LJMU Research Online

Wilding, CS

 Genetic diversity of the African malaria vector Anopheles gambiae

http://researchonline.ljmu.ac.uk/id/eprint/7775/

Article

For more information please contact researchonline@ljmu.ac.uk

Genetic diversity of the African malaria vector *Anopheles gambiae*

The *Anopheles gambiae* 1000 Genomes Consortium[*]

The sustainability of malaria control in Africa is threatened by the rise of insecticide resistance in *Anopheles* mosquitoes that transmit the disease[1]. To gain a deeper understanding of how mosquito populations are evolving, we sequenced the genomes of 765 specimens of *Anopheles gambiae* and *Anopheles coluzzii* sampled from 15 locations across Africa, identifying over 50 million single nucleotide polymorphisms within the accessible genome. These data revealed complex population structure and patterns of gene flow, with evidence of ancient expansions, recent bottlenecks, and local variation in effective population size. Strong signals of recent selection were observed in insecticide resistance genes, with multiple sweeps spreading over large geographical distances and between species. The design of novel tools for mosquito control using gene drive will need to take account of high levels of genetic diversity in natural mosquito populations.

Blood-sucking mosquitoes of the *Anopheles gambiae* species complex are the principal vectors of *Plasmodium falciparum* malaria in Africa. Substantial reductions in malaria morbidity and mortality have been achieved by the use of insecticide-based interventions[2], but increasing levels of insecticide resistance and other adaptive changes in mosquito populations threaten to reverse these gains[1]. A better understanding of the molecular, ecological and evolutionary processes driving these changes is essential to maximize the active lifespan of existing insecticides, and to accelerate the development of

---

[*] Lists of participants and their affiliations appear at the end of the paper

20   new strategies and tools for vector control. The *Anopheles gambiae* 1000 Genomes Project[*] (Ag1000G)

21   was established to provide a foundation for detailed investigation of mosquito genome variation and

22   evolution. Here we report the first phase of the project which analysed 765 wild-caught specimens of

23   *Anopheles gambiae sensu stricto* and *Anopheles coluzzii*. These two species account for the majority of

24   malaria transmission in Africa, and are morphologically indistinguishable and often sympatric, but are

25   genetically distinct[3,4] and differ in geographical range[5], larval ecology[6], behaviour[7] and strategies for

26   surviving the dry season[8]. The specimens were collected at 15 locations across 8 African countries,

27   spanning a range of ecologies including rainforest, inland savanna and coastal biomes, and thus provide

28   a broad sample in which to explore factors shaping mosquito population variation (Extended Data Fig. 1;

29   Supplementary Text 1).

30   Specimens were sequenced using the Illumina HiSeq platform and single nucleotide polymorphisms

31   (SNPs) were identified by alignment against the AgamP3 reference genome (Methods; Supplementary

32   Text 2). A rigorous evaluation of data quality, including the use of experimental genetic crosses to

33   quantify error rates, identified genomic regions totaling 141 Mbp (61% of the reference genome) that

34   were accessible for analysis of population variation (Supplementary Text 3; Extended Data Fig. 2). We

35   identified 52,525,957 high-quality SNPs, of which 21% had three or more alleles, an average of one

36   variant allele every 2.2 bases of the accessible genome (Fig. 1a). Individual mosquitoes carried between

37   1.7 and 2.7 million variant alleles, with no systematic difference observed between the two species

38   (Extended Data Fig. 3a). In most populations, nucleotide diversity was 1.5% on average (Extended Data

39   Fig. 3b) and >3% at synonymous coding sites (Extended Data Fig. 3c), confirming these are among the

40   most genetically diverse eukaryotic species[9].

---

[*] http://www.malariagen.net/ag1000g

41   High levels of natural diversity have practical implications for the development of gene drive

42   technologies for mosquito control[10]. CRISPR/Cas9 gene drives can be designed to edit a specific gene

43   and confer a phenotype such as female sterility, which could suppress mosquito populations and

44   thereby reduce disease transmission. However, naturally occurring polymorphisms within the ~21 bp

45   Cas9 target site could prevent target recognition, and thus undermine gene drive efficacy in the field.

46   We found viable Cas9 targets in 11,625 protein-coding genes, but only 5,474 genes remained after

47   excluding target sites with nucleotide variation in any of the 765 genomes sequenced here (Extended

48   Data Fig. 3d; Supplementary Text 5). Resistance to gene drive could be countered by designing

49   constructs that target multiple sites within the same gene, and we identified 863 genes that each

50   contain at least 10 non-overlapping conserved target sites, including 13 putative sterility genes[10]

51   (Supplementary Text 5.2). However, clearly more variants remain to be discovered (Extended Data Fig.

52   3d) and extensive sampling of multiple populations will be needed to inform the design of gene drives

53   that are robust to natural genetic variation.

54   *An. gambiae* and *An. coluzzii* have a geographical range spanning sub-Saharan Africa and encompassing

55   a variety of ecological settings[5]. Previous studies have found evidence that populations are locally

56   adapted, and that migration between populations is limited both by geographical distance and major

57   ecological discontinuities, notably the Congo Basin tropical rainforest and the East African rift system[11–

58   14]. As a starting point for analysis of population structure, we constructed neighbour-joining trees to

59   explore patterns of genetic similarity between individuals (Fig. 1b; Supplementary Text 6.1). We

60   observed four contrasting patterns of relatedness, associated with different regions of the genome.

61   Within pericentromeric regions of chromosomes X, 3 and arm 2R, mosquitoes segregated into two

62   highly distinct clades, largely corresponding to the two species as determined by conventional molecular

63   diagnostics, consistent with previous studies finding that genome regions of reduced recombination are

64   associated with stronger differentiation between closely-related species[15]. The large chromosomal

65   inversions 2La and 2Rb were each associated with a distinct pattern of relatedness, as expected if

66   recombination is reduced between inversion karyotypes. In most of the remaining genome, there was

67   evidence of clustering by geographical region but not by species. There were also some genome regions

68   where we found unusually short genetic distances between individuals from different populations and

69   species, indicating the influence of recent selective sweeps and adaptive gene flow.

