1  **The *Glossina* Genome Cluster: Comparative Genomic Analysis of the Vectors of African**

2  **Trypanosomes**

3  **Authorship:**

4  Geoffrey M. Attardo, (gmattardo@ucdavis.edu) *[22]; Adly M.M. Abd-Alla, (a.m.m.abd-

5  alla@iaea.org)[13]; Alvaro Acosta-Serrano, (alvaro.acosta-serrano@lstmed.ac.uk)[16]; James E.

6  Allen, (jallen@ebi.ac.uk)[6]; Rosemary Bateta, (batetarw@gmail.com)[2]; Joshua B. Benoit,

7  (joshua.benoit@uc.edu)[24]; Kostas Bourtzis, (K.Bourtzis@iaea.org)[13]; Jelle Caers,

8  (jellecrs@gmail.com)[15]; Guy Caljon, (Guy.Caljon@uantwerpen.be)[21]; Mikkel B. Christensen,

9  (mikkel@ebi.ac.uk)[6]; David W. Farrow, (farrowdw@mail.uc.edu)[24]; Markus Friedrich,

10  (friedrichwsu@gmail.com)[33]; Aurélie Hua-Van, (aurelie.hua-van@egce.cnrs-gif.fr)[5]; Emily C.

11  Jennings, (jenninec@mail.uc.edu)[24]; Denis M. Larkin, (dmlarkin@gmail.com)[19]; Daniel Lawson,

12  (daniel.lawson@imperial.ac.uk)[10]; Michael J. Lehane, (mike.lehane@lstmed.ac.uk)[16]; Vasileios

13  P. Lenis, (vasilis.lenis@plymouth.ac.uk)[30]; Ernesto Lowy-Gallego, (ernesto@ebi.ac.uk)[6];

14  Rosaline W. Macharia, (rwanjiru@icipe.org, rslnmacharia1@gmail.com)[27,12]; Anna R. Malacrida,

15  (malacrid@unipv.it)[29]; Heather G. Marco, (heather.marco@uct.ac.za)[23]; Daniel Masiga,

16  (dmasiga@icipe.org)[12]; Gareth L. Maslen, (gmaslen@ebi.ac.uk)[6]; Irina Matetovici,

17  (imatetovovici@itg.be)[11]; Richard P. Meisel, (rpmeisel@uh.edu)[25]; Irene Meki

18  (irene_meki@yahoo.com)[13], Veronika Michalkova, (vmichalk@fiu.edu,

19  vmichalkova@yahoo.com)[7,20]; Wolfgang J. Miller, (wolfgang.miller@meduniwien.ac.at)[17]; Patrick

20  Minx, (pminx@genome.wustl.edu)[32]; Paul O. Mireji, (mireji.paul@gmail.com)[2,14]; Lino Ometto,

21  (lino.ometto@unipv.it)[8,29]; Andrew G. Parker, (a.g.parker@iaea.org)[13]; Rita Rio,

22  (Rita.Rio@mail.wvu.edu)[34]; Clair Rose, (clair.rose@lstmed.ac.uk)[16]; Andrew J. Rosendale,

23  (andrew.rosendale@msj.edu)[18,24]; Omar Rota-Stabelli, (omar.rota@fmach.it)[8]; Grazia Savini,

24  (grazia.savini01@universitadipavia.it)[29]; Liliane Schoofs, (liliane.schoofs@kuleuven.be)[15];

25  Francesca Scolari, (francesca.scolari@unipv.it)[29]; Martin T. Swain, (mts11@aber.ac.uk)[1]; Peter

26    Takáč, (peter.takac@savba.sk, Peter@scientica.sk)[31]; Chad Tomlinson,

27    (ctomlins@wustl.edu)[32]; George Tsiamis, (gtsiamis1@gmail.com)[28]; Jan Van Den Abbeele,

28    (jvdabbeele@itg.be)[11]; Aurelien Vigneron, (aurelien.vigneron@yale.edu)[35]; Jingwen Wang,

29    (jingwenwang@fudan.edu.cn)[9]; Wesley C. Warren, (warrenwc@missouri.edu)[32,36]; Robert M.

30    Waterhouse, (robert.waterhouse@unil.ch)[26]; Matthew T. Weirauch,

31    (matthew.weirauch@cchmc.org)[4]; Brian L. Weiss, (brian.weiss@yale.edu)[35]; Richard K Wilson,

32    (rwilson@genome.wustl.edu)[32]; Xin Zhao, (kitty.zhaoxin@foxmail.com)[3]; Serap Aksoy,

33    (serap.aksoy@yale.edu)*[35]

34    * - Corresponding Authors

35    **Author Affiliations:**

36    1: Institute of Biological, Environmental and Rural Sciences, Aberystwyth University,

37    Aberystwyth, Ceredigion, United Kingdom

38    2: Department of Biochemistry, Biotechnology Research Institute - Kenya Agricultural and

39    Livestock Research Organization, Kikuyu, Kenya

40    3: CAS Center for Influenza Research and Early-warning (CASCIRE), Chinese Academy of

41    Sciences, Beijing, China

42    4: Center for Autoimmune Genomics and Etiology and Divisions of Biomedical Informatics and

43    Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United

44    States

45    5: Laboratoire Evolution, Genomes, Comportement, Ecologie, CNRS, IRD, Univ. Paris-Sud,

46    Université Paris-Saclay, Gif-sur-Yvette, France

47    6: VectorBase, European Molecular Biology Laboratory, European Bioinformatics Institute

48    (EMBL-EBI), Cambridge, Cambridgeshire, United Kingdom

49   7: Department of Biological Sciences, Florida International University, Miami, Florida, United

50   States

51   8: Department of Sustainable Ecosystems and Bioresources, Research and Innovation Centre,

52   Fondazione Edmund Mach, San Michele all'Adige (TN), Italy

53   9: School of Life Sciences, Fudan University, Shanghai, China

54   10: Department of Life Sciences, Imperial College London, London, United Kingdom

55   11: Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

56   12: Molecular Biology and Bioinformatics Unit, International Center for Insect Physiology and

57   Ecology, Nairobi, Kenya

58   13: Insect Pest Control Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food &

59   Agriculture, Vienna, Vienna, Austria

60   14: Centre for Geographic Medicine Research, Coast, Kenya Medical Research Institute, Kilifi,

61   Kenya

62   15: Department of Biology - Functional Genomics and Proteomics Group, KU Leuven, Leuven,

63   Belgium

64   16: Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, Merseyside,

65   United Kingdom

66   17: Department of Cell and Developmental Biology, Medical University of Vienna, Vienna,

67   Austria

68   18: Department of Biology, Mount St. Joseph University, Cincinnati, Ohio, United States

69   19: Department of Comparative Biomedical Sciences, Royal Veterinary College, London, United

70   Kingdom

71   20: Institute of Zoology, Slovak Academy of Sciences, Bratislava, Slovakia

72   21: Laboratory of Microbiology, Parasitology and Hygiene, University of Antwerp, Antwerp,

73   Belgium

74   22: Department of Entomology and Nematology, University of California, Davis, Davis,

75   California, United States

76   23: Department of Biological Sciences, University of Cape Town, Rondebosch, South Africa

77   24: Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio, United States

78   25: Department of Biology and Biochemistry, University of Houston, Houston, Texas, United

79   States

80   26: Department of Ecology & Evolution, and Swiss Institute of Bioinformatics, University of

81   Lausanne, Lausanne, Switzerland

82   27: Centre for Biotechnology and Bioinformatics, University of Nairobi, Nairobi, Kenya

83   28: Department of Environmental and Natural Resources Management, University of Patras,

84   Agrinio, Etoloakarnania, Greece

85   29: Department of Biology and Biotechnology, University of Pavia, Pavia, Italy

86   30: Schools of Medicine and Dentistry, University of Plymouth, Plymouth, United Kingdom

87   31: Department of Animal Systematics, Ústav zoológie SAV; Scientica, Ltd., Bratislava, Slovakia

88   32: McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri,

89   United States

90   33: Department of Biological Sciences, Wayne State University, Detroit, Michigan, United States

91   34: Department of Biology, West Virginia University, Morgantown, West Virginia, United States

92   35: Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New

93   Haven, Connecticut, United States

4

94    36: Bond Life Sciences Center, University of Missouri, Columbia, Missouri, United States

95

96    **Abstract:**

97    **Background:**

98    Tsetse flies (*Glossina sp.*) are the sole vectors of human and animal trypanosomiasis

99    throughout sub-Saharan Africa. Tsetse are distinguished from other Diptera by unique

100    adaptations, including lactation and the birthing of live young (obligate viviparity), a vertebrate

101    blood specific diet by both sexes and obligate bacterial symbiosis. This work describes

102    comparative analysis of six *Glossina* genomes representing three sub-genera: *Morsitans* (*G.*

103    *morsitans morsitans (G.m. morsitans)*, *G. pallidipes*, *G. austeni*), *Palpalis* (*G. palpalis*, *G.*

104    *fuscipes*) and *Fusca* (G*. brevipalpis*) which represent different habitats, host preferences and

105    vectorial capacity.

106    **Results:**

107    Genomic analyses validate established evolutionary relationships and sub-genera. Syntenic

108    analysis of *Glossina* relative to *Drosophila melanogaster* shows reduced structural conservation

109    across the sex-linked X chromosome. Sex linked scaffolds show increased rates of female

110    specific gene expression and lower evolutionary rates relative to autosome associated genes.

111    Tsetse specific genes are enriched in protease, odorant binding and helicase activities.

112    Lactation associated genes are conserved across all *Glossina* species while male seminal

113    proteins are rapidly evolving. Olfactory and gustatory genes are reduced across the genus

114    relative to other characterized insects. Vision associated Rhodopsin genes show conservation

115    of motion detection/tracking functions and significant variance in the Rhodopsin detecting colors

116    in the blue wavelength ranges.

117    **Conclusions:**

118    Expanded genomic discoveries reveal the genetics underlying *Glossina* biology and provide a

119    rich body of knowledge for basic science and disease control. They also provide insight into the

120    evolutionary biology underlying novel adaptations and are relevant to applied aspects of vector

121    control such as trap design and discovery of novel pest and disease control strategies.

122    **Keywords:**

123    Tsetse, trypanosomiasis, hematophagy, lactation, disease, neglected, symbiosis

124

## Background:

Flies in the genus *Glossina* (tsetse flies) are vectors of African trypanosomes, which are of great medical and economic importance in Africa. Sleeping sickness (Human African Trypanosomiasis or HAT) is caused by two distinct subspecies of the African trypanosomes transmitted by tsetse. In East and Southern Africa, *Trypanosoma brucei rhodesiense* causes the acute *Rhodesiense* form of the disease, while in Central and West Africa *T. b. gambiense* causes the chronic *Gambiense* form of the disease, which comprises about 95% of all reported HAT cases. Devastating epidemics in the 20<sup>th</sup> century resulted in hundreds of thousands of deaths in sub-Saharan Africa [1], but more effective diagnostics now indicate that data concerning sleeping sickness deaths are subject to gross errors due to under-reporting [2]. With hindsight, it is thus reasonable to infer that millions died from sleeping sickness during the colonial period. Loss of interest and funding for control programs within the endemic countries resulted in a steep rise in incidence after the post-independence period of the 1960s. In an ambitious campaign to control the transmission of Trypanosomiasis in Africa, multiple groups came together in a public/private partnership. These include the WHO, multiple non-governmental organizations, Sanofi Aventis and Bayer. The public sector groups developed and implemented multi-country control strategies and the companies donated the drugs required for treatment of the disease. The campaign reduced the global incidence of *Gambiense* HAT to <3,000 cases in 2015 [3]. Based on the success of the control campaign there are now plans to eliminate *Gambiense* HAT as a public health problem by 2030 [4]. In contrast, control of *Rhodesiense* HAT has been more complex as disease transmission involves domestic animals, which serve as reservoirs for the parasite. Hence elimination of the *Rhodesiense* disease will require treatment or elimination of domestic reservoirs, and/or reduction of tsetse vector populations. These strategies play a key part while medical interventions are used largely for humanitarian purposes. In addition to the public health impact of HAT, Animal African

7

150   Trypanosomiasis (AAT or Nagana) limits the availability of meat and milk products in large

151   regions of Africa. It also excludes effective cattle rearing from ten million square kilometers of

152   Africa [5] with wide implications for land use, i.e. constraints on mixed agriculture and lack of

153   animal labor for ploughing [6]. Economic losses in cattle production are estimated at 1-1.2 billion

154   dollars US and total agricultural losses caused by AAT are estimated at 4.75 billion dollars US

155   per year [7, 8].

156   Achieving disease control in the mammalian host has been difficult given the lack of vaccines.

157   This is due to the process of antigenic variation the parasite displays in its host. Hence,

158   accurate diagnosis of the parasite and staging of the disease are important. This is of particular

159   importance due to the high toxicity of current drugs available for treatment of late-stage disease

160   although introduction of a simpler and shorter nifurtimox and eflornithine combination therapy

161   (NECT)[9] and discovery of new oral drugs, such as fexinidazole [10] and acoziborole, are

162   exciting developments. Although powerful molecular diagnostics have been developed in

163   research settings, few have yet to reach the patients or national control programs [11]. Further

164   complicating control efforts, trypanosomes are showing resistance to available drugs for

165   treatment [12, 13]. While vector control is essential for zoonotic *Rhodesiense* HAT, it has not

166   played a major role in *Gambiense* HAT as it was considered too expensive and difficult to

167   deploy in the resource poor settings of HAT foci. However, modelling, historical investigations

168   and practical interventions demonstrate the significant role that vector control can play in the

169   control of *Gambiense* HAT [14-16], especially given the possibility of long-term carriage of

170   trypanosomes in both human and animal reservoirs [17, 18]. The African Union has made

171   removal of trypanosomiasis via tsetse fly control a key priority for the continent [19].

172   Within the *Glossinidae*, 33 extant taxa are described from 22 species in 4 subgenera. The first

173   three sub-genera *Austenina Townsend*, *Nemorhina Robineau-Desvoidy* and *Glossina*

174   *Wiedemann* correspond to the *Fusca*, *Palpalis*, and *Morsitans* species groups, respectively [20].

175    The fourth subgenus *Machadomia* was established in 1987 to incorporate *G. austeni*. The

176    relationship of *G. austeni* Newstead with respect to the *Palpalis* and *Morsitans* complex flies

177    remains controversial [21]. While molecular taxonomy shows that *Palpalis-* and *Morsitans-*

178    species groups are monophyletic, the *Fusca* species group emerges as sister group to all

179    remaining Glossinidae. *Morsitans* group taxa are adapted to drier habitats relative to the other

180    two subgenera [22]. *Palpalis* group flies tend to occur in riverine and lacustrine habitats. *Fusca*

181    group flies largely inhabit moist forests of West Africa. The host-specificity of the different

182    species groups vary, with the *Palpalis* group flies displaying strong anthrophilicity while the

183    others are more zoophilic in preference. The principal vectors of HAT include *G. palpalis* s.l., *G.*

184    *fuscipes* and *G. m. morsitans* s.l. The riverine habitats of *Palpalis* group flies and their

185    adaptability to peridomestic environments along with human blood meal preferences make them

186    excellent vectors for HAT. Other species belonging to the *Morsitans* group (such as *G.*

187    *pallidipes*) can also transmit human disease, but principally play an important role in AAT

188    transmission. In particular, *G. pallidipes* has a wide distribution and a devastating effect in East

189    Africa. Also, of interest is *G. brevipalpis*, an ancestral tsetse species within the Fusca species

190    complex. This species exhibits poor vectorial capacity with *T. brucei* relative to *G. m. morsitans*

191    in laboratory infection expriments using colonized fly lines [23]. Comparison of the susceptibility

192    of *G. brevipalpis* to *Trypanosoma congolense* (a species that acts as a major causative agent of

193    AAT) also showed it has a much lower rate of infection relative to *Glossina austeni* [24].

194    To expand the genetic/genomic knowledge and develop new and/or improved vector control

195    tools a consortium, the International Glossina Genome Initative (IGGI), was established in early

196    2004 to sequence the *G. m. morsitans* genome [25]. In 2014 the first tsetse fly genome from the

197    *Glossina m. morsitans* species was produced [26]. This project facilitated the use of modern

198    techniques such as transcriptomics and enabled functional investigations at the genomic level

199    into tsetse's viviparous reproductive physiology, obligate symbiosis, trypanosome transmission

200    biology, olfactory physiology and the role of saliva in parasite transmission. The species

201    *Glossina m. morsitans* was chosen for the first genome discovery effort as it is relatively easy to

202    maintain under laboratory conditions and many physiological studies had been based on this

203    species. To study the genetics underlying tsetse species-specific traits, such as host preference

204    and vector competence, we have now assembled five additional representative genomes from

205    the species complexes of *Glossina*: Morsitans (*G. m. morsitans*, G. *pallidipes,*),

206    Morsitans/Machadomia (*G. austeni*), Palpalis *(G. palpalis, G. fuscipes)* and Fusca (*G.*

207    *brevipalpis).* These species represent flies with differences in geographical localization,

208    ecological preferences, host specificity and vectorial capacity (Summarized in Figure 1). Here

209    we report on the evolution and genetics underlying this genus by comparison of their genomic

210    architecture and predicted protein-coding sequences and highlight some of the genetic

211    differences that hold clues to the differing biology between these species.

## Results and Discussion:

**Multiple genetic comparisons confirm *Glossina* phylogenetic relationships and the inclusion of *G. austeni* as a member of the *Morsitans* sub-Genus**

215    Sequence similarity between the genomes was analyzed using whole genome nucleotide

216    alignments of supercontigs and predicted coding sequences from the five new *Glossina*

217    genomes as well as those from the *Musca domestica* genome using *G. m. morsitans* as a

218    reference (Figure 2A). The results indicate that *G. pallidipes* and *G. austeni* are most similar at

219    the sequence level to *G. m. morsitans*. This is followed by the species in the *Palpalis* sub-genus

220    (*G. fuscipes* and *G. palpalis*). The remaining species (*G. brevipalpis*) shows the least sequence

221    conservation relative to *G. m. morsitans* followed by the outgroup species *M. domestica*. The

222    lower sequence similarity between *G. brevipalpis* and the other tsetse species reinforces its

223    status as a distant relative to the *Morsitans* and *Palpalis* sub-genera.

10

224    Alignment of the predicted coding sequences produced a similar result to that observed in the

225    whole genome alignment in terms of similarity to *G. m. morsitans* (Figure 2A). Of interest is that

226    more than 25% of the *G. m. morsitans* exon sequences were not align-able with *G. brevipalpis*,

227    indicating that they were either lost, have diverged beyond alignability or were in an

228    unsequenced region in *G. brevipalpis*. In addition, *G. brevipalpis* has on average ~5000 fewer

229    predicted protein-coding genes than the other species. Given the low GC content of the *G.*

230    *brevipalpis* sequenced genome it is possible that some of the regions containing these

231    sequences lie within heterochromatin. The difficulties associated with sequencing

232    heterochromatic regions may have excluded these regions from our analysis; however, it also

233    implies that if these protein coding genes are indeed present they are located in a region of the

234    genome with low transcriptional activity.

