

1 **Ultra-conserved sequences in the genomes of highly diverse *Anopheles* mosquitoes,**  
2 **with implications for malaria vector control.**

3

4 Samantha M. O’Loughlin\*, Annie J. Forster\*, Silke Fuchs\*, Tania Dottorini†, Tony  
5 Nolan\*‡, Andrea Crisanti\*, Austin Burt\*.

6

7 \*Department of Life Sciences, Imperial College London, UK.

8 †School of Veterinary Medicine and Science, Sutton Bonington Campus, University of  
9 Nottingham, Leicestershire, UK

10 ‡Liverpool School of Tropical Medicine, Liverpool, UK

11

12 Running title: *Anopheles* ultra-conserved sequences

13 Keywords: *Anopheles*, gene drive, conserved, malaria

14 Corresponding author: S. M. O’Loughlin, Department of Life Sciences, Imperial  
15 College London, Silwood Park Campus, Buckhurst Road, Ascot, SL5 7PY, UK

16

17 **ABSTRACT**

18 DNA sequences that are exactly conserved over long evolutionary time scales have  
19 been observed in a variety of taxa. Such sequences are likely under strong functional  
20 constraint and they have been useful in the field of comparative genomics for  
21 identifying genome regions with regulatory function. A potential new application for  
22 these ultra-conserved elements has emerged in the development of gene drives to  
23 control mosquito populations. Many gene drives work by recognising and inserting at  
24 a specific target sequence in the genome, often imposing a reproductive load as a  
25 consequence. They can therefore select for target sequence variants that provide

26 resistance to the drive. Focusing on highly conserved, highly constrained sequences  
27 lowers the probability that variant, gene drive-resistant alleles can be tolerated.  
28  
29 Here we search for conserved sequences of 18bp and over in an alignment of 21  
30 *Anopheles* genomes, spanning an evolutionary timescale of 100 million years, and  
31 characterise the resulting sequences according to their location and function. Over 8000  
32 ultra-conserved elements were found across the alignment, with a maximum length of  
33 164 bp. Length-corrected gene ontology analysis revealed that genes containing  
34 *Anopheles* ultra-conserved elements were over-represented in categories with structural  
35 or nucleotide binding functions. Known insect transcription factor binding sites were  
36 found in 48% of intergenic *Anopheles* ultra-conserved elements. When we looked at  
37 the genome sequences of 1142 wild-caught mosquitoes we found that 15% of the  
38 *Anopheles* ultra-conserved elements contained no polymorphisms. Our list of  
39 *Anopheles* ultra-conserved elements should provide a valuable starting point for the  
40 selection and testing of new targets for gene-drive modification in the mosquitoes that  
41 transmit malaria.

42

43

## INTRODUCTION

44 DNA sequences that are highly conserved over long evolutionary timescales have been  
45 identified in many organisms. Some of these sequences show complete conservation at  
46 the nucleotide level and are often known as ultra-conserved elements (UCEs).  
47 Originally, UCEs were defined as sequences of at least 200bp that were identical  
48 between human, mouse and rat genomes (Bejerano *et al.* 2004). Subsequently the  
49 search for UCEs has been extended to other vertebrates, insects and plants (e.g.

50 Makunin *et al.* 2013; Siepel *et al.* 2005; Baxter *et al.* 2012), and to sequences of length  
51 50bp or more.

52

53 There are several reasons why UCEs are of interest. First, in the field of comparative  
54 genomics, UCEs are thought to represent functionally important regions. While there  
55 is still some mystery around why sequences might be conserved at the nucleotide level  
56 over long evolutionary timescales, it has been shown that UCEs 1) often are involved  
57 in regulation of transcription of genes, especially essential genes involved in  
58 development (e.g. Visel *et al.* 2008); 2) may have a role in chromosomal structure (e.g.  
59 Chiang *et al.* 2008); and 3) are sometimes non-coding RNA genes (e.g. Kern *et al.*  
60 2015). Even UCEs in protein coding regions may have multi-functional roles  
61 (Warnefors *et al.* 2016). Second, UCEs can act as probes to facilitate genomic  
62 sequencing of non-model organisms using sequence-capture methods (Faircloth *et al.*  
63 2012). Third, alterations in UCEs have been shown to have an association with human  
64 cancers (e.g. Calin *et al.* 2007; Lin *et al.* 2012).

65

66 A new potential role for UCEs has recently emerged in the fight against malaria using  
67 gene-drive mosquitoes (Kyrou *et al.* 2018). *Anopheles* mosquitoes are the vectors of  
68 malaria parasites, and mosquito control has been responsible for much of the recent  
69 success in reduction of malaria cases (78% of the 663 million malaria cases averted  
70 globally since 2000 (Bhatt *et al.* 2015)). Progress in reducing malaria cases has stalled  
71 (WHO 2018), probably in part due to resistance of the mosquitoes against commonly  
72 used pesticides. One novel method under consideration is the development of  
73 mosquitoes containing gene drives that either reduce the population size (Windbichler  
74 *et al.* 2011; Hammond *et al.* 2016) or make them unable to transmit the malaria parasite

75 (Gantz *et al.* 2015). Both methods currently rely on nuclease-based synthetic gene drive  
76 systems that introduce a desired trait at a precise genomic location, spreading it in a  
77 target population at such a rate that outweighs fitness costs associated with the trait  
78 (Burt 2003). The technologies include RNA-guided endonucleases (such as  
79 CRISPR/Cas9) and homing endonucleases (Jinek *et al.* 2012; Windbichler *et al.* 2011).  
80 These enzymes recognise and cleave a particular target size of about 18 bp. When the  
81 sequence coding for these enzymes is engineered into its own target site in the genome  
82 and is expressed in the germline, it creates a double-strand break in the homologous  
83 chromosome. The break will usually be repaired by homology-directed repair using the  
84 drive-containing chromosome as a template which results in conversion of the repaired  
85 to also contain the drive element in greater than the usual 50% inheritance rate among  
86 the gametes. An efficient gene drive can be inherited by almost 100% of progeny  
87 (Hammond *et al.* 2015). Theoretical and laboratory studies have shown that changes to  
88 the recognition site can result in alleles that cannot be recognised or cleaved. If these  
89 alleles confer increased fitness compared to the wild type allele in the presence of the  
90 gene drive they can be expected to spread and retard the spread of the gene drive  
91 (Deredec *et al.* 2008; Unckless *et al.* 2017; Hammond *et al.* 2017). For population  
92 suppression gene drives that are designed to impair essential genes the selection  
93 pressure for resistance alleles to arise is high. These alleles can arise from standing  
94 variation at the target site in a wild population, or may come about from the action of  
95 the endonuclease. This is because non-homologous end joining can sometimes repair  
96 the double-strand break, and random insertions and deletions can be introduced to the  
97 target site.  
98

99 Two of the most important vector species in sub-Saharan Africa are the close relatives  
100 *Anopheles gambiae* and *An. coluzzii*, both of which are highly genetically diverse. A  
101 study of 765 mosquitoes in phase 1 of Ag1000G project, which looked to sample  
102 genetic diversity in the wild through the resequencing of wild caught individuals across  
103 Africa (*Anopheles gambiae* 1000 Genomes Consortium 2017), found a polymorphism  
104 on average every 2.2 bases of the accessible genome. Nucleotide diversity ( $\pi$ ) ranged  
105 from ~0.008 to ~0.015 per population sampled, and even non-degenerate sites (which  
106 are expected to be strongly constrained) had an average  $\pi$  of ~0.0025.

107

108 Proof of principle for retarding the evolution of resistance to nuclease-based gene-drive  
109 by targeting an evolutionarily conserved sequence has recently been demonstrated. A  
110 strain of mosquitoes with a CRISPR/Cas9 gene-drive targeting the *doublesex* gene fully  
111 suppressed laboratory caged populations of *An. gambiae* (Kyrou *et al.* 2018) without  
112 selecting for resistance. The CRISPR/Cas9 target sequence in this strain is an  
113 intron/exon junction that is highly conserved across the *An. gambiae* species complex,  
114 and only one rare single nucleotide polymorphism was found in the sequence in *An.*  
115 *gambiae* and *An. coluzzii* in the Ag1000G data. Consistent with the target site being a  
116 region of high functional constraint, monitoring of potential resistant mutations during  
117 the cage experiment revealed that although some indels had been introduced by the  
118 endonuclease, none of them showed signs of positive selection.

119

120 This strong constraint at the nucleotide level may exist at other loci in *An. gambiae*.  
121 The Ag1000G project looked for conserved putative CRISPR/Cas9 target sites (18  
122 invariant bases followed by the -NGG motif necessary for Cas9 cleavage) in the 765

123 mosquitoes of Phase 1 of the project, and found 5474 genes containing such sequences.