70   To investigate geographical sub-divisions in more detail, we focused on euchromatic regions of

71   Chromosome 3, which are free from polymorphic inversions and regions of reduced recombination

72   (Supplementary Text 6). ADMIXTURE models and principal components analysis (PCA) supported five

73   major ancestral populations, corresponding to: (i) *An. gambiae* from Guinea, Burkina Faso, Cameroon

74   and Uganda; (ii) *An. gambiae* from Gabon; (iii) Kenya; (iv) Angola *An. coluzzii*; (v) Burkina Faso *An.*

75   *coluzzii* and Guinea-Bissau (Fig. 2; Extended Data Figs. 4, 5). Within each species, we found relatively

76   high allele frequency differentiation across the Congo Basin rainforest, exceeding differentiation

77   between the two species at a single location (Extended Data Fig. 5b). There were also more subtle

78   distinctions within and between populations. For example, in Cameroon mosquitoes were sampled

79   along a cline from savanna into forest, and there was some population structure associated with these

80   different ecologies. However, among *An. gambiae* populations north of the Congo Basin, differentiation

81   was extremely weak overall, despite considerable distances between populations, suggesting substantial

82   gene flow.

83   Earlier studies concluded that purposeful movement of *Anopheles* mosquitoes is limited to short-range

84   dispersal up to 5 km[16]; however, recent evidence has emerged for long-distance seasonal migration in

85   *An. gambiae*[8]. To explore evidence for migration, we computed joint site frequency spectra for selected

86   population pairs and fitted models of population history (Methods; Supplementary Text 8). For all pairs

87   examined, models with migration provided a better fit than models without migration (Supplementary

88   Table 2). The inferred rate of migration was high between *An. gambiae* savanna populations, but some

89   migration was also inferred between species and across both the Congo Basin rainforest and the East

90   African rift. Although these analyses do not allow us to infer the timing or direction of gene flow events,

91   they suggest that mosquito migration between different parts of the continent could impact on the

92   spread of insecticide resistance and dynamics of disease transmission.

93   A key question in mosquito evolution concerns the extent and impact of gene flow between species, and

94   *An. gambiae* and *An. coluzzii* are known to undergo hybridization at a rate that varies over space and

95   time[17]. To study this phenomenon, we analyzed 506 SNPs previously found to be highly differentiated

96   between the two species[18] (Extended Data Fig. 6; Supplementary Text 6.6). These ancestry-informative

97   markers (AIMs) showed that a genomic region on chromosome arm 2L has introgressed from *An.*

98   *gambiae* into *An. coluzzii* in Burkina Faso and Angola. This region spans the *Vgsc* gene where

99   introgression of insecticide resistance alleles has been reported in Ghana[19] and Mali[20], although this is

100 the first evidence that introgressed alleles have spread to *An. coluzzii* south of the Congo Basin. AIMs

101 also highlighted two populations with uncertain species status. In Guinea-Bissau, mosquitoes carried a

102 mixture of alleles from both species on all chromosomes. These individuals were sampled from the

103 coast, within a region of West Africa that is believed to be a zone of secondary contact because previous

104 studies have found evidence for extensive introgression[21,22]. We also found that mosquitoes from

105 coastal Kenya carried a mixture of both species' alleles on all chromosomes. This was unexpected, as the

106 geographical range of *An. coluzzii* is not thought to extend beyond the East African rift. There are several

107 possible explanations for the Kenyan data, including historical admixture between species and retention

108 of ancestral variation, and further analysis and population sampling are required. However, our data

109 demonstrate that a simple *gambiae/coluzzii* dichotomy is not adequate for describing malaria vector

110 species composition in some parts of Africa, and caution against the use of any single marker to infer

111 species ancestry or recent hybridization.

112   Historical fluctuations in effective population size (N_e) can be inferred from the genomes of extant

113   individuals. Analysis of our genome variation data indicated a major expansion in all populations north

114   of the Congo Basin and west of the East African rift (Fig. 3a; Extended Data Fig. 7; Methods;

115   Supplementary Text 8). Knowledge of the *Anopheles* mutation rate is required to date this expansion,

116   and this has not yet been determined, but assuming it is similar to *Drosophila* then the onset of

117   expansion would be within the range 7,000 to 25,000 years ago (Fig. 3a; Methods). Since *An. gambiae*

118   and *An. coluzzii* are highly anthropophilic, mosquito population expansion could be linked to that of

119   humans, and particularly to the expansion of agricultural Bantu-speaking groups originating from north

120   of the Congo Basin beginning ~5,000 years ago[23]. It is possible to reconcile this theory with our data if

121   *Anopheles* has a higher mutation rate than *Drosophila*, causing us to over-estimate the age of the

122   expansion, but it is also possible that mosquito populations benefited from earlier human population

123   growth, or that other factors such as climate change played a role.

124   We also observed genomic signatures of a major recent population decline of *An. gambiae* in coastal

125   Kenya. All Kenyan specimens (but no specimens from other locations) had long runs of homozygosity

126   comprising 10-60% of the genome, indicating high levels of inbreeding consistent with a recent

127   population bottleneck (Fig. 3b). In Kenya, free mass distribution of insecticide-treated nets (ITNs)

128   starting in 2006 resulted in a major increase in ITN coverage[24]. The specimens in this study were

129   collected in 2012, raising the question of whether the population decline of *An. gambiae* can be

130   attributed to ITN usage. To address this question, we analysed sharing of genome regions that are

131   identical by descent (IBD) (Methods; Extended Data Figs. 8a, 8b). We estimated that the *An. gambiae*

132   population in Kenya has fallen in size by at least two orders of magnitude, to N_e <1,000 (Extended Data

133   Fig. 8c; Supplementary Text 8.4). The beginning of this inferred decline occurred approximately 200

134   generations before the date of sampling, which would pre-date mass ITN distributions, assuming ~11

135   generations per year. This is consistent with other studies that have found evidence for low N_e[11] and

136 changes in mosquito species abundance[25] in the region prior to high levels of ITN coverage.

