Genome sequence of the tsetse fly (Glossina morsitans)

Swain, Martin Thomas

Published in:
Science

DOI:
10.1126/science.1249656

Publication date:
2014

Citation for published version (APA):
Swain, M. T. (2014). Genome sequence of the tsetse fly (Glossina morsitans): Vector of African trypanosomiasis. *Science*, *344*(6182), 380-386 . https://doi.org/10.1126/science.1249656

Title: Genome Sequence of the Tsetse Fly (*Glossina morsitans*): Vector of African Trypanosomiasis

Authors: International *Glossina* Genome Initiative[1]

Affiliations:
[1]Membership of the International *Glossina* Genome Initiative is provided in the Acknowledgments

*Correspondence to:

Serap Aksoy (serap.aksoy@yale.edu) and Geoffrey Attardo (geoffrey.attardo@yale.edu), Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510

Matt Berriman (mb4@sanger.ac.uk), Wellcome Trust Sanger Institute

Abstract: Tsetse flies are the sole vectors of African trypanosomiasis throughout Sub-Saharan Africa. Both sexes of adult tsetse rely on a vertebrate blood specific diet and in doing so contribute to disease transmission. Notable differences between tsetse and other disease vectors include their symbioses with multiple microbes, viviparous reproduction and lactation. Here we describe the sequence and annotation of the *Glossina morsitans morsitans* genome with an emphasis on findings that highlight the differences between tsetse and its dipteran relatives, and on aspects of their biology that have potential for disease control. This analysis has uncovered multiple discoveries including the chromosomal insertions of bacterial (*Wolbachia*) gene sequences, a novel family of lactation specific proteins and a reduction in the number of pathogen recognition receptors and olfaction/chemosensory associated genes. The availability of this genome data provides a foundation for research into trypanosomiasis prevention and yields important insights with broad implications for multiple aspects of biology.

One Sentence Summary: Annotation of the tsetse fly genome reveals novel genetic adaptations associated with the unique biology and vector capacity of this insect.

Main Text: African trypanosomiasis affects humans and livestock throughout Sub-Saharan Africa with an estimated 70 million people at risk of infection (*1*). Rearing livestock in tsetse-infested areas is difficult to impossible, and results in an estimated economic loss of 4-4.5 billion US dollars per year (*2*). Human infections are fatal if untreated, and tools for disease control are limited. There are no vaccines, and current trypanosidal drug treatments have undesirable side effects with growing reports of parasite drug resistance (*3*). The sole vector of African trypanosomes is the tsetse fly, and approaches that reduce or eliminate vector populations are highly effective for disease control (*4*).

Tsetse flies belong to the order of true flies (Diptera), and are members of the superfamily Hippoboscoidea, which are defined by their ability to nourish intrauterine offspring from glandular secretions and give birth to fully developed larvae (obligate adenotrophic viviparity). All members of the Hippoboscoidea are exclusive blood feeders (*5*, *6*). Tsetse are specific to the Glossinidae family (fig. S1) (*7*). These flies acquire trypanosome infections by blood feeding from an infected vertebrate host. Trypanosome transmission via tsetse is a complex process as the parasite must overcome multiple host immune barriers to establish an infection within the fly. As a result trypanosome infection prevalence is low in field populations and experimentally infected tsetse (*8*), indicating presence of a strong natural resistance against parasite infection and transmission. Tsetse also carry obligate microbes, which compensate for their restricted diet and influence multiple aspects of their immune and reproductive physiology (*9-12*).

In 2004, the International *Glossina* Genome Initiative (IGGI) was formed (*13*) to develop research capacity for *Glossina*, particularly in sub-Saharan Africa, through the generation and distribution of molecular resources, bioinformatics training, and the expansion of the *Glossina* research community. An outcome of the effort undertaken by IGGI is the production of the annotated *Glossina morsitans* genome presented here and several satellite papers on genomic and functional biology findings that reflect the unique biology of this disease vector (see Tsetse Biology Collection in PLoS NTDs).

Characteristics of the *Glossina* genome:

The 366 Mb *Glossina morsitans morsitans* genome was assembled into 13,807 scaffolds of up to 25.4 Mb (with mean and N50 sizes of 27 and 120 kb, respectively) and is more than twice the size of the *Drosophila melanogaster* genome (Fig. 1a and table S3). When using a 10 kb resolution threshold for detecting conserved synteny, blocks of synteny comprise at least 63 Mb and 28 Mb in the *Glossina* and *Drosophila* genomes, respectively, with the *Glossina* blocks tending to be twice the size of their equivalents in *Drosophila.* The larger regions of synteny in *Glossina* may be attributed to larger introns and an increase in the size of intergenic sequences as a result of possible transposon activity and or repetitive sequence expansions. The *Glossina* genome is estimated to contain 12,220 protein encoding genes based on automated and manual annotations. Although this number is slightly less than *Drosophila*, the average gene size in *Glossina* is almost double that of *Drosophila* (Fig. 1b). The number of exons and their average size is roughly equivalent in both fly species (Fig. 1c) but the average intron size in *Glossina* appears roughly twice that of *Drosophila* (Fig. 1d).

Orthologous clusters of proteins were generated by comparing the Glossina protein sequences to 5 other complete Dipteran genomes (*Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus*, and *Phlebotomus papatasi* peptides). Each cluster contained proteins or groups of paralogs from at least two taxon; groups from single taxa where considered species specific paralogs.

In total 9,172 (74%) of *Glossina* genes (from 8,374 orthologous clusters) were found to have a Diperan ortholog; 2,803 genes (23%) had no ortholog/paralog and 482 (4%) had a unique duplication/paralog in Glossina. The analysis of genes in orthologous gene clusters across the Diptera (Fig. 2A.) shows that 94% (7,867/8374) of clusters containing a Glossina gene also contained an ortholog with Drosophila (Fig. 2B).

Blood feeding and nutrition: A major difference between *Glossina* and its blood feeding relatives in the sub-order Nematocera (such as mosquitoes and sand flies) is that both male and female *Glossina* utilize blood as their sole source of nutrients and energy. Blood feeding in *Glossina* evolved independently from these other flies. This is reflected in the differing architecture of the mouthparts of tsetse versus mosquitoes to enable pool versus capillary feeding respectively (*14*), as well as in the development of different salivary and digestive physiologies to deal with the challenges associated with blood feeding (*15, 16*).

Adult tsetse have an armament of salivary molecules which are essential for efficient blood feeding and digestion. These molecules counteract the complex physiological responses by the host that impede blood feeding (i.e. coagulation, blood platelet aggregation and vasoconstriction) (table S4) (*15, 17*). The *tsal* gene family, encodes high affinity nucleic acid binding proteins without strong endonuclease activity (*18*) and are the most abundant proteins in the *Glossina* sialome (*15*). Orthologs to *tsal* are not found in *Drosophila* but are present in sand fly and mosquito species (*Culex* but not *Aedes* or *Anopheles*). In mosquitoes and sand flies, a single gene is responsible for the production of salivary endonucleases (*19*). Genomic analysis of the *Glossina tsal* genes reveals three distinct genes (*GMOY012071*, *GMOY012361* and *GMOY012360*) that co-localize to a single 10 kb genomic locus. It remains unknown why *Glossina*, unlike other blood feeding insects, has developed a highly expressed salivary gene family specialized in nucleic acid binding rather than hydrolysis.

Another family of abundant salivary gland proteins is related to adenosine deaminases and insect growth factors (ADGFs), which are thought to reduce the inflammation/irritation resulting from adenosine and inosine-induced mast cell activation. In tsetse, the ADGF genes are uniquely organized as a cluster of 4 genes in a 20 kb genomic locus (*GMOY002973-1,2,3,4*). An adenosine deaminase (ADA) gene (*GMOY008741*) without the putative growth factor domain is encoded elsewhere in the genome. In *Drosophila*, 5 ADGF genes can be found in various loci and have been associated with developmental regulation (*20*), while nematoceran insects such as sand flies and mosquitoes have a reduced number with a maximum of 3 genes. In arthropods such as *Ixodes scapularis*, *Rhodnius prolixus* and *Pediculus humanus* only *bona fide* ADAs can be found.

