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

Authors' e-mail addresses (in order of authorship): mhinsley@ebi.ac.uk; iarmean@ebi.ac.uk; rhiannon.silk@pirbright.ac.uk; lara.harrup@pirbright.ac.uk;

asier.gonzalez@rothamsted.ac.uk; eva.veronesi@uzh.ch; lcampbell@ebi.ac.uk; Dana.Nayduch@ARS.USDA.GOV; saski@clemson.edu; wjt@ufl.edu; pkersey@ebi.ac.uk; simon.carpenter@pirbright.ac.uk; mark.fife@pirbright.ac.uk

Abstract

Background: The use of the new genomic technologies has led to major advances in control of several arboviruses of medical importance such as Dengue. 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 genus of vectors. In the present study, the genome of *Culicoides sonorensis,* a vector of bluetongue virus (BTV) in the USA, has been sequenced to provide the first reference genome for these vectors. In this study, we also report the use of the reference genome to perform initial transcriptomic analyses of vector competence for BTV.

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

Conclusions: The reference genome of *C. sonorensis* has enabled preliminary analyses of the gene expression profiles of vector competent and refractory individuals. The genome and transcriptomes generated in this study provide suitable tools for future research on arbovirus transmission. These provide a significant resource for these vector lineage, which diverged from other major Dipteran vector families over 200

million years ago. The genome will be a valuable source of comparative data for other important Dipteran vector families including mosquitoes (Culicidae) and sandflies (Psychodidae), and yield potential targets for transgenic modification in vector control and functional studies.

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

86 transcriptomic and genome manipulation studies [14]. Only one of 14 confirmed vector
87 species of *Culicoides* is currently colonized, *Culicoides sonorensis* Wirth and Jones [8, 15].
88 *Culicoides sonorensis* colonies have already provided significant insights into the genetic
89 basis of vector competence for BTV [16], a *de novo* transcriptome [17], and have been
90 used to construct a physical map of the *C. sonorensis* genome which consists of four
91 chromosomes [18, 19].
92 Susceptibility to infection and transmission of arboviruses by *Culicoides* is determined
93 in part by the heritability of barriers to virus dissemination following ingestion of the
94 bloodmeal [13, 20, 21]. Experimental evidence of a midgut infection barrier, a midgut
95 escape barrier and a haemocoel dissemination barrier in *C. sonorensis* (formerly *C.*
96 *variipennis* or *C. v. sonorensis* [22]) have been defined [23]. There is no evidence to date
97 of salivary gland barriers preventing transmission in any species of *Culicoides* [24], as
98 has been inferred in several species of mosquitoes[20]. Previous studies of the genetic
99 basis for vector competence of *C. sonorensis* for BTV, using maternal inheritance
100 properties, identified a 90 kd protein and used antibodies to this protein to isolate and
101 characterize a cDNA clone encoding a glutathione S-transferase class delta enzyme [25].
102 Differential gene regulation in response to BTV infection, however, has not been
103 investigated to date.
104 In this study, we sequence, *de novo* assemble, annotate and explore the first full genome
105 of *C. sonorensis*. We then use transcriptomic analyses both to improve gene prediction
106 within the genome build, and to elucidate differential gene expression associated with
107 *Culicoides* competence for a BTV serotype 1 strain. The full genome sequence and
108 transcriptome analyses are important resources for further studies of this phylogenetic
109 group which is separated from the other major Dipteran vector families by at least 220
110 million years [26]. In addition to being a valuable resource for comparative study of

111 vector phylogenomics, the provision of *Culicoides* genomes is of interest as a group that
112 demonstrates several unique features, including a significant capacity for long-distance
113 dispersal by semi-passive flight and the ability to reach huge population density under
114 suitable conditions. In addition, their close association with livestock raises questions
115 concerning both host preference and vector competence for arboviruses. Many of these
116 questions can be readily addressed by understanding genetic diversity within
117 populations which will be enhanced by the provision of comparative genomic data, but
118 also in the long term by yielding targets for transgenic modification. The present
119 genome provides a resource to facilitate all these studies.
120

121 **Methods**

122 **Samples**

123 All *C. sonorensis* used in this study originated from the 'AA' colony which was originally
124 established in 1955 at the Kerrville, Texas laboratory of the US Department of
125 Agriculture [27]. Since 1969, this colony has been maintained at The Pirbright Institute
126 without any additional outbreeding [28]. Within this period, two selection bottlenecks
127 of ≤10 individuals were performed to increase susceptibility to BTV serotype 4 and
128 African horse sickness virus (AHSV) serotype 9 infection (Mellor, Pers Comm).
129 Following the second selection in the 1990's [29], the colony strain was renamed as
130 PIR-s-3. No attempt was made to further reduce heterozygosity prior to using
131 individuals from the PIR-s-3 line in this study. All *C. sonorensis* used were 3-4 days old
132 and had not been fed sucrose following emergence from pupae.
133

134 **DNA extraction**

135 The genomes of *C. sonorensis* males and females were sequenced in separate. Adults
136 were separated by sex under a stereomicroscope and stored in 95% ethanol at room
137 temperature prior to use. In order to obtain sufficient amount of DNA for genomic
138 sequencing, DNA was extracted from 375 males in three pools of 100 and one pool of 75
139 individuals, and from 150 females in one pool of 100 and one pool of 50 individuals.
140 Pooled samples were homogenised twice for 30 seconds in 75 µl of phosphate buffered
141 saline using a Tissuelyser™ (Qiagen, UK) at a frequency of 25 Hz/s and a 3mm stainless-
142 steel ball bearing (DeJay Distribution Ltd, UK) within 2ml screw-topped tubes. Genomic
143 DNA extraction was conducted using gravity-flow anion-exchange tips (Qiagen's
144 Genomics Tip 20/G) with extraction of DNA up to 150kb in size and a maximum of 20µg
145 of product. A user-developed protocol for mosquitoes and other insects was followed
146 (see Supplementary Methods). The DNA obtained from each extraction was pooled by
147 sex and the concentration of the resulting samples was evaluated using a Qubit® 2.0
148 Fluorometer (ThermoFisher Scientific, UK) with the Qubit® DNA BR Assay Kit
149 (ThermoFisher Scientific, UK). The integrity of the genomic DNA was visualized using a
150 5 µl sample on a 1% (*w/v*) agarose gel.
151
152 **Genome sequencing and *de novo* assembly**
153 8 – 9 µg of high molecular weight (HMW) genomic DNA were used to sequence the
154 genome of male and female *C. sonorensis* separately. Two paired-end (PE) libraries of
155 insert sizes 200 bp and 400 bp and a mate-pair (MP) library with an average fragment
156 size of 4.4 kb were sequenced with 150 bp (200 bp and 4.4 kb libraries) and 100 bp
157 (400 bp library) paired-end from both female and male pools were run on Illumina
158 HiSeq 2000 and 2500 sequencing systems. Library construction and sequencing was
159 performed at The Earlham Institute (Norwich, UK). Read pair quality was assessed with

160 FastQC v0.10.1. Results suggested that 6 bp should be removed from the start of each
161 read, and in the case of MP data the last 40-50 bp had unexpectedly high levels of k-mer
162 duplication. Both PE and MP reads were filtered to limit their size to 100bp after
163 removal of the first 6 bp. MP data was processed with NextClip v1.0 [30] to select reads
164 which contained both ends of the original fragment. This created a dataset of reads with
165 variable length, which were subsequently limited to 100bp.
166 PE and MP libraries were initially assembled using Velvet v1.2.08 [31] with a k-mer
167 length of 91. The N50 of the assembly was maximized using an expected coverage of 27,
168 a coverage cut-off of six, and with scaffolding option set to 'no'. This initial assembly was
169 filtered to remove short (<4kb) and repeated contigs (> 90% sequence ID to a larger
170 contig) using BLAT v34 and a custom script to count identical bases, considering any
171 overlaps between matches. The assembly was then scaffolded using SSPACE v2.0 [32]
172 and the 400bp PE library, using three passes of decreasing stringency. Any remaining
173 contigs smaller than 500bp were removed (9913 total). The resulting assembly was
174 then scaffolded a second time using the 4.4kb MP library with SSPACE v2.0 (Figure 1).
175 This assembly was assessed using FRCbam v1.0 [33] and Reapr v1.0.15 [34], and the
176 output from Reapr was processed with GapFiller v1.11 [35]. The final assembly was
177 assessed using FRCbam v1.0 to confirm that an improvement in the reported error rate
178 had occurred. Genome size was estimated using the number of total trimmed
179 nucleotides which were used in the assembly divided by the maximum of the per-
180 position coverage frequency distribution, after reads were mapped back to the
181 assembly using BWA-MEM [36]. Kmer spectra were assessed using KAT
182 (https://kat.readthedocs.io/en/latest). Redundant contigs were identified using
183 Redundans [37].

184 Data (reads, annotation and assembly) have been deposited in the ENA database under
185 the accession number PRJEB19938. The genome used in this present study is the latest,
186 v1.
187
188 **Gene model prediction and annotation**
189 Prediction and annotation of the genome assembly was conducted using MAKER v2.31.6
190 [38]. The genome annotation was done using the transcriptome data reported in this
191 study and transcriptome data from previous studies on *C. sonorensis* [17, 39]. In
192 addition, we used the genome annotations of ten species of *Anopheles*, *Ae. aegypti,*
193 *Belgica antarctica* Jacobs, *Culex quinquefasciatus* Say, *Drosophila melanogaster* Mg.,
194 *Lutzomyia longipalpis* (Lutz & Neiva, 1912) and *Phlebotomus papatasi* (Scopoli) (see
195 Supplementary Methods for the species and genome assembly versions used). SNAP
196 v2006-7-28 [40] and AUGUSTUS v2.5.5 [41] were used to predict *ab initio* gene models.
197 SNAP v2006-7-28 was trained using 500 models and AUGUSTUS v2.5.5 using the best
198 1000 models extracted from the initial MAKER v2.31.6 output, as recommended in the
199 software documentation. Coverage of conserved proteins was assessed using the
200 CEGMA v2.4 [42] and BUSCO v1 [43] pipelines. Analysis of Gene Ontology annotation
201 was performed with Blast2GO v3 [44]. The orthology of the annotated genes was
202 established using the Ensembl Compara Gene Trees pipeline [45], using the genomes of
203 *C. sonorensis* and 17 other arthropod species as input (see Supplementary Metods).
204 Genes were functionally annotated using InterProScan, which computationally
205 identifies the presence of domain, motif or family signatures within a protein sequence
206 and infers functional descriptors (taken from the Gene Ontology (GO)).
207
208 **Infection experiments with bluetongue virus**

