyne *et al.,* 2000) and have been reportedly expressed in various tissues in *Drosophila* (Scott *et al.*, 2001). In contrast, ORs and IRs are primarily expressed in the sensory dendrites (Benton *et al.,* 2009) and the olfactory receptor neurons (Leal, 2011), which are found within the insect antennae and/or maxilary palpi (Vosshall, 2003).

G. m. morsitans, a Savannah species shows preference to warthog as its host for blood meal relative to its other hosts (e.g. ox, buffalo and human) (Liu *et al.*, 2010). Traps treated

with chemicals that mimic host odors were used to successfully control the population of Savannah tsetse sub-group species (Gikonyo *et al.,* 2003). However, control of other tsetse sub-groups such as the riverine (palpalis group) species that cause sleeping sickness in humans remains a challenge. Potentially, the host preference exhibited by different tsetse species could be due to forces of natural selection that confer fitness advantage and adaptation to an organism's niche.

An earlier study showed correlation between starvation and increased senstivity of electroantennogram (EAG) in *G. m. morsitans* and *G. tachinoides*, but no clear effect on *G. f. fuscipes* and *G. austeni* (Otter & Schutte, 1991). These findings suggest differential expression patterns of chemosensory proteins in female and male flies. Recently, identification and characterization of OBPs and CSPs in *G. m. morsitans* have provided information on their putative functions in host searching based on their expression patterns in the sensory organ. In *G. m. morsitans*, three OBPs (GmmOBP8/9 and GmmOBP14) and one CSP (GmmCSP2) were reported to have high expression in the anntenae, suggesting their involvement in host-seeking (Liu *et al.*, 2010, 2012). Further, the results of the study on expression of CSPs suggested involvement of GmmCSP1/3 in non-olfactory functions, similar to its homolog in *D. melanogaster*, DmelPebIII (Liu *et al.,* 2012).

Diet specialization has been attributed to contraction of tsetse chemosensory repetoire compared to that of other insects (Attardo *et al.,* 2014; Obiero *et al.,* 2014). Nevertheless, the conserved orthologs have shown close phylogenetic relationship with those in *Drosophila*, suggesting conservation of their functions. Determination of expression profiles for tsetse chemosensory genes is an important step in understanding their functions, differences across species and their potential application in development of molecular-based control approaches. To date, expression patterns of tsetse's chemosensory related genes has been described in olfactory organs and whole bodies (Liu *et al.*, 2010, 2012), but not in specific non-olfactory tissues such as salivary glands and reproductive organs.

This study hypothesized that some of the annotated chemosensory proteins may be involved in non-olfactory functions such as development and reproduction and thus would have high transcript abundance in the corresponding non-olfactory transcriptomes. To investigate this hypothesis, *in silico* transcriptome analysis of the *G. m. morsitans* chemosensory genes (Liu *et al.*, 2010, 2012; Obiero *et al.,* 2014) was carried out in non-olfactory tissues using CLC Genomics workbench 8 (CLC Bio, Cambridge, MA). Findings of this study will inform putative functions played by chemosensory proteins in tsetse and how they can be exploited to control tsetse populations.

4.2 Materials and Methods

4.2.1 Transcriptome Data

The transcriptome data used in the current study were obtained from the following previous studies conducted in Aksoy's laboratory, school of public health, Yale univeristy: wild type and aposymbiotic larvae (Weiss *et al.,* 2012), male testes and accesory glands (Attardo., unpublished data), and parasitised and uninfected female salivary glands (Telleria *et al.,* 2014), and dry and lactating females (Benoit *et al.*, 2014).
The number of reads for each transcriptome were as follows: N= 42,325,367 for dry (non-lactating) females, N= 42,085,623 for lactating females, wild-type larvae; N= 50,407,071, Apo-larvae, N= 51,751,183, testes; N= 131,973,344 and N= 116,622,394 for male accessory glands.

4.2.2 Retrieval of Gene Sequences

Nucleotide sequences of all chemosensory genes annotated in *G. m. morsitans*, OBPs (n=30), GRs (n=14) ORs (n=46), IRs (n=30), SNMPs (n=2) and CSPs (n=5) were retrieved from the VectorBase database (Lawson et al., 2009). All the genes were renamed after their best matching orthologs in *D. melanogaster* for easier functional comparison. *Drosophila* was chosen due to its close phylogenetic relationship established with *Glossina* (Liu et al., 2010; Obiero *et al.*, 2014) and the fact that a lot of functional studies have been conducted in its chemosensory genes (Robertson *et al.,* 1999; Zhou *et al.,* 2009; Isono *et al.,* 2010). The best matching ortholog was identified using BLASTp searches (Altschul *et al.*, 1997) on non-redundant NCBI database using an e-value threshold of 0.001. For the *G. m. morsitans* IRs, there were no published names at the time this study was done and thus their *Drosophila* homolog names were adopted.

4.2.4 Analysis of Transciptome Data

The quality of the transcriptomes was verified using FastQC software http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. The reads were then mapped onto the annotated gene sequences using CLC Genomics Workbench 8 (CLC Bio, Cambridge, MA) allowing only single/unique match per read with an identity of 80% which is the default percentage identity cut-off used in CLC Genomics Workbench, RNA-Seq analysis tool. The transcript abundance was determined using reads aligned per kilobase

mapped (RPKM) (Mortazavi *et al.,* 2008). Only genes that were supported by at least 100 unique reads were considered for purposes of reporting. Numbers (1-6) were used to represent the RPKM values according to scale. Genes with RPKM values 1 :<=20,000, 2: 20,001-40,000, respectively were considered to be lowly abundant, those with RPKM ranging between 3:40,001-60,000 were considered to have an average abudance while those with RPKM value ranging between ,4: 60,001-80,000, 5: 80,001-100,000, and 6 >100,000 were considered to be highly abundant.

4.3 Results

Expression values of abundant genes (supported by at least 100 unique reads) are summarized in Figure 4.1. There were no values of expression recorded in either the parasitized or uninfected salivary glands. None of the two SNMPs were expressed in any of the evaluated data sets.

The OBPs showed diverse expression patterns. Among them, GmmObp44a, GmmObp99b/c were abundant in all of the analyzed datasets. The three were highly expressed in dry and lactating females. On the other hand, GmmObp99d showed high expression in aposymbiotic larvae and male testes. In contrast, GmmObp8a and GmmObp19d showed higher expression and average expression in dry and lactating females respectively. Further, three OBPs (GmmObp28a, GmmObp83g/cd) were only found to be abundant in the larval tissues. GmmObp83g showed average expression in wild type larvae, while GmmObp83cd showed low expression in the larval transcriptome. The GmmObp28a showed high expression in wild type and low expression abundance in aposymbiotic larvae. Additionally, four OBPs (GmmObp19, the two copies of GmmObp56e, GmmObp56d and GmmObp56i) showed expression in the male reproductive organs (testes and acessory glands).The expression of GmmObp56e and GmmObp19 was high both in the male testes and accesory glands while that of GmmObp56i and GmmOb56d was lower in the testes as compared to accessory glands.

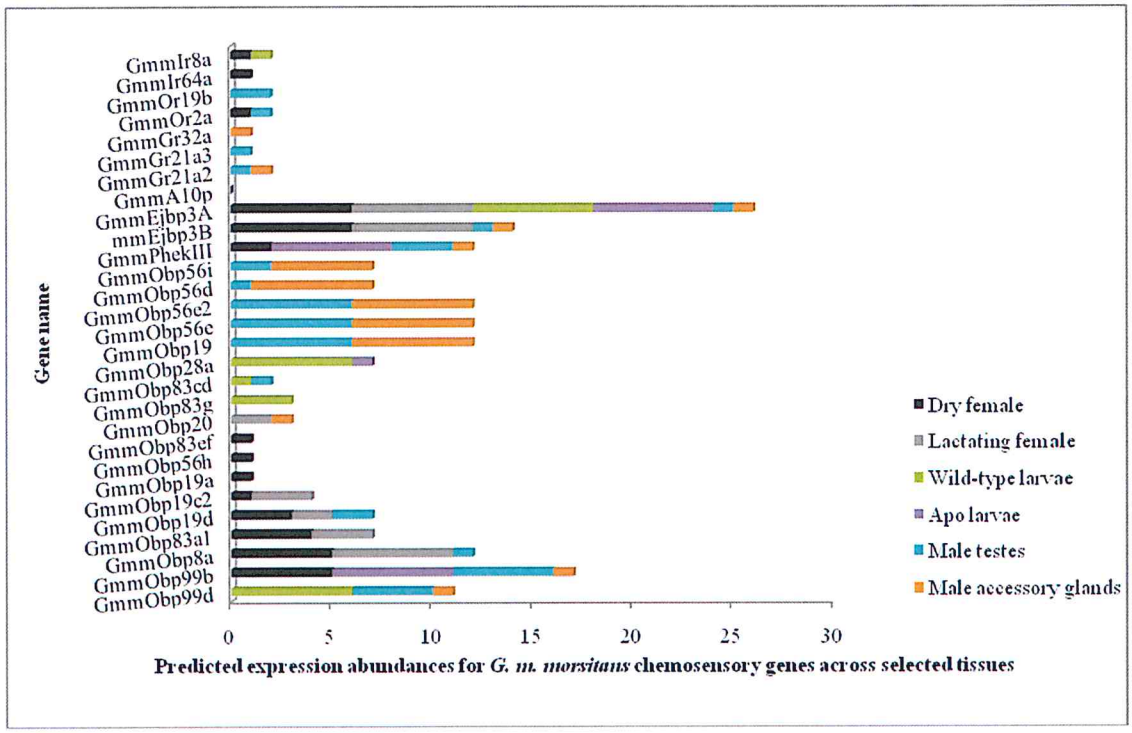


Figure 4.1: A summary of the abundance of expression of *G. m. morsitans* chemosensory genes in non-olfactory tissues

The two copies of ejaculatory bulb protein 3 (GmmEjbp3A and GmmEjbp3B) showed high expression in both dry and lactating females and low expression in testes and in accessory glands. In contrast, only GmmEjbp3A showed high expression abundance in both wild type and aposymbiotic larvae. On the other hand, GmmPhekIII showed average expression abundance in male testes, low expression in the male accessory glands and dry females and high expression in the Aposymbiotic larvae.

Low expression abundance of two copies of GmmGr21a (GmmGr21a2/3) was observed in male testes. Similar abundance was observed for GmmGr32a and Gr21a2 in the male accessory glands. On the other hand, two GmmIr64a and and a co-receptor GmmIr8a showed low expression abundance (RPKM <20,000) in dry females. The latter also showed low expression in wild type larvae. Like the IRs, only two ORs (GmmOr2a/19b) qualified as expressed based on our inclusion criteria. GmmOr2 had low expression in dry females and male tsetse while GmmOr19b showed low expression in male testes.

4.4 Discussion

Results of the current study demonstrate expression of CSPs in the non-olfactory tissues of *G. m. morsitans*, a result that is consistent with reports in other insects such as *Drosophila* and *A. aegypti* (Shanbhag *et al.*, 2001; Li *et al.*, 2008). In a recent study, Liu and colleagues (2012) described differential expression patterns of CSPs in male and female tsetse antennae which linked them to host-searching by female tsetse (Liu *et al.*, 2012). Their results also showed ubiquituos expression of GmmCsp3 throughout the body which corresponds to expression of GmmEjbp3 in all the non-olfactory tissues evaluated here (Fugure 4.1). Our observation thus, provides further evidence of multi-functionality of some insect CSPs. Of interest is GmmA10p which showed low abundance in dry females unlike in the lactating females. This gene has been described as antennal-specific both in *Glossina* (Liu *et al.*, 2012) and in *Drosophila* (McKenna *et al.*, 1994; Pikielny *et al.*, 1994). Potentially, the expression of GmmA10p in dry females implies roles in host seeking after parturation. It should be noted that the dry and lactating female transcriptome libraries were prepared with their heads intact and thus the expression of GmmA10p in dry females could be linked to the antennae.

On ther other hand, the expression patterns of the OBPs were observed to be similar in both dry (non-lactating) and lactating females. Majority of the OBPs with high abundances in the dry and lactating transcriptomes were those reportedly enriched in olfactory organs of *G. m. morsitans* (Liu *et al.*, 2010). Given that both the lactacting and dry female datasets were processed with their heads intact, it is likely that the observed expression OBPs patterns in the two groups is cumulative of the olfactory organs, the body, and the larvae in case of the lactating females. Among the genes showing high abundance were GmmOBP99b and GmmOBP44a, which have been reported to decrease after acquisation of a blood meal, thus confirming their active role in regulating the feeding process (Liu *et al.*, 2010). Unlike in *Drosophila* where expression of Obp99b is confined to adults, its high abundance was noted in aposymbiotic tsetse larvae. Further, high expression of GmmObp99b and GmmObp44a in the male testes suggests their involvement in mating. In *Drosophila*, OBP99b, which is known as turn-on-specificity (Tsx), is usually expressed more in males than in females, and has been attributed to affect mating behavior of females while expressed in high amounts (Fujii and Amrein, 2002). Similarly, Obp44a has been reported in male reproductive tissues in *Drosophila* (Yamamoto and Takemori, 2010), suggesting its involvement in reprodction. Similar to *Drosophila*, expression of GmmObp99c was abundant in both adult and larvae. Its

expression in lactating females is consistent with that observed in mated *Drosophila* females (McGraw *et al.,* 2004) and its expression in dry females is indicative of its involvement in olfaction, more so, host seeking.

On the other hand, GmmObp83cd was found to be present in larvae, an observation that is in agreement with previous reports by Liu and colleagues (2010). Taken together, these data implicates the involvement of GmmObp83cd in larval development. In contrast, the abundance of GmmObp83ef and GmmObp19c2 was not observed in larvae in this study. However, the two genes were reported in larval tissues using qRT-PCR (Liu *et al.,* 2010), probably because qRT-PCR is more sensitive compared to the *in silico* approach used in the current study. Further, GmmObp8a was highly abundant compared to GmmObp19d (with average expression abundance) in dry and lactating females. The two genes were reported to be enriched in the heads of *C. capitata* (Gomulski *et al.,* 2012), and have been implicated in hexanol response and nutrient sensing and/or starvation stress in *Drosophila*, respectively (Arya et al., 2010). Similarly, their abundance in dry and lactating tsetse females suggests their role in nourishment both for the female and developing larvae. On the other hand, GmmObp83a1 was only found in dry and lactating females. The observed abundance could be linked to antennae which supports the findings of Liu and colleagues (2010) that this gene is olfactory specific.

Although not clearly understood, the high expression of GmmObp28a in wild type larvae potentially suggests its involvement in larval development. GmmObp28a in *Drosophila* has been linked to mitigation of bitter tastant intake by that adult flies *(Swarup et al.,* 2011). Conversely, Swarup *et al.,* (2011) observed that low expression of Obp56h increased the intake of bitter tastants such as quinine. In the current study, though lowly abundant, the expression of GmmObp56h was confined to the dry (non-lactating females). Given that the dry female libraries were prepared 48 hrs post partutation, their need to feed may be responsible for the abundance of GmmObp56h, thus suggesting its involment in host-seeking.

Expression of GmmObp56d/e/i and GmmObp19 in both male testes and accessory gland is suggestive of their participation in reproduction functions. Our observation is similar to that reported in *Drosophila*, where the expression of these OBPs was observed in testes, accessory glands and ovaries (Chintapalli *et al.*, 2007).

Unlike the chemosensory proteins, majority of chemoreceptor encoding genes did not show abundance in most of the evaluated tissues. The low abundance of Gr21a observed in male reproductive organs is not clear given that this gene is associated to CO_2 detection in

tsetse (Torr *et al.*, 2006). On the other hand, Gr32a is linked to pheromone detection and male aggression in *Drosophila* (Andrews *et al.*, 2014). The expression of Gr32a in *Drosophila* has been reported in the mouth and legs (Fan et al., 2013; Andrews *et al.*, 2014); the observed abundance in the accesory glands of *G. m. morsitans* in the current study may imply that the gene is also expressed in low quantities in male acessory glands. The expression of GRs observed in this study are different from what has been reported in insects such as *Drosophila*. High abundance of Gr22d/e has been reported in *Drosophila* larvae (Zhou *et al.*, 2009). Homologs of most *Drosophila* GRs including Gr22d/e were not found in *Glossina* (Obiero *et al.*, 2014). Similarly, low expression abudance of two ORs and two IRs observed in the non-olfactory tissues may imply their pleitropic functionality. Unlike in this study, expression of some ORs has been reported in non-olfactory functions in *Drosphila*. They include: Or13a, Or45a, Or65c, Or67a and Or85b, which showed differential expression abundances in virgin and mated female flies (Zhou *et al.*, 2009). Further, expression of the two IRs observed in dry females could be linked to the antennae. Homolog of Ir64a in *Drosophila* has not shown any evidence of antennal expression but shows close phylogenetic relationship with antennal IRs (Croset *et al.*, 2010). On the other hand, Ir8a has been characterized as a co-receptor that is ubiquitously co-expressed with specifc ionotropic receptors (Rytz *et al.*, 2013).