One critical point of tsetse fly interaction with African trypanosomes is in the establishment within the salivary glands. Recent studies have shown specific genes and proteins are suppressed within salivary glands during parasite infection and these are critical to trypansome transmission (*21*). By utilizing RNA-seq, a complete analysis of salivary gland gene expression during parasite infection has been determined, see satellite Telleria et al. (2013)(*22*). This analysis confirms the reduced transcript abundance of previously identified genes such as adenosine deaminases, tsal1 + 2 and 5' nucleotidase, as well as predicting the reduction of many other secreted salivary peptides of unknown function by trypanosome infection (*22*). Genes with increased expression are those involved in stress tolerance and cell repair, suggesting significant damage to the salivary glands due to the presence of trypanosomes.

Upon blood meal ingestion, the peritrophic matrix (PM) separates and protects the midgut epithelium from damaging or toxic dietary elements, allows for controlled digestion and metabolism of the blood meal and is a barrier against trypanosome infection and establishment (*9*). *Glossina* produces a type-II PM, which is secreted continuously as concentric sleeves by the proventriculus and separates the lumen of the midgut (endoperitrophic space) from the monolayer of epithelial cells (*23, 24*). Type-II PMs are generally composed of chitin, peritrophins proteins, glycosaminoglycans (GAGs) and mucin-like molecules. We identified seven peritrophin genes from the *Glossina* genome, three of which are exclusively expressed by the proventriculus (table S5).

Glossina takes a blood meal that is equivalent to its own weight and to mitigate this burden, blood is rapidly concentrated and excess water excreted. The aquaporin family of transport proteins is critical for selectively transporting large volumes of water molecules and other solutes across cellular membranes (*25*). Ten aquaporin genes (AQPs) were identified in *Glossina*, compared to eight and six in *Drosophila* and mosquitoes, respectively (*26*) (table S6). Although no one class of AQPs has undergone expansion in *Glossina*, two individual genes have been duplicated: the AQP2 gene and the homologue of the *Drosophila* integral protein (Drip).

Experimental evidence indicates that multiple AQPs are critical for water homeostasis during blood feeding and milk synthesis (*27*).

The mobilization, utilization and metabolism of nutrients also differ in *Glossina* in comparison to mosquitoes and sandflies. *Glossina* has a marked reduction in genes associated with carbohydrate metabolism (Fig. 3), instead utilizing a proline-alanine shuttle system for energy distribution and triglycerides/diglycerides for energy storage within the fat body and milk secretions. Little to no sugar or glycogen is detectable in these flies (*28*). Genes involved in lipid metabolism are generally conserved with expansions in genes associated with fatty acid synthase, fatty acyl-CoA reductase and 3-keto acyl-CoA synthase functions. In addition, 3 multivitamin transporters from the Solute:Sodium Symporter (SSS) family are found in *Glossina* and mosquitoes, but not in *Drosophila* suggesting that they may assist in blood meal metabolism (table S31).

Microbiome: *Glossina* harbor multiple maternally transmitted microorganisms, the relationships of which range from mutualistic to parasitic. The most vital of these is the mutualistic *Wigglesworthia glossinidia,* which resides intracellularly in the midgut-associated bacteriome organ, but extracellularly in the milk produced by accessory glands of females. The putative proteome of *Wigglesworthia* indicates the capacity for B vitamin biosynthesis including: biotin (B7), thiazole (B1), riboflavin (B2), pantothenate (B5), pyridoxine (B6) and folate (B9) (*29*). The nutritional supplementation of the host's restricted diet is an essential role in this symbiosis as females without *Wigglesworthia* prematurely abort their larval offspring. Larva that have undergone intrauterine development in the absence of *Wigglesworthia* (facilitated by blood meal supplementation with yeast extract) give rise to adults that lack immune cells (phagocytes) in the hemolymph (*10, 12*), have a compromised midgut peritrophic matrix barrier (*9*) and are highly susceptible to infection with parasites (*30*). This suggests an additional role for the symbiont in the development of host immune physiology (*11*).

Laboratory lines as well as some natural populations of tsetse also harbor the commensal bacteria *Sodalis glossinidius,* which are found both intra and extracellularly in the fly. The thiamine biosynthetic capacity differs between the *Wigglesworthia* and *Sodalis* genomes. *Wigglesworthia* are capable of synthesizing thiamine in the form of thiamine monophosphate (TMP). *Sodalis* and *Glossina* lack this capability; however they have thiamine transporters. *Glossina* carries a gene for a member of the reduced folate carrier family which has thiamine binding capabilities (GMOY009200) and a folate transporter (GMOY005445). Sodalis has a thiamine ABC transporter (tbpAthiPQ) capable of scavenging free thiamine produced by *Wigglesworthia* (*31*).

The third endosymbiont present in some natural *Glossina* populations (and in the strain sequenced here) is *Wolbachia,* which resides in gonadal tissues. Laboratory studies have shown that this *Wolbachia* strain induces cytoplasmic incompatibility (CI) in tsetse (*32*). In addition to cytoplasmic infection, multiple horizontal transfer events (HTEs) from *Wolbachia* were detected in *Glossina* chromosomes. Examination of *Glossina* contigs indicated the presence of at least three different HTEs (A, B and C). Insertions A and B are the largest in size respectively carrying a total of 197 and 159 putative functional protein-coding genes. *In situ* staining of *Glossina* mitotic chromosomes with *Wolbachia* specific DNA probes localized multiple insertions on the X, Y and B chromosomes (table S7), see satellite paper Brelsfoard et al., (*33*). In addition, HTEs representing sequences from most of the major groups of both retrotransposons and DNA transposons were identified in the *Glossina* genome contigs (table

S8). These sequences comprised approximately 14% of the assembled genome, in contrast to only 3.8% of the *Drosophila* euchromatic genome (*34*).

Many *Glossina* species, including the strain sequenced here, harbor a large DNA hytrosavirus, the *Glossina pallidipes* Salivary Gland Hypertrophy Virus GpSGHV (*35*). The virus can reduce fecundity and lifespan in *Glossina* and cause salivary hypertrophy at high densities. Strong evidence of viral exposure was discovered during analysis of a group of genes lacking Dipteran orthologs. The analysis resulted in identification of many putative bracoviral genes (BLAST E-values of <1E-50) spread over 151 genomic scaffolds. The putative bracoviral sequences bear highest homology to those identified from the parasitic braconid wasps *Glyptapanteles flavicoxis* and *Cotesia congregata*. This suggests that *Glossina* was parasitized by an unidentified braconid wasp. The natural history of this relationship remains unknown and requires further study.

Immunity: Multiple factors including age, sex, nutritional status and the presence of symbiotic fauna have been shown to influence tsetse's vector competence at the time of parasite acquisition (*36*). Among the pathways and effectors validated as important to tsetse's observed resistance to parasites are the peptidoglycan recognition proteins (PGRPs) (*37, 38*), the innate immune signaling pathway IMD (Immune deficiency) produced effector antimicrobial peptides (AMPs) (*39, 40*), midgut lectins (*41*), antioxidants (*42*), EP-protein (*43*) and the gut peritrophic matrix structure (*9*).

Microbial detection is a multistep process that requires direct contact between host pattern recognition receptors (PRRs) and pathogen associated molecular patterns (PAMPs). *Drosophila* has 13 peptidoglycan recognition proteins (PGRPs), which play a role in the recognition of peptidogycan (PGN), an essential component of the cell wall of virtually all bacteria (*44*). In *Glossina*, only seven PGRPs were identified, four in the long subfamily (PGRP-LB, -LC, -LD and –LA) and two in the short subfamily (PGRP-SB and –SA), while *Drosophila* has a gene duplication resulting in two related forms of PGRP-SB (Figs. 4a and 4b). *Glossina* also lacks homologs of receptors LE, SD, SC, and LF, based on both genome annotation and transcriptome data. The reduced PGRP repertoire of *Glossina* may reflect the nature of its sterile blood diet, which likely exposes the tsetse gut to fewer microbes relative to *Drosophila*. In the *Drosophila* gut, PGRP-LE functions as the master bacterial sensor, which induces balanced responses to infectious bacteria and tolerance to microbiota by up-regulation of negative regulators of the IMD pathway including PGRP-LB (*45*). In the case of *Glossina*, loss of amidase -SC1 along with PGRP-LE may indicate the presence of a streamlined gut immune response, possibly to protect symbiosis with intracellular *Wigglesworthia*. A reduced immune capacity is also observed in Aphids, another group of insects harboring obligate symbionts (*46*). A complete listing of orthologs to *Drosophila* immune genes is presented in table S9.

