# The genome of the biting midge *Culicoides sonorensis* and gene expression analyses of vector competence for Bluetongue virus

Ramiro Morales-Hojas[1,2,*], Malcolm Hinsley[3], Irina M. Armean[3], Rhiannon Silk[1], Lara E. Harrup[1], Asier Gonzalez-Uriarte[4], Eva Veronesi[1,5], Lahcen Campbell[3], Dana Nayduch[6], Christopher Saski[7], Walter J. Tabachnick[8], Paul Kersey[3], Simon Carpenter[1] and Mark Fife[1]

[1] The Pirbright Institute, Ash Road, Woking, Surrey GU24 0NF, UK

[2] Rothamsted Insect Survey, Rothamsted Research, West Common, Harpenden, Hertfordshire, AL5 2JQ, UK

[3] EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[4] Bioinformatics group, Rothamsted Research, West Common, Harpenden, Hertfordshire, AL5 2JQ, UK

[5] National Centre for Vector Entomology, Institute of Parasitology, Vetsuisse Faculty, University of Zürich, Zürich, Switzerland

[6] USDA-ARS, Center for Grain and Animal Health Research, Arthropod Borne Animal Diseases Research Unit, 1515 College Avenue, Manhattan, KS 66502, USA

[7] Clemson University Genomics Institute, Department of Genetics and Biochemistry, BRC #310, 105 Collins Street, Clemson, SC 29634, USA

[8] Florida Medical Entomology Laboratory, Department of Entomology and Nematology, University of Florida, IFAS, 200 9th St., SE, Vero Beach, FL 32962 USA

* Corresponding author: Ramiro.morales-hojas@rothamsted.ac.uk

**Abstract**

**Background:** The use of the new genomic technologies has led to major advances in control of several arboviruses of medical importance. However, the development of tools and resources available for vectors of non-zoonotic arboviruses remains neglected. Biting midges of the genus *Culicoides* transmit some of the most important arboviruses of wildlife and livestock worldwide, with a global impact on economic productivity, health and welfare. The absence of a suitable reference genome has hindered genomic analyses to date in this important group of vectors. In the present study, the genome of *Culicoides sonorensis,* a vector of bluetongue virus (BTV) in the USA, has been sequenced to be used as reference genome for transcriptomic analyses of vector competence for BTV.

**Results:** The genome has a size of 197.4 Mb and has been assembled in 7,974 scaffolds. Its annotation using the transcriptomic data has identified 15,629 genes. Gene expression analyses of *C. sonorensis* females infected with BTV have identified 165 genes that were differentially expressed between vector competent and refractory females. Two candidate genes, *glutathione S-transferase* (*gst*) and the antiviral helicase *ski2*, were previously recognized as involved in vector competence for BTV in *C. sonorensis* (*gst*) and repressing dsRNA virus propagation (*ski2*).

**Conclusions:** The genome of *C. sonorensis* has enabled preliminary analyses of the genetic profiles of vector competent and refractory individuals. The *C. sonorensis* genome and transcriptomes generated in this study provide suitable platforms for future research on arbovirus transmission. They will be a valuable source of comparative data for other important Dipteran vector families including mosquitoes (Culicidae) and sandflies (Psychodidae).

**Keywords:** Culicoides, biting midges, bluetongue virus, genome, genomics, transcriptomics, vector competence

## Background

Arboviruses (arthropod-borne viruses) are a taxonomically diverse group that include some of the most important emerging and re-emerging pathogens of wildlife, livestock and human beings worldwide [1-3]. Among arbovirus vectors, the majority of recent genomic studies have been carried out on the mosquitoes *Aedes* (*Stegomyia*) *aegypti* (L.) and *Aedes* (*Stegomyia*) *albopictus* (Skuse) due to their involvement in human to human transmission of a wide-range of arboviruses, their relative ease of colonization and recent technical advances in genome sequencing and annotation technologies [4]. While these studies have led to major advances in control of the arboviruses these species transmit and our understanding of what drives susceptibility to infection [5-7], tools and resources for use with many other vector groups remain neglected.

To date, no genomic analyses have been carried out for vectors of non-zoonotic arboviruses that are pathogenic to livestock and wildlife. Among the most important of these vector groups are *Culicoides* biting midges (Diptera: Ceratopogonidae), species of which transmit internationally important arboviruses of ruminants, equines and deer [8-10]. In recent years, unprecedented outbreaks of *Culicoides*-borne arboviruses such as bluetongue virus (BTV) have inflicted huge economic losses on the livestock sector of Europe (e.g. estimate of US$ 1.4 billion in France in 2007) through clinical disease and accompanying restrictions against the movement of livestock imposed to limit virus spread [11, 12]. Globally, BTV outbreaks initiate non-tariff trade barriers restricting the movement of livestock and livestock germplasm, and cause a decrease in ruminant productivity in regions as diverse as India and the USA [13]. Globally, the economic impact of bluetongue has been estimated to US$ 3 billion [12].

*Culicoides* are notoriously difficult to culture under laboratory conditions and their small size of approximately 1.5mm body length renders them far from ideal subjects for transcriptomic and genome manipulation studies [14]. Only one of 14 confirmed vector species of *Culicoides* is currently colonized, *Culicoides sonorensis* Wirth and Jones [8, 15]. *Culicoides sonorensis* colonies have already provided significant insights into the genetic basis of vector competence for BTV [16], a *de novo* transcriptome [17], and have been used to construct a physical map of the *C. sonorensis* genome which consists of four chromosomes [18, 19].

Susceptibility to infection and transmission of arboviruses by *Culicoides* is determined in part by the heritability of barriers to virus dissemination following ingestion of the bloodmeal [13, 20, 21]. Experimental evidence of a midgut infection barrier, a midgut escape barrier and a haemocoel dissemination barrier in *C. sonorensis* (formerly *C. variipennis* or *C. v. sonorensis* [22]) have been defined [23]. There is no evidence to date of salivary gland barriers preventing transmission in any species of *Culicoides* [24], as has been inferred in several species of mosquitoes[20]. Previous studies of the genetic basis for vector competence of *C. sonorensis* for BTV, using maternal inheritance properties, identified a 90 kd protein and used antibodies to this protein to isolate and characterize a cDNA clone encoding a glutathione S-transferase class delta enzyme [25]. Differential gene regulation in response to BTV infection, however, has not been investigated to date.

In this study, we sequence, *de novo* assemble, annotate and explore the first full genome of *C. sonorensis* and then use transcriptomic analyses both to improve gene prediction within the genome build, and to elucidate differential gene regulation associated with

*Culicoides* competence for a BTV serotype 1 strain. The full genome sequence and transcriptome analyses are important resources for further studies of this phylogenetic group which is separated from the other major Dipteran vector families by at least 220 million years [26]. In addition to being a valuable resource for comparative study of vector phylogenomics, the provision of *Culicoides* genomes is of interest as a group that demonstrates several unique features, including a significant capacity for long-distance dispersal by semi-passive flight and the ability to reach huge population density under suitable conditions. In addition, their close association with livestock raises questions concerning both host preference and vector competence for arboviruses. Many of these questions can only be addressed by understanding genetic diversity within populations which will be enhanced by the provision of comparative genomic data, but also in the long term by yielding targets for transgenic modification. The present genome provides a resource to facilitate all these studies.

## Results

### Genome sequencing, construction and annotation

Assembly of the Illumina reads resulted in a genome of 197,417,436 bp assembled in 7,974 scaffolds. The N50 of the genome assembly was 89,502 and the proportion of Ns was of 2.63% (Table 1). Analysis with Repeat Modeller [27] identified a total of 793 repeat regions encompassing ~14% of the genome. When low-complexity regions are included, the total repeat content is 29.7% including 2 Mb of Type II transposons and 5 Mb of Type I transposons (mostly LINE elements). The GC content of the *C. sonorensis* genome is 28% in the complete assembly and 34% in the exons (Table 1). This is low when compared to the closest fully sequenced relative to *C. sonorensis* (*Belgica antarctica*: total GC content = 39%) and a range of other Dipteran genomes (Table 1).

Analyses on the completeness of the assembly identified 98.79% (100% including partial matches) of 248 Core Eukaryotic Genes Mapping Approach (CEGMA) genes and 88.33% (95.88% including partial genes) of 2,675 Benchmarking Universal Single-Copy Orthologs (BUSCO) arthropod genes. These methods define gene sets that are widely conserved in single copy and whose can therefore be used to ascertain the completeness and non-redundancy of a genome sequence. Annotation of the genome resulted in 15,629 protein-coding genes and 21,336 transcripts, which is comparable to the number of protein-coding genes in other species of Diptera (Table 1). The mean gene, intron and exon lengths in *C. sonorensis* are 5,040 bp, 828 bp, and 480 bp, respectively. The average number of introns per gene in *C. sonorensis* is 6.7. A total of 2,991 genes have been estimated to be alternatively spliced and the maximum number of splice variants per gene (19) is comparable to *A. aegypti* and *An. gambiae*.

Homology to genes annotated in other species was detected for 93% of the genes annotated in *C. sonorensis*, with an average of 19 homologous genes and nine orthologues identified per gene; these numbers are comparable to the other species used in the analysis. Over 90% of the *C. sonorensis* genes were linked to their homologues via the ancestral nodes of Arthropoda (6,609 genes), Neoptera (6,361 genes) or Diptera (1,225 genes).

Of 149 high-level categories in the GO ("GO Slim"), 137 were present in *C. sonorensis* as opposed to 140 categories represented in the genome of *An. gambiae* and 144 in the model Dipteran, *D. melanogaster*. GO terms with large numbers of annotated proteins compared to *An. gambiae* and *D. melanogaster* include: ion binding (2,869 genes),

150 oxidoreductase activity (703 genes), kinase activity (367 genes), hydrolase activity
151 acting on glycosyl bonds (112 genes), methyltransferase activity (111 genes) and tRNA
152 metabolic process (101 genes).
153
154 To identify tandemly arrayed genes, we looked for genes that contained common
155 domain architectures (as defined by the set of InterPro-defined domains found within
156 each gene) and were located within 100,000 nucleotides of each other. 1,281 genes
157 were found in such arrays. The longest such array contained 10 copies of a gene
158 containing a chitin-binding domain. The longest array of a protein with this architecture
159 comprises just two copies in *D. melanogaster*, and there are no such arrays in *A. aegypti*
160 or *An. gambiae*, although there are various longer arrays of proteins containing this
161 domain alongside others in these species. The second longest such array contains 9
162 copies of a gene encoding a protein with arrestin and immunoglobulin domains; the
163 homologous proteins are found in 10 tandem copies in *A. aegypti* and *D. melanogaster*,
164 and four tandem copies in *An. gambiae*.
165
166 **Transcriptome sequencing and analysis**
167 Paired-end sequencing of transcriptomes resulted in an average of 50 million reads
168 with a length of 50 bp for each sample sequenced. Phred scores exceeded 34 in all cases.
169 The transcriptomes from the different conditions were merged with Cuffmerge prior to
170 a comparative analysis of their gene expression profiles. The number of genes on the
171 merged transcriptome is 17,263 and the number of transcripts is 35,813. Of the genes
172 present in the transcriptome, 15,630 had already been annotated in the genome. 1,816
173 transcripts, mapping to 1,635 loci, were novel. The proportion of gene loci with a single
174 transcript is 62.7% (10,826), but 93.3% among the genes that had not been annotated

