

1 **Title:** A new species in the *Anopheles gambiae* complex reveals new evolutionary
2 relationships between vector and non-vector species

3

4 **Running title:** Vectorial evolution in the *An. gambiae* complex.

5

6 **Authors**

7 Maite G Barron¹, Christophe Paupy², Nil Rahola^{2,3}, Ousman Akone-Ella³, Marc F.
8 Ngangue^{3,4}, Theodel A. Wilson-Bahun³, Marco Pombi⁵, Pierre Kengne², Carlo
9 Costantini², Frédéric Simard², Josefa Gonzalez^{1,*} & Diego Ayala^{2,3,*}.

10 **Affiliations**

11 ¹ IBE (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

12 ² MIVEGEC, IRD, CNRS, Univ. Montpellier, Montpellier, France

13 ³ CIRMF, Franceville, Gabon.

14 ⁴ ANPN, Libreville, Gabon

15 ⁵ Università di Roma “Sapienza”, Rome, Italy

16

17 *** Co-last and co-corresponding authors:**

18 Diego Ayala, MIVEGEC, IRD, CNRS, Univ. Montpellier, 911 av Agropolis, BP
19 64501, 34394 Montpellier, France;

20 email: diego.ayala@ird.fr.

21 Josefa González, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra),
22 Passeig Marítim de la Barceloneta 37-49. 08003 Barcelona, Spain;

23 email: josefa.gonzalez@ibe.upf-csic.es.

24

25 **Key-words:** *Anopheles gambiae*, speciation, local adaptation, evolution, malaria

26

1 **Abstract**

2 Complexes of closely related species provide key insights about the rapid and
3 independent evolution of adaptive traits. Here, we described and studied a presumably
4 new species in the *Anopheles gambiae* complex, *Anopheles fontenillei*, recently
5 discovered in the forested areas of Gabon, Central Africa. Our analysis placed the new
6 taxon in the phylogenetic tree of the *An. gambiae* complex, revealing important
7 introgression events with other members of the complex. In particular, we detected
8 recent introgression with *An. gambiae* and *An. coluzzii* of genes directly involved in
9 vectorial capacity. Moreover, genome analysis of the new species also allowed us to
10 resolve the evolutionary history of inversion 3La. Overall, *Anopheles fontenillei* has
11 implemented our understanding about the relationship of species within the *gambiae*
12 complex and provides insight into the evolution of vectorial capacity traits, relevant
13 for the successful control of malaria in Africa.

14

15

16

1 **Introduction**

2 Species at earlier stages of speciation provide unique insights about the evolutionary
3 forces involved in the origin of new species before demographic and selective
4 processes blur the signals. However, the closer we are to the first signals of
5 divergence, the harder it is to define the species concept, or even to intuit if this
6 process will end in speciation [1]. Complex of species, closely related taxa where the
7 species boundaries are uncertain, offers a compelling opportunity to study the
8 “speciation continuum” [1, 2]. Unfortunately, the reproductive isolation is still
9 incomplete to fully prevent introgression between taxa, hindering the true
10 phylogenetic relationships [3]. On the other hand, the genetic exchange across
11 backcrossed hybrids can favor adaptation [4]. Indeed, advantageous alleles can be
12 selected in one species and introgressed in other favouring range expansions [5],
13 altitudinal adaptation [6], or insecticide resistance [7], among others traits.
14 Most of the major malaria vectors across the world belong to species complexes with
15 other non-vector species [8], providing a compelling opportunity to understand the
16 rapid and independent evolution of their vectorial capacity [9, 10]. While malaria
17 mosquitoes exhibit a wide ecological plasticity, preference for feeding on humans,
18 and large population size, the non-vectors species display narrower geographical
19 range, zoophilic host preference, and strong seasonal-dependence or reduced
20 population size [11]. In Africa, three of the six major malaria vectors belong to the
21 same complex, the *Anopheles gambiae*: *Anopheles gambiae*, *Anopheles coluzzii* and
22 *Anopheles arabiensis* [12]. The complex is comprised of eight cryptic species [13-15],
23 which differ in many ecological aspects, particularly in host feeding preference,
24 breeding sites, feeding behavior and their role in malaria transmission [13, 16]. Most
25 of the species inhabit natural habitats with none or a secondary role in malaria
26 transmission. Indeed, adaptation to anthropogenic habitats, and therefore
27 implementing their role in human malaria transmission, is an exception rather than the
28 rule within the complex [16]. The *An. gambiae* complex is an example of speciation
29 with gene flow, where species exhibit extensive genomic introgression, evidencing
30 permeable gene flow barriers among them [3, 10], sustained by heterogenic patterns
31 of reproductive isolation [17]. Consequently, pervasive introgression has hindered the
32 elucidation of the correct phylogenetic relationships [18]. Besides, gene exchange
33 between species in the complex has modulated their local adaptation capacity. For
34 instance, the ability of *Anopheles arabiensis* to live in desiccating environments has

1 been conferred by the introgression of the inversion 2La from *An. gambiae*/*An.*
2 *coluzzii* [10, 19]. *Anopheles coluzzii* has developed resistance to insecticide treatments
3 due to the introgression of the *kdr* mutation from *An. gambiae* [7]. Onwards,
4 insecticide resistance introgressions have repeatedly occurred during the last decades
5 [20, 21]. Thus, introgression has accelerated local adaptation and range expansion
6 within the complex. Complex of closely related species also offer a compelling
7 opportunity to study locally adapted alleles. Comparative genomics in recent species
8 radiations allows unraveling the genetic basis of the traits involved in their ecological,
9 behavioral or genetic divergence [22]. In *Anopheles*, these comparative studies has
10 contributed to elucidate some traits involved in vectorial capacity, that in turn could
11 be used to improve vector control strategies [9]. For instance, antennal transcriptomic
12 comparisons between *Anopheles gambiae* and *An. quadriannulatus* have provided
13 genomic insights on host preference evolution to humans [23]. Moreover, genome
14 wide analysis between one fresh-water (*An. gambiae*) and one salt water (*An. melas*)
15 species has allowed to identify regions involved into the salinity tolerance within the
16 complex [24]. Therefore, understanding the origin and the mechanisms underlying
17 vectorial capacity within the *Anopheles gambiae* complex is decisive for the
18 successful control of malaria in Africa [16, 25].

19
20 During an exploratory survey at La Lope National Park (Gabon) in 2014, we
21 discovered mosquitoes morphologically identified as *An. gambiae*. Further bio-
22 ecological, behavioral, taxonomic, cytogenetic, and preliminary molecular analysis
23 suggested the probable existence of a new taxon in the *An. gambiae* complex. Then,
24 genome-wide phylogenetic analysis placed this potential new taxon in the phylogeny
25 of the complex as a sister species of *Anopheles bwambae*, and in the same clade as
26 *Anopheles quadriannulatus*, *An. arabiensis*, and *An. melas*. Comparative genomic
27 analysis indicated the existence of recent introgression between the potential new
28 species and *An. gambiae*/*An. coluzzii*. Genes involved were enriched for
29 detoxification, desiccation, and olfactory perception functions, directly linked to local
30 adaptation and host preference. These analyses also elucidated the evolutionary
31 history of the 3La inversion within the complex that entailed multiple lost events.
32 Overall, the discovery of a probable new taxon has evidenced the importance of new
33 species for the understanding of evolutionary relationships of species in the *An.*

1 *gambiae* complex with potential implications for a better understanding of vectorial
2 capacity traits and ultimately malaria control.

3

4 **Results**

5 All the specimens morphologically identified as *An. gambiae* belonged to a potential
6 unknown taxon, within the *An. gambiae* complex, hereafter called *An. fontenillei* n.sp.

7 This species is dedicated to our colleague Didier Fontenille and his wife Marielle,
8 medical entomologist, who has greatly contributed to the study of mosquitoes and the
9 development of medical entomology in Africa.

10

11 *Bio-ecology of An. fontenillei*

12 We prospected 22 sites in the National Park of La Lope in Gabon: 17 sites in the park
13 and 5 sites in the village of La Lope, 10-15 km away from the park sites. In total, we
14 collected more than 1,500 mosquitoes, belonging to 13 different species. Of them, 45
15 adults and two larvae were morphologically identified as *An. gambiae* but presented
16 an unexpected DNA band in the PCR assay for identification of the *Anopheles*
17 complex species [26]. In Gabon, only three species of the *gambiae* complex have been
18 recorded, and all of them can be identified based on specific PCR bands [27, 28]. The
19 individuals of the unknown species were found in six natural sites across the park,
20 away of any human activity or presence (Fig. 1, Table S1). All the positive sites were
21 in the edge of forest patches and close to natural marshes frequented by wild animals
22 (*e.g.* African forest buffalos and other ungulates). The presence of larvae belonging to
23 the potential new species in two of these marshes suggested a particular affinity for
24 sunny clay soil water collections containing turbid fresh-water of rain origin. Another
25 *Anopheles* species, *Anopheles maculipalpis*, was collected in the same breeding site.
26 This species is known to breed in sunny, low oxygen and generally stagnant water and
27 it has already been found breeding in sympatry with *An. gambiae* [29]. This typology
28 of larval habitat is very similar to that of *An. gambiae*, *An. coluzzii*, *An. arabiensis*
29 [30], but different to the other members of the complex, such as *An. merus* or *An.*
30 *melas* (mangrove swamps) or *An. bwambae* (hot thermal springs), which place the
31 new taxon in the fresh-water group of species within the *An. gambiae* complex [16].
32 Although no blood-fed mosquitoes were found, we assumed a preference for feeding
33 on animals (zoophily) due to the lack of human hosts in the sylvatic sites. Mosquitoes
34 were sampled using BG® traps baited with BG-lure a source of CO₂ [31] and Human

1 Landing Catches –HLC– (Fig. 1B), revealing that the potential new species can feed
2 on humans as well. Moreover, our collections in the village of La Lope (~10-15 km
3 away of the park sites) revealed the presence of two other members of the complex
4 (*An. gambiae* and *An. coluzzii*). No specimen of *An. fontenillei* was found in the
5 village (HLC and larva prospections).