The expression profiles observed in tsetse's non-olfactory tissues evaluated in this study suggest pleitropic functions of OBPs and CSPs. In the cases, expression abundances of tsetse OBPs and CSPs are consistent with what has been reported in *Drosophila* among other insects. For example, expression of Obp56d-i in male reproductive organs provides more evidence in their involvement in reproduction as documented in *Drosophila* (Zhou *et al.*, 2009). Similarly, abundance of Obp99b in adult females and larvae is an indication of its participation in unrelated functions. In contrast, we did not find similar expression patterns for the chemoreceptors including GRs, IRs and ORs. This could imply that tseste's chemoreceptors unlike chemosensory proteins have diverged to mainly play olfactory roles. Otherwise, it could mean that the level of chemoreceptor abundance was too low for consideration based on our inclusion criteria. The latter was observed in *Drosophila* where the magnitude of GRs and ORs was reported to be significantly lower compared to that of OBPs (Zhou *et al.*, 2009).

4.5 Conclusions

This study provides expression profile of tsetse's chemosensory genes in non-olfactory tissues. Our findings suggest multiple functionality of OBPs and CSPs; playing pleitropic roles ranging from development, reproduction and olfaction. Further, this study emphasizes the need to conduct functional studies on the tsetse chemosensory genes to evaluate their suitability in development of novel control stategies. For example, knock-down of genes such as Obp99b putatively involved in larval development could provide more insight on how control of tsetse could be achieved through arresting larvae development. In addition, targeting genes that are actively involved in reproduction such as Obp56d-i could reduce tsetse populations. These strategies combined with baited-traps that expliot host odor responses would achieve better results in controlling tsetse populations.

CHAPTER 5

5. 0 Binding Properties of *Glossina* Homologs of Obp83a1: Olfactory Specific Odorant Binding Protein

5.1 Introduction

Recognition of airborne semiochemicals mediated by various protein families is crucial for insect survival and reproduction. Insect odorant binding proteins (OBPs) are believed to be the first biological molecules that interact with hydrophobic semiochemicals, relaying them to their specific odorant receptors, which in turn initiate signal transduction (Leite *et al.*, 2009; Venthur *et al.*, 2014). OBPs are characterized as small globular proteins that weigh between 13-16 kDa and mainly found in high concentration in the insect's senillium lymph (Leal, 2011; Venthur *et al.*, 2014). Some OBPs, else known as pheromone binding proteins (PBPs) are involved in selective binding of pheromones (Venthur *et al.*, 2014) while others are known to bind a wide variety of odorants, hence referred to as general odorant binding proteins (GOBPs) (Honson *et al.*, 2005; Zhou, 2010). The number of known OBPs has greatly increased with the advent of high throughput sequencing technologies. Nevertheless, only few studies have been carried out to determine their functions.

Numerous chemical ecology studies have been undertaken in the past to identify host odors used in tsetse fly control (Voskamp & Otter, 1999). Among them, 1-octen-3-ol (octenol) 4-methylphenol (p-cresol), 3-n-propylphenol, CO_2, and acetone were shown to increase fly catches in the field (Hall *et al.*, 1984; Späth, 1995; Voskamp *et al.*, 1999). On the other hand, compounds such as acetophenone, lactic acid, 2-methoxyphenol and waterbuck derivatives depicted repellency against tsetse flies (Vale *et al.*, 1988; Torr *et al.*, 1996; Voskamp *et al.*, 1999; Gikonyo *et al.*, 2002). Voskamp and colleagues (1999) identified more than 50% of the tsetse's antennae cells as generalists that respond to more than one compound suggesting existence of multiple receptor sites on the olfactory cells.

Despite the identification of various host odors applied in tsetse control, molecular mechanism involved in their interaction with chemosensory proteins remains unknown to date.

Various forces including, electrostatic forces, van der Walls forces, hydrogen bonding and hydrophobic interactions are believed to participate in the binding of odorants/ligands onto the OBPs (Venthur *et al.*, 2014). Study of such protein-ligand interactions is necessary

to understand their biological roles as well as in drug-discovery. In the past, expensive approaches have been used to define protein specificities and to determine their active sites. These include X-ray crystallography (Smyth and Martin, 2000) and nuclear magnetic resonance (NMR) (Wüthrich, 1990) which are based on protein structural information (Venthur et al., 2014). As an alternative, homology modeling is becoming popular for determining the tertiary (3D) structures of the proteins (Paas *et al.,* 2000; Venthur et al., 2014). Homology modeling relies on a target protein (whose structure has been determined experimentally) to predict the conformation of a new protein sequence (Bishop *et al.,* 2008). Once a protein structure has been determined, its binding affinity to its selected ligands can then be determined. This is achieved through molecular docking (Morris and Lim-Wilby, 2008) and molecular dynamic simulations (Hansson *et al.,* 2002) that determine preferred binding conformations and the dynamic characteristics of protein complexes respectively.

There has been heightened research in characterization of OBPs in disease vectors and crop pests with an aim of targeting them for control through behavior manipulation *(Leite et al.,* 2009). Recent annotation of OBP in five tsetse genomes (Liu *et al.*, 2010; Macharia *et al.*, 2016), has yield four paralogs of Obp83a in each of the tsetse genomes. The paralogs were found to be under positive selection (see Chapter 1) supporting their importance in tsetse's olfaction. Earlier, Liu and colleagues (2010) determined the expression profiles of *G. m. morsitans* OBPs under different starvation periods. Their results suggested that Obp83a1, (previously named GmmOBP8 - Liu *et al.,* 2010) is an olfactory specific protein that plays a key role in host seeking (Liu *et al.*, 2010). As such, we hypothesized that: (i) *Glossina* homologs of Obp83a1 have varied ligand binding properties to tsetse attractants and/or repellents (ii) these binding properties are responsible for observed differential responses to known attractants and/or repellents exhibited by different *Glossina* species. To investigate these hypotheses, we determined the 3D structures of *Glossina* OBP83a1 homologs and compared their binding affinities to known tsetse attractants and repellents through molecular docking and simulations. Findings of this study will inform on structural features important for binding potential ligands complementary to tsetse's OBPs. The study serves as a demonstration of application of *in silico* docking in genomics research to investigate interactions at molecular level.

5.2 Materials and Methods

5.2.1 Homology Modeling

To predict the 3D structures of *Glossina* Obp83a1, four best structures with the highest similarity to sequences of Obp83a1 annotated in the genomes of *G. austeni, G. brevipalpis, G. f. fuscipes, G. m. morsitans* and *G. pallidipes* (Liu *et al.,* 2010, Chapter 1) were selected as templates. The HHpred web server (Biegert and Lupas, 2005) was used to search for closest structural homologs in Protein Data Bank (PDB)- release 70 (Berman, 2000) The homology models were predicted using Modeller (Eswar *et al.,* 2008) available at: http://toolkit.tuebingen.mpg.de/hhpred. It should be noted that the Modeller was preferred for modeling because it combines spatial restrains with stereochemistry to model proteins including loop regions (Bishop *et al.*, 2008). The four selected structural templates included 3R72- *Apis mellifera* Obp5, 3Q8I- *An. gambiae* Obp4, 3V2L- *An. gambiae* Obp20, 100H- *D. melanogaster* LUSH (Obp76a) (See Supplementary, Table S5.1).

5.2.2 Structure Validation

To determine the quality of *Glossina* Obp83a1 3D structures, the models generated as described above were analyzed using PROCHECK (Laskowski *et al.,* 1993) and PROSA (Wiederstein and Sippl, 2007). PROCHECK evaluates the general stereochemistry of the protein while PROSA checks for potentials errors in the 3D structures using Z-score as a means of scoring. To identify regions of similarity/ dissimilarity, multiple sequence alignment of the Obp83a1 amino acid sequences against their predicted structures was done in PROMALS3D (Pei *et al.,* 2008). Further, similarity across the five structures was determined through their superimposition using MATRAS (Kawabata, 2003) and viewed using Jmol (The Jmol Team, 2007). Regions encoding the definitive domain of the proteins (PBP/GOBP) were determined using Delta BLAST searches against Conserved Domain Database (CDD) (Marchler-Bauer *et al.,* 2005) and compared among the five *Glossina* sequences (Supplementary, Table S5.2).

5.2.3 Ligand Selection

To perform docking experiments, the following five attractants were selected as test ligands: 4-methylphenol (p-cresol), phenol, acetone, 3-n-propyphenol and 1-octen-3-ol. These compounds were selected because it has been demonstrated that they have strong

effects in increasing fly catches (Voskamp *et al.,* 1999). In addition, these compounds are currently used in tsetse control by The Pan African Tsetse and Trypanosomiasis Eradication Campaign (PATTEC). Similarly the following repellents: 2-methylphenol, acetophenone (Torr et al., 1996) and nine waterbuck odor derivatives (Table 5.2, Gikonyo *et al.*, 2002) which have been reported as potential repellents and whose structures were available in ChemSpider were used as test repellents. The molecular structures of the ligands were downloaded from the ChemSpider database (Pence and Williams, 2010), a free online resource that provides access to known chemical compounds. Open Babel version 2.3.2 (O'Boyle *et al., 2011*) was used to convert the chemical structures from Mol2 format into PDB format for compatibility with the docking software.

5.2.4 Binding Site Analysis and Ligand Docking

The binding sites of the predicted OBP83a1 3D structures were determined using web-based Metapocket version 2.0 (Huang, 2009), which uses a combined approach in prediction of the sites. The predicted binding sites were visualized in PyMOL version 1.6.x (DeLano, 2002) and their corresponding coordinates rewritten into configuration files used for docking. To reduce the computational cost, the non-binding N-terminal amino acid residues were stripped off from the receptors prior to docking. Scripts within AutoDock tools were used to optimize the receptor and ligand inputs prior to docking. Optimization of the inputs involved removal of solvent molecules and co-factors from the receptor, and addition of missing atoms and partial charges. To predict the orientation of ligands in the receptor binding sites, rigid docking was carried out using Autodock Vina version 1.1.2 (Trott and Olson, 2010) limiting the number of binding modes to 1 (with lowest scoring pose). One attractant (3-n-propylphenol) and two repellents (d-octalactone and (E)-6, 10-Dimethyl-5-9-undecadein -2-one) showing lowest scores were selected for further analysis. Their amino acids residues participating in docking was visualized in LIGPLOT (Wallace *et al.,* 1995).

5.2.5 Molecular Dynamic Simulations

To resolve the interaction of the three selected ligands with the five homologs of Obp83a1 receptors, molecular dynamic simulations were carried out using GROMACS version 4.5.7 (Pronk *et al.,* 2013) applying the AMBER 96 force field. Starting coordinates of Obp83a1-ligand complexes were extracted from the docked complexes. Protonation state of all ionisable amino acid groups was assigned to pH 5.0, which is an average pH at which selective binding of some insect OBPs has been reported (Katre *et al.*, 2009). The protonation

state was achieved using pdb2gmx script within GROMACS and the ligands topologies were automatically parametized using ANTECHAMBER program implemented within ACPYPE package (Silva *et al.,* 2012). Solvation of the Obp83a1-ligand complexes was done using water in a triclinic box of dimension 17.5 Å and the SPC water model (Berendsen *et al.,* 1981) applied to maintain explicit solvation. Water molecules were replaced using counter Na^+ and Cl^- ions to achieve a neutral system. Energy minimization was performed in a vacuo of 1000 steps to a tolerance level of 1000 kJ mol-1 to avoid steric clashes. The minimized complex was then equilibrated for 200 pico seconds (ps) under the NVT conditions where the temperatures were fixed to 300K. Afterward, equilibration was performed under the NPT conditions where reference pressure was set to 1 bar in all directions with a pressure coupling time of 2.0 ps. The equilibration was achieved through canonical sampling using the velocity-rescaling thermostat (Bussi *et al.,* 2007) and the Parrinelo-Rahman barostat (Parrinello and Rahman, 1981) respectively. Molecular dynamic simulations of up to 15ns were conducted with integration time of 2 femtoseconds (fs) under constant temperature and pressure. LINCS algorithm (Hess et al., 1997) was applied to constrain the bond lengths during the simulation process. Electrostatic and van der Waals interactions were determined using particle-mesh Ewald algorithm (Darden *et al.,* 1993).

5.2.6 Analysis of Simulations

The trajectories produced in the molecular dynamic simulations were visualized in Visual Molecular Dynamics (VMD) version 1.9.2 (Humphrey *et al.,* 1996). In built GROMACS analysis tools were used to calculate root mean square deviations (RMSD) and the root mean square fluctuations (RMSF) of the protein backbone. Similarly, the potential and kinetic energies of the system, temperature and pressure were determined. Further, g-mmpbsa tool (Kumari *et al.,* 2014) which employs Molecular Mechanics–Poisson Boltzmann Surface Area (MM-PBSA) approach was adopted in calculation of bonded and non-bonded interaction energies to aid in understanding the strength of interactions at the receptor-ligand interface.

5.3 Results

5.3.1 Similarity Analysis of *Glossina* Obp83a1 Homologs

Similarity searches against the conserved domain database yield high identities (>60%) among the *Glossina* Obp83a1 homologs (Table 5.1). All the five homologs had a

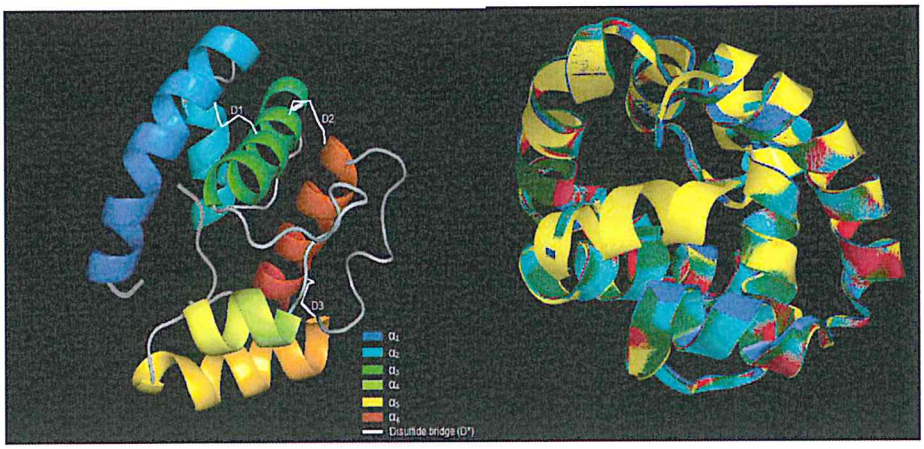
complete PBP-GOBP domain signature made up of 112 amino acid residues, majority of which are hydrophobic. Their closest homologs were identified in three dipteran species including *S. calcitrans, D. melanogaster* and *M. domestica.*

Table5.1: Domain structure and BLASTp analysis of *Glossina* OBP83a homologs

Gene id	Length (aa)	Domain co-ordinates	Closest homolog	e-value	Coverage (%)	Identity (%)
Gau_Obp83a1	147	32-144	Pheromone binding protein6: *S. calcitrans*	3e-33	74	77
Gbr_Obp83a1	158	40-152	Pheromone binding protein6:*S. calcitrans*	2e-34	80	73
Gff_Obp83a1	148	32-142	Obp83a: *D. melanogaster*	1e-33	62	63
Gmm_Obp83a1	150	32-144	Pheromone binding protein 6:*M. domestica*	7e-37	76	76
Gpd_Obp83a1	156	32-144	Pheromone binding protein 6: *S. calcitrans*	7e-34	74	74