209 The virus used to infect *C. sonorensis* was a western topotype strain of BTV serotype 1
210 (GIB2007/01) [46]. Virus stock was mixed 1:1 with defibrinated horse blood (TCS
211 Biosciences, Buckingham, UK) with 6.2 Log_{10} $TCID_{50}$/ml being the final infectious dose.
212 All blood-feeding of *C. sonorensis* was conducted using a Hemotek membrane-based
213 system (Discovery Workshops, Accrington, UK). A total of 150 and 145 *C. sonorensis* in
214 two replicates were fed on the blood:BTV-1 suspension and survived the extrinsic
215 incubation period of eight days at 25 °C. During this period of incubation, 10% (*w*/v)
216 sucrose solution was offered via a cotton wool pad. Each individual was then
217 decapitated using disposable needles and the head and remainder of the body of each
218 were stored separately in RNAlater™ (Thermofisher Scientific, UK) at -20 °C.
219
220 **RNA extraction**
221 RNA extraction from *C. sonorensis* was done following the TRIzol® (ThermoFisher
222 Scientific, UK) protocol (see Supplementary Methods). In brief, samples were
223 homogenised in 100 µl of Schneider's *Drosophila* media (Gibco™, Thermofisher
224 Scientific, UK) and RNA was extracted using TRIzol® and chloroform, followed by a
225 precipitation with Isopropanol. The integrity of the RNA was analysed using a
226 Bioanalyser 2100 and the concentration was estimated on a Qubit® 2.0 Fluorometer
227 (ThermoFisher Scientific, UK) with the Qubit® DNA BR Assay Kit (ThermoFisher
228 Scientific, UK).
229 RNA was extracted from individual heads of *C. sonorensis* fed with blood:BTV-1 as
230 described for the bodies (Supplementary Methods). The corresponding bodies
231 (abdomen and thorax) of the BTV-1-positive and BTV-1-negative heads were pooled
232 separately for RNA extraction. In addition, the RNA from 100 *C. sonorensis* fed on horse
233 blood three days post-emergence and left without access to sucrose for eight days, and

234 from 100 *C. sonorensis* fed on a 10% (*w*/v) sucrose solution from 3-8 days following
235 emergence, were also extracted and their transcriptomes sequenced as controls.
236
237 **Quantification of infection with BTV-1**
238 We used RT-qPCR to detect and quantify the presence of disseminated BTV-1 viral RNA
239 in the heads of blood:BTV-1 fed *C. sonorensis*. For this, the RNA was first reverse-
240 transcribed into DNA using the ProtoScript II First Strand cDNA Synthesis kit (New
241 England Biolabs) (see Supplementary Methods for a detailed protocol). A reaction with
242 no enzyme was included as a no-RT negative control. The resulting cDNA was used as
243 template in RT-qPCRs for the detection of BTV using SYBR Green assays
244 (Supplementary Methods). The reactions were performed using the BTV specific
245 primers BTVuni 291-311F 5' GCTTTTGAGGTGTACGTGAAC 3' and BTVuni 381-357R 5'
246 TCTCCCTTGAAACTCTATAATTACG 3' [47]. No-template and no-RT reactions were
247 included as negative controls. To verify that negative reactions were not due to a lack of
248 cDNA in the sample, we also performed RT-qPCR reactions for each sample using *C.*
249 *sonorensis* specific primers for the Vacuolar ATPase gene (Vac-ATPase forward 5'
250 GCTGCTGCTGCCATCATTTT 3' and Vac-ATPase reverse 5' CCGGTCGCATCACTGACATA
251 3'). All reactions were performed in duplicate. Identification of samples with and
252 without BTV-1 RNA was done using the quantification cycle (C_q) values of the reactions.
253 For the purposes of this study, we considered *vector competent* individuals to be those
254 with disseminated infections that include replication in the head capsule, as this has
255 been demonstrated to allow the isolation of BTV-1 [46]. We refer as *refractory*
256 *individuals* to those without detectable BTV after a suitable incubation period (eight
257 days), those with BTV midgut infections that had not disseminated within the insect,
258 those where the process of dissemination had not been completed (i.e. those with

259 haemocel dissemination barriers, midgut escape and infection barriers) and those still
260 retaining inactivated BTV following the infective bloodmeal [24]. In the case of this
261 study, we used a C_q value of 27 to differentiate between vector competent and vector
262 refractory (see Results).
263

264 **Transcriptome sequencing and analyses**

265 In total, eight transcriptomes were sequenced: two biological replicates each for i) BTV-
266 competent; ii) BTV-refractory; iii) blood-fed; and, iv) sucrose-fed. Library construction
267 and sequencing was performed at Edinburgh Genomics (The University of Edinburgh,
268 Scotland). A total of 2.5 µg of RNA from each sample was used to construct TrueSeq
269 libraries, which were sequenced using 50 base paired-end (PE) reads on an Illumina
270 HiSeq 2500. Each library was run in two lanes.
271 The quality of the raw sequence data was analyzed in FastQC v0.11.2. Reads were
272 aligned to the genome assembly (Illumina assembly, without removal of the redundant
273 contigs) with TopHat v2.0.6 [48] using the gene model annotations generated from the
274 genome as reference. The resulting bam files were assembled with Cufflinks v2.2.1 [49]
275 and merged into a single transcriptome with Cuffmerge v2.2.1. Quantification of gene
276 and transcript expression and comparison of the expression levels was performed with
277 Cuffquant v2.2.1 and Cuffdiff v2.2.1, respectively. The Cuffquant results for each
278 transcriptome (two biological + two sequencing lane runs) were grouped per
279 experimental conditions (blood-fed, sucrose-fed, vector competent and vector
280 refractory) to compare the expression profiles between these different experimental
281 conditions with Cuffdiff. Differential expression analyses were then explored with the R
282 package CummeRbund v2.9.3 [50]. Differentially expressed (DE) genes with a
283 significant change in expression level were identified using an α value of 0.05, which

284 establishes the filtering value of the multiple-testing corrected *q*-values. Identification
285 and functional classification of the DE genes was done using BLAST and InterProScan
286 searches in Blast2Go version 4.0.2. Enrichment and gene set enrichment analyses were
287 performed on the sets of DE genes between the different conditions using the interface
288 provided in Blast2Go to the Fisher's Exact Test implemented in FatiGO [51] and the
289 Gene Set Enrichment Analysis (GSEA) package [52]. To perform the GSEA, all genes
290 were ranked per the logarithmic fold-change, removing the infinite values.
291 Transcriptome data have been deposited in the ENA database under the accession
292 numbers ERR2171964-ERR2171979.
293
294 **Validation of differentially expressed genes**
295 Quantitative reverse transcription PCR (RT-qPCR) was used to validate the change in
296 expression of four differentially expressed (DE) genes between the vector competent
297 and refractory pools of *C. sonorensis*. These DE genes were selected based on previous
298 identification in vector competence studies of *C. sonorensis* or because they have a direct
299 functional link with antiviral response. The genes were: a gene sharing 73% identity
300 with the antiviral helicase *ski2* of *Cx. quinquefasciatus* (*ski2*; XP_001845019); a gene
301 sharing 63% identity with *glutathione S-transferase* (*gst*; XP_001654620), a gene
302 sharing 96% identity with *glutathione S-transferase-1* from *C. sonorensis* (*gst-1*;
303 AAB94639) and a gene sharing 49% identity with a gene encoding a *Toll* protein in
304 *Acyrthosiphon pisum* (Hempitera: Aphididae) (XP_001948700). The RNA used to
305 validate the change in expression were the two biological replicas of vector competent
306 and refractory *C. sonorensis* transcriptomes sequenced.
307 Expression levels were normalized against three reference genes whose stability was
308 ranked using three approaches, BestKeeper [53], geNorm [54], and NormFinder [55]

309 (Supplementary Table S3). Gene-specific primers and probes were designed using
310 Primer3Plus [56] (Supplementary Table S4; Supplementary Methods). Primers were
311 compared against the *C. sonorensis* genome using BLAST to ensure specificity to the
312 target region. Amplification efficiency was tested by RT-qPCR with hydrolysis probes by
313 generating a standard curve using triplicate technical replicates of five-fold serial
314 dilutions of RNA template from teneral female *C. sonorensis* specimens. All RNA
315 templates were treated with Turbo DNA-Free™ Kit (ThermoFisher Scientific, UK) prior
316 to RT-qPCR to remove any DNA from the sample following the manufacture
317 recommended protocol. Starting template total RNA concentration was evaluated using
318 a Qubit® 3.0 Fluorometer (ThermoFisher Scientific, UK) with the Qubit® RNA HS Assay
319 Kit (ThermoFisher Scientific, UK).
320 cDNA synthesis and quantitative amplification was done in one reaction using the
321 Superscript® III One-Step RT-qPCR System with Platinum® Taq DNA polymerase
322 (ThermoFisher Scientific, UK) (see Supplementary Methods for a detailed protocol).
323 Reactions were performed following the fast cycling programme as described in the
324 manual (Supplementary Methods). PCR amplification efficiency of each primer-probe-
325 target combinations was calculated via the linear regression of C_q as the log2 of the
326 relative RNA template concentration using the ggplot2 package [57] in R. Reactions
327 which exhibited amplification efficiency between 90-110% with an R^2 value of ≥0.98
328 were considered valid. Relative expression levels for the gene of interest between
329 vector competent and refractory *C. sonorensis* were calculated using the ΔΔCq method
330 [58].
331
332 **Molecular evolution analyses**

333 Homologues of the *ski2* and *gst-1* genes in other insect species were identified using
334 BLAST in Blast2Go. The most significant BLAST hits from *Ae. aegypti*, *Cx.*
335 *quinquefasciatus* or *An. gambiae* were then used to identify the orthologues in other
336 vector species in VectorBase [59]. We verified that the sequences used started with
337 Methionine and ended with a stop codon, whenever possible. The protein sequences
338 identified were aligned using the mode expresso of T-Coffee [60]. Poorly aligned regions
339 (positions with score 0-6) were trimmed from the alignment and the resulting
340 alignments were used in the evolutionary analyses (see Supplementary Methods for
341 more details and commands used to align the sequences; Supplementary Alignment files
342 which include the sequences before and after alignment in fasta format). Genetic
343 distances and the diversity were estimated using the software MEGA v7.0.26
344 [61].Phylogenetic analyses were performed in MrBayes v3.2 [62] using the protein
345 alignment as input and the models of evolution that best fitted our alignments were
346 identified using ProtTest 3.4.2 [63]. Two independent analyses of two million
347 generations with four chains (one cold and three heated) were conducted. Trees and
348 parameters were sampled every 50th generation and the first 10,000 were discarded.
349 The remaining trees were used to estimate the consensus and Bayesian posterior
350 probabilities of each branch. The *ski2* from *Saccharomyces cerevisiae* Meyen ex E.C.
351 Hansen and the *gst-1* from *Pediculus humanus* L. were used as outgroups.
352