Reproduction and Developmental Biology: The reproductive biology of tsetse is unique to the Hippoboscoidea superfamily (*6*). The evolution of adenotrophic viviparity (intrauterine larval development and nourishment by glandular secretions) required dramatic adaptations to reproductive physiology that included ovarian follicle reduction (2 follicles per ovary relative to 30-40 in *Drosophila*), expansion and adaptation of the uterus to accommodate developing larvae and adaptation of the female accessory gland to function as a nutrient synthesis and delivery system (*47*).

Glossina and *Drosophila* both use lipase derived yolk proteins for vitellogenesis, unlike non-brachyceran flies that utilize the vitellogenin family of yolk proteins (*48, 49*). However, *Glossina* has a much lower rate of oogenesis than *Drosophila* and other flies in the Brachycera suborder. Unlike *Drosophila*, which has 3 yolk protein genes (*yp1, yp2* and *yp3*) localized on the X chromosome, *Glossina* has only a single yolk protein gene yp2 ortholog (GMOY002338) that is expressed only in the ovaries and lacks fat body associated expression. Multiple yolk proteins have been identified in other cyclorrhaphan flies, suggesting that *Glossina* may have lost these genes in association with its reduction in reproductive capacity (*48, 50*).

In *Drosophila*, the *male specific lethal* (*MSL*) complex is required for X chromosome dosage compensation (*51, 52*). Glossina is thought to utilize a similar dosage compensation system. Orthologs of the five MSL proteins are present in the *Glossina* genome. Protein motifs identified as important for interaction between the MSL proteins (*53, 54*) are also well conserved in the *Glossina* orthologs. However, the motifs associated with X chromosome binding in *Drosophila* (e.g. MSL1 amino terminal end (*55*)) are not well conserved. This suggests that the *Glossina* MSL complex is likely binding to quite a different DNA sequence than that recognized by the *Drosophila* complex (*56*) (table S10).

A critical process in development is the determination of embryo anterior/posterior polarity. Absent from the *Glossina* genome are both the *bicoid* and the *nanos* genes, which are responsible for the well-defined anterior and posterior embryonic polarity system in *Drosophila* (*57, 58*). Orthologs for these genes were not found in the genomic scaffolds or in *de novo* assemblies created using Illumina data from reproductively active whole female flies. Orthologs to genes immediately flanking the 5' and 3' ends of the *bicoid* and *nanos* loci in *Drosophila* are present in the *Glossina* assembly. This polarity mechanism is thought to be specific to the Brachycera (*59*). These findings suggest that the conservation of this system between *Drosophila* and other Brachycera may not be as well defined as previously thought. Other insects determine embryonic polarity through a gradient of maternal RNA for orthologs of the *ocelliless/orthodenticle* (*oc/otd*) (GMOY006617) and *hunchback* (*hb*) (GMOY004735) genes both of which are present in *Glossina (60).*

Glossina larvae are dependent upon their mother's accessory gland (milk gland) secretions for their nutrition as well as for transfer of symbiotic fauna (*61*). This gland is highly specialized and is responsible for integrating a complex mixture of stored lipids and milk proteins. The water in the milk is provided by two aquaporins (DripA and DripB), and RNAi knockdown of these genes results in dehydration of the intrauterine larva, see satellite paper Benoit et al. (*27*). Characterized milk proteins include a Lipocalin (*mgp1*) (*62*), Transferrin (*trf*) (*63*), an acid sphingomyelinase (*asmase*) (*64*), milk proteins mgp2 and -3 (*65*) and peptidoglycan recognition protein LB (*PGRP-LB*) (*37*) (Fig. 5). Many of these proteins are functional analogs of milk proteins identified in placental mammals and marsupials. Annotation of the mgp2 and -3 genomic loci identified an additional cluster of 7 genes that appear to be paralogs, with identical tissue and stage specific expression patterns to *mgp2* and *mgp3*, see satellite paper Benoit et al. (*66*). The milk proteins may function as lipid emulsification agents, sources of amino acids and possibly phosphate (table S11)(*66*). The 12 milk genes accounts for nearly 50% of the transcriptional investment during lactation, which is a source of substantial oxidative stress to the mother (*66*). This stress is counteracted by an antioxidant response which is critical to allow fecundity late into the tsetse lifetime, see satellite paper Michalkova et al. (*67*). Analysis of the predicted promoter sequences of the milk proteins revealed the conservation of homeodomain protein binding sites. Annotation of *Glossina* homeodomain factors (table S35) revealed the presence of a homeodomain protein *ladybird late*, which is expressed exclusively in the milk gland of adult female flies. Knockdown of this factor results in a global reduction of milk gland protein expression, see satellite paper Attardo et al., (*68*) suggesting that this factor is an important regulator of these genes during pregnancy.

Sensory genes as targets for *Glossina* control strategies: Different species of *Glossina* display strong host preferences and vary in their response to cues from different mammalian hosts. The primary hosts of *G. m. morsitans* are ungulates. *Glossina* utilizes both chemical and visual cues to find vertebrate hosts and potential mates. In insects, including tsetse, chemical cues are detected by a suite of proteins which include odorant binding proteins (OBP), chemosensory proteins (CSP), odorant receptors (OR), gustatory receptors (GR) ligand gated ionotropic receptors (IR), sensory neuron membrane proteins (SNMP), and CD36-like pheromone sensors (*69-73*). These proteins capture and decode ecological signals to drive appropriate behavioral responses including host-seeking, oviposition, mate searching, and detection of predators.

Glossina has an overall reduction in olfactory proteins relative to *Drosophila, Anopheles gambiae* and *Apis mellifera* (Table 1 and see satellite paper Obiero et al.) (*74*), that could result from the less complex ecology of tsetse and their restricted food preference (vertebrate blood). Their narrow host range has probably negated the need for an expanded array of chemical sensors. This is in contrast to mosquitoes, which in addition to feeding on blood also use plant sugars for energy, thus requiring greater complexity in these sensory systems.

The visual system of *Glossina* conforms to that of other well-characterized calyptrate Diptera, such as the house fly *Musca domestica* and the blow fly *Calliphora vicina*, all of which are fast flying species (*75*). *Glossina* is readily attracted to blue/black colors, a behavior which has been widely exploited in targets and traps to reduce vector populations. There is a great degree of conservation of retinal morphology throughout the Brachycera, allowing for direct comparisons with *Drosophila* (for review see *76*). The lack of sexual dimorphism in tsetse eye morphology (*75, 77*) is consistent with the fact that both sexes employ vision for host identification and pursuit (*78*). The males, however, also depend on vision for long-distance identification and pursuit of female mating partners (*79*).

Glossina has orthologs of four of the five opsin genes that are expressed in the *Drosophila* retina: Rh1, Rh3, Rh5 and Rh6. The finding of a Rh5 opsin ortholog in *Glossina* is the first experimental evidence for the presence of blue-sensitive R8p cells, which were missed in earlier experimental studies (*80*). The *Glossina* genome also contains the ortholog of the *Drosophila* Rh7 opsin gene. The role of Rh7 in eye development and vision has yet to be determined in *Drosophila*. An ortholog of the *Drosophila* ocellus specific Rh2 was not detected. *Glossina* genome data correspond well with the study of opsin conservation and expression in the retina of *C. vicina* (*81*), which has also retained orthologs of Rh1, Rh3, Rh5 and Rh6. The structural/function analysis of these proteins could yield important insights into tsetse's attraction to blue/black. The expanded search for vision associated genes revealed that all of the core components of the photo transduction cascade downstream of the opsin transmembrane receptors are conserved in *Glossina* (table S12).