5.3.2 Overall structures of *Glossina* OBP83a1

Homology modeling yielded similar structures for the five Obp83a homologs, each of them having six alpha (α) helices and three disulphide bridges: D1, D2 and D3 between the helices α1/α3, α3/α6 and α5/α6, respectively (Figure 5.1, panel A_1). Hydrophobic amino acids residues were found to be closely packed, thus forming the binding cavities of the proteins. The similarity was further confirmed through superimposition (Figure 5.1, panel A_2). Additionally, alignment of their secondary structures with corresponding amino acid sequences showed highest variation occurring between *G. brevipalpis* and *G. m. morsitans* orthologs (in regions highlighted in red: Figure 5.1, panel B). The homologs of the other three species including *G. austeni, G. f. fuscipes* and *G. pallidipes* showed the least variation.




```
Conservation:                    9 55  5     5       55   5 55599 5999 5         9995  55 5 5 95 995 99 5995999999995995999999995999   5999 9559 959    59
GbrObp83a1.pdb_chain__s00    1 MTLSVKC-----CFLTVYTILIVLLSTWTRAQQPRRDDE---------YPPPQILKLARPFHDICVEQTGVKEEAIKEFSDGEIHEDEALKCYMNCLFHEFDVVDDNGDVHLEKLFSKIP  106
GffObp83a1                   1 MLL--KCDWLLLIFLTKLSLLIRIHST-----PLRRDNN---------YPPKELITGAKPLHEKCVKETGVTEEAIKEFSDGEVHEDEALKCYMNCFFHELGAVDDKGDVHLETLNLIMP  104
Gff_Obp83a1.pdb_chain__s0    1 MLL--KCDWLLLIFLTKLSLLIRIHST-----PLRRDNN---------YPPKELITGAKPLHEKCVKETGVTEEAIKEFSDGEVHEDEALKCYMNCFFHELGAVDDKGDVHLETLNLIMP  104
GmmObp83a1                   1 MKKY--H-----IYIVTFAITLLMSFGLNNAQKPRRDEN---------YPPPDFLKSFKIIHDVCVEKTGATEEAIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDAGELHFEKLVRMIP  104
GaObp83a1                    1 MKKY--H-----IYYVTFTITLLLSFGLNNAQKPRRDEN---------YPPPDFLKSFQIIHDVCVEKTGATEEAIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDAGELHFEKLVRMIP  104
Ga_Obp83a1.pdb_chain__s00    1 MKKY--H-----IYYVTFTITLLLSFGLNNAQKPRRDEN---------YPPPDFLKSFQIIHDVCVEKTGATEEAIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDAGELHFEKLVRMIP  104
Gmm_Obp83a1.pdb_chain__s0    1 MKKY--H-----IYIVTFAITLLMSFGLNNAQKPRRDEN---------YPPPDFLKSFKIIHDVCVEKTGATEEAIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDAGELHFEKLVRMIP  104
GpdObp83a1                   1 MEKY--H-----IYIVTFAMTLLLSFGLNNAQKPRRDEN---------YPPPDFLKSFKIIHDVCVEKTGATEEAIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDAGELHFEKLVRMIP  104
Gpd_Obp83a1.pdb_chain__s0    1 MEKY--H-----IYIVTFAMTLLLSFGLNNAQKPRRDEN---------YPPPDFLKSFKIIHDVCVEKTGATEEAIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDAGELHFEKLVRMIP  104
GbrObp83a1                   1 MKKY--------ICISTFAITLLLSFALNNAQKPRRDENMLNVLSVFQYPPPDFLKYFKILHDVCVEKTGTTEESIKEFSDGEIHEDPALKCYMNCLFHEVNVVDDNGELHYEKLKRMIP  112
Consensus aa:                        M.b...h.....h@lh.hslhl.l..shspAQ..RRDpp.........YPP.phlp.h+.hE-.CVcpTGhpEEAIKEFSDGE!HED.ALKCYMNChFHEhshVDD.G-lHhKpL..bhP
Consensus ss:                          hhhh    hhhhhhhhhhhhhhhhhhhhhhhh           hhhhhhhhhhhhhhhhh   hhhhhhhh       hhhhhhhhhhhhhh         hhhhhhh

Conservation:                      5 5       9 5  959 559559955959 5995      5
GbrObp83a1.pdb_chain__s00  107 AAIRDLLINASKGCVH--PEGDTLCHKAWWFHQCWKKADPVV------CY-K-  149
GffObp83a1                 105 GSFVDAILKPAQHCIH--PEGDTLCHKAWWFHQCWKKADPEH------YFLL-  148
Gff_Obp83a1.pdb_chain__s0  105 GSFVDAILKPAQHCIH--PEGDTLCHKAWWFHQCWKKADPEH------YFLL-  148
GmmObp83a1                 105 EPFLEMVKHIIDACESHIPKGETQCDRAWSWHVCFKQTDPVL------YF-LP  150
GaObp83a1                  105 EPFLEMVKHIIDACESHIPKGETQCDRAWSWHVCFKQTDPVI----------C  147
Ga_Obp83a1.pdb_chain__s00  105 EPFLEMVKHIIDACESHIPKGETQCDRAWSWHVCFKQTDPVI----------C  147
Gmm_Obp83a1.pdb_chain__s0  105 EPFLEMVKHIIDACESHIPKGETQCDRAWSWHVCFKQTDPVL------YF-LP  150
GpdObp83a1                 105 EPFLEMVKHIIDACESHIPKGETQCDRAWSWHVCFKQTDPVLKSIKIFYL-SN  156
Gpd_Obp83a1.pdb_chain__s0  105 EPFLEMVKHIIDACESHIPKGETQCDRAWSWHVCFKQTDPVLKSIKIFYL-SN  156
GbrObp83a1                 113 ESFMETVQHIVDACESHVPKGVDQCDRAWSWHVCFKTTDPVL------YF-LP  158
Consensus aa:                      .sh.-hlbp..p.C.p..PcG-TbCc+AW.@H.C@KphDP.h......h@.b.
Consensus ss:                      hhhhhhhhhhhhhhhh      hhhhhhhhhhhhhh
```

Figure 5.1: Homology model and alignment of *Glossina* OBP83a1. Panel A_1 shows the cartoon representation of OBP83a1 structure with six α-helices and 3 disulphide bridges visualized in PyMOL (DeLano, 2002) which are critical in forming the ligand binding sites. Panel A_2 shows superimposition of the five OBP83a1 homologs. The rainbow colours indicate aligned regions. Panel B, of Figure 5.1 shows PROMALS3D (Pei et al., 2008) sequence-structure alignment of all the five OBP83a1 homologs. The "-hhhhhhhh-" pattern in panel B represents α helices and regions colored in red contain different amino acid residues.

The quality of the modeled Obp83a1 as analyzed using PROCHECK, PROSA and Verify3D is summarized in Table 5.2. Under PROCHECK, a model is considered to be good if >90% of its amino acid residues fall within the allowed core region. A negative Z-score under PROSA analysis is indicative of high accuracy in model folding (Wiederstein and Sippl, 2007) similar to a least 65% of 3D structure is mapping correctly on its primary sequence (Eisenberg, Lüthy and Bowie, 1997).

Table 5.2: Empirical distribution of amino acid residues into defined regions (column 2), the quality of 3D protein models (Column 3) and the potentially correctly folded regions as determined by mapping the 3D structure onto its corresponding amino acid residues (column 4)

Protein Name †	PROCHECK analysis (% residues in different regions of Ramachandran plot)	PROSA analysis (Z-score)	Verify3D analysis (% residues with >=0.2 3D-1D mapping score)
Gmm_Obp83a1	92.5 core , 6.0 allowed, 1.5 general 0.0disallowed	-5.73	80.67
Gpd_Obp83a1	94.3 core,5.0 allowed,0.0 general,0.7disallowed	-5.17	73.72
Gau_Obp83a1	1.6 core, 7.6 allowed, 0.0 general, 0. disallowed	-4.34	40.14
Gff_Obp83a1	96.2 core,3.8 allowed,0.0 general,0.0disallowed	-6	79.73
Gbr_Obp83a1	96.3 core, 3.7 allowed,0.0 general,0.0 disallowed	-5.64	80.54

† Gau, *G. austeni*; Gbr, *G. brevipalpis*; Gff, *G. fuscipes fuscipes*; Gmm, *G. morsitans morsitans* and Gpd, *G. pallidipes*.

5.3.3 Obp83a1-Ligand(s) Docking

Scores of the predicted best poses for the docked ligands are tabulated in Table 5.2. Lower scores indicate low free energies for the receptor-ligand complexes and vice versa. The lowest scoring attractant (3-n-propylphenol) and repellent ((E)-6, 10-Dimethyl-5-9 undecadein-2-one) were visualized in LIGPLOT to visualized the amino acid residues participating in binding. LIGPLOT reads in both the receptor and ligand 3D structures and flattens them into a 2D structure showing the hydrogen bonds and hydrophobic interactions involved.

Table 5.3: Summary of best pose scores for different ligands docked to OBP83a1 using AutoDock Vina (Trott and Olson, 2010). The Values in bold correspond to the ligands with lowest scores

Ligand	ChemSpiderID	Docking Scores †				
		Gmm_OBP83a1	Gau_OBP83a1	Gpd_OBP83a1	Gbr_OBP83a1	Gff_OBP83a1
Attractant						
p-cresol(4-Methylphenol)	1380	-6	-6	-5.4	-4.7	-6.2
Phenol	97	-5.5	-5.5	-5.3	-4.5	-5.5
Acetone	15	-3.1	-3.1	-3.1	-2.8	-3.4

Ligand	ChemSpiderID	Docking Scores †				
		Gmm_OBP83a1	Gau_OBP83a1	Gpd_OBP83a1	Gbr_OBP83a1	Gff_OBP83a1
3-n-propyphenol	2513	**-6.8**	**-6.8**	**-6.2**	**-5.6**	**-6.7**
1-Octen-3-ol (Octenol)	23253598	-5.6	-5.6	-5.5	-4.8	-5.6
Repellent						
2-Methylphenol	13835772	-6	-6	-5.4	-4.7	-6.2
Acetophenone	7132	-6.6	-6.6	-6	-5.3	-6.6
2-Octanone	7802	-5.7	-5.7	-5.3	-4.6	-5.7
2-Nonanone	12632	-6.1	-6.1	-5.4	-5.1	-6.1
2-Undecanone	7871	-6.5	-6.5	-5.9	-5.3	-6.5
Delta- octalactone	12252	-6.2	-6.2	-5.8	-5.4	-6.2
(E)-6,10-Dimethyl5,9-undecadien-2-one	1266569	**-7.9**	**-7.9**	**-7.5**	**-6.5**	**-7.9**
(E)-2-Heptenal	7838	-5.3	-5.3	-7.5	-4.4	-5.4
Nonanal	29029	-5.7	-5.7	-5.3	-4.8	-5.7
Undecanal	7894	-6.4	-6.4	-5.9	-5.1	-6.2

† Gau, *G. austeni*; Gbr, *G. brevipalpis*; Gff, *G. fuscipes fuscipes*; Gmm, *G. morsitans morsitans* and Gpd, *G.pallidipes*

5.3.4 Molecular dynamic Analysis

Overall, the potential and kinetic energies of the each Obp83a1 homologs remained stable during the molecular dynamic production for the complexes with the three ligands (Figure 5.2). The global RMSD of the proteins varied with the binding of different ligands (Figure 5.3). Nonetheless, the RMSD of the five homologs remained below 0.5 A° which is relatively good for close homologs. The RMSD averages and standard deviation (STDEV) for the three complexes are provided as footnotes under Figure 5.2.

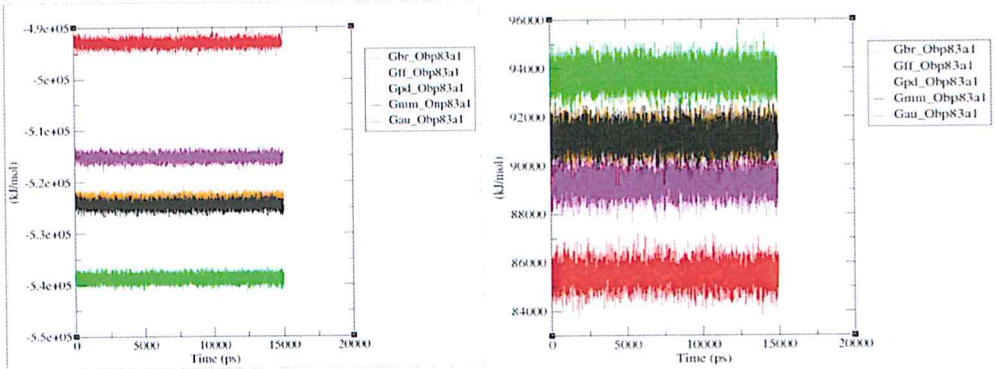


Figure 5.2: Time evolution of system energies: Potential energy of the system (A) and Kinetic energy of the system (B).

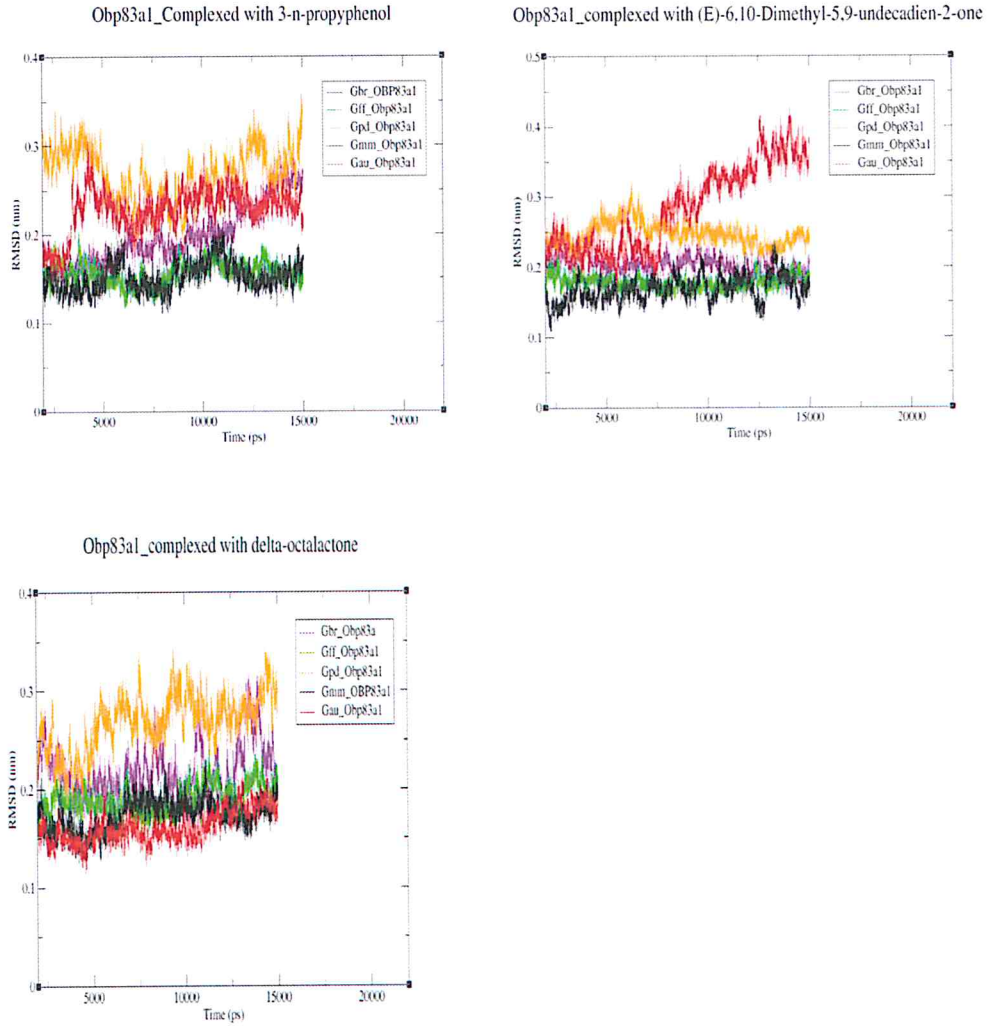


Figure 5.3: Time evolution of Root mean square deviation (RMSD). RMSD was determined for the backbone atoms of OBP83a1 complexed with 3-n-propyphenol (A), (E)-6, 10-Dimethyl-5, 9-undecadien-2-one (B) and d-octalactone (C)

Comparison of energies contributing to the binding and interaction of Obp83a1 with the ligands is summarized in Table 5.3. The van der Waals forces, electrostatic interactions and polar salvation energies were found to be the key players contributing to the binding energy of the ligands. All the five Obp83a1 homologs, showed stronger affinity for repellent (E)-6, 10-Dimethyl-5, 9-undecadien-2-one as compared to d-octalactone with Gau_Obp83a1 showing the strongest affinity with binding energy of -444.389 and Gff_Obp83a1 showing the least affinity (binding energy = -367.697). On the other hand Gpd_Obp83a1 depicted

relatively higher affinity for 3-n-propoylphenol (Table 5.3) as compared to the other four homologs. The binding energy of Gpd_Obp83a1 was to determine to be three fold that of Gff_Obp83a1, fivefold that of Gmm_Obp83a1 and Gbr_Obp83a1 and fiftyfold that of Gau_Obp83a1. Further, different amino acids at the C-terminal were identified as key contributors to the binding energies. The major players include Isoleucine (I-146) in Gau_Obp83a1, Threonine (T-149) in Gbr_Obp83a1, Glutamine (Q-135) in Gff_Obp83a1, Cysteine (C-138) in Gmm_Obp83a1 and Lysine (K-154) in Gpd_Obp83a1.