6

7 *Brief taxonomic description*

8 Five *Anopheles fontenillei* specimens were preserved for taxonomic purposes (Table
9 S1, holotypes deposited at the IRD in Montpellier, France). In general, *An. fontenillei*
10 presents the classical morphotype of species within the *An. gambiae* complex [15, 32,
11 33]: three white-scaled bands in the maxillary palpus, irregularly shaped speckling in
12 femora and tibiae and a pale interruption on vein R₁ (Fig. 1C) (for further details see
13 Text S1). However, small differences were detected. In particular, the maxillary
14 palpus exhibited a large white-scaled band covering completely the palpomere 5 and
15 part of the palpomere 4 (Fig. 1C), similarly to *An. bwambiae* [34].

16

17 *Cytogenetic analysis*

18 In order to confirm the species status and its phylogenetic relationships within the
19 *gambiae* complex, we collected 270 sylvatic *Anopheles* for cytogenetic purposes.
20 Forty mosquitoes survived to attain the correct stage (half-gravid) to observe polytene
21 chromosomes. Among them, four mosquitoes were morphologically identified as
22 belonging to the *An. gambiae* complex, but only three revealed readable polytene
23 chromosome preparations. According to the classical nomenclature for chromosomal
24 rearrangements in the *An. gambiae* complex [35], all the specimens exhibited the X
25 chromosome and the 2L arm standard arrangements, and the inversions 3Rb and 3La
26 were fixed. In addition, the inversion 2Rl was polymorphic: inverted in one specimen
27 and standard in the other two mosquitoes (Fig. 1D, Fig. S1). For the 2La inversion, a
28 molecular karyotyping test is available [36]. We then used five additional specimens
29 to validate the status of the inversion 2La [36]. All the specimens revealed a PCR-
30 band consistent with the 2La standard arrangement, confirming our cytogenetic
31 karyotype. Globally, *Anopheles fontenillei* revealed a karyotype similar to *An.*
32 *bwambiae* [34], except for the possibility that the inversion 3Rb is fixed in the new
33 taxon, while it is polymorphic in *An. bwambiae*. Further cytogenetic works with a

1 bigger number of individuals will be necessary to confirm the inversion
2 polymorphisms of this species.

3

4 *Preliminary phylogenetic analysis*

5 Sixteen specimens were used to obtain sequences for the nuclear ITS2 and IGS and
6 the mitochondrial ND5 and COI regions, routinely used for *Anopheles* phylogenetic
7 studies. Nevertheless, ITS2 and ND5 regions were successfully amplified and
8 sequenced only for nine and five specimens respectively (Table S1). Overall, all the
9 genes exhibited a low diversity with an unique haplotype, except the COI gene, which
10 presented 5 haplotypes. The phylogenetic trees showed that *An. fontenillei* sequences
11 always clustered with *An. bwambae* within a monophyletic clade (Figure S2),
12 corroborating the previous cytogenetic results but in contrast with the ecological
13 observations. Two of the four genes, ITS2 and ND5, revealed differences between *An.*
14 *fontenillei* and *An. bwambae* (Figure S2). These results are in congruence with
15 previous studies revealing that most of the classical molecular markers are not
16 discriminant among species in the complex due to their extensive introgression [10].

17

18 Overall, the new taxon revealed important similarities to *An. bwambae*, a thermal
19 spring breeding species from a forested area of Uganda (Semliki valley). Taxonomic
20 (large band in the palpomeres 4 and 5), cytogenetic (chromosomal inversions) and
21 molecular (sequence divergence) criteria could not differentiate between the two
22 species. On the other hand, ecological (freshwater marshes vs thermal springs) and
23 geographical (allopatric distribution: Gabon vs Uganda) results clearly discriminated
24 between *An. fontenillei* and *An. bwambae*. Therefore, further genomics studies are
25 needed to elucidate the true phylogenetic place of *An. fontenillei* within the *An.*
26 *gambiae* complex.

27

28 *Anopheles fontenillei* is a potential new species of the *Anopheles gambiae* complex.

29 We conducted a genome-wide analysis in order to accurately locate the new species in
30 the *An. gambiae* complex phylogenetic tree. According to previous studies [10], we
31 considered that the true *An. gambiae* complex species tree is mainly observed in the X
32 chromosome. Hence, we initially focused on analyzing the X chromosome arm. For
33 this purpose, we made a genome assembly of one *An. fontenillei* individual sequenced
34 at high coverage (~112X) (Table S2). This assembly was nearly complete, according

1 to the 96% of completely found BUSCO genes, but highly fragmented with a N50 of
2 21kb (Materials and methods, Table S3B and Table S3C). We then added to the
3 available multiple alignment file (MAF), based on six described gambiae complex
4 species [10], our *An. fontenillei* assembly and the highest coverage *An. bwambiae*
5 individual publicly available (see Material and Methods). Maximum likelihood (ML)
6 phylogenetic trees were built for each non-overlapping 50kb windows (see Materials
7 and Methods, Table S4).

8 Following this approach, the relationship among species observed in the X
9 chromosome was as shown in Figure 2. Similarly to previous studies, the relative
10 position of the basal node, *An. merus* and *An. coluzzii*-*An. gambiae* clade, was not
11 clearly determined due to incomplete lineage sorting (ILS, [10]). In our analysis of the
12 X chromosome, *An. fontenillei* appeared as the sister species of *An. bwambiae* in 83%
13 of the trees (264 out of 319). However, there exists a certain ambiguity to determine
14 the ancestral taxon of the clade. While, the clade branches with *An. quadriannulatus*
15 in 78 of 319 windows (Figure 2), it branches with *An. arabiensis* in 59 of 319
16 windows (Fig. S3). Assuming that the X chromosome shows the true species tree, we
17 then presume that either *An. quadriannulatus* or *An. arabiensis*, shared a common
18 ancestor with *An. fontenillei* and *An. bwambiae*.

19 To investigate if we find a stable distinction between *An. fontenillei* and *An.*
20 *bwambiae*, we repeated the analysis creating a new MAF adding 3 additional
21 individuals of *An. fontenillei*, and 2 additional individuals of *An. bwambiae* (see
22 Material and Methods). Out of the 343 analyzed windows, 278 (81%) showed trees
23 where individuals of *An. fontenillei* and *An. bwambiae* clustered together and these
24 two species were permanently separated, indicating that they are different populations
25 and/or species (Fig. S4).

26 We also estimated the pairwise genetic distance between *An. fontenillei* and *An.*
27 *bwambiae* and compared it with the pairwise genetic distance between *An. coluzzii* and
28 *An. gambiae*, the most recently diverged species within the complex (Fig. S5A) [10,
29 37]. The pairwise genetic distance was significantly larger in the *An. fontenillei* - *An.*
30 *bwambiae* clade compared with the *An. gambiae* - *An. coluzzii* clade (bootstrapping
31 analysis, median 0.0117 and 0.0067 respectively, Figure S5B, Table S5). If we
32 assumed a substitution rate of 1.1×10^{-9} per site, per generation, and 10 generation per
33 year [38], there had been 0.53 Ma since the *An. fontenillei* - *An. bwambiae* clade split,
34 and 0.31 Ma since the *An. gambiae* - *An. coluzzii* clade split (Fig. 2 and Figure S3).

1 This result together with the clear ecological distinction between *An. fontenillei* and
2 its closest species within the complex, *An. bwambae*, suggested that *An. fontenillei* is
3 a new species in the *An. gambiae* complex rather than a sub-population of the *An.*
4 *bwambae* species.

5

6 *Recent and ancestral relationship of An. fontenillei with other members of the*
7 *complex.*

8 We extended our analysis from the X chromosome to the whole genome. In 84% of
9 the analyzed genome, *An. bwambae* is the closest species to *An. fontenillei*, forming
10 the *An fontenillei- An. bwambae* (FB) clade (Fig. 3, R line, Figure S6, Table S5). This
11 proportion is similar in every chromosome arm ranging from 78.4% in the 3R
12 chromosome arm to the 86.6% in the 3L chromosome arm. The proportion of FB
13 clade in the autosomes, 84.1%, is in concordance with the FB clade proportion in the
14 X chromosome, 82.8%, *i.e* the species tree, indicating that *An. fontenillei* has not
15 extensively introgressed with other members of the complex in a recent period.

16 However, the relationship of the FB clade with its closest species or other clades,
17 showed a very different pattern between the X chromosome and the autosomes
18 (Figure 3, A line). In the X chromosome, the majority of windows showed the species
19 tree, as it was previously described [10]. Accordingly, FB clade is closely related to
20 *An. quadriannulatus* (27.5%) or *An. arabiensis* (24.5%). While the autosomes, the
21 majority of windows showed the recent introgression between *An. arabiensis* and *An.*
22 *gambiae – An. coluzzii* clade, the A(GC) clade[10]. In the autosomes, the FB clade is
23 branching with the A(GC) clade for the majority of windows, 27.6% (Fig. 3, Figure
24 S6, Table S6). The next more frequent topology, 16.4%, shows the FB clade with *An.*
25 *quadriannulatus* as the closest species. However, if we do not take into account the
26 2La inversions (see below), which shows its own topology, this proportion was
27 reduced to 9.6%. According to the X chromosome analysis, we could not conclude
28 whether *An. quadriannulatus* or *An. arabiensis* is the closest species to the FB clade
29 due to similar number of windows showing one or the other topology (Fig. 2, Fig.
30 S3). However, in the autosomes we can clearly observed that the FB clade is more
31 frequently branching with A(GC) clade. Hence, if *An. quadriannulatus* is the closest
32 species to the FB clade, the FB common ancestor must have suffered introgressions
33 with *An. arabiensis* prior to the *An. arabiensis* and *An. gambiae-An. coluzzii*
34 introgressions. On the contrary, if *An. arabiensis* is the closest species to the FB clade,

1 there is no need of additional introgression to explain the observed results. Only that,
2 the split of the FB common ancestor with *An. arabiensis* must have been prior to the
3 introgressions between *An. arabiensis* and the GC clade.