124 However, they note that more variation is likely to be found with further sampling.

125

126 Here we take an approach that is likely to be more stringent in identifying functionally

127 constrained sequences by searching for regions that are ultra-conserved across the

128 whole *Anopheles* genus, which has a most recent common ancestor ~100 million years

129 ago (Neafsey *et al.* 2015). Although sequence constraint across such a long time scale

130 is not necessary for a good target (as indicated by the *doublesex* locus, which is ultra-

131 conserved within the *An. gambiae* species complex, but shows less conservation outside

132 the complex), we are hypothesising that such highly conserved sequences will contain

133 few polymorphisms in the wild *Anopheles gambiae* population, and any

134 polymorphisms that do arise (either spontaneously or due to the action of the

135 endonuclease) are likely to have strong fitness costs. We also do not confine our

136 analysis to sequences compatible with any single nuclease architecture (e.g the 5'-

137 NGG-3' PAM sequence required by the SpCas9 nuclease) since the range and

138 flexibility of nuclease architectures is constantly expanding, meaning that these

139 requirements may be relaxed (Anders *et al.* 2016; Chatterjee *et al.* 2018; Hu *et al.* 2018).

140 We extracted UCEs from an alignment of the genomes of 21 *Anopheles* species and

141 strains that was constructed by the *Anopheles* 16 genomes consortium (Neafsey *et al.*

142 2015). We characterise the UCEs according to their locations in the genome, and use

143 data from *Drosophila* orthologues to group genic UCEs according to potential

144 phenotype. We then use the Ag1000G data (765 *An. gambiae* and *An. coluzzii*) to see

145 whether these conserved elements contain any variation in natural populations of

146 potential target mosquito species.

147

148

## METHODS

### 149 **Data.**

150 Two sources of genomic data were used in this study: a multi-species alignment from  
151 the *Anopheles* 16 genomes project (Neafsey *et al.* 2015) and variation data from phase  
152 1 of the MalariaGEN *An. gambiae* 1000 genomes project (Anopheles gambiae 1000  
153 Genomes Consortium 2017). The *Anopheles* 16 genomes project multi-species  
154 alignment contains reference genomes from 21 *Anopheles* species and strains: *An.*  
155 *gambiae* PEST, *An. gambiae* s.s., *An. coluzzii*, *An. merus*, *An. arabiensis*, *An.*  
156 *quadriannulatus*, *An. melas*, *An. christyi*, *An. epiroticus*, *An. minimus*, *An. culicifaces*,  
157 *An. funestus*, *An. stephensi* S1, *An. stephensi* I2, *An. maculatus*, *An. farauti*, *An. dirus*,  
158 *An. sinensis*, *An. atroparvus*, *An. darlingi*, *An. albimanus*. A description of the methods  
159 used to create the alignment is found in Neafsey *et al.* 2015. Phase 1 of the Ag1000G  
160 project comprises 590 *An. gambiae* and 129 *An. coluzzii* collected from 9 countries in  
161 Africa, plus 46 hybrid individuals from Guinea Bissau. A detailed description of the  
162 samples and data is given in Anopheles gambiae 1000 Genomes Consortium 2017.

163

### 164 **Identifying UCEs.**

165 To identify invariant regions we used only parts of the multi-species alignment where  
166 sequence data was available for all 21 strains. We used Variscan v2.03 (Vilella *et al.*  
167 2005) to find regions of the alignment of 18bp or longer containing no variation. We  
168 mapped the resulting regions back to the PEST reference genome using BWA-aln with  
169 strict mapping parameters (exact parameters can be provided on request; bwa-0.7.10  
170 (Li and Durbin 2010)). Sequences that mapped at multiple places in the genome were  
171 included in the analysis, but flagged as ‘repeat sequences’ as these would not be suitable  
172 for use as CRISPR targets. We used BEDTools (Quinlan and Hall 2010) to classify the

173 genomic location of the UCEs (such as exonic, intronic etc). The AgamP4.12  
174 basefeatures file was used from VectorBase (Giraldo-Calderón *et al.* 2015). Genic  
175 sequences were defined as those with an AGAP gene annotation so include exons,  
176 UTRs and introns. UCEs that partly or wholly fell within genes were classified by us  
177 as genic, and those outside genes were classified as intergenic.

178

179 For comparison, we used the same method to identify invariant sequences of 18bp or  
180 more just in the *An. gambiae* complex species (*An. gambiae PEST*, *An. gambiae s.s.*,  
181 *An. coluzzii*, *An. merus*, *An. arabiensis*, *An. quadriannulatus*, *An. melas*). We also  
182 looked for conservation beyond the *Anopheles* genus by performing a BLAST with  
183 default search parameters (Altschul *et al.* 1990) search of the UCEs against *Culex*  
184 *quinquefasciatus* and *Aedes aegypti* reference genomes. From the BLAST results we  
185 extracted sequences of 18bp or more with no substitutions, insertions or deletions.

186

#### 187 **Random control sequences.**

188 So that we could compare the location of UCEs with non-UCEs we extracted 10  
189 randomly distributed sets of control sequences from the multi-species alignment file  
190 that were matched to give the same number of sequences with the same base-lengths.

191 To compare variation in the Ag1000G data in UCEs and non-UCEs, we also extracted  
192 10 sets of control sequences from the AgamP4 genome but also matching for genic and  
193 intergenic locations.

194

#### 195 **Orthology between species.**

196 For UCEs that fell within genes, we compared the orthology identifiers between  
197 AgamP4 and *An. arabiensis* Dongola references genomes, and between *An. gambiae*



198 PEST and *An. funestus* FUMOS reference genomes. We chose these species because  
199 *An. gambiae* (and its sister species *An. coluzzii*), *An. arabiensis* and *An. funestus* are the  
200 most important malaria vectors in sub-Saharan Africa. *An. gambiae* PEST is a hybrid  
201 strain of *An. gambiae* and *An. coluzzii* (previously known as S and M forms of *An.*  
202 *gambiae*). *An. gambiae* and *An. arabiensis* are closely related (in the same species  
203 complex) and *An. funestus* is more distantly related. Genic UCEs were checked for  
204 orthology between *An. gambiae* and *An. arabiensis* and between *An. gambiae* and *An.*  
205 *funestus*. Coordinates of UCEs were extracted from the multiple-alignment file for *An.*  
206 *arabiensis* and *An. funestus* reference genomes, and annotated with gene names from  
207 the basefeatures files *Anopheles-arabiensis-Dongola\_BASEFEATURES\_AaraD1* and  
208 *Anopheles-funestus-FUMOS\_BASEFEATURES\_AfunF1.3* (from VectorBase).  
209 Orthology identifiers for each gene in each species were found from the  
210 ODBMOZ2\_Anophelinae database at OrthoDb.org (Kriventseva *et al.* 2019).  
211 Orthology identifiers that match between species indicated that the genes were  
212 orthologous. We could not use orthology to directly compare intergenic UCEs, so  
213 instead we identified flanking genes for each intergenic UCE in the reference genome  
214 of each species, and then compared the orthology identifiers for these genes as before.

215

#### 216 **Ontology analysis of genes containing UCEs.**

217 PANTHER software (version 14.0) (Mi *et al.* 2016) was used to categorise the gene  
218 ontology (GO-Slim) terms of the genes containing UCEs. A gene was represented in  
219 the analysis once, regardless of how many UCEs it contained. The genes were clustered  
220 by GO-Slim molecular function, biological process and cellular component terms.

221

222 Because the Panther categorisation does not take into account how much of the genome  
223 is covered by each GO term, we used GSeq (Young *et al.* 2010) to carry out length-  
224 bias corrected gene ontology (GO) enrichment analysis, implemented in Galaxy (Afgan  
225 *et al.* 2018). GSeq corrects for gene length using a Wallenius non-central hyper-  
226 geometric distribution. We used GO-Slim terms extracted from VectorBase (Giraldo-  
227 Calderón *et al.* 2015) for AgamP4.12 gene set. GO terms with a Benjamini-Hochberg  
228 corrected false discovery rate (FDR) of less than or equal to 0.05 were considered over-  
229 represented. We also looked for over-representation of GO-Slim terms in the genes  
230 flanking intergenic UCEs. We were interested to see how our set of UCEs compared to  
231 UCEs from *Drosophila* studies, so as well as our full data set, we also performed the  
232 GO term analysis on a subset of genes that contained at least one UCE over 50bp long,  
233 to make the data comparable.