137 Nevertheless, our data show that major demographic events leave genetic signatures that could be used

138 to gain important information about the impact of vector control interventions.

139 Many genes have been associated with insecticide resistance in *Anopheles*, but different genetic variants

140 may be responsible for resistance in different populations, and it is not yet clear where or how

141 resistance is spreading. Genomic data can help address these questions by identifying genes with

142 evidence of recent evolutionary adaptation in one or more mosquito populations. We found strong

143 signals of recent positive selection at several genes that are known to play a role in resistance, including:

144 *Vgsc*, the target site for DDT and pyrethroid insecticides[26]; *Gste*, a cluster of glutathione S-transferase

145 genes including *Gste2*, previously implicated in metabolism of DDT and pyrethroids[27]; and *Cyp6p*, a

146 cluster of genes encoding cytochrome P450 enzymes, including *Cyp6p3* which is upregulated in

147 permethrin and bendiocarb resistant mosquitoes[28] (Extended Data Fig. 9; Supplementary Text 9). We

148 also observed strong signals of selection at multiple loci with no known resistance genes, and these

149 merit detailed investigation in future studies.

150 Mutations in *An. gambiae Vgsc* codon 995 (orthologous to *Musca domestica Vgsc* codon 1014), known

151 as "*kdr*" due to their knock-down resistance phenotype, reduce susceptibility to DDT and pyrethroids[26].

152 We found the Leucine→Phenylalanine (L995F) *kdr* variant at high frequency in West and Central Africa

153 (Guinea 100%; Burkina Faso 93%; Cameroon 53%; Gabon 36%; Angola 86%). A second *kdr* allele,

154 Leucine→Serine (L995S), was present in Central and East Africa (Cameroon 15%; Gabon 65%; Uganda

155 100%; Kenya 76%). To investigate the evolution and spread of the two *kdr* alleles, we analyzed the

156 genetic backgrounds on which they were carried (Fig. 4; Supplementary Text 9.3). L995F occurred within

157 five distinct haplotype clusters (labeled F1-F5 in Fig. 4), while L995S was found in a further 5 haplotype

158 clusters (labeled S1-S5 in Fig. 4). Cluster F1 contained individuals of both species and from 4 countries

159   spanning the Congo Basin, proving that recent gene flow has carried resistance alleles between these

160   populations. Three *kdr* haplotypes (F4, F5, S2) were found in both Cameroon and Gabon, providing

161   multiple examples of recent gene flow between these two populations. The S3 haplotype was present in

162   both Uganda and coastal Kenya, thus resistance alleles can reach populations on both sides of the rift

163   system.

164   While the evolution of resistance in the *Vgsc* gene is clearly driven primarily by the two *kdr* alleles, we

165   also found 15 other non-synonymous variants at a frequency above 1% in our cohort (Fig. 4). 13 of these

166   variants occurred almost exclusively on haplotypes carrying the L995F allele (D' > 0.96). These included

167   N1570Y, previously found on L995F haplotypes in West and Central Africa and shown to confer

168   increased resistance[29]. Overall there was a highly significant enrichment for non-synonymous mutations

169   on haplotypes carrying the L995F allele, indicating secondary selection on multiple variants that either

170   enhance or compensate for the L995F phenotype (Supplementary Text 9.5).

171   Resistance due to genes that enhance insecticide metabolism is also a serious concern, as it has been

172   implicated in extreme resistance phenotypes in some *Anopheles* populations[27,28]. Although several

173   metabolic genes have been shown to be upregulated in resistant mosquitoes, only a single molecular

174   marker of metabolic resistance (*Gste2*-I114T) has previously been identified in *An. gambiae* or *An.*

175   *coluzzii*[27]. At both *Gste* and *Cyp6p* we found evidence that resistance has emerged on multiple genetic

176   backgrounds and is spreading between species and over considerable distances. At the *Gste* locus we

177   found at least four distinct haplotypes under selection (Extended Data Fig. 10a). One of these

178   haplotypes carried the known *Gste2*-I114T resistance allele, and this haplotype was found in all

179   populations except Guinea-Bissau and Uganda, indicating a continent-wide spread. However, the other

180   three haplotypes did not carry this allele, thus other genetic variants with a resistance phenotype must

181   be present at this locus. At the *Cyp6p* locus we found at least eight distinct haplotypes under selection,

182   but limited spread between populations (Extended Data Fig. 10b). At both loci, we found multiple SNPs

183   associated with haplotypes under selection which could be used as markers to track the spread of

184   resistance and characterize resistance phenotypes (Extended Data Fig. 10).

185   In 1899 Ronald Ross proposed that malaria could be controlled by destroying breeding sites of the

186   mosquitoes that transmit the disease[30]. *An. gambiae*, identified in the same year by Ross as a vector of

187   malaria in Africa, has proved resilient to a century of attempts to repress it. The vector control

188   armamentarium needs to be expanded, not only with new classes of insecticide and novel genetic

189   control strategies, but also with tools for gathering intelligence, to enable those responsible for planning

190   and executing interventions to stay ahead of the mosquito's remarkable capacity for rapid evolutionary

191   adaptation. There remain major knowledge gaps concerning the ecology and life history of *Anopheles*

192   mosquitoes, such as the rate and range of migration, which are fundamental to understanding both

193   malaria transmission and the spread of insecticide resistance, and which will require spatiotemporal

194   analysis of mosquito populations. Most importantly, it is essential to start collecting population genomic

195   data prospectively as an integral part of vector control interventions, to identify which strategies are

196   causing increased insecticide resistance, or what it takes to cause a population crash of the magnitude

197   observed in our Kenyan data. By treating each intervention as an experiment, and by analyzing its

198   impact on both mosquito and parasite populations, we can aim to improve the efficacy and

199   sustainability of future interventions, while at the same time learning about basic processes in ecology

200   and evolution.