235    We inferred the phylogeny and divergence times of *Glossina* using a concatenated alignment of

236    286 single-copy gene orthologs (478,000 nucleotide positions) universal to *Glossina* (Figure

237    2B). The tree recovered from this analysis has support from both Maximum Likelihood and

238    Bayesian analyses, using respectively homogeneous and heterogeneous models of

239    replacement. A coalescent-aware analysis further returned full support, indicating a speciation

240    process characterized by clear lineage sorting (Supplemental Figure 1). These results suggest

241    an allopatric speciation process characterized by a small founder population size followed by

242    little to no introgression among newly formed species.

243    Furthermore, we assembled complete mitochondrial (mtDNA) genome sequences for each

244    species as well as *Glossina morsitans centralis* as references for use in distinguishing samples

245    at the species, sub-species or haplotype levels. All the mtDNA genomes encode large (16S

246    rRNA) and small (12S rRNA) rRNAs, 22 tRNAs and 13 protein-coding genes. Phylogenetic

247    analysis of the resulting sequences using the Maximum Likelihood method resulted in a tree

248    with congruent topology to that produced by analysis of the concatenated nuclear gene

11

249  alignment (Figure 2C). A comparative analysis of the mtDNA sequences identified variable

250  marker regions with which to identify different tsetse species via traditional sequencing and/or

251  high-resolution melt analysis (HRM) (Supplemental Figure 2). Analysis of the amplicons from

252  this region using HRM facilitated the discrimination of these products based on their

253  composition, length and GC content. Use of HRM on these variable regions successfully

254  resolved differences between test samples consisting of different tsetse species as well as

255  individuals with different haplotypes or from different populations (Supplemental Figure 3). This

256  method provides a rapid, cost effective and relatively low-tech way of identifying differences in

257  field caught tsetse for the purposes of population genetics and measurement of population

258  diversity.

259  The trees derived from the nuclear and mitochondrial phylogenetic analyses agree with

260  previously published phylogenies for tsetse [27-29] and the species delineate into groups

261  representing the defined *Fusca*, *Palpalis* and *Morsitans* sub-genera.

262  A contentious issue within the taxonomy of *Glossina* is the placement of *G. austeni* within the

263  *Machadomia* sub-Genus. Comparative anatomical analysis of male genitalia places *G. austeni*

264  within the *morsitans* sub-Genus. However, female *G. austeni* genitalia bear anatomical

265  similarities to members of the *Fusca* sub-genus. In addition, *G. austeni's* habitat preferences

266  and some external morphology resemble those of the *palpalis* sub-Genus [28]. Recent

267  molecular evidence suggests that *G. austeni* are closer to the *morsitans* sub-genus [27, 29].

268  The data generated via the three discrete analyses described above all support the hypothesis

269  that *G. austeni* is a member of the *Morsitans* sub-genus rather than the *Palpalis* sub-genus and

270  belongs as a member of the *Morsitans* group rather than its own discrete sub-Genus.

271  **Comparative analysis of *Glossina* with *Drosophila* reveals reduced synteny and female**

272  **specific gene expression on X-linked scaffolds**

12

273    The scaffolds in each *Glossina spp.* genome assembly were assigned to chromosomal arms

274    based on orthology and relative position to protein-coding sequences in the *D. melanogaster*

275    genome (*Drosophila*) [30]. The *Glossina* and *Drosophila* genomes contain six chromosome

276    arms (Muller elements A-F) [31-33]. We assigned between 31-52% of annotated genes in each

277    species to a Muller element, which we used to assign >96% of scaffolds to Muller elements in

278    each species (Figure 3 and Supplemental table 1). From these results, we inferred the relative

279    size of each Muller element in each species by counting the number of annotated genes

280    assigned to each element and calculating the cumulative length of all assembled scaffolds

281    assigned to each element. Using either measure, we find that element E is the largest and

282    element F is the shortest in all species, consistent with observations in *Drosophila* [34].

283    Mapping of the *Glossina* scaffolds to the *Drosophila* Muller elements reveals differing levels of

284    conservation of synteny (homologous genomic regions with maintained orders and orientations)

285    across these six-species relative to *Drosophila*. In *G. m. morsitans*, the X chromosome is

286    composed of Muller elements A, D, and F as opposed to the *Drosophila* X which only contains A

287    and sometimes D [33], and all other *Glossina* species besides *G. brevipalpis* have the same

288    karyotype [35].  We therefore assume that the same elements are X-linked in the other *Glossina*

289    species (apart from *G. brevipalpis*). This analysis reveals that scaffolds mapping to *Drosophila*

290    Muller element A show a reduced overall level of syntenic conservation relative to the other

291    Muller elements while the scaffolds mapping to *Drosophila* Muller element D (part of the

292    *Glossina* X chromosome, but not the *D. melanogaster* X) retain more regions of synteny

293    conservation. We hypothesize that the lower syntenic conservation on element A reflects a

294    higher rate of rearrangement because it has been X-linked for more time (both in the *Drosophila*

295    and *Glossina* lineages) than element D (only in *Glossina*) and rearrangement rates are higher

296    on the X chromosome (element A) in *Drosophila* [34]

13

297    To examine the relationship between gene expression and DNA sequence evolution, we

298    compared gene expression levels between the X chromosome and autosomes using sex

299    specific RNA-seq libraries derived from whole males, whole non-lactating females and whole

300    lactating females for all the *Glossina* species apart from *G. pallidipes.* Consistent with previous

301    results from *G. m. morsitans* [33], the ratio of female:male expression is greater on the X

302    chromosome than autosomes across species (Supplemental Figure 4). In addition, there is a

303    deficiency of genes with male-biased expression (up-regulated in males relative to females) on

304    the X-linked elements in all species (Supplemental Figure 5). Reduced levels of male-biased

305    gene expression have also been observed in mosquitoes and is a conserved feature of the

306    *Anopheles* genus [36]. The X chromosome is hemizygous in males, which exposes recessive

307    mutations to natural selection and can accelerate the rate of adaptive substitutions and facilitate

308    the purging of deleterious mutations on the X chromosome [37, 38]. Using dN/dS values for

309    annotated genes, we fail to find any evidence for this faster-X effect across the entire phylogeny

310    or along any individual lineages (Supplemental Figure 6). The faster-X effect is expected to be

311    greatest for genes with male-biased expression because they are under selection in males [37],

312    but we find no evidence for faster-X evolution of male-biased genes in any of the *Glossina*

313    species. In contrast, there is some evidence for "slower-X" evolution amongst female-biased

314    genes (Supplemental Figure 7), suggesting that purifying selection is more effective at purging

315    deleterious mutations on the X chromosome [39]. Genes with female-biased expression tend to

316    be broadly expressed [40], suggesting that pleiotropic constraints on female-biased genes

317    increase the magnitude of purifying selection and produce the observed slower-X effect [41].

318    The exception to these observations is element F. Element F, the smallest X-linked element,

319    has low female expression and an excess of genes with male-biased expression (Supplemental

320    Figure 8). In contrast with the other X-linked Muller elements in *Glossina*, the dN/dS ratios of all

321    Element F associated genes (male biased and unbiased) suggest that they are evolving faster

322 than the rest of the genome across all tsetse lineages (Supplemental Figure 9). The F elements

323 in *Drosophila* species, while not being X-linked, show similar properties in that they have lower

324 levels of synteny, increased rates of inversion, and higher rates of protein coding sequence

325 evolution, suggesting that the F element is rapidly evolving in flies within Schizophora [42].

326 **The *G. austeni* genome contains *Wolbachia* derived chromosomal insertions (Figure 4)**

327 A notable feature of the *G. m. morsitans* genome was the integration of large segments of the

328 *Wolbachia* symbiont genome via horizontal gene transfer (HGT). Characterization of the *G. m.*

329 *morsitans* HGT events revealed that the chromosomal sequences with transferred material

330 contained a high degree of nucleotide polymorphisms, coupled with insertions, and deletions

331 [43]. These observations were used in this analysis to distinguish cytoplasmic from

332 chromosomal *Wolbachia* sequences during the *in-silico* characterization of the tsetse genomes.

333 Analysis of the six assemblies revealed that all contain *Wolbachia* sequences although *G.*

334 *pallidipes, G. fuscipes*, *G. palpalis* and *G. brevipalpis* had very limited DNA sequence that

335 displayed homology with *Wolbachia*. Furthermore, analysis of fly lines from which the

336 sequenced DNA was obtained with *Wolbachia* specific primers PCR-amplification was negative.

337 This is in line with PCR-based screening of *Wolbachia* infections in natural populations, further

338 indicating that these short segments could be artifacts or contaminants [44]. However, *G.*

339 *austeni* contains more extensive chromosomal integrations of *Wolbachia* DNA (Supplemental

340 Table 2).

341 All *Wolbachia* sequences, chromosomal and cytoplasmic, identified in *G. austeni* were mapped

342 against the reference genomes of *Wolbachia* strains *wMel*, *wGmm*, and the chromosomal

343 insertions A & B in *G. m. morsitans* (Figure 4). The *G. austeni* chromosomal insertions, range in

344 size from 500 - 95,673 bps with at least 812 DNA fragments identified *in silico*. Sequence

345 homology between *w*Mel, *w*Gmm, and the chromosomal insertions A and B in *G. m. morsitans*

346 *morsitans* varied between 98 – 63%, with the highest sequence homologies observed with

15

347    chromosomal insertions A and B from *G. m. morsitans*. The similarity between the genomic

348    insertions in *G. m. morsitans* and *G. austeni* relative to cytoplasmic *Wolbachia* sequences

349    suggests they could be derived from an event in a common ancestor. The absence of

350    comparable insertions in *G. pallidipes* (a closer relative to *G. m. morsitans*) indicate that either

351    these insertions occurred independently or that the region containing the insertions was not

352    assembled in *G. pallidipes*. Additional data from field based *Glossina* species/sub-species is

353    required to determine the true origin of these events.

354    **Analysis of *Glossina* genus and sub-genus specific gene families reveals functional**

355    **enrichments.**

356    All annotated *Glossina* genes were assigned to groups (orthology groups - OGs) containing

357    predicted orthologs from other insect and arthropod species represented within Vectorbase. A

358    global analysis of all the groups containing *Glossina* genes was utilized to determine the gene

359    composition of these flies relative to their Dipteran relatives and between the Glossina sub-

360    genera. An array of twelve Diptera are represented within this analysis including *Anopheles*

361    *gambiae* (Nematocera), *Aedes aegypti* (Nematocera), *Lutzomyia longipalpis* (Nematocera),

362    *Drosophila melanogaster* (Brachycera), *Stomoxys calcitrans* (Brachycera) and *Musca*

363    *domestica* (Brachycera).

364    The tsetse associated OGs are represented by groups ranging from universal to the Diptera

365    included in the analysis to species specific to the individual tsetse species. The composition of

366    these OGs breaks down to a core of 3,058 OGs with constituents universal to Diptera (93430

367    genes), 299 OGs specific and universal to Brachyceran flies (4975 genes) and 162 OGs specific

368    and universal to *Glossina* (1548 genes). A dramatic feature identified by this analysis is the

369    presence of 2,223 OGs specific and universal to the *Palpalis* sub-genus (*G. fuscipes* and *G.*

370    *palpalis* 4948 genes). This contrasts with the members of the *Morsitans* sub-genus (*G. m.*

16

371    *morsitans*, *G. pallidipes* and *G. austeni*) in which there are 137 specific and universal OGs (153

372    genes) (Figure 5, Supplemental table 3, Supplemental data 2+3).

373    To understand the functional significance of the *Glossina* specific OGs, we performed an

374    analysis of functional enrichment of gene ontology (GO) terms within these groups. Many of the

375    *Glossina* specific genes are not currently associated with GO annotations as they lack

376    characterized homologs in other species. As such these sequences were not included in this

377    analysis. However, ~60% of the genes within the combined *Glossina* gene repertoire are

378    associated with GO annotations, which allowed for analysis of a sizable proportion of the

379    dataset.

380    **_Glossina_ genus universal and specific genes are enriched in genes coding for proteases**

381    **and odorant binding proteins.**

382    The orthology groups containing genes specific and universal to the *Glossina* genus are

383    enriched in odorant binding and serine-type endopeptidase activities. The universality of these

384    genes within *Glossina* and their absence from the other surveyed Dipteran species suggests

385    they are currently associated with tsetse specific adaptations.

386    The ontology category with the lowest p-value represents proteolysis associated genes. This

387    category encompasses 92 *Glossina* specific proteases with predicted serine-type

388    endopeptidase activity. The abundance of this category may be an adaptation to the protein-rich

389    blood specific diet of both male and female flies. A similar expansion of serine proteases is

390    associated with blood feeding in mosquitoes and the presence of an equivalent expansion in

391    tsetse may represent an example of convergent evolution [45]. This class of peptidases is also

392    associated with critical functions in immunity, development and reproduction in Diptera [46-49].

393    The other enriched GO term common to all *Glossina* is for genes encoding odorant binding

394    proteins (OBPs). Of the 370 OBPs annotated within *Glossina*, 55 lack orthologs in species

17

395    outside of *Glossina*. The primary function of OBPs is to bind small hydrophobic molecules to

396    assist in their mobilization in an aqueous environment. These proteins are primarily associated

397    with olfaction functions as many are specifically expressed in chemosensory associated

398    tissues/organs where they bind small hydrophobic molecules and transport them to odorant

399    receptors [50, 51]. However, functional analyses in *G. m. morsitans* have associated an OBP

400    (OBP6) with developmental activation of hematopoiesis during larvigenesis in response to the

401    mutualistic *Wigglesworthia* symbiont [52]. In addition, many of the OBPs identified in this

402    analysis are characterized as *Glossina* specific seminal proteins with male accessory gland

403    specific expression patterns. They are primary constituents of the spermatophore structure

404    produced by the male tsetse during mating [53]. The genus specific nature of these OBPs

405    suggests that they are key components of reproductive adaptations of male tsetse.

406    **The *Palpalis* sub-genus contains a large group of sub-genus specific genes.**

407    A large group of genes specific and universal to members of the Palpalis sub-genus (*G. palpalis*

408    and *G. fuscipes*) was a defining feature of the orthology analysis. The expansion includes 2223

409    OGs and encompasses 4948 genes between *G. palpalis* and *G. fuscipes*. Homology based

410    analysis of these genes by comparison against the NCBI NR database revealed significant (e-

411    value < $1x10^{-10}$) results for 603 of the genes. Within this subset of genes, ~ 5% represent

412    bacterial contamination from tsetse's obligate endosymbiont *Wigglesworthia*. Sequences

413    homologous to another well-known bacterial symbiont *Spiroplasma* were found exclusively in *G.*

414    *fuscipes*. This agrees with previous observations of *Spiroplasma* infection of colonized and field

415    collected *G. fuscipes* flies [54].

416    Four genes bear homology to viral sequences (*GPPI051037/GFUI045295* and

417    *GPPI016422/GFUI028200*). These sequences are homologus to genes from Ichnoviruses.

418    These symbiotic viruses are transmitted by parasitic Ichneumonid wasps with their eggs to

419   suppress the immune system of host insects [55]. These genes may have originated from a

420   horizontal transfer event during an attempted parasitization.

421   Another feature of note is the abundance of putative proteins with predicted helicase activity. Of

422   the 603 genes with significant hits, 64 (10.5%) are homologous to characterized helicases.

423   Functional enrichment analysis confirms the enrichment of helicase activity in this gene set.

424   These proteins are associated with the production of small RNA's (miRNAs, siRNAs and

425   piRNAs) which mediate posttranscriptional gene expression and the defensive response against

426   viruses and transposable elements. Of the 64 genes, 41 were homologous to the armitage

427   (*armi*) helicase. Recent work in *Drosophila* shows that *armi* is a reproductive tissue specific

428   protein and is responsible for binding and targeting mRNAs for processing into piRNAs by the

429   PIWI complex [56]. The reason for the accumulation of this class of genes within the *Palpalis*

430   sub-genus is unknown. However, given the association of these proteins with small RNA

431   production they could be associated with a defensive response against viral challenges or

432   overactive transposable elements. A similar phenomena is seen in *Aedes aegypti* where

433   components of the PIWI pathway have been amplified and function outside of the reproductive

434   tissues to generate piRNAs against viral genes [57].

435   **Analysis of gene family variations reveals sub-genus specific expansions and**

436   **contractions of genes involved in sperm production and chemosensation**

437   In addition to unique gene families, we identified orthology groups showing significant variation

438   in gene numbers between *Glossina* species. Of interest are groups showing significant sub-

439   genus specific expansions or contractions, which may represent lineage specific adaptations.

440   General trends that we observed in these groups show the largest number of gene family

441   expansions within the *Palpalis* sub-genus and the largest number of gene family contractions

442   within *G. brevipalpis* (a member of the *Fusca* sub-genus) (Figure 6 - For completely annotated

443   figures with descriptions and group IDs see Supplemental Figures 11 and 12, respectively).

19

444    *Palpalis sub-genus specific expansion of sperm associated genes* (Supplemental data 4)

445    Members of the *Palpalis* sub-genus had a total of 29 gene family expansions and 1 contraction

446    relative to the other 4 tsetse species. Of the three sub-Genera, this represents the largest

447    number of expansions and parallels with the large number of *Palpalis* specific orthology groups.

448    Two gene families expanded within the *Palpalis* group (VBGT00770000031191 and

449    VBGT00190000014373) encode WD repeat containing proteins. The *Drosophila* orthologs

450    contained within these families (*cg13930*, *dic61B*, *cg9313*, *cg34124*) are testes specific and

451    associated with cilia/flagellar biosynthesis and sperm production [58]. Alteration/diversification of

452    sperm associated proteins could explain the split of the *Palpalis* sub-genus from the other

453    *Glossina* and the potential incipient speciation documented between *G. palpalis* and *G. fuscipes*

454    [59].

455    *The Morsitans sub-genera shows reductions in chemosensory protein genes*

456    Within the *Morsitans* sub-genus six gene families are expanded and two are contracted relative

457    to the other tsetse species. Of interest, one of the contracted gene families encodes

458    chemosensory proteins (VBGT00190000010664) orthologous to the CheB and CheA series of

459    proteins in *D. melanogaster*. The genes encoding these proteins are expressed exclusively in

460    the gustatory sensilla of the forelegs of male flies and are associated with the detection of low

461    volatility pheromones secreted by the female in higher flies [60]. Of interest is that the number of

462    genes in *G. palpalis* (14), *G. fuscipes* (15) and G. *brevipalpis* (14) are expanded within this

463    family relative to *D. melanogaster* (12), *M. domestica* (10) and *S. calcitrans* (4). However, the

464    *Morsitans* group flies *G. m. morsitans* (7), *G. pallidipes* (7) and *G. austeni* (5) all appear to have

465    lost some members of this family. The functional significance of these changes is unknown.

