



## 12 **Abstract**

13 Hybridization plays a central role during the evolution of species boundaries, but the  
14 relative impact of gene flow on genomic divergence and vice versa remains largely  
15 unknown. The genome architecture of populations and emerging species exhibiting various  
16 levels of divergence along the speciation continuum should provide insights into the events  
17 that promote or prevent speciation. In this work, we have used a combination of population  
18 genomic approaches to examine the genomic signatures of hybridization between  
19 *Anopheles nili sensu stricto* and *An. ovengensis*, two malaria mosquitoes that have split ~3-  
20 Myr ago. Despite this substantial time since divergence, the two species hybridize  
21 extensively in nature, giving rise to a unique population of differentiated hybrids in contact  
22 zones. Using genomic clines and Bayesian models, we showed that signatures of  
23 introgression are widespread across the genome suggesting that recent hybridization  
24 between *Anopheles nili sensu stricto* and *An. ovengensis* involves multiple fitness traits and  
25 functional classes. Linkage Disequilibrium analyses allowed us to identified a block of 39  
26 linked loci that segregated between hybrids and parental species and may harbour genes  
27 responsible for reproductive isolation. Our results demonstrate that genome-wide  
28 admixtures can persist in the face of species divergence over long periods of time during  
29 speciation due to increased gene flow at loci providing selective advantage.

## 30 **Introduction**

31 Most cases of speciation occur in a gradual manner so that emerging species continue to  
32 mate and exchange genes before the onset of complete reproductive isolation (Nosil 2012;  
33 Nadeau *et al.* 2013). As a result, secondary contacts between diverging lineages are  
34 pervasive in nature and can lead to several scenarios including extensive hybridization and  
35 the creation of a new hybrid species or genetic homogenization (James 2007; Ozerov *et al.*  
36 2016). The role of hybridization either in the creation of biodiversity through speciation or  
37 in its reduction through genetic homogenization has been under increasing scrutiny over  
38 the last decades and has been recognized as one major force driving the evolution of plant  
39 and animal species (James 2007; Abbott *et al.* 2013; Mallet *et al.* 2015). However, although  
40 there is consistent evidence that hybridization can provide the raw material for  
41 evolutionary diversification (Grant 2015), examples of homoploid hybrid species (where  
42 the hybrid offspring has the same ploidy level as the two parental species) are rather rare in  
43 nature (James 2007; Abbott *et al.* 2013). Most often, hybrids have reduced fitness relative to  
44 parental populations and barely persist as a reproductively isolated unit in the wild.

45  
46 It is particularly difficult to predict the outcome of secondary contacts between diverging  
47 lineages and specifically the rate of hybridization (Payseur & Rieseberg 2016). One main  
48 issue is that the level of divergence between species supposed to encourage or prevent  
49 hybridization remains obscure. Until a recent past, estimates of the population  
50 differentiation parameter  $F_{ST}$  have been used to predict the rate of hybridization between  
51 incomplete species. Variations in  $F_{ST}$  among groups of related species were often

52 interpreted as indicative of differences in rates of gene flow across populations, high  $F_{ST}$   
53 values being considered as evidence of limited gene flow. However, instances of  
54 hybridization and introgression between highly divergent taxa are accumulating, thereby  
55 indicating that estimates of population differentiation are particularly poor predictors of  
56 hybridization between two species (Nydam & Harrison 2011; Roux *et al.* 2013; Parchman *et*  
57 *al.* 2013; Martin *et al.* 2013; Canestrelli *et al.* 2014).

58  
59 Likewise, genomic regions showing high differentiation among populations have been  
60 thought to harbor hybrid incompatibility or other reproductive isolation (RI) factors. In  
61 support of this view, genome scans in many couples of incipient species have identified  
62 genomic regions of extreme  $F_{ST}$  values qualified as “speciation islands”, considered to be  
63 enriched in RI factors and resistant to gene flow (Turner *et al.* 2005; Hohenlohe *et al.* 2010;  
64 Lawniczak *et al.* 2010; Ellegren *et al.* 2012). Yet, not only cases where genes involved in RI  
65 have been explicitly identified in  $F_{ST}$  outliers regions of the genome are extremely rare, but  
66 also extensive introgression has been found to sometimes coincide with regions of high  
67 differentiation. For example, (Parchman *et al.* 2013) used genomic clines and Bayesian  
68 models to study the correlation between regions of exceptional introgression and genetic  
69 differentiation manakins birds (*Manacus*). Contrary to expectations, they found that loci of  
70 strong introgression were relatively positively correlated to the genetic differentiation. The  
71 most plausible interpretation of this intriguing result is that divergent selection may  
72 promote introgression of select genes or genomic regions, resulting in differential patterns  
73 of introgression across the genome. A relative concordance between locus-specific

74 divergence and locus-specific measures of introgression has also been described in the  
75 *Lycaeides* butterfly (Gompert *et al.* 2012, 2013) and in *Drosophila melanogaster* (Pool *et al.*  
76 2012). Overall, the relationship between introgressive hybridization and genetic divergence  
77 at the species or the genome level is more complex than previously thought. Whether  
78 exceptionally differentiated regions of the genome harbor genes that can, paradoxically,  
79 introgress easily is an empirical question that remains opened and should be answered by  
80 studying closely related species at hybrid zones.

81  
82 Mosquitoes of the genus *Anopheles* are ideal system in which to address the evolutionary  
83 processes at hybrid zones due to the prevalence of adaptive speciation (Ndo *et al.* 2013;  
84 Neafsey *et al.* 2015; Kamdem *et al.* 2016). In this paper, we focused on *An. nili*, a group of  
85 African malaria vectors characterized by a reticulate evolution leading to complex  
86 phylogenies that have been challenging to clarify (Kengne *et al.* 2003; Awono-Ambene *et al.*  
87 2004, 2006; Ndo *et al.* 2010, 2013; Peery *et al.* 2011; Sharakhova *et al.* 2013). The group  
88 harbour four known species identified by slight morphological differences: *An. nili sensu*  
89 *stricto*, *An. ovengensis*, *An. carnevalei*, and *An. somalicus*. *An. nili sensu stricto* (thereafter *An.*  
90 *nili*) is the most widespread across the continent while the three other species are more  
91 patchily distributed primarily in the equatorial rainforest. *An. nili* and *An. ovengensis* are  
92 important vectors of human malaria parasites (Antonio-Nkondjio *et al.* 2006). We sampled  
93 populations across the ranges of the four species of the *An. nili* group in Cameroon and we  
94 used a population resequencing approach to develop genome-wide SNP markers that we  
95 genotyped in 145 individuals. Our first aim was to clarify the evolutionary relationships

96 between populations. We discovered new cryptic species within *An. ovengensis* and *An. nili*  
97 as well as a hybrid subgroup resulting from massive hybridization in sympatric areas.  
98 Finally, we took advantage of the recent implementation of genomic cline models that  
99 enables the investigation of footprints of introgression across genomes of non-model  
100 organisms without the need of a high-quality reference genome (Gompert & Alex 2011;  
101 Gompert & Buerkle 2012), to critically evaluate the interplays between the locus-specific  
102 divergent selection and introgression.

## 103 **Materials and methods**

### 104 **Mosquito sample**

105 We surveyed a total of 28 geographic locations that were representative of the main  
106 habitats of species of the *An. nili* group previously described in Cameroon (Fig 1A) (Awono-  
107 Ambene *et al.* 2004, 2006; Antonio-Nkondjio *et al.* 2009; Ndo *et al.* 2010, 2013). The  
108 different species were identified using reference morphological identification keys and a  
109 diagnostic PCR that discriminates the four currently known members of the *An. nili* group  
110 on the basis of a point mutation of the ribosomal DNA (Gillies & De Meillon 1968; Gillies &  
111 Coetzee 1987; Kengne *et al.* 2003).

### 112 **Library preparation, sequencing and SNP discovery**

113 We created double-digest RAD (ddRAD) libraries using a modified version of the protocol  
114 described by Peterson *et al.*, 2012 (Peterson *et al.* 2012). Genomic DNA of mosquitoes was  
115 extracted using the DNeasy Blood and Tissue kit (Qiagen) and the Zymo Research MinPrep  
116 kit on larvae and adult samples respectively. Approximately 50 ng (10µl) of DNA of each  
117 mosquito sample was digested simultaneously with *MluC1* and *NlaIII* restriction enzymes.  
118 Digested products were then ligated to adapter and barcode sequences to enable the unique  
119 identification of the individual associated with each sequencing read. The samples were  
120 pooled, purified, and 400-bp fragments selected. The resulting libraries were then amplified  
121 via PCR and purified. The distribution of library fragment-size was checked on a  
122 BioAnalyzer (Agilent Technologies, Inc., USA). The PCR products were quantified and  
123 diluted for sequencing on Illumina HiSeq2000 (Illumina Inc., USA) to yield single-end reads  
124 of 101 bp.

## 125 **Bioinformatics pipeline**

126 The *An. nili* Dinderesso draft genome assembly comprises 51,048 contigs, varying between  
127 100 and 26775kb in length, to which short reads can be aligned and SNP called. However,  
128 members of the *An. nili* group found in Cameroon diverged ~0.2-6 million years ago (Ndo *et*  
129 *al.* 2013). Thus, alignments of our samples to *An. nili* Dinderesso reference genome might be  
130 subjected to an important bias associated with the inconsistent mapping of reads from  
131 highly divergent populations. To make sure that this potential reference sequence bias  
132 didn't undermined our analyses, we compared our results across two SNP sets that were  
133 identified within RAD loci created using two distinct approaches: a *de novo* assembly and an  
134 assembly of reads aligned onto the reference genome. Recent studies using RAD sequencing  
135 on *Heliconius* butterflies showed that combining *de novo* assemblies and reference  
136 alignments provided a robust approach to perform rigorous test on introgression and  
137 phylogenetic relationships in distantly related species (The Heliconius Genome Consortium  
138 2012). The *process\_radtags* program of the Stacks v 1.35 pipeline was used to demultiplex  
139 and clean raw reads. Reads without the *NlaIII* restriction site and those bearing ambiguous  
140 barcode sequences or having low-quality score (average Phred score < 33) were discarded.  
141 Reads were trimmed to 96-bp by removing index and barcode sequences. We aligned the  
142 short reads to the *An. nili* Dinderesso draft genome assembly using Gsnap (Wu & Nacu  
143 2010) with a maximum of five nucleotide mismatches allowed. The *ref\_map.pl* and  
144 *denovo\_map.pl* programs in Stacks were used to identify consensus RAD loci and to call  
145 SNPs within these loci across our populations using respectively the Gsnap-aligned SAM  
146 files or the individual fastq files as input. For both analyses, we set the minimum number of

147 reads required to form a stack to 3. In the *denovo* assembly, we allowed a maximum of three  
148 mismatches when creating loci in every individual (M parameter in *denovo\_map.pl*) and two  
149 mismatches when building the catalogue of loci across individuals (n parameter). In  
150 reference-based assembly, we specified n = 2 in *ref\_map.pl* to allow two mismatches during  
151 catalogue creation. We generated SNP files in different formats for further downstream  
152 analyses using the *populations* program of Stacks and PLINK v1.09 (Purcell *et al.* 2007).