References

202   1.   Hemingway, J. *et al.* Averting a malaria disaster: will insecticide resistance derail malaria control?

203   *Lancet* (2016). doi:10.1016/S0140-6736(15)00417-1

204   2.   Bhatt, S. *et al.* The effect of malaria control on Plasmodium falciparum in Africa between 2000

205   and 2015. *Nature* **526,** 207–211 (2015).

206   3.   Torre, A. della *et al.* Molecular evidence of incipient speciation within Anopheles gambiae s.s. in

207   West Africa. *Insect Mol. Biol.* **10,** 9–18 (2001).

208   4.   Lawniczak, M. K. N. *et al.* Widespread divergence between incipient Anopheles gambiae species

209   revealed by whole genome sequences. *Science* **330,** 512–4 (2010).

210   5.   Tene Fossog, B. *et al.* Habitat segregation and ecological character displacement in cryptic African

211   malaria mosquitoes. *Evol. Appl.* n/a-n/a (2015). doi:10.1111/eva.12242

212   6.   Diabate, A. *et al.* Larval development of the molecular forms of Anopheles gambiae (Diptera:

213   Culicidae) in different habitats: a transplantation experiment. *J Med Entomol* **42,** 548–553 (2005).

214   7.   Gimonneau, G. *et al.* A behavioral mechanism underlying ecological divergence in the malaria

215   mosquito Anopheles gambiae. *Behav. Ecol.* **21,** 1087–1092 (2010).

216   8.   Dao, A. *et al.* Signatures of aestivation and migration in Sahelian malaria mosquito populations.

217   *Nature* **516,** 387–90 (2014).

218   9.   Leffler, E. M. *et al.* Revisiting an Old Riddle: What Determines Genetic Diversity Levels within

219   Species? *PLoS Biol.* **10,** e1001388 (2012).

220   10.   Hammond, A. *et al.* A CRISPR-Cas9 gene drive system targeting female reproduction in the

221   malaria mosquito vector *Anopheles gambiae*. *Nat. Biotechnol.* 1–8 (2015). doi:10.1038/nbt.3439

222   11.   Lehmann, T. *et al.* The Rift Valley Complex as a Barrier to Gene Flow for Anopheles gambiae in

223   Kenya. *J. Hered.* **91,** 165–168 (1999).

224   12.   Lehmann, T. Population Structure of Anopheles gambiae in Africa. *J. Hered.* **94,** 133–147 (2003).

225   13.   Slotman, M. A. *et al.* Evidence for subdivision within the M molecular form of Anopheles

226   gambiae. *Mol. Ecol.* **16,** 639–649 (2006).

227   14.   Pinto, J. *et al.* Geographic population structure of the African malaria vector Anopheles gambiae

228   suggests a role for the forest-savannah biome transition as a barrier to gene flow. *Evol. Appl.* **6,**

229   910–24 (2013).

230   15.   Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due

231   to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23,** 3133–57 (2014).

232   16.   Service, M. W. Mosquito (Diptera: Culicidae) dispersal--the long and short of it. *J Med Entomol*

233   **34,** 579–588 (1997).

234   17.   Lee, Y. *et al.* Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S

235   forms of the malaria mosquito, Anopheles gambiae. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 19854–9

236   (2013).

237   18.   Neafsey, D. E. *et al.* SNP genotyping defines complex gene-flow boundaries among African

238   malaria vector mosquitoes. *Science* **330,** 514–7 (2010).

239   19.   Clarkson, C. S. *et al.* Adaptive introgression between Anopheles sibling species eliminates a major

240   genomic island but not reproductive isolation. *Nat. Commun.* **5,** 4248 (2014).

241   20.   Norris, L. C. *et al.* Adaptive introgression in an African malaria mosquito coincident with the

242   increased usage of insecticide-treated bed nets. *Proc. Natl. Acad. Sci.* 201418892 (2015).

243   doi:10.1073/pnas.1418892112

244   21.   Vicente, J. L. *et al.* Massive introgression drives species radiation at the range limit of Anopheles

245   gambiae. *Sci. Rep.* **7,** 46451 (2017).

246   22.   Nwakanma, D. C. *et al.* Breakdown in the process of incipient speciation in Anopheles gambiae.

247   *Genetics* **193,** 1221–31 (2013).

248   23.   Li, S., Schlebusch, C. & Jakobsson, M. Genetic variation reveals large-scale population expansion

249   and migration during the expansion of Bantu-speaking peoples. *Proc. R. Soc. London B Biol. Sci.*

250   **281,** (2014).

251   24.   Noor, A. M. *et al.* Increasing Coverage and Decreasing Inequity in Insecticide-Treated Bed Net

252   Use among Rural Kenyan Children. *PLoS Med.* **4,** e255 (2007).

253   25.   Mwangangi, J. M. *et al.* Shifts in malaria vector species composition and transmission dynamics

254   along the Kenyan coast over the past 20 years. *Malar. J.* **12,** 13 (2013).

255   26.   Davies, T. G. E., Field, L. M., Usherwood, P. N. R. & Williamson, M. S. A comparative study of

256   voltage-gated sodium channels in the Insecta: Implications for pyrethroid resistance in

257   Anopheline and other Neopteran species. *Insect Mol. Biol.* (2007). doi:10.1111/j.1365-

258   2583.2007.00733.x

259   27.   Mitchell, S. N. *et al.* Metabolic and target-site mechanisms combine to confer strong DDT

260   resistance in Anopheles gambiae. *PLoS One* **9,** e92662 (2014).