Future Directions: The assembly and annotation of the *Glossina* genome highlights specific adaptations to the unique biology of this organism (Fig. 6) and provides a foundation to better understand the biology of this unique vector. It also facilitates the application of powerful high throughput technologies in a way that was previously impossible. In addition, genomic and transcriptomic data on five more *Glossina* species (*fuscipes, brevipalpis, palpalis, austeni* and *gambiensis*) is being generated to produce additional genome assemblies. This will allow detailed evolutionary and developmental analyses to study genomic differences associated with host specificity, vectorial capacity and evolutionary relationships.

Acknowledgements

The genome sequence and associated information can be found at GenBank under the accession number ######## and are also hosted on VectorBase (*https://www.**vectorbase**.org/*).

The *Glossina morsitans* genome project was funded by the Wellcome Trust (grants 085775/Z/08/Z and 098051) and by World Health Organisation (WHO) Special Programme for Research and Training in Tropical Diseases (TDR) (Project No. A90088). BAC libraries were generated through NIAID resources. We thank the staff within the library construction, sequence production and informatics support groups at the Wellcome Trust Sanger Institute.

1. P. P. Simarro, J. Jannin, P. Cattand, *PLoS medicine* **5**, e55 (Feb, 2008).
2. P. Holmes, *J. Invertebr. Pathol.,* (Jul 24, 2012).
3. R. Brun, J. Blum, F. Chappuis, C. Burri, *Lancet* **375**, 148 (Jan 9, 2010).
4. S. C. Welburn, I. Maudlin, P. P. Simarro, *Parasitology* **136**, 1943 (Dec, 2009).
5. E. S. Krafsur, *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **9**, 124 (Jan, 2009).
6. R. Meier, M. Kotrba, P. Ferrar, *Biol. Rev. Camb. Philos. Soc.* **74**, 199 (1999).
7. F. T. Petersen, R. Meier, S. N. Kutty, B. M. Wiegmann, *Mol Phylogenet Evol* **45**, 111 (Oct, 2007).
8. M. J. Lehane, S. Aksoy, E. Levashina, *Trends Parasitol.* **20**, 433 (Sep, 2004).
9. B. L. Weiss, J. Wang, M. A. Maltz, Y. Wu, S. Aksoy, *PLoS Path.* **9**, e1003318 (Apr, 2013).
10. B. L. Weiss, M. Maltz, S. Aksoy, *J Immunol* **188**, 3395 (Apr 1, 2012).
11. B. Weiss, S. Aksoy, *Trends Parasitol.* **27**, 514 (Nov, 2011).
12. B. L. Weiss, J. Wang, S. Aksoy, *PLoS Biol.* **9**, e1000619 (May, 2011).
13. S. Aksoy *et al.*, *Trends Parasitol.* **21**, 107 (Mar, 2005).
14. H. W. Krenn, H. Aspock, *Arthropod structure & development* **41**, 101 (Mar, 2012).
15. J. Alves-Silva *et al.*, *BMC Genomics* **11**, 213 (2010).
16. M. J. Lehane *et al.*, *Genome biology* **4**, R63 (2003).
17. J. Van Den Abbeele *et al.*, *Insect Biochem. Mol. Biol.* **37**, 1075 (Oct, 2007).
18. G. Caljon *et al.*, *PloS one* **7**, e47233 (2012).
19. J. M. Ribeiro, B. J. Mans, B. Arca, *Insect Biochem Mol Biol* **40**, 767 (Nov, 2010).
20. T. Dolezal, E. Dolezelova, M. Zurovec, P. J. Bryant, *PLoS Biol* **3**, e201 (Jul, 2005).
21. J. Van Den Abbeele, G. Caljon, K. De Ridder, P. De Baetselier, M. Coosemans, *PLoS Path.* **6**, e1000926 (2010).
22. E. L. Telleria *et al.*, *PLoS Negl Trop Dis* **Submitted**, (2013).
23. M. J. Lehane, P. G. Allingham, P. Weglicki, *Cell Tissue Res.* **283**, 375 (1996).
24. M. J. Lehane, *Annu. Rev. Entomol.* **42**, 525 (1997).
25. E. M. Campbell, A. Ball, S. Hoppler, A. S. Bowman, *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology* **178**, 935 (Nov, 2008).
26. L. L. Drake *et al.*, *PloS one* **5**, e15578 (2010).
27. J. B. Benoit *et al.*, *PLoS Neglected Tropical Diseases* **In press**, (2013).
28. D. A. Norden, D. J. Paterson, *Comp Biochem Physiol* **31**, 819 (Dec 1, 1969).
29. R. V. Rio *et al.*, *MBio* **3**, (2012).
30. R. Pais, C. Lohs, Y. Wu, J. Wang, S. Aksoy, *Appl. Environ. Microbiol.* **74**, 5965 (Oct, 2008).
31. A. K. Snyder, J. W. Deberry, L. Runyen-Janecky, R. V. Rio, *Proc Biol Sci* **277**, 2389 (Aug 7, 2010).
32. U. Alam *et al.*, *PLoS Path.* **7**, e1002415 (Dec, 2011).
33. C. Brelsfoard *et al.*, *PloS Neglected Tropical Diseases* **Submitted**, (2013).
34. J. S. Kaminker *et al.*, *Genome Biol* **3**, RESEARCH0084 (2002).
35. A. M. Abd-Alla *et al.*, *J Virol* **82**, 4595 (May, 2008).
36. S. Aksoy, W. C. Gibson, M. J. Lehane, *Adv. Parasitol.* **53**, 1 (2003).
37. J. Wang, S. Aksoy, *Proc. Natl. Acad. Sci. USA* **109**, 10552 (Jun 26, 2012).
38. J. Wang, Y. Wu, G. Yang, S. Aksoy, *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12133 (Jul 21, 2009).
39. Z. Hao, I. Kasumba, S. Aksoy, *Insect Biochem. Mol. Biol.* **33**, 1155 (Nov, 2003).
40. C. Hu, S. Aksoy, *Mol Microbiol* **60**, 1194 (Jun, 2006).
41. S. C. Welburn, I. Maudlin, D. S. Ellis, *Med. Vet. Entomol.* **3**, 77 (Jan, 1989).
42. E. T. Macleod, I. Maudlin, A. C. Darby, S. C. Welburn, *Parasitology*, 1 (Feb 19, 2007).
43. L. R. Haines, S. M. Lehane, T. W. Pearson, M. J. Lehane, *PLoS Path.* **6**, e1000793 (Mar, 2010).
44. Z. Markiewicz, M. Popowska, *Pol J Microbiol* **60**, 181 (2011).
45. V. Bosco-Drayon *et al.*, *Cell host & microbe* **12**, 153 (Aug 16, 2012).
46. C. G. Elsik, *Genome Biol* **11**, 106 (2010).
47. S. S. Tobe, P. A. Langley, *Annu. Rev. Entomol.* **23**, 283 (1978).
48. K. Hens, P. Lemey, N. Macours, C. Francis, R. Huybrechts, *Insect Mol. Biol.* **13**, 615 (Dec, 2004).
49. P. Romans, Z. Tu, Z. Ke, H. H. Hagedorn, *Insect Biochem. Mol. Biol.* **25**, 939 (1995).
50. M. J. Scott *et al.*, *Genetica* **139**, 63 (Jan, 2011).
51. M. E. Gelbart, M. I. Kuroda, *Development* **136**, 1399 (May, 2009).
52. T. Conrad, A. Akhtar, *Nat Rev Genet* **13**, 123 (Feb, 2012).