Table 5.4: Summary of interaction energies for five *Glossina* OBP83a1 homologs complexed with three different ligands: Attractant (3-n-propylphenol) and two repellents (delta- octalactone and (E)-6, 10-Dimethyl-5, 9-undecadien-2-one)

Protein Name†	Interaction Energies kJ/mol				
	van der Waal	Electrostatic	Polar-solvation	SASA	Binding
Gau_Obp83a1	-2.837	-0.587	1.229	0.000	-4.673
Gbr_Obp83a1	-79.108	-14.755	57.990	-10.844	-46.717
Gff_Obp83a1	-101.458	-27.949	69.363	-9.012	-69.058
Gmm_Obp83a1	-74.362	-18.840	60.082	-10.358	-43.475
Gpd_Obp83a1	-8.277	-285.590	65.572	-2.232	-230.577
Gau_Obp83a1	-70.112	-6.768	37.864	-10.379	-49.395
Gbr_Obp83a1	-68.542	-5.405	34.514	-10.045	-49.478
Gff_Obp83a1	-99.026	-1.124	45.239	-10.325	-65.242
Gmm_Obp83a1	-78.478	-11.940	60.663	-11.065	-40.821
Gpd_Obp83a1	-86.447	-23.158	71.539	-10.887	-48.950
Gau_Obp83a1	-88.217	-572.428	230.577	-14.329	-444.389
Gbr_Obp83a1	-82.795	-384.980	111.314	-12.068	-368.523
Gff_Obp83a1	-96.530	-588.071	331.484	-14.594	-367.697
Gmm_Obp83a1	-94.107	-566.194	270.123	-13.890	-404.050
Gpd_Obp83a1	-91.244	-356.382	136.219	-11.614	-323.015

† Gau, *G. austeni*; Gbr, *G. brevipalpis*; Gff, *G. fuscipes fuscipes*; Gmm, *G. morsitans morsitans* and Gpd, *G. pallidipes.*

5.4 Discussion

Conservation of sequence length and functional domain structure across the five homologs (Table 5.1) suggest conserved role in the binding of odorants. Their corresponding

homology models generated in this study conformed to the alpha helical structure with three disulphide bridges similar to what has been described in other insects including silk moth (Sandler *et al.,* 2000), mosquitoes such as *A. aegypti* (Leite *et al.,* 2009), the honey bee, *A. mellifera* (Lartigue *et al.,* 2004) and *Drosophila* (Kruse et al., 2003).

Except for the *G. austeni* (Gau_Obp83a1), the other four protein models surpassed the default scores for a good protein model. This included at least 90% of residues falling in the allowed core region under the Ramachandran plot (Laskowski *et al.*, 1993), over 65% of a 3D structure being mapped back onto its corresponding residues (Eisenberg *et al.,* 1997) and a negative Z-score (Wiederstein and Sippl, 2007). Nevertheless, superimposition suggested same conformation of the five models as most of the regions aligned without mismatches (Figure 5.1, panel A_2).

Molecular docking analysis showed affinity of the five OBP83a1 homologs against the 15 ligands tested in this study (Table 5.2). This suggests that *Glossina* Obp83a1 is a GOBP that does not bind selectively (Zhou *et al.,* 2009; Venthur *et al.*, 2014). LIGPLOT analysis performed on all Obp83a1complexes revealed participation of different hydrophobic amino acid residues in rigid docking (Figure S5.1. and Figure S5.2, respectively). The docking scores obtained suggest that 3-n-propylphenol is the best tsetse attractant with highest binding affinity across the five species. It is however worth noting that calculated docking scores do not necessarily correlate to the actual binding affinity of ligands. In such cases, third party scoring software could be used to independently determine the binding affinities. Dynamic simulations of Obp83a1 homologs complexed with 3-n-propylphenol revealed strong electrostatic interactions with *G. pallidipes*' Gpd_Obp83a1, which corresponds to its high RMSD variation (Figure 5.2, panel A). This is in contrast with complexes of the other four homologs from *G. austeni* (Gau_Obp83a1), *G. brevipalpis* (Gbr_Obp83a1), *G. f. fuscipes* (Gff_Obp83a1) and *G. m. morsitans* (Gmm_Obp83a1) where van der Waal forces was the key contributor of binding energy. Relatively low binding energy of 3-n-propylphenol to *G. pallidipes* (Gpd_Obp83a1) suggested its high affinity for the attractant compared to the other four homologs. This observation is consistent with earlier findings where 3-n-propylphenol was found to catch twice the number of *G. pallidipes* compared to *G. m. morsitans* flies despite the two being sympatric. On the other hand, the *G. austeni's* Gau_Obp83a1 showed least affinity for 3-n-propylphenol. This result is supportive of earlier report that *G. austeni* is less responsive to known host kairomones (except CO_2) relative to other species of the Morsitans group (Gibson and Torr, 1999). Interestingly, *G. f. fuscipes* that belong to the palplis group, known for its non-responsive behavior to host odors

had similar binding energies with *G. m. morsitans* and *G. breviplapis.* This observation rules out non-functionality of any of the tsetse Obp83a1 but leaves its non-responsive behavior unexplained. It is possible that differential response to host odors by different tsetse species is partially due to underlying unknown differences in the downstream processing of odor chemicals, rather than their ability to bind onto OBPs. Further analysis on amino acids involved in the binding of 3-n-propylphenol may provide an explanation for the differential responses observed across tsetse species, ultimately providing direction on how to optimize control methods for each species.

Simulations of d-octalactone complexes showed relatively similar RMSD variations of the five Obp83a1 homologs. However, that of Gau_Obp83a1 underwent a sharp decrease at 750 ps rising to ~0.4A° afterward. The electrostatic interactions were found to have higher contribution to the binding energies of d-octalactone that were relatively similar suggesting potency of d-octalactone against all tsetse species sampled here. Its repellency has been demonstrated against *G. m. morsitans* (Gikonyo *et al.*, 2003; Mwangi, Gikonyo and Ndiege, 2008) and *G.pallidipes* (Gikonyo *et al.*, 2002) but not in any other tsetse species. On the other hand, (E)-6,10-Dimethyl-5,9-undecadien-2-one, has been reported among the waterbuck compounds active on *G. pallidipes* (Gikonyo *et al.,* 2002) though its potential in reducing fly catches has not been reported. The waterbuck derivative, (E)-6,10-Dimethyl-5,9-undecadien-2-one showed the lowest docking score suggesting that it could be a better repellent as compared to d- octalactone. Interaction energies for (E)-6,10-Dimethyl-5,9-undecadien-2-one with OBP83a1 homologs (Table 5.3) were lower as compared to those of d- octalactone. In addition, unlike the binding of d-octalactone , which was seen to have strong van der Waals forces, electrostatic energies were seen to contribute more to the binding energies of (E)-6,10-Dimethyl-5,9-undecadien-2-one.

This study has shown high similarity in structure and relative binding energies of the five Obp83a1 homologs. However, the mechanism involved in ligand release, their transfer to corresponding odorant receptors and the speed of at which these processes occur remains unknown. In mosquitoes, change of pH has been linked to conformational changes that lead to release of bound ligand (Leite *et al.,* 2009) but has not been demonstrated in other insects. There is need to undertake protein-ligand assays to validate the results generated in this study as well as study the tsetse behavior towards dose-dependent responses of pure (E)-6,10-Dimethyl-5,9-undecadien-2-one. Further, resolution of these proteins using experimental methods may reveal differences in their binding properties.

5.5 Conclusion

The 3D structures of *Glossina* Obp83a1 homologs determined in this study have revealed high structural similarity among them. All the five homologs have active sites made up of hydrophobic amino acid residues which are involved in protein-ligand interactions. Their docking dynamics studied here support their involvement of tsetse's Obp83a1 in olfaction, specifically host seeking. Further, results of this study suggest that this protein is a generalist (GOBPs) that binds to attractants and/or repellents. More so, conformation of the proteins remains relatively stable upon binding of the ligands. The relative binding energies calculated for the selected attractants and repellents suggest differential affinity for 3-n-propyphenol across the five tsetse species. *G. pallidipes* showed highest affinity for the attractant as compared to the other four species, similar to what has been observed in the field. On the other hand, similar binding energies were determined for the water buck derived compounds across the five species with (E)-6,10-Dimethyl-5,9-undecadien-2-one depicting higher binding potential as compared to d-octalactone. Given that all the five Obp83a1 homologs showed affinity for known tsetse baits, it could be that differential responses exhibited by tsetse species arise during the downstream processing of odors. Therefore it is recommendable to undertake studies on the odor processing and signaling machinery in tsetse species to investigate this hypothesis.

Supplementary Data

Table S5.1 Template used in Homology modeling of OBP83a1

PDB code	Description	Resolution (A°)	Reference
3Q8I	*An. gambiae* OBP4 complexed with indole	2.00	(Davrazou et al., 2011)
3R72	*A. mellifera* OBP5	1.15	(Spinelli et al., 2012)
3V2L	*An. gambiae* OBP 20 bound to polyethylene glycol	1.8	(De Val et al., 2012)
100H	Complex of *Drosophila* lush with butanol	1.25	(Kruse et al., 2003)

Table S5.2: Domain structure and BLASTp analysis of *Glossina* OBP83a homologs

Gene id	Length (aa)	Domain co-ordinates start-end	Closest homolog	e-value	Percentage identity
Gau_Obp83a1	147	32-144	Pheromone binding protein6:*S.calcitrans*	3e-33	77
Gbr_Obp83a1	158	40-152	Pheromone binding protein6:*S.calcitrans*	2e-34	73
Gff_Obp83a1	148	32-142	Obp83a: *D. melanogaster*	1e-33	63
Gmm_Obp83a1	150	32-144	Pheromone binding protein6:*M. domestica*	7e-37	76
Gpd_Obp83a1	156	32-144	Pheromone binding protein 6:*S. calcitrans*	7e-34	74

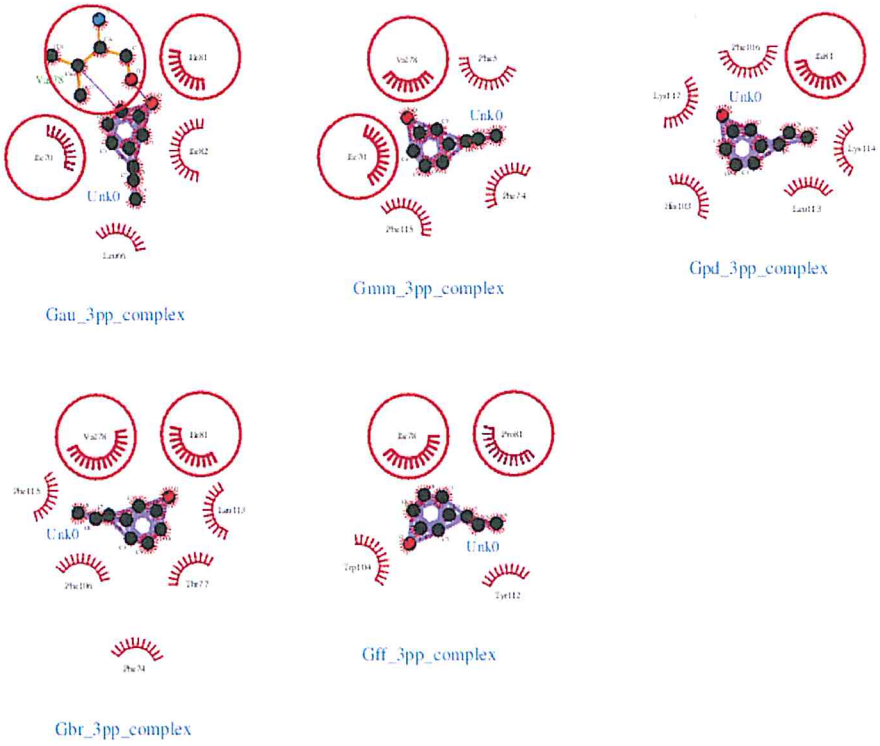


Figure S5.1: LigPLOT snapshot showing the OBP83a1 amino acid residues interacting with tsetse attractant 3-n-propylphenol

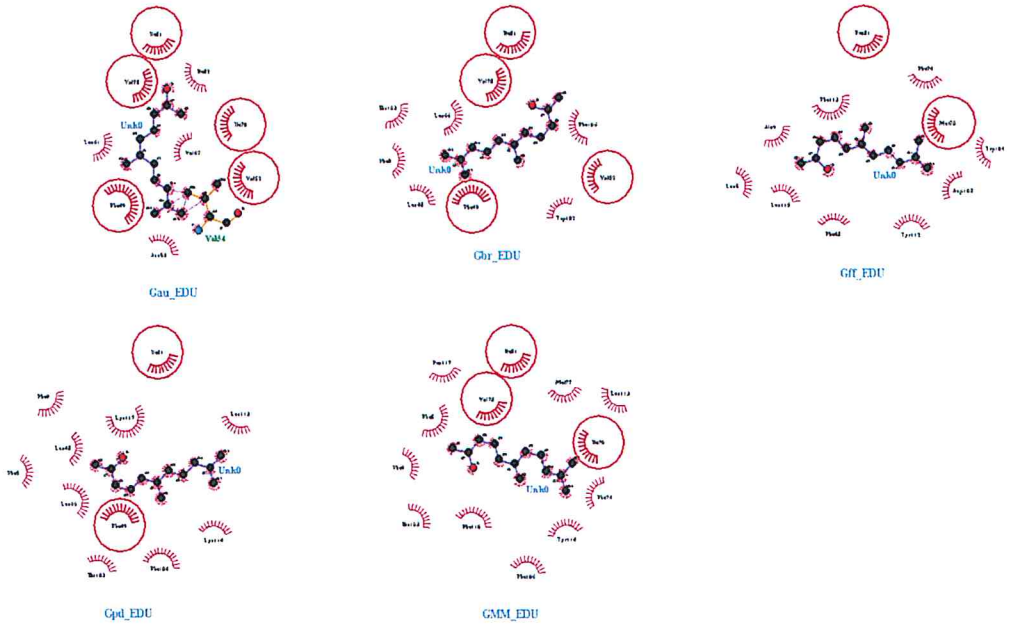


Figure S5.2: LigPLOT snapshot showing the OBP83a1 amino acid residues interacting with tsetse repellent (E)-6, 10-Dimethyl-5, 9-undecadien-2-one

CHAPTER 6

6.0 Study Summary and Future Perspectives

African trypanosomiases continue to impact negatively on human and animal health as well as income generating activities such as tourism. Eradication of African trypanosomiases through chemotherapeutic measures remains infeasible due the cost involved and the side effects of drugs. Scientists have thus resorted to using baited traps and natural repellents to mitigate host-tsetse interactions in order to minimize trypanosome infections. Despite the continued control efforts, tsetse flies have survived eradication mainly because they show varied responses towards available baits and have evolved differential tactics for locating their hosts. Furthermore, tsetse control efforts are concentrated in areas with high tsetse populations such as national parks, ignoring areas with low tsetse populations. This has resulted into re-invasion of previously cleared areas from time to time.

The work described in this thesis was designed to: (i) provide insights into the gene structure and evolutionary relationships among chemosensory proteins in five tsetse fly species in relation to other closely related dipteran insect species, (ii) determine the potential non-olfactory roles played by chemosensory genes and (iii) determine binding properties of Obp83a1, which is potentially involved in host seeking, to known attractants and repellents. Together, the data obtained from the various aspects of chemosensory gene families investigated in this study greatly enhances our understanding of tsetse biology, particularly chemosensation in the major tsetse species.