175 on the genome. BLAST searches against NCBI's Arthropod non-redundant (nr) protein
176 database resulted in 92.4% of the isoforms (33,095) and 90% of the genes matching to a
177 database entry (15,268 of the genes had one or more hits).
178
179 **Differential expression of genes between competent and refractory *Culicoides***
180 ***sonorensis***
181 Eighty-five (56%) *C. sonorensis* in replicate one and 53 (37%) in replicate two produced
182 $C_q$ values of ≤ 27 (Additional file 2: Figure S1). Sixty-five (44%) *C. sonorensis* in replicate
183 one and 92 (63%) in replicate two produced $C_q$ values of > 27 or no $C_q$ value. Replicate
184 one had a significantly greater proportion of disseminated infections than replicate two
185 (Fishers exact test; df=1, $P$ = 0.0007). Three and nine samples within those produced
186 from the first and second replicates demonstrated failed RNA extraction from bodies
187 and were discarded. Transcriptome studies were therefore conducted using 147 and
188 136 female *C. sonorensis*, from the two biological replicates of the infection study,
189 respectively.
190
191 Gene expression profiles demonstrated a trend towards upregulation of expression in
192 refractory *C. sonorensis* when compared to competent individuals (Fig. 1a). A total of
193 165 genes demonstrated significant differential expression (DE) (Fig. 1b; Additional file
194 1: Table S1) of which 123 (75%) had been annotated and 42 (25%) represented novel
195 genes (Additional file 1: Table S2). A total of 94 (57%) were upregulated in refractory
196 individuals, while 71 were downregulated (43%) (Additional file 1: Table S1). 120 of
197 the 165 genes showing DE had one or more BLAST hits in the non-redundant (nr)
198 protein database from NCBI (Additional file 1: Table S2) and 88% of these
199 corresponded closely to genes from other members of the suborder Nematocera. Of the

165 genes, 63 were annotated with gene ontology terms. The distribution of these terms has a mean maximum depth of six (Additional file 2: Figure S2). However, enrichment analysis showed no significant results.

*Ski2* (XLOC_006435) and *glutathione S-transferase-1* (CSON011559) showed an increase in expression levels in the females that were resistant to full dissemination, the log2 fold change being -1.10 ($P$ = 0.00035; $q$-value = 0.00381) and -1.07 ($P$ = 0.00055, q-value = 0.00555), respectively (Fig. 2; Additional file 1: Table S1). In contrast, the *Toll* (XLOC_000522) gene and the gene matching *glutathione s transferase* (CSON010973) were expressed at higher levels in the samples that were susceptible to a fully disseminated BTV infection (log2 fold change 1.21, $P$ = 0.00005 and $q$-value = 0.00072) and (log2 fold change 2.07, $P$ = 0.00005 and $q$-value = 0.00072) (Fig. 2; Additional file 1: Table S1).

The direction and degree of change in expression levels of the four genes identified as potentially influencing *C. sonorensis* competence for BTV (*Ski2*, *gst-1*, *gst* and *Toll*) was confirmed using RT-qPCR and the same samples. Housekeeping genes used as reference were *CytB5*, *rpl13* and *rps8*, which showed the highest level of stability of those tested (Additional file 1: Table S3 and Table S4). The direction of up and down regulation of the genes and the log-2 fold change was similar in all cases as that detected in the RNAseq data (Additional file 1: Table S5).

**Molecular evolution of ski2 antiviral helicase and glutathione s transferase-1**

The most significant BLAST hit for the *C. sonorensis ski2* was the gene from *Cx. quinquefasciatus* (XP_001845019.1 = CPIJ003293). There are 37 genes identified as orthologous to *Cx. quinquefasciatus ski2* in 36 species in VectorBase. 35 are 1-to-1

225 orthologues and there are two orthologous genes in *An. maculatus* (Additional file 1:
226 Table S6a). The evolutionary relationship of this protein is consistent with species-level
227 phylogeny (Fig. 3a). In the *gst-1* gene from *C. sonorensis*, the most similar BLAST hit is
228 that of *C. variipennis* (synonym of *C. sonorensis*; AAB94639.1) [25]. The second closest
229 was the *gst-1* copy of *An. gambiae* (AGAP004164), for which 61 orthologues have been
230 identified in 31 other species in VectorBase (Additional file 1: Table S6b). The
231 phylogeny of the *gst-1* protein recovers the two main clades in Diptera, Brachycera and
232 Nematocera, but there is little resolution within these clades (Fig. 3b).
233
234 **Immune response related genes in *C. sonorensis***
235 All *D. melanogaster* genes from FlyBase that had InterPro hits containing Toll, Imd and
236 JAK/STAT associated to their names and all GO terms matching to the Toll, Imd and
237 JAK/STAT pathway terms were used as reference to identify the homologues in *C.*
238 *sonorensis* (Additional file 1: Table S7, and S8; Additional file 2: Figure S3). 36 of 42
239 genes identified in the Toll pathway of *D. melanogaster* were also identified in *C.*
240 *sonorensis*; eight of the 11 genes of the Imd signalling pathway in *D. melanogaster* were
241 identified in the *C. sonorensis* genome and 10 of the 13 *D. melanogaster* JAK/STAT
242 pathway genes were identified. Nineteen genes of the Toll pathway are duplicated in
243 this species, in comparison with 13 in *M. domestica* and *Ae. aegypti*, which are the
244 species with the second largest number of duplicated genes (13) (Additional file 1:
245 Table S9a). In the Imd pathway, seven genes are duplicated in comparison to five genes
246 in *Cx. quinquefasciatus* and *A. aeqypti* (Chi square $P = 0.009$; Additional file 1: Table S9b).
247 In the JAK/STAT pathway the number of duplicated genes is within that observed in
248 other species (Chi square $P = 0.77$; Additional file 1: Table S9c). However, the mean
249 number of copies per gene in these pathways was not significantly different among

250 Nematoceran species for Toll and the Imd pathways (single factor ANOVA with *d.f.* = 1,
251 11 for Toll: $P = 0.4$; Imd: $P = 0.45$), but was significantly different in the JAK/STAT
252 pathway ($P = 0.03$).
253
254 Additional searches for immune related genes in *C. sonorensis* were carried out using
255 BLASTp with *A. aegypti* and *Cx. quinquefasciatus* gene sequences from the ImmunoDB
256 database (http://cegg.unige.ch/Insecta/immunodb) as queries (Additional file 1: Table
257 S10 and S11). Four annotated genes in *C. sonorensis* were identified as potential anti-
258 microbial peptides (AMPs) belonging to the gene subfamilies Attacin, Cecropin,
259 Defensin and Diptericin (Additional file 1: Table S10a); however, the analyses did not
260 identify more than a single gene within each of the AMP subfamilies. Eight potential
261 Toll-like receptors were identified using BLASTp although there were three instances in
262 which the most significant identities were different depending on the species used to
263 provide input data (Additional file 1: Table S11), reflecting the large difference in
264 sequence between the homologous genes in *A. aegypti* and *Cx. quinquefasciatus*. All five
265 genes belonging to the Toll pathway (Additional file 1: Table S11) and all genes of the
266 JAK/STAT signal transduction pathway were identified in *C. sonorensis* (Additional file
267 1: Table S11); while of the Imd pathway, one gene was identified in five of the eight
268 subfamilies (Additional file 1: Table S11).
269
270 The Compara approach reconciles the gene tree (derived from sequence similarity
271 metrics) with the species tree to identify likely orthologous relationships; but requires
272 an initial clustering step (which can lead to the inclusion of sequences into the wrong
273 trees). The only instance in which the two approaches identify a different *C. sonorensis*
274 gene and the orthologue of a gene in the above pathways is that of *Pelle* (which activates

275 the Toll Receptor), for which BLAST identifies CSON012655 as its homologue, and
276 Compara identifies CSON013584 and CSON013585. There were other cases in which
277 BLAST identifies a potential homologue while Compara does not (e.g. TNF-receptor-
278 associated factor 6, Traf6, or tube) and genes Imd and Fadd, neither of the two
279 approaches identify a homologue in *C. sonorensis*.
280
281 **Discussion**
282 This study has produced the first *de novo* genome assembly of the Dipteran family
283 Ceratopogonidae, which includes important vectors of emerging and re-emerging
284 veterinary arboviruses [8, 28]. The closest relative to this group for which a genome of
285 comparable quality is available is the non-biting midge, *Polypedilum vanderplanki*
286 Hinton (Family: Chironomidae) [29], which diverged from the Ceratopogonidae some
287 220 million years ago. This study provides a primary resource for comparative studies
288 with other arbovirus vector species (e.g. mosquitoes [30]). It also will facilitate efforts to
289 produce transgenic *Culicoides* to control *Culicoides*-borne pathogens, although the
290 limitations on colony production for the other epidemiologically relevant species of
291 *Culicoides* remains a major constraining factor [14, 31].
292
293 The genome assembly of *C. sonorensis* is not wholly contiguous, however, its general
294 features are consistent with similarly sized genomes from other species. Thus, features
295 such as the repeat content, gene count and level of alternative splicing are within the
296 ranges observed in *Ae. aegypti*, *An. gambiae* and *D. melanogaster*. Coverage of well-
297 conserved gene families is complete and comparable with other published assemblies
298 (98.79% of CEGMA genes and 88.33% of 2,675 BUSCO genes). The gene set has been
299 annotated with a comparable number of GO Slim terms in *An. gambiae*, with the

molecular functions of ion binding, oxidoreductase activity, kinase activity, hydrolase activity acting on glycosyl bonds, methyltransferase activity all highly represented, as is the tRNA metabolic process. In addition, the genes of the immune-related pathways have been identified and annotated, revealing complete or nearly-complete orthologous pathways. Overall, these observations indicate that the present assembly and annotation of the *C. sonorensis* genome represents a reliable resource to use for some genetic studies. Nevertheless, it is still highly fragmented and the use of long-read sequencing technologies like PacBio would help improving the scaffolding. In addition, the use of this genomic resource by the community will help improve the annotation of the genes. This is the case for many of the genomes of vector species that have been neglected over the years.

A striking feature of the *C. sonorensis* genome is its GC content, which is among the lowest reported for Dipteran species (although close to identical to the non-biting midge *P. vanderplanki* [29] and other non-Dipteran arthropods like the pea aphid, *Acyrthosiphon pisum* Harris [32]). This is not just a feature of the repeated content, but extends into the protein-coding genes (which have a higher GC content than the non-coding regions, but less than in comparable species: Table 1). Recent comparative analyses of codon usage in Diptera and Hymenoptera showed a significant association between codon bias and high GC content in Diptera, but low GC content in Hymenoptera [33]. Thus, the low GC content observed in *C. sonorensis* could reflect a codon bias in comparison to other Diptera species. Functional investigations of GC content have also revealed that it has an impact on genome functioning and species ecology in microbes [34], vertebrates [35] and plants [36]. It is also known that the GC content has an impact in the efficiency of sequencing technologies [37]. Regions with low GC content have less

coverage than more GC balanced ones. This characteristic of the *C. sonorensis* genome should be accounted for in future genomic studies, and appropriate experimental designs that take into account the low GC content should be used [e.g. 38].