53. M. J. Scott, L. L. Pan, S. B. Cleland, A. L. Knox, J. Heinrich, *EMBO J.* **19**, 144 (2000).
54. J. Kadlec *et al.*, *Nat Struct Mol Biol* **18**, 142 (Feb, 2011).
55. F. Li, D. A. Parry, M. J. Scott, *Mol Cell Biol* **25**, 8913 (Oct, 2005).
56. A. A. Alekseyenko *et al.*, *Cell* **134**, 599 (Aug 22, 2008).
57. M. Hulskamp, C. Pfeifle, D. Tautz, *Nature* **346**, 577 (Aug 9, 1990).
58. V. Irish, R. Lehmann, M. Akam, *Nature* **338**, 646 (Apr 20, 1989).
59. M. Stauber, S. Lemke, U. Schmidt-Ott, *Dev Genes Evol* **218**, 81 (Feb, 2008).
60. R. Schroder, *Nature* **422**, 621 (Apr 10, 2003).
61. G. M. Attardo *et al.*, *J. Insect Physiol.* **54**, 1236 (Aug, 2008).
62. G. M. Attardo, N. Guz, P. Strickler-Dinglasan, S. Aksoy, *J. Insect Physiol.* **52**, 1128 (Nov-Dec, 2006).
63. N. Guz, G. M. Attardo, Y. Wu, S. Aksoy, *J. Insect Physiol.* **53**, 715 (Jul, 2007).
64. J. B. Benoit *et al.*, *Biol. Reprod.* **87**, 17 (2012).
65. G. Yang, G. M. Attardo, C. Lohs, S. Aksoy, *Insect Mol. Biol.* **19**, 253 (Jan 28, 2010).
66. J. B. Benoit *et al.*, *PLoS Genet.* **In Press**, (2013).
67. V. Michalkova, J. B. Benoit, G. M. Attardo, J. Medlock, S. Aksoy, *PLoS One* **Submitted**, (2013).
68. G. M. Attardo *et al.*, *PLoS Biol.* **Submitted**, (2013).
69. R. Benton, K. S. Vannice, C. Gomez-Diaz, L. B. Vosshall, *Cell* **136**, 149 (Jan 9, 2009).
70. C. Liu *et al.*, *PLoS Biol.* **8**, (2010).
71. R. Benton, *Curr. Opin. Neurobiol.* **18**, 357 (Aug, 2008).
72. R. Liu *et al.*, *Insect Mol. Biol.* **21**, 41 (Feb, 2012).
73. R. Liu *et al.*, *Cellular and molecular life sciences : CMLS* **67**, 919 (Mar, 2010).
74. G. F. O. Obiero *et al.*, *PLoS Neglected Tropical Diseases* **Submitted**, (2013).
75. R. Hardie, K. Vogt, A. Rudolph, *J. Insect Physiol.* **35**, 423 (1989).
76. M. Friedrich, in *Encyclopedia of Life Sciences*. (John Wiley & Sons, Ltd, Chichester, 2010), pp. DOI: 10.1002/9780470015902.a0022898.
77. D. A. Turner, J. F. Invest, *Bull. Ent. Res.* **62**, 343 (1973).
78. G. Gibson, S. J. Torr, *Med Vet Entomol* **13**, 2 (1999).
79. J. Brady, *Physiol. Entomol.* **14**, 153 (1991).
80. E. K. Buschbeck, N. J. Strausfeld, *The Journal of comparative neurology* **383**, 282 (Jul 7, 1997).
81. A. Schmitt, A. Vogt, K. Friedmann, R. Paulsen, A. Huber, *J Exp Biol* **208**, 1247 (Apr, 2005).

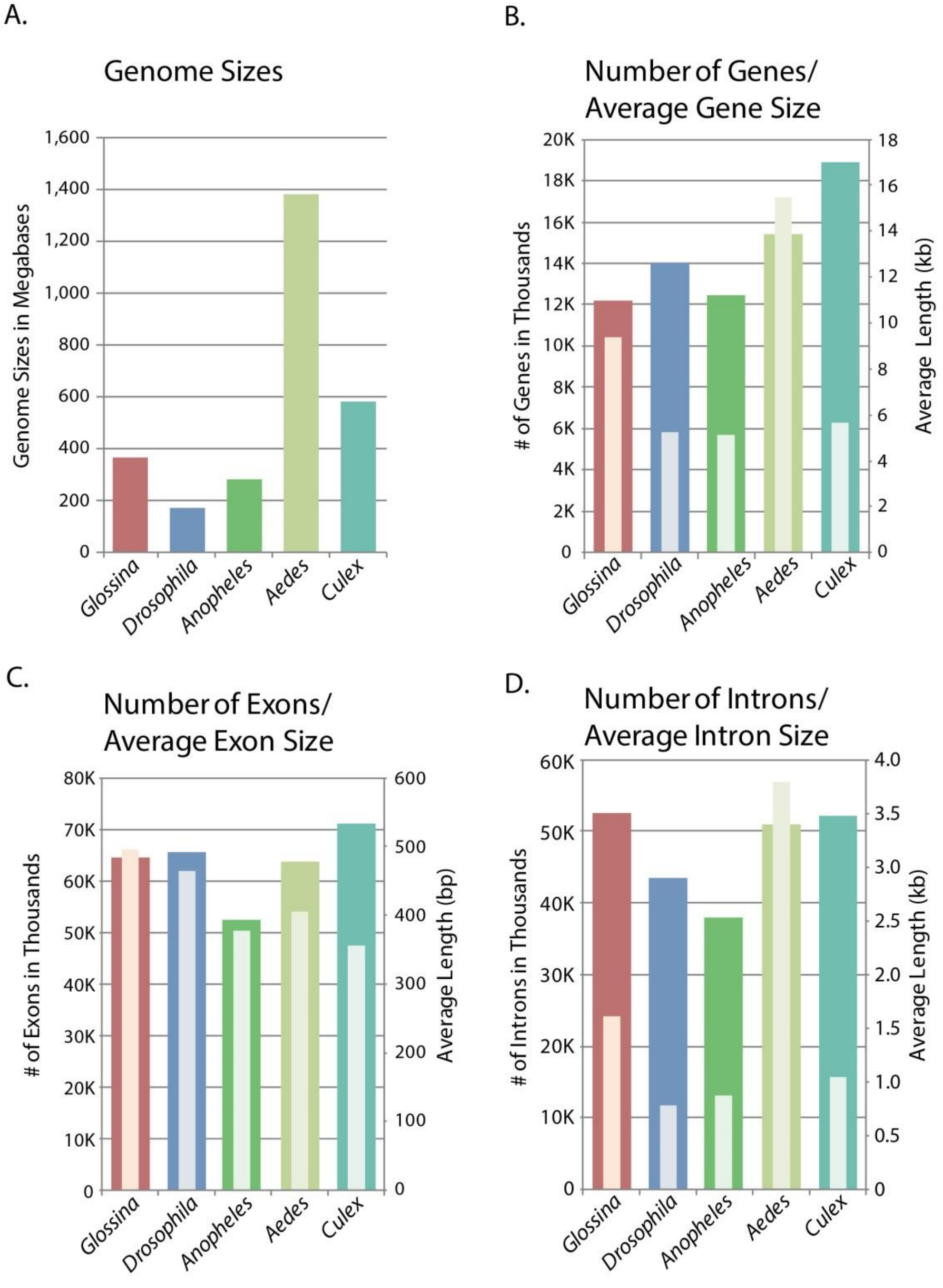


Figure 1: Overview comparing genomic statistics from *Glossina* with *Drosophila melanogaster*, *Aedes aegypti*, *Culex quinquefaciatis*, and *Anopheles gambiae*. In figures B-D thick bars are associated with the left axis and thin bars are associated with the right axis. A. Comparison of genome sizes, B. Comparison of the number and length of gene predictions, C. Comparison of the number and length of exons, D. Comparison of the number and length of introns.