234

### 235 **Targets for mosquito control.**

236 One form of gene drive aimed at population suppression looks to disrupt essential  
237 mosquito genes and thereby impose a strong reproductive load on the population as it  
238 spreads. UCEs may offer good targets for control of *An. gambiae* by a gene drive  
239 method; if any sequence variation at these sites results in high fitness costs there would  
240 be little selective advantage to a mosquito having the variant allele over the gene drive  
241 allele. We searched the functional annotations of genes containing UCEs to find genes  
242 that may have a suitable function to be targeted for control. Gene descriptions were  
243 obtained from VectorBase (Giraldo-Calderón *et al.* 2015). Gene drives that confer  
244 recessive female sterility are particularly potent since both sexes can transmit the drive  
245 at very high rates to offspring yet only females homozygous for the drive display the  
246 phenotype, which results in a drastic reduction of the population's reproductive

247 capacity (Burt 2003, Burt and Deredec 2008). P-sterile values were available for some  
248 genes from (Hammond *et al.* 2016). P-sterile is a sterility index based on a logistic  
249 regression model that correlates gene expression features in *Anopheles* with the  
250 likelihood that mutations of the gene produce female sterile alleles in the model  
251 dipteran *Drosophila melanogaster* (Baker *et al.* 2011).

252

253 To narrow down the gene list to potential vector control targets, we leveraged the large  
254 amount of phenotype data already available for *Drosophila* mutants. Where possible,  
255 *Drosophila* orthologues were identified for genes containing UCEs (in Vectorbase).  
256 We used ID converter in FlyBase (Gramates *et al.* 2017) to batch convert *Drosophila*  
257 gene identifiers into alleles associated with the genes (FBal numbers). The alleles have  
258 associated phenotype data provided by the research community; we searched for  
259 phenotypes conferring female sterility or recessive lethality.

260

#### 261 **Transcription factor binding site motifs in UCEs.**

262 We used the ‘Find Individual Motif Occurrences’ (FIMO, Grant *et al.* 2011) scanning  
263 module (MEME suite 4.12.0, Bailey *et al.* 2009) to look for transcription factor binding  
264 motifs in UCEs and controls. The UCEs were scanned for known insect transcription  
265 factor binding sites using weighted matrices from the JASPER CORE collection (Insect  
266 position frequency matrices 8th release (2020), Khan *et al.* 2018). The results were  
267 filtered by q-value to account for multiple tests. A cut-off of  $q < 0.05$  was used.

268

#### 269 **Variation at UCE locations in Ag1000G data.**

270 Using the final filtered variant file from phase 2 of the Ag1000G project (described at  
271 <https://www.malariagen.net/data/ag1000g-phase-2-ar1>) we extracted single nucleotide

272 polymorphisms for the UCEs identified above, and for matched non-UCE regions.  
273 Diversity statistics were calculated in scikit-allel v1.3.2 (Miles et al 2020): number of  
274 segregating sites (s), nucleotide diversity ( $\pi$ ) and the neutrality test Tajima's D (Tajima  
275 1989).  
276  
277 Data used in this study are publicly available from the *Anopheles* 16 genomes  
278 consortium and the *Anopheles gambiae* 1000 Genomes project. Data generated in this  
279 study are given in the supplementary tables.

280

## 281 RESULTS

### 282 **Ultra-conserved regions from the multi-species alignment.**

283 Much of the MAF file does not include alignments of all 21 species and strains (see  
284 Table S8 in Neafsey *et al.* 2015). The total number of aligned bases from which we  
285 extracted the UCEs was 17,095,206 (7.4% of the AgamP4 reference genome (Suppl  
286 Table 1). A total of 8338 invariant regions of 18bp or more were identified; 1675 on  
287 chromosome arm 2L, 3015 on chromosome arm 2R, 1375 on chromosome arm 3L,  
288 2188 on chromosome arm 3R and 85 on chromosome X (Table 1; we have also included  
289 the same metrics at different evolutionary timescales for comparison). The longest UCE  
290 was 164bp. Genomic coordinates of the UCEs relative to the *Anopheles gambiae* PEST  
291 reference genome are given in Supplementary table 2. The UCEs were distributed  
292 throughout the chromosomes, but were significantly under-represented on the X  
293 chromosome (See Supplementary figure 1 and Supplementary Table 1). The X  
294 chromosome is already under-represented in the MAF as it was less alignable than other  
295 chromosomes (see Figure 2 in Neafsey et al 2015). It is well established that the X  
296 chromosome shows higher differentiation between species than autosomes (due to

297 ‘Haldanes Rule’ and the ‘Large X effect’) and genomic studies have reinforced this  
 298 observation (Presgraves 2018). However, the under-representation in the MAF is not  
 299 sufficient to explain the paucity of UCEs on the X. In the *Anopheles* genus, the X  
 300 chromosome was observed to have undergone particularly dynamic evolution, with  
 301 chromosome rearrangements at a rate of 2.7 times higher than the autosomes, and a  
 302 significant degree of observed gene movement from X to other chromosomes relative  
 303 to *Drosophila* (Neafsey et al 2015). This dynamic evolution of the chromosome may  
 304 explain why it would be less likely to contain functional sequences that require  
 305 conservation at the nucleotide level.

306

307 Size distributions of the UCEs are shown in Supplementary figure 2. In the autosomal  
 308 genic UCEs there is pattern of a jump in frequency every 3 bases, indicating the  
 309 tendency for runs of ultra-conserved bases to neither start nor end on third codon  
 310 positions in coding regions. UCEs are significantly more AT-rich than random  
 311 sequences (64% and 54% respectively, t-test  $p < 0.001$ ).

312

	2L	2R	3L	3R	X
<i>Gambiae</i> complex					
No. UCEs	452,281	612,824	376,383	498,473	99,561
No. Invariant bases					
within UCEs	15,365,491	21,350,270	12,886,437	17,278,830	3,338,454
<i>Anopheles</i>					
No. UCEs	1,675	3,015	1,375	2,188	85

No. Invariant bases					
within UCEs	45,916	81,186	37,102	59,055	2,299
<hr/> <hr/>					
<i>Anopheles+Aedes</i>					
No. UCEs	278	344	193	293	15
No. Invariant bases					
within UCEs	8,161	10,275	5,499	8,339	456
<hr/> <hr/>					
<i>Anopheles+Culex</i>					
No. UCEs	279	350	202	310	16
No. invariant bases					
within UCEs	8,201	10,184	5,716	8,691	503
<hr/> <hr/>					
<i>Anopheles+Aedes+Culex</i>					
No. UCEs	192	247	133	217	12
No. invariant bases					
within UCEs	5,995	7,579	3,989	6,391	393
<hr/> <hr/>					

313

314 Table 1. Number of ultra-conserved sequences of 18bp or more, and total number of  
 315 invariant sites within these sequences. *Gambiae* complex = 7 species and strains;  
 316 *Anopheles* = 21 species and strains; *Culex* = *Culex quinquefasciatus* reference genome;  
 317 *Aedes* = *Aedes aegypti* reference genome.

318

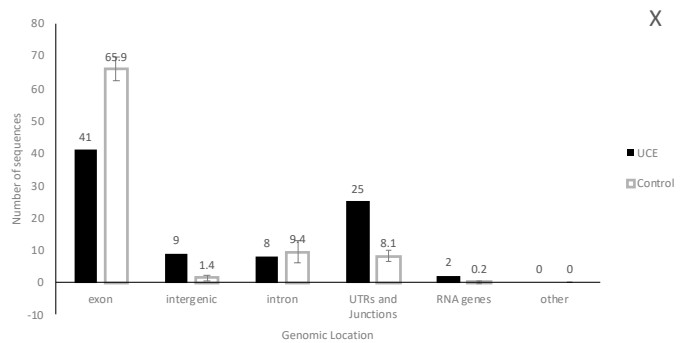
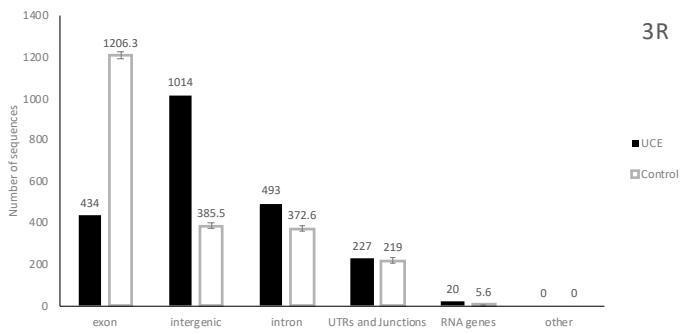
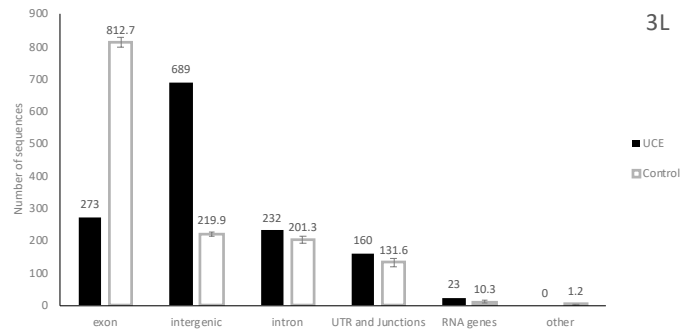
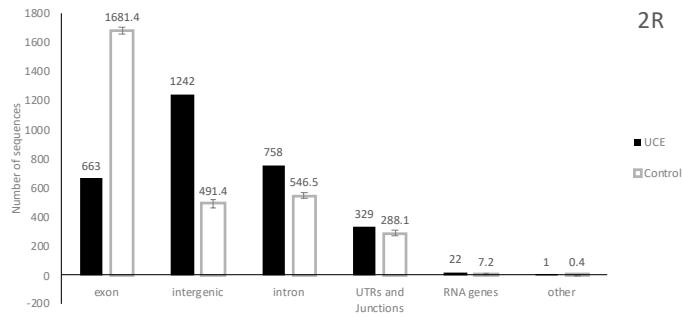
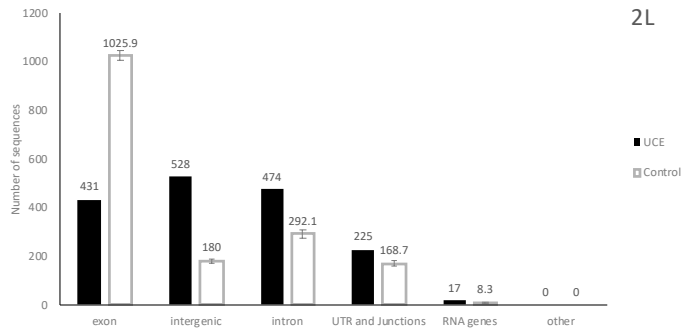
319 We annotated the UCEs in BEDtools to identify where they were found in the genome  
 320 with regards to exons, introns, UTRs, intergenic regions etc (Figure 1). The 21-genome  
 321 aligned parts of the MAF file from which we extracted the UCEs is not representative  
 322 of the reference genome with respect to these features, so we extracted randomly