Studies investigating gene expression during infection with BTV identified four candidate genes whose regulation was correlated with vector competence, despite only using a single time-point at eight days post-infection and pools of *C. sonorensis* exhibiting different degrees of virus dissemination. Interestingly, the elevated expression of *gst-1* in refractory *Culicoides* observed in this study is consistent with a previously proposed genetic mechanism for *C. sonorensis* vector competence for BTV [13, 15, 24]. These studies identified a locus controlling vector competence in the AA colony, the origin colony of PIR-s-3 colony used in the present study. The former study used a maternally inherited 90kd protein consistent with genetic studies that suggested maternal inheritance of the competence controlling factor identified in unfertilized eggs of refractory females to isolate a clone that upon sequencing was identified as *gst-1* [25]. Though the current study evaluated gene expression after exposure to BTV and the previous study compared genetically selected resistant and susceptible families using individuals from these families who were not exposed to BTV, both independently identified involvement of *gst-1*. Thus, the differential expression of the gene in the current study is confirmation of its involvement in BTV infection in AA colony, although its role requires functional validation.

Future studies of vector competence in *C. sonorensis* must characterize genes that facilitate BTV competence and differentiate these from DE genes that are the consequence of BTV infection. Studies of other *C. sonorensis* colonies from different

350 populations and other populations of other *Culicoides* vectors and non-vector species
351 are needed to determine variation in mechanisms that control competence and factors
352 that be different from AA and PIR-s-3 colonies. Once controlling genes are identified it
353 will be essential to assess the effect of polymorphisms, the various functions of the
354 identified genes in the absence of the pathogen to evaluate factors that influence their
355 frequencies, and compare these processes between different species of *Culicoides* [23].
356
357 The present paper presents the genome of *C. sonorensis* and this has been used in
358 transcriptomic analyses to provide information that will help facilitate the transmission
359 of viruses by vectors and provide new avenues of research to understand vector
360 competence for BTV. We present a list of candidate genes that will be further explored
361 to show their involvement in the transmission of BTV by *C. sonorensis*. Furthermore, the
362 results also provide evidence for the involvement of genes belonging similar functional
363 families in response to virus infection in different species belonging to different families
364 of Diptera. These results show the importance of comparative analyses to interrogate
365 the evolution of host-pathogen interactions, and the *C. sonorensis* genome will facilitate
366 such studies.
367
368 **Conclusions**
369 Here we present the first annotated genome of the biting midge, *C. sonorensis*, a vector
370 of economically important viruses of livestock, such as Bluetongue virus (BTV) and
371 Schmallenberg virus. This genome has been fundamental in gaining information about
372 the genetics of BTV transmission. Thus, gene expression comparison between females
373 susceptible to a full infection by BTV and those females refractory to full BTV infection,
374 identified 165 genes that are candidates to be involved in vector competence in this

375 species. Of these genes, the *gst-1* gene was identified previously to be involved in BTV
376 transmission, while the antiviral helicase *ski2* involved in suppressing the replication of
377 dsRNA viruses. Overall, the publication of the genome and transcriptomes will lead to
378 advances in the control of economically relevant arboviruses of livestock. In addition,
379 this genome fills a phylogenetic gap of more than 200 million years, and will be a useful
380 source of comparative data for studies of vector competence across Diptera.
381

382 **Methods**

383 **Insects used and their infection with bluetongue virus**

384 All *C. sonorensis* originated from the 'AA' colony which was originally produced in 1955
385 at the Kerrville, Texas laboratory of the US Department of Agriculture [39]. Since 1969,
386 this colony has been maintained at The Pirbright Institute without any additional
387 outbreeding [40]. Within this period, two selection bottlenecks of ≤10 individuals were
388 performed to increase susceptibility to BTV serotype 4 and African horse sickness virus
389 (AHSV) serotype 9 infection (Mellor, Pers Comm). Following the second selection in the
390 1990's [41], the colony strain was renamed as PIR-s-3. No attempt was made to further
391 reduce heterozygosity prior to using individuals from the PIR-s-3 line in this study. All
392 *C. sonorensis* used in genomic and transcriptomic studies were 3-4 days old and had not
393 been fed sucrose following emergence from pupae.
394

395 The virus used to infect *C. sonorensis* was a western topotype strain of BTV serotype 1
396 (GIB2007/01) [42]. Virus stock was mixed 1:1 with defibrinated horse blood (TCS
397 Biosciences, Buckingham, UK) with 6.2 $Log_{10}$ $TCID_{50}$/ml being the final infectious dose.
398 All blood-feeding of *C. sonorensis* was conducted using a Hemotek membrane-based
399 system (Discovery Workshops, Accrington, UK).

400 ***Full genome sequencing***: *C. sonorensis* adults were separated by sex under a
401 stereomicroscope and stored in 95% ethanol at room temperature prior to use. A total
402 of 375 male (three pools of 100 and one pool of 75 individuals) and 150 females (one
403 pool of 100 and one pool of 50 individuals) were homogenized twice for 30 seconds in
404 75 µl of phosphate buffered saline using a Tissuelyser™ (Qiagen, UK) at a frequency of
405 25 Hz/s and a 3mm stainless-steel ball bearing (DeJay Distribution Ltd, UK) within 2ml
406 screw-topped tubes. Genomic DNA was extracted from these pools separately and was
407 pooled afterwards according to sex.
408
409 ***Transcriptome sequencing*** *Vector competent vs refractory:* A total of 150 and 145 *C.*
410 *sonorensis* in two replicates were fed on the blood: BTV-1 suspension and survived the
411 extrinsic incubation period of eight days at 25 $^{o}$C. During this period of incubation, 10%
412 (*w*/v) sucrose solution was offered via a cotton wool pad. Each individual was then
413 decapitated using disposable needles and the head and remainder of the carcass of each
414 were stored separately in RNAlater™ (Thermofisher Scientific, UK) at -20 $^{o}$C.
415 Homogenization was carried out on individual heads and remaining carcasses in 100µl
416 Schneider's *Drosophila* media (Gibco™, Thermofisher Scientific, UK), but was otherwise
417 identical as for full genome sequencing. *Additional transcriptomes:* A further 100 *C.*
418 *sonorensis* that had fed on horse blood at three days post-emergence and then stored
419 without access to sucrose for 8 days and 100 *C. sonorensis* that had been offered 10%
420 (*w*/v) sucrose solution from 3-8 days following emergence, were also used to create
421 transcriptomes. In this case individuals were stored whole in RNAlater™ at -20 $^{o}$C and
422 homogenized as for full genome sequencing in a pool of 100 for each of the two
423 treatments.
424

**DNA/RNA extraction**

***Whole genome sequencing (WGS)***: Genomic DNA extraction was conducted using gravity-flow anion-exchange tips (Qiagen's Genomics Tip 20/G) with extraction of DNA up to 150kb in size and a maximum of 20µg of product. A user-developed protocol was followed (https://www.qiagen.com/gb/resources/resourcedetail?id=b45c3cc3-7f2b-4f4a-aa37-21d814ed3730&lang=en), initially mixing the 75µl sample with 1.8 ml of lysis buffer. DNA extracted from the pools of male and female specimens were pooled by sex and the concentration of the resulting samples was evaluated using a Qubit® 2.0 Fluorometer (ThermoFisher Scientific, UK) with the Qubit® DNA BR Assay Kit (ThermoFisher Scientific, UK). The integrity of the genomic DNA was visualized using a 5 µl sample on a 1% (*w/v*) agarose gel.

***BTV Quantification:*** The ProtoScript II Reverse Transcriptase kit (New England Biolabs) was used to quantify BTV in individual decapitated heads. A reaction mixture of 5 µl of homogenized sample, 2 µl of Oligo $d(T)_{18}$ primer and 1 µl of nuclease free water (NFW) was used. The reaction mixture was incubated at 65 $^{o}$C for 5 minutes and then 10 µl of ProtoScript II reaction mix and 2 µl of ProtoScript enzyme mix were added. The cDNA synthesis reaction was carried out at 42 $^{o}$C for one hour with a final incubation at 80 $^{o}$C for five minutes and a reaction with no enzyme was included as a control. The resulting cDNA was used as template in an RT-qPCR for the detection of BTV. The reactions included 5 µl of cDNA, 10 µl of PrimerDesign master mix with SYBR green dye, 20 pmol of each BTV specific primer BTVuni 291-311F 5' GCTTTTGAGGTGTACGTGAAC 3' and BTVuni 381-357R 5' TCTCCCTTGAAACTCTATAATTACG 3' [43] and 4 µl NFW. Controls with no template and with no reverse transcriptase were included. To test that negative reactions were not

450 due to a lack of cDNA in the sample, RT-qPCR for each sample were performed with *C.*
451 *sonorensis* specific primers for the Vacuolar ATPase gene (Vac-ATPase forward 5'
452 GCTGCTGCTGCCATCATTTT 3' and Vac-ATPase reverse 5' CCGGTCGCATCACTGACATA
453 3'). All reactions were performed in duplicate.
454
455 For the purposes of this study, *vector competent* individuals possess disseminated
456 infections that include replication in the head capsule and quantification threshold ($C_q$)
457 values (≤27) that have been demonstrated to allow isolation of BTV-1 [42]. *Refractory*
458 *individuals* refer to those without detectable BTV after a suitable incubation period,
459 those with BTV midgut infections that had not disseminated within the insect and those
460 where the process of dissemination had not been completed (i.e. those with haemocel
461 dissemination barriers, midgut escape and infection barriers) and those still retaining
462 inactivated BTV following the infective bloodmeal [24]. Pools of vector competent and
463 refractory *C. sonorensis* were created from decapitated bodies linked to the infection
464 status of the head as above.
465
466 ***Transcriptome sequencing***: TRIzol® (ThermoFisher Scientific, UK) was added to each
467 homogenized pool of *C. sonorensis* to a final volume of 400 µl. Homogenized samples
468 were incubated at room temperature for 5 minutes after which 80 µl of chloroform
469 were added. Samples were inverted for 15 seconds and left at room temperature for 2-3
470 minutes. The samples were centrifuged at 18000 *x g* for 15 minutes at 4 °C and the clear
471 upper aqueous layer transferred to a new microcentrifuge tube. RNA was precipitated
472 with 200 µl of isopropanol incubated at room temperature for 10 minutes. The RNA was
473 centrifuged at 18000 *x g* for 10 minutes at 4 °C and the supernatant removed. The pellet
474 was washed with 500 µl of 75% ethanol and centrifuged for 5 minutes at 18000 *x g* at 4