4 Most of the windows that do not show the FB clade are located close to the
5 centromeric ends. This is mainly observed, in the 2R and 3R last ~11Mb close to the
6 centromere and, a similar pattern is observed in the first ~10Mb close to the
7 centromere of the 2L chromosome arm (Fig. 3, R line). In these regions, the
8 proportion of windows showing the FB clades is smaller than in the rest of the
9 chromosome. Interestingly, another difference is that the proportion of trees showing
10 *An. fontenillei* close to GC clade or either *An. gambiae* or *An. coluzzii* is substantially
11 bigger than in the rest of the chromosome. Specifically, in regions close to the
12 centromeres, the FB clade proportion is ~40% for 2R and 3R, and, 51% for 2L
13 chromosome arm while in the rest of the three chromosome arms, the FB clade
14 proportion is > 80%. The proportion of trees showing *An. fontenillei* close to GC
15 clade or *An. gambiae* or *An. coluzzi* is ~20% for 2R and 3R, and, 7% for 2L
16 chromosome arm while on the rest of the three chromosome arms this proportion is <
17 1%. We checked that the alignment quality of these regions were not different from
18 other regions in the chromosome to exclude possible biases due to low quality
19 alignments (Fig. S7, Materials and Method). The alignments in these regions are
20 shorter but still have on average 16,482 informative positions per window and they
21 showed better alignment qualities than in the other regions of the chromosome arm, in
22 terms of the proportion of gaps or alignment fragmentation. Hence, neither low
23 quality nor short alignments are likely to be the cause of the observed differences
24 (Fig. S7, Materials and Methods).

25 Although we cannot discard that these regions are a consequence of incomplete
26 lineage sorting, it is difficult to explain why the FB clade appears close to the GC
27 clade repeatedly. If we remove the FB clade from the analysis, we cannot observe any
28 differences in these regions compared to the rest of the genome. We argue that these
29 windows may indicate a very recent introgression between *An. fontenillei* and *An.*
30 *gambiae* or *An. coluzzii*, or both.

31

32 *Recent introgressed genes are enriched in metabolic detoxification, desiccation, and*
33 *olfactory perception*

1 We analyzed the gene content of windows were *An. fontenillei* instead of branching
2 with its closer species, *An. bwambae*, clustered with the major malaria vectors i) *An.*
3 *gambiae*, ii) *An. coluzzii*, or iii) the GC clade. These species occur in sympatry at La
4 Lope area, so it could be possible that they share DNA through secondary contact. We
5 analyzed the three ML tree topologies related with this possible recent introgression
6 separately because the presence of the 2La polymorphic inversion may affect the
7 results: the inversion breaks apart the more frequently observed GC clade, because the
8 *An. coluzzii* individuals used for this study were predominantly inversely oriented
9 while *An. gambiae* individuals were predominantly standardly oriented.

10 i) There were 64 windows harboring 198 genes were *An. fontenillei* branched with *An.*
11 *gambiae*. We performed a functional enrichment analysis of the 198 genes, with
12 DAVID, using *An. gambiae* genome as background [39, 40]. There were four
13 significant clusters (Table 1). The first three clusters were related with cuticle
14 proteins, membrane transporter activity, peptidases and proteases. All these protein
15 families had been related to metabolic detoxification of insecticides in high-
16 throughput genome-wide studies in several mosquito species (reviewed in [41]).
17 Additionally, the cuticle proteins had also been described as being critical for the
18 desiccation tolerance in embryos [42]. Interestingly, the GO term of the peptidases
19 and proteases cluster had been previously related with high evolutionary rates [9]. The
20 last cluster, was related to *heat shock protein 70*, a conserved protein related to heat
21 stress but also to oxidative stress and detoxification of some toxins [43, 44]. The
22 InterPro domain in this cluster has also been shown to be a rapid evolving gene family
23 (Table1, [9]).

24 ii) There were 25 windows containing 62 genes where *An. fontenillei* clusters with *An.*
25 *coluzzii*. In these case, none of the clusters were significantly enriched for a particular
26 functional term. Finally, iii) there were 35 autosomal windows containing 89 genes in
27 which *An. fontenillei* branched with the GC clade. If these windows are actually
28 regions of recent introgression, this would mean that those genes were introgressed
29 between *An. gambiae* and *An. coluzzii* common ancestor with *An. fontenillei*. The
30 functional term enrichment analysis, showed two significantly enriched clusters
31 (Table 1). The most significant cluster was enriched in Flavin monooxygenase, which
32 share function similarity with the cytochrome P450-monooxygenases [45]. The P450
33 proteins are one of the main protein families related to metabolic detoxification of
34 insecticides in mosquito species (reviewed in [41]). The other significant cluster is

1 related to olfaction. Three of the four GO terms in the cluster (GO:0050911,
2 GO:0004984 and GO:0005549) and the InterPro domain (Table 1) had been described
3 to show high evolutionary rates [9].

4 Following a reverse complementary approach, we also checked whether known
5 mutations that confer resistance to insecticides or to some infections, both traits
6 relevant for malaria transmission, were present in *An. fontenillei*. Specifically, we
7 checked 42 mutations in 14 genes related to insecticide resistance [20] and five
8 mutations in one gene related to immunity and infection resistance [46]. We map the
9 four *An. fontenillei* individuals to the reference genome (AgamP3) with *bwa-mem*
10 (Text S2.4). Only two mutations in two genes related with insecticide resistance were
11 found in this analysis (Table S7). GSTE6 mutation (E89D) and GSTE3 mutation
12 (N73I) were observed in the four sequenced *An. fontenillei* individuals. The GST
13 protein family, together with the P450 family, are considered to be determinant in the
14 metabolic detoxification of insecticides in mosquitos (reviewed [41]). We checked
15 whether this mutation was present in the other members of the complex. All the
16 available genome references, *An. gambiae* Pimperena, *An. coluzzii*, *An.*
17 *quadriannulatus*, and *An. arabiensis*, showed the susceptible mutation. However, in
18 the MAF made with wild specimens, all the species that could map to those regions,
19 *An. gambiae*, *An. coluzzii* and the three *An. bwambae* individuals showed the resistant
20 alleles, as did the four *An. fontenillei* individuals. This showed that this mutation is
21 polymorphic in several species within the complex and suggest that the resistant
22 phenotype should thus also be shared by all these species.

23

24 *Chromosome inversions reveal putative introgression events in the Anopheles*
25 *gambiae* complex.

26 There are two main inversions in the *An. gambiae* complex, which emerged in our
27 phylogenetic analysis, and shaped the chromosomal evolution within the complex: the
28 2La inversion and the 3La inversion. The 2La inversion has only been described in
29 *An. arabiensis*, *An. gambiae*, and *An. coluzzii* [35]. Neither *An. bwambae* nor *An.*
30 *fontenillei* have this inversion. Hence, in this region of the 2L chromosome arm the
31 FB clade is closer to *An. quadriannulatus*, defining a well-determined different block
32 easily distinguishable in Fig. 3 (line A). On the contrary, *An. fontenillei* had the 3La
33 inversion fixed as well as *An. bwambae* and *An. melas*, as shown by the cytogenetic
34 results. In the 3L chromosome arm the inverted region can also be easily identified

1 because in those windows, the FB clade is closely related to *An. melas* (Figure 3, line
2 A). The inferred breakpoints based on the ML tree topology of the 2La and 3La are
3 inside the known cytological breakpoint ranges, except for the 2L telomeric
4 breakpoint, which was 400kb shorter (Table S8, [35], VectorBase.org).
5 The majority of windows, 45%, in the 3L chromosome arm showed the three known
6 *Anopheles gambiae* complex species with the 3La inversion, *An. fontenillei*, *An.*
7 *bwambae* and *An. melas*, together and separated from the species without the
8 inversion: *An. arabiensis*, *An. quadriannulatus*, *An. merus*, *An. gambiae*, and, *An.*
9 *coluzzii* (Figure 4, Figure S6). Additionally, this topology also suggests two events of
10 introgression, i) *An. arabiensis* with the *An. gambiae* and *An. coluzzii* common
11 ancestor and ii) *An. merus* with *An. quadriannulatus* (Figure 4). To date the 3La
12 inversion, we estimated the pairwise distances between *An. fontenillei* and *An.*
13 *quadriannulatus* in the 3L chromosome arm outside and inside the inversion (3L: 14.5
14 - 35.9 Mb+/- the 500Kb flanking region). Outside the inversion, the divergence
15 between *An. fontenillei* and *An. quadriannulatus* was 1.4 Ma (+/- 0.91), which is
16 similar to the one estimated in the more common X chromosome phylogenetic tree
17 between *An. fontenillei* and *An. quadriannulatus* (1.25 Ma (+/- 0.54), Figure S8, Table
18 S9). We then estimated the divergence between *An. bwambae* and *An.*
19 *quadriannulatus* outside the inversion and again it was similar to the one previously
20 estimated for the more common X chromosome phylogenetic tree between these two
21 species: 1.24 Ma (+/- 0.6). However, the divergence estimated inside the inversion
22 between *An. fontenillei* and *An. quadriannulatus*, and *An. bwambae* and *An.*
23 *quadriannulatus* were 2.53Ma (+/- 0.97), and 2.23 Ma (+/- 0.76), respectively. These
24 estimates are on the range of the *Anopheles gambiae* complex origin around 2 (+/-
25 0.64) Ma ago. We repeated this analysis using *An. arabiensis* instead of *An.*
26 *quadriannulatus* and we obtained similar results (Table S9). We could not accurately
27 date the 3La inversion with this method due to the high uncertainty, but we could
28 show that the inversion is at least older than *An. melas*, *An. arabiensis*, *An.*
29 *quadriannulatus*, *An. bwambae*, and *An. fontenillei* group. However, *An. arabiensis*
30 and *An. quadriannulatus* showed the standard karyotype of the 3La inversion.
31 According to the phylogenetic trees, we thus hypothesized that the ancestral
32 karyotype of the group is the 3La inversion and that *An. quadriannulatus* lost the
33 inversion in the introgression from *An. merus*, as has already been suggested by

1 Fontaine et al. [10], and that *An. arabiensis* lost the inversion in the introgression with
2 *An. gambiae/An. coluzzii*.