The findings of this study revealed a conserved gene repertoire across the *Glossina* genus. Notably, approximately the same numbers of chemosensory genes ware annotated in all the five *Glossina* species that were analyzed. With exception of three gustatory receptor genes (Gr66a, Gr32a and Gr58c) that were not found in *G. brevipalpis* (considered as the ancestral tsetse species). These three genes play a role in recognition of bitter compounds and their role in *G. brevipalpis* could imply its avoidance of bitter tastants. The rest of the genes showed little to no sequence variation. Although the majority of chemosensory genes that were modeled revealed the right structures, a few of these genes were either incomplete and/or fused with neighboring gene models. In an attempt to improve the incomplete gene models, manual curation was performed. Although manual curation achieved up to 70 % of correct models, the downside of this approach is that manual curation is time consuming,

especially when dealing with huge data sets as was the case in the studies described in this thesis. A potential approach to circumvent manual curation would be to improve the gene finding algorithms for tsetse among other eukaryotes.

A few chemosensory gene loci (Obp83a, Gr21a, Gr28b and GluRIIA), which depicted rapid evolution, were found to be under significant positive selection possibly to confer adaptive advantage to the specific species. These genes could therefore be key in determining suitable habitat for the different species. Nevertheless, the selection analyses conducted under this study were limited by the number of available sequences leading to discrepancy in the results obtained with the two software (PAML and HyPhy) used for selection analysis. This could be improved by inclusion of more sequences as more genomes become available. In addition, the 3D structures of proteins identified could be predicted and the specific amino acids under pressure mapped. This could help determine if they play a significant role in ligand binding and/or in protein interactions. Knowledge on how well these genes fit as

High abundance of genes in non-olfactory tissues and larvae suggest their involvement in reproduction and development. Manipulation of such genes through transgenesis or gene knock out techniques could offer an alternative approach to control of testse hence mitigating their interaction with hosts. The *in silico* RNA-Seq approach used here was faced with a challenge of the criteria of selecting appropriate cut-off values to unequivocally qualify a gene as significantly expressed. A potential approach to solve this problem would have been to conduct RT-qPCR on the identified genes to validate the results obtained. This method was however out of the scope of this thesis, and is highly recommended in future studies.

Further, functionality of tsetse OBPs was demonstrated through molecular docking of an olfactory specific protein (Obp83a1) that showed affinity with selected compounds used as baits for control of tsetse flies. The Obp83a1 homologs depicted a conserved 3D structure and binding site similar to that of close relatives such as *Drosophila*. The calculated relative binding energies suggest high affinity for attractant (3-n-propylphenol) to *G. pallidipes'* Obp83a1 homolog relative to the other four species. This could explain why the current baits are effective in trapping of *G. pallidipes* flies as compared to other species. In contrast, similar binding energies were calculated for the five homologs when bound to waterbuck odor derivatives suggesting similar affinity for the repellent. Further studies into the binding properties of tsetse attractants and repellents are necessary to provide a roadmap to improve tsetse intervention startegies.

The size of *Glossina* chemosensory protein families (CSPs, OBPs and SNMPs) was found to be relatively similar to those of close dipterans such as house fly, fruit fly and the malaria causing mosquito. On the other hand, these studies revealed that the *Glossina* chemoreceptor repertoire (GRs, IRs and ORs) is relatively reduced in relation to that of other insects. It is interesting to note that this study did not identify any sugar receptors (Gr5a, Gr64a-f, and Gr43a) in any of the five tsetse species, presumably due to the hematophagous nature of these insects. Rather, gene expansions were noted in CO_2 responsive gustatory receptor (Gr21a), supporting the fact that tsetse flies like other insects locate their vertebrate hosts through this volatile odorant. In addition, genes accrued to pheromone reception (Or67d) and larval-stress response (Or45a) were expanded across *Glossina spp.* similar to other insects. Observed expression profiles for the *Glossina* chemosensory genes in non-olfactory tissues is similar to what has been reported in other insects. This observation supports pleitropic roles of insect OBPs and CSPs.

Expression profiles predicted here hint at an unexplored alternative approach to vector control which may target different stages of tsetse such as larvae for control. The study has also provided support for potential molecular targets that could be used to improve odor-baited control approaches. For instance genetic engineering of genes involved in larval development could arrest their growth hence reducing tsetse populations. Similarly, compounds that mimic host odors could be developed to interference with tseste's reception to host odors. These approaches though costly could contribute significantly to vector control.

Taken together, the objectives envisioned at the onset of this study were sufficiently met by the methods used. Nevertheless, there are still a number of outstanding questions that came along during the course of the study. For instance, although the tsetse fly species investigated here had similar repertoire of chemosensory gene families, it still remains to be elucidated why these different tsetse species show differences in host preference. Further, the *in silico* studies suggested olfactory-specific activity of Obp83a in *G. f. fuscipes*, yet this *Glossina* species is not known to respond to any known baits. Again, this study could not provide answers to this question. It is still not clear why *G. pallidipes* responds better to known baits as compared its sympatric species and if the tsetse's down-stream odor processing is conserved across the species. Even though this study could not provide answers to these and other outstanding questions, new research hypotheses could be inferred from these questions as a means to foster further research into tsetse biology in relation to host seeking behavior.

The findings presented in this thesis necessitate undertaking further studies to reveal any functional differences among the *Glossina* orthologs. In future, focus should be on functional genomics of the genes identified in this study. It will be time worthy to undertake selection analysis with genes from more species to increase the power of detecting genes evolving under influence of natural selection. In addition, future work will include studying the odor-processing and signaling mechanisms that may be responsible for species' behavioral divergence. More studies should be undertaken to ascertain the binding properties and molecular interactions of tsetse's chemosensory proteins.

Improved bait technology for control of tsetse need to be coupled with active surveillance of trypanosome infection in order to ensure timely treatment as well as provide guided control strategies. The results of this study may prompt a change of tactic in vector control management. Compounds inhibiting binding of tsetse attractants should be developed and used to mitigate tsetse-host interactions. Similarly, genetic engineering of key chemosensory proteins involved in non-olfactory functions should be explored for their suitability in controlling tsetse populations. Targeting the identified targets for transgenesis is a feasible option in this case. However, its application remains limited due to the public opinion and regulation by government agencies. It is therefore important for scientists to work together with relevant authorities in policy making, provision of funds and coordination of control efforts.

Finally, the knowledge garnered in this study could be extended to combating other disease vectors whose biology is closely related to that of tsetse fly. It will contribute to the efforts of developing an integrated vector management (IVM) strategy which aims at reducing cost of vector-borne disease management while minimizing detrimental effects on the environment. To achieve this, there will be need to integrate these genomics data with evolutionary ecology of the tsetse vector, molecular epidemiology of African trypanosomiases, and mathematical modeling of the processes involved in the transmission of the disease.

References

Abascal, Zardoya and Posada (2005) 'ProtTest : selection of best-fit models of protein evolution', *Bioinformatics*, 21(9), pp. 2104–2105.

Abuin, Bargeton, Ulbrich, Isacoff, Kellenberger and Benton (2011) 'Functional architecture of olfactory ionotropic glutamate receptors.', *Neuron*, 69(1), pp. 44–60.

Aebersold and Mann (2003) 'Mass spectrometry-based proteomics.', *Nature*, 422(6928), pp. 198–207.

Aksoy (2003) 'Control of tsetse flies and trypanosomes using molecular genetics', *Veterinary Parasitology*, 115(2), pp. 125–145.

Aksoy (2010) 'A proposal for tsetse fly (Glossina) genome projects', *Trends in Parasitology* ,(3), pp. 1–18.

Aksoy, Hao and Strickler (2002) 'What can we hope to gain for trypanosomiasis control from molecular studies on tsetse biology ?', *Kinetoplastid biology and disease*, 1, p. 4.

Aksoy, Maudlin, Dale, Robinson and O'Neill (2001) 'Prospects for control of African trypanosomiasis by tsetse vector manipulation', *Parasitology Today*, 17(1), pp. 29–35.

Almeida, Sánchez-Gracia, Campos and Rozas (2014) 'Family size evolution in Drosophila chemosensory gene families: a comparative analysis with a critical appraisal of methods.', *Genome biology and evolution*, 6(7), pp. 1669–1682.

Alsford, Kelly, Baker and Horn (2013) 'Genetic dissection of drug resistance in trypanosomes.', *Parasitology*, 140(12), pp. 1478–1491.

Altschul, Madden, Schäffer, Zhang, Zhang, Miller and Lipman (1997) 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Research*,25(17), pp. 3389–3402.

Alwine, Kemp and Stark (1977) 'Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5350–5354.

Andersson, Grosse-Wilde, Keeling, Bengtsson, Yuen, Li, Hillbur, Bohlmann, Hansson and Schlyter (2013) 'Antennal transcriptome analysis of the chemosensory gene families in the tree killing bark beetles, Ips typographus and Dendroctonus ponderosae (Coleoptera: Curculionidae: Scolytinae).', *BMC genomics*, 14(1), p. 198.

Andrews, Fernández, Yu, Leary, Leung, Kavanaugh, Kravitz and Certel (2014) 'Octopamine Neuromodulation Regulates Gr32a-Linked Aggression and Courtship Pathways in Drosophila Males', *PLoS Genetics*, 10(5),e1004356.

Anene, Onah and Nawa (2001) 'Drug resistance in pathogenic African trypanosomes: What hopes for the future?', *Veterinary Parasitology*, 96(2), pp. 83–100.

Angeli, Ceron, Scaloni, Monti, Monteforti, Minnocci, Petacchi and Pelosi (1999) 'Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from Schistocerca gregaria', *European Journal of Biochemistry*, 262(3), pp. 745–754.

Apweiler, Bairoch, Wu, Barker, Boeckmann, Ferro, Gasteiger, Huang, Lopez, Magrane, Martin, Natale, O'Donovan, Redaschi and Yeh (2004) 'UniProt: the Universal Protein knowledgebase.', *Nucleic acids research*, 32 suppl, pp. D115–D119.

Arya, Weber, Wang, Magwire, Serrano Negron, Mackay and Anholt (2010) 'Natural variation, functional pleiotropy and transcriptional contexts of odorant binding protein genes in Drosophila melanogaster', *Genetics*, 186(4), pp. 1475–1485.

Attardo, Abila, Auma, Baumann, Benoit, Brelsfoard, Ribeiro, Cotton, Pham, Darby, Abbeele, Denlinger, Field, Nyanjom, Gaunt, Geiser, Gomulski, Haines, Hansen, Jones, Kibet, Kinyua, Larkin, Lehane, Rio, Macdonald, Macharia, Malacrida, Marco, Marucha, Masiga, Meuti, Mireji, Obiero, Koekemoer, Okoro, Omedo, Osamor, Balyeidhusa, Peyton, Price, Quail, Ramphul, Rawlings, Riehle, Robertson, Sanders, Scott, Dashti, Snyder, Srivastava, Stanley, Swain, Hughes, Tarone, Taylor, Telleria, Thomas, Walshe, Wilson, Winzerling, Acosta-Serrano, Aksoy, Arensburger, Aslett, Bateta, Benkahla, Berriman, Bourtzis, Caers, Caljon, Christoffels, Falchetto, Friedrich, Fu, Gäde, Githinji, Gregory, Hall, Harkins, Hattori, Hertz-Fowler, Hide, Hu, Imanishi, Inoue, Jonas, Kawahara, Koffi, Kruger, Lawson, Lehane, Lehväslaiho, Luiz, Makgamathe, Malele, Manangwa, Manga, Megy, Michalkova, Mpondo, Mramba, Msangi, Mulder, Murilla, Mwangi, Okedi, Ommeh, Ooi, Ouma, Panji, Ravel, Rose, Sakate, Schoofs, Scolari, Sharma, Sim, Siwo, Solano, Stephens, Suzuki, Sze, Touré, Toyoda, Tsiamis, Tu, Wamalwa, Wamwiri, Wang, Warren, Watanabe, Weiss, Willis, Wincker, Zhang and Zhou (2014) 'Genome Sequence of the Tsetse Fly (Glossina morsitans): Vector of African Trypanosomiasis', *Science*, 344(6182), pp. 380–386.

Baker, de Koning, Mäser and Horn (2013) 'Drug resistance in African trypanosomiasis: The melarsoprol and pentamidine story', *Trends in Parasitology*,29(3) pp. 110–118.

Baker and Sali (2001) 'Protein structure prediction and structural genomics.', *Science (New York, N.Y.)*, 294(5540), pp. 93–96.

Balasegaram, Young, Chappuis, Priotto, Raguenaud and Checchi (2009) 'Effectiveness of melarsoprol and eflornithine as first-line regimens for gambiense sleeping sickness in nine Médecins Sans Frontières programmes', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103(3), pp. 280–290.

Barrett, Boykin, Brun and Tidwell (2007) 'Human African trypanosomiasis: pharmacological re-engagement with a neglected disease.', *British journal of pharmacology*, 152(8), pp. 1155–1171.

Becker-André and Hahlbrock (1989) 'Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY).', *Nucleic acids research*, 17(22), pp. 9437–9446.

Bellmann, Richardt, Freyberger, Nuwal, Schwärzel, Fiala and Störtkuhl (2010) 'Optogenetically Induced Olfactory Stimulation in Drosophila Larvae Reveals the Neuronal Basis of Odor-Aversion behavior.', *Frontiers in behavioral neuroscience*, 4, p. 27.

Benoit, Attardo, Michalkova, Krause, Bohova, Zhang, Baumann, Mireji, Takáč, Denlinger, Ribeiro and Aksoy (2014) 'A Novel Highly Divergent Protein Family Identified from a Viviparous Insect by RNA-seq Analysis: A Potential Target for Tsetse Fly-Specific Abortifacients', *PLoS Genetics*, 10(4), e1003874.

Benson, Karsch-Mizrachi, Lipman, Ostell and Sayers (2011) 'GenBank', *Nucleic Acids Research*, 39(suppl 1): pp.D48-D53.

Benton, Sachse, Michnick and Vosshall (2006) 'Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo', *PLoS Biology*, 4, pp. 240–257, e20.

Benton, Vannice, Gomez-diaz and Vosshall (2009) 'Variant Ionotropic Glutamate Receptors as Chemosensory Receptors in Drosophila', *Cell*. Elsevier Inc., 136(1), pp. 149–162.

Berendsen, Postma, van Gunsteren and Hermans (1981) 'Interaction models for water in relation to protein hydration', in *Intermolecular forces*. Springer, pp. 331–342.

Berman (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28(1|), pp. 235–242.

Besemer and Borodovsky (2005) 'GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses', *Nucleic Acids Research*, *33*(suppl 2), W451-W454.

Biegert and Lupas (2005) 'The HHpred interactive server for protein homology detection and structure prediction', *Nucleic Acid Research,* 33(suppl 2), W244-W248.

Blomquist and Vogt (2003) Insect Pheromone Biochemistry and Molecular Biology: The Biosynthesis and Detection of Pheromones and Plant Volatiles, *Academic Press*. pp. 3-18

Boratyn, Schäffer, Agarwala, Altschul, Lipman and Madden (2012) 'Domain enhanced lookup time accelerated BLAST', *Biology Direct*,7(1), p. 12.

Borodovsky and McIninch (1993) 'GENMARK: Parallel gene recognition for both DNA strands', *Computers & Chemistry*,17(2), pp. 123–133.

Van den Bossche, de La Rocque, Hendrickx and Bouyer (2010) 'A changing environment and the epidemiology of tsetse-transmitted livestock trypanosomiasis.', *Trends in parasitology*, 26(5), pp. 236–43.

Bouyer, Pruvot, Bengaly, Guerin and Lancelot (2007) 'Learning influences host choice in tsetse.', *Biology letters*, 3(2), pp. 113–116.

Briand, Swasdipan, Nespoulous, Bézirard, Blon, Huet, Ebert and Pernollet (2002) 'Characterization of a chemosensory protein (ASP3c) from honeybee (Apis mellifera *European Journal of Biochemistry*, 269(18), pp. 4586–4596.

Brooks, Bruccoleri, Olafson, States, Swaminathan and Karplus (1983) 'CHARMM: A program for macromolecular energy, minimization, and dynamics calculations', *Journal of Computational Chemistry*, 4(2), pp. 187–217.

Brun, Blum, Chappuis and Burri (2010) 'Human African trypanosomiasis.', *Lancet*, 375(9709),, pp. 148–159.

De Bruyne and Baker (2008) 'Odor detection in insects: Volatile codes', *Journal of Chemical Ecology*,34(7), pp. 882–897.

Bussi, Donadio and Parrinello (2007) 'Canonical sampling through velocity rescaling', *Journal of Chemical Physics*, 126(1).