261   28.   Edi, C. V. *et al.* CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple

262   insecticide resistance in the malaria mosquito Anopheles gambiae. *PLoS Genet.* **10,** e1004236

263   (2014).

264   29.   Jones, C. M. *et al.* Footprints of positive selection associated with a mutation (N1575Y) in the

265   voltage-gated sodium channel of Anopheles gambiae. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 6614–9

266   (2012).

267   30.   Ross, R. Inaugural Lecture on the Possibility of Extirpating Malaria from Certain Localities by a

268   New Method. *Br. Med. J.* **2,** 1–4 (1899).

Supplementary information

270   Further information is given in the Supplementary Text.

Acknowledgments

Author information

Corresponding authors

285   Alistair Miles[1,2], Mara K. N. Lawniczak[1], Martin J. Donnelly[3,1], Dominic P. Kwiatkowski[1,2].

The Anopheles gambiae 1000 Genomes Consortium

Data analysis group. Alistair Miles[1,2] (project lead), Nicholas J. Harding[2], Giordano Bottà[4,2], Chris S. Clarkson[1,3], Tiago Antão[5,3,2], Krzysztof Kozak[1], Daniel R. Schrider[6], Andrew D. Kern[6], Seth Redmond[7], Igor Sharakhov[8,9], Richard D. Pearson[1,2], Christina Bergey[10], Michael C. Fontaine[11], Martin J. Donnelly[3,1], Mara K. N. Lawniczak[1], Dominic P. Kwiatkowski[1,2] (chair).

Partner working group. Martin J. Donnelly[3,1] (chair), Diego Ayala[12,13], Nora J. Besansky[10], Austin Burt[14], Beniamino Caputo[4], Alessandra della Torre[4], Michael C. Fontaine[11], H. Charles J. Godfray[15], Matthew W. Hahn[16], Andrew D. Kern[6], Dominic P. Kwiatkowski[1,2], Mara K. N. Lawniczak[1], Janet Midega[17], Daniel E. Neafsey[7], Samantha O'Loughlin[14], João Pinto[18], Michelle M. Riehle[19], Igor Sharakhov[8,9], Kenneth D. Vernick[20], David Weetman[3], Craig S. Wilding[21,3], Bradley J. White[22].

Sample collections. Angola: Arlete D. Troco[23], João Pinto[18]; **Burkina Faso:** Abdoulaye Diabaté[24], Samantha O'Loughlin[14], Austin Burt[14]; **Cameroon:** Carlo Costantini[13,25], Kyanne R. Rohatgi[10], Nora J. Besansky[10]; **Gabon:** Nohal Elissa[12], João Pinto[18]; **Guinea:** Boubacar Coulibaly[26], Michelle M. Riehle[19], Kenneth D. Vernick[20]; **Guinea-Bissau:** João Pinto[18], João Dinis[27]; **Kenya:** Janet Midega[17], Charles Mbogo[17], Philip Bejon[17]; **Uganda:** Craig S. Wilding[21,3], David Weetman[3], Henry D. Mawejje[28], Martin J. Donnelly[3,1]; **Crosses:** David Weetman[3], Craig S. Wilding[21,3], Martin J. Donnelly[3,1].

Sequencing and data production. Jim Stalker[1], Kirk Rockett[2], Eleanor Drury[1], Daniel Mead[1], Anna Jeffreys[2], Christina Hubbart[2], Kate Rowlands[2], Alison T. Isaacs[3], Dushyanth Jyothi[1], Cinzia Malangone[1].

Web application development. Paul Vauterin[2], Ben Jeffrey[2], Ian Wright[2], Lee Hart[2], Krzysztof Kluczyński[2].

Project coordination. Victoria Cornelius[2], Bronwyn MacInnis[29], Christa Henrichs[2], Rachel Giacomantonio[1], Dominic P. Kwiatkowski[1,2].

Affiliations

1. Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

2. MRC Centre for Genomics and Global Health, University of Oxford, Oxford OX3 7BN, UK

3. Department of Vector Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK

4. Istituto Pasteur Italia – Fondazione Cenci Bolognetti, Dipartimento di Sanita Pubblica e Malattie Infettive, Università di Roma SAPIENZA, Rome, Italy

5. University of Montana, Missoula, MT 59812, USA

6. Department of Genetics, Rutgers University, 604 Alison Road, Piscataway, NJ 08854, USA

7. Genome Sequencing and Analysis Program, Broad Institute, 415 Main Street, Cambridge, MA 02142, USA

8. Department of Entomology, Virginia Tech, Blacksburg, VA 24061, USA

9. Laboratory of Ecology, Genetics and Environmental Protection, Tomsk State University, Tomsk 634050, Russia

10. Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, IN 46556, USA

11. Groningen Institute for Evolutionary Life Sciences (GELIFES), Nijenborgh 7, 9747 AG Groningen, The Netherlands

12. Unité d'Ecologie des Systèmes Vectoriels, Centre International de Recherches Médicales de Franceville, Franceville, Gabon

328   13. Institut de Recherche pour le Développement (IRD), UMR MIVEGEC (UM1, UM2, CNRS 5290, IRD

329   224), Montpellier, France

330   14. Department of Life Sciences, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, UK

331   15. Department of Zoology, University of Oxford, The Tinbergen Building, South Parks Road, Oxford OX1

332   3PS, UK

333   16. Department of Biology and School of Informatics and Computing, Indiana University, Bloomington,

334   IN 47405, USA

335   17. KEMRI-Wellcome Trust Research Programme, PO Box 230, Bofa Road, Kilifi, Kenya

336   18. Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT,

337   Universidade Nova de Lisboa, UNL, Rua da Junqueira 100, 1349-008 Lisbon, Portugal

338   19. Department of Microbiology and Immunology, Microbial and Plant Genomics Institute, University of