323 distributed sets of ‘control’ sequences from the MAF, and only from sequences where  
324 all 21 genomes were aligned. These control sequences were matched to give the same  
325 number of sequences with the same base-lengths as the UCEs, and were compared with  
326 the UCE locations to see whether the UCEs were randomly distributed. The UCE  
327 sequences were significantly over-represented (compared to control sequences) in  
328 intergenic regions (42% vs. 15%, ANOVA,  $p < 0.05$ ) and in RNA genes (1% vs. 0.4%,  
329  $p < 0.05$ ), and less frequent in exons (22% vs. 57%,  $p < 0.05$ ). The MAF itself is heavily  
330 skewed to including exonic sequences, as only about 7% of the *An. gambiae* genome  
331 as a whole is exonic (Holt et al. 2002).

332

333

334



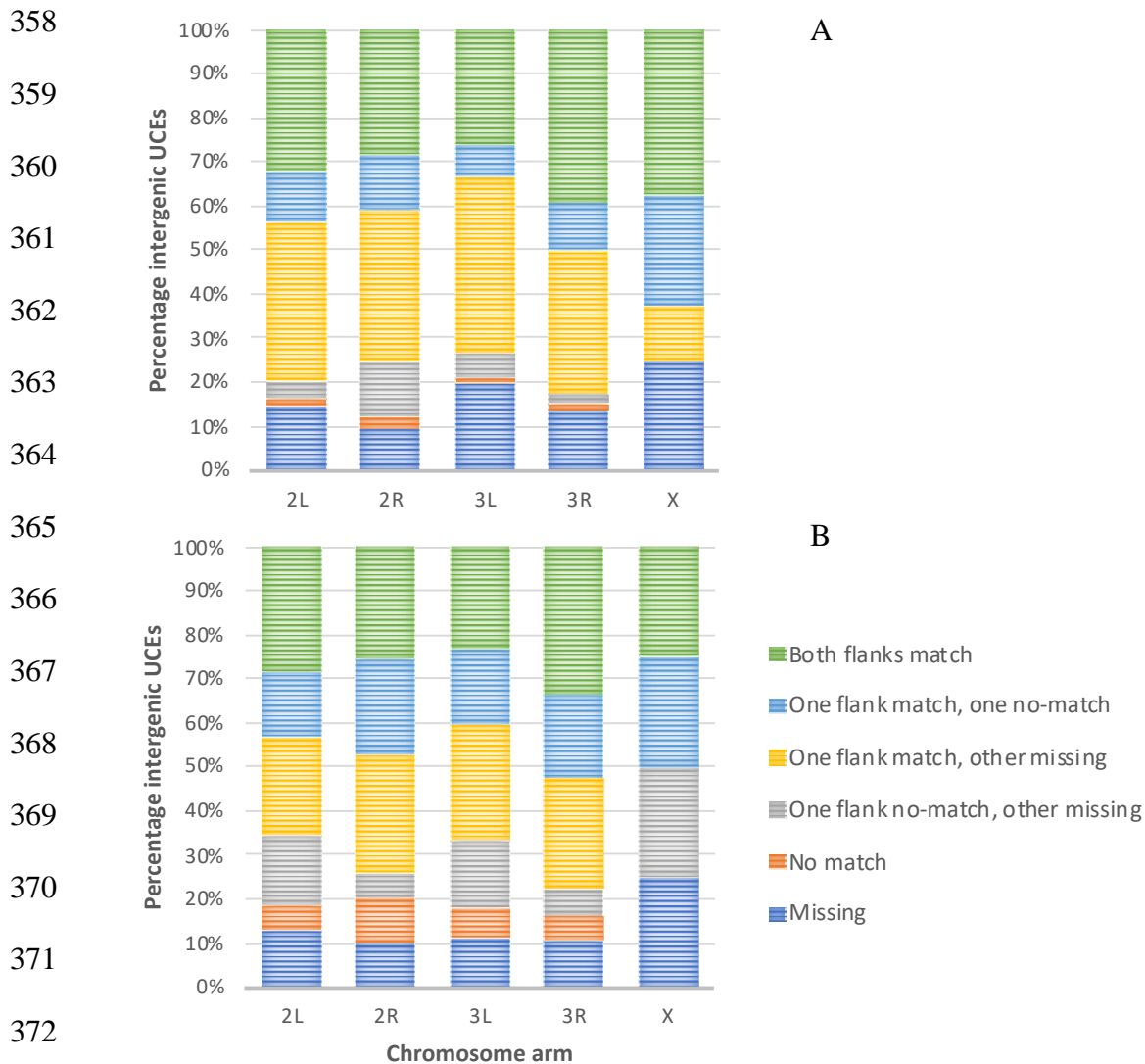


336 **Figure 1. Distribution of UCE and non-UCE control sequences according to**  
337 **genomic location.** Genomic locations annotated with BEDtools. Control error bars =  
338 standard deviation for 10 control data sets of sequences of matched length and number  
339 to the UCEs, extracted randomly from the MAF, only from regions where data for all  
340 21 genomes is present.

341

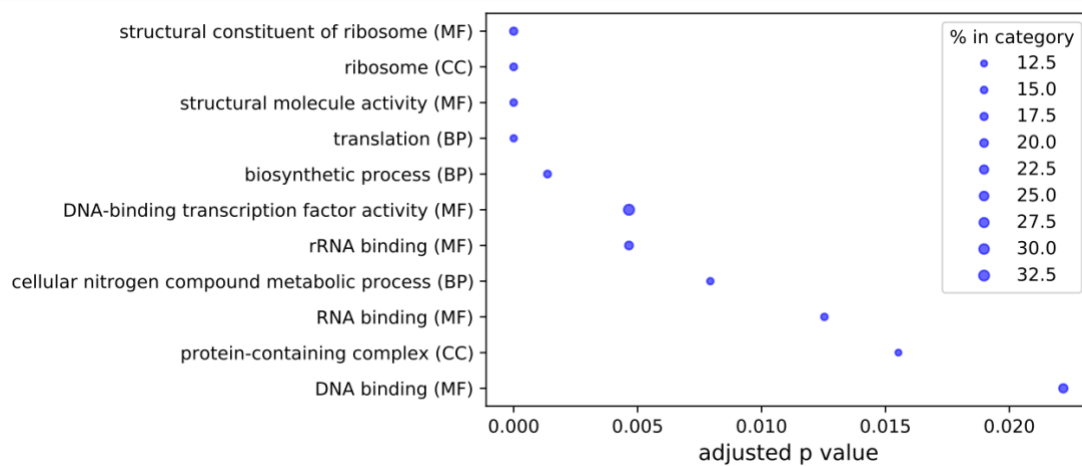
342 **Orthology between important vector species.**

343 To ensure that the UCEs were not falling in sequences that had aligned across the 21  
344 genomes by chance, we checked for orthology between some species in the UCEs. For  
345 UCEs that fell within genes this was done by simply by comparing orthology identifiers  
346 (from OrthDB.org) between *An. gambiae* and *An. arabiensis*, and between *An. gambiae*  
347 and *An. funestus*. For *An. gambiae* and *An. arabiensis*, 94% of autosomal genes  
348 containing UCEs shared orthology. For *An. gambiae* and *An. funestus*, this number was  
349 87%. The amount of orthology fell to 54% and 63% for genes containing UCEs on the  
350 X chromosome. For UCEs that were intergenic, we looked at orthology of the flanking  
351 genes. The results fell into six categories: orthology of both flanking genes, orthology  
352 of one flanking gene with no orthology on the other flank, orthology of one flanking  
353 gene with missing data on the other flank, no orthology on one flank with missing data  
354 on the other flank, missing data on both flanks, and no orthology of either flanking  
355 gene. Ignoring missing data, 92% of intergenic UCEs showed full or half synteny  
356 between *An. gambiae* and *An. arabiensis*, and 77% of UCEs showed full or half synteny  
357 between *An. gambiae* and *An. funestus* (Figure 2).