Cantarel, Korf, Robb, Parra, Ross, Moore, Holt, Alvarado and Yandell (2008) 'MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes', *Genome Research*, 18(1), pp. 188–196.

Carey and Carlson (2011) 'Insect olfaction from model systems to disease control', 108(32), pp. 12987–12995.

Cherry, Adler, Ball, Chervitz, Dwight, Hester, Jia, Juvik, Roe, Schroeder, Weng and Botstein (1998) 'SGD: Saccharomyces genome database', *Nucleic Acids Research*, 26(1), pp. 73–79.

Chintapalli, Wang and Dow (2007) 'Using FlyAtlas to identify better Drosophila melanogaster models of human disease.', *Nature genetics*, 39(6), pp. 715–720.

Chothia and Lesk (1986) 'The relation between the divergence of sequence and structure in proteins.', *The EMBO journal*, 5(4), pp. 823–826.

Clyne, Warr and Carlson (2000) 'Candidate taste receptors in *Drosophila*.', *Science (New York, N.Y.)*, 287(5459),, pp. 1830–1834.

Conesa, Götz, García-Gómez, Terol, Talón and Robles (2005) 'Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research', *Bioinformatics*, 21(18), pp. 3674–3676.

Costa, Angelini, Feis and Ciccodicola (2010) 'Uncovering the Complexity of Transcriptomes with RNA-Seq', *Journal of Biomedicine and Biotechnology*,27(2010), pp. 19.

Croset, Rytz, Cummins, Budd, Brawand, Kaessmann, Gibson and Benton (2010) 'Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction', *PLoS Genetics*, 6(8), e1001064.

Curwen, Eyras, Andrews, Clarke, Mongin, Searle and Clamp (2004) 'The Ensembl automatic gene annotation system.', *Genome research*, 14(5), pp. 942–50.

Darden, York and Pedersen (1993) 'Particle mesh Ewald: An N log(N) method for Ewald sums in large systems', *The Journal of Chemical Physics*, 98(12), p. 10089.

Davrazou, Dong, Murphy, Johnson and Jones (2011) 'New insights into the mechanism of odorant detection by the malaria - transmitting mosquito *anopheles gambiae*', *The Journal of biological chemistry*,286(39),pp. 34175-34183.

DeLano (2002) 'Pymol: An open-source molecular graphics tool', *CCP4 Newsletter On Protein Crystallography,* 40, pp 82-92.

Delport, Poon, Frost and Kosakovsky Pond (2010) 'Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology', *Bioinformatics*, 26, pp. 2455–2457.

Dionne and Dubin (1994) 'Transduction diversity in olfaction.', *The Journal of experimental biology*, 194(1), pp. 1–21.

Dong and Chen (2013) 'Transcriptomics: Advances and approaches', *Science China Life Sciences*,56(10), pp. 960–967.

Dowell, Jokerst, Day, Eddy and Stein (2001) 'The distributed annotation system.', *BMC bioinformatics*, 2(1), p. 7.

Dransfield, Brightwell, Kyorku and Williams (1990) 'Control of tsetse fly (Diptera: Glossinidae) populations using traps at Nguruman, south-west Kenya', *Bulletin of Entomological Research*,80(3), pp. 265-276.

Drysdale (2008) 'FlyBase : a database for the Drosophila research community.', *Methods in molecular biology (Clifton, N.J.)*, 420, pp. 45–59.

Duggan and Hutchinson (1966) 'Sleeping sickness in Europeans: a review of 109 cases.', *The Journal of tropical medicine and hygiene*, 69(6), pp. 124–131.

Dyer, Lawton, Ravel, Choi, Lehane, Robinson, Okedi, Hall, Solano and Donnelly (2008) 'Molecular Phylogenetics and Evolution Molecular phylogenetics of tsetse flies (Diptera : Glossinidae) based on mitochondrial (COI , 16S , ND2) and nuclear ribosomal DNA sequences , with an emphasis on the palpalis group', 49(1), pp. 227–239.

Eddy (1996) 'Hidden Markov models.', *Current opinion in structural biology*, 6(3), pp. 361–365.

Edgar, Drive and Valley (2004) 'MUSCLE : multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797.

Eisenberg, Lüthy and Bowie (1997) 'VERIFY3D: Assessment of protein models with three-dimensional profiles', *Methods in Enzymology*, 277, pp. 396–406.

Eswar, Eramian, Webb, Shen and Sali (2008) 'Protein structure modeling with MODELLER.', *Methods in molecular biology (Clifton, N.J.)*, 426, pp. 145–159.

Fan, Manoli, Ahmed, Chen, Agarwal, Kwong, Cai, Neitz, Renslo, Baker and Shah (2013) 'Genetic and neural mechanisms that inhibit drosophila from mating with other species', *Cell*, 154(1), pp. 89–102.

FAO (2013) *The state of food and agriculture, 2013.*, *Lancet*.

FAO (2014) 'Food and Agriculture Organization of the United Nations. FISHSTAT.', *Global Aquaculture Production*.

Feldmann, Maudlin, Holmes and Miles (2005) 'The sterile insect technique as a component of area-wide integrated pest management of tsetse', in *The Trypanosomiases* , pp. 565–582.

Fishilevich and Vosshall (2005) 'Genetic and functional subdivision of the Drosophila antennal lobe', *Current Biology*, 15(17), pp. 1548–1553.

Fly, Siciliano, Scolari, Gomulski, Falchetto, Gabrieli, Field, Zhou, Gasperi and Malacrida (2014) 'Sniffing Out Chemosensory Genes from the', 9(1), pp. 20–23.

Forêt, Wanner and Maleszka (2007) 'Chemosensory proteins in the honey bee: Insights from the annotated genome, comparative analyses and expressional profiling', *Insect Biochemistry and Molecular Biology*, 37(1), pp. 19–28.

Fox, Pitts, Robertson, Carlson and Zwiebel (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 of the United States of America*, 98(25), pp. 14693–14697.

Fujii and Amrein (2002) 'Genes expressed in the Drosophila head reveal a role for fat cells in sex-specific physiology', *EMBO Journal*, 21(20), pp. 5353–5363.

Fuss and Ray (2009) 'Mechanisms of odorant receptor gene choice in Drosophila and vertebrates', *Molecular and Cellular Neuroscience*,41(2), pp. 101–112.

Gardiner, Barker, Butlin, Jordan and Ritchie (2008) 'Drosophila chemoreceptor gene evolution: selection, specialization and genome size.', *Molecular ecology*, 17(7), pp. 1648–57.

Geerts, Holmes, Diall and Eisler (2001) 'African bovine trypanosomiasis: The problem of drug resistance', *Parasitology Today*, 17(1), pp. 25–28.

Gelbart, Rindone, Chillemi, Russo, Crosby, Mathews, Ashburner, Drysdale, De Grey, Whitfield, Kautman, Mathews, Gilbert and Tolstoshev (1996) 'FlyBase: The Drosophila database', *Nucleic Acids Research*, pp. 53–56.

Gibson and Torr (1999) 'Visual and olfactory responses of haematophagous Diptera to host stimuli', *Medical and Veterinary Entomology*, 13(1), pp. 2–23.

Gikonyo, Hassanali, Njagi, Gitu and Midiwo (2002) 'Odor composition of preferred (buffalo and ox) and nonpreferred (waterbuck) hosts of some savanna tsetse flies', *Journal of Chemical Ecology*, 28(5), pp. 969–981.

Gikonyo, Hassanali, Njagi and Saini (2003) 'Responses of Glossina morsitans morsitans to blends of electroantennographically active compounds in the odors of its preferred (buffalo

and ox) and nonpreferred (waterbuck) hosts', *Journal of Chemical Ecology*, 29(10), pp. 2331–2345.

Ginalski (2006) 'Comparative modeling for protein structure prediction', *Current Opinion in Structural Biology*, 16(2), pp. 172–177.

Goldman-Huertas, Mitchell, Lapoint, Faucher, Hildebrand and Whiteman (2015) 'Evolution of herbivory in Drosophilidae linked to loss of behaviors, antennal responses, odorant receptors, and ancestral diet.', *Proceedings of the National Academy of Sciences of the United States of America*, 112(10), pp. 3026–3031.

Gomez-Diaz, Martin and Alcorta (2004) 'The cAMP transduction cascade mediates olfactory reception in Drosophila melanogaster', *Behavior Genetics*, 34(4), pp. 395–406.

Gomulski, Dimopoulos, Xi, Scolari, Gabrieli, Siciliano, Clarke, Malacrida and Gasperi (2012) 'Transcriptome profiling of sexual maturation and mating in the mediterranean fruit fly, ceratitis capitata', *PLoS ONE*, 7(1),e30857.

Gong, Zhang, Zhao, Lin, Xia and Xiang (2007) 'Identification and expression pattern of the chemosensory protein gene family in the silkworm, Bombyx mori.', *Insect biochemistry and molecular biology*, 37(3), pp. 266–277.

Gong, Zhang, Zhao, Xia and Xiang (2009) 'The odorant binding protein gene family from the genome of silkworm, Bombyx mori.', *BMC genomics*, 10(1), p. 332.

Gooding and Krafsur (2005) 'Tsetse genetics: contributions to biology, systematics, and control of tsetse flies.', *Annual review of entomology*, 50, pp. 101–123.

Grabherr, Haas, Yassour, Levin, Thompson, Amit, Adiconis, Fan, Raychowdhury, Zeng, Chen, Mauceli, Hacohen, Gnirke, Rhind, di Palma, Birren, Nusbaum, Lindblad-Toh, Friedman and Regev (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome.', *Nature biotechnology*, 29(7), pp. 644–652.

Green (1994) 'Bait methods for tsetse fly control', *Advances in Parasitology*, 34, pp. 229–291.

Groh-Lunow, Getahun, Grosse-Wilde and Hansson (2014) 'Expression of ionotropic receptors in terrestrial hermit crab's olfactory sensory neurons', *Frontiers in Cellular Neuroscience*, 8(February), pp. 1–12.

Grosjean, Rytz, Farine, Abuin, Cortot, Jefferis and Benton (2011) 'An olfactory receptor for food-derived odours promotes male courtship in Drosophila', *Nature*,478(7368), pp. 236–240.

Guttman, Garber, Levin, Donaghey, Robinson, Adiconis, Fan, Koziol, Gnirke, Nusbaum, Rinn, Lander and Regev (2010) 'Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.', *Nature biotechnology*, 28(5), pp. 503–10.

Halgren, Murphy, Friesner, Beard, Frye, Pollard and Banks (2004) 'Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening', *Journal of Medicinal Chemistry*, 47(7), pp. 1750–1759.

Hall, Beevor, Cork, Nesbitt and Vale (1984a) '1-Octen-3-ol. A potent olfactory stimulant and attractant for tsetse isolated from cattle odours', *International Journal of Tropical Insect Science*, pp. 335–339.

Hall, Beevor, Cork, Nesbitt and Vale (1984b) '1-Octen-3-ol. A potent olfactory stimulant and attractant for tsetse isolated from cattle odours', *International Journal of Tropical Insect Science*,5(05), pp. 335–339.

Hallem and Carlson (2006) 'Coding of Odors by a Receptor Repertoire', *Cell*, 125(1), pp. 143–160.

Hallem, Dahanukar and Carlson (2006) 'Insect odor and taste receptors.', *Annual review of entomology*, 51, pp. 113–135.

Hamilton and Norgren (1984) 'Central projections of gustatory nerves in the rat.', *The Journal of comparative neurology*, 222(4), pp. 560–577.

Hansson, Oostenbrink and Van Gunsteren (2002) 'Molecular dynamics simulations', *Current Opinion in Structural Biology*, 12(2), pp. 190–196.

Heger and Ponting (2007) 'Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes', *Genome Research*, 17(12), pp. 1837–1849.

Hekmat-scafe, Scafe, Mckinney and Mark (2002) 'Genome-Wide Analysis of the Odorant-Binding Protein Gene Family in Drosophila melanogaster', *Genome Research* , 12, pp. 1357–1369.

Hernández-santoyo, Tenorio-barajas, Altuzar, Vivanco-cid and Mendoza-barrera 'Protein-Protein and Protein-Ligand Docking'.*Biochemsitry, Genetics and Molecular Biology.*

Hess, Bekker, Berendsen and Fraaije (1997) 'LINCS: A linear constraint solver for molecular simulations', *Journal of Computational Chemistry*, 18(12), pp. 1463–1472.

Hill, Fox, Pitts, Kent, Tan, Chrystal, Cravchik, Collins, Robertson and Zwiebel (2002) 'G protein-coupled receptors in Anopheles gambiae.', *Science (New York, N.Y.)*, 298(5591), pp. 176–178.

Hocking, Lamerton and Lewis (1963) 'Tsetse-fly control and eradication.', *Bulletin of the World Health Organization*, 28(5-6), pp. 811–823.

Holt and Yandell (2011) 'MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects', *BMC Bioinformatics*,12(1), p. 491.

Honson, Gong and Plettner (2005) 'Chapter Nine-Structure and Function of Insect Odorant and Pheromone-Binding Proteins (OBPs and PBPs) and Chemosensory-Specific Proteins (CSPs)', *Recent Advances in Phytochemistry*. Elsevier, 39, pp. 227–268.

Horn and McCulloch (2010) 'Molecular mechanisms underlying the control of antigenic variation in African trypanosomes', *Current Opinion in Microbiology*, pp. 700–705.

Huang (2009) 'MetaPocket: a meta approach to improve protein ligand binding site prediction.', *Omics : a journal of integrative biology*, 13(4), pp. 325–330.

Humphrey, Dalke and Schulten (1996) 'VMD: Visual molecular dynamics', *Journal of Molecular Graphics*, 14(1), pp. 33–38.

Inohara and Nuez (2002) 'ML - A conserved domain involved in innate immunity and lipid metabolism', *Trends in Biochemical Sciences*,27(5), pp. 219–221.

Ishida and Leal (2005) 'Rapid inactivation of a moth pheromone.', *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), pp. 14075–14079.

Isono, Morita, Bickmeyer and Wegener (2010) 'Molecular and cellular designs of insect taste receptor system', *Front Cell Neuroscience*, pp. 1–16.

Jacquin-joly and Merlin (2004) 'Review article insect olfactory receptors : contributions of molecular biology to chemical ecology', *Journal of Chemical Biology* 30(12), pp. 2359–2397.

Jin, Ha and Smith (2008) 'SNMP is a signaling component required for pheromone sensitivity in Drosophila', *105*(31), 10996-11001.

Jones, Cayirlioglu, Kadow and Vosshall (2007) 'Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*.', *Nature*, 445(7123), pp. 86–90.

Kabasa (2007) 'Public-private partnership works to stamp out sleeping sickness in Uganda', *Trends in Parasitology*, 23(5), pp. 191–192.

Katre, Mazumder, Prusti and Mohanty (2009) 'Ligand binding turns moth pheromone binding protein into a pH sensor: Effect on the antheraea polyphemus PBP1 conformation', *Journal of Biological Chemistry*, 284(46), pp. 32167–32177.

Kaupp (2010) 'Olfactory signalling in vertebrates and insects: differences and commonalities.', *Nature reviews. Neuroscience*, 11(3), pp. 188–200.

Kawabata (2003) 'MATRAS: A program for protein 3D structure comparison', *Nucleic Acids Research*, 31(13), pp. 3367–3369.

Kent, Sugnet, Furey, Roskin, Pringle, Zahler and Haussler (2002) 'The human genome browser at UCSC', *Genome Res*, 12(6), pp. 996–1006.

Kioy, Jannin and Mattock (2004) 'Focus: Human African trypanosomiasis', *Nat Rev Micro*, 2(3), pp. 186–187.

Korf (2004) 'Gene finding in novel genomes.', *BMC bioinformatics*, 5(1), p. 59.

Kosakovsky Pond, Frost and Muse (2005) 'HyPhy: Hypothesis testing using phylogenies', *Bioinformatics*, 21(5), pp. 676–679.

Krafsur (2009) 'Tsetse flies: Genetics, evolution, and role as vectors', *Infection, Genetics and Evolution*, 9(1), pp. 124–141.

Krang, Knaden, Steck and Hansson (2012) 'Transition from sea to land: olfactory function and constraints in the terrestrial hermit crab Coenobita clypeatus', *Proceedings of the Royal Society B: Biological Sciences*,279(1742), pp. 3510–3519.

Kruse, Zhao, Smith and Jones (2003) 'Structure of a specific alcohol-binding site defined by the odorant binding protein LUSH from Drosophila melanogaster.', *Nature structural biology*, 10(9), pp. 694–700.