339   Minnesota, St. Paul, MN 55108

340   20. Unit for Genetics and Genomics of Insect Vectors, Institut Pasteur, Paris, France

341   21. School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool L3 3AF, UK

342   22. Department of Entomology, University of California, Riverside, CA, USA

343   23. Programa Nacional de Controle da Malária, Direcção Nacional de Saúde Pública, Ministério da Saúde,

344   Luanda, Angola

345   24. Institut de Recherche en Sciences de la Santé (IRSS), Bobo Dioulasso, Burkina Faso

346   25. Laboratoire de Recherche sur le Paludisme, Organisation de Coordination pour la lutte contre les

347   Endémies en Afrique Centrale (OCEAC), Yaoundé, Cameroon

348   26. Malaria Research and Training Centre, Faculty of Medicine and Dentistry, University of Mali

349   27. Instituto Nacional de Saúde Pública, Ministério da Saúde Pública, Bissau, Guiné-Bissau

350   28. Infectious Diseases Research Collaboration, 2C Nakasero Hill Road, P.O. Box 7475, Kampala, Uganda

351   29. The Broad Institute of Massachusetts Institute of Technology and Harvard, 415 Main Street,

352   Cambridge, MA 02142, USA

Contributions

354   Details of author contributions are given in the consortium author list.

Competing financial interests

356   The authors declare no competing financial interests.

Data availability

358   Sequence read alignments and variant calls from Ag1000G phase 1 are available from the European

359   Nucleotide Archive (ENA - http://www.ebi.ac.uk/ena) under study PRJEB18691. Variant and haplotype

360   calls and associated data from Ag1000G phase 1 can be explored via an interactive web application or

361   downloaded via the MalariaGEN website (https://www.malariagen.net/projects/ag1000g#data).

Figure legends

363   **Figure 1. Patterns of genomic variation. a**, Density of nucleotide variation in 200 kbp windows over the

364   genome. **b**, Variation in the pattern of relatedness between individual mosquitoes over the genome. The

365   three chromosomes are painted using colours to represent the major pattern of relatedness found

366   within each 100 kbp window. Below, neighbour-joining trees are shown from a selection of genomic

367   windows that are representative of the four major patterns of relatedness found, as well as for the

368   window spanning the *Vgsc* gene. AO=Angola; BF=Burkina Faso; GW=Guinea-Bissau; GN=Guinea;

369   CM=Cameroon; GA=Gabon; UG=Uganda; KE=Kenya.

370   **Figure 2. Geographical population structure and migration.** In the upper panel, each mosquito is

371   depicted as a vertical bar painted by the proportion of the genome inherited from each of K=8 inferred

372   ancestral populations. Pie charts on the map depict the same ancestry proportions summed over all

373   individuals for each population. Text in white shows average F_{ST} followed in parentheses by estimates of

374   the population migration rate (2Nm).

375   **Figure 3. Population size history. a**, Stairway Plot of changes in population size over time. Absolute

376   values of time and N_e are shown on alternative axes as a range of values, assuming lower and upper

377   limits for the mutation rate μ as 2.8x10^{-9} and 5.5x10^{-9} respectively and T=11 generations per year. **b**,

378   Runs of homozygosity (*ROH*) in individual mosquitoes, highlighting recent inbreeding in Kenyan (grey)

379   and colony mosquitoes (black; P=Pimperena, M=Mali, K=Kisumu, G=Ghana).

380   **Figure 4. Evolution and spread of insecticide resistance in the *Vgsc* gene.** The upper panel shows a

381   dendrogram obtained by hierarchical clustering of haplotypes from wild-caught individuals. The colour

382   bar below shows the population of origin for each haplotype. The lower panel shows alleles carried by

383   each haplotype at 17 non-synonymous SNPs with alternate allele frequency > 1% (white=reference

384   allele, black=alternate allele, red=previously known resistance allele). At the lower margin, we label 10

385   haplotype clusters carrying a *kdr* allele (either L995F or L995S). The inset map depicts haplotypes shared

386   between populations, demonstrating the spread of insecticide resistance.

Methods

388   **Population sampling.** Mosquitoes were collected from natural populations at 15 sampling sites in 8

389   African countries (Extended Data Fig. 1). Sampling locations, dates, specimen collection methods and

390   DNA extraction methods are given in Supplementary Text 1.1. We also performed genetic crosses

391   between adult mosquitoes obtained from lab colonies (Supplementary Text 1.2). Parents and progeny of

392   four crosses were contributed to Ag1000G phase 1 (Extended Data Fig. 1).

393   **Whole genome sequencing.** Sequencing was performed on the Illumina HiSeq 2000 platform at the

394   Wellcome Trust Sanger Institute. Paired-end multiplex libraries were prepared using the manufacturer's

395   protocol, with the exception that genomic DNA was fragmented using Covaris Adaptive Focused

396   Acoustics rather than nebulization. Multiplexes comprised 12 tagged individual mosquitoes and three

397   lanes of sequencing were generated for each multiplex to even out variation in yield between

398   sequencing runs. Cluster generation and sequencing were undertaken per the manufacturer's protocol

399   for paired-end 100 bp sequence reads with insert size in the range 100-200 bp.

400   **Sequence analysis and variant calling.** Sequence reads were aligned to the AgamP3 reference genome[31]

401   using bwa[32] and SNPs were discovered using GATK following best practice recommendations[33,34]

402   (Supplementary Text 3.1, 3.2). After sample quality control, we analyzed data on 765 wild-caught

403   specimens and a further 80 specimens comprising parents and progeny from the four lab crosses

404   (Supplementary Text 3.3). The alignments were also used to identify genome regions accessible to SNP

405   calling, where short reads could be uniquely mapped and there was minimal evidence for structural

406   variation (Supplementary Text 3.4). Mendelian errors in the crosses were used to guide the design of

407   filters to remove poor quality variant calls (Supplementary Text 3.5). We performed capillary sequencing

408   of five genes in 58 individual mosquitoes to provide an estimate for the SNP false discovery rate (FDR),

409   sensitivity and genotyping accuracy (Supplementary Text 3.6). We also performed genotyping by primer-

410   extension mass spectrometry using the Sequenom MassARRAY® platform at 158 SNPs in 229 individual

411   mosquitoes to provide a second estimate for genotyping accuracy (Supplementary Text 3.7).