### 153 **Population genetic structure**

154 We analyzed the genetic structure of *An. nili s.l.* populations using Principal Component  
155 Analysis (PCA) and Neighbor-Joining trees (NJ). We also examined ancestry proportions  
156 and admixtures between populations in ADMIXTURE v1.23 (Alexander *et al.* 2009) and  
157 STRUCTURE v2.3.4 (Pritchard *et al.* 2000). We performed these tests using filtered SNPs  
158 identified in RAD loci present in every population and in at least 50% of individuals by the  
159 *populations* program in Stacks. We used the R package *adegenet* (Jombart 2008) to  
160 implement the PCA. Neighbor-Joining trees were generated from matrixes of Euclidian  
161 distance computed from allele frequencies at genome-wide SNPs using the R package *ape*  
162 (Paradis *et al.* 2004). We ran ADMIXTURE with 10-fold cross-validation for values of k from  
163 1 through 20. We analyzed patterns of ancestry from k ancestral populations in  
164 STRUCTURE, testing five replicates of k = 1-10. We used 200000 iterations and discarded  
165 the first 50000 iterations as burn-in for each STRUCTURE run. CLUMPP v1.1.2 (Jakobsson &  
166 Rosenberg 2007) was used to summarize assignment results across independent runs and  
167 DISTRUCT v1.1 (Rosenberg 2004) to plot STRUCTURE results. To identify the optimal  
168 number of genetic clusters in our samples, we used simultaneously the lowest cross-

169 validation error in ADMIXTURE, the ad-hoc statistic deltaK (Evanno *et al.* 2005; Earl &  
170 VonHoldt 2012) and the Discriminant Analysis of Principal Component (DAPC) method in  
171 *adegenet*.

## 172 **Population genomics analyses**

173 To quantify the genetic differentiation between species and putative subgroups within  
174 species, we used a subset of 1000 high-quality SNPs to calculate the overall pairwise  
175 differentiation index  $F_{ST}$  (Weir & Cockerham 1984) in Genodive v 1.06 (Meirmans & Van  
176 Tienderen 2004). Statistical significance was assessed with 10000 permutations. To further  
177 examine the genomic footprints of selection and introgression, we used ANGSD v 0.612  
178 (Korneliussen *et al.* 2014) and ngsTools (Fumagalli *et al.* 2014) to derive locus-specific  
179 estimates of the nucleotide diversity (measured as  $\theta_w$  and  $\theta_\pi$ ), Tajima's  $D$ , absolute  
180 sequence divergence ( $d_{xy}$ ) and  $F_{ST}$  across 9622 sites using Gnap alignments without SNP  
181 calling. To identify statistical outliers of  $F_{ST}$ , we used the software LOSITAN which applies  
182 the coalescent simulation method (FDIST2) (Beaumont & Nichols 1996) to identify loci with  
183 exceptionally high  $F_{ST}$  values relative to a "neutral" genome-wide  $F_{ST}$  value expected under  
184 neutral evolution. We ran LOSITAN using the neutral  $F_{ST}$  option, 50,000 simulations and a  
185 false discovery rate of 0.05%. Finally, we used the R package LDna (Kemppainen *et al.*  
186 2015) for identifying clusters of linkage disequilibrium in our samples. This analysis was  
187 conducted primarily to assess the possible influence of inversion polymorphism on the  
188 genomic architecture of divergence and hybridization. Precisely, we wished to test if loci  
189 with great significance in the genetic divergence and/or introgression were independent  
190 loci scattered throughout the genome or clusters of linked loci encapsulated in low

191 recombination regions like chromosomal inversions. We used PLINK to calculate the  
192 linkage disequilibrium (estimated as the  $r^2$  correlation coefficient) between all pairs of SNPs  
193 and we applied the graph-based method implemented in LDna to search for clusters of  
194 strongly correlated SNPs.

### 195 **Tests of introgression**

196 In addition to the interpretation of patterns of ancestry provided by clustering analyses in  
197 STRUCTURE and ADMIXTURE, we conducted formal tests to infer the history of population  
198 splits and admixtures. We used the three-population ( $f_3$ ) and the four-population ( $f_4$ )  
199 statistics, introduced in (Reich *et al.* 2009) as methods to estimate mixture proportions in  
200 an admixed group, to test for introgression among species. The  $f_3$  and  $f_4$  statistics exploit the  
201 idea that the genetic drift, defined as a function of allele frequency, should be uncorrelated  
202 in unadmixed populations. As a result, correlations detected between allele frequencies of  
203 three ( $f_3$ ) or four ( $f_4$ ) populations joined by an unrooted tree indicate episodes of gene flow.  
204 Specifically, the  $f_3$  (X; Y, W) tests for admixture between a test population X and two  
205 reference populations Y and W. The expected value of  $f_3$  is positive in case of no mixture and  
206 negative if X is admixed with Y or W, or both. Similarly, the  $f_4$  for an unrooted tree (A,B;C,D),  
207 tests whether allele frequency differences between A and B are correlated with differences  
208 between C and D.  $f_4$  is equal to zero if there are no correlations and no admixtures across  
209 branches of the tree. In contrast,  $f_4$  is significantly positive if admixture occurs between A  
210 and C, or between B and D, or both, and significantly negative if admixture occurs between  
211 A and D, or between B and C, or both. The statistical significance of  $f_3$  and  $f_4$  values is  
212 assessed using a Z-score: an  $f_3$  or  $f_4$  value divided by its standard deviation. We used a

213 threshold Z-score of 2.5 corresponding to a p-value of 0.05 as suggested in (Reich *et al.*  
214 2009). The *threepop* and *fourpop* programs of the software *TreeMix* (Pickrell & Pritchard  
215 2012) were used to calculate  $f_3$  and  $f_4$  statistics and Z-scores for all possible triples and  
216 quadruples of populations we described within *An. nili s.l.*

217 We next applied the graph-based method developed in *TreeMix* to determine the  
218 directionality and quantify the extent of gene flow among species. The *TreeMix* approach  
219 uses allele frequencies at genome-wide polymorphisms and a Gaussian approximation of  
220 the genetic drift among populations to first construct a Maximum Likelihood (ML)  
221 phylogeny connecting sampled populations by simple bifurcations. The model then  
222 compares the covariance structure modeled by this dendrogram and adds edges to the  
223 phylogeny to account for admixtures. We first conducted one *TreeMix* run without  
224 migration. We noted the percentage of explained variance of the models and visually  
225 inspected the residuals of covariance matrixes among populations. We next ran *TreeMix*  
226 100 times without migration using different numbers of random seed and we built a ML  
227 consensus tree from the 100 trees in *SumTrees* (Sukumaran & Holder 2010) using the 95%  
228 majority rule. We finally added 1 and 4 migration edges to the ML trees and examined the  
229 changes in the percentage of explained variance and the residual fit. For all  $f_3$ ,  $f_4$  and  
230 *TreeMix* analyses, we used 4343 SNPs that were present in all populations and in at least  
231 50% of individuals in every population.

### 232 **Genomic clines**

233 Loci of exceptional introgression compared with genome-wide average admixture may be  
234 important for local adaptation or the maintenance of species barriers. To identify sites with

235 the greatest contribution to introgression between *An. nili* and *An. ovengensis*, we used the  
236 Bayesian model of genomic clines as implemented in *bgc* (Gompert & Buerkle 2012). The  
237 model first estimates a genome-wide average of a hybrid index ( $h$ ) that varies between zero  
238 and one for every potentially admixed individual, using allele frequencies of “pure” parental  
239 populations (Gompert & Alex 2011). The value zero corresponds to pure individuals of the  
240 alternative species (*An. ovengensis* in our case) and the value one to pure individuals of the  
241 reference species (*An. nili* in our case). Two genomic clines parameters  $\alpha$  and  $\beta$  are then  
242 used to evaluate the deviation of individual loci in admixed individuals from the expected  
243 genome-wide hybrid index ( $h$ ). The cline parameter  $\alpha$  reflects the probability of *An. nili*  
244 ancestry relative to the base expectation ( $h$ ), whereas the genomic cline parameter  $\beta$   
245 denotes the rate of transition from low to high probability of *An. ovengensis* ancestry as a  
246 function of hybrid index (Gompert & Alex 2011). Deviations of loci from the genome-wide  
247 average of  $\alpha$  and  $\beta$  were examined on the basis of departures from the 95% confidence  
248 envelope.

249 The “pure” parental populations we used in genomic clines analyses were apparently  
250 allopatric *An. nili* and *An. ovengensis* populations collected in Nyabessan and Nkoteng,  
251 respectively (Fig. 1A). The two locations are separated by ~420 km, which presumably  
252 reduces the rate of gene flow between *An. nili* and *An. ovengensis*. Hybrids were admixed  
253 individuals collected from the sympatric area in Mbébé (Fig. 1A). These individuals had  
254 almost equivalent ancestry proportions from *An. nili* and *An. ovengensis* as revealed by both  
255 STRUCTURE and ADMIXTURE clustering analyses. In *bgc*, we estimated genomic cline  
256 parameters using 9622 SNPs deriving from aligned RAD-loci in a dataset consisting of 11

257 admixed individuals, 8 *An. nili* parents and 15 *An. ovengensis* parents. We calculated average  
258 values of  $\alpha$  and  $\beta$  across five runs of *bgc*, each run including 50000 steps with samples from  
259 the posterior distribution recorded every 25th step following a 25000 step burn-in. We  
260 visually inspected the MCMC output to assess convergence to the stationary distribution.  
261 To examine the relationship between the strength of selection and introgression, we tested  
262 whether loci with extreme or outlier genomic clines were enriched in genomic regions that  
263 were targets of selection. To do so, we studied the correlation between locus-specific cline  
264 parameters ( $\alpha$  and  $\beta$ ) and several divergence and diversity statistics ( $F_{ST}$ ,  $d_{xy}$ ,  $\theta_w$ ,  $\theta_\pi$ ,  
265 Tajima's  $D$ ) across 9622 SNPs in the genome. The strength of correlation was assessed  
266 using Pearson's product moment correlation coefficient.  
267

## 268 **Results**

### 269 **SNP genotyping**

270 We collected mosquitoes from four locations (Fig. 1A, Table S1) and sequenced 145  
271 individuals belonging, according to morphological identifications and PCR, to two species  
272 (*An. nili* (n = 24) and *An. ovengensis* (n = 121)). We aligned reads from all 145 individuals to  
273 the reference genome and we assembled 197724 96-bp RAD loci that mapped to unique  
274 positions throughout the genome. We retained 408 loci present in all populations and in at  
275 least 50% of individuals in each population and identified 4343 high-quality biallelic  
276 markers within these loci. We also identified another set of 704408 unique loci by building  
277 consensus RAD loci *de novo* without aligning reads to the reference genome. We applied the  
278 same stringent filtration as with aligned reads to identify 3071 high-quality SNPs.

### 279 **Genetic structure of populations**

280 PCA and NJ trees showed that the genotype variation at 4343 genome-wide SNPs among the  
281 145 sequenced individuals is best explained by more than two clusters, implying cryptic  
282 subdivisions among *An. nili* and *An. ovengensis* populations (Fig. 1C-D). The first three PCA  
283 axes and NJ trees clearly distinguished three subgroups within *An. nili* and two clusters in  
284 *An. ovengensis*. The five different clusters were associated with the sampling locations  
285 suggesting a strong correlation between the genetic structure and a local distinctness of  
286 populations. This marked geographic structure of *An. nili* and *An. ovengensis* populations  
287 can be explained by the ongoing adaptive divergence and ecological speciation within the  
288 two species. Importantly, both STRUCTURE and ADMIXTURE analyses revealed that, at k =  
289 3, a subgroup containing 11 individuals corresponded to a cluster of hybrids with almost

290 half ancestry from *An. nili* and half from *An. ovengensis* (Fig. 1B). This result is surprising  
291 given the substantial time since divergence. We applied three different methods to identify  
292 the optimal number of genetic clusters in our samples. The DAPC suggested the presence of  
293 five clusters as indicated by PCA and NJ trees (Fig. 1E). However, the method of Evanno et  
294 al. indicated two probable ancestors while the distribution of the cross-validation error as a  
295 function of the number of putative populations in ADMIXTURE failed to unambiguously  
296 reveal an optimal number of genetic clusters (Fig. 1F,G). These conflicting results between  
297 methods are sometimes observed when the history of subdivisions and admixtures events  
298 is very complex as it is the case in *An. nili s.l* (Decker *et al.* 2014). The Evanno et al. method  
299 is overwhelmed by early divergence between *An. nili* and *An. ovengensis* while the results of  
300 DAPC and the ADMIXTURE cross-validation error reflect recent hierarchical population  
301 subdivisions within the two species. The inferred genetic structure was consistent when we  
302 used respectively 4343 and 3071 SNPs identified from reference-based and *de novo*  
303 assemblies (Fig. 1 and Fig. S1).