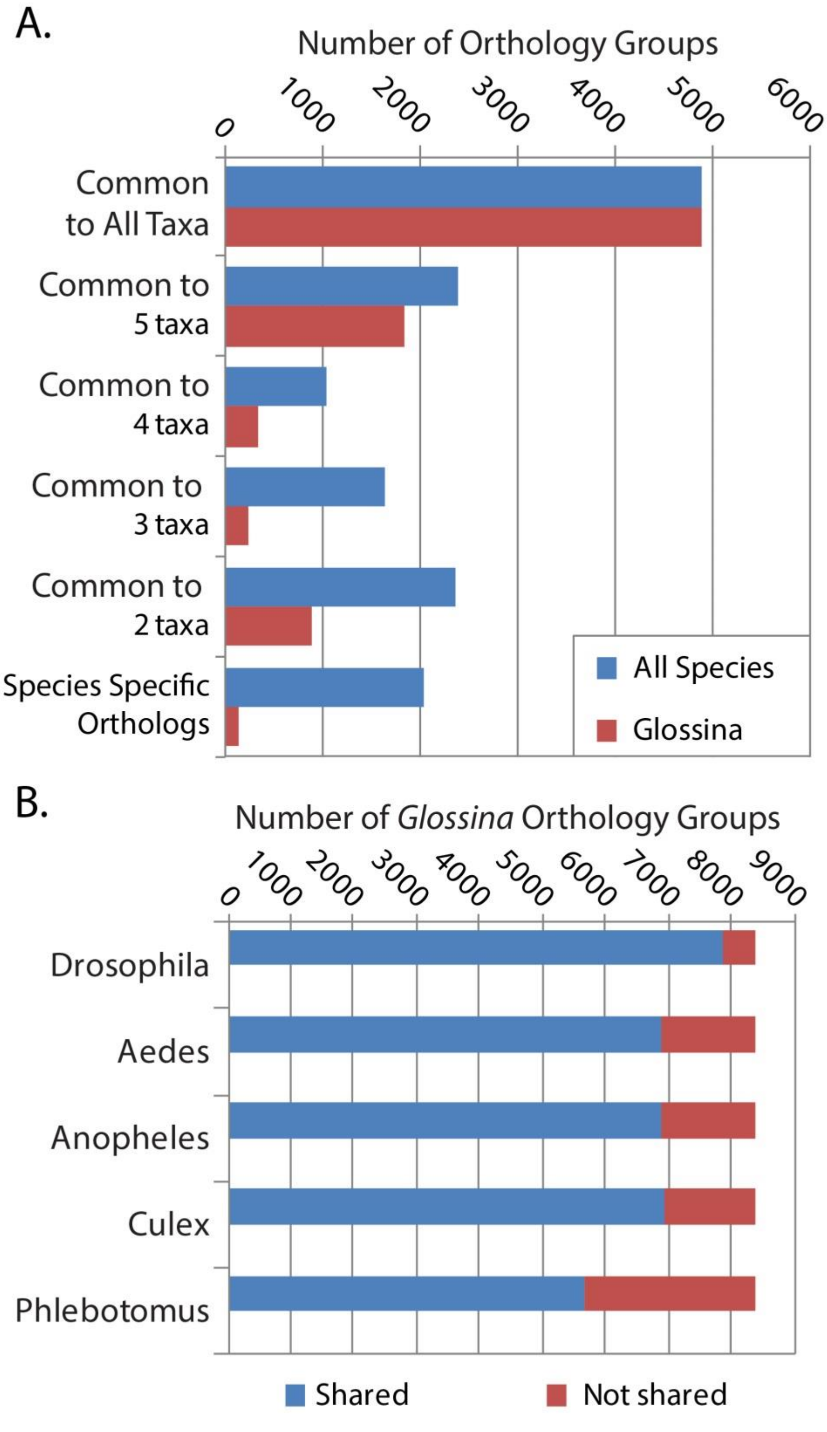


Figure 2: Orthology Analysis.

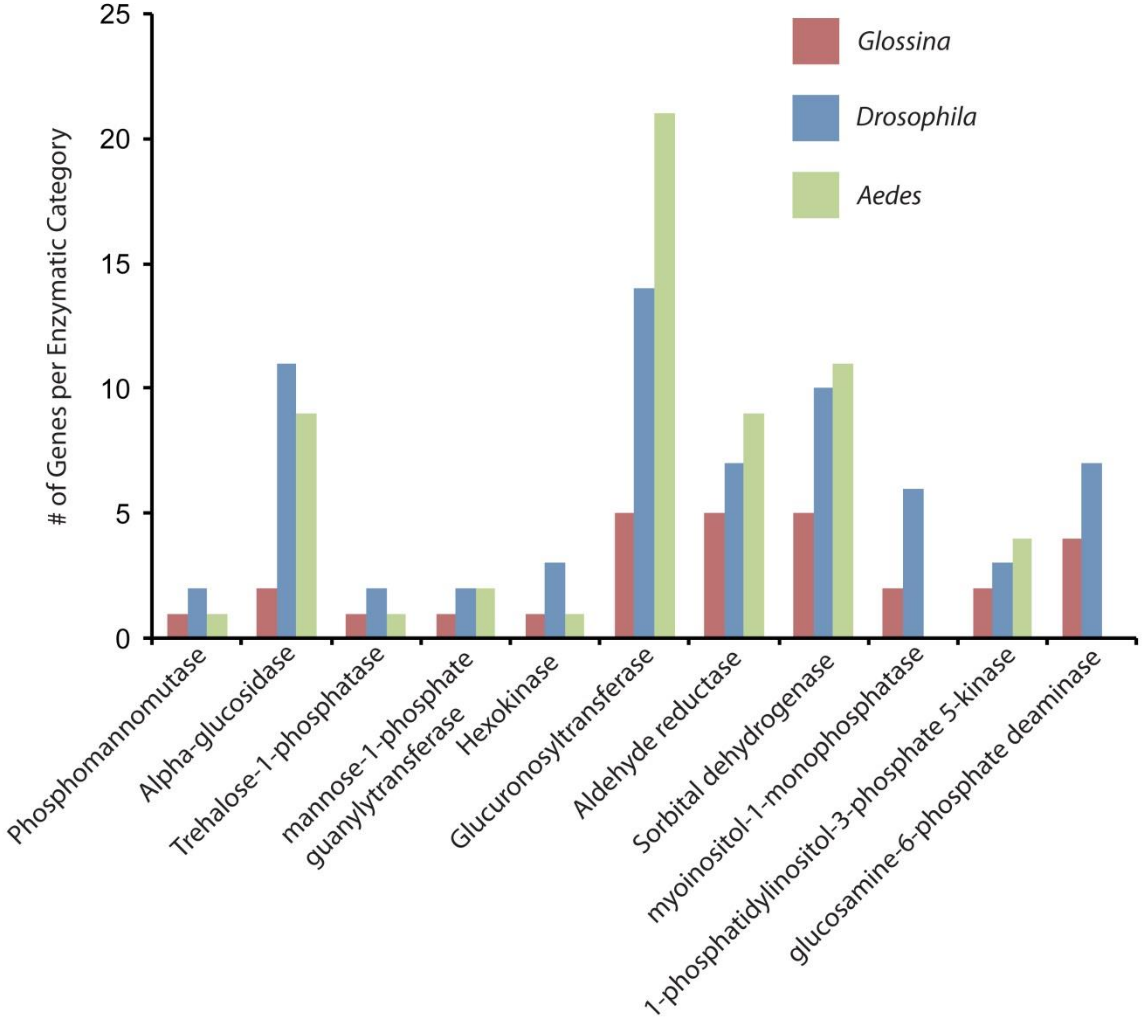


Figure 3: Comparison of carbohydrate metabolism and vitamin transporter genes between fly species. Number of genes associated with different carbohydrate metabolic enzyme activities from *Glossina*, *Drosophila melanogaster* and *Aedes aegypti*.