466    However, it could represent an optimization of the male chemosensory repertoire within the

467    *Morsitans* sub-genus.

468   In terms of expanded gene families in *Morsitans*, we find two encoding enzymes associated with

469   the terpenoid backbone biosynthesis pathway (VBGT00190000010926 -farnesyl pyrophosphate

470   synthase and VBGT00840000047886 – farnesol dehydrogenase). This pathway is essential for

471   the generation of precursors required for the synthesis of the insect hormone Juvenile Hormone

472   (JH). In adult *G. m. morsitans*, JH levels play an important role in regulating nutrient balance

473   before and during pregnancy. High JH titers activate lipid biosynthesis and accumulation in the

474   fat body prior to lactation. During lactation, JH titers fall, resulting in the catabolism and

475   mobilization of stored lipids for use in milk production [61].

476   **Comparative analysis of the immune associated genes in *Glossina* species reveals**

477   **specific expansions, contractions and losses relative to *Musca domestica* and**

478   ***Drosophila melanogaster***

479   Tsetse flies are exposed to bacterial, viral, protozoan and fungal microorganisms exhibiting a

480   broad spectrum of beneficial, commensal, parasitic and pathogenic phenotypes within their

481   host. Yet, the diversity and intensity of the microbial challenge facing tsetse flies is limited

482   relative to that of related Brachyceran flies such as *D. melanogaster* and *M. domestica* in terms

483   of level of exposure, microbial diversity and host microbe relationships. While tsetse larvae live

484   in a protected environment (maternal uterus) feeding on maternally produced lactation

485   secretions, larval *D. melanogaster* and *M. domestica* spend their entire immature development

486   in rotting organic materials surrounded by and feeding on a diverse array of microbes. The adult

487   stages also differ in that tsetse feed exclusively on blood which exposes them to a distinct yet

488   limited array of microbial fauna. The immune function and genetic complement of *D.*

489   *melanogaster* is well characterized and provides the opportunity to compare the constitution of

490   orthologous immune gene sequences between *M. domestica* and the *Glossina* species [62].

491   Orthology groups containing *Drosophila* genes associated with the 'Immune System Process'

21

492    GO tag (GO:0002376) were selected and analyzed to measure the presence/absence or

493    variance in number of orthologous sequences in *Glossina* (Figure 7 + Supplemental Table 5).

494    Several orthologs within this ontology group are highly conserved across all species and are

495    confirmed participants with the fly's antimicrobial immune response. These genes include the

496    peptidoglycan recognition proteins (PGRPs) (with the exception of the PGRP SC1+2 genes)

497    [63], prophenoloxidase 1, 2 and  3 [52], the reactive oxygen intermediates *dual oxidase* and

498    *peroxiredoxin 5* [64, 65], and antiviral (RNAi pathway associated) *dicer 2* and *argonaute 2*. The

499    antimicrobial peptide encoding genes *attacin* (variants A and B) and *cecropin* (variants A1, A2,

500    B and C) are found within *Glossina*, but have diverged significantly (the highest % identity based

501    on blastx comparison = 84%) from closely related fly taxa [66-68].

502    *Glossina species are missing immune gene families present in D. melanogaster and M.*

503    *domestica*

504    Several gene families are missing within the *Glossina* species although expanded within *M.*

505    *domestica* (Figure 7). These include *lysozyme E, defensin, elevated during infection*, and the

506    *PGRP-SC1+2* gene families. These may be adaptations to the microbe rich diet and

507    environment in which *M. domestica* larvae and adults exist. The expansion of immune gene

508    families in *M. domestica* relative to *D. melanogaster* was previously documented in the

509    publication of the *M. domestica* genome [69]. However, the added context of the *Glossina*

510    immune gene complement highlights the significance of the expansion of these families relative

511    to their loss in all *Glossina* species. The loss of these families may represent the reduced

512    dietary and environmental exposure to microbial challenge associated with the dramatic

513    differences in life history between these flies.

514    *Glossina species show immune gene family expansions associated with the Toll and IMD*

515    *pathways*

22

516    In contrast, we observed several *Glossina* immune related gene families which are expanded

517    relative to orthologous families in *Drosophila* and *M. domestica* (Figure 7). Duplications of this

518    nature often reflect evolutionarily important aspects of an organism's biology, and in the case of

519    tsetse, may have resulted from the fly's unique association with parasitic African trypanosomes.

520    Prominent among the expanded immune related *Glossina* genes are those that encode *Attacin*

521    *A* and *Attacin B*, which are IMD pathway produced effector antimicrobial molecules, and *Cactus*,

522    a component of the Toll signaling pathway. Similarly, the most highly expanded immune related

523    gene across *Glossina* species are the orthologs of *Drosophila* CG4325. RNAi-based studies in

524    *Drosophila* indicate that CG4325 is a regulator of both the Toll and IMD signaling pathways [70].

525    Significant expansion of this gene family in *Glossina* substantiates previously acquired data that

526    demonstrated the functional importance of the Toll and IMD pathways in tsetse's response to

527    trypanosome challenge [71, 72]. Finally, all six *Glossina* genomes encode multiple copies of

528    *moira*. This gene, which is involved in cell proliferation processes [73], is differentially expressed

529    upon trypanosome infection when compared to uninfected *G. m. morsitans* [74]. In an effort to

530    eliminate parasite infections, tsetse flies produce reactive oxygen intermediates that cause

531    collateral cytotoxic damage [64]. Additionally, trypanosome infection of tsetse's salivary glands

532    induces expression of fly genes that encode proteins associated with stress and cell division

533    processes, further indicating that parasite infection results in extensive damage to host cells.

534    Expansion of *moira* gene copy number in *Glossina's* genome may reflect the fly's need to

535    maintain epithelial homeostasis in the face of damage caused by trypanosome infections.

536    *G. brevipalpis has a species-specific expansion of immune associated proteins*

537    An interesting highlight from this analysis is the identification of a gene expansion associated

538    with alpha-mannosidase activity (VBGT00190000009892). An orthologous *Drosophila* gene (*α-*

539    *Man-Ia*) is an essential component in the encapsulation response by hemocytes to attack by

540    parasitoid wasps. This enzyme modifies lamellocyte surface glycoproteins to facilitate the

23

541  recognition and encapsulation of foreign bodies. As described in the *G. m. morsitans* genome

542  paper and here, there is evidence of parasitization by parasitoid wasps in the genomes of these

543  flies in the form of integrated gene sequences homologous to polynavirus genes [26]. The

544  expansion of these proteins could be an evolutionary response to pressure induced by

545  parasitization although the current status of tsetse associated parasitoids is unknown.

546  **Tsetse reproductive genetics**

547  *Milk protein genes are universal and tightly conserved in Glossina (Figure 8 + Supplemental*

548  *table 6)*

549  The intrauterine development and nourishment of individual larval offspring is a defining

550  characteristic of the *Hippoboscoidea* superfamily, which includes the *Glossinidae* (Tsetse flies),

551  *Hippoboscidae* (Ked flies), *Nycteribiidae* (Bat flies), and *Streblidae* (Bat flies) families [75].

552  Nutrient provisioning is accomplished by the secretion of a milk-like substance from specialized

553  glands into the uterus where the larval flies consume the milk. Dry weight of tsetse milk is

554  roughly 50% protein and 50% lipids [76]. A compiled list of the milk protein orthologs from the

555  six species of tsetse have been assembled (Supplemental Table 6).

556  Milk protein genes 2-10 (*mgp2-10*) in *G. m. morsitans* are the largest milk protein gene family.

557  These genes are tsetse specific, lack conserved functional protein domains and their origin is

558  currently unknown. However, experimental evidence suggests they act as lipid emulsification

559  agents and possible phosphate carrier molecules in the milk [77]. Search for orthologous

560  sequences to these genes revealed 1:1 orthologs to each of the 9 genes in the 5 new *Glossina*

561  species except for *G. brevipalpis* which lacks an orthologous sequence for the *mgp2* gene.

562  These genes are conserved at the levels of both synteny and sequence (Figure 8A+B).

563  Comparative expression analysis of these genes (and the other characterized milk protein

564  orthologs: *milk gland protein 1*, *acid sphingomyelinase* and *transferrin* [78, 79]) in male, non-

565  lactating and lactating females shows sex and lactation specific expression profiles across the

24

566   five species for which sex-specific RNA-seq data was available (Figure 8C+D). Comparison of

567   sequence variation across species for these genes by dN/dS analysis indicates that they are

568   under heavy negative selective pressure (Figure 8D). Enrichment analysis based on

569   comparison of lactation-based RNA-seq data confirms that these 12 orthologous sequences are

570   enriched in lactating flies across all *Glossina* (Figure 8E). The *mgp2-10* gene family is a unique

571   and conserved adaptation that appears essential to the evolution of lactation in the *Glossina*

572   genus. Determination of the origins of this protein family requires genomic analyses of other

573   members of the Hippoboscoidea superfamily that exhibit viviparity along with other species

574   closely related to this group.

575   *Tsetse seminal protein genes are rapidly evolving and vary in number and sequence*

576   *conservation between species (Figure 9)*

577   Recent proteomic analysis of male seminal proteins in *G. m. morsitans* revealed an array of

578   proteins transferred from the male to the female as components of the spermatophore [80].

579   Cross referencing of the proteomic data with tissue specific transcriptomic analyses of the

580   testes and male accessory glands (MAGs) allowed us to identify the tissues from which these

581   proteins are derived. Many of the MAG associated proteins are *Glossina*-specific and are

582   derived from gene families with multiple paralogs. These sequences were used to identify and

583   annotate orthologous sequences in the other five *Glossina* species. In contrast to the milk

584   proteins, sequence variance and differences in paralog numbers varies in male reproductive

585   genes between the six *Glossina* species.

586   This is particularly evident in the genes with MAG biased/specific expression. MAG

587   biased/specific genes are represented by 22 highly expressed gene families encoding

588   characterized seminal fluid proteins (SFP). We investigated the evolutionary rate of reproductive

589   genes over-expressed in the MAGs and testes, relative to a set of 5,513 *G. m. morsitans* genes,

590   orthologous between the six species (Figure 9A). The average dN/dS ratio is higher in MAG

591     biased genes than in testes biased genes or the entire *Glossina* ortholog gene set suggesting

592     that the MAG genes are under relaxed selective constraints. In addition, we found high

593     heterogeneity in the selective pressure across MAG genes. This is specifically evident in the

594     tsetse specific genes *GMOY002399*, *GMOY007759*, *GMOY004505* and *GMOY005874* (a

595     protein with OBP like conserved cysteine residues) as well as the OBP ortholog *GMOY007314*.

596     All five genes encode seminal fluid proteins as confirmed by the proteomic analysis of the

597     spermatophore [80].

598     In addition to sequence variability the number of paralogs per species differs as well (Figure

599     9B). This is similar to comparative analysis observations in *Anopheles* and *Drosophila* species

600     [81, 82]. This variance is especially evident in *Glossina* specific protein families (i.e.

601     *GMOY002399*, *GMOY004505/4506*, *GMOY005771*). In particular, there are a large number of

602     gene orthologs/paralogs to the *GMOY005771* gene across all *Glossina* species revealing a

603     large family of MAG genes of unknown function. The number of orthologs/paralogs differs

604     significantly between *Glossina* species. In addition, the two *G. m. morsitans* paralogs

605     *GMOY004724* and *GMOY004725* (predicted peptidase regulators), appear to display a higher

606     number of putative gene duplications in the *Morsitans* sub-genus relative to the *Palpalis* and

607     *Fusca* sub-Genera. Conservation appears instead to be more evident across testes genes that

608     code for proteins associated with conserved structural and functional components of sperm.

609     Overall, comparison of the MAG biased genes across *Glossina* reveals that this group shows

610     substantial variability in terms of genomic composition and rate of evolution. This is in

611     agreement with other studies indicating that male accessory proteins evolve at high rates due to

612     intraspecific competition between males or sexually antagonistic coevolution between males

613     and females [83].

614     **Olfactory associated protein-coding genes are conserved and reduced in number relative**

615     **to other Diptera.**

616   Comparative analyses of genes responsible for perireceptor olfaction activities revealed high

617   conservation of the repertoire among the six species. The genes appear to scatter across their

618   respective genomes with only a few duplicates occurring in clusters [84]. *Glossina* species

619   expanded loci that include Gr21a (responsible for $CO_2$ detection) [85], Or67d (mediates *cis-*

620   vaccenyl acetate reception) and Obp83a, (thought to be olfactory specific) [86]. The expanded

621   loci suggest involvement of gene duplication and/or transposition in their emergence [84]. All six

622   species lack sugar receptors likely as a result of tsetse's streamlined blood-feeding behavior.

623   Although our analysis did not reveal major discrepancies among the species, *G. brevipalpis* has

624   lost three key gustatory receptors (Gr58c, Gr66a and Gr32a) compared to other species. In

625   addition, *G. brevipalpis* showed higher structural gene rearrangements that could be attributed

626   to its evolutionary distance relative to the other tsetse species [87].

627   **A salivary protein gene shows sub-genus specific repeat motifs (Figure 10)**

628   Efficient acquisition of a blood meal by tsetse relies on a broad repertoire of physiologically

629   active saliva components inoculated at the bite site. These proteins modulate early host

630   responses, which, in addition to facilitating blood feeding can also influence the efficacy of

631   parasite transmission [88, 89]. The differences in the competence of different tsetse fly species

632   to develop mature *T. brucei* salivary gland infections may also be correlated with species-

633   specific variations in saliva proteins. Tsetse saliva raises a species-specific IgG response in

634   their mammalian hosts [90]. This response could potentially function as a biomarker to monitor

635   exposure of host populations to tsetse flies [91].

636   The *sgp3* gene [92] is characterized in all the tsetse species by two regions: a

637   metallophosphoesterase/5`nucleotidase and a repetitive glutamate/aspartate/asparagine-rich

638   region (Figure 10A). The complete sequence for this gene from *G. brevipalpis* could not be

639   obtained due to a gap in the sequence. The metallophosphoesterase/5`nucleotidase region is

640   highly conserved between all tsetse species. However, the sequences contain sub-genus

27

641    specific (*Morsitans* and *Palpalis*) repeat motifs within the glutamate/aspartate/asparagine

642    region. The motifs differ in size (32 amino acids in the *Morsitans* group and 57 amino acids in

643    the *Palpalis* group) and amino acid composition (Figure 10B). Moreover, within each sub-genus,

644    there are differences in the number of repetitive motifs. Within the *Morsitans* group, *G. m.*

645    *morsitans* and *G. pallidipes* have five motifs while *G. austeni* has only four. In the *Palpalis*

646    group, *G. palpalis* has three repetitive motifs and *G. fuscipes* five. Between the

647    metallophosphoesterase/5`nucleotidase and the glutamate/aspartate/asparagine-rich regions

648    there are a series of amino acids doublets comprising a lysine at the first position followed on

649    the second position by another amino acid (glutamic acid, glycine, alanine, serine, asparagine

650    or arginine). These differences may account for the differential immunogenic 'sub-Genus-

651    specific' antibody response caused by Sgp3 in *Morsitans* and *Palpalis* group flies [90].

652    **Comparison of vision associated Rhodopsin genes reveals conservation of motion**

653    **tracking receptors and variation in receptors sensitive to blue wavelengths (Figure 11).**

654    Vision plays an important role in host and mate seeking by flies within the *Glossina* genus. This

655    aspect of their biology is a critical factor in the optimization and development of trap/target

656    technologies [93, 94]. Analysis of the light sensitive Rhodopsin proteins across the *Glossina*

657    species reveals orthologs to those described in the *G. m. morsitans* genome (Figure 11A). The

658    expanded analysis provided by these additional genomes corroborates observations made for

659    the original *G. m. morsitans* genome, including the conservation of the blue sensitive *Rh5*

660    rhodopsin and the loss of one of the two dipteran UV-sensitive Rhodopsins: *Rh4* [26]. The

661    availability of the new genomes provides complete sequences for an additional long wavelength

662    sensitive Rhodopsin gene, *Rh2*. Prior to this analysis the recovery of a complete sequence from

663    *G. m. morsitans* was not possible due to poor sequence quality at its locus.

664    Rhodopsin protein sequence divergence among the six *Glossina* species and *M. domestica* (as

665    an outgroup) was investigated by calculating pairwise sequence divergence. As expected, the

666    average pairwise sequence divergence between *M. domestica* and any *Glossina* species is

667    higher than maximum sequence divergence among *Glossina* species for any of the five

668    investigated Rhodopsin subfamilies, ranging between 0.13 to 0.3 substitutions per 100 sites.

669    Average sequence divergence of *G. brevipalpis* to other *Glossina* is consistently lower than

670    *Musca* vs *Glossina* but also higher than the average pairwise distances between all other

671    *Glossina*, suggesting the older evolutionary lineage of *G. brevipalpis* (Figure 11B).

672    Three interesting aspects emerge in the comparison between subfamilies at the level of

673    sequence divergence between *Glossina* species. The *Rh1* subfamily, which is deployed in

674    motion vision, has the lowest average sequence divergence suggesting the strongest level of

675    purifying selection. *Rh2*, which is expressed in the ocelli, and *Rh5*, which is expressed in color-

676    discriminating inner photoreceptors, are characterized by conspicuously higher than average

677    sequence divergence among *Glossina* species. This observation could account for the varying

678    attractivity of trap and targets to different tsetse species.

## Conclusions

680    The comparative genomic analysis of these six *Glossina* species highlights important aspects of

681    *Glossina* evolution and provides further insights into their unique biology. Additional comparative

682    analyses of the genome assemblies, repetitive element composition, genes coding for

683    neuropeptides and their receptors, cuticular protein genes, transcription factor genes and

684    peritrophic matrix protein genes are available in the associated supplemental materials text,

685    figures and data files (Supplemental Material; Supplemental Figures 10 and 13, Supplemental

686    Tables 15, 16, 17, 18, 19; Supplemental Data 6 and 7). The results derived from the analysis of

687    these genomes are applicable to many aspects of tsetse biology including host seeking,

688    digestion, immunity, metabolism, endocrine function, reproduction and evolution. This expanded

689    knowledge has important practical relevance. Indeed, tsetse control strategies utilize trapping as

690    a key aspect of population management. These traps use both olfactory and visual stimuli to

691    attract tsetse. The findings of a reduced contingent of olfactory associated genes and the

692    variability of color sensing Rhodopsin genes provide research avenues into improvements of

693    trap efficacy. Deeper understanding of the important chemosensory and visual stimuli

694    associated with the different species could facilitate the refinement of trap designs for specific

695    species. The findings associated with *Glossina* digestive biology, including the enrichment of

696    proteolysis-associated genes and identification of *Glossina* specific expansions of immune

697    associated proteins provide new insights and avenues of investigation into vector competence

698    and vector/parasite relationships. Analysis of the female and male reproduction associated

699    genes reveals the differential evolutionary pressures on females and males. The conservation of

700    female milk proteins across species highlights the fact that this unique biology is optimized and

701    under strong negative evolutionary pressure. In counterpoint, male accessory gland derived

702    seminal proteins appear to have evolved rapidly between *Glossina* species and with little

703    conservation relative to other Diptera in gene orthology and functional conservation. Tsetse

704    reproduction is slow due to their unique viviparous adaptations, making these adaptations a

705    potential target for the development of new control measures. The knowledge derived from

706    these comparisons provide context and new targets for functional analysis of the genetics and

707    molecular biology of tsetse reproduction. In addition to the practical aspects of the knowledge

708    derived from these analyses, they also provide a look at the genetics underlying the evolution of

709    unique adaptive traits and the resources to develop deeper understanding of these processes.