475 °C. The supernatant was discarded and the pellet was left to dry for 10-15 minutes
476 before being resuspended in 30 µl NFW at 55 °C for 10 minutes. The integrity of the
477 RNA was analyzed using a Bioanalyser 2100 and the concentration was evaluated using
478 a Qubit® 2.0 Fluorometer (ThermoFisher Scientific, UK) with the Qubit® DNA BR Assay
479 Kit (ThermoFisher Scientific, UK).
480
481 **Genome sequencing and *de novo* assembly**
482 8 – 9 µg of high molecular weight (HMW) genomic DNA were used to sequence the
483 genome of male and female *C. sonorensis* separately. Two paired-end (PE) libraries of
484 insert sizes 200 bp and 400 bp and a mate-pair (MP) library with an average fragment
485 size of 4.4 kb were sequenced with 150 bp (200 bp and 4.4 kb libraries) and 100 bp
486 (400 bp library) paired-end from both female and male pools were run on Illumina
487 HiSeq 2000 and 2500 sequencing systems. Library construction and sequencing was
488 performed at The Earlham Institute (Norwich, UK). Read pair quality was assessed with
489 FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Results
490 suggested that 6 bp should be removed from the start of each read, and in the case of
491 MP data the last 40-50 bp had unexpectedly high levels of k-mer duplication. Both PE
492 and MP reads were filtered to limit their size to 100bp after removal of the first 6 bp.
493 MP data was processed with NextClip [44] to select reads which contained both ends of
494 the original fragment. This created a dataset of reads with variable length, which were
495 subsequently limited to 100bp.
496
497 PE and MP libraries were initially assembled using Velvet [45] with a k-mer length of
498 91. The N50 of the assembly was maximized using an expected coverage of 27, a
499 coverage cut-off of six, and with scaffolding option set to 'no'. This initial assembly was

filtered to remove short (<4kb) and repeated contigs (> 90% sequence ID to a larger contig) using a BLAT and a custom script to count identical bases, considering any overlaps between matches. The assembly was then scaffolded using SSPACE [46] and the 400bp PE library, using three passes of decreasing stringency. Any remaining contigs smaller than 500bp were removed (9913 total). The resulting assembly was then scaffolded a second time using the 4.4kb MP library with SSPACE (Figure 1). This assembly was assessed using FRCbam [47] and Reapr [48], and the output from Reapr was then processed with GapFiller [49]. The final assembly was assessed using FRCbam to confirm that an improvement in the reported error rate had occurred. Data (reads, annotation and assembly) have been deposited in the ENA database under the accession number PRJEB19938. The genome used in this present study is the latest, v1.

**Gene model prediction and annotation**

Prediction and annotation of the genome assembly was conducted using MAKER [50]. The genome annotation was done using the proteomes of nine species of *Anopheles* (*An. albimanus* Wiedemann, *An. arabiensis* Patton, *An. christyi* (Newstead & Carter, 1911), *An. dirus* Peyton & Harrison, *An. epiroticus* Linton & Harbach, *An. funestus* Giles, *An. minimus* Theobald, *An. quadrimaculatus* Say and *An. stephensi* Liston) together with the proteomes of *A. aegypti, Belgica antarctica* Jacobs, *Culex quinquefasciatus* Say, *Drosophila melanogaster* Mg., *Lutzomyia longipalpis* (Lutz & Neiva, 1912) and *Phlebotomus papatasi* (Scopoli). In addition, the transcriptome data from the BTV infection experiments reported in this study and transcriptome data from previous studies on *C. sonorensis* [17, 51] were used. SNAP [52] and AUGUSTUS [53] were used to predict *ab initio* gene models. SNAP was trained using 500 models and AUGUSTUS using the best 1000 models extracted from the initial MAKER output, as recommended in the

525 software documentation. Coverage of conserved proteins was assessed using the
526 CEGMA [54] and BUSCO [55] pipelines. Analysis of Gene Ontology annotation was
527 performed with Blast2GO [56].
528
529 The orthology of the annotated genes was established using the Ensembl Compara Gene
530 Trees pipeline [57]. Genes were functionally annotated using InterProScan, which
531 computationally identifies the presence of domain, motif or family signatures within a
532 protein sequence and infers functional descriptors (taken from the Gene Ontology
533 (GO)).
534
535 **Transcriptome sequencing and analyses**
536 In total, eight transcriptomes were sequenced: two biological replicates each for i) BTV-
537 competent; ii) BTV-refractory; iii) blood-fed; and, iv) sucrose-fed. Library construction
538 and sequencing was performed at Edinburgh Genomics (The University of Edinburgh,
539 Scotland). A total of 2.5 μg of RNA from each sample was used to construct TrueSeq
540 libraries, which were sequenced using 50 base paired-end (PE) reads on an Illumina
541 HiSeq 2500. Each library was run in two lanes.
542
543 The quality of the raw sequence data was analyzed in FastQC v 0.11.2. Reads were
544 aligned to the genome assembly with TopHat v2.0.6 [58] using the gene model
545 annotations generated from the genome as reference. The resulting bam files were
546 assembled with Cufflinks v 2.2.1 [59] and merged into a single transcriptome with
547 Cuffmerge. Quantification of gene and transcript expression and comparison of the
548 expression levels was performed with Cuffquant and Cuffdiff, respectively. The
549 comparative analysis of expression profiles between conditions used grouped replicates

550 of each condition (two biological + two lane runs). Differential expression analyses were
551 performed with the R package CummeRbund [60]. Differentially expressed (DE) genes
552 with a significant change in expression level were identified using an $\alpha$ value of 0.05,
553 which establishes the filtering value of the multiple-testing corrected $q$-values.
554 Identification and functional classification of the DE genes was done using BLAST and
555 InterProScan searches in Blast2Go version 4.0.2. Enrichment and gene set enrichment
556 analyses were performed on the sets of DE genes between the different conditions using
557 the interface provided in Blast2Go to the Fisher's Exact Test implemented in FatiGO [61]
558 and the Gene Set Enrichment Analysis (GSEA) package [62]. To perform the GSEA, all
559 genes were ranked per the logarithmic fold-change, removing the infinite values.
560 Transcriptome data have been deposited in the ENA database under the accession
561 numbers ERR2171964-ERR2171979.
562
563 **Validation of differentially expressed genes**
564 Quantitative reverse transcription PCR (RT-qPCR) was used to validate the change in
565 expression of four selected genes between the vector competent and refractory pools of
566 *C. sonorensis*. These were selected on the basis of previous identification in vector
567 competence studies of *C. sonorensis* or by having a direct functional link with antiviral
568 response. The genes were: a gene sharing 73% identity with the antiviral helicase *ski2*
569 of *Cx. quinquefasciatus (ski2)*; a gene sharing 63% identity with *glutathione S-transferase*
570 *(gst)*, a gene sharing 96% identity with *glutathione S-transferase-1* (*gst-1*) and a gene
571 sharing 49% identity with a gene encoding a *Toll* protein.
572
573 Expression levels were normalized against three reference genes whose stability was
574 ranked using three approaches, BestKeeper [63], geNorm [64], and NormFinder [65]

575 (Additional file 1: Table S3). Gene-specific primers and probes were designed using
576 Primer3Plus [66], to meet the following conditions: amplicon 80-120bp, Tm ~60 °C, GC
577 ~40%, 3' GC clamp, minimal hairpin loops/primer dimers (Additional file 1: Table S4).
578 Primers were compared against the *C. sonorensis* genome using BLAST to ensure
579 specificity to the target region. Amplification efficiency was tested by RT-qPCR with
580 hydrolysis probes by generating a standard curve using triplicate technical replicates of
581 five-fold serial dilutions of RNA template from teneral female *C. sonorensis* specimens.
582 All RNA templates were treated with Turbo DNA-Free™ Kit (ThermoFisher Scientific,
583 UK) prior to RT-qPCR following the manufacture recommended protocol. Starting
584 template total RNA concentration was evaluated using a Qubit® 3.0 Fluorometer
585 (ThermoFisher Scientific, UK) with the Qubit® RNA HS Assay Kit (ThermoFisher
586 Scientific, UK).
587
588 cDNA synthesis and quantitative amplification was done in one reaction using the
589 Superscript® III One-Step RT-qPCR System with Platinum® Taq DNA polymerase
590 (ThermoFisher Scientific, UK). Reactions were performed in a total volume of 25 µl
591 containing 5.5 µl NFW, 12.5 µl 2x reaction mix, 0.5 µl SuperScript III/PlatinumTaq Mix,
592 0.5 µl ROX reference dye (1:10 dilution), 0.5 µl 10 µM forward primer, 0.5 µl 10 µM
593 reverse primer, 1.0 µl hydrolysis probe and 4 µl RNA template. Reaction cycles
594 consisted of the cDNA synthesis at 50 °C for 5 minutes, followed by a denaturation step
595 at 94 °C for 2 minutes and amplification for 40 cycles at 95 °C for 3s each and 56 °C for
596 30s. PCR amplification efficiency of each primer-probe-target combinations was
597 calculated via the linear regression of $C_q$ as the log2 of the relative RNA template
598 concentration using the ggplot2 package [67] in R. Primers which exhibited
599 amplification efficiency between 90-110% with an $R^2$ value of ≥0.98 were considered

600 valid. Relative expression levels for the gene of interest between vector competent and
601 refractory *C. sonorensis* were calculated using the ΔΔCq method [68].
602
603 **Molecular evolution analyses**
604 Orthologues of the *ski2* and *gst-1* genes in other insect species were identified using
605 Blast2Go. The most significant BLAST hit from *A. aegypti*, *Cx. quinquefasciatus* or *An.*
606 *gambiae* was then used to identify the orthologues in other vector species in VectorBase
607 [69]. The protein sequences identified were aligned using the mode expresso of T-
608 Coffee [70]. Poorly aligned regions (positions with score 0-6) were trimmed and the
609 resulting alignments were used to construct a phylogenetic tree. Phylogenetic analyses
610 were performed in MrBayes v3.2 [71] and the models of evolution that best fitted our
611 alignments were identified using ProtTest 3.4.2 [72]. Two independent analyses of two
612 million generations with four chains (one cold and three heated) were conducted. Trees
613 and parameters were sampled every 50$^{th}$ generation and the first 10,000 were
614 discarded. The remaining trees were used to estimate the consensus and Bayesian
615 posterior probabilities of each branch. The s*ki2* from *Saccharomyces cerevisiae* Meyen
616 ex E.C. Hansen and the *gst-1* from *Pediculus humanus* L. were used as outgroups.
617
618 **Declarations**
619 **Ethics approval and consent to participate**
620 Not applicable.
621 **Availability of data and material**
622 All sequencing data, genomic and transcriptomic, generated in this study has been
623 deposited at the European Nucleotide Archive (ENA) database under the accession
624 number PRJEB19938.

625 **Competing interests**

626 The authors declare that they have no competing interests.

627 **Funding**

628 This work was supported by the Biotechnology and Biological Sciences Research

629 Council (BBSRC) grant BB/J016721/1 and the core Strategic Programme Grant to The

630 Pirbright Institute.

631 **Author's contributions**

632 MF, SC and PK conceived the research project. RMH, MH, IMA, RS, LEH, AGU, EV and LC

633 have performed the laboratory experiments and bioinformatic analyses. DN and CS have

634 helped with the annotation of the immune-related genes. RMH, MF and SC have written

635 the manuscript and PK, WT, ND and CS have provided insightful revision. All authors

636 read and approved the final manuscript.