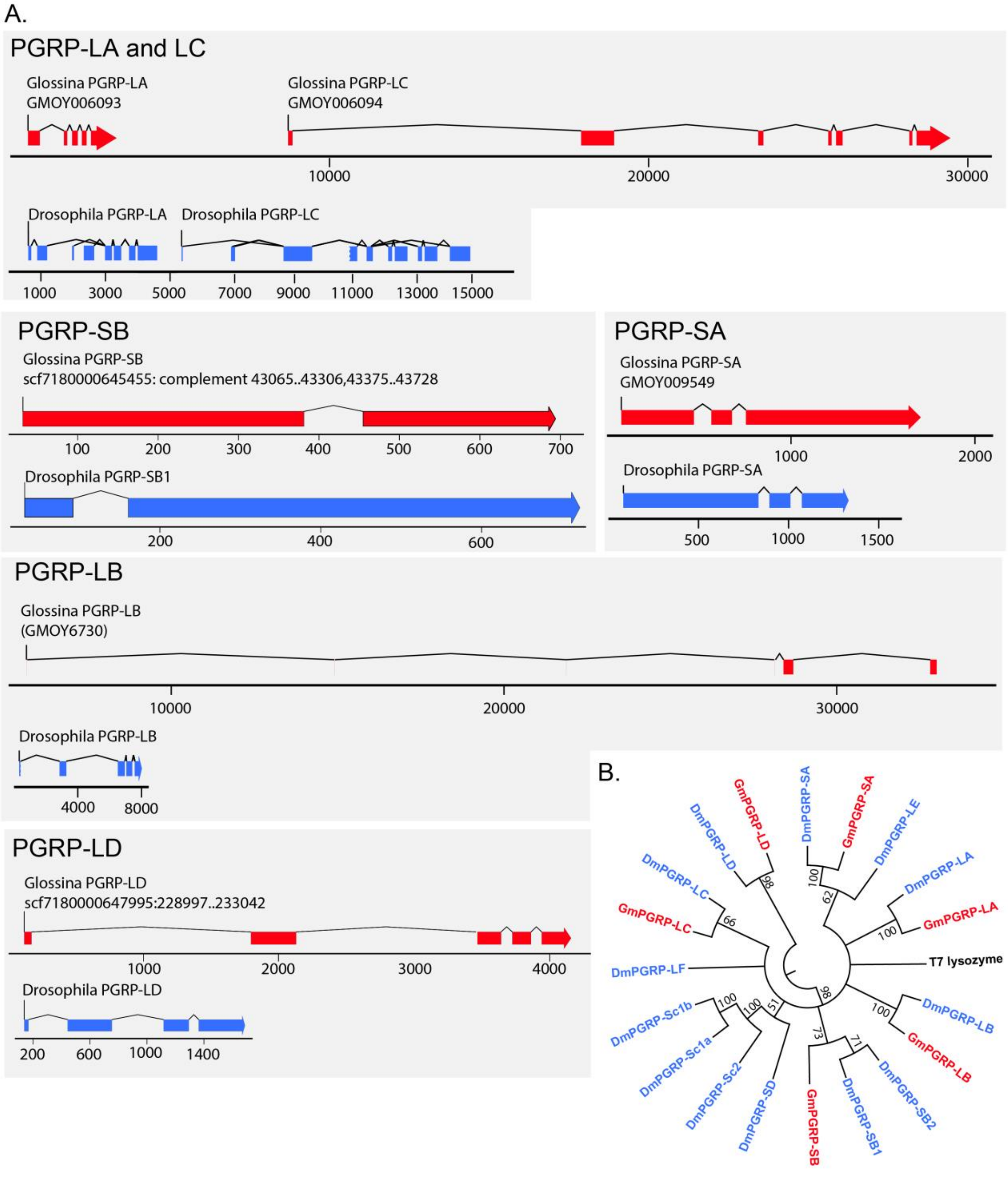


Figure 4: Gene structure and phylogeny of Glossina PGRP genes. A. Schematic of gene structure of the *Glossina* PGRP genes. **B.** Phylogenic comparison of *Glossina* and *Drosophila* PGRPs. The tree was generated using MEGA5 following a hand edited MUSCLE alignment. The tree was generated using neighbor joining based on p-distance using partial deletion with a site coverage cutoff of 50%. Bootstrap analysis was performed with 1000 replications. The tree is condensed to only show bootstrap values over 50%.

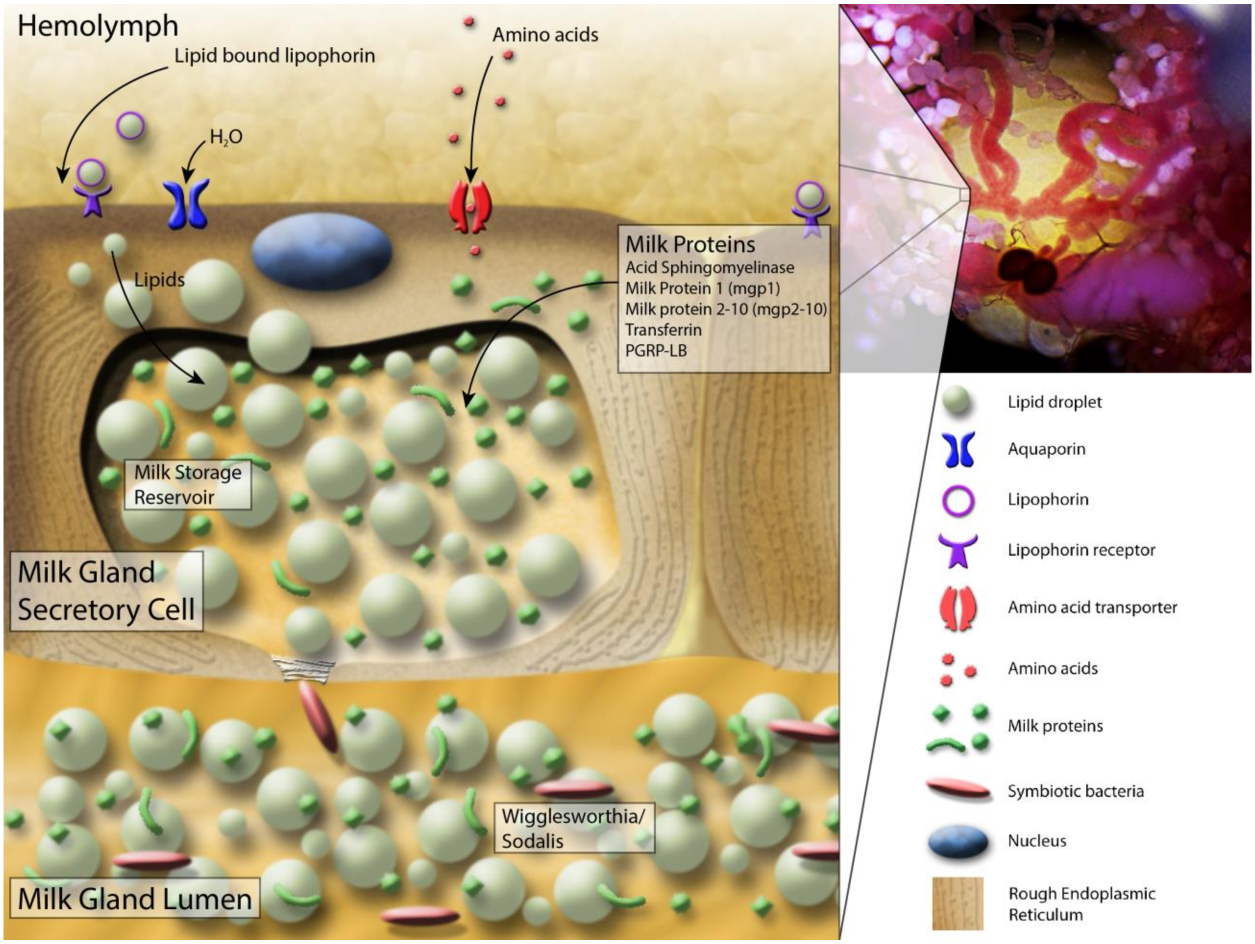


Figure 5: Overview schematic of milk gland secretory cell physiology and milk production with associated milk proteins and nutrient transporters.