353 **Results**

354 **Genome sequencing, construction and annotation**

355 Assembly of the Illumina reads resulted in a genome of 197,417,436 bp assembled in
356 7,974 scaffolds. In contrast, the total genome size estimated using the method of Schell
357 *et al.* [36] (from the maximum depth of read coverage) was 294.45 Mb (Supplementary

358 Table S11). The N50 of the genome assembly was 89,502 and the proportion of Ns was
359 of 2.63% (Table 1). Analysis with Repeat Modeller [64] identified a total of 793 repeat
360 regions encompassing ~14% of the genome. When low-complexity regions are
361 included, the total repeat content is 29.7% including 2 Mb of Type II transposons and 5
362 Mb of Type I transposons (mostly LINE elements). The GC content of the *C. sonorensis*
363 genome is 28% in the complete assembly and 34% in the exons (Table 1). This is low
364 when compared to the closest fully sequenced relative to *C. sonorensis* (*Belgica*
365 *antarctica* (Diptera: Chironomidae): total GC content = 39%) and a range of other
366 Diptera genomes (Table 1).
367 Analyses on the completeness of the assembly identified 98.79% (100% including
368 partial matches) of 248 Core Eukaryotic Genes Mapping Approach (CEGMA) genes and
369 over 97% of the Benchmarking Universal Single-Copy Orthologs (BUSCO) gene set for
370 Insecta (97.1%), Arthropoda (97.7%) and Metazoa (97.3%), with additional genes from
371 these sets (0.7-0.9%) found in fragmentary form. Of these genes, 65.5% (Metazoa)-
372 66.3% (Insecta, Arthropoda) were found in single copy, compared with 89.3% of Insecta
373 BUSCO genes that were present in single copy in the *Aedes aegypti* assembly, and a
374 higher number for some other assemblies (Supplementary table S12). This indicates
375 that there is some level of duplication within the *Culicoides* genome assembly, which
376 could be due to a variation among/within the sequenced genomes and the separate
377 representation within the assembly of alternative alleles.
378 Annotation of the Illumina assembled genome resulted in 15,629 protein-coding genes
379 and 21,336 transcripts, which is comparable to the number of protein-coding genes in
380 other species of Diptera (Table 1). The mean gene, intron and exon lengths in *C.*
381 *sonorensis* are 5,040 bp, 828 bp, and 480 bp, respectively. The average number of
382 introns per gene in *C. sonorensis* is 6.7. A total of 2,991 genes have been estimated to be

383 alternatively spliced and the maximum number of splice variants per gene (19) is
384 comparable to *Ae. aegypti* and *An. gambiae*.
385 To assess redundancy, we generated the K-mer spectra of the assembly and removed
386 potentially redundant contigs using Redundans. This generated an alternative assembly
387 of 156 Mb, with 3,839 contigs, and an N50 scaffold length of 109,184 bp. Coverage of the
388 BUSCO data sets was not changed, with values for Insecta, Arthropoda, Metazoa and
389 Eukaryota within +/- 0.1% of the original assembly. The proportion of duplicate
390 BUSCOs, however, was reduced from 30.8% (unreduced Illumina assembly) to 12.1%-
391 13.2% (reduced assembly) (Supplementary Table S13), confirming the presence of
392 BUSCO gene family members within the redundant contigs.
393 Homology to genes annotated in other species was detected for 93% of the genes
394 annotated in *C. sonorensis*, with an average of 19 homologous genes and nine
395 orthologues identified per gene; these numbers are comparable to the other species
396 used in the analysis. Over 90% of the *C. sonorensis* genes were linked to their
397 homologues via the ancestral nodes of Arthropoda (6,609 genes), Neoptera (6,361
398 genes) or Diptera (1,225 genes).
399 InterProScan was used to detect protein domains and assign GO terms to genes. Overall,
400 a similar number of domains and high-level GO terms were associated with at least one
401 gene as in most other sequenced insects (Supplementary table S14), although a larger
402 number of specific GO terms tend to have been applied in genomes that have been
403 manually annotated for gene function. In total, 46,329 GO term assignments were made.
404 This compares to 45,600 assignments for An. *gambiae*, and 103,897 assignments for *D.*
405 *melanogaster* (in which many terms have been manually assigned based on the
406 literature, in addition to automatic assignments). Of 149 high-level categories in the GO
407 ("GO Slim"), 137 were present in *C. sonorensis* as opposed to 140 categories represented

408 in the genome of *An. gambiae* and 144 in *D. melanogaster*. GO terms assigned to more
409 genes in *C. sonorensis* than either *An. gambiae* and *D. melanogaster* include: ion binding
410 (2,869 genes, as opposed to 2543 in *An. gambiae* and 2,349 in *D. melanogaster*),
411 oxidoreductase activity (703 genes, as opposed to 622 in *An. gambiae* and 632 in *D.*
412 *melanogaster*), kinase activity (367 genes as opposed to 315 in *An. gambiae* and 363 in
413 *D. melanogaster*), hydrolase activity acting on glycosyl bonds (112 genes as opposed to
414 105 in *An. gambiae* and 107 in *D. melanogaster*) and tRNA metabolic process (101 genes
415 as opposed to 86 in *An. gambiae* and *D. melanogaster*). These results suggest an increase
416 in the number of genes encoding these functions in *C. sonorensis* compared with other
417 species, although the possibility that such genes are selectively over-represented in the
418 current *C. sonorensis* assembly cannot be ruled out.
419 To identify tandemly arrayed genes, we looked for genes that contained common
420 domain architectures (as defined by the set of InterPro-defined domains found within
421 each gene) and were located within 100,000 nucleotides of each other. 1,281 genes
422 were found in such arrays. The longest such array contained 10 copies of a gene
423 containing a chitin-binding domain. The longest array of a protein with this architecture
424 comprises just two copies in *D. melanogaster*, and there are no such arrays in *Ae. aegypti*
425 or *An. gambiae*, although there are various longer arrays of proteins containing this
426 domain alongside others in these species (Supplementary Table S15). The second
427 longest such array contains 9 copies of a gene encoding a protein with arrestin and
428 immunoglobulin domains; the homologous proteins are found in 10 tandem copies in
429 *Ae. aegypti* and *D. melanogaster*, and four tandem copies in *An. gambiae* (Supplementary
430 Table S15).
431
432 **Transcriptome sequencing and analysis**

433 Paired-end sequencing of transcriptomes resulted in an average of 50 million reads
434 with a length of 50 bp for each sample sequenced. Phred scores exceeded 34 in all cases.
435 The transcriptomes from the different conditions were merged with Cuffmerge prior to
436 a comparative analysis of their gene expression profiles. The number of genes on the
437 merged transcriptome is 17,263 and the number of transcripts is 35,813. Of the genes
438 present in the transcriptome, 15,630 had already been annotated in the genome. 1,816
439 transcripts, mapping to 1,635 loci, were novel. The proportion of gene loci with a single
440 transcript is 62.7% (10,826), but 93.3% among the genes that had not been annotated
441 on the genome. BLAST searches against NCBI's Arthropod non-redundant (nr) protein
442 database resulted in 92.4% of the isoforms (33,095) and 90% of the genes matching to a
443 database entry (15,268 of the genes had one or more hits).
444

445 **Differential expression of genes between competent and refractory *Culicoides***
446 ***sonorensis***

447 Eighty-five (56%) *C. sonorensis* in replicate one and 53 (37%) in replicate two produced
448 C_q values ≤ 27 (Supplementary Figure S1). We used the C_q value 27 as a cut off for
449 differentiating between vector competent and refractory individuals as the distribution
450 of values was bimodal and 27 was the local minimum (Supplementary Figure S1). Sixty-
451 five (44%) *C. sonorensis* in replicate one and 92 (63%) in replicate two produced C_q
452 values of > 27 or no C_q value. Replicate one had a significantly greater proportion of
453 disseminated infections than replicate two (Fishers exact test; df=1, P = 0.0007). Three
454 and nine individuals from the first and second experiment replicates were false
455 negatives, that is, samples that were negative for BTV-1 but that also failed to amplify
456 the *C. sonorensis* genes. This indicated that the RNA extraction failed and were discarded
457 from further analyses. Transcriptome studies were therefore conducted using 147 and

458 136 female *C. sonorensis*, from the two biological replicates of the infection study,
459 respectively.
460 Gene expression profiles showed that more genes had their expression upregulated in
461 refractory *C. sonorensis* than in competent individuals (Fig. 1a). A total of 165 genes
462 demonstrated significant differential expression (DE) (Fig. 1b; Supplementary Table S1)
463 of which 123 (75%) had been annotated and 42 (25%) represented novel genes
464 (Supplementary Table S2). A total of 94 (57%) were upregulated in refractory
465 individuals, while 71 were downregulated (43%) (Supplementary Table S1). 120 of the
466 165 genes showing DE had one or more BLAST hits in the non-redundant (nr) protein
467 database from NCBI (Supplementary Table S2) and 88% of these corresponded closely
468 to genes from other members of the suborder Nematocera. Of the 165 genes, 63 were
469 associated with gene ontology (GO) terms (Supplementary Figure S2). However,
470 enrichment analysis showed no significant results.
471 *Ski2* (XLOC_006435) and *glutathione S-transferase-1* (CSON011559) showed an increase
472 in expression levels in the females that were resistant to full dissemination, the log2 fold
473 change being -1.10 (P = 0.00035; q-value = 0.00381) and -1.07 (P = 0.00055, q-value =
474 0.00555), respectively (Fig. 2; Supplementary Table S1). In contrast, the *Toll*-like
475 (XLOC_000522) gene and the gene matching *glutathione s transferase* (CSON010973) were
476 expressed at higher levels in the samples that were susceptible to a fully disseminated
477 BTV infection (log2 fold change 1.21, P = 0.00005 and q-value = 0.00072) and (log2 fold
478 change 2.07, P = 0.00005 and q-value = 0.00072) (Fig. 2; Supplementary Table S1).
479 The direction and degree of change in expression levels of the four genes identified as
480 potentially influencing *C. sonorensis* competence for BTV (*Ski2*, *gst-1*, *gst* and *Toll*-like)
481 was confirmed using RT-qPCR and the same samples. Housekeeping genes used as
482 reference were *CytB5*, *rpl13* and *rps8*, which showed the highest level of stability of