375 **Figure 2. Number of intergenic UCEs that show synteny between A: *An. gambiae***  
 376 **and *An. arabiensis* and B: *An. gambiae* and *An. funestus*.** The results are shown in  
 377 six categories: matching orthology of both flanking genes, matching orthology of one  
 378 flanking gene with no orthology on the other flank, matching orthology of one flanking  
 379 gene with missing data on the other flank, no orthology on one flank with missing data  
 380 on the other flank, no orthology of either flanking gene, and missing data on both flanks.  
 381  
 382 **Functional profile analysis of the genes containing UCEs via GO term enrichment**

383 Of the 13,796 genes annotated in the *Anopheles gambiae* PEST gene set Agam4.12,  
384 1,601 (12.9%) had at least one UCE. We clustered the genes based on GO-Slim terms  
385 for molecular function, biological process and cellular component (Suppl Figure 3).  
386  
387 Because the clustering does not take into account the amount of the genome covered  
388 by each GO class, we carried out length-bias corrected GO term enrichment analysis.  
389 This showed that certain functional groups were over-represented compared with the  
390 whole *Anopheles* PEST reference gene set (Figure 3).



391  
392 **Figure 3. Goseq GO term enrichment analysis with length-bias correction.** GO-  
393 Slim categories were extracted from the AgamP4.12 gene set. Results are shown for  
394 categories that were enriched with an FDR adjusted p-value<0.05. Bubble size is  
395 relative to the percentage of AgamP4.12 genes in a GO category that were present in  
396 the UCE gene set. MF=molecular function; BP=biological process; CC=cellular  
397 component.

398  
399 In the genes containing UCEs over 50bp long, only 4 categories were over-represented:  
400 transmembrane transporter activity (MF), transmembrane transport (BP), transport

401 (BP) and protein-containing complex (CC), (adjusted p values 0.0047, 0.0047, 0.0272,  
402 0.0272 respectively).

403

404 Genes flanking intergenic UCEs were enriched for the GO-Slim categories DNA  
405 binding (MF), DNA-binding transcription factor activity (MF) and anatomical structure  
406 development (BP) (adjusted p values 4.16E-06, 1.46E-05 and 0.016 respectively).

407

#### 408 **Potential targets for vector control.**

409 AGAP001189 (odorant-binding protein 10) contained the highest number of invariant  
410 bases in UCEs (1215/135306). Nine genes contained UCEs longer than 100bp, of which  
411 3 are annotated as being involved in ion transport. These include the voltage gated  
412 sodium channel gene (VGSC, AGAP004707), which is a target for (and therefore has  
413 a significant role in conferring resistance to) some of the main classes of insecticides  
414 used for malaria vector control. VGSC is one of the most conserved genes we found,  
415 containing 13 UCEs with a total of 507 invariant bases, of which 91% were in exons  
416 and most coded for trans-membrane domains. A total of 357 genes contained 100 or  
417 more invariant bases. A full list of genes containing UCEs is given in Supplementary  
418 table 3.

419

420 Eleven genes containing UCEs had a p-sterile score of greater than 0.5 implying that  
421 they could be good targets to affect female fertility.

422

423 *Drosophila* orthologues were identified for 1309 of the 1601 genes containing UCEs.  
424 Allele and phenotype classes for these genes were extracted from Flybase where  
425 available. For an effective population suppression gene-drive, the target would affect

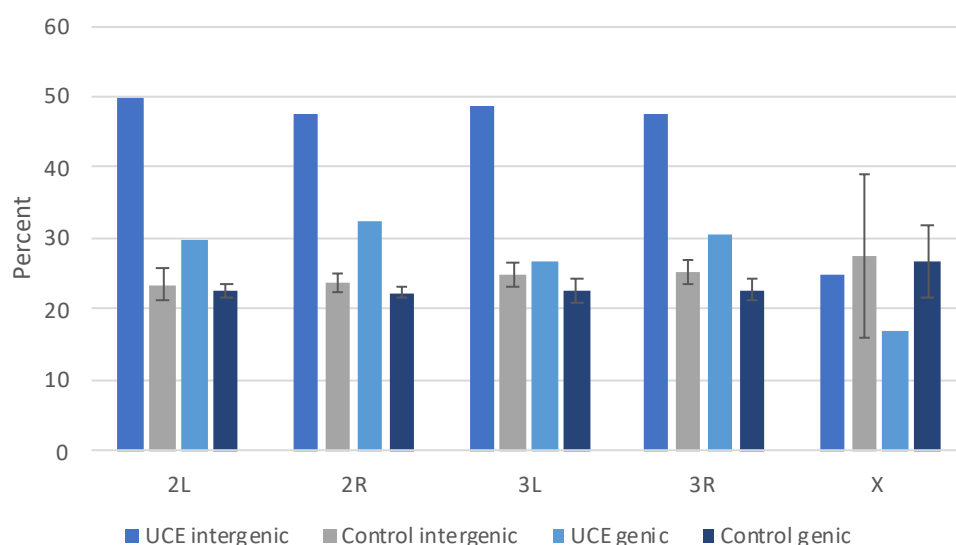
426 female fertility or impose a genetic load as a homozygote, so we extracted UCE  
427 containing genes that have *Drosophila* orthologues annotated with a female sterile term  
428 or a lethal recessive term (shown in Supplementary table 3). 177 genes containing  
429 UCEs have *Drosophila* orthologues with an allele phenotype affecting female fertility,  
430 and 367 genes have *Drosophila* orthologues with an allele conferring a lethal recessive  
431 phenotype.

432

### 433 **Transcription factor binding motifs in UCEs.**

434 DNA binding motifs recognised by transcription factors might be expected to be  
435 constrained and hence enriched for UCEs since this protein:DNA interaction is  
436 sequence-specific. The FIMO search found that 38% of UCEs contained hits for insect  
437 transcription factor binding sites with a q value <0.05 (48% of intergenic and 30% of  
438 genic UCEs). This compared with 23% for control (non-conserved) sequences of the  
439 same number and length. On the X chromosome, where data is sparse (only 8 intergenic  
440 and 75 genic UCEs (Figure 4)), the numbers of transcription factors binding sites were  
441 not significantly different between UCEs and controls.

442



443

444 **Figure 4. Percent of UCEs and control sequences that contain at least one insect**  
445 **transcription factor binding motif.** Control error bars = standard deviation for 10  
446 control data sets. UCEs were searched for known insect transcription factor binding  
447 sites from the JASPER CORE collection (Insect position frequency matrices 8th release  
448 (2020), Khan *et al.* 2018). The results were filtered by q-value to account for multiple  
449 tests. A cut-off of  $q < 0.05$  was used.