#### 304 **Genetic differentiation**

305 We estimated the level of population differentiation between the five genetic clusters  
306 identified in our samples using the overall pairwise  $F_{ST}$  (Table 1). We found strong pairwise  
307 differentiation characterized by extreme  $F_{ST}$  values, including between populations  
308 classified as belonging to the same species. The level of differentiation is even higher  
309 between some pairs of populations within the same species than between *An. nili* and *An.*  
310 *ovengensis*. The highest level of within-species genetic differentiation was recorded  
311 between “allopatric” populations of *An. ovengensis* collected from locations separated by

312 ~350km ( $F_{ST} = 0.896$ ,  $p < 0.005$ ). These findings strongly suggest that, in addition to  
313 hybridization, local differentiation associated with late stages of adaptive divergence within  
314 species are overwhelming current taxonomic descriptions in *An. nili s.l.* populations. The  
315 results obtained with reference alignments perfectly mirrored those of *de novo* assemblies  
316 (Table 1, S2). Taken as a whole, the population genetic structure and divergence of *An. nili*  
317 and *An. ovengensis* depicts a radiating group involving a collection of species whose  
318 phylogenetic relationships are blurred by ongoing hybridizations. As a result, current  
319 taxonomic classifications based on morphological characters and point mutations on the  
320 ribosomal DNA cannot effectively describe the actual reproductive units and the vectorial  
321 capacity within the *An. nili* species group.

### 322 **Evidence for recent gene flow between *An. nili* and *An. ovengensis***

323 We used *TreeMix* to construct a tree connecting the five species and to effectively describe  
324 and visualize the mixture(s) event(s). The consensus of 100 ML *TreeMix* tree inferred from  
325 4343 SNPs without migration resumes population-level relationships described by  
326 neighbor-joining analysis, PCA and clustering analyses (Fig. 2A). The long terminal  
327 branches leading to *An. nili* group 1, *An. nili* group 2 and *An. ovengensis* group 1 in the  
328 inferred tree reflect the signatures of strong bottlenecks in the history of these species.  
329 Interestingly, the residual fit from the ML model without migration suggested some  
330 correlations between species that were consistent with the admixture events inferred from  
331 ADMIXTURE and STRUCTURE analyses, particularly between the two species *An. nili* group  
332 2 and *An. ovengensis* group 2 (Fig. 1B). Additional support to admixture was provided by  
333 the four-population test. Across all possible combinations of populations, there was a highly

334 significant allele frequency correlation between *An. nili* group 2 and *An. ovengensis* group 1  
335 confirming that *An. nili* group 1 is the result of an admixture event between *An. nili* group 2  
336 and *An. ovengensis* group 1 (Table 2 and S3). Nevertheless, five independent runs of *TreeMix*  
337 including one or two migration edges captured none of these migration events. Also, the  
338 explained variance of the *TreeMix* ML model without migration edge was very high (99.8%)  
339 indicating that a bifurcation tree can also match the phylogenetic relationships among  
340 sampled populations. Finally, we found no significant  $f_3$  statistics in all triplets of  
341 populations, but this test may be underpowered because of the low-level and the precocity  
342 of admixture or because of the complex demographic history of our populations (Decker *et*  
343 *al.* 2014). As we showed previously with the population structure and genetic divergence,  
344 the admixture events we described between *An. nili* and *An. ovengensis* also hold across  
345 assembly methods (*de novo* or reference alignments)(Fig. 2 and S2).

#### 346 **Genomic signatures of divergent selection and differential introgression**

347 We retained 9622 SNP loci identified from a reference-based assembly that we used to  
348 examine the genomic footprints of divergent selection and introgression in a dataset  
349 containing 8 “pure” *An. nili* s.s. individuals, 15 “pure” *An. ovengensis* and 11 admixed  
350 individuals (see Materials and methods). The values of  $F_{ST}$  among SNP loci between parent  
351 species *An. nili* and *An. ovengensis* were heterogeneous across the genome (Fig. 3). The  
352 empirical distribution of locus-specific  $F_{ST}$  also revealed an unusually bimodal shape  
353 featuring  $F_{ST}$  peaks centered on values around 0 and 1. The great majority of sites have low  
354 to moderate differentiation, but a substantial number of loci are extremely differentiated  
355 between the two species (Fig. 3A). We thinned our SNP set to 4003 using more stringent

356 filtration criteria and identified 42 statistical outliers of  $F_{ST}$  in LOSITAN, which were all  
357 above the 99<sup>th</sup> percentile of the empirical distribution of  $F_{ST}$ . The bimodality of the genetic  
358 divergence between *An. nili* and *An. ovengensis* is also evident in the empirical distribution  
359 of  $d_{xy}$  (Fig. 3B). As exemplified by some empirical cases like the introgressive hybridization  
360 observed between two closely related species of monkeyflowers (*Mimulus*) (Brandvain *et*  
361 *al.* 2014), the bimodal divergence is due to the coexistence between sites of abrupt  
362 differentiation reflecting the real level of divergence between the two species and sites of  
363 low divergence resulting from recent gene flow. We conducted Linkage Disequilibrium (LD)  
364 analyses to understand if highly differentiated sites are clustered into blocks of linked loci  
365 or dispersed throughout the genome. Particularly, polymorphic chromosomal inversions  
366 are important sources of genetic variation in *Anopheles* species and play a key role in local  
367 adaptation and speciation. High  $F_{ST}$  values can aggregate in genomic regions containing  
368 chromosomal inversions polymorphisms that are under strong divergent selection and  
369 drive adaptive segregation as often observed in other African anopheline mosquitoes  
370 (Ayala *et al.* 2011; Cheng *et al.* 2012; Kamdem *et al.* 2016). Cytogenetic studies have  
371 detected no polymorphic chromosomal inversion among *An. ovengensis* samples. Two  
372 polymorphic inversions (*2Rb*, *2Rc*) have been identified in *An. nili*, but samples from  
373 Cameroon bore only the *2Rc* at a low frequency (Sharakhova *et al.* 2013). Rearranged  
374 regions of the genome are characterized by reduced recombination and increased LD  
375 relative to the genome-wide average. To effectively examine the extent of LD in the  
376 genomes of *An. nili* and *An. ovengensis* and to test for the effect of chromosomal inversion  
377 polymorphisms in adaptive evolution, we used the package LDna to identify clusters of LD.

378 To prevent spurious clusters due to LD between SNPs located on the same RAD locus, we  
379 used a dataset in which only one SNP was randomly selected within each RAD locus (1330  
380 SNPs in total). We optimized LDna clustering parameters, which indicated the presence of 6  
381 LD blocks (Single Outlier Clusters (SOCs)) among our samples (Fig. 4). These clusters  
382 represent signals of independent or compound events in the evolutionary history that left  
383 imprints on LD across the genome (Kemppainen *et al.* 2015). To further understand the role  
384 of the 6 LD clusters in the evolutionary history of *An. nili* and *An. ovengensis*, we conducted  
385 downstream analysis using PCA to examine the population structure of SNPs within each  
386 SOC. We found that one SOC (containing 39 SNPs) clearly discriminated a cluster  
387 encompassing all hybrid individuals from the two parent species (Fig. 4B). Two other SOC  
388 (containing respectively 383 and 112 SNPs) consistently separated one parent species from  
389 a group comprising the other parent species and all hybrid individuals. The last three SOC  
390 revealed no clear clustering patterns. None of the SOC differentiated clusters that could be  
391 associated with the three alternative karyotypes (inverted homozygotes, heterozygotes and  
392 uninverted homozygotes) expected in case of polymorphic inversion. As suggested by  
393 cytogenetic observations, polymorphic inversions likely play only a moderate role in the  
394 genomic divergence in *An. nili s.l.*, but a more intensive sampling is needed to make any  
395 definitive conclusion. Interestingly, the three most important blocks of LD found in the  
396 genomes of *An. nili* and *An. ovengensis* are instead associated with hybridization, which  
397 emphasizes the central role played by interspecific gene flow in the genomic architecture of  
398 adaptive speciation in the two species. Our results also suggest that loci resistant to  
399 introgression and genes responsible for reproductive isolation between *An. nili* and *An.*

400 *ovengensis* are in strong linkage disequilibrium. This finding is consistent with recent  
401 introgression whereby long-range haplotypes that are generated by recent gene flow  
402 between genetically distinct populations have not have sufficient time to be broken down  
403 by recombination (admixture-induced LD) (Martin *et al.* 2013). However, LDna estimates  
404 the LD irrespective of the physical linkage between markers and LD blocks can also result  
405 from a strong correlation between allelic frequencies of SNPs scattered throughout the  
406 genome. A detailed characterization of long-range haplotypes across genomes of *An. nili* and  
407 *An. ovengensis* will provide a more powerful examination of the genomic architecture of  
408 reproductive isolation between the two species.

409 Here we have used another innovative approach (genomic cline) to further understand the  
410 relationship between divergence, selection and differential introgression at the genomic  
411 level. First, consistent with STRUCTURE and ADMIXTURE results, the empirical distribution  
412 of hybrid index among admixed individuals (average hybrid index  $\sim 0.3$ ) shows a slight  
413 predominance of *An. ovengensis* ancestry (Fig. 5A, Fig. 1B). Using estimates of genomic cline  
414 parameters, we noted that introgression was very heterogeneous across loci (Fig. 5B).  
415 Indeed, the level of introgression at 1297 SNP loci differed significantly from the genome-  
416 wide average (outliers) and we detected an excess *An. ovengensis* ancestry for 2506 loci  
417 (lower bound of 95% CI for  $\alpha > 0$ ) and excess *An. nili* ancestry for 2635 loci (upper bound of  
418 95% CI for  $\alpha < 0$ ). Estimates of the genomic cline parameter  $\alpha$  range from -5 to 5.25 while  
419 our values of  $\beta$  are low overall ( $\beta$  varies from -0.13 to 0.09). Moreover, we identified no loci  
420 with significantly elevated estimates of genomic cline rate. The  $\beta$  parameter assesses the  
421 rate of transition from one ancestry to the other and thereby scores the steepness of the

422 genomic cline at each locus. Extreme values of  $\beta$  are expected when there is population  
423 structure in the hybrid zone, selection against hybrids or gene flow among parent species  
424 (Parchman *et al.* 2013). Our hybrid population is highly differentiated ( $F_{ST} > 0.8$ ) from both  
425 parent species and is well adapted in hybrid zone, which certainly explain the low  $\beta$  values  
426 observed. One key aspect in the Bayesian implementation of genomic clines is the  
427 relationship between  $\alpha$  and  $\beta$ , which is crucial to understand the rate of transition of sites of  
428 exceptional introgression from one side to the other of the genome. As we have shown  
429 previously with  $F_{ST}$  and  $d_{xy}$ , recent introgression in divergent genomes of *An. nili* and *An.*  
430 *ovengensis* has resulted in bimodal genomic divergence featuring two blocks of sites with  
431 extreme divergence values. This pattern presumes that most variant sites will have very  
432 steep genomic clines because the transition from one ancestry to the other is very abrupt.  
433 In agreement with this prediction, the scatterplot of  $\beta$  as a function of  $\alpha$  indicates the  
434 presence of two blocks of SNPs with either high probability of *An. nili* or of *An. ovengensis*  
435 ancestry (Fig. 5B). The genomes of both species are split into two compartments of ancestry  
436 due to recent introgression resulting in steep genomic clines at hybrid zone (Fig. 5B). To  
437 better understand the biological significance of these outlier loci of introgression in  
438 admixed individuals, we assessed the correlation between  $\alpha$  and  $\beta$  and locus-specific  
439 estimates of divergence and selection parameters. We found a detectable negative  
440 correlation between  $F_{ST}$  and values of  $\alpha$  ( $r = -0.15$ ,  $p < 0.005$ ). Among the 32 LOSITAN  $F_{ST}$   
441 outliers, 13 loci had extreme  $\alpha$  estimates, and 11 were outlier of  $\beta$ , but the average values of  
442 both  $\alpha$  and  $\beta$  were not significantly different between  $F_{ST}$  outliers and the 9622 genome-  
443 wide SNP loci (Mann-Whitney U-test,  $P < 0.001$ ). Erroneous correlation between  $F_{ST}$  and