## 710    Materials and Methods:

711    **Aim**

712    The aim of these studies was to generate and mine the genomic sequences of six species of

713    tsetse flies with different ecological niches, host preferences and vectorial capacities. The goals

714    of the analyses performed here are to identify novel genetic features specific to tsetse flies and

715    to characterize differences between the *Glossina* species to correlate genetic changes with

30

716    phenotypic differences in these divergent species. This was accomplished by the analyses

717    described below.

### *Glossina* strains

719    All genomes were sequenced from DNA obtained from 2-4 lines of flies originating from

720    individual pregnant females and their female offspring. Species collections were derived from

721    laboratory strains with varied histories (See Supplemental Table 8). The *G. pallidipes, G.*

722    *palpalis* and *G. fuscipes* flies were maintained in the laboratory at the Slovak Academy of

723    Sciences in Bratislava, Slovakia. The *G. brevipalpis* strain were maintained in the Insect Pest

724    Control Laboratory of the Joint FAO/IAEA Division of Nuclear Techniques in Food and

725    Agriculture, Seibersdorf, Austria. Finally, *G. austeni* were obtained from the Tsetse

726    Trypanosomiasis Research Institute in Tanga, Tanzania. Females were given two blood meals

727    supplemented with 20 mg/ml tetracycline to cure them of symbionts to eliminate non-tsetse

728    derived DNA.

### Genomic sequencing and assembly

730    Total genomic DNA was isolated from female pools for each species. High quality/ high

731    molecular weight DNA was isolated from individual flies using Genomic-tip purification columns

732    (QIAGEN) and the associated buffer kit. Samples were treated according to the protocol for

733    tissue-based DNA extraction. The pooled individual DNA isolates were utilized for sequencing

734    on Illumina HiSeq2000 instruments. The sequencing plan followed the recommendations

735    provided in the ALLPATHS-LG assembler [95]. Using this model, we targeted 45x sequence

736    coverage each of fragments (overlapping paired reads ~180bp length) and 3kb paired end (PE)

737    sequences as well as 5x coverage of 8kb PE sequences. The first draft assembly scaffold gaps

738    of each species were closed where possible with mapping of the same species assembly input

739    sequences (overlapping paired reads ~180bp length) and local gap assembly [96].

740    Contaminating sequences and contigs 200bp or less were removed (Supplemental Table 9).

741 **Scaffold mapping to Muller Elements and Sex Specific Muller Element Expression Biases**

742 We mapped scaffolds in each *Glossina spp.* genome assembly to chromosomes using

743 homology relationships with *D. melanogaster* (Supplemental Table 1). This method exploits the

744 remarkable conservation of chromosome arm (Muller element) gene content across flies [33,

745 97, 98]. We used the 1:1 orthologs between each *Glossina* species and *D. melanogaster* from

746 OrthoDB [99] to assign scaffolds from each species to Muller elements, applying an approach

747 previously developed for house fly [30]. For each species, a gene was assigned to a Muller

748 element if it was a 1:1 ortholog with a *D. melanogaster* gene. Then, each scaffold was assigned

749 to a Muller element if the majority (>50%) of genes with 1:1 orthologs on that scaffold were

750 assigned to a single Muller element.

751 We used the RNA-seq data (described below) to compare gene expression in males and

752 females. Expression comparisons were between male flies and either lactating (L) or non-

753 lactating (NL) females.

754 **Repeat feature annotation**

755 Repeat libraries for each species were generated using RepeatModeler [100]. The resultant

756 libraries were used to annotate the genome with RepeatMasker [101], alongside tandem and

757 low complexity repeats identified with TRF [102] and DUST [103]. The proportion of the genome

758 covered by repeats is shown in (Supplemental Table 10), with the figures for *G. m. morsitans*

759 provided for comparison.

760 **Automated gene annotation**

761 Gene annotation was performed with MAKER [104], using the first 2 rounds to iteratively

762 improve the training of the *ab initio* gene predictions derived from the combined Benchmarking

763 Universal Single-Copy Orthologs (BUSCO) [105] and Core Eukaryotic Genes Mapping

764 Approach (CEGMA) [106] HMMs, which were aligned to the genome assemblies using

765    GeneWise [107]. RNA-seq data for each species (described below) were used to build a

766    reference-guided transcriptome assembly with Tophat [108] and Cufflinks [109]. The initial

767    MAKER analysis produced unrealistically high numbers of gene models, so InterProScan [110]

768    and OrthoMCL [111] were used to identify gene predictions which lacked strong evidence. Only

769    gene models that met one or more of the following criteria were retained: (a) an annotation edit

770    distance < 1 [112]; (b) at least one InterPro domain (other than simple coils or signal peptides);

771    (c) an ortholog in the *Glossina* species complex. This process resulted in a reduction of 12-25%

772    in the number of gene models for each species (Supplemental Table 11). Genes from all six

773    species were assigned to 15,038 orthology groups via the Ensembl Compara 'GeneTrees'

774    pipeline [113].

775    For all types of ncRNA except tRNA and rRNA genes, we predicted RNA gene models by

776    aligning sequences from Rfam [114] against the genome using BLASTN [115]. The BLAST

777    results were then used to seed Infernal [116] searches of the aligned regions with the

778    corresponding Rfam covariance models. rRNA genes were predicted with RNAmmer [117], and

779    tRNA genes with tRNAScan-SE [118].

780    **Manual gene annotation**

781    *Glossina* sequence data and annotation data were loaded into the Apollo [119] community

782    annotation instances in VectorBase [120]. Manual annotations, primarily from a workshop held

783    in Kenya in 2015, underwent both manual and automated quality control to remove incomplete

784    and invalid modifications, and then merged with the automated gene set. Gene set versions are

785    maintained at (www.vectorbase.org) for each organism. All highlighted cells relate to the current

786    gene set version indicated in the table. Statistics for older gene set versions are provided along

787    with the relevant version number.

788    **Genome completeness analysis (BUSCO and CEGMA Analysis)**

33

789    Quality of the genome assembly and training of the *ab initio* predictors used in the gene

790    prediction pipeline was determined using the diptera_odb9 database which represents 25

791    Dipteran species and contains a total of 2799 BUSCO (Benchmarking Universal Single-Copy

792    Orthologs) genes derived from the OrthoDB v9 dataset [105] (Supplemental Table 12).

793    **Identification of Horizontal Gene Transfer Events**

794    All genome sequence files for *G. pallidipes*, *G. palpalis*, *G. fuscipes*, *G. austeni*, and *G.*

795    *brevipalpis* used for the whole genome assembly were also introduced into a custom pipeline for

796    the identification of putative Horizontal Gene Transfer (HGT) events between *Wolbachia* and

797    tsetse. *Wolbachia* sequences were filtered out from WGS reads using a combination of MIRA

798    [121] and NextGenMap [122] mapping approaches. The reference sequences used were *w*Mel

799    (AE017196), *w*Ri (CP001391), *w*Bm (AE017321), *w*Gmm (AWUH01000000), *w*Ha

800    (NC_021089), *w*No (NC_021084), *w*Oo (NC_018267), *w*Pip (NC_010981), and the

801    chromosomal insertions A and B in *G. morsitants morsitans.* All filtered putative *Wolbachia*-

802    specific sequences were further examined using blast and custom-made databases.

803    To identify the chromosomal *Wolbachia* insertions, the following criteria were used: Sequences

804    that (relative to the reference genomes): (a) exhibit high homology to insertion sequences A & B

805    from *G. m. morsitans*, (b) exhibit a high degree of nucleotide polymorphisms (at least 10

806    polymorphisms/100bp) with the reference genomes, and (c) contain a high degree of

807    polymorphism coupled with insertions and/or deletions. *Wolbachia* specific sequences for each

808    *Glossina* species were assembled with MIRA using a *de novo* approach. For *G. pallidipes, G.*

809    *palpalis, G. fuscipes,* and *G. brevipalpis* assembled sequences corresponding only to

810    cytoplasmic *Wolbachia* were identified*.* Genomic insertions were only observed in assembled

811    sequences from *G. austeni* (Supplemental Table 2). The statistics for the *G. austeni* assembled

812    sequences are as follows: N50 4493, N90 1191, and mean contig length, 2778bps. During the

34

813    process of identifying HGT events in *G. fuscipes*, we also recovered *Spiroplasma* sequences

814    but none of the recovered sequences were chromosomal.

815    **Whole-genome pairwise alignment**

816    We generated all possible pairwise alignments between the six *Glossina* species (including *G.*

817    *m. morsitans*) and an outgroup, *M. domestica*, using the Ensembl Compara software pipeline

818    [123]. LASTZ [124] was used to create pairwise alignments, which were then joined to create

819    'nets' representing the best alignment with respect to a reference genome [125]. *G. m.*

820    *morsitans* was always used as the reference for any alignment of which it was a member,

821    otherwise the reference genome was randomly assigned. Coverage statistics and configuration

822    parameters for all alignments are available at

823    https://www.vectorbase.org/compara_analyses.html.

824    **Glossina phylogeny prediction**

825    We identified orthologous genes across the six *Glossina* species and six outgroups (*M.*

826    *domestica*, *D. melanogaster*, *D. ananassae*, *D. grimshawi*, *L. longipalpis*, and *A. gambiae*) by

827    employing a reciprocal-best-hit (RBH) approach in which *G. m. morsitans* was used as focal

828    species. We identified 286 orthologs with a clear reciprocal relationship among the 12 species.

829    All orthologs were aligned individually using MAFFT [126] and concatenated in a super-

830    alignment of 478.617 nucleotide positions. The nucleotide alignment was translated in the

831    corresponding amino acids and passed through Gblocks [127] (imposing "half allowed gap

832    positions" and leaving remaining parameters at default) to obtain a dataset of 117.783 amino

833    acid positions. This dataset was used for a Maximum Likelihood analysis in RAxML [128]

834    employing the LG+G+F model of replacement, and for a Bayesian analysis using Phylobayes

835    [129, 130] employing the heterogeneous CAT+G model of replacement. We further performed a

836    coalescent-aware analysis using Astral [131] and the 286 single gene trees obtained using

35

837    Raxml [128] and analyzing the alignments at the nucleotide level with the GTR+G model of

838    replacement.

839    **Mitochondrial genome analysis and phylogeny**

840    The mtDNA genomes of *G. m. centralis* and *G. brevipalpis* were sequenced using the Illumina

841    HiSeq system and about 15 kb of mitochondrial sequence of each species was obtained. These

842    sequences were used to identify the mtDNA sequences within the sequenced tsetse genomes

843    (*G. pallidipes, G. m. morsitans, G. p. gambiensis, G. f. fuscipes* and *G. austeni)* from the

844    available genomic data. Sanger sequencing confirmed the mtDNA genome sequence of each

845    tsetse species. This involved PCR amplification of the whole mtDNA genome using fourteen

846    pairs of degenerate primers designed to cover the whole mitochondrial genomes of the

847    sequences species (Supplemental Table 13). The PCR products were sent for Sanger

848    sequencing. The sequences obtained by Sanger and Illumina sequencing for each species were

849    assembled using the SegMan program from the lasergene software package (DNAStar Inc.,

850    Madison, USA). The phylogenetic analysis based on these sequences was performed using

851    maximum likelihood method with the MEGA 6.0 [132].

852    **Synteny analysis**

853    The synteny analysis was derived from whole genome alignments performed as follows using

854    tools from the UCSC Genome Browser [133]. The LASTZ software package (version 1.02.00)

855    generated the initial pairwise sequence alignments with the parameters: *E=30, H=2000,*

856    *K=3000, L=2200, O=400* and the default substitution matrix. From these alignments, Kent's

857    toolbox (version 349) [133] was used to generate chain and nets (higher-level abstractions of

858    pairwise sequence alignments) with the parameters: *-verbose=0 -minScore=3000 and -*

859    *linearGap=medium.* The multiple alignment format (MAF) files were built with MULTIZ for TBA

860    package (version 01.21.09) [134], using the chains and nets, along with the phylogenetic

861    relationships and distances between species. Using the MAF files, pairwise homologous

36

862  synteny blocks (HSBs) were automatically defined using the SyntenyTracker software [135].

863  Briefly, the SyntenyTacker software defines an HSB as a set of two or more consecutive

864  orthologous markers in homologous regions of the two genomes, such that no other defined

865  HSB is within the region bordered by these markers. There are two exceptions to this rule: the

866  first involves single orthologous markers not otherwise defined within HSBs; and the second

867  involves two consecutive singleton markers separated by a distance less than the resolution

868  threshold (10 kb for this analysis). As the 10 Kb blocks were too small for visualisation in Circos

869  [136], they were aggregated into larger 100 Kb histogram blocks, where each 100 Kb Circos

870  block shows the fraction of sequence identified as syntenic for a particular species when aligned

871  to *D. melanogaster*. Synteny blocks are available for visualisation from the Evolution Highway

872  comparative chromosome browser: http://eh-demo.ncsa.uiuc.edu/drosophila/.

873  **Orthology and paralogy inference and analysis**

874  Phylogenetic trees were inferred with the Ensembl Compara 'GeneTrees' pipeline [123, 137]

875  using all species from the VectorBase database of arthropod disease vectors [120]. The trees

876  include 33 non-*Glossina* species, such as *D. melanogaster*, which act as outgroup comparators.

877  All analyses are based on the VectorBase April 2016 version of the phylogenetic trees.

878  Representative proteins from all genes were clustered and aligned, and trees showing orthologs

879  and paralogs were inferred with respect to the NCBI taxonomy tree

880  (http://www.ncbi.nlm.nih.gov/taxonomy).

881  The 15,038 predicted gene trees containing *Glossina* sequences were parsed to quantify the

882  trees based on their constituent species. Raw tree files (Supplemental Data 1) were parsed

883  using a custom PERL script (Supplemental File 1) to determine gene counts for representative

884  Dipteran species for each gene tree. Count data were imported into Excel and filtered using

885  pivot tables to categorize orthology groups based upon species constitution (Supplemental Data

886  2).

887   The orthology groups were broken into cohorts based on the phylogenetic composition of

888   species within each group. The *Glossina* containing orthology groups were categorized as

889   follows: common to Diptera (including the Nematocera sub-order), Brachycera sub-order

890   specific, *Glossina* genus specific, *Glossina* sub-genus specific (*Morsitans* and *Palpalis*) or

891   *Glossina* species specific. Each category is sub-divided into two groups, universal groups that

892   contain representative sequences from all species within the phylogenetic category or partial

893   orthology groups containing sequences from some but not all members of the phylogenetic

894   category. *Glossina* gene IDs and associated FASTA sequences associated with groups of

895   interest were extracted using a custom Perl script for gene ontology analysis (Supplemental File

896   2)

**Gene Ontology (GO) analysis**

898   Gene associated GO terms were obtained from the VectorBase annotation database via the

899   BioMart interface. Genes from *Glossina* genus, and sub-genus specific orthology groups were

900   isolated and tested for enrichment of GO terms. Analysis for GO terms for enrichment was

901   performed with the R package "topGO". The enriched genes were separated into species

902   specific lists compared against the entirety of predicted protein-coding genes from the

903   respective species. Significance of enrichment was determined using Fisher's Exact Test

904   (Supplemental Table 3 + Supplemental Data 3).

**Identification and analysis of gene expansions/contractions**

906   Gene trees containing orthologs/paralogs representing each of the six *Glossina* species were

907   analyzed to identify sub-genus associated gene expansions/contractions. Gene trees were

908   considered for analysis if the variance in the number of orthologs/paralogs between the 6

909   species was greater than 2. Variable gene trees were tested for phylogenetic significance

910   relative to the predicted *Glossina* phylogeny using the CAFE software package [138] to reject

911   potentially inaccurate variance predictions due to erroneous gene annotations. Gene trees with

912    a CAFE score of <0.05 were considered significant (Supplemental table 4 + Supplemental data

913    4).

914    Sequences from gene trees satisfying the variance and CAFE thresholds were extracted with a

915    custom PERL script (Supplemental File 2) and analyzed by BLASTP analysis against an insect

916    specific subset of the NCBI NR database. Gene trees were annotated with the most common

917    description associated with the top BLAST hits of its constituent sequences. Gene trees were

918    subjected to PCA analysis in R using the FactoMineR and Factoextra packages using species

919    specific gene counts as input data. The results were plotted and annotated with their associated

920    BLAST derived descriptions.

921    **RNA-seq data**

922    Total RNA was isolated for each of the six tsetse species from whole male and whole female

923    (non-lactating and lactating) for RNA-seq library construction. Poly(A)+ RNA was isolated, then

924    measured with an Agilent Bioanalyzer for quality. Samples were considered to be of high quality

925    if they showed intact ribosomal RNA peaks and lacked signs of degradation. Samples passing

926    quality control were used to generate non-normalized cDNA libraries with a modified version of

927    the Nu-GEN Ovation® RNASeq System V2 (http://www.nugeninc.com). We sequenced each

928    cDNA library (0.125 lane) on an Illumina HiSeq 2000 instrument (~36 Gb per lane) at 100 bp in

929    length.

930    RNA-seq analyses were conducted based on methods described in Benoit et al. [77],

931    Rosendale et al.[139], and Scolari et al.[140] with slight modifications. RNA-seq datasets were

932    acquired from whole males, whole dry females, and whole lactating females. The SRA numbers

933    for each of the libraries are listed in (Supplemental Table 14).

934    RNA-seq datasets were quality controlled using the FastQC (Babraham Bioinformatics)

935    software package. Each set was trimmed/cleaned with CLC Genomics (Qiagen) and quality was

936    re-assessed with FastQC. Each dataset was mapped to the predicted genes from each

937    *Glossina* genome with CLC Genomics. Each read required at least 95% similarity over 50% of

938    length with three mismatches allowed. Transcripts per million (TPM) was used as a proxy for

939    gene expression. Relative transcript abundance differences were determined as the TPM in one

940    sample relative to the TPM of another dataset (*e.g.* male/lactating Female). A proportion based

941    statistical analysis [141] followed by Bonferroni correction at 0.05 was used to identify genes

942    with significant sex and stage specific transcript enrichment. This stringent statistical analysis

943    was used as only one replicate was available for each treatment.