450

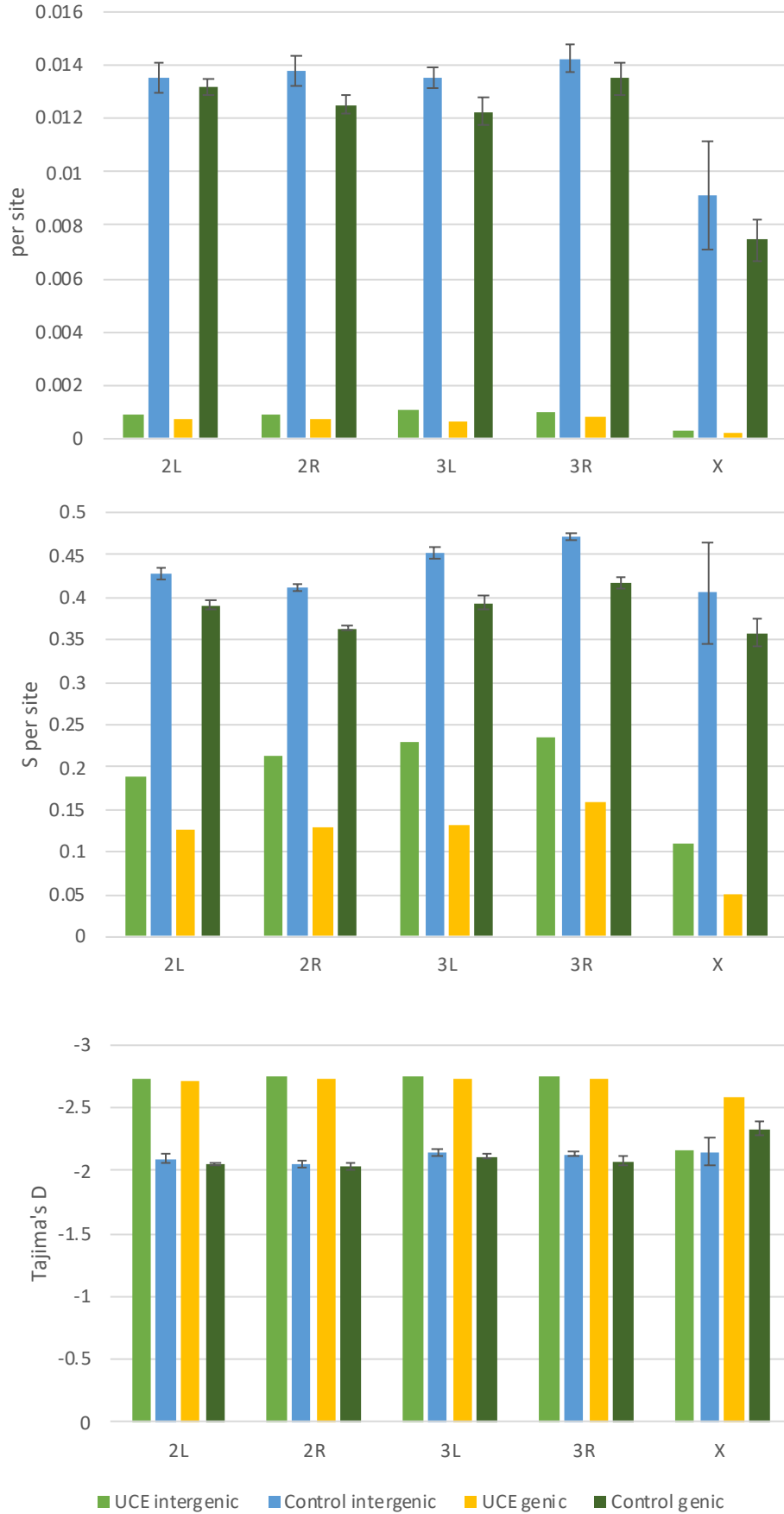
451 **Genetic variation at UCE locations in Ag1000G data.**

452 In order to see whether sequences that are ultra-conserved across the *Anopheles* genus  
453 show variation in wild mosquito populations, we searched for single nucleotide  
454 polymorphisms (SNPs) in the 1142 samples from phase 2 of the Ag1000G project.  
455 There were significantly fewer sites containing polymorphisms in UCEs than control  
456 sequences (Figure 5 middle), and those SNPs that were present were at significantly  
457 lower frequency (Figure 5 top). Of the 8338 UCEs, 1213 (15%) contained no SNPs in  
458 the 1142 samples (229 on 2L, 470 on 2R, 226 on 3L, 259 on 3R and 29 on X). Tajima's  
459 D is significantly more negative for UCEs than controls, with the exception of X  
460 chromosome intergenic sequences (Figure 5 bottom). Negative values of Tajima's D  
461 are expected for sequences under purifying selection.

462

463 The Ag1000G study (*Anopheles gambiae* 1000 genomes consortium 2017), performed  
464 a search within the Phase 1 data to look for potential Cas9 targets (non-overlapping  
465 exonic invariant sequences of 21bp, ending in the 'NGG' motif). They identified 13  
466 genes containing sequences corresponding to this motif. None of these 13 genes  
467 contained UCEs as defined by our study, so these genes also did not overlap with  
468 invariant sequences in Ag1000G data found here. We did not confine our search for

469 UCEs to current Cas9 target site restrictions because of the growing possibility of  
470 relaxation of these constraints. However, for completeness we looked within our final  
471 set of UCEs for the Cas9 motif (18bp followed by -NGG, or CC- followed by 18bp).  
472 We found 1997 (24%) UCEs contained suitable targets for Cas9.  
473  
474





476 **Figure 5. Genetic diversity per chromosome arm in 1142 *Anopheles gambiae* s.l.**  
477 **samples in UCE locations. Top: nucleotide diversity ( $\pi$ ); middle: segregating sites**  
478 **(s); bottom: Tajima's D.** Calculations were made in scikit-allel v1.3.2 (Miles et al  
479 2020).

480

481

## DISCUSSION

482 **Similarities and differences of *Anopheles* UCEs with UCEs from *Drosophila*.**

483 Despite approximately 100 million years since their most recent common ancestor, we  
484 identified in the *Anopheles* genus over 8000 sequences of 18bp or more where there  
485 was no nucleotide variation across the alignment of 21 species and strains. By  
486 coincidence, this is approximately the same span of evolutionary time covered in the  
487 human/mouse/rat data set in which UCEs were originally identified (Bejerano *et al.*  
488 2004). 481 UCEs of more than 200bp were observed between these genomes, but the  
489 longest we found in the *Anopheles* genus was 164bp. This is consistent with previous  
490 reports that UCEs are fewer and shorter in insects (mainly *Drosophila*) than vertebrates  
491 (Makunin *et al.* 2013; Glazov *et al.* 2005). Our criteria for identifying UCEs were  
492 somewhat different than those used previously. First, we only considered sequences  
493 that were present in all 21 species/strains in the alignment; some of these species have  
494 poorly assembled genomes, so this may have reduced the number of UCEs that we  
495 uncovered. Second, we also included invariant stretches of 18bp or more, whereas  
496 *Drosophila* studies have used cut-offs of 50bp (Glazov *et al.* 2005, Warnefors *et al.*  
497 2016), 80bp (Kern *et al.* 2015) or 100bp (Makunin *et al.* 2013). Despite this we see  
498 some similarities between our UCEs and UCEs found in *Drosophila*. UCEs are located  
499 in all parts of the genome and, like *Drosophila*, the majority are found in intergenic  
500 regions and introns. We also found that junction locations (e.g. intron-exon, exon-

501 intergenic etc) are over-represented compared to random sequences, which in  
502 *Drosophila* has been linked to conservation of splice-sites (Glazov *et al.* 2005;  
503 Warnefors *et al.* 2016). Another similarity with *Drosophila* is the high proportion of  
504 genes with the GO terms ‘binding’ and ‘transporter activity’ (Kern *et al.* 2015; Glazov  
505 *et al.* 2005). In *Drosophila*, ion channel/transporter genes have been shown to undergo  
506 extensive RNA editing (Hanrahan *et al.* 2000; Hoopengardner *et al.* 2003; Rodriguez  
507 *et al.* 2012) which is thought to explain the high level of conservation. This is because  
508 RNA adenosine deaminases require double stranded RNA as a substrate, which means  
509 that there is likely to be strong selection at the nucleotide level. The high number of  
510 UCEs in *Anopheles* ion channel/transporter genes suggests a similar mechanism is  
511 responsible for the high conservation in the *Anopheles* genus. However, these genes  
512 are extremely long and are not over-represented in the UCE data when a length-bias  
513 corrected analysis is carried out in GOseq. In the GOseq analysis, the most over-  
514 represented molecular functions are mostly involved in binding or structure.  
515 Transcription factor binding, enzyme binding and nucleic acid binding have also been  
516 shown to be associated with ultra-conservation in both invertebrates and mammals  
517 (Bejerano *et al.* 2004; Glazov *et al.* 2005). A noteworthy addition to highly represented  
518 GO terms in *Anopheles* that has not been reported in *Drosophila*, is the category of  
519 ‘catalytic activity’ genes, although again, these were not over-represented when gene  
520 length was taken into account. When the GO term clustering was carried out on genes  
521 containing UCEs of 50bp or more in length, we found that the category reduced from  
522 28% to 18% suggesting that these shorter ultra-conserved regions most likely code for  
523 a small number of key residues around an active site.  
524

525 The high number of UCEs that we observe in intergenic regions and introns suggests  
526 that we have found numerous unannotated locations in the *Anopheles* PEST reference  
527 genome with putative regulatory functions. At least 70% were syntenic between *An.*  
528 *gambiae*/*An. arabiensis* and *An. gambiae*/*An. funestus* so the location of these highly  
529 conserved sequences is likely to be important. A GOseq analysis of the genes flanking  
530 these intergenic sequences showed significant over-representation of genes with DNA  
531 binding GO terms. Sequences that are ultra-conserved at the nucleotide level across a  
532 long evolutionary time have been shown to be linked to regulatory functions such as  
533 cis-regulation of genes (e.g. enhancers, insulators, silencers) and RNA genes (e.g.  
534 miRNA, snRNA), likely because of the sequence-specific nature of protein:nucleotide  
535 or nucleotide:nucleotide interactions. 19 of the 77 miRNA genes that are annotated in  
536 the *Anopheles* PEST genome were included in our set of UCEs (other miRNAs may  
537 contain ultra-conserved regions that did not meet our criteria). We also found known  
538 insect transcription binding factors in 48% of the intergenic UCEs.