483 those tested (Supplementary Table S3 and Table S4). The direction of up and down
484 regulation of the genes and the log-2 fold change was similar in all cases as that detected
485 in the RNAseq data (Supplementary Table S5).
486
487 **Molecular evolution of *ski2 antiviral helicase* and *glutathione s transferase-1***
488 The most significant BLAST hit for the *C. sonorensis ski2* was the gene from *Cx.*
489 *quinquefasciatus* (XP_001845019.1 = CPIJ003293). There are 37 genes identified as
490 orthologous to *Cx. quinquefasciatus ski2* in 36 species in VectorBase. 35 are 1-to-1
491 orthologues and there are two orthologous genes in *An. maculatus* (Supplementary
492 Table S6a). The mean evolutionary diversity estimated using JTT model with gamma
493 distribution value $\alpha = 0.5$ (as estimated with ProtTest) and pairwise deletion across the
494 sequences from Diptera species (no outgroups included) was 0.513, and pairwise
495 distances between protein sequences ranged between 0.000 (*An. coluzzi* v. *An. gambiae*)
496 and 1.198 (*G. fuscipes fuscipes* v. *An. maculatus*). The distance of the ski2 protein
497 between *C. sonorensis* and mosquitoes was 0.752, with sandflies was 0.834, and with
498 Brachycera had a mean distance of 0.928. The mean evolutionary distance within the
499 Nematocera was 0.272, and 0.171 within the Brachycera. The evolutionary relationship
500 of this protein is consistent with species-level phylogeny (Fig. 3a).
501 In the *gst-1* gene from *C. sonorensis*, the most similar BLAST hit is that of *C. variipennis*
502 (synonym of *C. sonorensis*; AAB94639.1) [25]. The second closest was the *gst-1* copy of
503 *An. gambiae* (AGAP004164), for which 61 orthologues have been identified in 31 other
504 species in VectorBase (Supplementary Table S6b). The mean evolutionary distance
505 between the gst-1 protein of the Diptera species, estimated with the same model
506 parameters as for ski2 protein, was 0.647. The mean distance between the gst-1 protein
507 of *C. sonorensis* and that of mosquitoes was 0.199, while its evolutionary mean distance

508 to Brachycera was 0.866. The pairwise distances ranged between 0.000 (observed
509 between several sequences) and 2.914 (between the copy of *M. domestica*
510 MDOA012670 and *An. coluzzi* ACOM0333224). The mean evolutionary distance within
511 the Nematocera was 0.141, and 0.722 within the Brachycera. The phylogeny of the *gst-1*
512 protein recovers the two main clades in Diptera, Brachycera and Nematocera, but there
513 is little resolution within these clades (Fig. 3b).
514
515 **Immune response related genes in *C. sonorensis***
516 All *D. melanogaster* genes from FlyBase that had InterPro hits containing Toll, Imd and
517 JAK/STAT associated to their names and all GO terms matching to the Toll, Imd and
518 JAK/STAT pathway terms were used as reference to identify the homologues in *C.*
519 *sonorensis* (Supplementary Table S7, and S8; Supplementary Figure S3). 36 of 42 genes
520 identified in the Toll pathway of *D. melanogaster* were also identified in *C. sonorensis*;
521 eight of the 11 genes of the Imd signalling pathway in *D. melanogaster* were identified in
522 the *C. sonorensis* genome and 10 of the 13 *D. melanogaster* JAK/STAT pathway genes
523 were identified. There are four instances in which the same *C. sonorensis* genes were
524 identified as different immune-related genes, and we have verified them. Two *C.*
525 *sonorensis* genes, CSON012766 and CSON015181, were predicted by the Compara
526 workflow to be at the same time *Dorsal* and *Dif*. These two genes had 99.6% similarity
527 at protein sequence level and 96% similarity at DNA sequence level, and are located in
528 different scaffolds. The similarity to *D. melanogaster Dorsal* was 65% and to *Dif* was
529 61%. Four *C. sonorensis* genes, CSON001282, CSON001790, CSON007335, CSON011712,
530 were predicted to be both *Toll* (FBgn0262473) and *Tehao* (FBgn0026760) at the same
531 time. Blastp searches identified all four genes as having leucine-rich repeats and
532 Toll/Interleukin-1 receptor homology (TIR) domains, which suggests they could be

533 identified as *Toll* rather than *Tehao*. However, with the similarity with *Toll* is about 38-
534 40% at the protein sequence level. Two *C. sonorensis* genes, CSON003218, CSON008584,
535 have been identified as both *Peptidoglycan recognition protein LC* (Fbgn0035976) and
536 *Peptidoglycan recognition protein LF* (Fbgn0035977). Using Blastp searches, the two
537 genes showed partial similarity higher than 50% to several PGRPs at the protein level,
538 not being able to differentiate between them using Blast. Finally, one single *C. sonorensis*
539 gene (CSON004570) was identified as Unpaired 1, 2 and 3 (FBgn0004956,
540 FBgn0030904, FBgn0053542). Using Blastp, the *C. sonorensis* protein had partial
541 identities of 25-30% to these three Unpaired proteins of *D. melanogaster*. Similar to the
542 cases above, it is not possible to identify the exact identity of the gene without further
543 analyses.
544 Additional searches for immune related genes in *C. sonorensis* were carried out using
545 BLASTp with *Ae. aegypti* and *Cx. quinquefasciatus* gene sequences from the ImmunoDB
546 database (http://cegg.unige.ch/Insecta/immunodb) as queries (Supplementary Table
547 S9 and S10). Four annotated genes in *C. sonorensis* were identified as potential anti-
548 microbial peptides (AMPs) belonging to the gene subfamilies Attacin, Cecropin,
549 Defensin and Diptericin (Supplementary Table S9a); however, the analyses did not
550 identify more than a single gene within each of the AMP subfamilies. Eight potential
551 Toll-like receptors were identified using BLASTp although there were three instances in
552 which the most significant identities were different depending on the species used to
553 provide input data (Supplementary Table S10), reflecting the large difference in
554 sequence between the homologous genes in *Ae. aegypti* and *Cx. quinquefasciatus*. All five
555 genes belonging to the Toll pathway (Supplementary Table S10) and all genes of the
556 JAK/STAT signal transduction pathway were identified in *C. sonorensis* (Supplementary

557 Table S10); while of the Imd pathway, one gene was identified in five of the eight
558 subfamilies (Supplementary Table S10).
559 The Compara approach reconciles the gene tree (derived from sequence similarity
560 metrics) with the species tree to identify likely orthologous relationships; but requires
561 an initial clustering step (which can lead to the inclusion of sequences into the wrong
562 trees) [45]. The only instance in which the two approaches identify a different *C.*
563 *sonorensis* gene and the orthologue of a gene in the above pathways is that of *Pelle*
564 (which activates the Toll Receptor), for which BLAST identifies CSON012655 as its
565 homologue, and Compara identifies CSON013584 and CSON013585. There were other
566 cases in which BLAST identifies a potential homologue while Compara does not (e.g.
567 TNF-receptor-associated factor 6, Traf6, or tube) and genes Imd and Fadd, neither of the
568 two approaches identify a homologue in *C. sonorensis*.
569 It should be noted that the results presented in this study come from automated
570 pipelines. Although some cases have been manually verified, there are still genes which
571 have not been identified in this study and other that need improvement in their
572 annotation. It is expected that the annotation and curation of the genome will advance
573 as it is used by the research community.
574
575 **Discussion**
576 This study has produced the first *de novo* genome assembly of the Dipteran family
577 Ceratopogonidae, which includes important vectors of emerging and re-emerging
578 veterinary arboviruses [8, 65]. The closest relative to this group for which a genome of
579 comparable quality is available is the non-biting midge, *Polypedilum vanderplanki*
580 Hinton (Family: Chironomidae) [66], which diverged from the Ceratopogonidae some
581 220 million years ago. This study provides a primary resource for comparative studies

582 with other arbovirus vector species (e.g. mosquitoes [67]). It also will facilitate efforts to
583 produce transgenic *Culicoides* to control *Culicoides*-borne pathogens, although the
584 limitations on colony production for the other epidemiologically relevant species of
585 *Culicoides* remains a major constraining factor [14, 68].
586 The genome assembly of *C. sonorensis* is not wholly contiguous, however, its general
587 features are consistent with similarly sized genomes from other species. Thus, features
588 such as the repeat content, gene count and level of alternative splicing are within the
589 ranges observed in *Ae. aegypti*, *An. gambiae* and *D. melanogaster*. Coverage of well-
590 conserved gene families is complete and comparable with other published assemblies
591 (Supplementary Table S12). Various measures (the high proportion of BUSCO genes
592 present in multiple copy, and estimates of genome size and redundancy using read
593 mapping and alignment approaches) indicate the presence of redundancy in the
594 assembly (Supplementary Table S13), which could result from genetic variation within
595 and amongst individuals from which sequenced material was extracted. However, the
596 individuals come from a colony that has not had any outbreeding for over 40 years,
597 which should have reduced the genetic variation among the individuals of the colony.
598 Long read-sequencing approaches would likely help reduce this redundancy. However,
599 the gene set has been annotated with a comparable number of InterPro domains and
600 high-level GO terms as *An. gambiae* and other Diptera, while the number of specific GO
601 terms is comparable to other genomes that have not been extensively manually
602 annotated with functional descriptions. The molecular functions of ion binding,
603 oxidoreductase activity, kinase activity, hydrolase activity acting on glycosyl bonds,
604 methyltransferase activity all highly represented, as is the tRNA metabolic process. In
605 addition, the genes of the immune-related pathways have been identified and
606 annotated, revealing complete or nearly-complete orthologous pathways. Overall, these

607 observations indicate that the present assembly and annotation of the *C. sonorensis*
608 genome represents a reliable resource to use in genetic studies of biting midges.
609 Nevertheless, it is still highly fragmented and the use of long-read sequencing
610 technologies like PacBio would help improving the scaffolding. In addition, the use of
611 this genomic resource by the community will help improve the annotation of the genes.
612 This is the case for many of the genomes of vector species that have been neglected over
613 the years.
614 A striking feature of the *C. sonorensis* genome is its GC content, which is among the
615 lowest reported for Dipteran species (although close to identical to the non-biting
616 midge *P. vanderplanki* [66] and other non-Dipteran arthropods like the pea aphid,
617 *Acyrthosiphon pisum* Harris [69]). This is not just a feature of the repeated content, but
618 extends into the protein-coding genes (which have a higher GC content than the non-
619 coding regions, but less than in comparable species: Table 1). Recent comparative
620 analyses of codon usage in Diptera and Hymenoptera showed a significant association
621 between codon bias and high GC content in Diptera, but low GC content in Hymenoptera
622 [70]. Thus, the low GC content observed in *C. sonorensis* could reflect a codon bias in
623 comparison to other Diptera species. Functional investigations of GC content have also
624 revealed that it has an impact on genome functioning and species ecology in microbes
625 [71], vertebrates [72] and plants [73]. It is also known that the GC content has an impact
626 in the efficiency of sequencing technologies [74]. Regions with low GC content have less
627 coverage than more GC balanced ones. This characteristic of the *C. sonorensis* genome
628 should be accounted for in future genomic studies, and appropriate experimental
629 designs that take into account the low GC content should be used [e.g. 75].
630 Studies investigating gene expression during infection with BTV identified four
631 candidate genes whose regulation was correlated with vector competence, despite only