637 **Acknowledgements**

638 We thank Eric Denison and James Barber in the insectary for their supply of reagents,

639 cell lines and insects for this study.

640

641 **References**

642 1. Weaver SC, Reisen WK: Present and future arboviral threats. Antiviral Res. 2010;

643 85:328-345.

644 2. Mlakar J, Korva M, Tul N, Popovic M, Poljsak-Prijatelj M, Mraz J, Kolenc M, Rus

645 KR, Vipotnik TV, Vodusek VF *et al*: Zika Virus Associated with Microcephaly. New

646 Engl J Med. 2016; 374:951-958.

647 3. Pialoux G, Gauzere BA, Jaureguiberry S, Strobel M: Chikungunya, an epidemic

648 arbovirosis. Lancet Infect Dis. 2007; 7:319-327.

649 4. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS,
650 Machol I, Lander ES, Aiden AP *et al*: De novo assembly of the *Aedes aegypti*
651 genome using Hi-C yields chromosome-length scaffolds. Sci. 2017; 356:92-95.
652 5. Alphey L: Genetic Control of Mosquitoes. Annu Rev Entomol. 2014; 59:205-224.
653 6. Franz AWE, Sanchez-Vargas I, Adelman ZN, Blair CD, Beaty BJ, James AA, Olson
654 KE: Engineering RNA interference-based resistance to dengue virus type 2 in
655 genetically modified *Aedes aegypti*. Proc Natl Acad Sci U S A. 2006; 103:4198-
656 4203.
657 7. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi ZY, Megy K,
658 Grabherr M *et al*: Genome sequence of Aedes aegypti, a major arbovirus vector.
659 Sci. 2007; 316:1718-1723.
660 8. Purse BV, Carpenter S, Venter GJ, Bellis G, Mullens BA: Bionomics of Temperate
661 and Tropical *Culicoides* Midges: Knowledge Gaps and Consequences for
662 Transmission of *Culicoides*-Borne Viruses. Annu Rev Entomol. 2015; 60:373-392.
663 9. Carpenter S, Mellor PS, Fall AG, Garros C, Venter GJ: African Horse Sickness Virus:
664 History, Transmission, and Current Status. Annu Rev Entomol. 2017; 62:343-
665 358.
666 10. Savini G, Afonso A, Mellor P, Aradaib I, Yadin H, Sanaa M, Wilson W, Monaco F,
667 Domingo M: Epizootic haemorragic disease. Res Vet Sci. 2011; 91:1-17.
668 11. Carpenter S, Wilson A, Mellor PS: *Culicoides* and the emergence of bluetongue
669 virus in northern Europe. Trends Microbiol. 2009; 17:172-178.
670 12. Rushton J, Lyons N: Economic impact of Bluetongue: a review of the effects on
671 production. Vet Ital. 2015; 51:401-406.
672 13. Tabachnick WJ: *Culicoides variipennis* and bluetongue-virus epidemiology in the
673 United States. Annu Rev Entomol. 1996; 41:23-43.

674 14. Nayduch D, Cohnstaedt LW, Saski C, Lawson D, Kersey P, Fife M, Carpenter S:
675 Studying *Culicoides* vectors of BTV in the post-genomic era: Resources,
676 bottlenecks to progress and future directions. Virus Res. 2014; 182:43-49.
677 15. Jones RH, Foster NM: Oral infection of *Culicoides variipennis* with bluetongue
678 virus - development of susceptible and resistant lines from a colony population. J
679 Med Entomol. 1974; 11:316-323.
680 16. Tabachnick WJ: Genetic control of oral susceptibility to infection of *Culicoides*
681 *variipennis* with bluetongue virus. Am J Trop Med Hyg. 1991; 45:666-671.
682 17. Nayduch D, Lee MB, Saski CA: The Reference Transcriptome of the Adult Female
683 Biting Midge (*Culicoides sonorensis*) and Differential Gene Expression Profiling
684 during Teneral, Blood, and Sucrose Feeding Conditions. PLoS One. 2014; 9.
685 18. Nunamaker RA, Brown SE, Murphy KE, Tabachnick WJ, Knudson DL: Physical
686 mapping of the genome of the biting midge, *Culicoides variipennis*. Mol Biol Cell.
687 1998; 9:452A-452A.
688 19. Nunamaker RA, Brown SE, McHolland LE, Tabachnick WJ, Knudson DL: First-
689 generation physical map of the *Culicoides variipennis* (Diptera : Ceratopogonidae)
690 genome. J Med Entomol. 1999; 36:771-775.
691 20. Mellor PS, Boorman J, Baylis M: *Culicoides* biting midges: Their role as arbovirus
692 vectors. Annu Rev Entomol. 2000; 45:307-340.
693 21. Tabachnick WJ: Nature, Nurture and Evolution of Intra-Species Variation in
694 Mosquito Arbovirus Transmission Competence. Int J Env Res Public Health.
695 2013; 10:249-277.
696 22. Holbrook FR, Tabachnick WJ, Schmidtmann ET, McKinnon CN, Bobian RJ, Grogan
697 WL: Sympatry in the *Culicoides variipennis* complex (Diptera : Ceratopogonidae):
698 a taxonomic reassessment. J Med Entomol. 2000; 37:65-76.

699 23. Carpenter S, Veronesi E, Mullens B, Venter G: Vector competence of Culicoides
700 for arboviruses: three major periods of research, their influence on current
701 studies and future directions. Revue Scientifique Et Technique-Office
702 International Des Epizooties. 2015; 34:97-112.
703 24. Fu H, Leake CJ, Mertens PPC, Mellor PS: The barriers to bluetongue virus
704 infection, dissemination and transmission in the vector, *Culicoides variipennis*
705 (Diptera : Ceratopogonidae). Arch Virol. 1999; 144:747-761.
706 25. Abdallah MA, Pollenz RS, Droog FN, Nunamaker RA, Tabachnick WJ, Murphy KE:
707 Isolation and characterization of a cDNA clone coding for a glutathione S-
708 transferase class delta enzyme from the biting midge Culicoides variipennis
709 sonorensis Wirth and Jones. Biochem Genet. 2000; 38:377-390.
710 26. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, Bertone
711 MA, Cassel BK, Bayless KM, Heimberg AM *et al*: Episodic radiations in the fly tree
712 of life. Proc Natl Acad Sci U S A. 2011; 108:5690-5695.
713 27. Smit AFA, Hubley R: RepeatModeler Open-1.0. 2008-2015
714 <http://www.repeatmasker.org >) 2015.
715 28. Carpenter S, Groschup MH, Garros C, Felippe-Bauer ML, Purse BV: *Culicoides*
716 biting midges, arboviruses and public health in Europe. Antiviral Res. 2013;
717 100:102-113.
718 29. Gusev O, Suetsugu Y, Cornette R, Kawashima T, Logacheva MD, Kondrashov AS,
719 Penin AA, Hatanaka R, Kikuta S, Shimura S *et al*: Comparative genome
720 sequencing reveals genomic signature of extreme desiccation tolerance in the
721 anhydrobiotic midge. Nature Communications. 2014; 5.
722 30. Severson DW, Behura SK: Mosquito Genomics: Progress and Challenges. Annu
723 Rev Entomol. 2012; 57:143-166.

724 31. Mills MK, Nayduch D, Michel K: Inducing RNA interference in the arbovirus
725 vector, Culicoides sonorensis. Insect Mol Biol. 2015; 24:105-114.
726 32. Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D, Tagu D,
727 Wilson ACC, Muzny D, Kovar C *et al*: Genome Sequence of the Pea Aphid
728 *Acyrthosiphon pisum*. Plos Biology. 2010; 8:24.
729 33. Behura SK, Severson DW: Comparative Analysis of Codon Usage Bias and Codon
730 Context Patterns between Dipteran and Hymenopteran Sequenced Genomes.
731 PLoS ONE. 2012; 7.
732 34. Foerstner KU, von Mering C, Hooper SD, Bork P: Environments shape the
733 nucleotide composition of genomes. EMBO Rep. 2005; 6:1208-1213.
734 35. Galtier N, Piganeau G, Mouchiroud D, Duret L: GC-content evolution in
735 mammalian genomes: The biased gene conversion hypothesis. Genetics. 2001;
736 159:907-911.
737 36. Smarda P, Bures P, Horova L, Leitch IJ, Mucina L, Pacini E, Tichy L, Grulich V,
738 Rotreklova O: Ecological and evolutionary significance of genomic GC content
739 diversity in monocots. Proc Natl Acad Sci U S A. 2014; 111:E4096-E4102.
740 37. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC: Effects of GC bias in next-
741 generation-sequencing data on de novo genome assembly. PLoS One. 2013;
742 8:e62856.
743 38. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B,
744 Kwiatkowski DP, Swerdlow HP *et al*: Optimizing Illumina next-generation
745 sequencing library preparation for extremely AT-biased genomes. BMC
746 Genomics. 2012; 13:1.
747 39. Jones RH: The laboratory colonization of *Culicoides variipennis* (Coq). J Econ
748 Entomol. 1957; 50:107-108.

749 40. Boorman J: Maintenance of laboratory colonies of *Culicoides variipennis* (Coq),
750 *Culicoides nubeculosus* (Mg) and *Culicoides riethi* Kieffer (Diptera,
751 Ceratopogonidae). Bull Entomol Res. 1974; 64:371-377.
752 41. Mellor PS, Rawlings P, Baylis M, Wellby MP: Effect of temperature on African
753 horse sickness virus infection in Culicoides. Arch Virol. 1998:155-163.
754 42. Veronesi E, Antony F, Gubbins S, Golding N, Blackwell A, Mertens PPC, Brownlie J,
755 Darpel KE, Mellor PS, Carpenter S: Measurement of the Infection and
756 Dissemination of Bluetongue Virus in Culicoides Biting Midges Using a Semi-
757 Quantitative RT-PCR Assay and Isolation of Infectious Virus. PLoS One. 2013; 8:8.
758 43. Shaw AE, Monaghan P, Alpar HO, Anthony S, Darpel KE, Batten CA, Guercio A,
759 Alimena G, Vitale M, Bankowska K *et al*: Development and initial evaluation of a
760 real-time RT-PCR assay to detect bluetongue virus genome segment 1. J Virol
761 Methods. 2007; 145:115-126.
762 44. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M: NextClip: an analysis and
763 read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics. 2014;
764 30:566-568.
765 45. Zerbino DR, Birney E: Velvet: Algorithms for de novo short read assembly using
766 de Bruijn graphs. Genome Res. 2008; 18:821-829.
767 46. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: Scaffolding pre-
768 assembled contigs using SSPACE. Bioinformatics. 2011; 27:578-579.
769 47. Vezzi F, Narzisi G, Mishra B: Reevaluating Assembly Evaluations with Feature
770 Response Curves: GAGE and Assemblathons. PLoS One. 2012; 7.
771 48. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD: REAPR: a
772 universal tool for genome assembly evaluation. Genome Biol. 2013; 14.