412   **Haplotype estimation.** We used SHAPEIT2 to perform statistical phasing with information from

413   sequence reads[35] for all wild-caught individuals (Supplementary Text 4.1). We assessed phasing

414   performance by comparison with haplotypes generated from the crosses and from male X chromosome

415   haplotypes (Supplementary Text 4.2; Extended Data Fig. 2b, 2c).

416   **Population structure.** To investigate variation in patterns of relatedness along the genome, we

417   performed a windowed analysis using genetic distance and neighbour-joining trees (NJT). We divided

418   the genome into 1,418 contiguous non-overlapping windows, where each window contained 100 kbp of

419   accessible positions. Within each window, we computed the city-block distance between all pairs of

420   individuals. We used these distance matrices to construct a NJT for each window. We then computed

421   the Pearson correlation coefficient between all pairs of distance matrices, and performed a singular

422   value decomposition (SVD) on the correlation matrix. The resulting SVD components were used to

423   identify major patterns of relatedness (Supplementary Text 6.1). We analysed geographical population

424   structure using ADMIXTURE[36] and PCA[37]. For these analyses, we used biallelic SNPs from within the

425   regions 3R:1-37Mbp and 3L:15-41Mbp and with minor allele frequency >= 1%, then each chromosome

426   arm was randomly down-sampled to 100,000 variants using 10 different random seeds to provide 10

427   replicate variant sets, then each set was pruned to remove variants in linkage disequilibrium

428   (Supplementary Text 6.2). For each of the 10 replicate variant sets, ADMIXTURE was run for K (number

429   of ancestral populations) from 2 to 11 with 5-fold cross-validation. Each ADMIXTURE analysis was

430   repeated 10 times with different seeds, resulting in a total of 100 runs for each value of K. We then used

431   CLUMPAK[38] to analyse the ADMIXTURE results and compute ancestry proportions (Supplementary Text

432   6.2). Average F_{ST} was computed using Hudson's estimator and the ratio of averages, and standard errors

433   were computed using a block-jackknife[39] (Supplementary Text 6.4). Ancestry informative markers (AIMs)

434   were ascertained by starting with SNPs previously discovered in Mali[18] with an allele frequency

435   difference between *An. gambiae* and *An. coluzzii* > 0.9, then taking the intersection with biallelic SNPs

436   discovered in this study, resulting in 506 AIMs (Supplementary Text 6.6).

437   **Population size history.** We inferred the scale and timing of historical changes in N_e using two methods,

438   Stairway Plot[40] and ∂a∂i[41], both using site frequency spectra but taking different modelling approaches.

439   To compute site frequency spectra, we used SNPs from within the regions 3R:1-37 Mbp and 3L:15-41

440   Mbp, taking only intergenic SNPs at least 5 kbp from the nearest gene (Supplementary Text 8). We

441   modified Stairway Plot to include an additional parameter representing the probability of ancestral

442   misclassification for each SNP (Supplementary Text 8.1). We fitted a three-epoch (two N_e changes) ∂a∂i

443   model for each population singly, and fitted joint population models for selected pairs of populations

444   (Supplementary Text 8.2). Scaling of parameters assumed that the *Anopheles* mutation rate is within the

445   range of values estimated for *Drosophila,* where estimates[42,43] range from 2.8×10^{-9} to 5.5×10^{-9}. For joint

446   population models, we computed the joint site frequency spectrum for each pair of populations from

447   the same set of SNPs used for single-population inferences. Joint population models allowed for a phase

448   of exponential size change in the ancestral population up until the time of the population split, after

449   which each of the daughter populations experienced their own exponential size change until the

450   present. We fitted these models with and without the addition of a symmetric, bidirectional migration

451   rate parameter following the split. To study recent population history in Kenya we used IBDseq[44] to infer

452   genome tracts identical by descent (IBD) then ran IBDN_e[45] to infer population size history

453   (Supplementary Text 8.4).

454   **Recent selection.** To scan the genome for signals of recent selection, we computed the H12 haplotype

455   diversity statistic[46] for each population, and the cross-population extended haplotype homozygosity (XP-

456   EHH) score[47] for selected pairs of populations. H12 was computed in non-overlapping windows over the

457   genome, where each window contained a fixed number of SNPs, and window-sizes were calibrated

458   separately for each population to account for differences in the extent of linkage disequilibrium

459   (Supplementary Text 9.1). XP-EHH was computed for all SNPs with a minor allele frequency ≥ 5% in the

460   union of both populations in each pair, and normalized within each chromosome (Supplementary Text

461   9.2). To study haplotype structure at the *Vgsc*, *Gste* and *Cyp6p* loci, we computed the Hamming distance

462   between all pairs of haplotypes, then performed hierarchical clustering of haplotypes (Supplementary

463   Text 9.3). To identify haplotype clusters resulting from recent selection, we cut the dendrograms at a

464   small genetic distance (0.0004 SNP differences per accessible bp) and studied the largest clusters

465   obtained after cutting. To look for evidence that the haplotype clusters we identified were related via

466   recombination events, we performed the same clustering analysis but in non-overlapping windows

467   upstream and downstream of the target region and compared the resulting clusters.