632 using a single time-point at eight days post-infection and pools of *C. sonorensis*
633 exhibiting different degrees of virus dissemination. Interestingly, the elevated
634 expression of *gst-1* in refractory *Culicoides* observed in this study is consistent with a
635 previously proposed genetic mechanism for *C. sonorensis* vector competence for BTV
636 [13, 15, 24]. These studies identified a locus controlling vector competence in the AA
637 colony, the origin colony of PIR-s-3 colony used in the present study. The former study
638 used a maternally inherited 90kd protein consistent with genetic studies that suggested
639 maternal inheritance of the competence controlling factor identified in unfertilized eggs
640 of refractory females to isolate a clone that upon sequencing was identified as *gst-1* [25].
641 Though the current study evaluated gene expression after exposure to BTV and the
642 previous study compared genetically selected resistant and susceptible families using
643 individuals from these families who were not exposed to BTV, both independently
644 identified involvement of *gst-1*. Thus, the differential expression of the gene in the
645 current study is confirmation of its involvement in BTV infection in AA colony, although
646 its role requires functional validation.
647 From the evolutionary perspective, the phylogenies of ski2 and gst-1 reflect the
648 accepted evolutionary history of the included Diptera species, although the phylogenetic
649 tree of ski2 shows more resolution within Nematocera and Brachycera than the gst-1. It
650 is also interesting to note the two contrasting modes of evolution of gst-1 between the
651 two main clades of Diptera. The gst-1 protein in the Brachycera clade shows multiple
652 duplications in comparison with the Nematocera and the estimated diversity in the
653 Brachycera is higher (0.722) than that in the Nematocera (0.141). Within the
654 Nematocera the level of diversity of the gst-1 protein is lower than in ski2.
655 Future studies of vector competence in *C. sonorensis* must characterize genes that
656 facilitate BTV competence and differentiate these from differentially expressed genes

657 that are the consequence of BTV infection. To determine variation in the mechanism
658 and factors controlling BTV-vector competence in the AA and PIR-s-3 colonies observed
659 in the present study, it will be necessary to carry out studies of different populations of
660 *C. sonorensis* and other *Culicoides* vectors and non-vector species. Once controlling
661 genes are identified it will be essential to assess the effect of polymorphisms, the
662 various functions of the identified genes in the absence of the pathogen to evaluate
663 factors that influence their frequencies, and compare these processes between different
664 species of *Culicoides* [23].
665 The present paper presents the genome of *C. sonorensis* and its use in transcriptomic
666 analyses to provide information that will help elucidate the transmission of viruses by
667 vectors and provide new avenues of research to understand vector competence for BTV.
668 We present a list of candidate genes that will be further explored to show their
669 involvement in the transmission of BTV by *C. sonorensis*. Furthermore, the results also
670 provide evidence for the involvement of genes belonging similar functional families in
671 response to virus infection in different species belonging to different families of Diptera.
672 These results show the importance of comparative analyses to interrogate the evolution
673 of host-pathogen interactions, and the *C. sonorensis* genome will facilitate such studies.
674

675 **Conclusions**

676 Here we present the first annotated genome of the biting midge, *C. sonorensis*, a vector
677 of economically important viruses of livestock, such as Bluetongue virus (BTV) and
678 Schmallenberg virus. This genome has been fundamental in gaining information about
679 the genetics of BTV transmission. Thus, gene expression comparison between females
680 susceptible to a full infection by BTV and those females refractory to full BTV infection,
681 identified 165 genes that are candidates to be involved in vector competence in this

682 species. Of these genes, the *gst-1* gene was identified previously to be involved in BTV
683 transmission, while the antiviral helicase *ski2* involved in suppressing the replication of
684 dsRNA viruses. Overall, the publication of the genome and transcriptomes will lead to
685 advances in the control of economically relevant arboviruses of livestock. In addition,
686 this genome fills a phylogenetic gap of more than 200 million years, and will be a useful
687 source of comparative data for studies of vector competence across Diptera.
688

689 **Declarations**

690 **Ethics approval and consent to participate**

691 Not applicable.

692 **Consent for publication**

693 Not applicable.

694 **Availability of data and material**

695 All sequencing data, genomic and transcriptomic, generated in this study has been
696 deposited at the European Nucleotide Archive (ENA) database under the accession
697 number PRJEB19938.

698 **Competing interests**

699 The authors declare that they have no competing interests.

700 **Funding**

701 This work was supported by the Biotechnology and Biological Sciences Research
702 Council (BBSRC) grant BB/J016721/1 and the core Strategic Programme Grant to The
703 Pirbright Institute.

704 **Author's contributions**

705 MF, SC and PK conceived the research project. RMH, MH, IMA, RS, LEH, AGU, EV and LC
706 have performed the laboratory experiments and bioinformatic analyses. DN and CS have

707 helped with the annotation of the immune-related genes. RMH, MF and SC have written
708 the manuscript and PK, WT, ND and CS have provided insightful revision. All authors
709 read and approved the final manuscript.

710 **Acknowledgements**

711 We thank Eric Denison and James Barber in the insectary for their supply of reagents,
712 cell lines and insects for this study. We would like to thank the three reviewers and the
713 editor for their positive comments and suggestions.

714

715 **References**

716 1. Weaver SC, Reisen WK: Present and future arboviral threats. Antiviral Research.
717 2010; 85:328-345.
718 2. Mlakar J, Korva M, Tul N, Popovic M, Poljsak-Prijatelj M, Mraz J, Kolenc M, Rus
719 KR, Vipotnik TV, Vodusek VF *et al*: Zika Virus Associated with Microcephaly. New
720 England Journal of Medicine. 2016; 374:951-958.
721 3. Pialoux G, Gauzere BA, Jaureguiberry S, Strobel M: Chikungunya, an epidemic
722 arbovirosis. Lancet Infectious Diseases. 2007; 7:319-327.
723 4. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS,
724 Machol I, Lander ES, Aiden AP *et al*: De novo assembly of the *Aedes aegypti*
725 genome using Hi-C yields chromosome-length scaffolds. Science. 2017; 356:92-
726 95.
727 5. Alphey L: Genetic Control of Mosquitoes. Annual Review of Entomology. 2014;
728 59:205-224.
729 6. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi ZY, Megy K,
730 Grabherr M *et al*: Genome sequence of Aedes aegypti, a major arbovirus vector.
731 Science. 2007; 316:1718-1723.

732 7. Olson KE, Blair CD: Arbovirus-mosquito interactions: RNAi pathway. Curr Opin
733 Virol. 2015; 15:119-126.
734 8. Purse BV, Carpenter S, Venter GJ, Bellis G, Mullens BA: Bionomics of Temperate
735 and Tropical *Culicoides* Midges: Knowledge Gaps and Consequences for
736 Transmission of *Culicoides*-Borne Viruses. Annual Review of Entomology. 2015;
737 60:373-392.
738 9. Carpenter S, Mellor PS, Fall AG, Garros C, Venter GJ: African Horse Sickness Virus:
739 History, Transmission, and Current Status. Annual Review of Entomology. 2017;
740 62:343-358.
741 10. Savini G, Afonso A, Mellor P, Aradaib I, Yadin H, Sanaa M, Wilson W, Monaco F,
742 Domingo M: Epizootic haemorragic disease. Research in Veterinary Science.
743 2011; 91:1-17.
744 11. Carpenter S, Wilson A, Mellor PS: *Culicoides* and the emergence of bluetongue
745 virus in northern Europe. Trends in Microbiology. 2009; 17:172-178.
746 12. Rushton J, Lyons N: Economic impact of Bluetongue: a review of the effects on
747 production. Vet Ital. 2015; 51:401-406.
748 13. Tabachnick WJ: *Culicoides variipennis* and bluetongue-virus epidemiology in the
749 United States. Annual Review of Entomology. 1996; 41:23-43.
750 14. Nayduch D, Cohnstaedt LW, Saski C, Lawson D, Kersey P, Fife M, Carpenter S:
751 Studying *Culicoides* vectors of BTV in the post-genomic era: Resources,
752 bottlenecks to progress and future directions. Virus Research. 2014; 182:43-49.
753 15. Jones RH, Foster NM: Oral infection of *Culicoides variipennis* with bluetongue
754 virus - development of susceptible and resistant lines from a colony population.
755 Journal of Medical Entomology. 1974; 11:316-323.

756 16. Tabachnick WJ: Genetic control of oral susceptibility to infection of *Culicoides*
757 *variipennis* with bluetongue virus. American Journal of Tropical Medicine and
758 Hygiene. 1991; 45:666-671.
759 17. Nayduch D, Lee MB, Saski CA: The Reference Transcriptome of the Adult Female
760 Biting Midge (*Culicoides sonorensis*) and Differential Gene Expression Profiling
761 during Teneral, Blood, and Sucrose Feeding Conditions. Plos One. 2014; 9.
762 18. Nunamaker RA, Brown SE, Murphy KE, Tabachnick WJ, Knudson DL: Physical
763 mapping of the genome of the biting midge, *Culicoides variipennis*. Molecular
764 Biology of the Cell. 1998; 9:452A-452A.
765 19. Nunamaker RA, Brown SE, McHolland LE, Tabachnick WJ, Knudson DL: First-
766 generation physical map of the *Culicoides variipennis* (Diptera : Ceratopogonidae)
767 genome. Journal of Medical Entomology. 1999; 36:771-775.
768 20. Mellor PS, Boorman J, Baylis M: *Culicoides* biting midges: Their role as arbovirus
769 vectors. Annual Review of Entomology. 2000; 45:307-340.
770 21. Tabachnick WJ: Nature, Nurture and Evolution of Intra-Species Variation in
771 Mosquito Arbovirus Transmission Competence. International Journal of
772 Environmental Research and Public Health. 2013; 10:249-277.
773 22. Holbrook FR, Tabachnick WJ, Schmidtmann ET, McKinnon CN, Bobian RJ, Grogan
774 WL: Sympatry in the *Culicoides variipennis* complex (Diptera : Ceratopogonidae):
775 a taxonomic reassessment. Journal of Medical Entomology. 2000; 37:65-76.
776 23. Carpenter S, Veronesi E, Mullens B, Venter G: Vector competence of Culicoides
777 for arboviruses: three major periods of research, their influence on current
778 studies and future directions. Revue Scientifique Et Technique-Office
779 International Des Epizooties. 2015; 34:97-112.