773 49. Boetzer M, Pirovano W: Toward almost closed genomes with GapFiller. Genome
774 Biol. 2012; 13.
775 50. Holt C, Yandell M: MAKER2: an annotation pipeline and genome-database
776 management tool for second-generation genome projects. BMC Bioinformatics.
777 2011; 12.
778 51. Nayduch D, Lee MB, Saski CA: Gene discovery and differential expression
779 analysis of humoral immune response elements in female *Culicoides sonorensis*
780 (Diptera: Ceratopogonidae). Parasites & Vectors. 2014; 7.
781 52. Stanke M, Schoffmann O, Morgenstern B, Waack S: Gene prediction in eukaryotes
782 with a generalized hidden Markov model that uses hints from external sources.
783 BMC Bioinformatics. 2006; 7.
784 53. Stanke M, Steinkamp R, Waack S, Morgenstern B: AUGUSTUS: a web server for
785 gene finding in eukaryotes. Nucleic Acids Res. 2004; 32:W309-W312.
786 54. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes
787 in eukaryotic genornes. Bioinformatics. 2007; 23:1061-1067.
788 55. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: BUSCO:
789 assessing genome assembly and annotation completeness with single-copy
790 orthologs. Bioinformatics. 2015; 31:3210-3212.
791 56. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M,
792 Talon M, Dopazo J, Conesa A: High-throughput functional annotation and data
793 mining with the Blast2GO suite. Nucleic Acids Res. 2008; 36:3420-3435.
794 57. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle
795 SMJ, Amode R, Brent S *et al*: Ensembl comparative genomics resources (vol 2016,
796 bav096, 2016). Database-the Journal of Biological Databases and Curation. 2016.

797 58. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate
798 alignment of transcriptomes in the presence of insertions, deletions and gene
799 fusions. Genome Biol. 2013; 14.
800 59. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL,
801 Rinn JL, Pachter L: Differential gene and transcript expression analysis of RNA-
802 seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562-578.
803 60. L G, C T, D K: cummeRbund: Analysis, exploration, manipulation, and
804 visualization of Cufflinks high-throughput sequencing data. R package version
805 2180. 2013.
806 61. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: a web tool for finding significant
807 associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004;
808 20:578-580.
809 62. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA,
810 Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: Gene set enrichment
811 analysis: A knowledge-based approach for interpreting genome-wide expression
812 profiles. Proc Natl Acad Sci U S A. 2005; 102:15545-15550.
813 63. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP: Determination of stable
814 housekeeping genes, differentially regulated target genes and sample integrity:
815 BestKeeper - Excel-based tool using pair-wise correlations. BiotL. 2004; 26:509-
816 515.
817 64. Hoerndli FJ, Toigo M, Schild A, Gotz J, Day PJ: Reference genes identified in SH-
818 SY5Y cells using custom-made gene arrays with validation by quantitative
819 polymerase chain reaction. AnBio. 2004; 335:30-41.
820 65. Ohl F, Jung M, Xu CL, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G,
821 Loening SA, Radonic A *et al*: Gene expression studies in prostate cancer tissue:

822 which reference gene should be selected for normalization? Journal of Molecular
823 Medicine-Jmm. 2005; 83:1014-1024.
824 66. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM:
825 Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007;
826 35:W71-W74.
827 67. Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. New York, USA;
828 2009.
829 68. Pfaffl MW: A new mathematical model for relative quantification in real-time RT-
830 PCR. Nucleic Acids Res. 2001; 29.
831 69. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R,
832 Campbell KS, Christophides GK, Christley S, Dialynas E *et al*: VectorBase: a data
833 resource for invertebrate vector genomics. Nucleic Acids Res. 2009; 37:D583-
834 D587.
835 70. Notredame C, Higgins DG, Heringa J: T-Coffee: A novel method for fast and
836 accurate multiple sequence alignment. J Mol Biol. 2000; 302:205-217.
837 71. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B,
838 Liu L, Suchard MA, Huelsenbeck JP: MrBayes 3.2: Efficient Bayesian Phylogenetic
839 Inference and Model Choice Across a Large Model Space. Syst Biol. 2012; 61:539-
840 542.
841 72. Darriba D, Taboada GL, Doallo R, Posada D: ProtTest 3: fast selection of best-fit
842 models of protein evolution. Bioinformatics. 2011; 27:1164-1165.
843

844 **Figure Legends**

845 **Figure 1.** Differential gene expression analyses between females susceptible to full BTV
846 infection and refractory females. **a** Scatterplot sof the pairwise comparison of the gene
847 expression levels between the two phenotypes. The average gene expression across all
848 genes is displayed by the blue line. **b** Volcano plot displaying fold changes in expression
849 of the 165 differentially expressed genes between *Culicoides sonorensis* that are
850 susceptible to infection and those that are refractory to full infection with BTV.

851 **Figure 2.** Expression bar plots for candidate genes identified as differentially expressed
852 at significant level between vector competent (orange bars) and refractory (blue bars)
853 samples to full BTV infection. *Toll* corresponds to the *Toll*-like gene, *ski2* is the antiviral
854 helicase *ski2*, *gst* is *glutathione s transferase*, and *gst-1* is the *glutathione s transferase-1*.

855 **Figure 3.** Bayesian phylogenies of *ski2* (a) and *gst-1* (b) genes. Tip labels correspond to
856 the Ensembl Identifier followed by the species name (in the case of *C. sonorensis*, the
857 number that precedes correspond to the RNA transcript); colouring corresponds to the
858 Nematocera in red and Brachycera in green. Labels on branches show the posterior
859 probability values.

860

861 **Supporting Information Legends**

862 **Supplementary Material Tables**

863 **Table S1.** Differentially expressed genes between vector competent and vector
864 refractory *C. sonorensis* females identified using CummeRbund. The table shows the
865 fpkm values for each phenotype, the log2 fold change of expression levels and the
866 significance level of the change. Genes that are not annotated in the *C. sonorensis*
867 genome are denoted with the name XLOC_XXXXXX. Genes belonging to an enriched GO
868 term are shown in bold.

869 **Table S2.** BLAST results for the differentially expressed genes between vector
870 competent and refractory transcriptomes showing the top hit as determined by BLAST.
871 **Table S3.** Ranking of the expression stability of the reference genes tested for use in the
872 quantitative reverse transcription PCR (RT-qPCR). Three different tests were used to
873 rank the expression stability of the genes, BestKeeper, gNorm and NormFinder. *r*-
874 correlation coefficient, Cq – quantification cycle, SE – standard error, SD – standard
875 deviation.
876 **Table S4.** Primers forward (F), reverse (R) and hydrolysis probe (P) sequences used for
877 RT-qPCR analysis of the expression levels of *ski2*, *gst*, *gst-1* and *Toll*.
878 **Table S5.** Difference of expression levels in four genes (*Toll*, *gst*, *gst-1* and *ski2*) between
879 vector-competent and vector-refractory females assessed by RNAseq and RT-qPCR. *
880 denote a difference in expression between refractory and competent females with a $P <$
881 0.05 (RT-qPCR significance estimated using a Welch Two Sample t-test with d.f. = 1).
882 **Table S6.** Orthology analysis of two of genes *Ski2* **(a)** and *gst-1* **(b)**.
883 **Table S7.** Immune pathways genes identified in *C. sonorensis* using the Ensembl
884 Compara pipeline and the *D. melanogaster* genes as reference. **a)** Toll pathway; **b)** Imd
885 pathway; **c)** Jak/Stat pathway.
886 **Table S8.** Homologue immune genes in several invertebrate species identified using *D.*
887 *melanogaster* genes as reference all with InterPro and GO terms associated with Toll
888 **(a)**, Imd **(b)** and Jak/Stat **(c)**, in the Ensembl Compara pipeline.
889 **Table S9.** Number of copies of the different genes identified in several Diptera species
890 as homologues to the *D. melanogaster* genes containing InterPro and GO terms related
891 to Toll **(a)**, Imd **(b)** and Jak/Stat **(c)** pathways. Ae - *A. aegypti;* Ag - *An. gambiae;* Ba - *B.*
892 *antarctica;* Cq - *Cx. quinquefasciatus;* Cs - *C. sonorensis;* Dm - *D. melanogaster;* Gm - *G.*

*morsitans;* Ll - *L. longipalpis;* Lc - *L. cuprina;* Md - *M. domestica;* Pp - *P. papatasi;* Sc - *S. calcitrans.*

**Table S10.** Immune related genes identified in *C. sonorensis* using BLASTp top hits with the gene copies from *A. aegypti* and *Cx. quinquefasciatus* available in the ImmunoDB database (http://cegg.unige.ch/Insecta/immunodb) as query. a) Anti-Microbial Peptides; b) Toll receptors; c) Toll path; d) jak/Stat path; e). Imd path. * - Top BLASTp hits different found when *A. aegypti* and *Cx. quinquefasciatus* were used as query.

**Table S11.** Details of the BLASTp results using *A. aegypti* (a) and *Cx. quinquefasciatus* (b) as query to identify the *C. sonorensis* genes of the immune pathways.

**Supplementary Material Figures**

**Figure S1.** Distribution of $C_q$ values from RT-qPCR used to detect BTV-1 virus infection in the two virus-feeding experiments (in blue and orange) of *C. sonorensis*. The vertical, dashed line corresponds to a $C_q$ value of 27 used to separate vector competent from refractory females.

**Figure S2.** Classification and functional distribution of the genes differentially expressed between vector competent and refractory *C. sonorensis* per the Gene Ontology level 6. Blue: Molecular Function; Green: Cellular Component; Pink: Biological Process.

**Figure S3.** Number of genes in the Toll (a), Imd (b) and Jak/Stat (c) pathways that have been duplicated or lost in different species of Diptera as identified using the Ensembl Compara pipeline.