Kulmuni and Havukainen (2013) 'Insights into the Evolution of the CSP Gene Family through the Integration of Evolutionary Analysis and Comparative Protein Modeling', *PLoS ONE*, 8(5), e63688.

Kumari, Kumar and Lynn (2014) 'G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations', *Journal of Chemical Information and Modeling*, 54(7), pp. 1951–1962.

Kwon, Dahanukar, Weiss and Carlson (2007) 'The molecular basis of CO_2 reception in *Drosophila*',*Proceedings of National Academy of Sciences,* 104(9), pp.3574-3578

Langmead (2010) 'Aligning short sequencing reads with Bowtie', *Current Protocols in Bioinformatics*, (SUPP.32),.

Larsson, Domingos, Jones, Chiappe, Amrein and Vosshall (2004) 'Or83b encodes a broadly expressed odorant receptor essential for Drosophila olfaction', *Neuron*, 43(5), pp. 703–714.

Lartigue, Gruez, Briand, Blon, Bézirard, Walsh, Pernollet, Tegoni and Cambillau (2004) 'Sulfur Single-wavelength Anomalous Diffraction Crystal Structure of a Pheromone-Binding Protein from the Honeybee Apis mellifera L', *Journal of Biological Chemistry*, 279(6), pp. 4459–4464.

Laskowski, MacArthur, Moss and Thornton (1993) 'PROCHECK: a program to check the stereochemical quality of protein structures', *Journal of Applied Crystallography*,26(2), pp. 283–291.

Law and Sansom (2004) 'Homology modelling and molecular dynamics simulations: Comparative studies of human aquaporin-1', *European Biophysics Journal*, 33(6), pp. 477–489.

Lawson, Arensburger, Atkinson, Besansky, Bruggner, Butler, Campbell, Christophides, Christley, Dialynas, Hammond, Hill, Konopinski, Lobo, Maccallum, Madey, Megy, Meyer, Redmond, Severson, Stinson, Topalis, Birney, Gelbart, Kafatos, Louis and Collins (2009) 'VectorBase: A data resource for invertebrate vector genomics', *Nucleic Acids Research*, *37*(suppl 1), D583-D587.

Leal, W. S. (2013). Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annual review of entomology*, *58*, 373-391.

Leite, Krogh, Xu, Ishida, Iulek, Leal and Oliva (2009) 'Structure of an Odorant-Binding Protein from the Mosquito Aedes aegypti Suggests a Binding Pocket Covered by a pH-Sensitive " Lid ',*PloS ONE*, 4(11), pp. 1–7.

Levitan (1994) 'Modulation of ion channels by protein phosphorylation and dephosphorylation.', *Annual review of physiology*, 56(1), pp. 193–212.

Lewis, Searle, Harris, Gibson, Lyer, Richter, Wiel, Bayraktaroglir, Birney, Crosby, Kaminker, Matthews, Prochnik, Smithy, Tupy, Rubin, Misra, Mungall and Clamp (2002) 'Apollo: a sequence annotation editor.', *Genome biology*, 3(12), p. 0082.

Li, Picimbon, Ji, Kan, Chuanling, Zhou and Pelosi (2008) 'Multiple functions of an odorant-binding protein in the mosquito Aedes aegypti', *Biochemical and Biophysical Research Communications*, 372(3), pp. 464–468.

Li, Stoeckert and Roos (2003) 'OrthoMCL: Identification of ortholog groups for eukaryotic genomes', *Genome Research*, 13(9), pp. 2178–2189.

Liu, He, Lehane, Lehane, Hertz-Fowler, Berriman, Field and Zhou (2012) 'Expression of chemosensory proteins in the tsetse fly Glossina morsitans morsitans is related to female host-seeking behaviour.', *Insect molecular biology*, 21(1), pp. 41–8.

Liu, Lehane, He, Lehane, Hertz-Fowler, Berriman, Pickett, Field and Zhou (2010) 'Characterisations of odorant-binding proteins in the tsetse fly Glossina morsitans morsitans', *Cellular and Molecular Life Sciences*, 67(6), pp. 919–929.

Logan and Birkett (2007) 'Semiochemicals for biting fly control: Their identification and exploitation', *Pest Management Science*, 63(7), pp. 647–657.

Löytynoja (2014) 'Phylogeny-aware alignment with PRANK', *Methods in Molecular Biology*, 1079, pp. 155–170.

Macharia, Mireji, Murungi, Murilla, Christoffels, Aksoy and Masiga (2016) 'Genome-Wide Comparative Analysis of Chemosensory Gene Families in Five Tsetse Fly Species', *PLoS Negl Trop Dis*. 10(2), p. e0004421.

Majoros, Pertea, Antonescu and Salzberg (2003) 'GlimmerM, Exonomy and Unveil: Three ab initio eukaryotic genefinders', *Nucleic Acids Research*, 31(13), pp. 3601–3604.

Mameli, Tuccini, Mazza, Petacchi and Pelosi (1996) 'Soluble proteins in chemosensory organs of phasmids', *Insect Biochemistry and Molecular Biology*, 26(8), pp. 875–882.

Mamidala, Wijeratne, Wijeratne, Poland, Qazi, Doucet, Cusson, Beliveau and Mittapalli (2013) 'Identification of Odor-Processing Genes in the Emerald Ash Borer, Agrilus planipennis', *PLoS ONE*, 8, e56555.

Marchler-Bauer, Anderson, Cherukuri, DeWeese-Scott, Geer, Gwadz, He, Hurwitz, Jackson, Ke, Lanczycki, Liebert, Liu, Lu, Marchler, Mullokandov, Shoemaker, Simonyan, Song, Thiessen, Yamashita, Yin, Zhang and Bryant (2005) 'CDD: A Conserved Domain Database for protein classification', *Nucleic Acids Research*, *33*(suppl 1), D192-D196.

Marshall, Warr and de Bruyne (2010) 'Detection of volatile indicators of illicit substances by the olfactory receptors of Drosophila melanogaster.', *Chemical senses*, 35(7), pp. 613–25.

Martin, Bruno, Fang, Meng, Blow, Zhang, Sherlock, Snyder and Wang (2010) 'Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads.', *BMC genomics*, 11(1), p. 663.

Martin and Wang (2011) 'assembly', *Nature Publishing Group*. Nature Publishing Group, 12(10), pp. 671–682.

Marti-Renom, Fiser, Madhusudhan, John, Stuart, Eswar, Pieper, Shen and Sali (2003) 'Modeling protein structure from its sequence', *Current Protocols in Bioinformatics*, 5, p. 0.

Masiga, Okech, Irungu, Ouma, Wekesa, Ouma, Guya and Ndung'u (2002) 'Growth and mortality in sheep and goats under high tsetse challenge in Kenya.', *Tropical animal health and production*, 34(6), pp. 489–501.

Mathé, Sagot, Schiex and Rouzé (2002) 'Current methods of gene prediction, their strengths and weaknesses.', *Nucleic acids research*, 30(19), pp. 4103–4117.

Matovu, Seebeck, Enyaru and Kaminsky (2001) 'Drug resistance in Trypanosoma brucei spp., the causative agents of sleeping sickness in man and nagana in cattle', *Microbes and Infection*, 3(9), pp. 763–770.

Mayer (2011) 'Emerging models of glutamate receptor ion channel structure and function.', *Structure (London, England : 1993)*, 19(10), pp. 1370–80.

McBride and Arguello (2007) 'Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily', *Genetics*, 177(3), pp. 1395–1416.

McGraw, Gibson, Clark and Wolfner (2004) 'Genes regulated by mating, sperm, or seminal proteins in mated female Drosophila melanogaster', *Current Biology*, 14(16), pp. 1509–1514.

McKenna, Hekmat-Scafe, Gaines and Carlson (1994) 'Putative Drosophila pheromone-binding proteins expressed in a subregion of the alfactory system', *Journal of Biological Chemistry*, 269(23), pp. 16340–16347.

Mireji, Mabveni, Dube, Ogembo, Matoka and Mangwiro (2003) 'Field Responses of Tsetse Flies (Glossinidae) and Other Diptera to Oils in Formulations of Deltamethrin', *International Journal of Tropical Insect Science*,23(04), pp. 317–323.

Mishra, Miyamoto, Rezenom, Broussard, Yavuz, Slone, Russell and Amrein (2013) 'The molecular basis of sugar sensing in drosophila larvae', *Current Biology*, 23(15), pp. 1466–1471.

Miyamoto, Slone, Song and Amrein (2012) 'A fructose receptor functions as a nutrient sensor in the drosophila brain', *Cell*, 151(5), pp. 1113–1125.

Montell (2009) 'A taste of the Drosophila gustatory receptors', *Current Opinion in Neurobiology*,19(4) pp. 345–353.

Morimoto, Nobechi, Komatsu, Miyakawa and Nose (2010) 'Subunit-specific and homeostatic regulation of glutamate receptor localization by CaMKII in Drosophila neuromuscular junctions', *Neuroscience*, 165(4), pp. 1284–1292.

Morozova, Hirst and Marra (2009) 'Applications of new sequencing technologies for transcriptome analysis.', *Annual review of genomics and human genetics*, 10, pp. 135–51.

Morris and Lim-Wilby (2008) 'Molecular docking.', *Methods in molecular biology (Clifton, N.J.)*, 443, pp. 365–382.

Mortazavi, Williams, Mccue, Schaeffer and Wold (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', 5(7), pp. 621–628.

Mugasa, Schoone, Ekangu, Lubega, Kager and Schallig (2008) 'Detection of Trypanosoma brucei parasites in blood samples using real-time nucleic acid sequence-based amplification', *Diagnostic Microbiology and Infectious Disease*, 61(4), pp. 440–445.

Mulder and Apweiler (2007) 'InterPro and InterProScan: tools for protein sequence classification and comparison.', *Methods in molecular biology (Clifton, N.J.)*, 396, pp. 59–70.

Murrell, Wertheim, Moola, Weighill, Scheffler and Kosakovsky Pond (2012) 'Detecting individual sites subject to episodic diversifying selection', *PLoS Genetics*, 8.

Muturi, Ouma, Malele, Ngure, Rutto, Mithöfer, Enyaru and Masiga (2011) 'Tracking the feeding patterns of tsetse flies (glossina genus) by analysis of bloodmeals using mitochondrial cytochromes genes', *PLoS ONE*, 6(2),e1002764.

Mwangi, Gikonyo and Ndiege (2008) 'Repellent properties of delta-octalactone against the tsetse fly, Glossina morsitans morsitans.', *Journal of insect science*, 8(1), pp. 1–4.

Nichols and Ã (2008) 'The SNMP / CD36 gene family in Diptera , Hymenoptera and Coleoptera : *Drosophila melanogaster* , *D . pseudoobscura* , *Anopheles gambiae* , *Aedes aegypti* , *Apis mellifera* , and *Tribolium castaneum*.' *Insect Biochem Mol Biol*, 38(4), pp 398-415

Niimura and Nei (2006) 'Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates', *Journal of Human Genetics*, 51(60, pp. 505–517.

O'Boyle, Banck, James, Morley, Vandermeersch and Hutchison (2011) 'Open Babel: An Open chemical toolbox', *Journal of Cheminformatics*, 3(10),p 33.

O'Brien, Remm and Sonnhammer (2005) 'Inparanoid: A comprehensive database of eukaryotic orthologs', *Nucleic Acids Research*, 33(Database issue), D476-D480.

Obiero (2014) 'Genome-wide annotation of chemosensory and glutamate-gated receptors, and related genes in Glossina morsitans morsitans tsetse fly.'

Obiero, Mireji, Nyanjom, Christoffels, Robertson and Masiga (2014) 'Odorant and gustatory receptors in the tsetse fly Glossina morsitans morsitans.', *PLoS neglected tropical diseases*, 8(4), p. e2663.

Ochieng, Hallberg and Hansson (1998) 'Fine structure and distribution of antennal sensilla of the desert locust, Schistocerca gregaria (Orthoptera: Acrididae).', *Cell and tissue research*, 291(3), pp. 525–536.

Odiit, Bessell, Fèvre, Robinson, Kinoti, Coleman, Welburn, McDermott and Woolhouse (2006) 'Using remote sensing and geographic information systems to identify villages at high risk for rhodesiense sleeping sickness in Uganda', *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100(4), pp. 354–362.

Omolo, Hassanali, Mpiana, Esterhuizen, Lindh, Lehane, Solano, Rayaisse, Vale, Torr and Tirados (2009) 'Prospects for developing odour baits to control Glossina fuscipes spp., the major vector of human African trypanosomiasis', *PLoS Neglected Tropical Diseases*, 3(5),e435.

Otter Tchicaya and Schutte (1991) 'Effects of age, sex and hunger on the antennal olfactory sensitivity of tsetse flies', *Physiological Entomology*. Blackwell Publishing Ltd, 16(2), pp. 173–182.

Ozaki, Utoguchi, Yamada and Yoshikawa (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(11), pp. 969–976.

Özlem Tastan Bishop, de Beer and Joubert (2008) 'Protein homology modelling and its use in South Africa', *South African Journal of Science*, 104(1-2), pp. 2–6.

Paas, Devillers-Thiéry, Teichberg, Changeux and Eisenstein (2000) 'How well can molecular modelling predict the crystal structure: the case of the ligand-binding domain of glutamate receptors.', *Trends in pharmacological sciences*, 21(3), pp. 87–92.

Parrinello and Rahman (1981) 'Polymorphic transitions in single crystals: a new molecular dynamics method.', *Journal of Applied Physics*, 52(12), pp. 7182–7190.

Pei, Tang and Grishin (2008) 'PROMALS3D web server for accurate multiple protein sequence and structure alignments.', *Nucleic acids research*, 36(7),pp 2295-2300.

Pence and Williams (2010) 'Chemspider: An online chemical information resource', *Journal of Chemical Education*, 87(11), pp. 1123–1124.

Pikielny, Hasan, Rouyer and Rosbash (1994) 'Members of a family of Drosophila putative odorant-binding proteins are expressed in different subsets of olfactory hairs', *Neuron*, 12(1), pp. 35–49.

Ponting, Mott, Bork and Copley (2001) 'Novel protein domains and repeats in Drosophila melanogaster: insights into structure, function, and evolution.', *Genome research*, 11(12), pp. 1996–2008.

Poon, Frost and Pond (2009) 'Detecting signatures of selection from DNA sequences using datamonkey', *Methods in Molecular Biology*, 537, pp. 163–183.

Prabakaran, Lippens, Steen and Gunawardena (2012) 'Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*,) pp. 565–583.

Pronk, Páll, Schulz, Larsson, Bjelkmar, Apostolov, Shirts, Smith, Kasson, Van Der Spoel, Hess and Lindahl (2013) 'GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit', *Bioinformatics*, 29(7), pp. 845–854.

Quevillon, Silventoinen, Pillai, Harte, Mulder, Apweiler and Lopez (2005) 'InterProScan: Protein domains identifier', *Nucleic Acids Research*, *33*(suppl 2), W116-W120.

Falk (1976) 'Labellar taste organs of Drosophila melanogaster.', *Journal of Morphology*, 150(2), pp. 327–342.

Reeves, Talavera and Thornton (2009) 'Genome and proteome annotation: organization, interpretation and integration.', *Journal of the Royal Society, Interface / the Royal Society*, 6, pp. 129–147.

Riesgo-Escovar, Piekos and Carlson, John R. (1997) 'The Drosophila antenna: Ultrastructural and physiological studies in wild-type and lozenge mutants', *Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology*, 180(2), pp. 151–160.

Riesgo-Escovar, Piekos and Carlson, J. R. (1997) 'The maxillary palp of Drosophila: Ultrastructure and physiology depends on the lozenge gene', *Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology*, 180(2), pp. 143–150.

Robertson, Martos, Sears, Todres, Walden and Nardi (1999) 'Diversity of odourant binding proteins revealed by an expressed sequence tag project on male Manduca sexta moth antennae.', *Insect molecular biology*, 8(4), pp. 501–18.

Robertson, Schein, Chiu, Corbett, Field, Jackman, Mungall, Lee, Okada, Qian, Griffith, Raymond, Thiessen, Cezard, Butterfield, Newsome, Chan, She, Varhol, Kamoh, Prabhu, Tam, Zhao, Moore, Hirst, Marra, Jones, Hoodless and Birol (2010) 'De novo assembly and analysis of RNA-seq data.', *Nature methods*, 7(11), pp. 909–912.