444 exceptional introgression can be inferred when the  $F_{ST}$  between parental populations is  $< 0.1$   
445 (Parchman *et al.* 2013). The high overall  $F_{ST}$  ( $\sim 0.8$ ) between *An. nili* and *An. ovengensis*  
446 minimizes such errors in our study. We also detected a negative correlation between  $\alpha$  and  
447  $d_{xy}$  ( $r = -0.13$ ,  $p < 0.005$ ), which confirms previous results with  $F_{ST}$ , suggesting that sites that  
448 diverge strongly resist to differential introgression because they likely contain reproductive  
449 isolation factors. The negative correlation between  $d_{xy}$  and estimates of locus-specific  
450 admixture indicates that perhaps gene flow has had sufficient time to reduce sequence  
451 divergence at admixed loci despite the relatively recent hybridization process. We next  
452 correlated cline parameters ( $\alpha$  and  $\beta$ ) to the estimates of locus-specific diversity ( $\theta_w$  and  $\theta_\pi$   
453 ) and the allele frequency spectrum (Tajima's  $D$ ). The results provided another clear  
454 illustration of the steep genomic clines described previously with alpha and beta values.  
455 Notably, there is a strong positive correlation between  $\theta_w$  and  $\alpha$  ( $r = 0.53$ ,  $p < 0.005$ ) in *An.*  
456 *ovengensis* and a negative correlation of the same magnitude with the diversity of *An. nili* ( $r$   
457  $= -0.49$ ,  $p < 0.005$ ), which translate the segregation of genomes of admixed individuals in  
458 two sources of ancestry that are at opposite ends of the genomic cline. Interestingly, the  
459 high correlation between alpha and diversity indicate that sites that are favored by  
460 introgression are not constrained by divergent selection that should have likely resulted in  
461 depression in nucleotide diversity. In contrast to what has been shown in manakin birds  
462 (Parchman *et al.* 2013) and in *Lycaeides* butterflies (Gompert *et al.* 2012, 2013), there is no  
463 evidence of increased introgression at loci under divergent selection between *An. nili* and  
464 *An. ovengensis*.

## 465 **Discussion**

### 466 **Hybridization at late stages of speciation**

467 We have described a complex case of speciation and hybridization in a group of related  
468 *Anopheles* mosquitoes endemic to Sub-Saharan Africa. Both incomplete speciation and  
469 pervasive hybridization are leading to a very dynamic pattern of genetic structure between  
470 and within species. We first analysed the population genetic structure and revealed cryptic  
471 subdivisions between the two malaria vectors *An. nili* and *An. ovengensis*. The exact number  
472 of demes or cryptic subgroups within each species in our samples was difficult to  
473 determine. We found conflicting results between estimates from three different genetic  
474 clustering methods. However, three different methods suggested the existence of 2 to 5  
475 clusters in our sample. *An. nili* is subdivided into three clusters while ongoing adaptive  
476 speciation in *An. ovengensis* results so far in two cryptic species. The extremely high genetic  
477 differentiation between populations indicates that all these cryptic subgroups are almost  
478 complete species. Our results therefore suggest that *An. nili* and *An. ovengensis*, contrary to  
479 the current taxonomy, represent probably two different complexes of cryptic species. The  
480 geographic origin of samples explains a great part of the genetic variance among individuals  
481 consistent with strong local differentiation and adaptive speciation. Significant population  
482 structure has been described among *An. nili* populations from Cameroon with 8  
483 microsatellite loci (Ndo *et al.* 2013). Ndo *et al.* 2013 also found  $F_{ST}$  values as high as 0.48  
484 between two populations from the forest area. Using genome-wide SNPs, we have identified  
485 new subdivisions and revealed that the population genetic structure of *An. nili* is more  
486 complex. Our work highlights the strength of Next Generation Sequencing (NGS)

487 approaches and the necessity of fine-scale genomic examinations in the resolution of  
488 intricate patterns of ancestry in this group of mosquito. By contrast, the ongoing speciation  
489 we have observed within *An. ovengensis* has never been described in the past, and this  
490 species has been sometimes considered as a sibling of *An. nili* (Kengne *et al.* 2003; Awono-  
491 Ambene *et al.* 2004, 2006; Ndo *et al.* 2013). Nevertheless, more recent studies have started  
492 to challenge the assumed relatedness between the two species (Ndo *et al.* 2013; Sharakhova  
493 *et al.* 2013). Precisely, analyses of polytene chromosomes revealed high karyotypic  
494 divergence of *An. ovengensis* from *An. nili* (Sharakhova *et al.* 2013) and estimates of time  
495 since divergence indicated that the two species split from one another 3 to 6-Myr ago (Ndo  
496 *et al.* 2013). Intriguingly, our work provides multiple lines of evidence supporting the  
497 existence of extensive ongoing gene flow between the two species despite this strong  
498 divergence. First, clustering methods and Bayesian implementation of genomic clines  
499 identified individuals with almost half ancestry from both species and formal tests of  
500 population admixture corroborated these ongoing admixture events. Second, perhaps the  
501 most compelling evidence for hybridization is the presence of admixture-induced linkage  
502 disequilibrium (ALD) characterized by blocks of linked SNPs that discriminate hybrid  
503 populations from parental species. In general, linkage disequilibrium (LD) or the  
504 correlation between allele frequencies of different loci across the genome, which can have  
505 multiple origins including selection, genetic drift, or population structure, is normally  
506 eroded by recombination in the course of time. ALD is caused by associations between  
507 nearby loci co-inherited on an intact chromosomal block from one of the ancestral mixing  
508 populations (Loh *et al.* 2013). Signatures of ALD are frequent in genomes of recently

509 admixed populations for which recombination has not yet broken the large introgressed  
510 chromosomal segments into smaller portions. ALD is a well-known feature in evolutionary  
511 history of humans and estimates of long-range LD have been proposed as an approach to  
512 measure the extent and the timing of admixture events that have shaped the genetic  
513 polymorphism across genomes of extant populations (Loh *et al.* 2013). In insects,  
514 signatures of ALD have been found for example in the *Heliconius* butterfly genome,  
515 especially around loci involved in mimicry of color patterns that circulate between species  
516 (Martin *et al.* 2013). *An. nili* provides another rare example of ALD across the genome of an  
517 insect species. Further studies using a more comprehensive genomic sequencing and a  
518 reference genome of better quality will help us to improve our knowledge of functional and  
519 sequential characteristics of admixed LD blocks found in the *An. nili* genome.

520 Although the concept of speciation-with-gene-flow has become the dominant paradigm in  
521 speciation studies, we remain ignorant about the conditions that prevent or motivate gene  
522 flow between divergent lineages before the onset of complete reproductive isolation.  
523 Moreover, even the notion of “complete reproductive isolation” is now challenged because  
524 an increasing number of examples from diverse taxa showed rampant gene exchange across  
525 strong reproductive barriers, sometimes between established species that diverged several  
526 million years ago (Nydam & Harrison 2011; Roux *et al.* 2013; Parchman *et al.* 2013; Martin  
527 *et al.* 2013; Canestrelli *et al.* 2014). These studies and ours suggest that processes  
528 underlying hybridization and introgression in the presence of clear genetic differentiation  
529 will be best addressed within speciation continuum rather than across couples of  
530 occasionally mating species. A continuum of speciation featuring a collection of taxa

531 occupying a gradient of genetic/ecological divergence provides ideal conditions where the  
532 relation between divergence and hybridization can be inferred. Contrary to what we  
533 initially thought, the *An. nili* group does not represent a speciation continuum but instead a  
534 collection of complete species sharing a common ancestry. Moreover, despite extensive  
535 efforts, we couldn't sample populations of the two remaining species of the group: *An.*  
536 *somalicus* and *An. carnevalei*. Therefore, patterns of admixtures and ancestry among  
537 populations of this group of mosquitoes are probably more complex than what we have  
538 shown. Nevertheless, instead of a comprehensive description of splits and admixtures in *An.*  
539 *nili s.l.*, we have focused our efforts on the examination of the genomic architecture of  
540 divergence and introgression between *An. nili* and *An. ovengensis*. The results we discuss  
541 below are among the rare cases that address the genomic signatures of gene flow and the  
542 relationship between divergence, selection and introgression at late stages of the speciation  
543 process.

#### 544 **Genomic architecture of adaptive introgression**

545 Owing to the increasing availability of high-throughput sequencing information, genomes of  
546 multiple taxa have been scanned and analyzed in comparative frameworks to search for  
547 genomic signatures of divergence and speciation. The prevailing idea behind these  
548 approaches is that regions of extreme differentiation between incipient or complete species  
549 contain factors that maintain reproductive isolation (RI) among their populations. In  
550 reality, the genomic distribution of highly differentiated regions has been more contentious  
551 (Nosil & Feder 2012, 2013). Overall, two models are well documented: the “speciation  
552 island” model whereby RI loci are thought to be caught up in in a few regions of the genome

553 where outliers of genetic differentiation cluster (e.g. (Andrew & Rieseberg 2013)) and the  
554 “heterogeneous” distribution model, which posits that genomic divergence is instead lead  
555 by numerous small genomic regions scattered throughout all chromosomes (e.g.  
556 (Lawniczak *et al.* 2010; Roesti *et al.* 2012)). A shift in this paradigm is envisioned as  
557 evidence is accumulating indicating that highly differentiated loci can paradoxically  
558 coincide with regions of elevated introgression between two species (Gompert *et al.* 2012,  
559 2013; Pool *et al.* 2012; Parchman *et al.* 2013). The genomic distribution of the genetic  
560 divergence between *An. ovengensis* and *An. nili* shows that the compound effects of strong  
561 divergence and recent introgression generates a bimodal pattern of divergence which  
562 assigns most sites into two main categories: a majority of low divergence sites and a small  
563 cluster of high divergence loci with  $F_{ST}$  values centered around 1. In most of the widespread  
564 *Anopheles* species, signatures of high divergence can be found in large chromosomal  
565 segments corresponding to rearranged regions of chromosomes where recombination is  
566 rare (Neafsey *et al.* 2015). This is the case for example in the 2La inversion locus, which  
567 depicts signatures of strong divergent selection along a latitudinal cline (Cheng *et al.* 2012).  
568 In agreement with cytogenetic studies, which found roughly no polymorphic inversions  
569 among *An. nili* samples from Cameroon, we observed no tendency for high  $F_{ST}$  or  $d_{xy}$  to  
570 cluster within regions that can be assimilated to chromosomal rearrangements. In contrast  
571 to what has been recently demonstrated in manakin birds (Parchman *et al.* 2013) and in  
572 *Lycaeides* butterflies (Gompert *et al.* 2012, 2013), there is a negative correlation between  
573 the introgression parameter  $\alpha$  and genetic divergence (estimated both as  $F_{ST}$  and  $d_{xy}$ )  
574 between *An. nili* and *An. ovengensis*. As expected, gene flow is favored across neutral loci

575 and those that provide selective advantage, but are presumably not under strong divergent  
576 selection. Moreover, the relationship between  $\alpha$  and  $\beta$  suggests a prevalence of steep clines  
577 in genomes of admixed individuals at contact zones. This also translates into a positive  
578 correlation between  $\alpha$  and the nucleotide diversity in *An. ovengensis* and a negative  
579 correlation of  $\alpha$  to the diversity of *An. nili*. In theory, SNPs with steep cline are hypothesized  
580 to be near genes involved in reproductive isolation and as such are possibly under selection  
581 in hybrids (Janoušek *et al.* 2015). Therefore, the abundance of steep clines provides another  
582 clear illustration of the mosaic genome characterized by coexistence between high  
583 divergence and consistent gene flow observed in the hybrid species.