944    Enriched transcripts in lactating and dry transcriptomes from the species examined were

945    compared to orthologous sequences in *G. m. morsitans* [142]. Overlap was determined by

946    comparison of the enrichment status of orthologous sequences in the *Glossina* species tested.

947    The results of this analysis are visualized in a Venn diagram

948    (http://bioinformatics.psb.ugent.be/webtools/Venn/). Determination of dN/dS values and

949    production of phylogenetic trees was conducted with the use of DataMonkey [143, 144] for

950    dN/dS analyses and MEGA5 for alignment and tree construction [145].

951    **Cuticular Protein Analysis**

952    The predicted peptide sequences from each species were analyzed by BLASTp analysis [115]

953    against characteristic sequence motifs derived from several families of cuticle proteins [146].

954    Predicted cuticle proteins were further analyzed with CutProtFam-Pred, a cuticle protein

955    prediction tool described in Ioannidou et al. [147], to assign genes to specific families of cuticle

956    proteins. To find the closest putative homolog to cuticle protein genes in Glossina, genes were

957    searched (BLASTp) against Refseq protein database from the National Center for

958    Biotechnology Information (NCBI). The protein sequences with the lowest e-value were

959    considered the closest putative homologs (Supplemental Data 5).

960    **Transcription factor identification and annotation**

40

961 Likely transcription factors (TFs) were identified by scanning the amino acid sequences of

962 predicted protein-coding genes for putative DNA binding domains (DBDs). When possible, we

963 predicted the DNA binding specificity of each TF using the procedures described in Weirauch *et*

964 *al.* [148]. Briefly, we scanned all protein sequences for putative DBDs using the 81 Pfam [149]

965 models listed in Weirauch and Hughes [150] and the HMMER tool [151], with the recommended

966 detection thresholds of Per-sequence Eval < 0.01 and Per-domain conditional Eval < 0.01. Each

967 protein was classified into a family based on its DBDs and their order in the protein sequence

968 (e.g., bZIPx1, AP2x2, Homeodomain+Pou). We then aligned the resulting DBD sequences

969 within each family using clustalOmega [152], with default settings. For protein pairs with multiple

970 DBDs, each DBD was aligned separately. From these alignments, we calculated the sequence

971 identity of all DBD sequence pairs (*i.e.* the percent of AA residues that are identical across all

972 positions in the alignment). Using previously established sequence identity thresholds for each

973 family [148], we mapped the predicted DNA binding specificities by simple transfer. For

974 example, the DBD of the *G. austeni* GAUT024062-PA protein is identical to the DBD of *D.*

975 *melanogaster* mirr (FBgn0014343). Since the DNA binding specificity of mirr has already been

976 experimentally determined, and the cutoff for Homeodomain family of TFs is 70%, we can infer

977 that GAUT024062-PA will have the same binding specificity as mirr. All associated data can be

978 found in (Supplemental Data 6)

979

980 **Abbreviations:**

981 DBD – DNA binding domain, HAT – Human African Trypanosomiasis, AAT – Animal African

982 Trypanosomiasis, mtDNA – mitochondrial DNA, rRNA – ribosomal RNA, tRNA- transfer RNA,

983 HGT – horizontal gene transfer, OG – orthology group, GO – gene ontology, OBP – odorant

984 binding protein, miRNA – micro RNA, siRNA – small interfering RNA, piRNA – piwi interacting

985 RNA, MGP – milk gland protein, MAG – male accessory gland, SFP – seminal fluid protein.

## Declarations:

**Ethics approval and consent to participate:**

Not applicable

**Consent for publication:**

Not applicable

**Availability of data and material:**

The genomes, transcriptomes and predicted protein-coding sequences are available from VectorBase via the following link https://www.vectorbase.org/taxonomy/glossina. The raw RNA-seq datasets generated and/or analyzed during the current study are available from the NCBI SRA database repository, at the following link https://www.ncbi.nlm.nih.gov/sra/SRP158014.

All data generated during the analyses of these dataset are included in this published article and its supplementary information files.

**Competing interests:**

**Funding:**

1009    **Contributions** ([&] = Annotation group leader, [$] = Project leader)

1010    Analysis of whole genomic sequences and database management: D.L., E.L., G.L.M., M.B.;

1011    BUSCO and Female reproductive gene analysis: E.C.J., J.B.B.[&], V.M.; Chemosensory gene

1012    analysis: D.M., P.O.M., R.W.M.; Cuticular protein gene analysis: A.J.R., D.W.F.; Gene orthology

1013    and expansion analyses: G.M.A.[$&], J.E.A.; Genome sequencing, assembly and analysis:

1014    WCW[$], C.T., P.M., R.K.W.; Horizontal gene transfer analysis: G.T., K.B.; Immune gene

1015    analysis: A.V., B.L.W., J.W., R.B.; Male reproductive gene analysis: A.R.M., F.S., G.S.;

1016    Mitochondrial DNA sequence analysis: A.M.M.A., I.M., A.G.P.; Molecular evolution and

1017    phylogenetic analyses: L.O., O R-S; Neuropeptide and G protein-coupled receptor analysis:

1018    H.G.M.[&], J.C.[&], L.S.; Rhodopsin gene analysis: M.F.; Orthology and comparative genomics

1019    analysis advice, manuscript editing: R.M.W.; Peritrophin gene analysis: A.A-S.[&], C.R.; Project

1020    conception: S.A.[$], M.J.L.[$]; Project funding: S.A. [$], Project management: S.A. [$], W.C.W.,

1021    Provision of experimental material: P.T., S.A. [$], A.M.M.A.; Salivary protein gene analysis:

1022    J.V.D.A., I.M.[&], G.C., X.Z.; Symbiont associated gene analysis: R.R.[&]; Syntenic analysis of

1023    genomes: M.T.S., D.M.L., V.P.E.L.; Transcription factor and DNA binding motif prediction:

1024    M.T.W.; Transposable element analysis: A.H.V., W.J.M.; X chromosome and sex linked

1025    expression analysis: R.P.M.

1026    **Acknowledgements**

43

1043

1044 **Figure 1: Geographic distribution, ecology and vectorial capacity of sequenced *Glossina***

1045 **species.** Visual representation of the geographic distribution of the sequenced *Glossina* species

1046 across the African continent. Ecological preferences and vectorial capacities are described for

1047 each associated group.

1048 **Figure 2: *Glossina* whole genome alignment, phylogenetic analysis of orthologous**

1049 **protein-coding nuclear genes and phylogenetic analysis of mitochondrial sequences.**

1050 A. Analysis of whole genome and protein-coding sequence alignment. The left graph reflects the

1051 percentage of total genomic sequence aligning to the *G. m. morsitans* reference. The right side

1052 of the graph represents alignment of all predicted coding sequences from the genomes with

1053 coloration representing matches, mismatches, insertions and uncovered exons. B. Phylogenic

1054 tree from conserved protein-coding sequences. Black dots at nodes indicate full support from

1055 Maximum likelihood (Raxml), Bayesian (Phylobayes), and coalescent-aware (Astral) analyses.

1056 Raxml and Phylobayes analyses are based on an amino acid dataset of 117,782 positions from

1057 286 genes from 12 species. Astral analyses is based on a 1125 nucleotide dataset of 478,617

1058 positions from the 6 *Glossina* (full trees are in Supplemental Figure 1A-C). C. Molecular

1059 phylogeny derived from whole mitochondrial genome sequences. The analysis was performed

1060 using the maximum likelihood method with MEGA 6.0.

1061 **Figure 3: Visualization of syntenic block analysis data and predicted Muller Element**

1062 **sizes**

1063 Level of syntenic conservation between tsetse scaffolds and *Drosophila* chromosomal

1064 structures (Muller Elements). The color-coded concentric circles consisting of bars represent the

1065 percent of syntenic conservation of orthologous protein-coding gene sequences between the

1066 *Glossina* genomic scaffolds and *Drosophila* Muller elements. Each bar represents 100 kb of

1067 aligned sequence and bar heights represent the percent of syntenic conservation. The graphs

1068 on the periphery of the circle illustrate the combined predicted length and number of genes

45

1069   associated with the Muller elements for each tsetse species. The thin darkly colored bars

1070   represent the number of 1:1 orthologs between each *Glossina* species and *D. melanogaster*.

1071   The thicker lightly colored bands represent the predicted length of each Muller element for each

1072   species. This was calculated as the sum of the lengths of all scaffolds mapped to those Muller

1073   elements.

1074   **Figure 4: Homology map of the *Wolbachia* derived cytoplasmic and horizontal transfer**

1075   **derived nuclear sequences.** Circular map of the *G. austeni Wolbachia* horizontal transfer

1076   derived genomic sequences (*wGau* - blue), the *D. melanogaster Wolbachia* cytoplasmic

1077   genome sequence (*wMel* - green), the *G. m. morsitans Wolbachia* cytoplasmic genome

1078   sequence (*wGmm* - red), and the *Wolbachia* derived chromosomal insertions A & B from *G. m.*

1079   *morsitans* (*wGmm* Insertion A and Insertion B yellow and light yellow respectively). The

1080   outermost circle represents the scale in kbp. Contigs for the *wGau* sequences, *wGmm* and the

1081   chromosomal insertions A & B in *G. m. morsitans* are represented as boxes. Regions of

1082   homology between the *G. austeni* insertions and the other sequences are represented by

1083   orange ribbons. Black ribbons represent syntenic regions between the *wGau* insertions and the

1084   cytoplasmic genomes of *wGmm* and *wMel*.

1085   **Figure 5: Constituent analysis of *Glossina* associated gene orthology groups.**

1086   Visualization of relative constitution of orthology groups containing *Glossina* gene sequences.

1087   Combined bar heights represent the combined orthogroups associated with each *Glossina*

1088   species. The bars are color-coded to reflect the level of phylogenetic representation of clusters

1089   of orthogroups at the order, sub-order, genus, sub-genus and species. Saturated bars represent

1090   orthology groups specific and universal to a phylogenetic level. Desaturated bars represent

1091   orthogroups specific to a phylogenetic level but lack universal representation across all included

1092   species. Gene ontology analysis of specific and universal groups can be found in Table 3.

46

1093 **Figure 6: Sub-genus specific gene family expansions/retractions.** Principal component

1094 analysis-based clustering of gene orthology groups showing significant differences in the

1095 number of representative sequences between the six *Glossina* species. Orthology groups

1096 included have sub-genus specific expansions/contractions as determined by CAFE test (P-value

1097 < 0.05). Detailed information regarding the functional associations of the unlabeled groups is

1098 provided in Supplemental Figures 10+11 and in Supplemental Table 4.

1099 **Figure 7: Heat map of counts of *Glossina* homologs to *Drosophila* immune genes.** Counts

1100 of *Glossina* sequences within ortholog groups containing *Drosophila* genes annotated with the

1101 "Immune System Process" GO tag (GO:0002376).

1102 **Figure 8: Conservation of synteny, sequence homology and stage/sex specific**

1103 **expression of tsetse milk proteins between species.** Overview of the conservation of tsetse

1104 milk protein genes and their expression patterns in males, non-lactating and lactating females.

1105 A.) Syntenic analysis of gene structure/conservation in the *mgp2-10* genetic locus across

1106 *Glossina* species. B.) Phylogenetic analysis of orthologs from the *mgp2-10* gene family. C.)

1107 Combined sex and stage specific RNA-seq analysis of relative gene expression of the 12 milk

1108 protein gene orthologs in males, non-lactating and lactating females of 5 *Glossina* species. D.)

1109 Visualization of fold change in individual milk protein gene orthologs across 5 species between

1110 lactating and non-lactating female flies. Gene sequence substitution rates are listed for each set

1111 of orthologous sequences. E.) Comparative enrichment analysis of differentially expressed

1112 genes between non-lactating and lactating female flies.

1113 **Figure 9: Comparative analysis of *Glossina* male accessory gland (MAG) protein family**

1114 **memberships**

1115 Graphical representation of the putative *Glossina* orthologs and paralogs to characterized MAG

1116 genes from *G. m. morsitans.* The genes are categorized by their functional classes as derived

1117 by orthology to characterized proteins from *Drosophila* and other insects.

47

1118 **Figure 10: 5'Nuc/apyrase salivary gene family organization and sequence features across**

1119 **Glossina species.** A.) Chromosomal organization of the 5'Nuc/apyrase family orthologs on

1120 genome scaffolds from the six *Glossina* species. The brown gene annotations represent 5'Nuc

1121 gene orthologs; purple gene annotations represent *sgp3* gene orthologs and the blue gene

1122 annotations an apyrase-like encoding gene. The broken rectangular bars on the *G. brevipalpis*

1123 scaffold indicate that the sequence could not be determined due to poor sequence/assembly

1124 quality. B.) Schematic representation of sgp3 gene structure in tsetse species. The K(.) denotes

1125 a repetition of a Lysine (K) and another amino acid (Glutamic acid, Glycine, Alanine, Serine,

1126 Asparagine or Arginine). The green oval represents a repetitive motif found in *Morsitans* sub-

1127 genus; the red oval represents a repetitive motif found in *Palpalis* group. The dashed line

1128 indicates a partial motif present. For each of the two motifs the consensus sequence is shown in

1129 the right by a Logo sequence. The poor sequence/assembly quality of the *G. brevipalpis*

1130 scaffold prevented inclusion of this orthology in the analysis.

1131 **Figure 11: Phylogenetic and sequence divergence analysis of Glossina vision associated**

1132 **proteins.** Phylogenetic and sequence conservation analysis of the vision associated *Rhodopsin*

1133 G-protein coupled receptor genes in *Glossina* and orthologous sequences in other insects. A.)

1134 Phylogenetic analysis of Rhodopsin protein sequences. B.) Pairwise analysis of sequence

1135 divergence between *M. domestica* and *Glossina* species and within the *Glossina* genus.

1136 **Supplemental Figure 1:** Maximum Likelihood, Bayesian and Astral based phylogenetic

1137 analysis of a concatenated single gene ortholog alignment.

1138 **Supplemental Figure 2:** Tsetse variable mitochondrial DNA sequence and species delineation

1139 by high resolution melt curve analysis

1140 **Supplemental Figure 3:** Application of high-resolution melt curve analysis to distinguish Tsetse

1141 haplotypes/populations

1142 **Supplemental Figure 4:** The percent of female-, male-, and un-biased genes that are on an X

1143 chromosome scaffold (Muller elements A, D, or F) is plotted for each species. Sex-biased

1144 expression was measured between males and either lactating (L) or non-lactating (NL) females.

1145 Asterisks indicate a significant difference between the percent of sex-biased genes that are X-

1146 linked when compared to unbiased genes (*$P<0.05$ in Fisher's exact test).

1147 **Supplemental Figure 5:** The distribution of the log2 (fold-change between females and males)

1148 is plotted for autosomal and X-linked genes in each species. Female-male gene expression

1149 comparisons are between males and either lactating or non-lactating females.

1150 **Supplemental Figure 6:** Rates of non-synonymous to synonymous substitution (dN/dS) along

1151 different evolutionary lineages within the *Glossina* genus.

1152 **Supplemental Figure 7:** Rates of non-synonymous to synonymous substitution (dN/dS) rates

1153 of female, male and non-sex biased genes across the X and autosomal muller elements.

1154 **Supplemental Figure 8:** The distribution of male biased gene expression across the predicted

1155 *Glossina* Muller Elements. Bar heights represent the percentage of genes per element with

1156 male biased gene expression.

1157 **Supplemental Figure 9:** Rates of non-synonymous to synonymous substitution (dN/dS) rates

1158 across the predicted muller elements.

1159 **Supplemental Figure 10: Repetitive element constitution of Glossina genomes.** Analysis of

1160 repetitive element composition across the six *Glossina* species. A.) Graphical representation of

1161 the constitution and sequence coverage by the various classes of identified repetitive elements.

1162 B.) Relative constitution of DNA Terminal Inverted Repeat (TIR) families across the *Glossina*

1163 genomes. C.) Relative constitution of Long Interspersed Nuclear Elements (LINEs) across the

1164 *Glossina* genomes.

1165 **Supplemental Figure 11: Sub-genus specific gene family expansions/retractions (with**

1166 **functional annotations).** Principal component analysis-based clustering of gene orthology

1167 groups showing significant differences in the number of representative sequences between the

1168 six *Glossina* species.

1169 **Supplemental Figure 12: Sub-genus specific gene family expansions/retractions (with**

1170 **orthology group number annotations).** Principal component analysis-based clustering of

1171 gene orthology groups showing significant differences in the number of representative

1172 sequences between the six *Glossina* species.

1173 **Supplemental Figure 13: Distribution of transcription factor families across insect**

1174 **genomes.** Heatmap depicting the abundance of transcription factor (TF) families across a

1175 collection of insect genomes. Each entry indicates the number of TF genes for the given family

1176 in the given genome, based on presence of DNA binding domains. Color key is depicted at the

1177 top (light blue means the TF family is completely absent) – note log (base 2) scale.

# 1178 References:

1179 1.    Lyons M: *The Colonial Disease. A social History of Sleeping Sickness in Norther Zaire, 1900-1940.*
1180       Cambridge UK: Cambridge University Press; 1992.
1181 2.    Odiit M, Coleman PG, Liu WC, McDermott JJ, Fevre EM, Welburn SC, Woolhouse ME:
1182       **Quantifying the level of under-detection of *Trypanosoma brucei rhodesiense* sleeping sickness**
1183       **cases.** *Trop Med Int Health* 2005, **10:**840-849.
1184 3.    Franco JR, Cecchi G, Priotto G, Paone M, Diarra A, Grout L, Simarro PP, Zhao W, Argaw D:
1185       **Monitoring the elimination of human African trypanosomiasis: Update to 2016.** *PLoS Negl Trop*
1186       *Dis* 2018, **12:**e0006890.
1187 4.    Franco JR, Simarro PP, Diarra A, Ruiz-Postigo JA, Jannin JG: **The journey towards elimination of**
1188       **gambiense human African trypanosomiasis: not far, nor easy.** *Parasitology* 2014, **141:**748-760.
1189 5.    Steelman CD: **Effects of external and internal arthropod parasites on domestic livestock**
1190       **production.** *Annual Review of Entomology* 1976, **21:**155-178.
1191 6.    Jordan A: *Trypanosomiasis Control and African Rural Development.* London: Longman; 1986.
1192 7.    Budd LT: *Tsetse and Trypanosomosis Research and Development since 1980: an Economic*
1193       *Analysis.* UK; 1999.
1194 8.    Alsan M: **The Effect of the TseTse Fly on African Development.** *American Economic Review*
1195       2015, **105:**382-410.
1196 9.    Opigo J, Woodrow C: **NECT trial: more than a small victory over sleeping sickness.** *Lancet* 2009,
1197       **374:**7-9.