539

#### 540 **Polymorphisms in UCEs in *Anopheles* populations.**

541 All of the UCEs discovered from the alignment of the reference genomes of 21  
542 *Anopheles* species were also found to be highly conserved in the sample of 1142 wild  
543 caught mosquitoes sequenced in phase 1 of Ag1000G. Although the majority of UCEs  
544 contained one or more polymorphisms, they were almost all rare. 1213 UCEs showed  
545 no polymorphisms at all in this sample. This does not rule out the existence of  
546 polymorphisms in the wild populations, but does imply that there may be strong  
547 constraint at a nucleotide level that means alteration of the sequence either naturally or  
548 by the action of a gene-drive may have a strong fitness cost. This would need to be  
549 tested experimentally as different levels of underlying functional constraint may have

550 different fitness costs. For instance, deletion of certain ultra-conserved sequences in  
551 mice gave no discernible fitness cost (Ahituv *et al.* 2007), but a similar experiment in  
552 *Drosophila* showed promise, with 4 out of 11 UCEs with inserted transposons having  
553 a lethal recessive phenotype (Makunin *et al.* 2013). For a resistance-proof gene drive,  
554 selecting target sites that show high levels of conservation is a good starting point, but  
555 the targets would need to be tested under selection pressure to ensure that functional  
556 mutants do not arise.

557

### 558 **UCEs and vector control**

559 UCEs occur within many genes that could have potential for vector control. Nearly 200  
560 genes have *Drosophila* orthologues with an allele phenotype affecting female fertility,  
561 and over three hundred genes have *Drosophila* orthologues with an allele conferring a  
562 lethal recessive phenotype. These phenotypes could both be used for a population  
563 suppression strategy i.e. to reduce the numbers of mosquitoes to a level where malaria  
564 could no longer be transmitted (Deredec *et al.* 2011). More investigation would be  
565 needed to see whether disrupting the genes at the ultra-conserved loci gives the same  
566 phenotype in *Anopheles*. There are also genes that confer recessive phenotypes in  
567 *Drosophila* such as ‘flightless’ or ‘behaviour defective’ that could also be used for  
568 population suppression, or for a population modification type of strategy, where instead  
569 of reducing the mosquito population it is replaced by a strain that cannot transmit  
570 malaria (Carballar-Lejarazú *et al.* 2018). Precise targeting of sequences using  
571 CRISPR/Cas9 gene editing had made testing for these phenotypes feasible.

572

573 Another potential mode of mosquito genetic alteration that has not yet been explored  
574 would be to target sequences involved in gene regulation. Many ultra-conserved

575 sequences in mammals and invertebrates are thought to be involved in regulation of  
576 genes important in development.

577

578 Targeting a sequence that is conserved between species means that the gene drive could  
579 spread between closely related species that hybridise in the wild. For this to happen the  
580 species would need to mate in the wild, produce some fertile offspring, and be able to  
581 express the CRISPR enzyme using the same promoter. Three species (*An. gambiae*, *An.*  
582 *coluzzii* and *An. arabiensis*) are responsible for the majority of malaria transmission in  
583 some parts of sub-Saharan Africa, and are known to hybridise in nature (e.g. Weetman  
584 *et al.* 2014, Anopheles gambiae 1000 Genomes Consortium 2017, Fontaine *et al* 2015).  
585 For effective vector control it would be desirable to be able to reduce or alter all three  
586 species with one construct. The gene drive would not spread to *Anopheles* species that  
587 do not mate in the wild, so would not spread beyond the *Anopheles gambiae* species  
588 complex. However, if a particular target site was proved to be effective for vector  
589 control in *An. gambiae*, a gene drive targeting an orthologous site could be developed  
590 in the laboratory for other important malaria vectors such as *An. funestus*.

591

592

## CONCLUSION

593 Thousands of short genomic regions exist that are conserved across the *Anopheles*  
594 genus. These sequences show many of the same traits as ultra-conserved elements  
595 found in *Drosophila* (such as an association with gene regulation and ion channel  
596 activity). Our list of ultra-conserved elements in the *Anopheles* genus should provide a  
597 valuable starting point for the selection and testing of new targets for gene-drive  
598 modification in the mosquitoes that transmit malaria. Focussing on sequences that have

599 been tightly conserved over long evolutionary time has promise for mitigating against  
600 or slowing the development of resistant alleles in the wild population.

601

## 602 REFERENCES

603

604

605 Afgan, E., Baker, D., Batut, B., *et al.* 2018 The Galaxy platform for accessible,  
606 reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*  
607 46:537-544.

608

609 Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., *et al.*, 2007 Deletion of  
610 ultraconserved elements yields viable mice. *PLoS Biol.* 5(9):234.

611

612 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D.J. 1990 Basic local  
613 alignment search tool. *J. Mol. Biol.* 215:403-410.

614

615 Anders, C., Bargsten, K., Jinek, M. 2016. Structural Plasticity of PAM Recognition by  
616 Engineered Variants of the RNA-Guided Endonuclease Cas9. *Molecular cell*  
617 61(6):895–902.

618

619 *Anopheles gambiae* 1000 Genomes Consortium, Data analysis group, Partner working  
620 group, *et al.*. Genetic diversity of the African malaria vector *Anopheles*  
621 *gambiae*. *Nature.* 2017;552(7683):96-100.

622

623 Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., *et al.* 2009 MEME  
624 SUITE: tools for motif discovery and searching. *Nucleic acids res.* 37:202–208.  
625  
626 Baker, D. A., Nolan, T., Fischer, B., Pinder, A., Crisanti, A., *et al.*, 2011 A  
627 comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector,  
628 *Anopheles gambiae*. *BMC Genomics.* 12:296  
629  
630 Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., *et al.* 2012 Conserved  
631 noncoding sequences highlight shared components of regulatory networks in  
632 dicotyledonous plants. *Plant Cell.* 24(10):3949-65.  
633  
634 Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S.,  
635 Haussler, D., 2004 Ultraconserved elements in the human genome. *Science.*  
636 304(5675):1321-5.  
637  
638 Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., *et al.*, 2015 The effect  
639 of malaria control on *Plasmodium falciparum* in Africa between 2000 and  
640 2015. *Nature.* 526(7572):207-211.  
641  
642 Burt, A., 2003 Site-specific selfish genes as tools for the control and genetic  
643 engineering of natural populations. *Proc Biol Sci.* 270(1518):921-8.  
644  
645 Calin, G. A., Liu, C., Ferracin, M., Hyslop, T., Spizzo, R., *et al.*, 2007 Ultraconserved  
646 regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer*  
647 *Cell.* 12(3):215-229

648

649 Carballar-Lejarazú, R., James, A. A., 2018 Population modification of Anopheline  
650 species to control malaria transmission. *Pathog Glob Health*. 111(8):424-435.

651

652 Chatterjee, P., Jakimo, N., Jacobson, J. M. 2018 Minimal PAM specificity of a highly  
653 similar SpCas9 ortholog. *Science advances*. 4(10).

654

655 Chiang, C. W., Derti, A., Schwartz, D., Chou, M. F., Hirschhorn, J. N., *et al.*, 2008  
656 Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries,  
657 *Genetics*. 180(4):2277-2293

658

659 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., *et al.* 2011 The variant  
660 call format and VCFtools. *Bioinformatics*. 27(15):2156-8.

661

662 Deredec, A., Burt, A., & Godfray, H. C., 2008. The population genetics of using  
663 homing endonuclease genes in vector and pest management. *Genetics*. 179(4), 2013–  
664 2026. <https://doi.org/10.1534/genetics.108.089037>

665

666 Deredec, A., Godfray, H. C., Burt, A., 2011 Requirements for effective malaria control  
667 with homing endonuclease genes. *Proc Natl Acad Sci U S A*. 108(43):874-80.

668

669 Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T.,  
670 *et al.*, 2012 Ultraconserved elements anchor thousands of genetic markers spanning  
671 multiple evolutionary timescales. *Systematic Biology*. 61, 717–726.