780 24. Fu H, Leake CJ, Mertens PPC, Mellor PS: The barriers to bluetongue virus
781 infection, dissemination and transmission in the vector, *Culicoides variipennis*
782 (Diptera : Ceratopogonidae). Archives of Virology. 1999; 144:747-761.
783 25. Abdallah MA, Pollenz RS, Droog FN, Nunamaker RA, Tabachnick WJ, Murphy KE:
784 Isolation and characterization of a cDNA clone coding for a glutathione S-
785 transferase class delta enzyme from the biting midge Culicoides variipennis
786 sonorensis Wirth and Jones. Biochemical Genetics. 2000; 38:377-390.
787 26. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, Bertone
788 MA, Cassel BK, Bayless KM, Heimberg AM *et al*: Episodic radiations in the fly tree
789 of life. Proceedings of the National Academy of Sciences of the United States of
790 America. 2011; 108:5690-5695.
791 27. Jones RH: The laboratory colonization of *Culicoides variipennis* (Coq). Journal of
792 Economic Entomology. 1957; 50:107-108.
793 28. Boorman J: Maintenance of laboratory colonies of *Culicoides variipennis* (Coq),
794 *Culicoides nubeculosus* (Mg) and *Culicoides riethi* Kieffer (Diptera,
795 Ceratopogonidae). Bulletin of Entomological Research. 1974; 64:371-377.
796 29. Mellor PS, Rawlings P, Baylis M, Wellby MP: Effect of temperature on African
797 horse sickness virus infection in Culicoides. Archives of Virology. 1998:155-163.
798 30. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M: NextClip: an analysis and
799 read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics. 2014;
800 30:566-568.
801 31. Zerbino DR, Birney E: Velvet: Algorithms for de novo short read assembly using
802 de Bruijn graphs. Genome Research. 2008; 18:821-829.
803 32. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: Scaffolding pre-
804 assembled contigs using SSPACE. Bioinformatics. 2011; 27:578-579.

805 33. Vezzi F, Narzisi G, Mishra B: Reevaluating Assembly Evaluations with Feature
806 Response Curves: GAGE and Assemblathons. Plos One. 2012; 7.
807 34. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD: REAPR: a
808 universal tool for genome assembly evaluation. Genome Biology. 2013; 14.
809 35. Boetzer M, Pirovano W: Toward almost closed genomes with GapFiller. Genome
810 Biology. 2012; 13.
811 36. Schell T, Feldmeyer B, Schmidt H, Greshake B, Tills O, Truebano M, Rundle SD,
812 Paule J, Ebersberger I, Pfenninger M: An Annotated Draft Genome for Radix
813 auricularia (Gastropoda, Mollusca). Genome Biol Evol. 2017; 9:585-592.
814 37. Pryszcz LP, Gabaldon T: Redundans: an assembly pipeline for highly
815 heterozygous genomes. Nucleic Acids Research. 2016; 44.
816 38. Holt C, Yandell M: MAKER2: an annotation pipeline and genome-database
817 management tool for second-generation genome projects. Bmc Bioinformatics.
818 2011; 12.
819 39. Nayduch D, Lee MB, Saski CA: Gene discovery and differential expression
820 analysis of humoral immune response elements in female *Culicoides sonorensis*
821 (Diptera: Ceratopogonidae). Parasites & Vectors. 2014; 7.
822 40. Stanke M, Schoffmann O, Morgenstern B, Waack S: Gene prediction in eukaryotes
823 with a generalized hidden Markov model that uses hints from external sources.
824 Bmc Bioinformatics. 2006; 7.
825 41. Stanke M, Steinkamp R, Waack S, Morgenstern B: AUGUSTUS: a web server for
826 gene finding in eukaryotes. Nucleic Acids Research. 2004; 32:W309-W312.
827 42. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes
828 in eukaryotic genornes. Bioinformatics. 2007; 23:1061-1067.

829 43. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: BUSCO:
830 assessing genome assembly and annotation completeness with single-copy
831 orthologs. Bioinformatics. 2015; 31:3210-3212.
832 44. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M,
833 Talon M, Dopazo J, Conesa A: High-throughput functional annotation and data
834 mining with the Blast2GO suite. Nucleic Acids Research. 2008; 36:3420-3435.
835 45. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle
836 SMJ, Amode R, Brent S *et al*: Ensembl comparative genomics resources (vol 2016,
837 bav096, 2016). Database-the Journal of Biological Databases and Curation. 2016.
838 46. Veronesi E, Antony F, Gubbins S, Golding N, Blackwell A, Mertens PPC, Brownlie J,
839 Darpel KE, Mellor PS, Carpenter S: Measurement of the Infection and
840 Dissemination of Bluetongue Virus in Culicoides Biting Midges Using a Semi-
841 Quantitative RT-PCR Assay and Isolation of Infectious Virus. Plos One. 2013; 8:8.
842 47. Shaw AE, Monaghan P, Alpar HO, Anthony S, Darpel KE, Batten CA, Guercio A,
843 Alimena G, Vitale M, Bankowska K *et al*: Development and initial evaluation of a
844 real-time RT-PCR assay to detect bluetongue virus genome segment 1. Journal of
845 Virological Methods. 2007; 145:115-126.
846 48. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate
847 alignment of transcriptomes in the presence of insertions, deletions and gene
848 fusions. Genome Biology. 2013; 14.
849 49. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL,
850 Rinn JL, Pachter L: Differential gene and transcript expression analysis of RNA-
851 seq experiments with TopHat and Cufflinks. Nature Protocols. 2012; 7:562-578.

852 50. L G, C T, D K: cummeRbund: Analysis, exploration, manipulation, and
853 visualization of Cufflinks high-throughput sequencing data. R package version
854 2180. 2013.
855 51. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: a web tool for finding significant
856 associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004;
857 20:578-580.
858 52. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA,
859 Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: Gene set enrichment
860 analysis: A knowledge-based approach for interpreting genome-wide expression
861 profiles. Proceedings of the National Academy of Sciences of the United States of
862 America. 2005; 102:15545-15550.
863 53. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP: Determination of stable
864 housekeeping genes, differentially regulated target genes and sample integrity:
865 BestKeeper - Excel-based tool using pair-wise correlations. Biotechnology
866 Letters. 2004; 26:509-515.
867 54. Hoerndli FJ, Toigo M, Schild A, Gotz J, Day PJ: Reference genes identified in SH-
868 SY5Y cells using custom-made gene arrays with validation by quantitative
869 polymerase chain reaction. Analytical Biochemistry. 2004; 335:30-41.
870 55. Ohl F, Jung M, Xu CL, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G,
871 Loening SA, Radonic A *et al*: Gene expression studies in prostate cancer tissue:
872 which reference gene should be selected for normalization? Journal of Molecular
873 Medicine-Jmm. 2005; 83:1014-1024.
874 56. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM:
875 Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research.
876 2007; 35:W71-W74.

877 57. Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. New York, USA;
878 2009.
879 58. Pfaffl MW: A new mathematical model for relative quantification in real-time RT-
880 PCR. Nucleic Acids Research. 2001; 29.
881 59. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R,
882 Campbell KS, Christophides GK, Christley S, Dialynas E *et al*: VectorBase: a data
883 resource for invertebrate vector genomics. Nucleic Acids Research. 2009;
884 37:D583-D587.
885 60. Notredame C, Higgins DG, Heringa J: T-Coffee: A novel method for fast and
886 accurate multiple sequence alignment. J Mol Biol. 2000; 302:205-217.
887 61. Kumar S, Stecher G, Tamura K: MEGA7: Molecular Evolutionary Genetics Analysis
888 Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 2016; 33:1870-
889 1874.
890 62. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B,
891 Liu L, Suchard MA, Huelsenbeck JP: MrBayes 3.2: Efficient Bayesian Phylogenetic
892 Inference and Model Choice Across a Large Model Space. Systematic Biology.
893 2012; 61:539-542.
894 63. Darriba D, Taboada GL, Doallo R, Posada D: ProtTest 3: fast selection of best-fit
895 models of protein evolution. Bioinformatics. 2011; 27:1164-1165.
896 64. Smit AFA, Hubley R: RepeatModeler Open-1.0. 2008-2015
897 <http://www.repeatmasker.org >) 2015.
898 65. Carpenter S, Groschup MH, Garros C, Felippe-Bauer ML, Purse BV: *Culicoides*
899 biting midges, arboviruses and public health in Europe. Antiviral Research. 2013;
900 100:102-113.

901 66. Gusev O, Suetsugu Y, Cornette R, Kawashima T, Logacheva MD, Kondrashov AS,
902 Penin AA, Hatanaka R, Kikuta S, Shimura S *et al*: Comparative genome
903 sequencing reveals genomic signature of extreme desiccation tolerance in the
904 anhydrobiotic midge. Nature Communications. 2014; 5.
905 67. Severson DW, Behura SK: Mosquito Genomics: Progress and Challenges. Annual
906 Review of Entomology. 2012; 57:143-166.
907 68. Mills MK, Nayduch D, Michel K: Inducing RNA interference in the arbovirus
908 vector, Culicoides sonorensis. Insect Molecular Biology. 2015; 24:105-114.
909 69. Richards S, Gibbs RA, Gerardo NM, Moran N, Nakabachi A, Stern D, Tagu D,
910 Wilson ACC, Muzny D, Kovar C *et al*: Genome Sequence of the Pea Aphid
911 *Acyrthosiphon pisum*. Plos Biology. 2010; 8:24.
912 70. Behura SK, Severson DW: Comparative Analysis of Codon Usage Bias and Codon
913 Context Patterns between Dipteran and Hymenopteran Sequenced Genomes.
914 Plos One. 2012; 7.
915 71. Foerstner KU, von Mering C, Hooper SD, Bork P: Environments shape the
916 nucleotide composition of genomes. Embo Reports. 2005; 6:1208-1213.
917 72. Galtier N, Piganeau G, Mouchiroud D, Duret L: GC-content evolution in
918 mammalian genomes: The biased gene conversion hypothesis. Genetics. 2001;
919 159:907-911.
920 73. Smarda P, Bures P, Horova L, Leitch IJ, Mucina L, Pacini E, Tichy L, Grulich V,
921 Rotreklova O: Ecological and evolutionary significance of genomic GC content
922 diversity in monocots. Proceedings of the National Academy of Sciences of the
923 United States of America. 2014; 111:E4096-E4102.