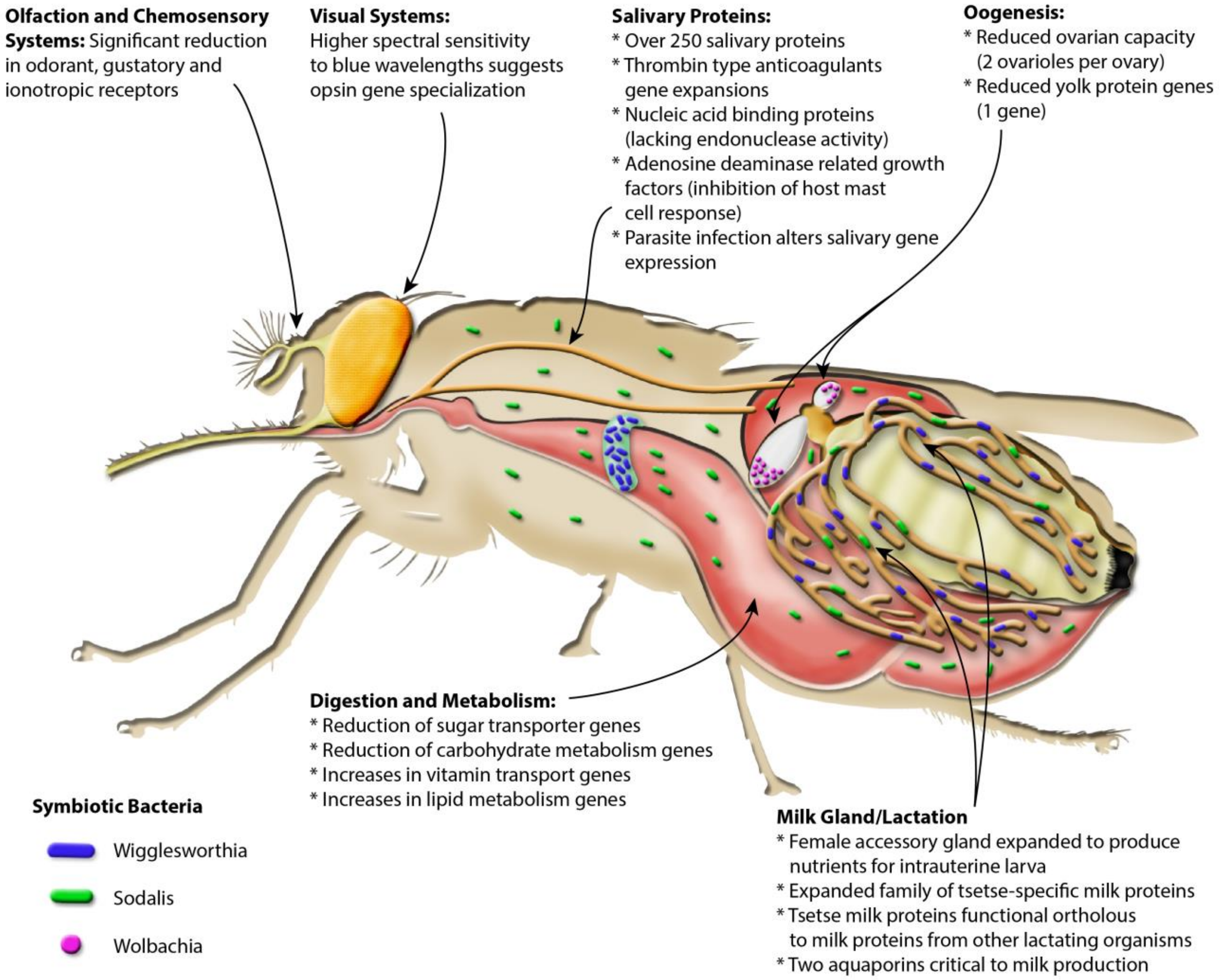
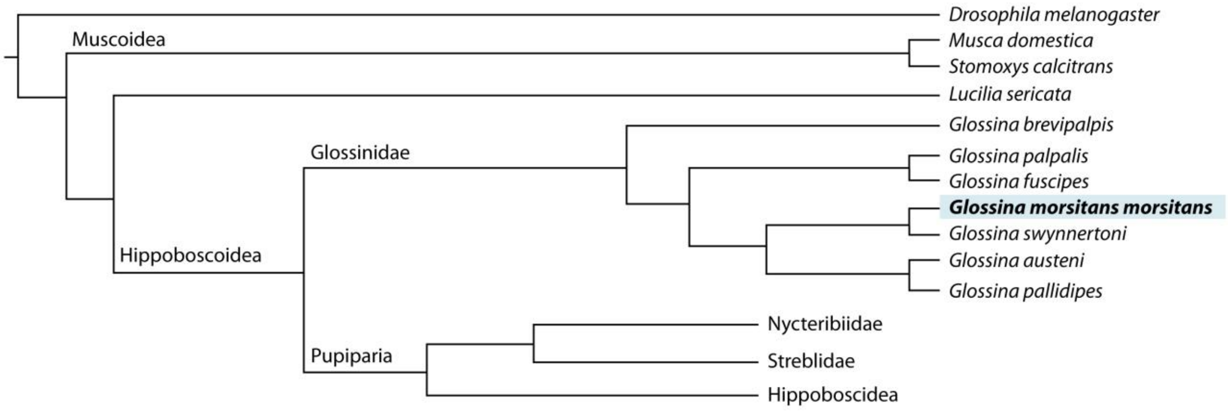

Olfaction and Chemosensory Systems: Significant reduction in odorant, gustatory and ionotropic receptors
Visual Systems: Higher spectral sensitivity to blue wavelengths suggests opsin gene specialization
Salivary Proteins:
* Over 250 salivary proteins
* Thrombin type anticoagulants gene expansions
* Nucleic acid binding proteins (lacking endonuclease activity)
* Adenosine deaminase related growth factors (inhibition of host mast cell response)
* Parasite infection alters salivary gene expression
Oogenesis:
* Reduced ovarian capacity (2 ovarioles per ovary)
* Reduced yolk protein genes (1 gene)
Digestion and Metabolism:
* Reduction of sugar transporter genes
* Reduction of carbohydrate metabolism genes
* Increases in vitamin transport genes
* Increases in lipid metabolism genes
Symbiotic Bacteria
Wigglesworthia
Sodalis
Wolbachia
Milk Gland/Lactation
* Female accessory gland expanded to produce nutrients for intrauterine larva
* Expanded family of tsetse-specific milk proteins
* Tsetse milk proteins functional ortholous to milk proteins from other lactating organisms
* Two aquaporins critical to milk production

Figure 6: Schematic overview of *Glossina* physiology and associated findings from the genome annotation



Phylogeny adapted from Petersen et.al. (2007) Molecular Phylogenetics and Evolution 45, 111-122

Supplemental Figure 1: Adapted phylogeny illustrating *Glossina morsitans morsitans* relationship within the Brachycera. The relative relationships between tsetse species and other selected members of the Brachycera. This tree was adapted from a Maximum parsimony tree based upon the combined sequence data from four genes: mitochondrial 16S ribosomal DNA (16s rDNA), nuclear 28S ribosomal DNA (28s rDNA), the carbamoylphosphate synthase (CPSase) domain of the nuclear CAD gene and the mitochondrial gene cytochrome oxidase I (COI). The full tree with additional species, bootstrap support values and posterior probabilities can be found in Petersen et.al. 2007 (*7*)

C**omparison of chemosensory gene homologs between species**				
Gene Family	***Glossina***	***Drosophila***	***Anopheles***	***Apis***
CSP	5	4	8	6
OBP	32	51	70	21
GR	14	68	76	10
OR	46	62	79	170
IR	17	61	70	10
SNMP	2	2	2	0
Total	**116**	**248**	**305**	**217**

Table 1: Comparison of chemoreceptor genes between *Glossina*, *Drosophila melanogaster*, *Anopheles gambiae* and *Apis mellifera*. CSPs: chemosensory proteins, GRs: Gustatory receptors, OBPs: Odorant Binding Proteins, ORs: Odorant Receptors, IRs: Ionotropic Receptors, SNMPs: Sensory Neuron Membrane Proteins.