1198 10. Mesu V, Kalonji WM, Bardonneau C, Mordt OV, Blesson S, Simon F, Delhomme S, Bernhard S,
1199      Kuziena W, Lubaki JF, et al: **Oral fexinidazole for late-stage African Trypanosoma brucei**
1200      **gambiense trypanosomiasis: a pivotal multicentre, randomised, non-inferiority trial.** *Lancet*
1201      2018, **391**:144-154.
1202 11. Buscher P, Deborggraeve S: **How can molecular diagnostics contribute to the elimination of**
1203      **human African trypanosomiasis?** *Expert Rev Mol Diagn* 2015, **15**:607-615.
1204 12. Anene BM, Onah DN, Nawa Y: **Drug resistance in pathogenic African trypanosomes: what**
1205      **hopes for the future?** *Vet Parasitol* 2001, **96**:83-100.
1206 13. Geerts S, Holmes PH, Eisler MC, Diall O: **African bovine trypanosomiasis: the problem of drug**
1207      **resistance.** *Trends Parasitol* 2001, **17**:25-28.
1208 14. Lehane M, Alfaroukh I, Bucheton B, Camara M, Harris A, Kaba D, Lumbala C, Peka M, Rayaisse JB,
1209      Waiswa C, et al: **Tsetse Control and the Elimination of Gambian Sleeping Sickness.** *PLoS Negl*
1210      *Trop Dis* 2016, **10**:e0004437.
1211 15. Solano P, Torr SJ, Lehane MJ: **Is vector control needed to eliminate *gambiense* human African**
1212      **trypanosomiasis?** *Front Cell Infect Microbiol* 2013, **3**:33.
1213 16. Courtin F, Camara M, Rayaisse JB, Kagbadouno M, Dama E, Camara O, Traore IS, Rouamba J,
1214      Peylhard M, Somda MB, et al: **Reducing Human-Tsetse Contact Significantly Enhances the**
1215      **Efficacy of Sleeping Sickness Active Screening Campaigns: A Promising Result in the Context of**
1216      **Elimination.** *PLoS Negl Trop Dis* 2015, **9**:e0003727.
1217 17. Ilboudo H, Jamonneau V, Camara M, Camara O, Dama E, Leno M, Ouendeno F, Courtin F,
1218      Sakande H, Sanon R, et al: **Diversity of response to Trypanosoma brucei gambiense infections**
1219      **in the Forecariah mangrove focus (Guinea): perspectives for a better control of sleeping**
1220      **sickness.** *Microbes Infect* 2011, **13**:943-952.
1221 18. Molyneux DH: **Animal reservoirs and Gambian trypanosomiasis.** *Ann Soc Belg Med Trop* 1973,
1222      **53**:605-618.
1223 19. Kabayo JP: **Aiming to eliminate tsetse from Africa.** *Trends Parasitol* 2002, **18**:473-475.
1224 20. Krafsur ES: **Tsetse flies: genetics, evolution, and role as vectors.** *Infect Genet Evol* 2009, **9**:124-
1225      141.
1226 21. Travassos Santos Dias J: **Contribuição para o estudo da sistemática do género Glossina**
1227      **Wiedamann 1830 (Insecta, Brachycera, Cyclorrphapha, Glossinidae) Proposta para a criação de**
1228      **um novo subgénero.** *Garcia de Orta, Ser Zool, Lisboa* 1987, **14**:67-78.
1229 22. Rogers D, Robinson T: **Tsetse distribution.** In *The Trypanosomiases.* Edited by Maudlin I, Holmes
1230      P, Miles M. Oxford: CAB International;; 2004: 139–179.
1231 23. Moloo SK, Kabata JM, Sabwa CL: **A study on the maturation of procyclic Trypanosoma brucei**
1232      **brucei in Glossina morsitans centralis and G. brevipalpis.** *Medical and Veterinary Entomology*
1233      1994, **8**:369-374.
1234 24. Motloang M, Masumu J, Mans B, Van den Bossche P, Latif A: **Vector competence of Glossina**
1235      **austeni and Glossina brevipalpis for Trypanosoma congolense in KwaZulu-Natal, South Africa.**
1236      *Onderstepoort Journal of Veterinary Research* 2012, **79**:E1-6.
1237 25. Aksoy S, Berriman M, Hall N, Hattori M, Hide W, Lehane MJ: **A case for a *Glossina* genome**
1238      **project.** *Trends in Parasitology* 2005, **21**:107-111.
1239 26. Initiative IGG: **Genome sequence of the tsetse fly (Glossina morsitans): vector of African**
1240      **trypanosomiasis.** *Science* 2014, **344**:380-386.
1241 27. Dyer NA, Lawton SP, Ravel S, Choi KS, Lehane MJ, Robinson AS, Okedi LM, Hall MJ, Solano P,
1242      Donnelly MJ: **Molecular phylogenetics of tsetse flies (Diptera: Glossinidae) based on**
1243      **mitochondrial (COI, 16S, ND2) and nuclear ribosomal DNA sequences, with an emphasis on**
1244      **the palpalis group.** *Molecular phylogenetics and evolution* 2008, **49**:227-239.

1245   28.   Gooding RH, Krafsur ES: **Tsetse genetics: contributions to biology, systematics, and control of**
1246         **tsetse flies.** *Annual review of entomology* 2005, **50**:101-123.
1247   29.   Petersen FT, Meier R, Kutty SN, Wiegmann BM: **The phylogeny and evolution of host choice in**
1248         **the Hippoboscoidea (Diptera) as reconstructed using four molecular markers.** *Mol Phylogenet*
1249         *Evol* 2007, **45**:111-122.
1250   30.   Meisel RP, Scott JG, Clark AG: **Transcriptome Differences between Alternative Sex Determining**
1251         **Genotypes in the House Fly, Musca domestica.** *Genome Biol Evol* 2015, **7**:2051-2061.
1252   31.   Brelsfoard C, Tsiamis G, Falchetto M, Gomulski L, Telleria E, Alam U, Ntountoumis E, Scolari F,
1253         Swain M, Takac P, et al: ***Wolbachia* symbiont genome sequence and extensive chromosomal**
1254         **insertions present in the host *Glossina morsitans morsitans* genome.** *PloS Neglected Tropical*
1255         *Diseases* 2014, **8**:e2728.
1256   32.   J. MH: **Bearings of the 'Drosophila' work on systematics.** In *The New Systematics*. Edited by J. H.
1257         Oxford: Clarendon; 1940: 185-268
1258   33.   Vicoso B, Bachtrog D: **Numerous transitions of sex chromosomes in Diptera.** *PLoS Biol* 2015,
1259         **13**:e1002078.
1260   34.   Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente
1261         VL, Aguade M, Anderson WW, et al: **Polytene chromosomal maps of 11 Drosophila species: the**
1262         **order of genomic scaffolds inferred from genetic and physical maps.** *Genetics* 2008, **179**:1601-
1263         1655.
1264   35.   Willhoeft U: **Fluorescence in situ hybridization of ribosomal DNA to mitotic chromosomes of**
1265         **tsetse flies (Diptera: Glossinidae: Glossina).** *Chromosome Research* 1997, **5**:262-267.
1266   36.   Papa F, Windbichler N, Waterhouse RM, Cagnetti A, D'Amato R, Persampieri T, Lawniczak MKN,
1267         Nolan T, Papathanos PA: **Rapid evolution of female-biased genes among four species of**
1268         **Anopheles malaria mosquitoes.** *Genome Res* 2017, **27**:1536-1548.
1269   37.   Meisel RP, Connallon T: **The faster-X effect: integrating theory and data.** *Trends Genet* 2013,
1270         **29**:537-544.
1271   38.   Charlesworth B, Coyne JA, Barton NH: **The Relative Rates of Evolution of Sex Chromosomes and**
1272         **Autosomes.** *The American Naturalist* 1987, **130**:113-146.
1273   39.   Mank JE, Vicoso B, Berlin S, Charlesworth B: **Effective population size and the Faster-X effect:**
1274         **empirical results and their interpretation.** *Evolution* 2010, **64**:663-674.
1275   40.   Meisel RP: **Towards a more nuanced understanding of the relationship between sex-biased**
1276         **gene expression and rates of protein-coding sequence evolution.** *Mol Biol Evol* 2011, **28**:1893-
1277         1900.
1278   41.   Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark
1279         AG: **Evolution of protein-coding genes in Drosophila.** *Trends Genet* 2008, **24**:114-123.
1280   42.   Leung W, Shaffer CD, Reed LK, Smith ST, Barshop W, Dirkes W, Dothager M, Lee P, Wong J, Xiong
1281         D, et al: **Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40**
1282         **Million Years of Evolution.** *G3: Genes|Genomes|Genetics* 2015, **5**:719.
1283   43.   Brelsfoard C, Tsiamis G, Falchetto M, Gomulski LM, Telleria E, Alam U, Doudoumis V, Scolari F,
1284         Benoit JB, Swain M, et al: **Presence of extensive Wolbachia symbiont insertions discovered in**
1285         **the genome of its host Glossina morsitans morsitans.** *PLoS Negl Trop Dis* 2014, **8**:e2728.
1286   44.   Doudoumis V, Alam U, Aksoy E, Abd-Alla AM, Tsiamis G, Brelsfoard C, Aksoy S, Bourtzis K:
1287         **Tsetse-Wolbachia symbiosis: comes of age and has great potential for pest and disease**
1288         **control.** *J Invertebr Pathol* 2013, **112 Suppl**:S94-103.
1289   45.   Wu DD, Wang GD, Irwin DM, Zhang YP: **A profound role for the expansion of trypsin-like serine**
1290         **protease family in the evolution of hematophagy in mosquito.** *Mol Biol Evol* 2009, **26**:2333-
1291         2341.

1292    46.    Gorman MJ, Paskewitz SM: **Serine proteases as mediators of mosquito immune responses.**
1293        *Insect Biochem Mol Biol* 2001, **31**:257-262.
1294    47.    LaFlamme BA, Ram KR, Wolfner MF: **The Drosophila melanogaster seminal fluid protease**
1295        **"seminase" regulates proteolytic and post-mating reproductive processes.** *PLoS Genet* 2012,
1296        **8**:e1002435.
1297    48.    Sirot LK, Findlay GD, Sitnik JL, Frasheri D, Avila FW, Wolfner MF: **Molecular characterization and**
1298        **evolution of a gene family encoding both female- and male-specific reproductive proteins in**
1299        **Drosophila.** *Mol Biol Evol* 2014, **31**:1554-1567.
1300    49.    Hamilton JV, Munks RJ, Lehane SM, Lehane MJ: **Association of midgut defensin with a novel**
1301        **serine protease in the blood-sucking fly Stomoxys calcitrans.** *Insect Mol Biol* 2002, **11**:197-205.
1302    50.    Larter NK, Sun JS, Carlson JR: **Organization and function of Drosophila odorant binding**
1303        **proteins.** *Elife* 2016, **5**.
1304    51.    Leal WS: **Odorant reception in insects: roles of receptors, binding proteins, and degrading**
1305        **enzymes.** *Annu Rev Entomol* 2013, **58**:373-391.
1306    52.    Benoit JB, Vigneron A, Broderick NA, Wu Y, Sun JS, Carlson JR, Aksoy S, Weiss BL: **Symbiont-**
1307        **induced odorant binding proteins mediate insect host hematopoiesis.** *Elife* 2017, **6**.
1308    53.    Scolari F, Benoit JB, Michalkova V, Aksoy E, Takac P, Abd-Alla AM, Malacrida AR, Aksoy S,
1309        Attardo GM: **The Spermatophore in** *Glossina morsitans morsitans*: **Insights into Male**
1310        **Contributions to Reproduction.** *Scientific Reports* 2016, **6**:20334.
1311    54.    Doudoumis V, Blow F, Saridaki A, Augustinos A, Dyer NA, Goodhead I, Solano P, Rayaisse JB,
1312        Takac P, Mekonnen S, et al: **Challenging the Wigglesworthia, Sodalis, Wolbachia symbiosis**
1313        **dogma in tsetse flies: Spiroplasma is present in both laboratory and natural populations.** *Sci*
1314        *Rep* 2017, **7**:4699.
1315    55.    Tschopp A, Riedel M, Kropf C, Nentwig W, Klopfstein S: **The evolution of host associations in the**
1316        **parasitic wasp genus Ichneumon (Hymenoptera: Ichneumonidae): convergent adaptations to**
1317        **host pupation sites.** *BMC Evol Biol* 2013, **13**:74.
1318    56.    Pandey RR, Homolka D, Chen KM, Sachidanandam R, Fauvarque MO, Pillai RS: **Recruitment of**
1319        **Armitage and Yb to a transcript triggers its phased processing into primary piRNAs in**
1320        **Drosophila ovaries.** *PLoS Genet* 2017, **13**:e1006956.
1321    57.    Miesen P, Joosten J, van Rij RP: **PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector**
1322        **Mosquitoes.** *PLoS Pathog* 2016, **12**:e1006017.
1323    58.    Avidor-Reiss T, Maer AM, Koundakjian E, Polyanovsky A, Keil T, Subramaniam S, Zuker CS:
1324        **Decoding cilia function: defining specialized genes required for compartmentalized cilia**
1325        **biogenesis.** *Cell* 2004, **117**:527-539.
1326    59.    Ravel S, de Meeus T, Dujardin JP, Zeze DG, Gooding RH, Dusfour I, Sane B, Cuny G, Solano P: **The**
1327        **tsetse fly Glossina palpalis palpalis is composed of several genetically differentiated small**
1328        **populations in the sleeping sickness focus of Bonon, Cote d'Ivoire.** *Infect Genet Evol* 2007,
1329        **7**:116-125.
1330    60.    Starostina E, Xu A, Lin H, Pikielny CW: **A Drosophila protein family implicated in pheromone**
1331        **perception is related to Tay-Sachs GM2-activator protein.** *J Biol Chem* 2009, **284**:585-594.
1332    61.    Baumann AA, Benoit JB, Michalkova V, Mireji PO, Attardo GM, Moulton JK, Wilson TG, Aksoy S:
1333        **Juvenile hormone and insulin suppress lipolysis between periods of lactation during tsetse fly**
1334        **pregnancy.** *Mol Cell Endocrinol* 2013, **372**:30-41.
1335    62.    Buchon N, Silverman N, Cherry S: **Immunity in Drosophila melanogaster--from microbial**
1336        **recognition to whole-organism physiology.** *Nat Rev Immunol* 2014, **14**:796-810.
1337    63.    Dziarski R, Gupta D: **The peptidoglycan recognition proteins (PGRPs).** *Genome Biol* 2006, **7**:232.

64. Vigneron A, Aksoy E, Weiss BL, Bing X, Zhao X, Awuoche EO, O'Neill MB, Wu Y, Attardo GM, Aksoy S: **A fine-tuned vector-parasite dialogue in tsetse's cardia determines peritrophic matrix integrity and trypanosome transmission success.** *PLoS Pathog* 2018, **14**:e1006972.

65. Macleod ET, Maudlin I, Darby AC, Welburn SC: **Antioxidants promote establishment of trypanosome infections in tsetse.** *Parasitology* 2007:1-5.

66. Hao Z, Kasumba I, Lehane MJ, Gibson WC, Kwon J, Aksoy S: **Tsetse immune responses and trypanosome transmission: implications for the development of tsetse-based strategies to reduce trypanosomiasis.** *Proceedings of the National Academy of Sciences, USA* 2001, **98**:12648-12653.

67. Aksoy S, Weiss BL, Attardo GM: **Trypanosome Transmission Dynamics in Tsetse.** *Curr Opin Insect Sci* 2014, **3**:43-49.

68. Hu C, Aksoy S: **Innate immune responses regulate trypanosome parasite infection of the tsetse fly *Glossina morsitans morsitans*.** *Molecular Microbiology* 2006, **60**:1194-1204.

69. Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK, Kasai S, Leichter CA, et al: **Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment.** *Genome Biol* 2014, **15**:466.

70. Valanne S, Myllymaki H, Kallio J, Schmid MR, Kleino A, Murumagi A, Airaksinen L, Kotipelto T, Kaustio M, Ulvila J, et al: **Genome-wide RNA interference in Drosophila cells identifies G protein-coupled receptor kinase 2 as a conserved regulator of NF-kappaB signaling.** *J Immunol* 2010, **184**:6188-6198.

71. Lehane MJ, Aksoy S, Gibson W, Kerhornou A, Berriman M, Hamilton J, Soares MB, Bonaldo MF, Lehane S, Hall N: **Adult midgut expressed sequence tags from the tsetse fly *Glossina morsitans morsitans* and expression analysis of putative immune response genes.** *Genome Biology* 2003, **4**:R63.

72. Aksoy E, Vigneron A, Bing X, Zhao X, O'Neill M, Wu YN, Bangs JD, Weiss BL, Aksoy S: **Mammalian African trypanosome VSG coat enhances tsetse's vector competence.** *Proc Natl Acad Sci U S A* 2016, **113**:6961-6966.

73. Nakamura K, Ida H, Yamaguchi M: **Transcriptional regulation of the Drosophila moira and osa genes by the DREF pathway.** *Nucleic Acids Res* 2008, **36**:3905-3915.

74. Gloria-Soria A, Dunn WA, Yu X, Vigneron A, Lee K-Y, Li M, Weiss BL, Zhao H, Aksoy S, Caccone A: **Uncovering Genomic Regions Associated with <em>Trypanosoma</em> Infections in Wild Populations of the Tsetse Fly <em>Glossina fuscipes</em>.** *G3: Genes|Genomes|Genetics* 2018.

75. Meier R, Kotrba M, Ferrar P: **Ovoviviparity and viviparity in the Diptera.** *Biological Reviews of the Cambridge Philosophical Society* 1999, **74**:199-258.

76. Cmelik SHW, Bursell E, Slack E: **Composition of Gut Contents of Third-Instar Tsetse Larvae (Glossina Morsitans Westwood).** *Comparative Biochemistry and Physiology* 1969, **29**:447-&.

77. Benoit JB, Attardo GM, Michalkova V, Krause TB, Bohova J, Zhang Q, Baumann AA, Mireji PO, Takac P, Denlinger DL, et al: **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* 2014, **10**:e1003874.

78. Benoit JB, Attardo GM, Michalkova V, Takac P, Bohova J, Aksoy S: **Sphingomyelinase activity in mother's milk is essential for juvenile development: a case from lactating tsetse flies.** *Biology of Reproduction* 2012, **87 (17)** 1-10.

79. Guz N, Attardo GM, Wu Y, Aksoy S: **Molecular aspects of transferrin expression in the tsetse fly (*Glossina morsitans morsitans*).** *Journal of Insect Physiology* 2007, **53**:715-723.