3 We found some interesting windows in the 3La inversion were *An. fontenillei* was
4 closer related to *An. melas* than to *An. bwambae* (Figure 3, line R). In this case, these
5 windows may be related with regions nearby the inversion breakpoints maintained
6 through positive selection. We made a functional enrichment analysis of these
7 windows using DAVID. There are 15 windows containing 25 genes that showed this
8 topology. There was only one enriched cluster with genes related to the stage-specific
9 breakdown of the larval midgut during metamorphosis, that allow replacement of
10 larval structures by tissues and structures that form the adult (Table 1).

11 Finally, the 3Rb inversion and the 2RI polymorphic inversions revealed by the
12 karyotyping of *An. fontenillei* individuals do not leave a clear pattern in the genomic
13 analysis performed here (see Fig. 3). Both inversions are only shared with *An.*
14 *bwambae*, the closest species to *An. fontenillei*, and hence it is not expected to
15 observed big differences in those regions.

16

17 **Discussion**

18 In 1975, the English entomologist G. B. White wrote: “As time passes, it becomes
19 increasingly less likely that other sibling species of this complex (*An. gambiae*) will
20 be found” [47]. Indeed, during the last 40 years, only one new species, *An.*
21 *quadriannulatus* B (recently called *An. amharicus*), has been discovered [15, 48], and
22 *An. coluzzii* has been separated from its sister species, *An. gambiae* [15]. In 2014, we
23 discovered a potential new species belonging to the *An. gambiae* species complex.
24 The species, *An. fontenillei*, was found in a mosaic savanna-forest area of Gabon,
25 Central Africa. This region is characterized by hosting the last vestige of savanna in
26 the Congo rainforest basin [49]. However, this habitat is not unique, and other parts of
27 Gabon and Central Africa could entertain the presence of this species. The new
28 mosquito seems to breed in rain dependent, sunlit, and open pools, evidencing similar
29 larval ecology to other fresh-water species within the complex [16]. According to its
30 ecology, we presumed a zoophilic host preference (Fig 1B). This behavior has already
31 been found in other members of the complex, such as *An. quadriannulatus* [50], and it
32 seems an ancestral character. However, *An. fontenillei* can also feeds on humans,
33 therefore, showing a generalist feeding habit with potential consequences on parasite
34 transfer between human and animals [51]. Indeed, the ancient and recent history of La

1 Lope provided multiple opportunities for *An. fontenillei* to adapt to humans [52]. In
2 the Neolithic age, La Lope was commonly colonized for hunting by nomad tribes, and
3 in the last century there was a forestry industry in the park. However, whether this
4 trait is ancestral or recently acquired (i.e. by introgression, see below) will need
5 further investigations (Table 1).

6

7 In order to disentangle its phylogenetic position within the complex, we sequenced
8 and *de novo* assembled *An. fontenillei* genome. The new genome allowed us to
9 determine that *An. fontenillei* and *An. bwambae* are sister species. Pairwise
10 comparisons revealed a higher divergence time between *An. fontenillei* and *An.*
11 *bwambae* than between *An. gambiae* and *An. coluzzii* (Fig. 2, [37]), corroborating the
12 geographical and ecological assumptions of two different species (Fig. 1). The *An.*
13 *fontenillei* - *An. bwambae* (FB) clade was placed together with *An. quadriannulatus*,
14 *An. arabiensis* and *An. melas*, being *An. quadriannulatus* or *An. arabiensis* the closest
15 species of the clade (Fig. 2). This is, to date, the most exhaustive phylogenetic tree of
16 the complex, including eight of the nine species described (no genome sequence is
17 available for *An. amharicus*).

18 Consistent with Fontaine *et al.*, [10], we found pervasive evidence of introgression in
19 *An. fontenillei*, confirming the permeable species boundaries in the *An. gambiae*
20 complex [37, 53]. Introgression within species complexes is common in nature,
21 challenging the possibility to trace the evolutionary history of species [3].
22 Interestingly, we observed patterns of recent introgression between *An. fontenillei* and
23 the clade *An. gambiae*-*An. coluzzii* (GC), particularly in the centromeric regions (20%
24 of the phylogenetic trees). These last two species were found in the village close to
25 the sylvatic sites where *An. fontenillei* was sampled (La Lope, Fig. 1A), indicating a
26 potential contact among them. The genomic windows introgressed were mostly
27 enriched for genes associated with detoxification, desiccation tolerance, and olfactory
28 perception (Table 1), which have been related with enhanced vectorial capacity [9].
29 These traits allow species to inhabit a broader range of habitats, and blood-feeding on
30 different hosts. The evidence of recent gene exchange between *An. gambiae*-*An.*
31 *coluzzii* with other species of the complex, may alter the evolution of these two major
32 malaria vectors, with potential consequences for malaria transmission (*i.e.* adaptation
33 to sylvatic habitats and/or preference for feeding on animals). However, we cannot
34 discard that this patterns of recent introgression in centromeric regions could be

1 affected by the low recombination rate in those areas, that could help to protect
2 introgressed haplotypes for a longer time compared with other genomic regions [54].
3 Finally, we resolved the evolution of the inversion 3La in the *An. gambiae* complex
4 While, this inversion was thought to be present in the ancestor of *An. melas* and *An.*
5 *bwambae*, we estimated that the origin of this inversion predated the radiation of the
6 *gambiae* complex [34]. Moreover, we evidenced that the inversion was independently
7 lost by *An. arabiensis* and *An. quadriannulatus* (Fig 3, Fig. 4). Although the 3La
8 inversion has not been associated yet to any trait, we observed functional enrichment
9 in larval midgut histolysis genes in recently introgressed regions between *An. melas*
10 and *An. fontillei* (Table 1). Again, these two species are present in Gabon, and
11 could envision potential gene exchange between them. Chromosomal rearrangements
12 have modulated the evolution of multiple species by affecting local adaptation or
13 speciation [5, 55-60]. In our genomic analysis (Fig. 3), we also observed the genomic
14 signature of the 2La inversion that affects the phylogenetic relationship between *An.*
15 *fontillei*, *An. arabiensis*, and *An. quadriannulatus*, highlighting the impact of fixed
16 inversions in chromosome evolution within the complex.

17 Besides the titanic collection effort led in Africa during the last century, the rainforest
18 of Central Africa has carefully hidden a new piece in the jigsaw puzzle of the *An.*
19 *gambiae* species complex. The discovery of a new species in the *An. gambiae*
20 complex has provided new insights into genome evolution (i.e. inversion 3La) and
21 local adaptation (i.e. salinity tolerance) in this group of closely related species.
22 Moreover, the new species has been an active actor in the evolution of *An. gambiae*-
23 *An. coluzzii*, exchanging genes involved in vectorial capacity. These introgressions
24 open new questions about how local populations of the major vectors, *An. gambiae*
25 and *An. coluzzii*, have been affected. Indeed, adaptation to rainforest habitats, host
26 preference or resting behavior could have been modified at La Lope. New studies may
27 provide important insights about how vectorial traits have evolved from wild to
28 domestic populations within the complex, with a direct impact in future malaria
29 control strategies.

30

31

32 **Material and Methods**

33 *Research and ethics statements*

1 We sampled *Anopheles* specimens under the national park entry authorization
2 AE16008/PR/ANPN/SE/CS/AEPN and the national research authorization
3 AR0013/16/MESRS/CENAREST/CG/CST/CSAR. Moreover, we obtained the
4 approval by National Research Ethics Committee of Gabon (0031/2014/SG/CNE) to
5 perform the human-landing catch (HLC) collections.

6

7 *Mosquito sampling and species identification*

8 Mosquitoes were sampled in the National Park of La Lopé in Gabon, Central Africa,
9 in an exploratory survey in November 2014. Since, several collections were carried
10 out in June 2015, February 2016 and November 2016. (Fig. 1, Table S1). Adults were
11 collected using BG traps with BG-lure and a source of CO₂ and HLC, while larvae
12 were sampled by the dipping method [61]. Collected *Anopheles* mosquitoes were
13 taxonomically identified according to standard morphological features [32, 33]. Then,
14 they were individually stored in 1.5 mL tubes at -20°C and sent to CIRMF for
15 molecular analysis. Total genomic DNA from specimens morphologically identified
16 as belonging to the *An. gambiae* complex was extracted using the DNeasy Blood and
17 Tissue Kit (Qiagen) according to the manufacturer's instructions. Genomic DNA was
18 eluted in 100 µL of TE buffer. A first molecular diagnostic (PCR-based) was
19 performed to molecularly identify species within the complex [26]. Surprisingly, an
20 unspecific fragment of 700 bp was amplified. This band does not correspond to any of
21 the species reported by the PCR-RFLP diagnostic test [26].

22

23 *Mosquito karyotyping*

24 Half-gravid females were sampled in November 2016 (Table S1) in forest sites where
25 we previously found the unspecified taxon. Females were collected by HLC and feed
26 to complete their blood-meal. Mosquitoes were allowed to develop follicles for 25 h
27 at field temperature. Then, ovaries were dissected and stored in Carnoy's fixative
28 solution (three parts 100% ethanol: one part glacial acetic acid, by volume). At the
29 CIRMF, we squashed the ovaries in a drop of 50% of propionic acid to obtain the
30 polytene chromosomes [62]. The banding patterns of polytene chromosomes were
31 examined using a Leica DM2000 and a camera system Leica DFC 450 (Leica
32 Microsystems GmbH, Wetzlar, Germany). Chromosomal arms and inversions were
33 recorded and scored according to *An. gambiae* chromosome map [63].

34

1 *Preliminary sequencing analysis*

2 In order to obtain further information about the unrecognized PCR band, we
3 sequenced three genes previously employed for phylogenetic studies in the complex
4 following the authors' instructions: internal transcribed spacer subunit 2 (ITS2~490
5 bp [64]); NADH dehydrogenase subunit 5 (ND5~300 bp [65, 66]); and cytochrome c
6 oxidase subunit I (COI~495 bp [67]). Moreover, we designed a new set of primers for
7 amplifying a fragment of the intergenic spacer gene (IGS~267 bp; IGSKPF 5'-
8 CTCTTGTGAGAGCAAGAGTGT-3' and IGSKPR 5'-
9 ATCAAGACAATCAAGTCGAGA-3') used also for species identification in the
10 complex. For the IGS gene, PCR reactions were carried out in 25µl reaction volume
11 than included 1X Qiagen PCR buffer (Qiagen, France), 1.5mM MgCl₂, 200µM each
12 dNTP (Eurogentec, Belgium), 10 pmol of each primer, 2.5 U Taq DNA polymerase
13 (Qiagen, France) and 1- 20 ng of template DNA. Amplifications were performed
14 using a Mastercycler Gradient thermocycler (Eppendorf) under the following
15 conditions: an initial step at 94°C for 5 minutes is followed by 35 cycles of 30
16 seconds at 94°C, 30 seconds at 54°C, 1 minute at 72°C and a final elongation step of
17 10 minutes at 72°C. Five microliters of the PCR product were analyzed by
18 electrophoresis on 1.5% agarose gels containing 0.5 µl/ml ethidium bromide and
19 photographed under UV light.