672



673 Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov,  
674 I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y. C.,  
675 Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M.  
676 W., & Besansky, N. J. (2015). Mosquito genomics. Extensive introgression in a malaria  
677 vector species complex revealed by phylogenomics. *Science (New York,*  
678 *N.Y.)*, 347(6217), 1258524. <https://doi.org/10.1126/science.1258524>  
679  
680 Gantz, V. M., Jasinskiene, N., Tatarenkova, O., *et al.*, 2015 Highly efficient Cas9-  
681 mediated gene drive for population modification of the malaria vector mosquito  
682 *Anopheles stephensi*. *Proc Natl Acad Sci U S A.* 112(49):6736-43.  
683  
684 Giraldo-Calderón, G. I., Emrich, S. J., MacCallum, R. M., Maslen, G., Dialynas, E., *et*  
685 *al.* 2014 VectorBase: an updated bioinformatics resource for invertebrate vectors and  
686 other organisms related with human diseases. *Nucleic Acids Res.* 43(Database  
687 issue):707-13.  
688  
689 Glazov, E. A., Pheasant, M., McGraw, E. A., Bejerano, G., Mattick, J. S., 2005  
690 Ultraconserved elements in insect genomes: a highly conserved intronic sequence  
691 implicated in the control of homothorax mRNA splicing. *Genome Res.* 15(6):800-8.  
692  
693 Gramates, L. S., Marygold, S. J., Santos, G. D., Urbano, J-M., Antonazzo, G., *et al.*  
694 2016 FlyBase at 25: looking to the future. *Nucleic Acids Res.* 45(D1):663-D671.  
695  
696 Grant, C. E., Bailey, T. L., Noble, W. S. 2011 FIMO: scanning for occurrences of a  
697 given motif. *Bioinformatics.* 27(7):1017–1018.

698

699 Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., *et al.*, 2015 A  
700 CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito  
701 vector *Anopheles gambiae*. *Nat Biotechnol.* 34(1):78-83.

702

703 Hammond, A., Kyrou, K., Bruttini, M., North, A., Galizi, R., *et al.*, 2017 The creation  
704 and selection of mutations resistant to a gene drive over multiple generations in the  
705 malaria mosquito. *PLoS Genetics* 13(10).

706

707 Hanrahan, C. J., Palladino, M. J., Ganetzky, B., Reenan, R. A., 2000 RNA editing of  
708 the *Drosophila* para Na(+) channel transcript. Evolutionary conservation and  
709 developmental regulation. *Genetics.* 155(3):1149-60.

710

711 Hoopengardner, B., Bhalla, T., Staber, C., Reenan, R., 2003 Nervous system targets of  
712 RNA editing identified by comparative genomics. *Science.* 301:832–836.

713

714 Hu, J. H., Miller, S. M., Geurts, M. H., Tang, W., Chen, L., *et al.* 2018 Evolved Cas9  
715 variants with broad PAM compatibility and high DNA specificity. *Nature*  
716 556(7699):57–63.

717

718 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., Charpentier, E., 2012  
719 A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial  
720 immunity. *Science.* 337(6096):816-21.

721

722 Kern, A. D., Barbash, D. A., Chang Mell, J., Hupaló, D., Jensen, A., 2015 Highly  
723 constrained intergenic *Drosophila* ultraconserved elements are candidate  
724 ncRNAs. *Genome Biol Evol.* 7(3):689-98.

725

726 Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., *et al.* 2017  
727 JASPAR 2018: update of the open-access database of transcription factor binding  
728 profiles and its web framework. *Nucleic Acids Res.* 46(1):260-D266.

729

730 Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., *et al.*, 2019  
731 OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral  
732 genomes for evolutionary and functional annotations of orthologs. *Nucleic acids*  
733 *research*, 47(1), 807–D811

734

735 Kyrou, K., Hammond, A., Galizi, R., Kranjc, N., Burt, A., *et al.*, 2018 A CRISPR–Cas9  
736 gene drive targeting doublesex causes complete population suppression in caged  
737 *Anopheles gambiae* mosquitoes. *Nature Biotech.* 36 (1062)

738

739 Li, H., and Durbin, R., 2010 Fast and accurate long-read alignment with Burrows-  
740 Wheeler Transform. *Bioinformatics*, Epub. [PMID: 20080505]

741

742 Lin, M., Eng, C., Hawk, E. T., Huang, M. S., Lin, J., *et al.*, 2012 Identification of  
743 polymorphisms in ultraconserved elements associated with clinical outcomes in locally  
744 advanced colorectal adenocarcinoma. *Cancer* 118(24):6188-6198

745

746 Makunin, I. V., Shloma, V. V., Stephen, S. J., Pheasant, M., Belyakin, S. N., 2013  
747 Comparison of ultra-conserved elements in drosophilids and vertebrates. PLoS One.  
748 8(12)  
749  
750 Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., *et al.* 2016 PANTHER  
751 version 11: expanded annotation data from Gene Ontology and Reactome pathways,  
752 and data analysis tool enhancements. Nucleic Acids Res. 45(1):183-D189.  
753  
754 Miles, A., pyup.io bot, Murillo R., Ralph, P., Harding, N., Pisupati, R., Millar, T., 2020  
755 cggh/scikit-allel: v1.3.2 (Version v1.3.2). Zenodo.  
756 <http://doi.org/10.5281/zenodo.3976233>  
757  
758 Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, M. A., Alekseyev, M. A.,  
759 *et al.*, 2015 Highly evolvable malaria vectors: the genomes of 16 *Anopheles*  
760 mosquitoes. Science. 347(6217):1258522.  
761  
762 Presgraves, D. C., 2018 Evaluating genomic signatures of “the large X-effect” during  
763 complex speciation. Mol Ecol. 27: 3822– 3830. <https://doi.org/10.1111/mec.14777>  
764  
765 Quinlan, A. R., Hall, I. M., 2010 BEDTools: a flexible suite of utilities for comparing  
766 genomic features. Bioinformatics. 26(6):841-2.  
767  
768 Rodriguez, J., Menet, J. S., Rosbash, M., 2012 Nascent-seq indicates widespread  
769 cotranscriptional RNA editing in *Drosophila*. Mol Cell. 47(1):27-37.  
770

771 Siepel, A., Bejerano, G., Pedersen, J. S., et al. 2005 Evolutionarily conserved elements  
772 in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15(8):1034-50.

773

774 Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA  
775 polymorphism. *Genetics.* 1989 Nov;123(3):585-95. PMID: 2513255; PMCID:  
776 PMC1203831.

777

778 Unckless, R. L., Clark, A. G., Messer, P. W., 2016 Evolution of resistance against  
779 CRISPR/Cas9 gene drive. *Genetics.* 205(2):827-841.

780

781 Villela, A., Blanco-Garcia, A., Hutter, S., Rozas, J., 2005 VariScan: Analysis of  
782 evolutionary patterns from large-scale DNA sequence polymorphism data.  
783 *Bioinformatics.* 21(11):2791-2793.

784

785 Warnefors, M., Hartmann, B., Thomsen, S., Alonso, C. R., 2016 Combinatorial gene  
786 regulatory functions underlie ultraconserved elements in *Drosophila*. *Mol Biol Evol.*  
787 33(9):2294-306.

788

789 Weetman, D., Steen, K., Rippon, E. J., Maweje, H. D., Donnelly, M. J., Wilding, C.  
790 S., 2014 Contemporary gene flow between wild *An. gambiae s.s.* and *An.*  
791 *arabiensis*. *Parasit Vectors.* 7:345.

792

793 WHO, World malaria report 2018. Geneva: World Health Organization; 2018.  
794 <https://www.who.int/malaria/publications/world-malaria-report-2018/en/>

795

796 Windbichler, N., Menichelli, M., Papathanos, P. A., Thyme, S. B., Li, H., Ulge, U. Y.,  
797 Hovde, B. T., Baker, D., Monnat, R. J., Burt, A., Crisanti, A., 2011 A synthetic homing  
798 endonuclease-based gene drive system in the human malaria mosquito. *Nature*.  
799 12;473(7346):212-5.

800

801 Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A., 2010 Gene ontology  
802 analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2), R14.

803

804

#### 805 *Authors' contributions*

806 SO and AB jointly devised the study; SO performed and guided data analysis and wrote  
807 the manuscript; AF carried out data analysis; SF designed and performed preliminary  
808 GO term analysis; TD assisted in bioinformatics; TN and AC gave advice on analysis;  
809 all authors provided editorial comments and read and approved the final manuscript.

810

#### 811 *Acknowledgements*

812 Many thanks to the Anopheles 16 genomes consortium and MalariaGEN for permission  
813 to use the 21 species alignment data and the Ag1000G variation data. Thanks also to  
814 Howard Lewis for providing several custom data handling scripts.

815

816

817

818