584 A substantial body of evidence indicates that genomic material coming from related species  
585 can confer an advantage to populations (adaptive introgression). Adaptive introgression of  
586 one or two loci has been widely studied over the last two decades and excellent examples  
587 have been described. The most prominent cases include the transfer of genes involved in  
588 mimicry of color patterns in *Heliconius* butterflies (Consortium 2012), the circulation of  
589 resistant alleles of insecticides in mosquitoes (Clarkson *et al.* 2014; Norris *et al.* 2015) and  
590 rodenticide in mice (Song *et al.* 2011). However, although convincing signatures of adaptive  
591 introgression around one gene or a few linked loci can now be described in an increasing  
592 number of species, the extent and the magnitude of introgressive hybridization across the  
593 genome remained unknown. Further, due in part to the fact that most species in which the  
594 concomitance of introgressive hybridization and high divergence has been observed are  
595 nonmodel species with little genomic resources, the knowledge of genomic characteristics  
596 of adaptive introgression in these species remains relatively modest. Meanwhile, a

597 consistent pattern has started to emerge from the few cases that have been studies with  
598 substantial genomic details. In general, genomic regions exhibiting non-random  
599 introgression are widely dispersed across the genome, rather than co-localized in a few  
600 discrete blocks. This heterogeneity of genome-wide introgression patterns has been  
601 observed for example between manakin birds (*Manacus candei* and *M. vitellinus*)  
602 (Parchman *et al.* 2013) in different mice subspecies (Liu *et al.* 2015), and among  
603 mosquitoes of the *Anopheles gambiae* species complex (Fontaine *et al.* 2014; Kamdem *et al.*  
604 2016). Our results also show that selective introgression can be widespread across the  
605 genome of two highly divergent species. However, in contrast to most of the reference cited,  
606 LD analyses in *An. nili* and *An. ovengensis* have revealed a LD cluster separating hybrids  
607 from the two parental species, which suggests that at least some of the recently  
608 introgression loci consist of relatively large chromosome segments that have yet to be  
609 further characterized sequentially and functionally with a high-quality reference genome.  
610 In addition to the lack of knowledge about the genomic architecture, the functional and  
611 phenotypic aspects of introgressive hybridization between established species remain  
612 obscures. In mice for example, introgression is associated with a polarization of GO terms,  
613 regions of elevated introgression exhibiting a disproportionate number of genes involved in  
614 signal transduction and olfactory receptor genes (Janoušek *et al.* 2015). Hybrids between  
615 *An. nili* and *An. ovengensis* were collected from an area of the equatorial forest whose  
616 environmental features were not apparently very divergent from those of the locations  
617 where the parent species were sampled (Fig. 1A). As a result, it is hard to pinpoint

618 environmental gradients and the life history traits that can be considered as the main  
619 drivers of speciation and introgression among species of *An. nili* and *An. ovengensis*.  
620

## 621 **Conclusions and implications**

622 Although hybridization has been recognized as one of the major forces that affect the  
623 evolution of living species, the detailed study of its fundamental and applied implications  
624 has been hampered by methodological limitations. Advances in high-throughput DNA  
625 sequencing and statistical genomics are revolutionizing experimental and conceptual  
626 approaches, allowing a very sensitive examination of the heterogeneity of hybridization  
627 across species and genomes. We have used a combination of tests tailored to infer patterns  
628 of introgression across genomes of nonmodel species. Although our results still need to be  
629 replicated in other contact zones across the distribution range of *An. nili* in Africa, they  
630 highlight the complex relationships between divergence, selection and introgression during  
631 the split of taxa. Our work has methodological, conceptual and applied implications. Most  
632 genome scans assume a negative correlation between genetic divergence and introgression.  
633 It has been suggested that that the opposite is possible (Parchman *et al.* 2013; Gompert *et*  
634 *al.* 2013), but our findings do not support this hypothesis. Climate change and anthropogenic  
635 disturbance are contributing to expand geographic ranges of mosquito species worldwide  
636 thereby increasing contact between previously isolated species that are capable of  
637 exchanging gene flow. This interspecific gene flow in mosquitoes often leads to the spread  
638 of insecticide resistance alleles and other epidemiologically significant genes. Our work  
639 provided a methodological validation of a cost-effective population genomic approach that  
640 can be applied to investigate the bases of introgressive hybridization in other mosquito  
641 species.

642

643 **Acknowledgements**

644 Funding for this project was provided by the University of California Riverside and NIH  
645 grants 1R01AI113248 and 1R21AI115271 to BJW.

646

647

## 648 **References**

- 649
- 650 Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of*  
651 *Evolutionary Biology*, **26**, 229–246.
- 652 Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in  
653 unrelated individuals. *Genome Research*, **19**, 1655–1664.
- 654 Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomic regions early in  
655 speciation: incipient speciation of sunflower ecotypes. *Evolution; international journal*  
656 *of organic evolution*, **67**, 2468–82.
- 657 Antonio-Nkondjio C, Kerah CH, Simard F *et al.* (2006) Complexity of the malaria vectorial  
658 system in Cameroon: contribution of secondary vectors to malaria transmission.  
659 *Journal of medical entomology*, **43**, 1215–1221.
- 660 Antonio-Nkondjio C, Ndo C, Costantini C *et al.* (2009) Distribution and larval habitat  
661 characterization of *Anopheles moucheti*, *Anopheles nili*, and other malaria vectors in  
662 river networks of southern Cameroon. *Acta Tropica*, **112**, 270–276.
- 663 Awono-Ambene HP, Kengne P, Simard F, Antonio-Nkondjio C, Fontenille D (2004)  
664 Description and bionomics of *Anopheles (Cellia) ovengensis* (Diptera: Culicidae), a new  
665 malaria vector species of the *Anopheles nili* group from south Cameroon. *Journal of*  
666 *medical entomology*, **41**, 561–568.
- 667 Awono-Ambene HP, Simard F, Antonio-Nkondjio C *et al.* (2006) Multilocus enzyme  
668 electrophoresis supports speciation within the *Anopheles nili* group of malaria vectors  
669 in Cameroon. *The American journal of tropical medicine and hygiene*, **75**, 656–658.
- 670 Ayala D, Fontaine MC, Cohuet A *et al.* (2011) Chromosomal inversions, natural selection and  
671 adaptation in the malaria vector *Anopheles funestus*. *Molecular biology and evolution*,  
672 **28**, 745–758.
- 673 Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of  
674 population structure. *Proceedings of the Royal Society B: Biological Sciences*, **263**, 1619–  
675 1626.
- 676 Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and Introgression  
677 between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics*, **10**.
- 678 Canestrelli D, Bisconti R, Nascetti G (2014) Extensive unidirectional introgression between  
679 two salamander lineages of ancient divergence and its evolutionary implications.  
680 *Scientific reports*, **4**, 6516.
- 681 Cheng C, White BJ, Kamdem C *et al.* (2012) Ecological genomics of *Anopheles gambiae* along  
682 a latitudinal cline: a population-resequencing approach. *Genetics*, **190**, 1417–32.
- 683 Clarkson CS, Weetman D, Essandoh J *et al.* (2014) Adaptive introgression between  
684 *Anopheles* sibling species eliminates a major genomic island but not reproductive  
685 isolation. *Nature communications*, **5**, 4248.
- 686 Consortium THG (2012) Butterfly genome reveals promiscuous exchange of mimicry  
687 adaptations among species. *Nature*, **487**, 94–8.
- 688 Decker JE, McKay SD, Rolf MM *et al.* (2014) Worldwide Patterns of Ancestry, Divergence,  
689 and Admixture in Domesticated Cattle. *PLoS Genetics*, **10**.
- 690 Earl DA, VonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for  
691 visualizing STRUCTURE output and implementing the Evanno method. *Conservation*

- 692           *Genetics Resources*, **4**.
- 693 Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in  
694 *Ficedula flycatchers*. *Nature*, **491**, 756–60.
- 695 Evanno G, Goudet J, Regnaut S (2005) Detecting the number of clusters of individuals using  
696 the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.
- 697 Fontaine MC, Pease JB, Steele A *et al.* (2014) Extensive introgression in a malaria vector  
698 species complex revealed by phylogenomics. *Science*, **347**, DOI:  
699 10.1126/science.1258524.
- 700 Fumagalli M, Vieira FG, Linderoth T (2014) ngsTools : methods for population genetics  
701 analyses from Next-Generation Sequencing data. *Bioinformatics*.
- 702 Gillies MT, Coetzee M (1987) *A supplement to the Anophelinae of Africa south of the Sahara*.  
703 The South African Institute for Medical Research, Johannesburg.
- 704 Gillies MT, De Meillon B (1968) *The Anophelinae of Africa South of the Sahara*. Publications  
705 of the South African Institute for Medical Research, Johannesburg.
- 706 Gompert Z, Alex BC (2011) Bayesian estimation of genomic clines. *Molecular ecology*, **20**,  
707 2111–2127.
- 708 Gompert Z, Buerkle C a. (2012) `bgc` : Software for Bayesian estimation of genomic  
709 clines. *Molecular Ecology Resources*, **12**, 1168–1176.
- 710 Gompert Z, Lucas L, Nice C (2012) Genomic regions with a history of divergent selection  
711 affect fitness of hybrids between two butterfly species. ..., 2167–2181.
- 712 Gompert Z, Lucas LK, Nice CC, Buerkle CA (2013) Genome divergence and the genetic  
713 architecture of barriers to gene flow between *Lycaeides idas* and *L. Melissa*. *Evolution;*  
714 *international journal of organic evolution*, **67**, 2498–514.
- 715 Grant BR (2015) Introgressive hybridization and natural selection in Darwin’s finches.  
716 *Biological Journal of the Linnean Society* *Biolog*, 812–822.
- 717 Hohenlohe P a, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation  
718 in threespine stickleback using sequenced RAD tags. *PLoS genetics*, **6**, e1000862.
- 719 Jakobsson M, Rosenberg N (2007) CLUMPP: a cluster matching and permutation program  
720 for dealing with multimodality in analysis of population structure. *Bioinformatics*, **23**,  
721 1801– 1806.
- 722 James M (2007) Hybrid speciation. *Nature*, **446**, 279–283.
- 723 Janoušek V, Munclinger P, Wang L, Teeter KC, Tucker PK (2015) Functional organization of  
724 the genome may shape the species boundary in the house mouse. *Molecular Biology*  
725 *and Evolution*, **32**, 1208–1220.
- 726 Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers.  
727 *Bioinformatics*, **24**, 1403–1405.
- 728 Kamdem C, Fouet C, Gamez S, White BJ (2016) Pollutants and insecticides drive local  
729 adaptation in African malaria mosquitoes. *bioRxiv*.
- 730 Kempainen P, Knight CG, Sarma DK *et al.* (2015) Linkage disequilibrium network analysis  
731 (LDna) gives a global view of chromosomal inversions, local adaptation and geographic  
732 structure. *Molecular Ecology Resources*, n/a–n/a.
- 733 Kengne P, Awono-Ambene, H. P. Antonio-Nkondjio C, Simard F, Fontenille D (2003)  
734 Molecular identification of the *Anopheles nili* group African malaria vectors. *Med Vet*  
735 *Entomol*, **17**, 67–74.