20

21 The sequences obtained for the four regions were analyzed using *Geneious* R10 [68].
22 We aligned the consensus sequences for each gene with randomly chosen sequences
23 of each species within the complex. We selected unique haplotypes to be included in
24 the phylogenetic analysis. The best substitution model for each gene was identified
25 using SMS [69]. The phylogenetic trees were then performed by maximum likelihood
26 (ML) method using PhyML [70], with nearest neighbour interchange (NNI) for tree
27 searching and approximate likelihood-ratio test (aLRT SH-like, [71]) for branch
28 support. Visualisation of trees was done using iTOL v.3.4.3 [72].

29

30 *Genome Sequencing and Assembly*

31 We sequenced four individuals of the unknown species using the Illumina platform at
32 the CNAG (Barcelona). To make a *de novo* genome assembly of this species one of
33 the individuals was deeply sequenced to ~112X. The other three individuals were

1 sequenced at an average coverage of ~29X. All reads were paired-end 126 bp long
2 (Table S2).

3 The genome assembly of the more deeply sequenced *An. fontenillei* individual was
4 performed at the Bioinformatics Unit, CRG (Barcelona) (Table S2, S3A). Reads were
5 trimmed and filtered using Skewer version 0.2.2 [73] to remove the adapter sequence
6 and trimming the low quality part. A FastQC analysis was performed to check the
7 quality of the trimmed reads. We looked at the presence of contaminants in a Kraken
8 database, which includes complete bacterial, archaeal, and viral genomes in RefSeq
9 [74]. We only found an enterobacteria phage phiX as contaminant (Table S3B). Then,
10 we assembled the trimmed reads by using Platanus software version 1.2.4 [75]
11 producing contigs and scaffolds using the paired-end information. To join the contigs
12 within the same scaffolds, stretches of N need to be added. To fill those gaps we used
13 Platanus *gap_close* function using the original reads (Table S3A). At this point,
14 improve the scaffolding of the assembled genome by using the proteins described for
15 AgamP4 reference in VectorBase (www.vectorbase.org). We used Blat [76] to map
16 the proteins to the assembled scaffolds and reorder and join scaffolds accordingly
17 with PEP_scaffolder [77]. We made another round of gap filling, and due to format
18 incompatibilities, this time we used *GapCloser* tool from the SOAPdenovo package
19 [78] (Table S3A). To evaluate the quality of our assembly, we scanned for the
20 presence of conserved genes among the diptera order by using BUSCO software [79].
21 We used 2,799 gene models conserved among diptera, and classify those genes as: i)
22 completely found in a single sequence, ii) fragmented in different sequences, or iii)
23 completely missing. Most of the BUSCO genes, 96%, were completely found in a
24 single sequence (Table S3C). We finally performed a polishing step by removing the
25 scaffolds mapping to previously found contaminants (Table S3A).

26

27 *Phylogenetic analysis*

28 To make the genome-wide phylogenetic tree by window analysis, we took advantage
29 of the available multiple alignment file (MAF) for six species of the *An. gambiae*
30 complex including two outgroup species: *An. christyi* and *An. epiroticus* [10]. Briefly,
31 we used the alignment formed by whole genome sequences from population samples
32 of multiple individuals of *An. gambiae*, *An. coluzzii*, *An. merus*, *An. melas*, *An.*
33 *quadriannulatus* and *An. arabiensis*. The *An. gambiae* PEST v3 (AgamP3) reference
34 genome obtained from VectorBase (www.vectorbase.org) was also included. Fontaine

1 et al., [10] made a whole genome alignment using ROAST [80] that represents
2 approximately 40% of the euchromatic genome. We download the resulting MAF
3 based on field-collected samples from
4 <http://datadryad.org/resource/doi:10.5061/dryad.f4114> [10]. We then added to this
5 MAF our *An. fontenillei* assembly, and the highest coverage *An. bwambae* genome
6 sequences available (see below).

7

8 ***Anopheles fontenillei***. We first generated a database with the scaffolds of the *An.*
9 *fontenillei* assembly. Then we run blastn for each region in the MAF using AgamP3
10 as query sequence against the *An. fontenillei* scaffold database. We then repeat this
11 blastn analysis using other species of the MAF regions as query; *An. arabiensis*, *An.*
12 *quadriannulatus*, *An. melas* and *An. merus* (Text S2.1.1 - Text S2.1.4). We did not
13 repeat the analysis using *An. coluzzii* and *An. gambiae* as queries, due to its similarity
14 to the reference genome AgamP3, or the two outgroup species, that are too divergent.
15 We then selected the MAF regions for which we get one unique hit in any of the
16 species, which represents the 63.2% of all the MAF regions for the eight species
17 (Table S10). For the additional MAF region that gave more than one hit we exclude
18 the multiple hits with e -value $> 10^{-4}$ or with $\leq 40\%$ of the query covered for each
19 region in each species (Text S2.1.5) and, we recovered the sequences that became a
20 unique hit after this filtering (Table S10). In total, we were able to include *An.*
21 *fontenillei* in 75.2% of the previous MAF regions, which represent the ~30% of the
22 euchromatic genome. For each of these MAF regions we cut the scaffolds according
23 to the blast result information (Text S2.1.6). Then, we added these sequences to the
24 corresponding MAF region using MAFFT as an aligner (v7.221, [81]). We use the
25 function `--add` to modify as less as possible the initial MAF [82]. Finally, we joined
26 each region of the MAF and generated the new MAF including the *An. fontenillei*
27 genome (Text S2.2).

28

29 ***Anopheles bwambae***. We downloaded the tree individual sequences of *Anopheles*
30 *bwambae* available at NCBI with fastq-dump; i) *An. bwambae* 1, SRR1255391,
31 SRR1255392, and, SRR1255303, ii) *An. bwambae* 3, SRR1255390, and, iii) *An.*
32 *bwambae* 4, SRR1255325. We then joined the SRR for individual 1 (Text S2.3). For
33 each individual, we evaluated read quality with fastQC and trimmed the reads using

1 cutadapt (v. 1.8.3; [83])(Text S2.4.1 - Text S2.4.3). After the trimming, the quality per
2 base was always higher than 24. We then map the trimmed read to AgamP3 reference
3 genome using bwa-mem [84]. We performed several post-mapping steps including
4 marking duplicates and realigning around indels using Picard (v.1.109;
5 <http://picard.sourceforge.net>), samtools (v. 1.3; [85]) and GATK (v3.4-46; [86])(Text
6 S2.4.4 - Text S2.4.8). Among the three available *An. bwambiae* individuals we
7 selected the one with the highest coverage, *An. bwambiae* 1 with 33.2X, to add it to the
8 MAF (the other two: *An. bwambiae* 3, 11.7X and *An. bwambiae* 4, 11.2X). We made a
9 consensus sequence of the *An. bwambiae* 1 reads mapping to the AgamP3 sequence of
10 every MAF regions with the 9 species (*An. gambiae*, *An. coluzzii*, *An. merus*, *An.*
11 *melas*, *An. quadriannulatus*, *An. arabiensis*, *An. fontenillei*, *An. christyi* and *An.*
12 *epiroticus*), using SAMtools mpileup (Text S2.5). If for a MAF region there were not
13 *An. bwambiae* reads, we added gaps so that we kept the same number of MAF regions
14 as before. This had a marginal effect as it only occurred in 0.06% of all the MAF
15 regions. Finally, we used MAFFT aligner, with the function `--add`, to add each
16 consensus sequence of *An. bwambiae* 1 to the MAF regions and then joined all these
17 regions in a new MAF (Text S2.2).

18
19 **MAF with four *An. fontenillei* individuals and three *An. bwambiae* individuals.** To
20 check the phylogenetic relationship between *An. fontenillei* and *An. bwambiae*, we also
21 created an additional MAF that included the 8 species previously available, the four
22 *An. fontenillei* individuals, and the three *An. bwambiae* individuals. We mapped each
23 one of the seven individuals to AgamP3 reference genome as described previously for
24 *An. bwambiae* (Text S2.4). Then, we generated a consensus sequence for each of the
25 new individuals for each of the MAF regions using SAMtools mpileup (Text S2.5).
26 Finally, we add sequentially each of the new sequences to the available multiple
27 alignment regions using MAFFT `--add` function as aligner (v7.221, [81, 82], Text
28 S2.2). Finally, we joined all these information in a new MAF file.

29
30 **Window-based phylogenies.** We generated 50 kb genome-wide non-overlapping
31 windows from the MAF (Text S2.6.1). For each window, we generated a maximum
32 likelihood (ML) phylogenetic tree using RAxML (v8.2.4, [87]) with GTRGAMMA
33 model and bootstrapping for 1,000 replicates (Text S2.6.2) [10]. We used the closer
34 related species, *An. christyi*, as an outgroup because Fontaine et al. [10] already

1 showed that the choice of the outgroup did not substantially alter the results. We
2 excluded the windows with less than 10% of informative base pairs (e.g. < 5,000 bp)
3 (following [10]). The different topologies obtained were sorted, counted, and analyzed
4 using ad-hoc perl scripts (Text S2.6.3).

5

6 *Pairwise distance and bootstrapping*

7 We used the R package ‘APE’ (v4.1, [88]) to estimate pairwise genetic distances
8 based in the ML phylogenetic trees. We then performed the bootstrap analysis using
9 the ‘boot’ package in R [89](Text S2.7).