Robertson, Warr and Carlson (2003) 'Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster.', *Proceedings of the National Academy of Sciences of the United States of America*, 100 Suppl , pp. 14537–14542.

Rogers, Steinbrecht and Vogt (2001) 'Expression of SNMP-1 in olfactory neurons and sensilla of male and female antennae of the silkmoth Antheraea polyphemus', *Cell and Tissue Research*, 303(3), pp. 433–446.

Rogers, Sun, Lerner and Vogt (1997) 'Snmp-1, a novel membrane protein of olfactory neurons of the silk moth Antheraea polyphemus with homology to the CD36 family of membrane proteins', *Journal of Biological Chemistry*, 272(23), pp. 14792–14799.

Rollmann, Wang, Date, West, Mackay and Anholt (2010) 'Odorant Receptor Polymorphisms and Natural Variation in Olfactory Behavior in *Drosophila melanogaster*', *Genetics*, 186(2), pp. 687–697.

Rombel, Sykes, Rayner and Johnston (2002) 'ORF-FINDER: A vector for high-throughput gene identification', *Gene*, 282(1), pp. 33–41.

Ronderos and Smith (2009) 'Diverse signaling mechanisms mediate volatile odorant detection in *Drosophila*', *Fly*, 3(4), pp 290-297.

Rutherford, Parkhill, Crook, Horsnell, Rice, Rajandream and Barrell (2000) 'Artemis: sequence visualization and annotation.', *Bioinformatics*, 16(10), pp. 944–945.

Rytz, Croset and Benton (2013) 'Ionotropic Receptors (IRs): Chemosensory ionotropic glutamate receptors in Drosophila and beyond', *Insect Biochemistry and Molecular Biology*. Elsevier Ltd, 43(9), pp. 888–897.

Salamov (2000) 'Ab initio Gene Finding in Drosophila Genomic DNA', *Genome Research*, 10(4), pp. 516–522.

Salzberg, Pertea, Delcher, Gardner and Tettelin (1999) 'Interpolated Markov models for eukaryotic gene finding.', *Genomics*, 59(1), pp. 24–31.

Sandler, Nikonova, Leal and Clardy (2000) 'Sexual attraction in the silkworm moth: Structure of the pheromone-binding-protein-bombykol complex', *Chemistry and Biology*, 7(2), pp. 143–151.

Sato, Pellegrino, Nakagawa, Nakagawa, Vosshall and Touhara (2008) 'Insect olfactory receptors are heteromeric ligand-gated ion channels.', *Nature*, 452(7190), pp. 1002–1006.

Scheffler, Martin and Seoighe (2006) 'Robust inference of positive selection from recombining coding sequences', *Bioinformatics*, 22(20), pp. 2493–2499.

Scott, Brady, Cravchik, Morozov, Rzhetsky, Zuker and Axel (2001) 'A chemosensory gene family encoding candidate gustatory and olfactory receptors in *Drosophila*.', *Cell*, 104(5), pp. 661–673.

Scott, Warren, Beukeboom, Bopp, Clark, Giers, Hediger, Jones, Kasai, Leichter, Li, Meisel, Minx, Murphy, Nelson, Reid, Rinkevich, Robertson, Sackton, Sattelle, Thibaud-Nissen, Tomlinson, van de Zande, Walden, Wilson and Liu (2014) 'Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment', *Genome Biology*, 15(10), p. 466.

Shanbhag, Hekmat-Scafe, Kim, Park, Carlson, Pikielny, Smith and Steinbrecht (2001) 'Expression mosaic of odorant-binding proteins in Drosophila olfactory organs', *Microscopy Research and Technique*, 55(5), pp. 297–306.

Shanbhag, Singh and Singh (1995) 'Fine structure and primary sensory projections of sensilla located in the sacculus of the antenna of Drosophila melanogaster.', *Cell and tissue research*, 282(2), pp. 237–49.

Siciliano, He, Woodcock, Pickett, Field, Birkett, Kalinova, Gomulski, Scolari, Gasperi, Malacrida and Zhou (2014) 'Identi fi cation of pheromone components and their binding af fi nity to the odorant binding protein CcapOBP83a-2 of the Mediterranean fruit fl y , *Ceratitis capitata'*, *Insect Biochem and Mol Biol.*,48, pp. 51–62.

Silbering, Rytz, Grosjean, Abuin, Ramdya, Jefferis and Benton (2011) 'Complementary Function and Integrated Wiring of the Evolutionarily Distinct Drosophila Olfactory Subsystems', *Journal of Neuroscience*, 31(38), pp. 13357–13375.

Silva, Alan and Vranken (2012) 'ACPYPE - AnteChamber PYthon Parser interfacE', *BMC Research* ,5(367),pp. 1–8.

Simarro, Diarra, Postigo, Franco and Jannin (2011) 'The human african trypanosomiasis control and surveillance programme of the world health organization 2000-2009: The way forward', *PLoS Neglected Tropical Diseases*, 5(2),e1007.

Simpson, Wong, Jackman, Schein, Jones and Birol (2009) 'ABySS: A parallel assembler for short read sequence data', *Genome Research*, 19(6), pp. 1117–1123.

Smit, Hubley and Green (1996) *RepeatMasker Open-3.0*, *RepeatMasker Open-3.0*.

Smyth and Martin (2000) 'x ray crystallography.', *Molecular pathology : MP*, 53(1), pp. 8–14.

Solano, Ravel and de Meeûs (2010) 'How can tsetse population genetics contribute to African trypanosomiasis control?', *Trends in Parasitology*, 26(5), pp. 255–263.

Sonnhammer, Eddy and Durbin (1997) 'Pfam: A comprehensive database of protein domain families based on seed alignments', *Proteins: Structure, Function and Genetics*, 28(3), pp. 405–420.

Sousa, Fernandes and Joa (2006) 'Protein – Ligand Docking : Current Status and Future', *Bioinformatics*, 65(1), pp. 15–26.

Späth (1995) 'Olfactory attractants for West African tsetse flies, glossina spp. (Diptera: Glossinidae).', *Tropical medicine and parasitology,* 46(4), pp. 253–257.

Späth (2000) 'Feeding patterns of three sympatric tsetse species (Glossina spp.) (Diptera: Glossinidae) in the preforest zone of Cote d'Ivoire', *Acta Tropica*, 75, pp. 109–118.

Spinelli, Lagarde, Iovinella, Legrand, Tegoni, Pelosi and Cambillau (2012) 'Crystal structure of Apis mellifera OBP14, a C-minus odorant-binding protein, and its complexes with odorant molecules', *Insect Biochemistry and Molecular Biology*, 42(1), pp. 41–50.

Van Der Spoel, Lindahl, Hess, Groenhof, Mark and Berendsen (2005) 'GROMACS: Fast, flexible, and free', *Journal of Computational Chemistry*, 26(16), pp. 1701–1718.

Stamatakis (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.', *Bioinformatics*, 30, pp. 1312–3.

Stanke and Morgenstern (2005) 'AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints.', *Nucleic acids research*, 33 (Web Server issue), pp. W465–7.

Stein (2001) 'Reviews Genome Annotation : From Sequence To Biology',*Nature Reviews Genetics,* 2(7), pp 493-503 .

Sun, Huang, Pelosi and Wang (2012) 'Expression in antennae and reproductive organs suggests a dual role of an odorant-binding protein in two sibling Helicoverpa species', *PLoS ONE*, 7(1), e300400.

Surget-Groba and Montoya-Burgos (2010) 'Optimization of de novo transcriptome assembly from next-generation sequencing data', *Genome Research*, 20(10), pp. 1432–1440.

Swarup, Williams and Anholt (2011) 'Functional dissection of Odorant binding protein genes in Drosophila melanogaster', *Genes, Brain and Behavior*, 10(6), pp. 648–657.

Tatusov, Galperin, Natale and Koonin (2000) 'The COG database: a tool for genome-scale analysis of protein functions and evolution.', *Nucleic acids research*, 28(1), pp. 33–36.

Taylor and Burnett (2000) 'DARWIN: A program for docking flexible molecules', *Proteins: Structure, Function and Genetics*, 41(2), pp. 173–191.

Telleria, Benoit, Zhao, Savage, Regmi, e Silva, O'Neill and Aksoy (2014) 'Insights into the Trypanosome-Host Interactions Revealed through Transcriptomic Analysis of Parasitized Tsetse Fly Salivary Glands', *PLoS Neglected Tropical Diseases*, 8(4),e2629.

The Jmol Team (2007) 'Jmol: an open-source Java viewer for chemical structures in 3D.'

Torr, Mangwiro and Hall (1996) 'Responses of Glossina pallidipes (Diptera: Glossinidae) to synthetic repellents in the field', *Bulletin of Entomological Research*. Cambridge Univ Press, 86(05), pp. 609–616.

Torr, Mangwiro and Hall (2006) 'The effects of host physiology on the attraction of tsetse (Diptera: Glossinidae) and Stomoxys (Diptera: Muscidae) to cattle.', *Bulletin of entomological research*, 96(1), pp. 71–84.

Trapnell, Williams, Pertea, Mortazavi, Kwan, Baren, Salzberg, Wold and Pachter (2010) 'letters Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology.*, 28(5), pp. 516–520.

Trapnell, Williams, Pertea, Mortazavi, Kwan, van Baren, Salzberg, Wold and Pachter (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.', *Nature biotechnology*, 28(5), pp. 511–515.

Trott and Olson (2010) 'AutoDock Vina improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading." ', *J. Comput. Chem.*, 31(2), pp. 445–461.

Tuck, Windmill and Robert (2009) 'Hearing in tsetse flies? Morphology and mechanics of a putative auditory organ.', *Bulletin of entomological research*, 99(02), pp. 107–119.

De Val, Declercq, Lim and Crichton (2012) 'Structural analysis of haemin demetallation by L-chain apoferritins', *Journal of Inorganic Biochemistry*, 112, pp. 77–84.

Vale, Hall and Gough (1988) 'The olfactory responses of tsetse flies, Glossina spp.(Diptera: Glossinidae), to phenols and urine in the field', *Bulletin of Entomological Research*. Cambridge Univ Press, 78(02), pp. 293–300.

Venthur, Mutis, Zhou and Quiroz (2014) 'Ligand binding and homology modelling of insect odorant-binding proteins', *Physiological Entomology*. Wiley Online Library, 39(3), pp. 183–198.

Vera, Wheat, Fescemyer, Frilander, Crawford, Hanski and Marden (2008) 'Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing', *Molecular Ecology*, 17(7), pp. 1636–1647.

Vieira and Rozas (2009) 'Molecular evolution of the major chemosensory gene families in insects' 103(3),, pp. 208–216.

Vieira, Sánchez-Gracia and Rozas (2007) 'Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution.', *Genome biology*, 8(11), p. R235.

Vogt, Miller, Litvack, Fandino, Sparks, Staples, Friedman and Dickens (2009) 'The insect SNMP gene family', *Insect Biochemistry and Molecular Biology*, 39 (7), pp. 448–456.

Voight, Kudaravalli, Wen and Pritchard (2006) 'A map of recent positive selection in the human genome', *PLoS Biology*, 4(3), pp. 0446–0458.

Voskamp, Everaarts and Den Otter (1999) 'Olfactory responses to attractants and repellents in tsetse', *Medical and Veterinary Entomology*, 13(4), pp. 386–392.

Vosshall (2003) 'Diversity and expression of odorant receptors in *Drosophila*', in *Insect Pheromone Biochemistry and Molecular Biology*, pp. 567–591.

Vosshall and Stocker (2007) 'Molecular architecture of smell and taste in Drosophila.', *Annual review of neuroscience*, 30, pp. 505–533.

Vreysen, Saleh, Ali, Abdulla, Zhu, Juma, Dyck, Msangi, Mkonyi and Feldmann (2000) 'Glossina austeni (Diptera: Glossinidae) eradicated on the island of Unguja, Zanzibar, using the sterile insect technique.', *Journal of economic entomology*, 93(1), pp. 123–135.

Wallace, Laskowski and Thornton (1995) 'Ligplot - a program to generate schematic diagrams of protein ligand interactions', *Protein Engineering*, 8(2), pp. 127–134.

Wang and Anderson (2010) 'Identification of an aggression-promoting pheromone and itw receptor neurons in *Drosophila'*, *Nature*, 463(7278), pp. 227–231.

Wang, Gerstein and Snyder (2009) 'RNA-Seq: a revolutionary tool for transcriptomics.'*Nat Rev Genet* 10(1):57-63.

Waterhouse, Procter, Martin, Clamp and Barton (2009) 'Jalview Version 2-A multiple sequence alignment editor and analysis workbench', *Bioinformatics*, 25(9), pp. 1189–1191.

Weiss, Maltz and Aksoy (2012) 'Obligate symbionts activate immune system development in the tsetse fly.', *Journal of immunology (Baltimore, Md. : 1950)*, 188(7), pp. 3395–403.

Welburn, Maudlin and Simarro (2009) 'Controlling sleeping sickness - a review.', *Parasitology*, 136(14), pp. 1943–1949.

Wicher, Schäfer, Bauernfeind, Stensmyr, Heller, Heinemann and Hansson (2008) *'Drosophila* odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels.', *Nature*, 452(7190), pp. 1007–1011.

Wiederstein and Sippl (2007) 'ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins', *Nucleic Acids Research*, 35(SUPPL.2), W407-W410.

Wiegmann, Trautwein, Winkler, Barr, Kim, Lambkin, Bertone, Cassel, Bayless, Heimberg, Wheeler, Peterson, Pape, Sinclair, Skevington, Blagoderov, Caravas, Kutty, Schmidt-Ott, Kampmeier, Thompson, Grimaldi, Beckenbach, Courtney, Friedrich, Meier and Yeates (2011) 'Episodic radiations in the fly tree of life.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), pp. 5690–5695.

Wu and Nacu (2010) 'Fast and SNP-tolerant detection of complex variants and splicing in short reads', *Bioinformatics*, 26(7), pp. 873–881.

Wüthrich (1990) 'Protein structure determination in solution by NMR spectroscopy', *J Biol Chem*, 265(36), pp. 22059–62.

Xie, Wu, Tang, Luo, Patterson, Huang, He, Gu, Li, Zhou, Li, Xu, Wong and Wang (2012) 'SOAPdenovo-Trans : De novo transcriptome assembly with short RNA-Seq reads',30(12), pp. 1–7.

Xu, He, Zhang, Fang, Dong, Zhang and Li (2009) 'Large-scale identification of odorant-binding proteins and chemosensory proteins from expressed sequence tags in insects', 10(1), pp. 1–13.

Yamamoto and Takemori (2010) 'Proteome profiling reveals tissue-specific protein expression in the male reproductive system of Drosophila melanogaster', *Fly*, 4(1), pp. 36–39.

Yandell and Ence (2012) 'A beginner's guide to eukaryotic genome annotation.', *Nature reviews. Genetics*, 13(5), pp. 329–42.

Yang (2007) 'PAML 4: Phylogenetic analysis by maximum likelihood', *Molecular Biology and Evolution*, 24, pp. 1586–1591.

Yang, Nielsen, Goldman and Pedersen (2000) 'Codon-substitution models for heterogeneous selection pressure at amino acid sites.', *Genetics*, 155(1), pp. 431–449.

Yao and Carlson (2010) 'Role of G-Proteins in Odor-Sensing and CO 2 -Sensing Neurons in *Drosophila*', 30(13), pp. 4562–4572.

Yao, Ignell and Carlson (2005) 'Chemosensory coding by neurons in the coeloconic sensilla of the Drosophila antenna.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(37), pp. 8359–8367.

Zerbino and Birney (2008) 'Velvet: Algorithms for de novo short read assembly using de Bruijn graphs', *Genome Research*, 18(5), pp. 821–829.

Zhang (2008) 'Progress and challenges in protein structure prediction', *Current Opinion in Structural Biology*, 18(3), pp. 342–348.

Zhao, Zhang, Hoon, Chandrashekar, Erlenbach, Ryba and Zuker (2003) 'The receptors for mammalian sweet and umami taste', *Cell*, 115, pp. 255–266.

Zhou (2010) 'Odorant-Binding Proteins in Insects', *Vitamins and Hormones*, 83(10),pp 241-272

Zhou, Stone, Mackay and Anholt (2009) 'Plasticity of the Chemoreceptor Repertoire in *Drosophila melanogaster*', *PloS Genetics* 5(10),e100681.

Zwiebel and Takken (2004) 'Olfactory regulation of mosquito – host interactions', *Insect Biochem and Mol Biol.* 34(7), pp. 645–652.