- 736 Korneliussen T, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation  
737 Sequencing Data. *BMC Bioinformatics*, **15**, 356.
- 738 Lawniczak MKN, Emrich SJ, Holloway a K *et al.* (2010) Widespread divergence between  
739 incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*  
740 (*New York, N.Y.*), **330**, 512–4.
- 741 Liu KJ, Steinberg E, Yozzo A *et al.* (2015) Interspecific introgressive origin of genomic  
742 diversity in the house mouse. *Proceedings of the National Academy of Sciences*, **112**,  
743 196–201.
- 744 Loh PR, Lipson M, Patterson N *et al.* (2013) Inferring admixture histories of human  
745 populations using linkage disequilibrium. *Genetics*, **193**, 1233–1254.
- 746 Mallet J, Besansky N, Hahn MW (2015) How reticulated are species? *BioEssays*, 140–149.
- 747 Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation  
748 with gene flow in *Heliconius* butterflies. *Genome research*, **23**, 1817–28.
- 749 Meirmans P, Van Tienderen P (2004) GENOTYPE and GENODIVE: two programs for the  
750 analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.
- 751 Nadeau NJ, Martin SH, Kozak KM *et al.* (2013) Genome-wide patterns of divergence and  
752 gene flow across a butterfly radiation. *Molecular ecology*, **22**, 814–26.
- 753 Ndo C, Antonio-Nkondjio C, Cohuet A *et al.* (2010) Population genetic structure of the  
754 malaria vector *Anopheles nili* in sub-Saharan Africa. *Malaria journal*, **9**, 161.
- 755 Ndo C, Simard F, Kengne P *et al.* (2013) Cryptic Genetic Diversity within the *Anopheles nili*  
756 group of Malaria Vectors in the Equatorial Forest Area of Cameroon (Central Africa).  
757 *PLoS ONE*, **8**, 1–12.
- 758 Neafsey DE, Waterhouse RM, Abai MR *et al.* (2015) Highly evolvable malaria vectors: The  
759 genomes of 16 *Anopheles* mosquitoes. *Science*, **347**, 1258522–1258522.
- 760 Norris LC, Main BJ, Lee Y *et al.* (2015) Adaptive introgression in an African malaria  
761 mosquito coincident with the increased usage of insecticide-treated bed nets.  
762 *Proceedings of the National Academy of Sciences*, 201418892.
- 763 Nosil P (2012) *Ecological Speciation*.
- 764 Nosil P, Feder JL (2012) Widespread yet heterogeneous genomic divergence. *Molecular*  
765 *Ecology*, **21**, 2829–2832.
- 766 Nosil P, Feder JL (2013) Genome evolution and speciation: toward quantitative descriptions  
767 of pattern and process. *Evolution; international journal of organic evolution*, **67**, 2461–  
768 7.
- 769 Nydam ML, Harrison RG (2011) Introgression Despite Substantial Divergence in a  
770 Broadcast Spawning Marine Invertebrate. *Evolution*, **65**, 429–442.
- 771 Ozerov MY, Gross R, Bruneaux M *et al.* (2016) Genome-wide introgressive hybridization  
772 patterns in wild Atlantic salmon influenced by inadvertent gene flow from hatchery  
773 releases. *Molecular Ecology*, n/a–n/a.
- 774 Paradis E, Claude J, Strimmer K (2004) Analyses of Phylogenetics and Evolution in R  
775 language. *Bioinformatics*, **20**, 289–290.
- 776 Parchman TL, Gompert Z, Braun MJ *et al.* (2013) The genomic consequences of adaptive  
777 divergence and reproductive isolation between species of manakins. *Molecular ecology*,  
778 **22**, 3304–17.
- 779 Payseur BA, Rieseberg LH (2016) A Genomic Perspective on Hybridization and Speciation.

- 780 *Molecular Ecology*, n/a–n/a.
- 781 Peery A, Sharakhova M V, Antonio-Nkondjio C *et al.* (2011) Improving the population  
782 genetics toolbox for the study of the African malaria vector *Anopheles nili*:  
783 microsatellite mapping to chromosomes. *Parasites & Vectors*, **4**, 202.
- 784 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An  
785 Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-  
786 Model Species. *PLoS ONE*, **7**, e37135.
- 787 Pickrell JK, Pritchard JK (2012) Inference of Population Splits and Mixtures from Genome-  
788 Wide Allele Frequency Data. *PLoS Genetics*, **8**.
- 789 Pool JE, Corbett-Detig RB, Sugino RP *et al.* (2012) Population Genomics of Sub-Saharan  
790 *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS Genetics*,  
791 **8**.
- 792 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using  
793 multilocus genotype data. *Genetics*, **155**, 945–959.
- 794 Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a toolset for whole-genome  
795 association and population-based linkage analysis. *American Journal of Human*  
796 *Genetics*, **81**.
- 797 Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian  
798 population history. *Nature*, **461**, 489–494.
- 799 Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during  
800 evolutionary diversification as revealed in replicate lake-stream stickleback population  
801 pairs. *Molecular ecology*, **21**, 2852–62.
- 802 Rosenberg N (2004) DISTRUCT: a program for the graphical display of population  
803 structure. *Molecular Ecology Resources*, **4**, 137–138.
- 804 Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: Genomic  
805 hotspots of introgression between two highly divergent *Ciona intestinalis* species.  
806 *Molecular Biology and Evolution*, **30**, 1574–1587.
- 807 Sharakhova M V., Peery A, Antonio-Nkondjio C *et al.* (2013) Cytogenetic analysis of  
808 *Anopheles ovengensis* revealed high structural divergence of chromosomes in the  
809 *Anopheles nili* group. *Infection, Genetics and Evolution*, **16**, 341–348.
- 810 Song Y, Endepols S, Klemann N *et al.* (2011) Adaptive introgression of anticoagulant rodent  
811 poison resistance by hybridization between old world mice. *Current Biology*, **21**, 1296–  
812 1301.
- 813 Sukumaran J, Holder MT (2010) DendroPy: A Python library for phylogenetic computing.  
814 *Bioinformatics*, **26**, 1569–1571.
- 815 Turner TL, Hahn MW, Nuzhdin S V. (2005) Genomic islands of speciation in *Anopheles*  
816 *gambiae*. *PLoS Biology*, **3**, 1572–1578.
- 817 Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population  
818 structure. *Evolution*, **38**, 1358–1370.
- 819 Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in  
820 short reads. *Bioinformatics*, **26**, 873–881.
- 821

822 **Author contributions**

823 Conceived and designed the experiments: CK CF BJW. Performed the experiments: CK CF SG  
824 BJW. Analyzed the data: CK CF BJW. Wrote the paper: CK CF BJW.

825

826 **Tables**

827

828 **Table 1:** Pairwise  $F_{ST}$  between *An. nili* and *An. ovengensis* populations.  $p < 0.005$  for all  
829 values.

| $F_{ST}$                         | <i>An. nili</i><br>group 1 | <i>An. nili</i><br>group 2 | <i>An. nili</i><br>group 3 | <i>An.</i><br><i>ovengensis</i><br>group 1 | <i>An.</i><br><i>ovengensis</i><br>group 2 |
|----------------------------------|----------------------------|----------------------------|----------------------------|--|--|
| <i>An. nili</i> group 1          | -                          |                            |                            |  |  |
| <i>An. nili</i> group 2          | 0.794                      | -                          |                            |  |  |
| <i>An. nili</i> group 3          | 0.838                      | 0.863                      | -                          |  |  |
| <i>An. ovengensis</i><br>group 1 | 0.655                      | 0.834                      | 0.873                      | -  |  |
| <i>An. ovengensis</i><br>group 2 | 0.857                      | 0.863                      | 0.858                      | 0.896                                      | -  |

830

831 **Table 2:** Results of the most significant  $f_4$  tests for gene flow.  
832

| Test  | $f_4 \pm \text{std err}$ | Z-score | P-value   |
|---|--------------------------|---------|-----------|
| $f_4(\text{Oveng 1, Nili 3 ; Nili 2, Nili 1})$  | -0.03506 $\pm$ 0.00287   | -12.20  | < 0.00001 |
| $f_4(\text{Oveng 1, Nili 3 ; Nili 2, Oveng 2})$ | 0.02241 $\pm$ 0.00263    | 8.52    | < 0.00001 |
| $f_4(\text{Oveng 1, Oveng 2 ; Nili 3, Nili 1})$ | -0.05775 $\pm$ 0.00355   | -16.25  | < 0.00001 |
| $f_4(\text{Oveng 1, Oveng 2 ; Nili 2, Nili 1})$ | -0.03718 $\pm$ 0.00292   | -12.72  | < 0.00001 |
| $f_4(\text{Nili 3, Nili 1 ; Nili 2, Oveng 2})$  | -0.02272 $\pm$ 0.00263   | -8.63   | < 0.00001 |

Oveng 1: *An. ovengensis* group 1 ; Oveng 2: *An. ovengensis* group 2 ; Nili 1: *An. nili* group 1 ; Nili 2: *An. nili* group 2 ; Nili 3: *An. nili* group 3

833

834 **Table 3:** Pearson's correlation coefficient assessing the relationship between cline  
 835 parameter ( $\alpha$  and  $\beta$ ) and locus-specific estimates of five population genomic parameters  
 836 (pairwise genetic divergence ( $F_{ST}$  and  $d_{xy}$ ), nucleotide diversity ( $\theta_w$  and  $\theta_\pi$ ) and allele  
 837 frequency spectrum (Tajima's  $D$ )).  
 838