10

11 *Centromeric regions alignment quality*

12 For each chromosomal arm, we choose randomly 30 windows from centromeric
13 regions and 30 regions from other genomic regions. Centromeric regions were defined
14 based on the observed ancestry pattern in Fig. 3: 2L: 0 to 10Mb, 2R: 50 to 61.3Mb,
15 3L: 0 to 10Mb, 3R: 40 to 53.1Mb and X: 15 to 20.2Mb. For these 60 windows by
16 chromosome arm, we gathered: i) the alignment length, ii) alignment length without
17 completely undetermined characters and gaps, iii) proportion of gaps, and iv) the
18 alignment patterns from the RAxML information file.

19

20 *Data analysis*

21 We used R v3.2.5 (R Development Core Team, <http://cran.r-project.org/>) to perform
22 all the statistical analysis. We used Inkscape software for figure edition
23 (<https://inkscape.org>).

24

25 **Acknowledgements**

26 We are grateful to P. Nosil and G. Lanzaro for comments on the manuscript. We
27 thank the “Agence Nationale de la Preservation de la Nature” (ANPN), the “Station
28 d’Etudes des Gorilles et Chimpanzes” (SEGC) and the “Centre National de la
29 Recherche Scientifique et Technologique of Gabon” (CENAREST) that authorized
30 this study and facilitated the access to the national parks of La Lopé. We specially
31 thank Vincenzo Petrarca for its help interpreting chromosome polymorphism.
32 Funding was provided by the “Institut de Recherche pour le Developpement”, the
33 “Agence Universitaire de la Francophonie” (grant: OKANDA), the “Centre National
34 de la Recherche Scientifique” (CNRS) and the “Consejo Superior de Investigaciones

1 Cientificas” (CSIC) (grant PICS ANCESTRAL to DA and JG), the “ANR” (grant
2 ANR--18-CE35-0002-01-WILDING to DA), and the “Ministerio de Ciencia,
3 Innovación y Universidades/AEI” (grant BFU2017-82937-P to JG).

4

5 **Author contribution**

6 **Conceptualization** DA, JG, CP, MGB

7 **Data Curation** MGB, DA, JG

8 **Formal Analysis** MGB, JG, DA,

9 **Funding Acquisition** JG, DA

10 **Investigation** PK, OAE, NR, PK, MFN, TWB, MP, DA

11 **Project Administration** DA, JG

12 **Resources** DA

13 **Supervision** JG, DA

14 **Validation** CC, FS

15 **Visualization** MGB, JG, NR, CP, DA

16 **Writing – Original Draft Preparation** MGB, JG, DA

17 **Writing – Review & Editing** MGB, JG, DA, CP

18

19 **Competing interests**

20 The authors declare no competing interests.

21

22 **Data accessibility:** DNA sequences have been deposited to GenBank under
23 accessions xxxxxx.

24

25 **References**

26 1. Coyne J, Orr H. Speciation: Sinauer Associates; 2004.

27 2. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow.
28 Trends in Genetics. 2012;28(7):342--50. doi: 10.1016/j.tig.2012.03.009. PubMed
29 PMID: Feder2012.

30 3. Mallet J, Besansky N, Hahn MW. How reticulated are species? BioEssays.
31 2016;38(2):140--9. doi: 10.1002/bies.201500149. PubMed PMID: Mallet2016.

32 4. Hedrick PW. Adaptive introgression in animals: examples and comparison
33 to new mutation and standing variation as sources of adaptive variation.
34 Molecular Ecology. 2013;22(18):4606-18. doi: 10.1111/mec.12415. PubMed
35 PMID: WOS:000324022600002.

- 1 5. Kirkpatrick M, Barrett B. Chromosome inversions, adaptive cassettes and
2 the evolution of species' ranges. *Molecular Ecology*. 2015;24(9):2046--55. doi:
3 10.1111/mec.13074. PubMed PMID: Kirkpatrick2015.
- 4 6. Huerta-Snchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al.
5 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA.
6 *Nature*. 2014;512(7513):194--7. doi: 10.1038/nature13408. PubMed PMID:
7 Huerta-Sanchez2014.
- 8 7. Weill M, Chandre F, Brengues C, Manguin S, Akogbeto M, Pasteur N, et al.
9 The *kdr* mutation occurs in the Mopti form of *Anopheles gambiae* s.s. through
10 introgression. *Insect molecular biology*. 2000;9(5):451--5. PubMed PMID:
11 Weill2000.
- 12 8. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, et al.
13 The dominant *Anopheles* vectors of human malaria in Africa, Europe and the
14 Middle East: occurrence data, distribution maps and bionomic pr \ ' e cis.
15 *Parasites \& Vectors*. 2010;3(1):117. doi: 10.1186/1756-3305-3-117. PubMed
16 PMID: Sinka2010.
- 17 9. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen
18 JE, et al. Highly evolvable malaria vectors: The genomes of 16 *Anopheles*
19 mosquitoes. *Science*. 2015;347(6217). doi: 10.1126/science.1258522. PubMed
20 PMID: Neafsey2015.
- 21 10. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov
22 IV, et al. Extensive introgression in a malaria vector species complex revealed by
23 phylogenomics. *Science (New York, NY)*. 2015;347(6217):1258524. doi:
24 10.1126/science.1258524. PubMed PMID: Fontaine2015.
- 25 11. Cohuet A, Harris C, Robert V, Fontenille D. Evolutionary forces on
26 *Anopheles*: what makes a malaria vector? *Trends in Parasitology*.
27 2010;26(3):130--6. PubMed PMID: Cohuet2010.
- 28 12. Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T,
29 Coetzee M, et al. A global map of dominant malaria vectors. *Parasites \& Vectors*.
30 2012;5(1):69. doi: 10.1186/1756-3305-5-69. PubMed PMID: Sinka2012.
- 31 13. Davidson G. *Anopheles gambiae* complex. *Nature*. 1962;196:907. PubMed
32 PMID: Davidson1962.
- 33 14. Davidson G, Hunt RH. The crossing and chromosome characteristics of a
34 new 6th species in the *Anopheles gambiae* complex. *Parassitologia*. 1973;15(1 -
35 2):121--8. PubMed PMID: Davidson1973.
- 36 15. Coetzee M, Hunt RH, Wilkerson Ra. *Anopheles coluzzii* and *Anopheles*
37 *amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa*.
38 2013;3619(3):246--74. PubMed PMID: Coetzee2013.
- 39 16. White BJ, Collins FH, Besansky NJ. Evolution of *Anopheles gambiae* in
40 Relation to Humans and Malaria. *Annual Review of Ecology, Evolution, and*

- 1 Systematics. 2011;42(1):111--32. doi: 10.1146/annurev-ecolsys-102710-
2 145028. PubMed PMID: White2011a.
- 3 17. Pombi M, Kengne P, Gimonneau G, Tene-Fossog B, Ayala D, Kamdem C, et
4 al. Dissecting functional components of reproductive isolation among closely
5 related sympatric species of the *Anopheles gambiae* complex. *Evolutionary*
6 *Applications*. 2017;10(10):1102-20. doi: 10.1111/eva.12517. PubMed PMID:
7 WOS:000414952000013.
- 8 18. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov
9 IV, et al. Data from: Extensive introgression in a malaria vector species complex
10 revealed by phylogenomics. *Science*. 2014. doi: doi:10.5061/dryad.f4114.
11 PubMed PMID: Fontaine2014.
- 12 19. Fouet C, Gray E, Besansky NJ, Costantini C. Adaptation to Aridity in the
13 Malaria Mosquito *Anopheles gambiae*: Chromosomal Inversion Polymorphism
14 and Body Size Influence Resistance to Desiccation. *PLoS ONE*. 2012;7(4):e34841.
15 doi: 10.1371/journal.pone.0034841. PubMed PMID: Fouet2012.
- 16 20. Miles A, Harding NJ, Bott. Genetic diversity of the African malaria vector
17 *Anopheles gambiae*. *Nature*. 2017;552(7683):96. doi: 10.1038/nature24995.
18 PubMed PMID: Miles2017.
- 19 21. Norris LC, Main BJ, Lee Y, Collier TC, Fofana A, Cornel AJ, et al. Adaptive
20 introgression in an African malaria mosquito coincident with the increased usage
21 of insecticide-treated bed nets. *Proceedings of the National Academy of Sciences*.
22 2015;112(3):815-20.
- 23 22. Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L. The genetic
24 architecture of reproductive isolation during speciation-with-gene-flow in lake
25 whitefish species pairs assessed by rad sequencing. *Evolution*. 2013;67(9):2483-
26 97. doi: 10.1111/evo.12075. PubMed PMID: WOS:000323828500003.
- 27 23. Rinker DC, Pitts RJ, Zhou X, Suh E, Rokas A, Zwiebel LJ. Blood meal-
28 induced changes to antennal transcriptome profiles reveal shifts in odor
29 sensitivities in *Anopheles gambiae*. *Proceedings of the National Academy of*
30 *Sciences*. 2013;110(20):8260--5. doi: 10.1073/pnas.1302562110. PubMed PMID:
31 Rinker2013a.
- 32 24. Smith HA, White BJ, Kundert P, Cheng C, Romero-Severson J, Andolfatto P,
33 et al. Genome-wide QTL mapping of saltwater tolerance in sibling species of
34 *Anopheles* (malaria vector) mosquitoes. *Heredity*. 2015;115(5):471--9. doi:
35 10.1038/hdy.2015.39. PubMed PMID: Smith2015.
- 36 25. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al.
37 The effect of malaria control on *Plasmodium falciparum* in Africa between 2000
38 and 2015. *Nature*. 2015;526(7572):207--11. doi: 10.1038/nature15535. PubMed
39 PMID: Bhatt2015.
- 40 26. Fanello C, Santolamazza F, della Torre A. Simultaneous identification of
41 species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP.