1384 80. Scolari F, Benoit JB, Michalkova V, Aksoy E, Takac P, Abd-Alla AM, Malacrida AR, Aksoy S,
1385 Attardo GM: **The Spermatophore in Glossina morsitans morsitans: Insights into Male**
1386 **Contributions to Reproduction.** *Sci Rep* 2016, **6**:20334.
1387 81. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arca B,
1388 Arensburger P, Artemov G, et al: **Mosquito genomics. Highly evolvable malaria vectors: the**
1389 **genomes of 16 Anopheles mosquitoes.** *Science* 2015, **347**:1258522.
1390 82. Wong A, Turchin MC, Wolfner MF, Aquadro CF: **Evidence for positive selection on Drosophila**
1391 **melanogaster seminal fluid protease homologs.** *Mol Biol Evol* 2008, **25**:497-506.
1392 83. Findlay GD, MacCoss MJ, Swanson WJ: **Proteomic discovery of previously unannotated, rapidly**
1393 **evolving seminal fluid genes in Drosophila.** *Genome Res* 2009, **19**:886-896.
1394 84. Macharia R, Mireji P, Murungi E, Murilla G, Christoffels A, Aksoy S, Masiga D: **Genome-Wide**
1395 **Comparative Analysis of Chemosensory Gene Families in Five Tsetse Fly Species.** *PLoS Negl*
1396 *Trop Dis* 2016, **10**:e0004421.
1397 85. Obiero GFO, Mireji PO, Nyanjom SRG, Christoffels A, Robertson HM, Masiga DK: **Odorant and**
1398 **gustatory receptors in the tsetse fly** *Glossina morsitans morsitans*. *PLoS Neglected Tropical*
1399 *Diseases* 2014, **8**:e2663.
1400 86. Liu R, Lehane S, He X, Lehane M, Hertz-Fowler C, Berriman M, Pickett JA, Field LM, Zhou JJ:
1401 **Characterisations of odorant-binding proteins in the tsetse fly** *Glossina morsitans morsitans*.
1402 *Cellular and Molecular Life Sciences* 2010, **67**:919-929.
1403 87. Rio RV, Symula RE, Wang J, Lohs C, Wu YN, Snyder AK, Bjornson RD, Oshima K, Biehl BS, Perna
1404 NT, et al: **Insight into the transmission biology and species-specific functional capabilities of**
1405 **tsetse (Diptera: Glossinidae) obligate symbiont** *Wigglesworthia*. *mBio* 2012, **3**.
1406 88. Caljon G, Van Reet N, De Trez C, Vermeersch M, Perez-Morga D, Van Den Abbeele J: **The Dermis**
1407 **as a Delivery Site of Trypanosoma brucei for Tsetse Flies.** *PLoS Pathog* 2016, **12**:e1005744.
1408 89. Caljon G, Van Den Abbeele J, Stijlemans B, Coosemans M, De Baetselier P, Magez S: **Tsetse fly**
1409 **saliva accelerates the onset of Trypanosoma brucei infection in a mouse model associated**
1410 **with a reduced host inflammatory response.** *Infect Immun* 2006, **74**:6324-6330.
1411 90. Zhao X, Silva TL, Cronin L, Savage AF, O'Neill M, Nerima B, Okedi LM, Aksoy S: **Immunogenicity**
1412 **and Serological Cross-Reactivity of Saliva Proteins among Different Tsetse Species.** *PLoS Negl*
1413 *Trop Dis* 2015, **9**:e0004038.
1414 91. Dama E, Cornelie S, Bienvenu Somda M, Camara M, Kambire R, Courtin F, Jamonneau V,
1415 Demettre E, Seveno M, Bengaly Z, et al: **Identification of Glossina palpalis gambiensis specific**
1416 **salivary antigens: towards the development of a serologic biomarker of human exposure to**
1417 **tsetse flies in West Africa.** *Microbes Infect* 2013, **15**:416-427.
1418 92. Van Den Abbeele J, Caljon G, Dierick JF, Moens L, De Ridder K, Coosemans M: **The** *Glossina*
1419 *morsitans* **tsetse fly saliva: general characteristics and identification of novel salivary proteins.**
1420 *Insect Biochemistry and Molecular Biology* 2007, **37**:1075-1085.
1421 93. Lindh JM, Goswami P, Blackburn RS, Arnold SE, Vale GA, Lehane MJ, Torr SJ: **Optimizing the**
1422 **colour and fabric of targets for the control of the tsetse fly Glossina fuscipes fuscipes.** *PLoS*
1423 *Negl Trop Dis* 2012, **6**:e1661.
1424 94. Green CH, Cosens D: **Spectral responses of the tsetse fly, Glossina morsitans morsitans.** *Journal*
1425 *of Insect Physiology* 1983, **29**:795-800.
1426 95. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP,
1427 Sykes S, et al: **High-quality draft assemblies of mammalian genomes from massively parallel**
1428 **sequence data.** *Proceedings of the National Academy of Sciences* 2011, **108**:1513-1518.
1429 96. Tsai IJ, Otto TD, Berriman M: **Improving draft assemblies by iterative mapping and assembly of**
1430 **short reads to eliminate gaps.** *Genome Biol* 2010, **11**:R41.

97. Weller GL, Foster GG: **Genetic maps of the sheep blowfly Lucilia cuprina: linkage-group correlations with other dipteran genera.** *Genome* 1993, **36**:495-506.
98. Foster TJ, Davis MA, Roberts DE, Takeshita K, Kleckner N: **Genetic organization of transposon Tn10.** *Cell* 1981, **23**:201-213.
99. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV: **OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.** *Nucleic Acids Res* 2017, **45**:D744-D749.
100. Smit A, Hubley R: **RepeatModeler Open-1.0.**; 2008-2015
101. Smit A, Hubley R, Green P: **RepeatMasker Open-3.0.**; 1996-2010.
102. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.
103. Morgulis A, Gertz EM, Schaffer AA, Agarwala R: **A fast and symmetric DUST implementation to mask low-complexity DNA sequences.** *J Comput Biol* 2006, **13**:1028-1040.
104. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18**:188-196.
105. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and phylogenomics.** *Mol Biol Evol* 2017.
106. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**:1061-1067.
107. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14**:988-995.
108. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105-1111.
109. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7**:562-578.
110. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30**:1236-1240.
111. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178-2189.
112. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nat Rev Genet* 2012, **13**:329-342.
113. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al: **Ensembl comparative genomics resources.** *Database* 2016, **2016**:bav096-bav096.
114. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD: **Rfam 12.0: updates to the RNA families database.** *Nucleic Acids Res* 2015, **43**:D130-137.
115. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC bioinformatics* 2009, **10**:421.
116. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics* 2013, **29**:2933-2935.
117. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35**:3100-3108.

| 1478 | 118. | Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**:955-964. |

1478  118.  Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in**
1479        **genomic sequence.** *Nucleic Acids Res* 1997, **25**:955-964.
1480  119.  Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG,
1481        Lewis SE: **Web Apollo: a web-based genomic annotation editing platform.** *Genome Biol* 2013,
1482        **14**:R93.
1483  120.  Giraldo-Calderon GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S,
1484        VectorBase C, Madey G, et al: **VectorBase: an updated bioinformatics resource for invertebrate**
1485        **vectors and other organisms related with human diseases.** *Nucleic Acids Res* 2015, **43**:D707-
1486        713.
1487  121.  Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST**
1488        **assembler for reliable and automated mRNA transcript assembly and SNP detection in**
1489        **sequenced ESTs.** *Genome Res* 2004, **14**:1147-1159.
1490  122.  Sedlazeck FJ, Rescheneder P, von Haeseler A: **NextGenMap: fast and accurate read mapping in**
1491        **highly polymorphic genomes.** *Bioinformatics* 2013, **29**:2790-2791.
1492  123.  Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R,
1493        Brent S, et al: **Ensembl comparative genomics resources.** *Database (Oxford)* 2016, **2016**.
1494  124.  Harris R: **Improved pairwise alignment of genomic DNA.** The Pennsylvania State University,
1495        College of Engineering; 2007.
1496  125.  Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication,**
1497        **deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci U S A* 2003,
1498        **100**:11484-11489.
1499  126.  Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements**
1500        **in performance and usability.** *Mol Biol Evol* 2013, **30**:772-780.
1501  127.  Castresana J: **Selection of conserved blocks from multiple alignments for their use in**
1502        **phylogenetic analysis.** *Mol Biol Evol* 2000, **17**:540-552.
1503  128.  Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large**
1504        **phylogenies.** *Bioinformatics* 2014, **30**:1312-1313.
1505  129.  Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-**
1506        **acid replacement process.** *Mol Biol Evol* 2004, **21**:1095-1109.
1507  130.  Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the**
1508        **animal phylogeny using a site-heterogeneous model.** *BMC Evol Biol* 2007, **7 Suppl 1**:S4.
1509  131.  Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: **The ASTRAL**
1510        **Compendium in 2004.** *Nucleic Acids Res* 2004, **32**:D189-192.
1511  132.  Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics**
1512        **Analysis version 6.0.** *Mol Biol Evol* 2013, **30**:2725-2729.
1513  133.  Kuhn RM, Haussler D, Kent WJ: **The UCSC genome browser and associated tools.** *Brief*
1514        *Bioinform* 2013, **14**:144-161.
1515  134.  Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K,
1516        Clawson H, Green ED, et al: **Aligning multiple genomic sequences with the threaded blockset**
1517        **aligner.** *Genome Res* 2004, **14**:708-715.
1518  135.  Donthu R, Lewin HA, Larkin DM: **SyntenyTracker: a tool for defining homologous synteny**
1519        **blocks using radiation hybrid maps and whole-genome sequence.** *BMC Res Notes* 2009, **2**:148.
1520  136.  Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos:**
1521        **an information aesthetic for comparative genomics.** *Genome Res* 2009, **19**:1639-1645.
1522  137.  Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees:**
1523        **Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res* 2009, **19**:327-335.
1524  138.  De Bie T, Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene**
1525        **family evolution.** *Bioinformatics* 2006, **22**:1269-1271.

1526  139.  Rosendale AJ, Romick-Rosendale LE, Watanabe M, Dunlevy ME, Benoit JB: **Mechanistic**
1527       **underpinnings of dehydration stress in ticks revealed through RNA-seq and metabolomics.**
1528       *Journal of Experimental Biology* 2016, **Submitted.** .
1529  140.  Scolari F, Benoit JB, Michalkova V, Aksoy E, Takac P, Abd-Alla AM, Malacrida AR, Aksoy S,
1530       Attardo GM: **The Spermatophore in Glossina morsitans morsitans: Insights into Male**
1531       **Contributions to Reproduction.** *Scientific reports* 2016, **6**.
1532  141.  Baggerly KA, Deng L, Morris JS, Aldaz CM: **Differential expression in SAGE: accounting for**
1533       **normal between-library variation.** *Bioinformatics* 2003, **19**:1477-1483.
1534  142.  Initiative IGG: **Genome sequence of the tsetse fly        (Glossina morsitans): vector of African**
1535       **trypanosomiasis.** *Science* 2014 **Submitted.** .
1536  143.  Pond SLK, Frost SDW: **Datamonkey: rapid detection of selective pressure on individual sites of**
1537       **codon alignments.** *Bioinformatics* 2005, **21**:2531-2533.
1538  144.  Delport W, Poon AFY, Frost SDW, Pond SLK: **Datamonkey 2010: a suite of phylogenetic analysis**
1539       **tools for evolutionary biology.** *Bioinformatics* 2010, **26**:2455-2457.
1540  145.  Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary**
1541       **genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony**
1542       **methods.** *Mol Biol Evol* 2011, **28**:2731-2739.
1543  146.  Willis JH: **Structural cuticular proteins from arthropods: annotation, nomenclature, and**
1544       **sequence characteristics in the genomics era.** *Insect Biochem Mol Biol* 2010, **40**:189-204.
1545  147.  Ioannidou ZS, Theodoropoulou MC, Papandreou NC, Willis JH, Hamodrakas SJ: **CutProtFam-**
1546       **Pred: detection and classification of putative structural cuticular proteins from sequence**
1547       **alone, based on profile hidden Markov models.** *Insect Biochem Mol Biol* 2014, **52**:51-59.
1548  148.  Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS,
1549       Lambert SA, Mann I, Cook K, et al: **Determination and inference of eukaryotic transcription**
1550       **factor sequence specificity.** *Cell* 2014, **158**:1431-1443.
1551  149.  Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G,
1552       Forslund K, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**:D211-222.
1553  150.  Weirauch MT, Hughes TR: **A catalogue of eukaryotic transcription factor types, their**
1554       **evolutionary origin, and species distribution.** *Subcell Biochem* 2011, **52**:25-73.
1555  151.  Eddy SR: **A new generation of homology search tools based on probabilistic inference.** *Genome*
1556       *Inform* 2009, **23**:205-211.
1557  152.  Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M,
1558       Soding J, et al: **Fast, scalable generation of high-quality protein multiple sequence alignments**
1559       **using Clustal Omega.** *Mol Syst Biol* 2011, **7**:539.

1560
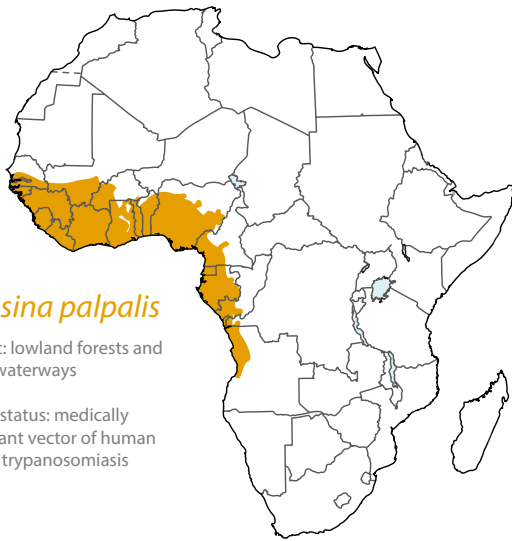
## Glossina palpalis

Habitat: lowland forests and along waterways

Vector status: medically important vector of human african trypanosomiasis

## Glossina fuscipes

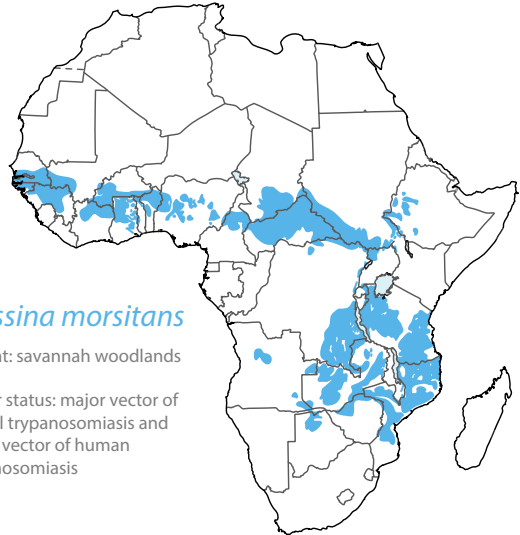Habitat: lowland forests and along waterways

Vector status: medically important vector of human african trypanosomiasis

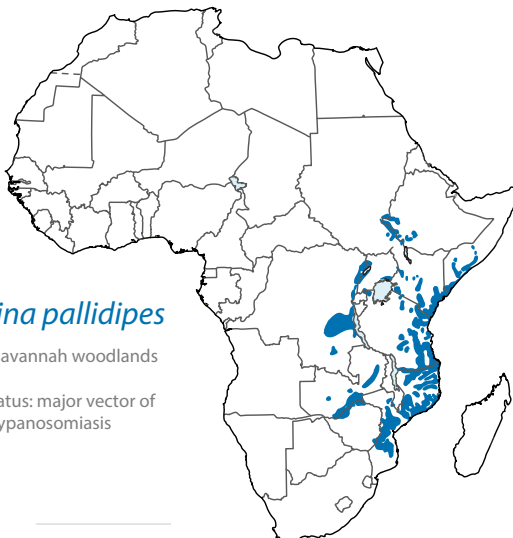## Glossina austeni

Habitat: savannah woodlands

Vector status: economically important vector of animal trypanosomiasis

## Glossina morsitans

Habitat: savannah woodlands

Vector status: major vector of animal trypanosomiasis and minor vector of human trypanosomiasis

## Glossina pallidipes

Habitat: savannah woodlands

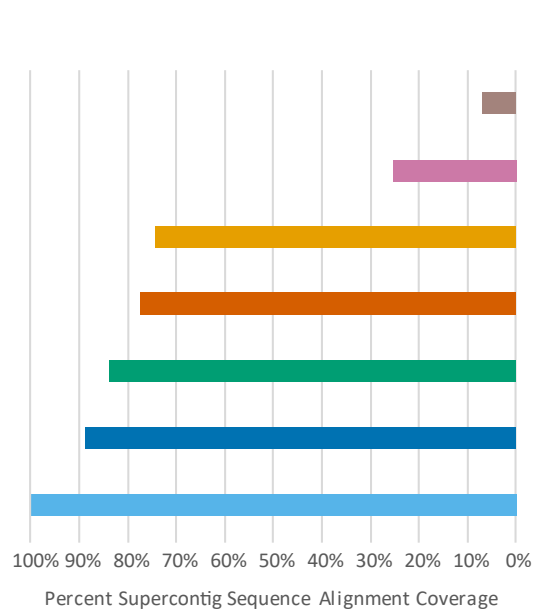Vector status: major vector of animal trypanosomiasis

## Glossina brevipalpis

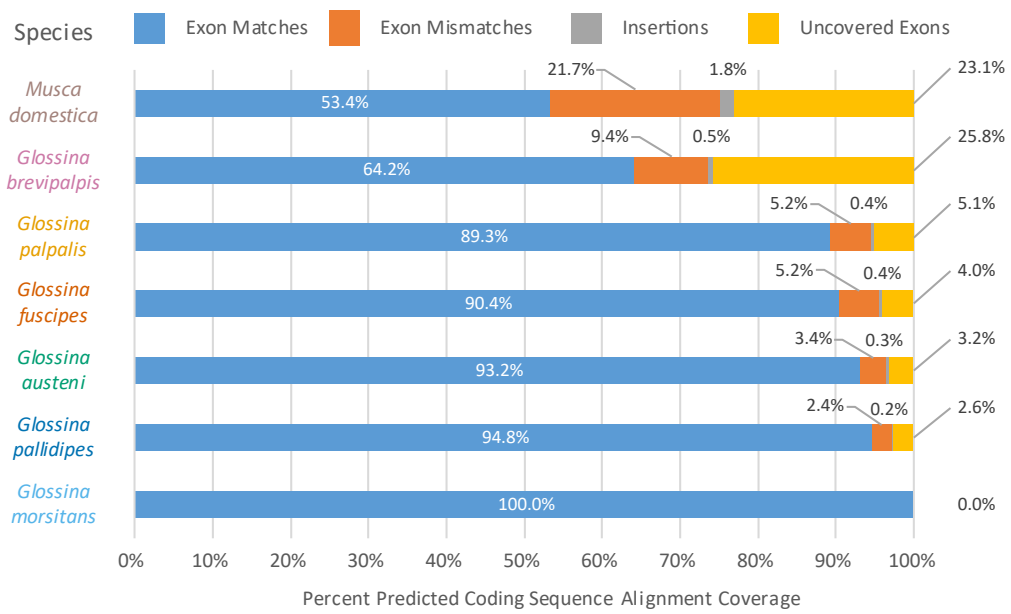Habitat: Forest dwelling

Vector status: Vector of animal trypanosomiasis. Resistant to Brucei type trypanosome infection