**Table 1.** A comparison of genome characteristics of the *C. sonorensis* genome with other selected Diptera species.

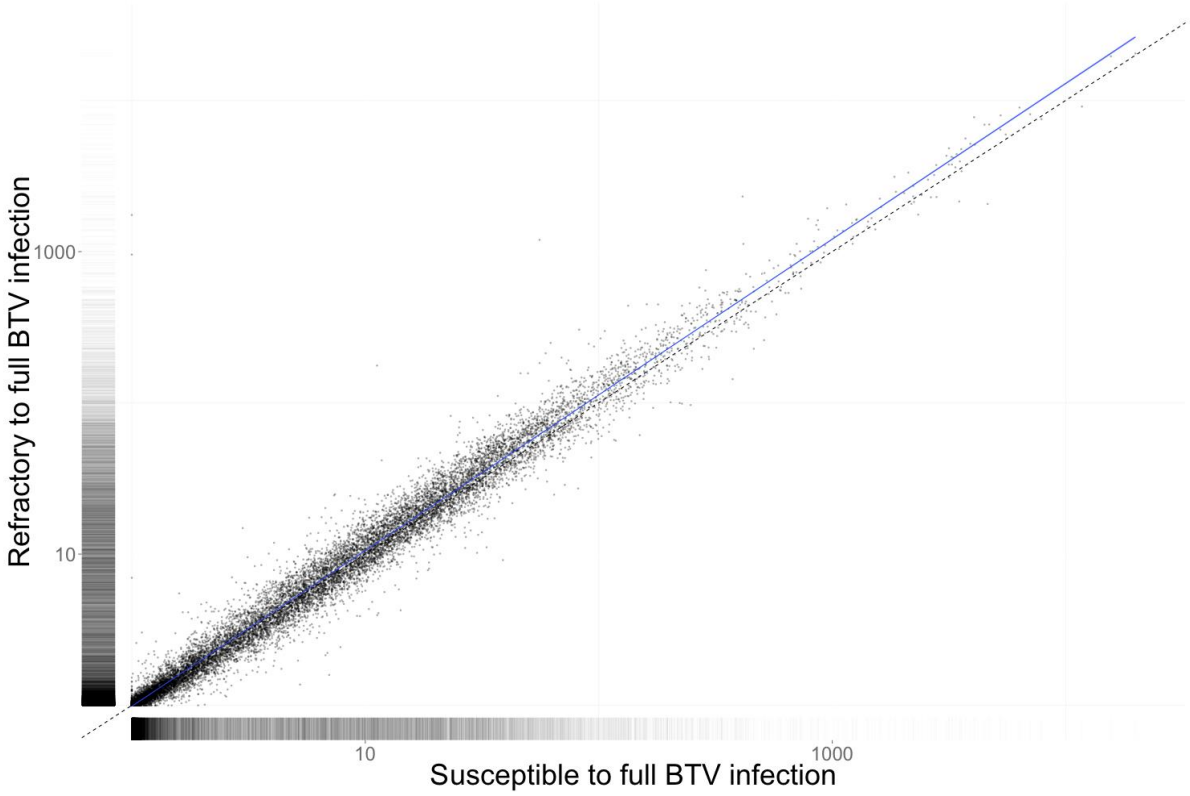
| Parameter | *Culicoides sonorensis* | *Aedes aegypti* L3 | *Aedes aegypti* L5 | *Anopheles gambiae* AGAM P4 | *Drosophila melanogaster* BDGP6 | *Belgica antarctica* GCA_000775305.1 |
|---|---|---|---|---|---|---|
| **Genome size (Mb)** | 197.4 | 1,384.1 | 1,362.0 | 278.3 | 143.7 | 89.7 |
| **Number of contigs** | 15,810 | 36,205 | 4,901 | 16,825 | 2,442 | 22,492 |
| **Contig N50** | 30,774 | 82,500 | 10,861,303 | 85,555 | 19,488,218 | 13,551 |
| **Contig N90** | 5,567 | 15.283 | 74,389 | 5,600 | 666,663 | 1,865 |
| **Number of scaffolds** | 7,974 | 4,757 | 4,616 | 10,523 | 1,870 | 4,997 |
| **Scaffold N50** | 89,077 | 1,547,048 | 409,777,670 | 49,364,325 | 25,286,936 | 98,164 |
| **Scaffold N90** | 14,215 | 324,062 | 83,687 | 2,684 | 23,513,712 | 18,418 |
| **Ns (%)** | 2.8 | 5.3 | 0.0 | 7.6 | 0.8 | 0.9 |
| **Maximum contig length** | 552,397 | 685,587 | 71,953,859 | 808,130 | 27,905,041 | 131,402 |
| **Repeats + low complexity regions** | 29.7% | 73% | 78% | 25% | 23% | 4% |

| | | | | | | |
|---|---|---|---|---|---|---|
| **GC content** | 28.4% | 38.3% | 38.2% | 44.3% | 42.0% | 38.9% |
| **GC exons** | 34.3% | 46.8% | 46.2% | 54.6% | 49.1% | 47.1% |
| **Protein-coding genes** | 15,629 | 15,796 | 14,626 | 13,008 | 13,918 | 13,510 |
| **Mean protein-coding gene length** | 5,040 bp | 18,126 bp | 46,920 bp | 6,442 bp | 6,960 bp | 2,555 bp |
| **Mean exon length (protein coding genes)** | 480 bp | 460 bp | 494 bp | 438 bp | 538 bp | 325 bp |
| **Mean intron length (protein-coding genes)** | 632 bp | 23,265 bp | 10,456 bp | 1,008 bp | 1,147 bp | 213 bp |
| **Mean introns per transcript (in protein-coding genes)** | 6.7 | 5.1 | 7.9 | 5.5 | 6.9 | 5.3 |
| **Maximum splice isoforms/protein-coding gene** | 19 | 20 | 50 | 20 | 75 | n.a. |

Contigs refer to the contiguous sequences assembled without Ns; scaffolds refer to the final top level sequence assembly used for annotation (they contain Ns).

**Fig. 1**

**a**


Refractory to full BTV infection
1000
10
10
1000
Susceptible to full BTV infection