924 74. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC: Effects of GC bias in next-
925 generation-sequencing data on de novo genome assembly. PLoS One. 2013;
926 8:e62856.
927 75. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B,
928 Kwiatkowski DP, Swerdlow HP *et al*: Optimizing Illumina next-generation
929 sequencing library preparation for extremely AT-biased genomes. BMC
930 Genomics. 2012; 13:1.
931

932 **Figure Legends**

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

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

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

948

949 **Supporting Information Legends**

950 **Supplementary Material Tables**

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

957 **Table S2.** BLAST results for the differentially expressed genes between vector
958 competent and refractory transcriptomes showing the top hit as determined by BLAST.
959 **Table S3.** Ranking of the expression stability of the reference genes tested for use in the
960 quantitative reverse transcription PCR (RT-qPCR). Three different tests were used to
961 rank the expression stability of the genes, BestKeeper, gNorm and NormFinder. *r*-
962 correlation coefficient, Cq – quantification cycle, SE – standard error, SD – standard
963 deviation.
964 **Table S4.** Primers forward (F), reverse (R) and hydrolysis probe (P) sequences used for
965 RT-qPCR analysis of the expression levels of *ski2*, *gst*, *gst-1* and *Toll*-like.
966 **Table S5.** Difference of expression levels in four genes (*Toll*-like, *gst*, *gst-1* and *ski2*)
967 between vector-competent and vector-refractory females assessed by RNAseq and RT-
968 qPCR. * denote a difference in expression between refractory and competent females
969 with a $P < 0.05$ (RT-qPCR significance estimated using a Welch Two Sample t-test with
970 d.f. = 1).
971 **Table S6.** Orthology analysis of two of genes *Ski2* **(a)** and *gst-1* **(b)**.
972 **Table S7.** Immune pathways genes identified in *C. sonorensis* using the Ensembl
973 Compara pipeline and the *D. melanogaster* genes as reference. **a)** Toll pathway; **b)** Imd
974 pathway; **c)** Jak/Stat pathway.
975 **Table S8.** Homologue immune genes in several invertebrate species identified using *D.*
976 *melanogaster* genes as reference all with InterPro and GO terms associated with Toll
977 **(a)**, Imd **(b)** and Jak/Stat **(c)**, in the Ensembl Compara pipeline.
978 **Table S9.** Immune related genes identified in *C. sonorensis* using BLASTp top hits with
979 the gene copies from *Ae. aegypti* and *Cx. quinquefasciatus* available in the ImmunoDB
980 database (http://cegg.unige.ch/Insecta/immunodb) as query. a) Anti-Microbial

981 Peptides; b) Toll receptors; c) Toll path; d) jak/Stat path; e). Imd path. * - Top BLASTp
982 hits different found when *Ae. aegypti* and *Cx. quinquefasciatus* were used as query.
983 **Table S10.** Details of the BLASTp results using *Ae. aegypti* (a) and *Cx. quinquefasciatus*
984 (b) as query to identify the *C. sonorensis* genes of the immune pathways.
985 **Table S11.** Read mapping and estimate of assembly size, by read mapping using BWA,
986 according to the method of Schell et al. [36].
987 **Table S12.** BUSCO analysis using the Insecta data set (version 3.0.2) against the
988 assembled genome sequence of six species, including *Culicoides sonorensis*.
989 **Table S13.** Comparison of the original, and redundancy-reduced, assemblies of *C.*
990 *sonorensis*.
991 **Table S14.** Number of distinct InterPro and GO terms annotated to annotated proteins
992 of *C. sonorensis* and six other Diptera species.
993 **Table S15.** Tandemly repeated gene arrays in *C. sonorensis* versus other insect species.
994 Longest gene arrays in *C. sonorensis* with matching InterPro domain architecture and
995 maximum 100,000 nucleotide between successive members. The longest array with the
996 same domain signature from 3 other insect species is given for comparison.
997
998 **Supplementary Material Figures**
999 **Figure S1.** Distribution of C_q values from RT-qPCR used to detect BTV-1 virus infection
1000 in the two virus-feeding experiments (in blue and orange) of *C. sonorensis*. The vertical,
1001 dashed line corresponds to a C_q value of 27 used to separate vector competent from
1002 refractory females.
1003 **Figure S2.** Classification and functional distribution of the genes differentially
1004 expressed between vector competent and refractory *C. sonorensis* per the Gene

1005 Ontology level 6. Blue: Molecular Function; Green: Cellular Component; Pink: Biological
1006 Process.
1007 **Figure S3.** Number of genes in the Toll (a), Imd (b) and Jak/Stat (c) pathways that have
1008 been duplicated or lost in different species of Diptera as identified using the Ensembl
1009 Compara pipeline.
1010
1011 **Supplementary Material Methods**
1012 **DNA extraction.** Protocol followed to extract DNA from *C. sonorensis* samples.
1013 **Genome annotation and comparative analysis.** Details about the genomes of species
1014 used in the gene annotation and in the comparative analyses.
1015 **RNA extraction.** Protocol used to extract RNA from *C. sonorensis*.
1016 **RT-qPCR for the detection of BTV-1 infection.** Description of the method used to
1017 detect the presence of BTV-1 in *C. sonorensis* females after the experimental infections.
1018 **RT-qPCR protocol for the validation of differentially expressed genes.** Description
1019 of the protocol used to verify the change in expression of *Toll-like protein*, *glutathione-s-*
1020 *transferase*, *glutathione-s-transferase 1* and *ski2 antiviral helicase* genes observed in the
1021 transcriptomes of vector-competent and refractory samples.
1022 **Alignment of gst-1 and ski2.** The commands used to align the protein sequences of
1023 gst-1 and ski2. These alignments were then used to run the phylogenetic analyses.
1024

1025 **Table 1.** A comparison of genome characteristics of the *C. sonorensis* genome with other selected Diptera species.

1026

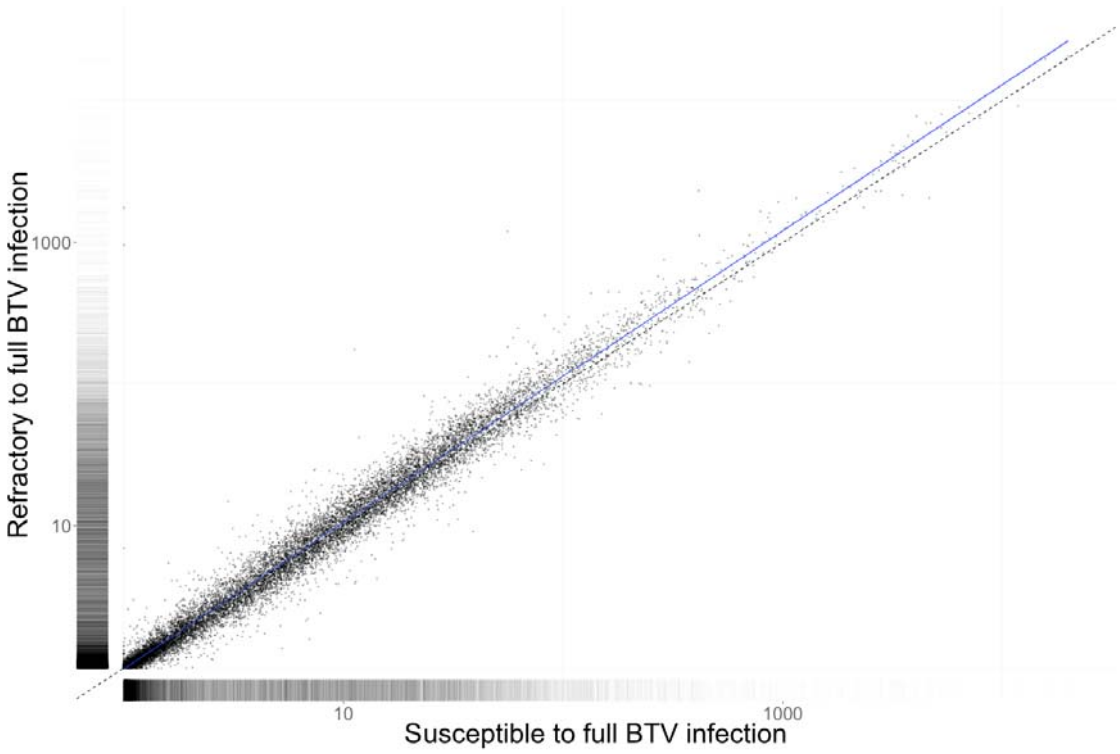
Parameter	*Culicoides sonorensis*	*Aedes aegypti* L3	*Aedes aegypti* L5	*Anopheles gambiae* AGAM P4	*Drosophila melanogaster* BDGP6	*Belgica antarctica* GCA_000775305.1
Genome size (Mb)	197.4	1,384.1	1,362.0	278.3	143.7	89.7
Number of contigs	15,810	36,205	4,901	16,825	2,442	22,492
Contig N50	30,774	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