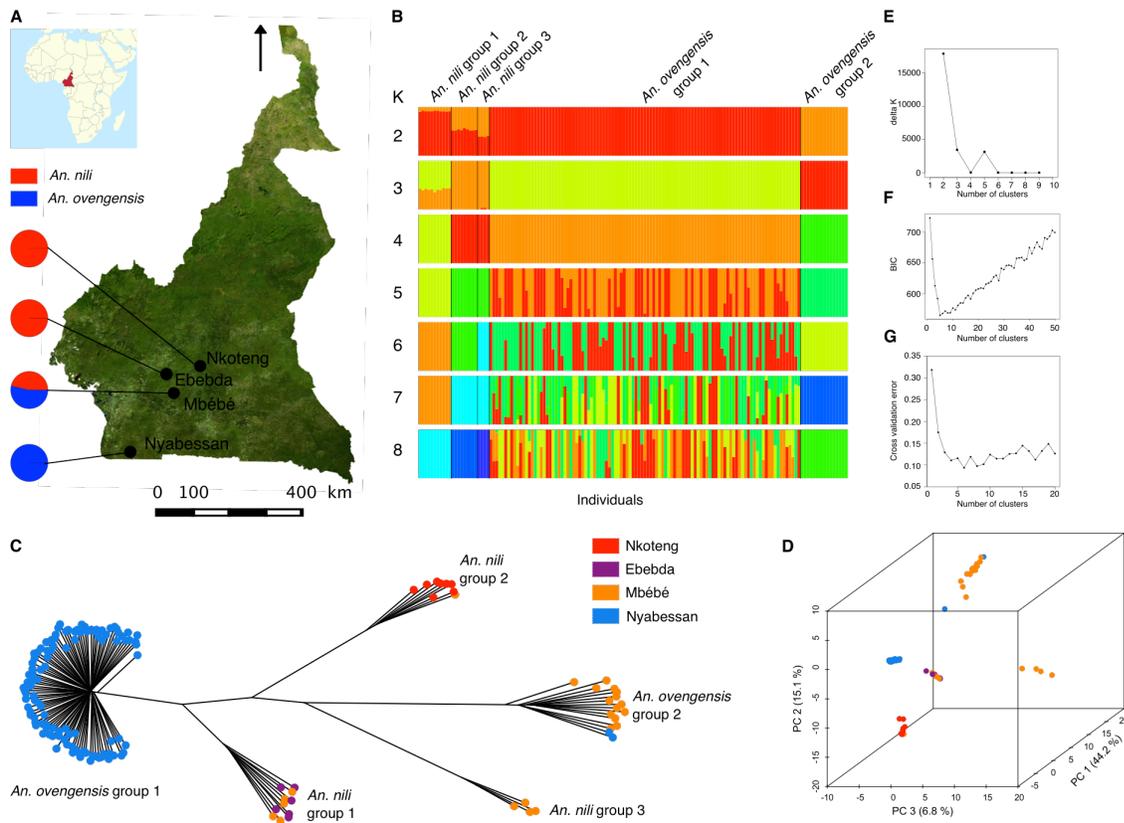
|          | Hybrids  |          |            |              |          |  | <i>An. nili</i> |              |        | <i>An. ovengensis</i> |              |          |
|----------|----------|----------|------------|--------------|----------|--|-----------------|--------------|--------|-----------------------|--------------|----------|
|          | $F_{ST}$ | $d_{xy}$ | $\theta_w$ | $\theta_\pi$ | $TD$     |  | $\theta_w$      | $\theta_\pi$ | $TD$   | $\theta_w$            | $\theta_\pi$ | $TD$     |
| $\alpha$ | -0.151   | -0.130   | 0.057      | 0.052        | 0.016 *  |  | -0.498          | -0.414       | 0.088  | 0.537                 | 0.432        | -0.033 * |
| $\beta$  | 0.060    | 0.048    | -0.016 *   | -0.017 *     | -0.011 * |  | 0.230           | 0.188        | -0.047 | -0.266                | -0.212       | 0.018 *  |

\* not significant ( $p > 0.005$ )

$TD$ : Tajima's  $D$

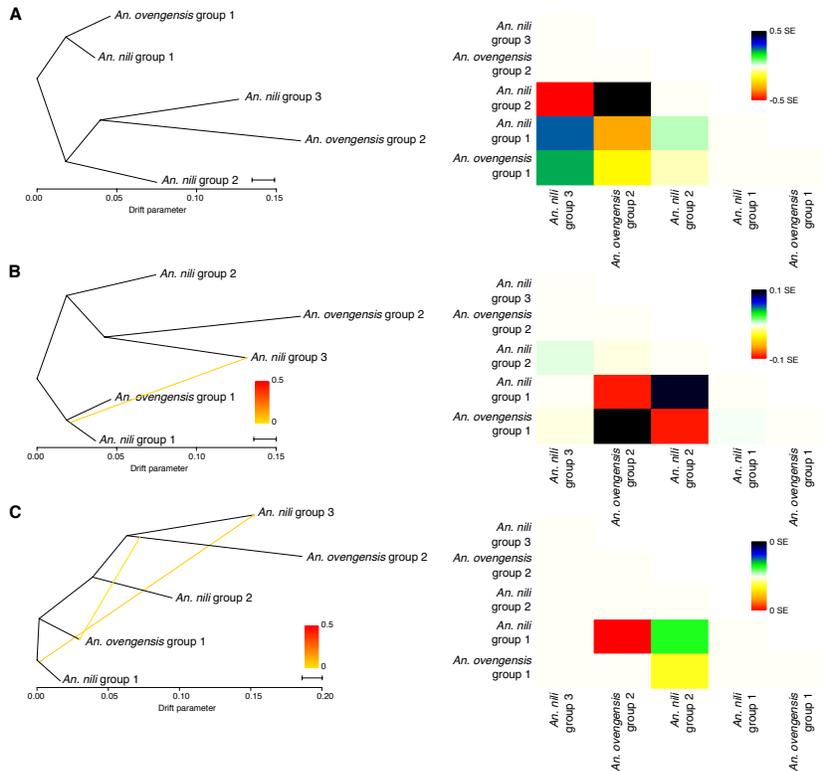
839 **Figures**

840 **Figure 1:** Population genetic structure of *An. nili sensu lato* inferred from 4343 SNPs  
841 identified with a reference-based assembly. (A) Map showing the sampling locations and  
842 relative frequencies of the two *An. nili* and *An. ovengensis*. (B) ADMIXTURE plots with k  
843 from 2 trough 8. (C) and (D) neighbor-joining tree and PCA. Each PCA axis is labeled with  
844 the percentage of variance explained. (E), (F) and (G) Identification of the optimal number  
845 of genetic clusters using the delta k method of Evanno et al, DAPC, and 10-fold cross-  
846 validation in ADMIXTURE. The lowest BIC and CV error and the highest delta k indicate the  
847 most probable number of genetic clusters.



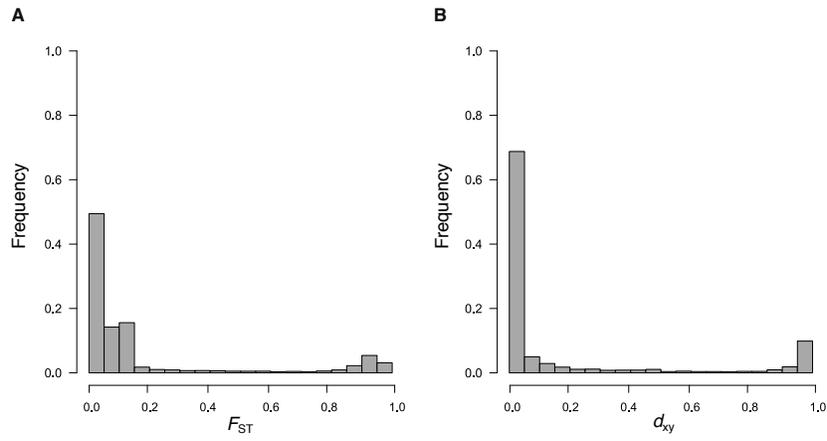
848

849 **Figure 2.** *TreeMix* Maximum Likelihood (ML) trees depicting the signals of gene flow  
850 between *An. nili* and *An. ovengensis*. ML tree and residual fit from the ML model inferred  
851 with (A) no migration edge, (B) a single migration edge and (C) two migration edges. The  
852 small arrow on each indicates the directionality of gene flow migration edge and the color  
853 of the edge reflect the intensity of admixture. Heat colors depict the residual covariance  
854 between each pair of populations. Darker colors indicate populations more closely related  
855 to each other than expected under a bifurcating maximum likelihood tree, suggestive of  
856 gene flow.



857

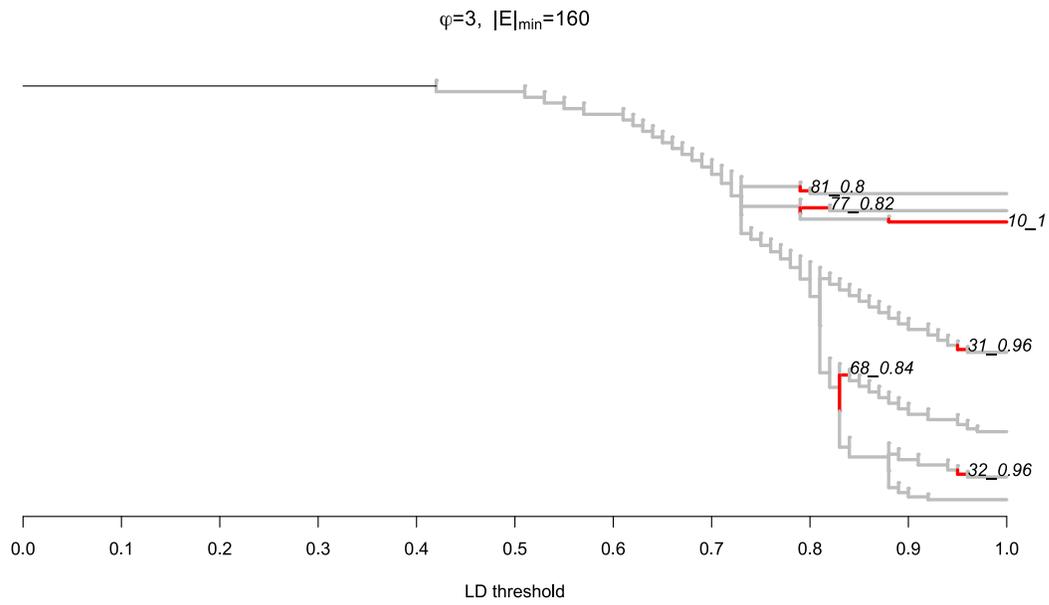
858 **Figure 3:** Frequency distribution of  $F_{ST}$  (A) and  $d_{xy}$  (B) based on 9622 variant sit



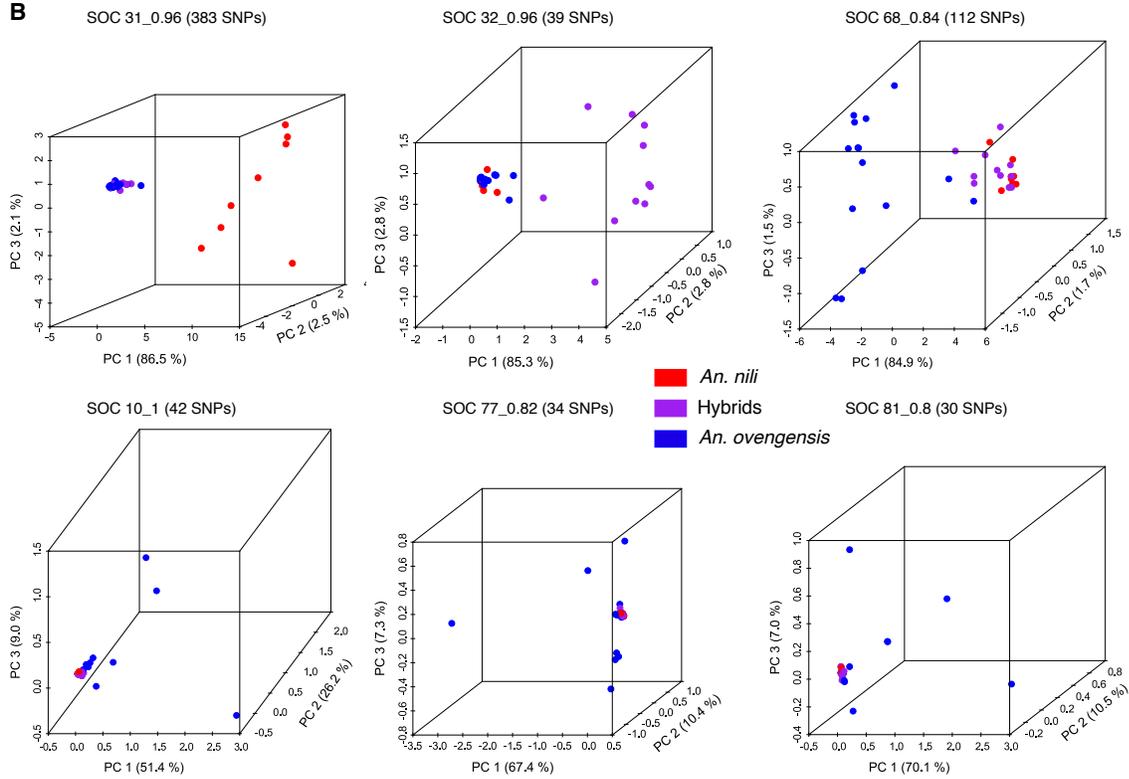
859

860 **Figure 4:** Results of Linkage disequilibrium analyses in LDna. (A) LDna graph suggesting  
861 the presence of 6 LD clusters (Single Outlier Clusters (SOCs)) based on 1330 SNPs in a  
862 dataset containing *An. nili*, *An. ovengensis*, and hybrids of the two species. Values of the two  
863 parameters:  $\varphi$  (which controls when clusters are defined as outliers) and  $|E|_{\min}$ , the  
864 minimum number of edges required for a LD cluster to be considered as an outlier, are  
865 indicated on top of the graph. Corresponding LD thresholds are shown on the x-axis. (B)  
866 Population genetic structure of the six SOC's identified.