- 1 Medical and veterinary entomology. 2002;16(4):461--4. PubMed PMID:
2 Fanello2002.
- 3 27. Elissa N, Karch S, Bureau P, Ollomo B, Lawoko M, Yangari P, et al. Malaria
4 transmission in a region of savanna-forest mosaic, Haut-Ogoou \ ' e , Gabon.
5 Journal of the American Mosquito Control Association. 1999;15(1):15--23.
6 PubMed PMID: Elissa1999.
- 7 28. Mourou J-R, Coffinet T, Jarjaval F, Pradines B, Amalvict R, Rogier C, et al.
8 Malaria transmission and insecticide resistance of *Anopheles gambiae* in
9 Libreville and Port-Gentil, Gabon. Malaria journal. 2010;9:321. doi:
10 10.1186/1475-2875-9-321. PubMed PMID: Mourou2010.
- 11 29. Hervy JFa. Les Anophèles de la region afrotropicale: Paris France; 1998.
- 12 30. White BJ, Collins FH, Besansky NJ. Evolution of *Anopheles gambiae* in
13 Relation to Humans and Malaria. Annual Review of Ecology, Evolution, and
14 Systematics. 2011;42(1):111-32. doi: doi:10.1146/annurev-ecolsys-102710-
15 145028.
- 16 31. Pombi M, Guelbeogo WM, Calzetta M, Sagnon NF, Petrarca V, La Gioia V, et
17 al. Evaluation of a protocol for remote identification of mosquito vector species
18 reveals BG-Sentinel trap as an efficient tool for *Anopheles gambiae* outdoor
19 collection in Burkina Faso. Malaria journal. 2015;14(1):161.
- 20 32. Gillies MT, de Meillon B. The anophelinae of Africa, south of the Sahara.
21 The South African Institute for Medical Research. 1968;54. PubMed PMID:
22 Gillies1968.
- 23 33. Gillies MT, Coetzee MC. A supplement to the Anophelinae of Africa south
24 of the Sahara (Afrotropical Region). Publications of the South African Institute
25 for Medical Research. 1987;55:143. PubMed PMID: Gillies1987.
- 26 34. White GB. *Anopheles bwambiae*, a malaria vector in the Semliki valley,
27 Uganda, and its relationships wiht other sibling species of the *An. gambiae*
28 complex (Diptera, Culicidae). Systematic Entomology. 1985;10:501--22. PubMed
29 PMID: White1985.
- 30 35. Coluzzi M, Sabatini A, della Torre Aa. A Polytene Chromosome Analysis of
31 the *Anopheles gambiae* Species Complex. Science. 2002;298(5597):1415--8. doi:
32 10.1126/science.1077769. PubMed PMID: Coluzzi2002.
- 33 36. White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, Mouline K, et al.
34 Molecular karyotyping of the 2La inversion in *Anopheles gambiae*. The American
35 journal of tropical medicine and hygiene. 2007;76(2):334--9. PubMed PMID:
36 White2007.
- 37 37. Neafsey DE, Lawniczak MKN, Park DJ, Redmond SN, Coulibaly MB, Traor.
38 SNP genotyping defines complex gene-flow boundaries among African malaria
39 vector mosquitoes. Science (New York, NY). 2010;330(6003):514--7. doi:
40 10.1126/science.1193036. PubMed PMID: Neafsey2010.

- 1 38. Tamura K, Subramanian S, Kumar S. Temporal Patterns of Fruit Fly
2 (*Drosophila*) Evolution Revealed by Mutation Clocks. *Molecular Biology and*
3 *Evolution*. 2004;21(1):36--44. doi: 10.1093/molbev/msg236. PubMed PMID:
4 Tamura2004.
- 5 39. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools:
6 paths toward the comprehensive functional analysis of large gene lists. *Nucleic*
7 *Acids Research*. 2009;37(1):1--13. doi: 10.1093/nar/gkn923. PubMed PMID:
8 Huang2009.
- 9 40. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis
10 of large gene lists using DAVID bioinformatics resources. *Nature Protocols*.
11 2009;4(1):44--57. doi: 10.1038/nprot.2008.211. PubMed PMID: Huang2009a.
- 12 41. Liu N. Insecticide Resistance in Mosquitoes: Impact, Mechanisms, and
13 Research Directions. *Annual Review of Entomology*. 2015;60(1):537--59. doi:
14 10.1146/annurev-ento-010814-020828. PubMed PMID: Liu2015.
- 15 42. Goltsev Y, Rezende GL, Vranizan K, Lanzaro G, Valle D, Levine M.
16 Developmental and evolutionary basis for drought tolerance of the *Anopheles*
17 *gambiae* embryo. *Developmental Biology*. 2009;330(2):462--70. doi:
18 10.1016/j.ydbio.2009.02.038. PubMed PMID: Goltsev2009.
- 19 43. Tavaría M, Gabriele T, Kola I, Anderson RL. A hitchhiker's guide to the
20 human Hsp70 family. *Cell stress & chaperones*. 1996;1(1):23--8. PubMed PMID:
21 Tavaría1996.
- 22 44. Morano KA. New tricks for an old dog: The evolving world of Hsp70.
23 *Annals of the New York Academy of Sciences*. 2007;1113(1):1--14. doi:
24 10.1196/annals.1391.018. PubMed PMID: Morano2007.
- 25 45. Cashman JR. Some distinctions between flavin-containing and cytochrome
26 P450 monooxygenases. *Biochemical and Biophysical Research Communications*.
27 2005;338(1):599--604. doi: 10.1016/j.bbrc.2005.08.009. PubMed PMID:
28 Cashman2005.
- 29 46. White BJ, Lawniczak MKN, Cheng C, Coulibaly MB, Wilson MD, Sagnon NF,
30 et al. Adaptive divergence between incipient species of *Anopheles gambiae*
31 increases resistance to *Plasmodium*. *Proceedings of the National Academy of*
32 *Sciences*. 2011;108(1):244. doi: 10.1073/pnas.1013648108/
33 /DCSupplemental.<http://www.pnas.org/cgi/doi/10.1073/pnas.1013648108>.
34 PubMed PMID: White2011.
- 35 47. White GB. Notes on a Catalogue of Culicidae of the Ethiopian Region.
36 *Mosquito Systematics*. 1975;7(4).
- 37 48. Hunt RH, Coetzee M, Fittene M. The *Anopheles gambiae* complex: a new
38 species from Ethiopia. *Transactions of the Royal Society of Tropical Medicine and*
39 *Hygiene*. 1998;92(2):231--5. PubMed PMID: Hunt1998.

- 1 49. Ngomanda A, Chepstow-Lusty A, Makaya M, Schevin P, Maley J, Fontugne
2 M, et al. Vegetation changes during the past 1300 years in western equatorial
3 Africa: a high-resolution pollen record from Lake Kamalee, Lope Reserve, Central
4 Gabon. *The Holocene*. 2005;15(7):1021--31. doi: 10.1191/0959683605hl875ra.
5 PubMed PMID: Ngomanda2005.
- 6 50. Dekker T, Takken W. Differential responses of mosquito sibling species
7 *Anopheles arabiensis* and *An. quadriannulatus* to carbon dioxide, a man or a calf.
8 *Medical and Veterinary Entomology*. 1998;12(2):136--40. doi: 10.1046/j.1365-
9 2915.1998.00073.x. PubMed PMID: Dekker1998.
- 10 51. Makanga B, Yangari P, Rahola N, Rougeron V, Elguero E, Boundenga L, et
11 al. Ape malaria transmission and potential for ape-to-human transfers in Africa.
12 *Proceedings of the National Academy of Sciences*. 2016;113(19):5329--34. doi:
13 10.1073/pnas.1603008113. PubMed PMID: Makanga2016.
- 14 52. Oslisly R, White L, Bentaleb I, Favier C, Fontugne M, Gillet JF, et al. Climatic
15 and cultural changes in the west Congo Basin forests over the past 5000 years.
16 *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1625):20120304. Epub
17 2013/07/24. doi: 10.1098/rstb.2012.0304. PubMed PMID: 23878334; PubMed
18 Central PMCID: PMC3720025.
- 19 53. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, Mukabayire O,
20 et al. Semipermeable species boundaries between *Anopheles gambiae* and
21 *Anopheles arabiensis*: evidence from multilocus DNA
22 sequence variation. *Proceedings of the National Academy of Sciences of the*
23 *United States of America*. 2003;100(19):10818--23. doi:
24 10.1073/pnas.1434337100. PubMed PMID: Besansky2003.
- 25 54. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of
26 speciation are due to reduced diversity, not reduced gene flow. *Molecular*
27 *Ecology*. 2014;23(13):3133--57. doi: 10.1111/mec.12796. PubMed PMID:
28 Cruickshank2014.
- 29 55. Rieseberg LH. Chromosomal rearrangements and speciation. *Trends in*
30 *ecology & evolution*. 2001;16(7):351--8. PubMed PMID: Rieseberg2001.
- 31 56. Noor MAF, Grams KL, Bertucci LA, Reiland J. Chromosomal inversions and
32 the reproductive isolation of species. *Proceedings of the National Academy of*
33 *Sciences*. 2001;98(21):12084--8. doi: 10.1073/pnas.221274498. PubMed PMID:
34 Noor2001.
- 35 57. Lowry DB, Willis JH. A Widespread Chromosomal Inversion
36 Polymorphism Contributes to a Major Life-History Transition, Local Adaptation,
37 and Reproductive Isolation. *PLoS Biology*. 2010;8(9):e1000500. doi:
38 10.1371/journal.pbio.1000500. PubMed PMID: Lowry2010.
- 39 58. Ayala D, Guerrero RF, Kirkpatrick M. Reproductive isolation and local
40 adaptation quantified for a chromosome inversion in a malaria mosquito.
41