Geographic data from: Training manual for tsetse control personnel, editor Pollock, J.N.
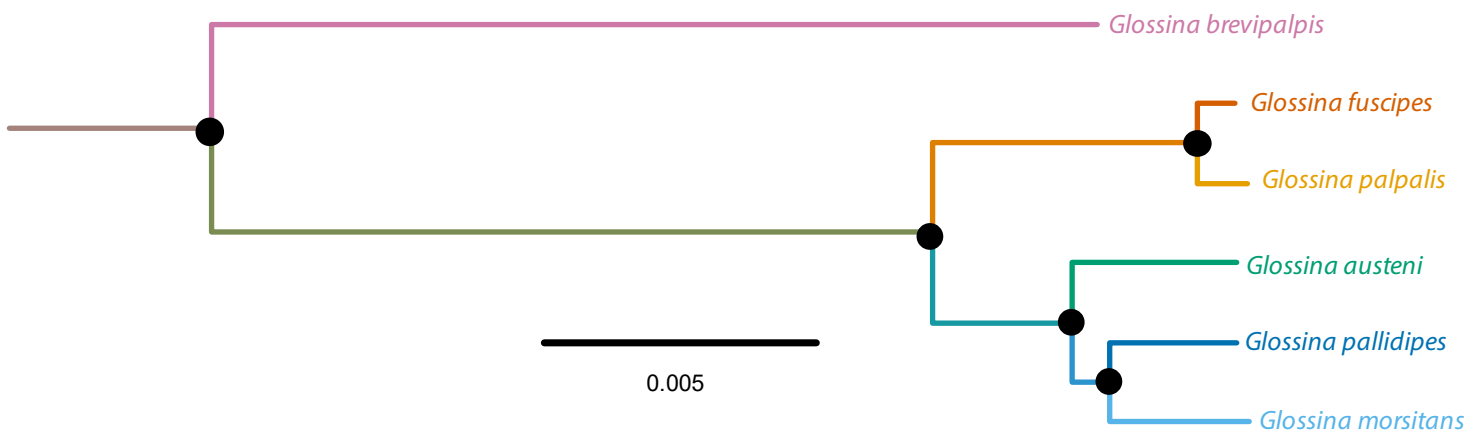Ecological and vectorial data from: The Biology of Blood Sucking Insects, Lehane and Lehane 2005

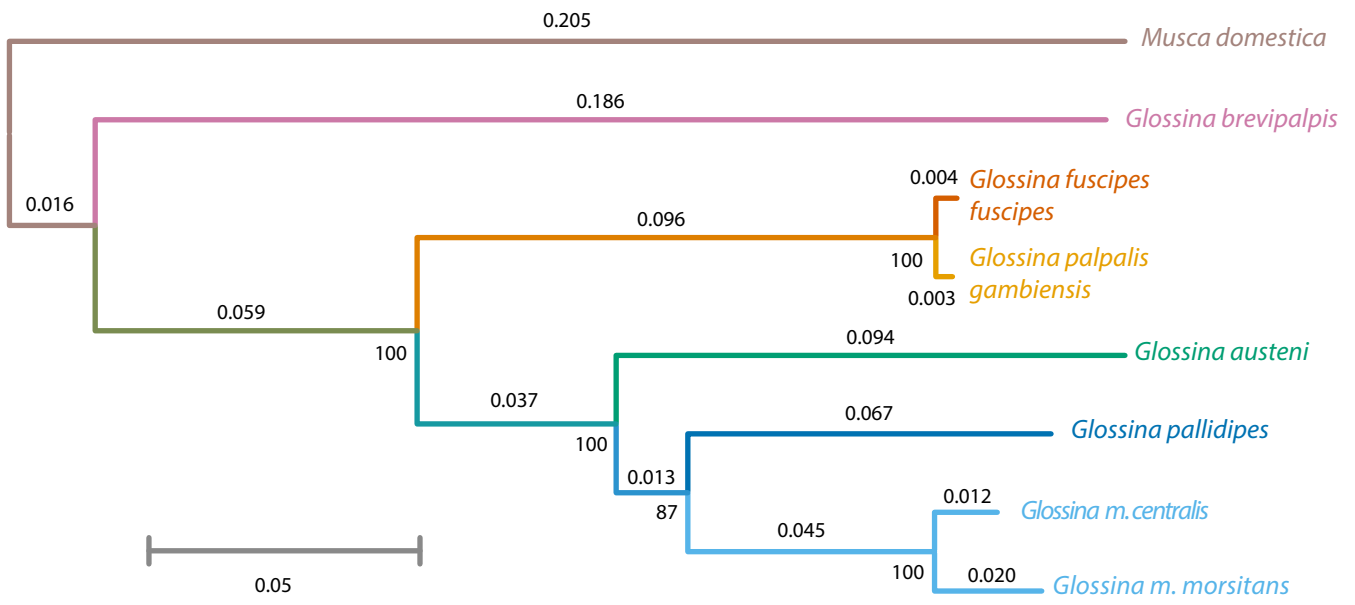A. Genomic Alignment of Glossina Cluster Species Supercontigs (Glossina morsitans reference)

Alignment of Predicted Coding Sequences from the Glossina Cluster Speces (Glossina morsitans reference)

*Drosophila* Muller Elements

X  A

2  B  C

3  D  E

4  F

*Glossina* Muller Elements

X  A  D  F

L1  B  C

L2  E

4 (Muller Element F)

X (Muller Element A)

3R (Muller Element E)

2L (Muller Element B)

3L (Muller Element D)

2R (Muller Element C)

length (Mb)

# of associated genes

*Glossina morsitans*
*Glossina pallidipes*
*Glossina austeni*
*Glossina fuscipes*
*Glossina palpalis*
*Glossina brevipalpis*

Thick bars represent Muller Element length

Thin bars represent number of genes per Muller Element

Each Bar in the Circle Represents 100 kb Chromosomal Alignment Height Represents % Synteny Relative to *Drosophila melanogaster*

# Analysis of Glossina Gene Orthology
## Using the Ensembl Compara Dataset



**39 Groups**
Specific & Universal
to *morsitans* Subgenus
**(Enriched in Proteasome
Activating ATPase Function)**

**2223 Groups**
Specific & Universal
to *Palpalis* Sub-Genus
**(Enriched in Helicase
Activity)**

**59 Groups**
Specific to
*Fusca* Sub-Genus
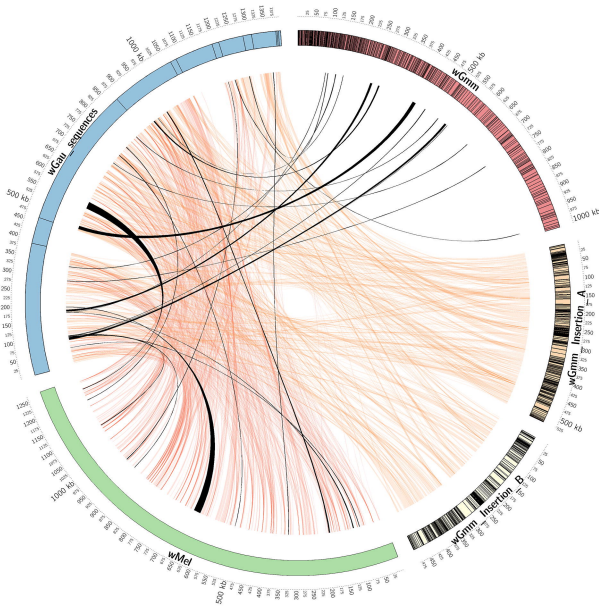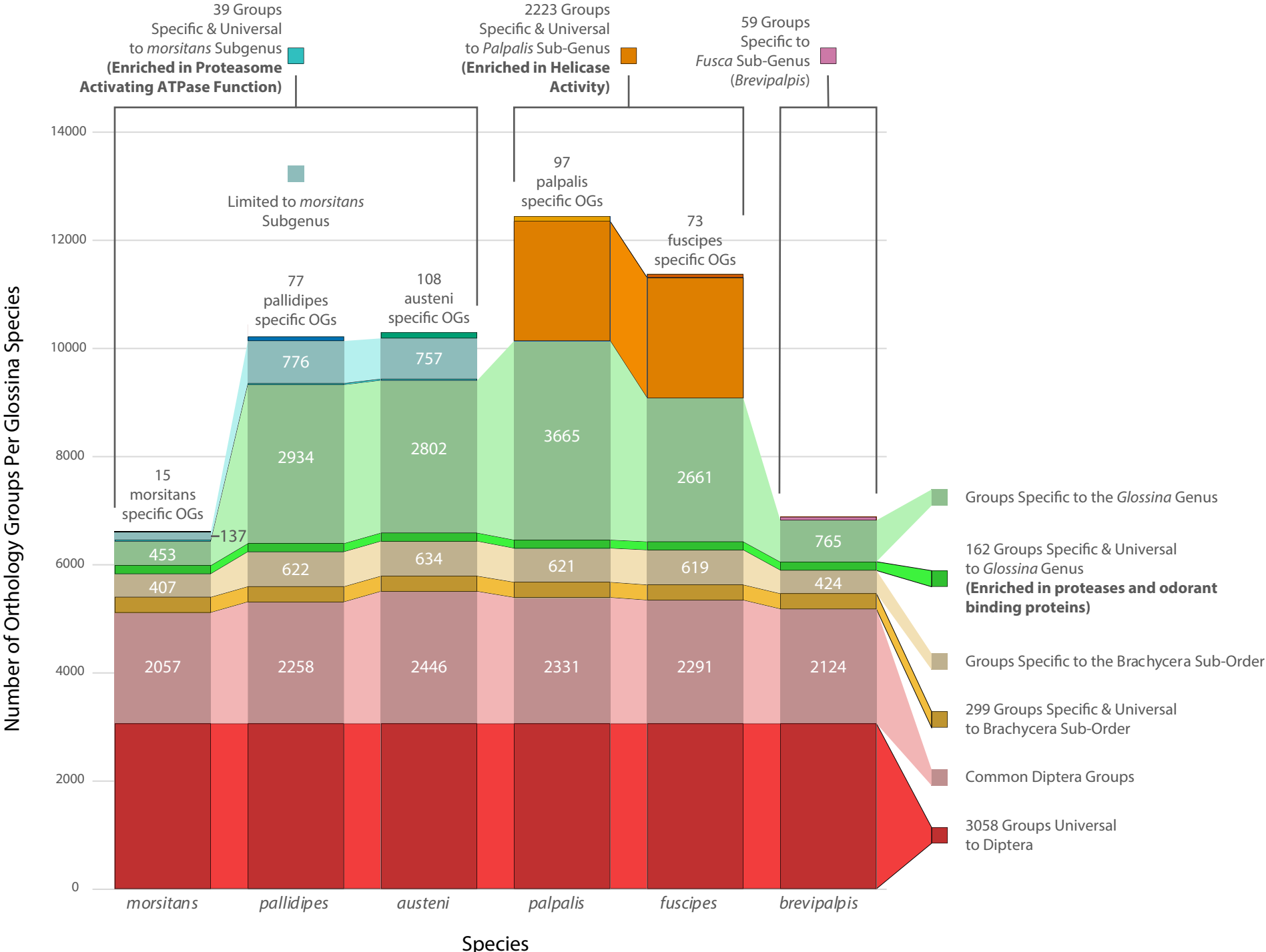(*Brevipalpis*)

Limited to *morsitans*
Subgenus

97
palpalis
specific OGs

73
fuscipes
specific OGs

77
pallidipes
specific OGs

108
austeni
specific OGs

15
morsitans
specific OGs

Number of Orthology Groups Per Glossina Species

| | morsitans | pallidipes | austeni | palpalis | fuscipes | brevipalpis |
|---|---|---|---|---|---|---|
| | | 776 | 757 | 3665 | | |
| | 453 | 2934 | 2802 | | 2661 | 765 |
| | -137 | | | 621 | | |
| | 407 | 622 | 634 | | 619 | 424 |
| | 2057 | 2258 | 2446 | 2331 | 2291 | 2124 |

Groups Specific to the *Glossina* Genus

162 Groups Specific & Universal
to *Glossina* Genus
**(Enriched in proteases and odorant
binding proteins)**

Groups Specific to the Brachycera Sub-Order

299 Groups Specific & Universal
to Brachycera Sub-Order

Common Diptera Groups

3058 Groups Universal
to Diptera

Species

Farnesyl pyrophosphate synthases
VBGT00190000010926

ER degradation-enhancing alpha-mannosidase-like proteins
VBGT00190000009892

Farnesol dehydrogenases
VBGT00840000047886

WD repeat-containing protein 78 like proteins
VBGT00770000031191

Male chemosensory proteins Associated with pheromone detection
VBGT00190000010664

cilia- and flagella-associated protein 44-like proteins
VBGT00190000014373

Groups

Fusca Expanded
Fusca Reduced
Morsitans Group Expanded
Morsitans Group Reduced
Palpalis Group Expanded
Palpalis Group Reduced
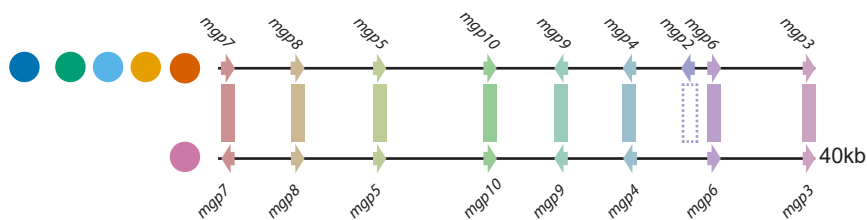
Dim1 (77.4%)

Dim2 (13.2%)

Variable Immune Gene Families Between *Glossina* Species, *Musca domestica* and *Drosophila melanogaster*

**A**

*G. m. morsitans* · *G. pallidipes* · *G. austeni* · *G. f. fuscipes* · *G. p. gambiensis* · *G. brevipalpis*
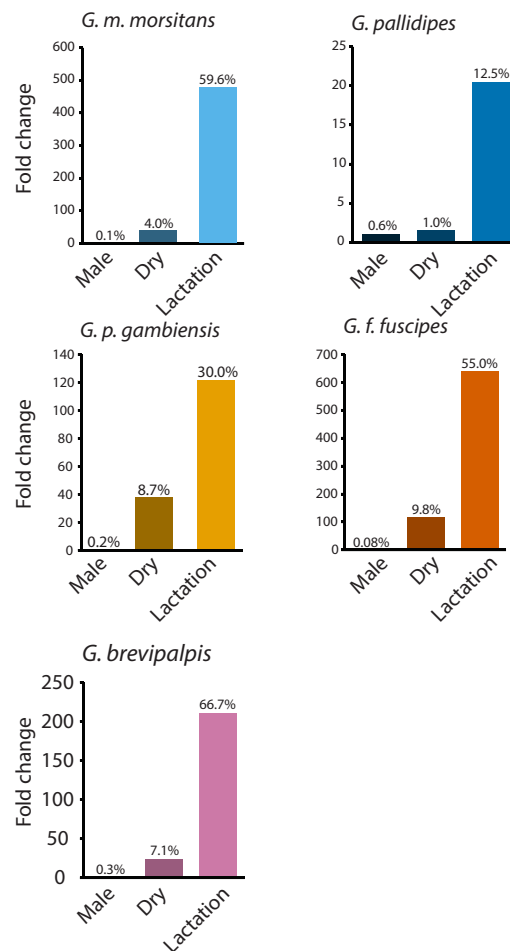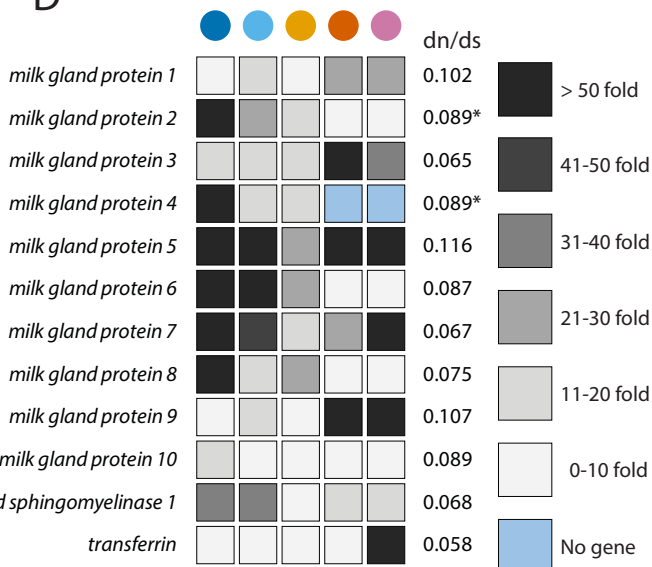
mgp7  mgp8  mgp5  mgp10  mgp9  mgp4  mgp2  mgp6  mgp3

mgp7  mgp8  mgp5  mgp10  mgp9  mgp4  mgp6  mgp3  40kb

**B**

Support: 100-90  89-80  79-70  69-60  59-50

MGP3  MGP2/4  MGP7  MGP9  MGP8  MGP10  MGP5  MGP6

**C**

*G. m. morsitans* — Male 0.1%, Dry 4.0%, Lactation 59.6%

*G. pallidipes* — Male 0.6%, Dry 1.0%, Lactation 12.5%

*G. p. gambiensis* — Male 0.2%, Dry 8.7%, Lactation 30.0%

*G. f. fuscipes* — Male 0.08%, Dry 9.8%, Lactation 55.0%

*G. brevipalpis* — Male 0.3%, Dry 7.1%, Lactation 66.7%

**D**

| | | | | | dn/ds |
|---|---|---|---|---|---|---|
| *milk gland protein 1* | | | | | | 0.102 |
| *milk gland protein 2* | | | | | | 0.089* |
| *milk gland protein 3* | | | | | | 0.065 |
| *milk gland protein 4* | | | | | | 0.089* |
| *milk gland protein 5* | | | | | | 0.116 |
| *milk gland protein 6* | | | | | | 0.087 |
| *milk gland protein 7* | | | | | | 0.067 |
| *milk gland protein 8* | | | | | | 0.075 |
| *milk gland protein 9* | | | | | | 0.107 |
| *milk gland protein 10* | | | | | | 0.089 |
| *acid sphingomyelinase 1* | | | | | | 0.068 |
| *transferrin* | | | | | | 0.058 |

> 50 fold · 41-50 fold · 31-40 fold · 21-30 fold · 11-20 fold · 0-10 fold · No gene

**E**

**Lactation-enriched**

Twelve genes are enriched in three or more species during the lactation phase.

**Dry-enriched**

Five genes are enriched in three or more species in nonlactating females.

A

B

# A

# B