**A**

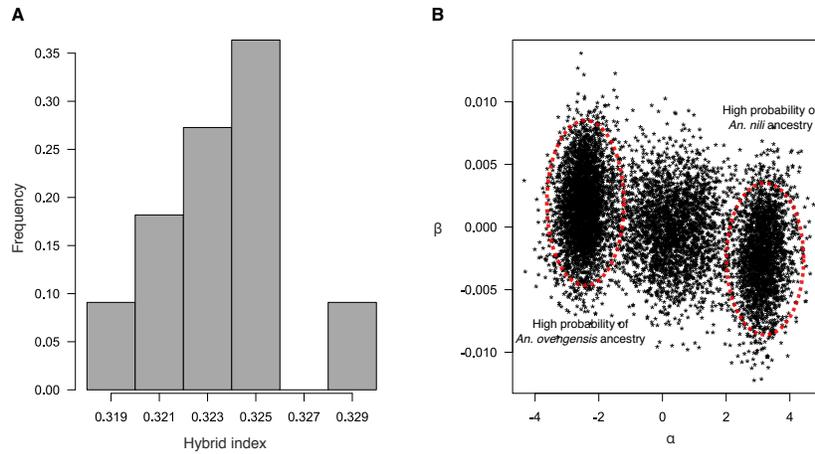


**B**



867

868 **Figure 5:** Genomic cline analysis. (A) Frequency distribution of the hybrid index in admixed  
869 individuals (hybrid index of pure *An. nili* = 1.0 and pure *An. ovengensis* = 0.0). (B)  
870 Prevalence of steep genomic clines illustrated by a scatterplot between the cline  
871 parameters  $\alpha$  and  $\beta$ .



872  
873

874 **Supplemental Material**

875

876 **Table S1:** Information on *An. nili sensu lato* mosquitoes included in this study.

| Sampling locations | Geographic coordinates | Sampling methods |        |    | Total |
|--------------------|------------------------|------------------|--------|----|-------|
|                    |                        | HLC-OUT          | HLC-IN | LC |       |
| Ebebda             | 4°20'00"N, 11°17'00"E  |                  |        | 6  | 6     |
| Nkoteng            | 4°31'00"N, 12°02'00"E  |                  |        | 8  | 8     |
| Nyabessan          | 2°24'00"N, 10°24'00"E  | 63               | 44     |    | 107   |
| Mbébé              | 4°10'00"N, 11°04'00"E  | 13               | 3      | 8  | 24    |
| Total              |                        | 76               | 47     | 22 | 145   |

HLC-OUT, human landing catches performed outdoor; HLC-IN, human landing catches performed indoor; LC, larval collection

877

878 **Table S2:** Pairwise  $F_{ST}$  estimated from a *de novo* assembly.  $p < 0.005$  for all values.

| $F_{ST}$                         | <i>An. nili</i><br>group 1 | <i>An. nili</i><br>group 2 | <i>An. nili</i><br>group 3 | <i>An. ovengensis</i><br>group 1 | <i>An. ovengensis</i><br>group 2 |
|----------------------------------|----------------------------|----------------------------|----------------------------|----------------------------------|----------------------------------|
| <i>An. nili</i> group 1          | -                          |                            |                            |                                  |                                  |
| <i>An. nili</i> group 2          | 0.791                      | -                          |                            |                                  |                                  |
| <i>An. nili</i> group 3          | 0.844                      | 0.862                      | -                          |                                  |                                  |
| <i>An. ovengensis</i><br>group 1 | 0.705                      | 0.861                      | 0.902                      | -                                |                                  |
| <i>An. ovengensis</i><br>group 2 | 0.854                      | 0.862                      | 0.867                      | 0.907                            | -                                |

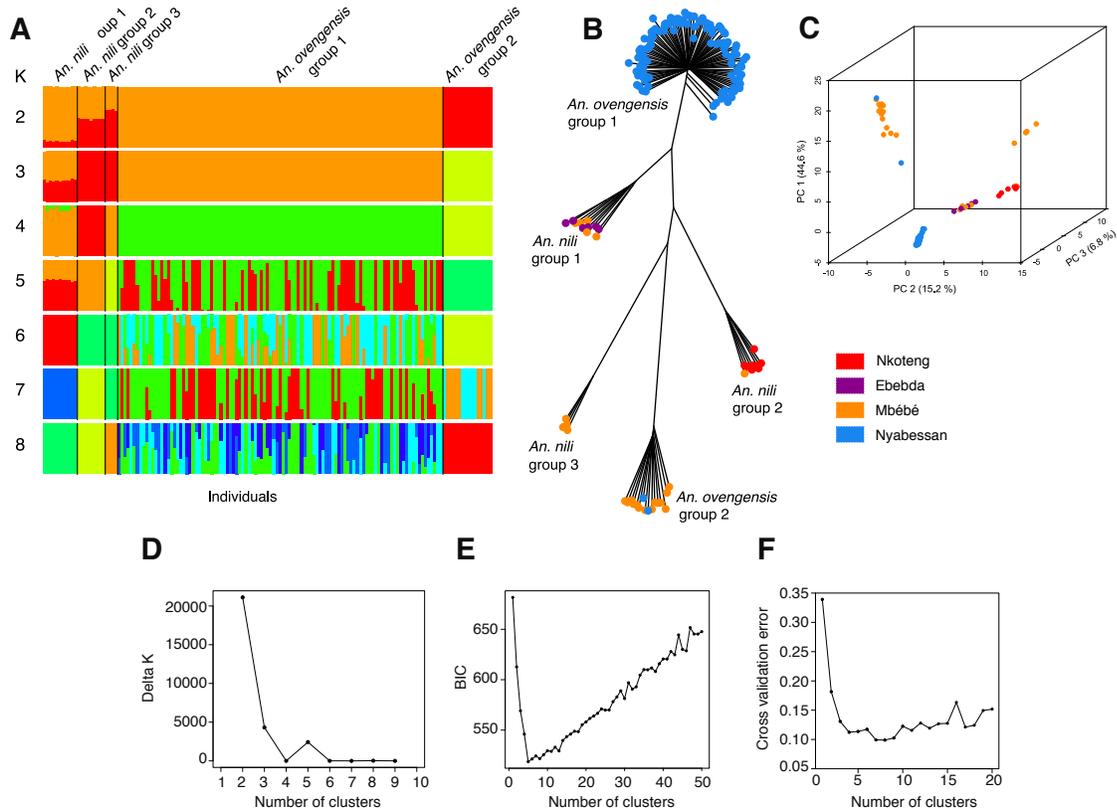
879

880 **Table S3:** Results of the most significant  $f_4$  tests for gene flow (*de novo* assembly).

| Test  | $f_4 \pm \text{std err}$ | Z-score | P-value  |
|---|--------------------------|---------|----------|
| $f_4(\text{Oveng 1, Nili 3 ; Nili 2, Nili 1})$  | -0.03692 $\pm$ 0.00360   | -10.25  | <0.00001 |
| $f_4(\text{Oveng 1, Nili 3 ; Nili 2, Oveng 2})$ | 0.02787 $\pm$ 0.00351    | 7.94    | <0.00001 |
| $f_4(\text{Oveng 1, Oveng 2 ; Nili 3, Nili 1})$ | -0.06610 $\pm$ 0.00451   | -14.64  | <0.00001 |
| $f_4(\text{Oveng 1, Oveng 2 ; Nili 2, Nili 1})$ | -0.04053 $\pm$ 0.00368   | -11.01  | <0.00001 |
| $f_4(\text{Nili 3, Nili 1 ; Nili 2, Oveng 2})$  | -0.02776 $\pm$ 0.00349   | -7.96   | <0.00001 |

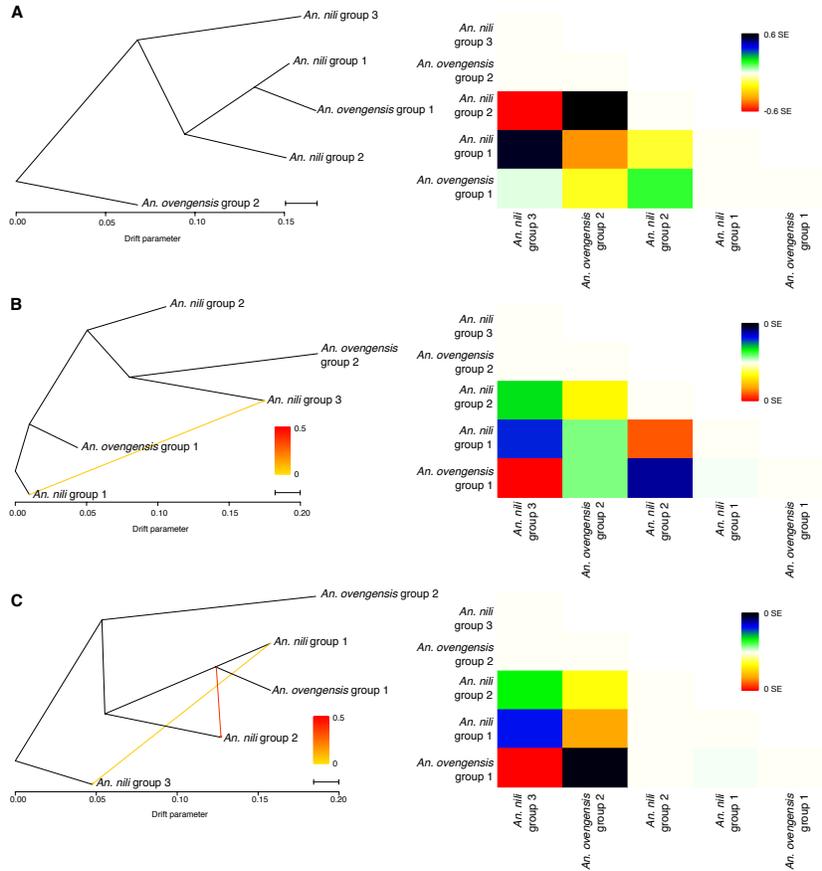
Oveng 1: *An. ovengensis* group 1 ; Oveng 2: *An. ovengensis* group 2 ; Nili 1: *An. nili* group 1 ; Nili 2: *An. nili* group 2 ; Nili 3: *An. nili* group 3

881 **Figure S1:** Population genetic structure of *An. nili sensu lato* inferred from 3071 SNPs  
882 identified with a *de novo* assembly. (A) ADMIXTURE plots with k from 2 through 8. (B) and  
883 (C) neighbor-joining tree and PCA. (D), (E) and (F) Identification of the optimal number of  
884 genetic clusters using the delta k method of Evanno et al, DAPC, and a 10-fold cross-  
885 validation in ADMIXTURE.



886

887 **Figure S2:** *TreeMix* Maximum Likelihood (ML) trees estimated from 3071 SNPs identified  
 888 with a *de novo* assembly. ML tree and residual fit from the ML model inferred with (A) no  
 889 migration, (B) a single migration and (C) two migration edges. See Fig. 2. in the main text  
 890 for additional description.



891