- 1 Evolution. 2013;67(4):946--58. doi: 10.1111/j.1558-5646.2012.01836.x.
2 PubMed PMID: Ayala2013.
- 3 59. Ayala D, Ullastres A, Gonzalez J. Adaptation through chromosomal
4 inversions in *Anopheles*. *Frontiers in Genetics*. 2014;5:129. doi:
5 10.3389/FGENE.2014.00129. PubMed PMID: Ayala2014.
- 6 60. Ayala D, Acevedo P, Pombi M, Dia I, Boccolini D, Costantini C, et al.
7 Chromosome inversions and ecological plasticity in the main African malaria
8 mosquitoes. *Evolution*. 2017;71(3):686--701. doi: 10.1111/evo.13176. PubMed
9 PMID: Ayala2017.
- 10 61. Service MW. Mosquito ecology field sampling methods. 2nd ed: Elsevier
11 Applied Science; 1993.
- 12 62. della Torre A. The Molecular Biology of Insect Disease Vectors : a Methods
13 Manual. 1997:329--36.
- 14 63. Pombi M, Caputo B, Simard Fa. Chromosomal plasticity and evolutionary
15 potential in the malaria vector *Anopheles gambiae sensu stricto*: insights from
16 three decades of rare paracentric inversions. *BMC Evolutionary Biology*.
17 2008;8(1):309. doi: 10.1186/1471-2148-8-309. PubMed PMID: Pombi2008.
- 18 64. Kengne P, Antonio-Nkondjio C, Awono-Ambene HP, Simard F, Awolola TS,
19 Fontenille D. Molecular differentiation of three closely related members of the
20 mosquito species complex, *Anopheles moucheti*, by mitochondrial and ribosomal
21 DNA polymorphism. *Medical and Veterinary Entomology*. 2007;21(2):177--82.
22 doi: 10.1111/j.1365-2915.2007.00681.x. PubMed PMID: Kengne2007.
- 23 65. Besansky NJ, Lehmann T, Fahey GT, Fontenille D, Braack LE, Hawley WA,
24 et al. Patterns of mitochondrial variation within and between African malaria
25 vectors, *Anopheles gambiae* and *An. arabiensis*, suggest extensive gene flow.
26 *Genetics*. 1997;147(4):1817--28. PubMed PMID: Besansky1997.
- 27 66. Thelwell NJ, Huisman RA, Harbach RE, Butlin RK. Evidence for
28 mitochondrial introgression between *Anopheles bwambae* and *Anopheles*
29 *gambiae*. *Insect molecular biology*. 2000;9(2):203--10. PubMed PMID:
30 Thelwell2000.
- 31 67. Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P. Evolution,
32 Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a
33 Compilation of Conserved Polymerase Chain Reaction Primers. *Annals of the*
34 *Entomological Society of America*. 1994;87(6):651--701. doi:
35 10.1093/aesa/87.6.651. PubMed PMID: Simon1994.
- 36 68. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al.
37 Geneious Basic: An integrated and extendable desktop software platform for the
38 organization and analysis of sequence data. *Bioinformatics*. 2012;28(12):1647--
39 9. doi: 10.1093/bioinformatics/bts199. PubMed PMID: Kearse2012.

- 1 69. Lefort V, Longueville J-E, Gascuel O. SMS: Smart Model Selection in PhyML.
2 Molecular Biology and Evolution. 2017;34(9):2422--4. doi:
3 10.1093/molbev/msx149. PubMed PMID: Lefort2017.
- 4 70. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O.
5 New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies:
6 Assessing the Performance of PhyML 3.0. Systematic Biology. 2010;59(3):307--
7 21. doi: 10.1093/sysbio/syq010. PubMed PMID: Guindon2010.
- 8 71. Anisimova M, Gascuel O. Approximate Likelihood-Ratio Test for Branches:
9 A Fast, Accurate, and Powerful Alternative. Systematic Biology. 2006;55(4):539--
10 52. doi: 10.1080/10635150600755453. PubMed PMID: Anisimova2006.
- 11 72. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for
12 phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127--8.
13 doi: 10.1093/bioinformatics/btl529. PubMed PMID: Letunic2007.
- 14 73. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter
15 trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics.
16 2014;15(1):182. doi: 10.1186/1471-2105-15-182. PubMed PMID: Jiang2014.
- 17 74. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence
18 classification using exact alignments. Genome Biology. 2014;15(3):R46. doi:
19 10.1186/gb-2014-15-3-r46. PubMed PMID: Wood2014.
- 20 75. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al.
21 Efficient de novo assembly of highly heterozygous genomes from whole-genome
22 shotgun short reads. Genome Research. 2014;24(8):1384--95. doi:
23 10.1101/gr.170720.113. PubMed PMID: Kajitani2014.
- 24 76. Kent WJ. BLAT---The BLAST-Like Alignment Tool. Genome Research.
25 2002;12(4):656--64. doi: 10.1101/gr.229202. PubMed PMID: Kent2002.
- 26 77. Zhu B-H, Song Y-N, Xue W, Xu G-C, Xiao J, Sun M-Y, et al. PEP _ scaffold:
27 using (homologous) proteins to scaffold genomes. Bioinformatics.
28 2016;32(20):3193--5. doi: 10.1093/bioinformatics/btw378. PubMed PMID:
29 Zhu2016.
- 30 78. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an
31 empirically improved memory-efficient short-read de novo assembler.
32 GigaScience. 2012;1(1):18. doi: 10.1186/2047-217X-1-18. PubMed PMID:
33 Luo2012.
- 34 79. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM.
35 BUSCO: assessing genome assembly and annotation completeness with single-
36 copy orthologs. Bioinformatics. 2015;31(19):3210-2. doi:
37 10.1093/bioinformatics/btv351. PubMed PMID: WOS:000362845400018.
- 38 80. Hou M. TOAST and ROAST. 2008.

- 1 81. Katoh K, Kuma K-i, Toh H, Miyata T. MAFFT version 5: improvement in
2 accuracy of multiple sequence alignment. *Nucleic Acids Research*.
3 2005;33(2):511--8. doi: 10.1093/nar/gki198. PubMed PMID: Katoh2005.
- 4 82. Katoh K, Frith MC. Adding unaligned sequences into an existing alignment
5 using MAFFT and LAST. *Bioinformatics*. 2012;28(23):3144--6. doi:
6 10.1093/bioinformatics/bts578. PubMed PMID: Katoh2012.
- 7 83. Martin M. Cutadapt removes adapter sequences from high-throughput
8 sequencing reads. *EMBnetjournal*. 2011;17(1):10. doi: 10.14806/ej.17.1.200.
9 PubMed PMID: Martin2011.
- 10 84. Li H. Aligning sequence reads, clone sequences and assembly contigs with
11 BWA-MEM. 2013. PubMed PMID: Li2013.
- 12 85. Li H, Durbin R. Fast and accurate short read alignment with Burrows-
13 Wheeler transform. *Bioinformatics*. 2009;25(14):1754--60. doi:
14 10.1093/bioinformatics/btp324. PubMed PMID: Li2009.
- 15 86. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et
16 al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-
17 generation DNA sequencing data. *Genome Research*. 2010;20(9):1297--303. doi:
18 10.1101/gr.107524.110. PubMed PMID: McKenna2010.
- 19 87. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-
20 analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312--3. doi:
21 10.1093/bioinformatics/btu033. PubMed PMID: Stamatakis2014.
- 22 88. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and
23 Evolution in R language. *Bioinformatics (Oxford, England)*. 2004;20(2):289--90.
24 PubMed PMID: Paradis2004.
- 25 89. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version
26 13-20. 2017. PubMed PMID: Canty2017.
27
28

1 **Table 1: Functional enrichment analysis result.** Significantly enriched clusters (>1.3 enrichment score) obtained with the analysis of
 2 functional terms with DAVID.
 3
 4

Tree topology	# of windows	# of genes	Term summary	Enrichment Score	GO terms	InterPro domains
(F,G)	64	198	Insect cuticle protein	5.6	GO:0042302	IPR000618
			Transmembrane transporter activity	1.6	GO:0042391, GO:0015701, GO:0015301, GO:0019531, GO:0015106, GO:0015116, GO:0008271, GO:0051453, GO:0005254, GO:1902476, GO:0005887	IPR001902, IPR011547, IPR002645
			Peptidase activity	1.6	GO:0004252	IPR001254, IPR018114, IPR009003
			Heat shock p70	1.4	-	IPR018181, IPR013126
(F,C)	25	62	None	<1.3	-	-
(F,(GC))	35	89	Flavin monooxygenase	3.6	GO:0004499, GO:0050661, GO:0055114, GO:0050660, GO:0004497	IPR000960, IPR020946, IPR023753
			Olfactory receptor	1.3	GO:0050911, GO:0004984, GO:0005549, GO:0005886	IPR004117
(F,L)	15	25	Larval midgut histolysis	4.9	GO:0035069, GO:0097200, GO:0097194, GO:0005737	IPR002138, IPR001309, IPR015917

5
6

1 **Figure legends**

2 **Figure 1. Overview of bionomic characteristics of *An. fontenillei*.** (A)

3 Geographical distribution of *An. fontenillei* within the National Park of La Lope.

4 Breeding site where the larva of the new species was found. (B) Mean (black dots) of

5 *An. fontenillei* collected by human landing catch (human, red) vs. BG traps (trap,

6 green) in the park. (C) Morphological features of *An. fontenillei*: Dorsal view of the

7 wing, maxillary palpus and hindleg with femur, tibia and tarsomeres. (D) Polytene

8 chromosomes from ovarian nurse cells of *An. fontenillei* with a contrast-phase

9 microscope (specimen n. 23). Chromosomal arms karyotypes are indicated following

10 the classical nomenclature [35]. Paracentric inversions are designed by lines (red and

11 blue) above the arms 3R(b) and 3L(a), respectively.

12

13 **Figure 2. Most common species tree.** 78 windows in the X chromosome show this

14 tree topology with a weak disagreement in the basal node. Black numbers represent

15 bootstrapping values and in red the millions years estimated based on the pairwise

16 distances of the ML phylogeny and assuming a substitution rate of 11×10^{-9} per site,

17 per generation and 10 generation per year.

18

19 **Figure 3. Recent (R) and ancestral (A) relationship of *An. fontenillei* with other**

20 **species in the *An. gambiae* complex according to the phylogenetic trees in 50kb**

21 **non-overlapping windows along each chromosome arm.** (R) *An. fontenillei* closer

22 species or clade on each tree. (A) When the closer species in the tree is *An. bwambae*,

23 then *An. fontenillei* - *An. bwambae* clade closer species or closest clade is shown for

24 each window.

25

26 **Figure 4. Species topology estimated from the X chromosome compared with the**

27 **topology of the 3La inversion.** *An. christyi* was used as outgroup species. Green

28 color: *An. arabiensis* - GC common ancestor possible introgression. Purple color:

29 species that share the inversion. Yellow color: *An. quadriannulatus* and *An. merus*

30 possible introgression event.

31

32