**b**

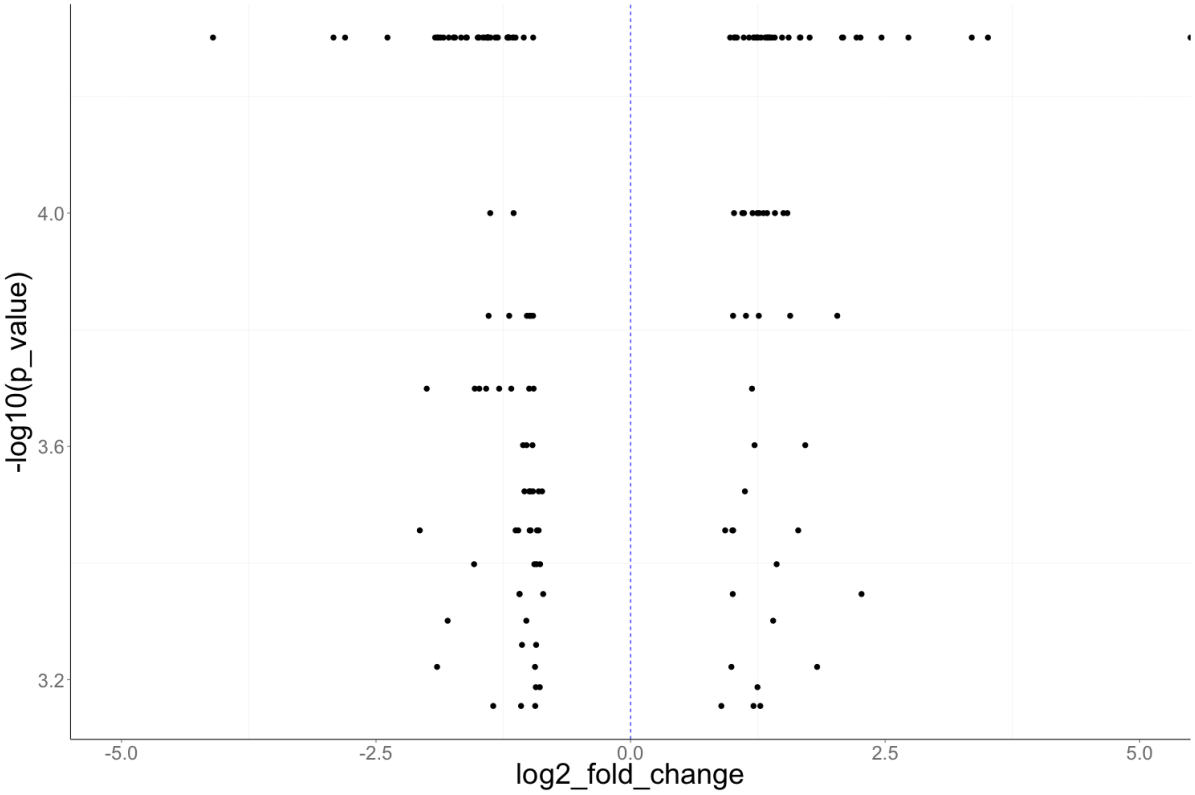

-log10(p_value)
4.0
3.6
3.2
-5.0
-2.5
0.0
2.5
5.0
log2_fold_change

**Fig. 2**

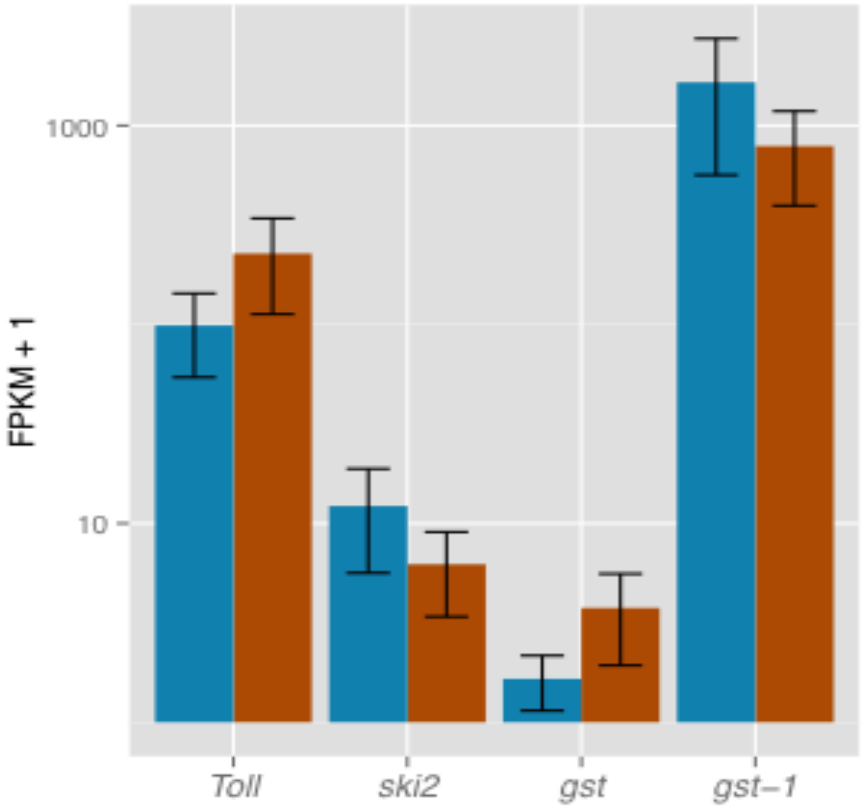
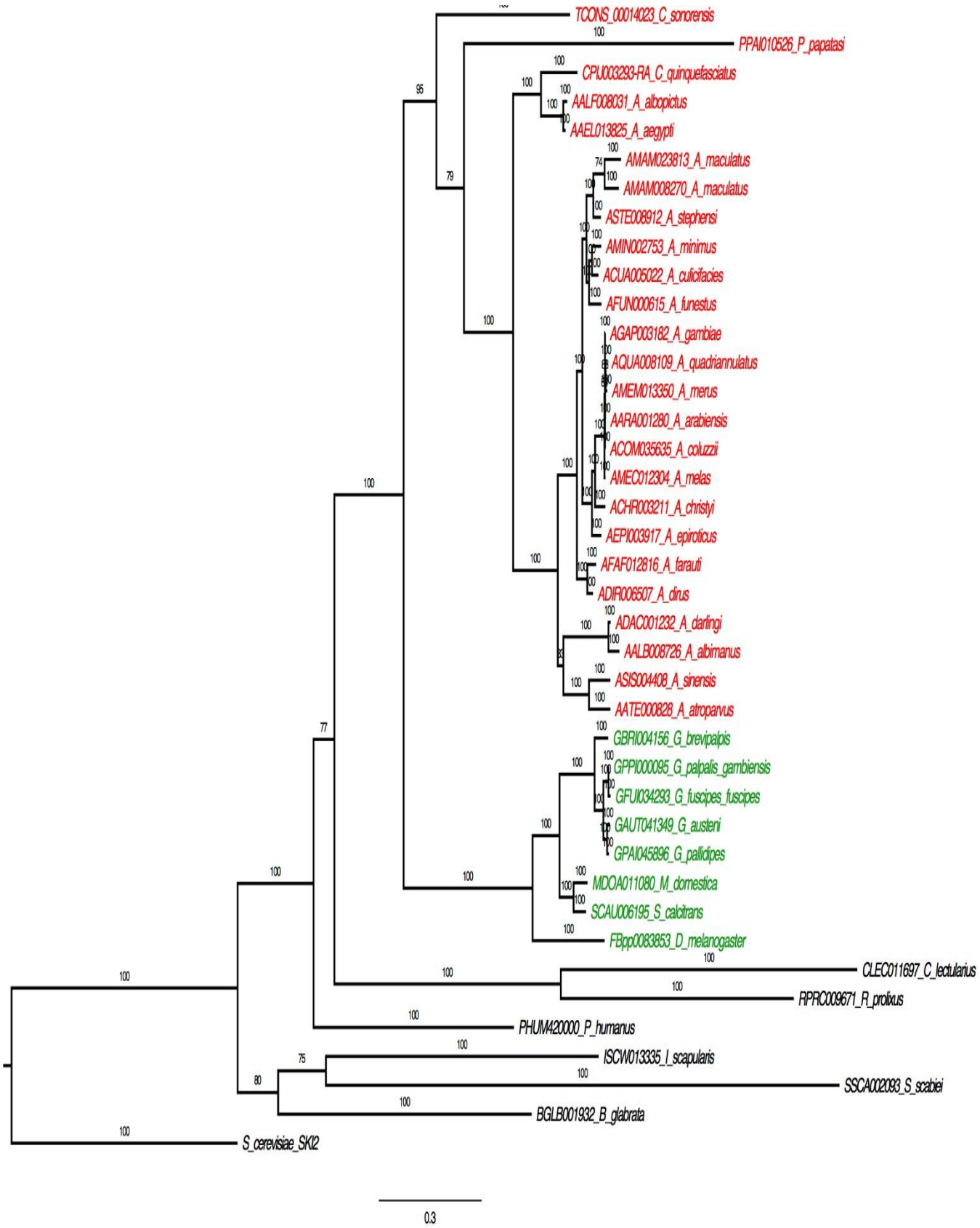
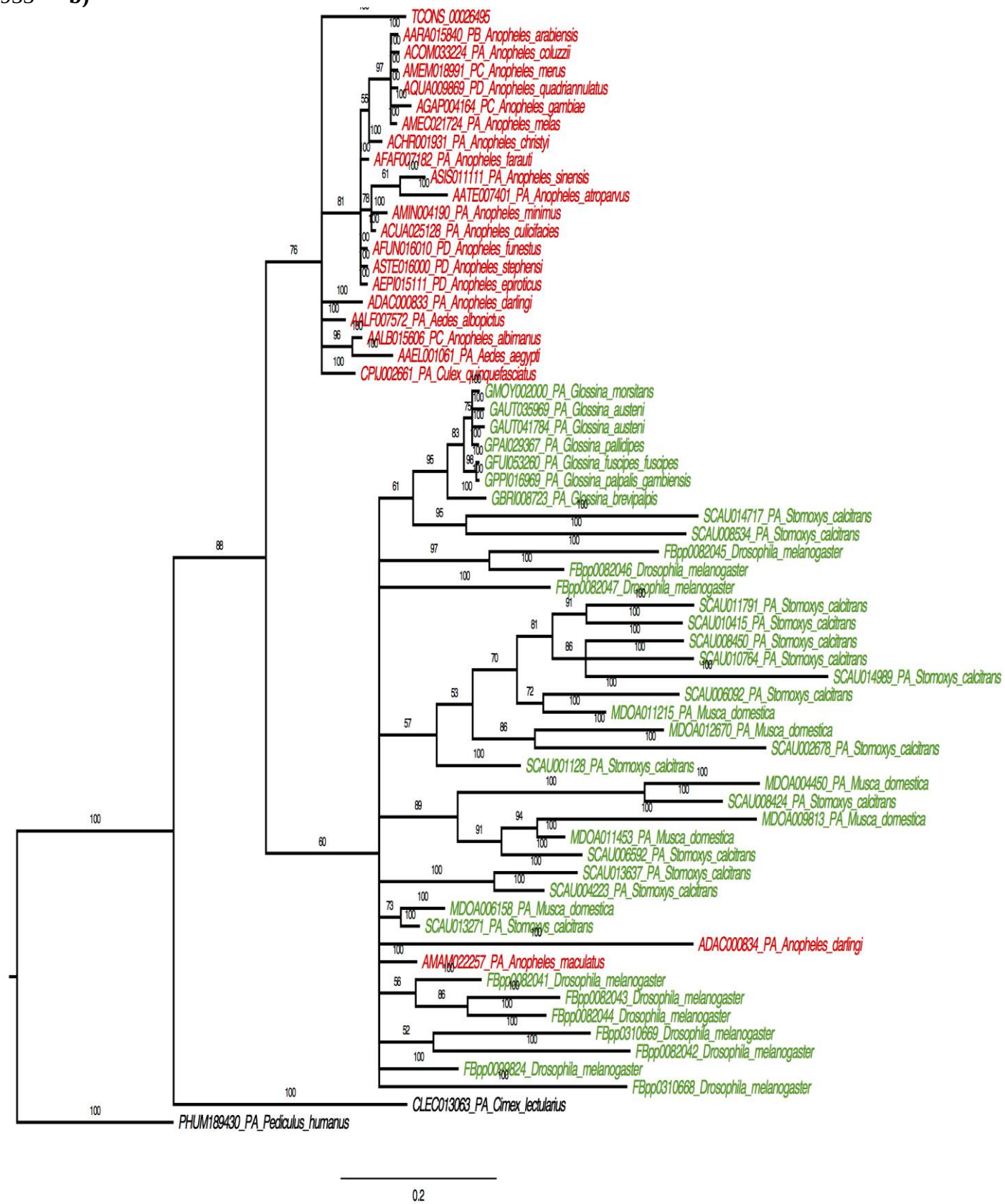


**Fig. 3**

**a)**


TCONS_00014023_C_sonorensis
PPAI010526_P_papatasi
CPIJ003293-RA_C_quinquefasciatus
AALF008031_A_albopictus
AAEL013825_A_aegypti
AMAM023813_A_maculatus
AMAM008270_A_maculatus
ASTE008912_A_stephensi
AMIN002753_A_minimus
ACUA005022_A_culicifacies
AFUN000615_A_funestus
AGAP003182_A_gambiae
AQUA008109_A_quadriannulatus
AMEM013350_A_merus
AARA001280_A_arabiensis
ACOM035635_A_coluzzii
AMEC012304_A_melas
ACHR003211_A_christyi
AEPI003917_A_epiroticus
AFAF012816_A_farauti
ADIR006507_A_dirus
ADAC001232_A_darlingi
AALB008726_A_albimanus
ASIS004408_A_sinensis
AATE000828_A_atroparvus
GBRI004156_G_brevipalpis
GPPI000095_G_palpalis_gambiensis
GFUI034293_G_fuscipes_fuscipes
GAUT041349_G_austeni
GPAI045896_G_pallidipes
MDOA011080_M_domestica
SCAU006195_S_calcitrans
FBpp0083853_D_melanogaster
CLEC011697_C_lectularius
RPRC009671_R_prolixus
PHUM420000_P_humanus
ISCW013335_I_scapularis
SSCA002093_S_scabiei
BGLB001932_B_glabrata
S_cerevisiae_SKI2
0.3

**b)**


TCONS_00026495
AARA015840_PB_Anopheles_arabiensis
ACOM033224_PA_Anopheles_coluzzii
AMEM018991_PC_Anopheles_merus
AQUA009869_PD_Anopheles_quadriannulatus
AGAP004164_PC_Anopheles_gambiae
AMEC021724_PA_Anopheles_melas
ACHR001931_PA_Anopheles_christyi
AFAF007182_PA_Anopheles_farauti
ASIS011111_PA_Anopheles_sinensis
AATE007401_PA_Anopheles_atroparvus
AMIN004190_PA_Anopheles_minimus
ACUA025128_PA_Anopheles_culicifacies
AFUN016010_PD_Anopheles_funestus
ASTE016000_PD_Anopheles_stephensi
AEPI015111_PD_Anopheles_epiroticus
ADAC000833_PA_Anopheles_darlingi
AALF007572_PA_Aedes_albopictus
AALB015606_PC_Anopheles_albimanus
AAEL001061_PA_Aedes_aegypti
CPIJ002661_PA_Culex_quinquefasciatus
GMOY002000_PA_Glossina_morsitans
GAUT035969_PA_Glossina_austeni
GAUT041784_PA_Glossina_austeni
GPAI029367_PA_Glossina_pallidipes
GFUI053260_PA_Glossina_fuscipes_fuscipes
GPPI016969_PA_Glossina_palpalis_gambiensis
GBRI008723_PA_Glossina_brevipalpis
SCAU014717_PA_Stomoxys_calcitrans
SCAU008534_PA_Stomoxys_calcitrans
FBpp0082045_Drosophila_melanogaster
FBpp0082046_Drosophila_melanogaster
FBpp0082047_Drosophila_melanogaster
SCAU011791_PA_Stomoxys_calcitrans
SCAU010415_PA_Stomoxys_calcitrans
SCAU008450_PA_Stomoxys_calcitrans
SCAU010764_PA_Stomoxys_calcitrans
SCAU014989_PA_Stomoxys_calcitrans
SCAU006092_PA_Stomoxys_calcitrans
MDOA011215_PA_Musca_domestica
MDOA012670_PA_Musca_domestica
SCAU002678_PA_Stomoxys_calcitrans
SCAU001128_PA_Stomoxys_calcitrans
MDOA004450_PA_Musca_domestica
SCAU008424_PA_Stomoxys_calcitrans
MDOA009813_PA_Musca_domestica
MDOA011453_PA_Musca_domestica
SCAU006592_PA_Stomoxys_calcitrans
SCAU013637_PA_Stomoxys_calcitrans
SCAU004223_PA_Stomoxys_calcitrans
MDOA006158_PA_Musca_domestica
SCAU013271_PA_Stomoxys_calcitrans
ADAC000834_PA_Anopheles_darlingi
AMAM022257_PA_Anopheles_maculatus
FBpp0082041_Drosophila_melanogaster
FBpp0082043_Drosophila_melanogaster
FBpp0082044_Drosophila_melanogaster
FBpp0310669_Drosophila_melanogaster
FBpp0082042_Drosophila_melanogaster
FBpp0098824_Drosophila_melanogaster
FBpp0310668_Drosophila_melanogaster
CLEC013063_PA_Cimex_lectularius
PHUM189430_PA_Pediculus_humanus
0.2