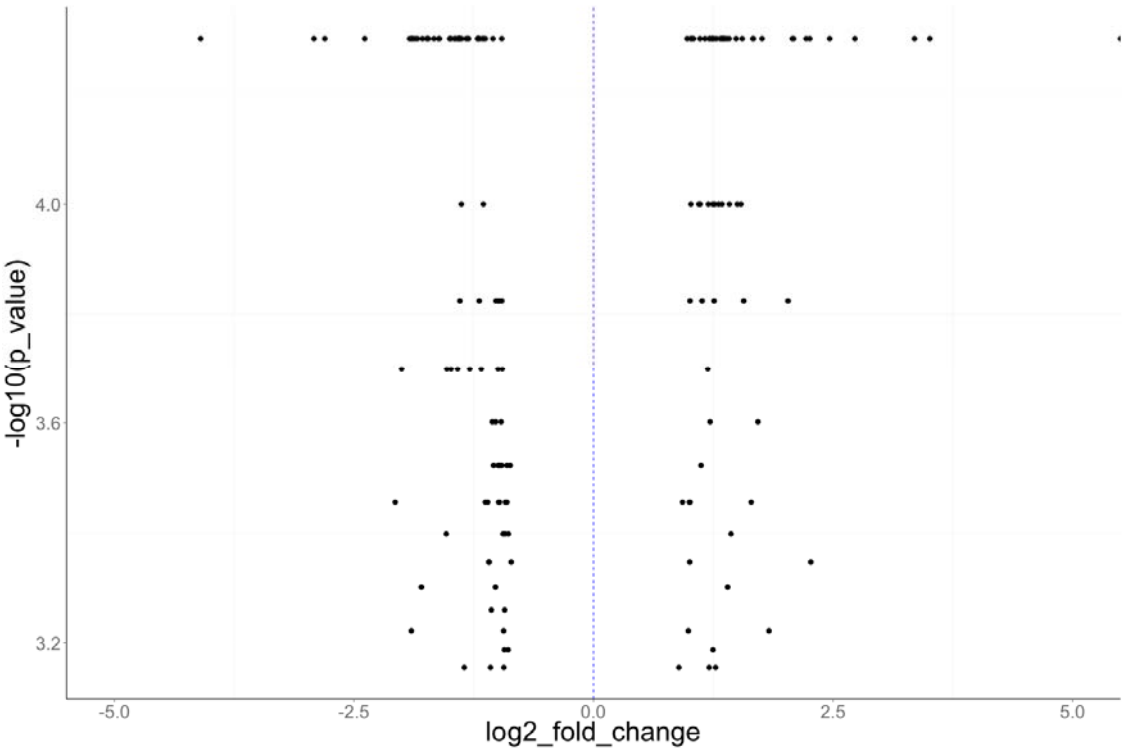

-log10(p_value)
4.0
3.6
3.2
-5.0
-2.5
0.0
2.5
5.0
log2_fold_change

Fig. 2

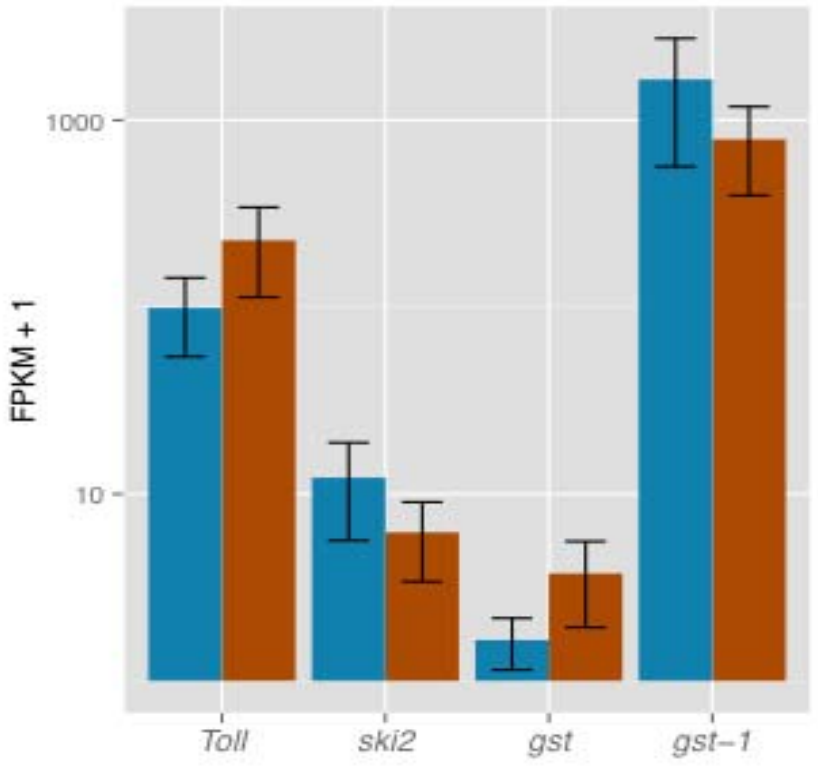
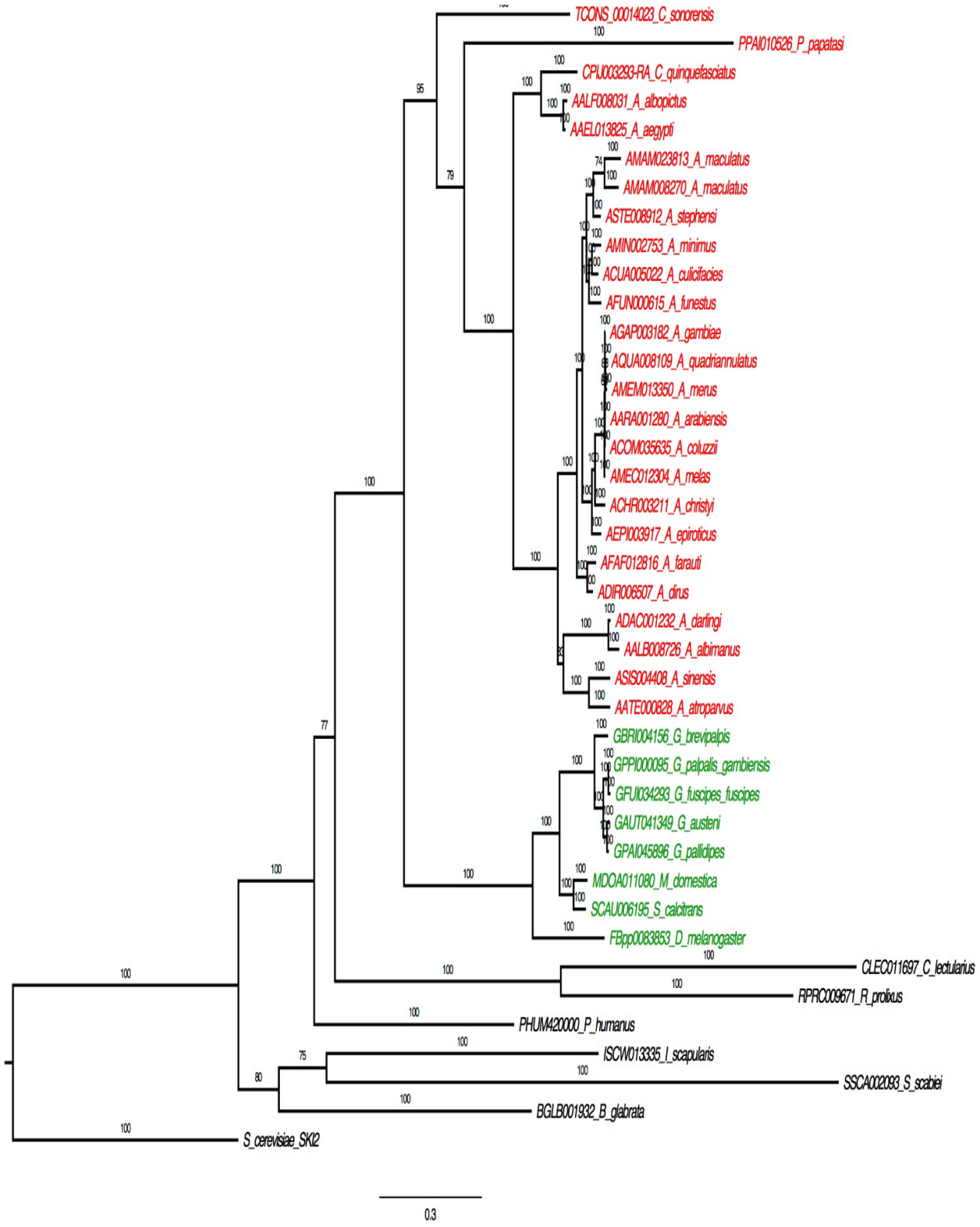
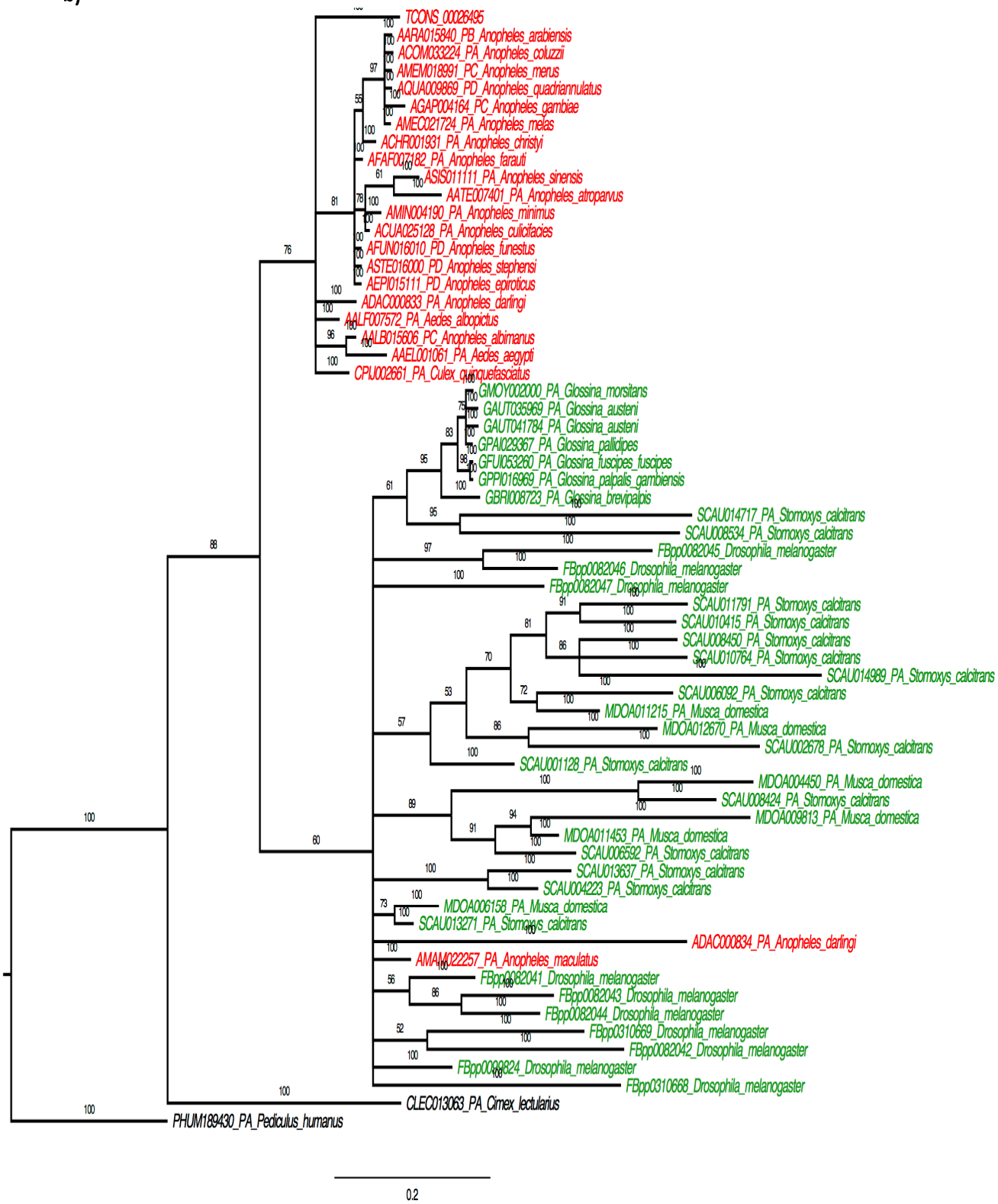


Fig. 3

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
FBgp0083853_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
CPIU002661_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
AMAN022257_PA_Anopheles_maculatus
FBpp0082041_Drosophila_melanogaster
FBpp0082043_Drosophila_melanogaster
FBpp0082044_Drosophila_melanogaster
FBpp0310669_Drosophila_melanogaster
FBpp0082042_Drosophila_melanogaster
FBpp0088824_Drosophila_melanogaster
FBpp0310668_Drosophila_melanogaster
CLEC013063_PA_Cimex_lectularius
PHUM189430_PA_Pediculus_humanus
0.2