468   **Plotting and maps.** All figures were produced using the matplotlib package for Python[48]. The map

469   component of Fig. 2 was produced via the matplotlib basemap package, using the NASA Blue Marble

470   image as the map background. The map components of Fig. 4 and Extended Data Fig. 10 were plotted

471   via the cartopy package, using the Natural Earth shaded relief raster as the map background. The map in

472   Extended Data Fig. 1 was plotted via the cartopy package, using data from the map of standardized

473   terrestrial ecosystems of Africa[49] as the map background.

474

475   31.   Sharakhova, M. V *et al.* Update of the Anopheles gambiae PEST genome assembly. *Genome Biol.*

476   **8,** R5 (2007).

477   32.   Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.

478   *Bioinformatics* **25,** 1754–60 (2009).

479   33.   DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation

480   DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

481   34.   Van der Auwera, G. A. *et al. Current Protocols in Bioinformatics*. *Current protocols in*

482   *bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **11,** (John Wiley & Sons, Inc.,

483   2013).

484   35.   Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using

485   sequencing reads. *Am. J. Hum. Genet.* **93,** 687–96 (2013).

486   36.   Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated

487   individuals. *Genome Res.* **19,** 1655–64 (2009).

488   37.   Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,**

489   2074–2093 (2006).

490   38.   Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak : a program

491   for identifying clustering modes and packaging population structure inferences across *K*. *Mol.*

492   *Ecol. Resour.* **15,** 1179–1191 (2015).

493   39.   Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the

494   impact of rare variants. *Genome Res.* **23,** 1514–21 (2013).

495   40.   Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47,**

496   555–559 (2015).

497   41.   Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint

498   demographic history of multiple populations from multidimensional SNP frequency data. *PLoS*

499   *Genet.* **5,** e1000695 (2009).

500   42.   Keightley, P. D., Ness, R. W., Halligan, D. L. & Haddrill, P. R. Estimation of the spontaneous

501   mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics* **196,** 313–

502   20 (2014).

503 43. Schrider, D. R., Houle, D., Lynch, M. & Hahn, M. W. Rates and genomic consequences of

504   spontaneous mutational events in Drosophila melanogaster. *Genetics* **194,** 937–54 (2013).

505 44. Browning, B. L. *et al.* Detecting identity by descent and estimating genotype error rates in

506   sequence data. *Am. J. Hum. Genet.* **93,** 840–51 (2013).

507 45. Browning, S. R. & Browning, B. L. Accurate Non-parametric Estimation of Recent Effective

508   Population Size from Segments of Identity by Descent. *Am. J. Hum. Genet.* **97,** 404–418 (2015).

509 46. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent Selective Sweeps in North

510   American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet.* **11,** 1–32 (2015).

511 47. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human

512   populations. *Nature* **449,** 913–8 (2007).

513 48. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).

514 49. Sayre, R. G. *et al.* A new map of standardized terrestrial ecosystems of Africa. *African Geogr. Rev.*

515   (2013).

516 50. Sharakhova, M. V *et al.* Genome mapping and characterization of the Anopheles gambiae

517   heterochromatin. *BMC Genomics* **11,** 459 (2010).

Extended data figure legends

519 **Extended Data Figure 1. Overview of population sampling.** Red circles show sampling locations for

520 wild-caught mosquitoes. Colours in the map represent ecosystem classes; dark green represents forest

521   ecosystems, see (49) Fig. 9 for a complete colour legend. The Congo Basin tropical rainforest is the large

522   region of dark green in Central Africa. Sampling details for each site are shown in light grey boxes,

523   including country (two-letter country code), location and year of collection, predominant ecosystem

524   classification for the local region, and number and sex of individuals sequenced. For colony crosses, the

525   direction of cross (colony of origin of mother and father) and number of offspring is shown. The inset

526   map depicts geological fault lines in the East African rift system[*]. Species assignment for Guinea-Bissau

527   and Kenya specimens is uncertain, see main text. Sequencing depth per individual is shown as median

528   (5th – 95th percentile) for each population.

529   **Extended Data Figure 2. Genome accessibility and haplotype validation. a**, Percentage of accessible

530   bases in non-overlapping 400 kbp windows. The schematic of chromosomes below shows chromatin

531   state predictions from (50). **b**, Haplotypes inferred in the crosses. Each panel shows either maternal or

532   paternal haplotypes from a single cross. Each row within a panel represents a single progeny haplotype.

533   Haplotypes are coloured by parental inheritance (blue=allele from parent's first chromosome, red=allele

534   from parent's second chromosome). Switches between colours along a haplotype indicate

535   recombination events. Regions that were within a run of homozygosity in the parent and thus not

536   informative for haplotype validation are masked in grey. **c**, Error rate estimates for haplotypes inferred

537   in wild-caught individuals. Upper plots show estimates for the mean switch distance (red line),

538   compared to the mean switch distance if heterozygotes were phased randomly (black line). Lower plots

539   show the switch error rate (probability of a switch error occurring between two adjacent heterozygous

540   genotype calls).

541   **Extended Data Figure 3. Variant discovery and nucleotide diversity. a**, Number of variant alleles

542   discovered per individual mosquito. Only females are plotted. **b**, Genetic diversity within populations.

---

[*] http://pubs.usgs.gov/publications/text/East_Africa.html

543 Nucleotide diversity (π) and Tajima's D were calculated in non-overlapping 20 kbp genomic windows.

544 SNP density depicts the distribution of allele frequencies (site frequency spectrum) for each population,

545 scaled such that a population with constant size over time is expected to have a constant SNP density

546 over all allele frequencies. **c**, Average nucleotide diversity (π) and ratio of diversity between sex-linked

547 (X) and autosomal (A) chromosomes in relation to gene architecture. **d**, Relationship between number of

548 individuals sampled and the cumulative number of variant sites discovered (left panel), availability of

549 conserved Cas9 target sites within genes (center panel), and number of genes containing at least 1

550 conserved Cas9 target site which could thus be "targetable" for gene drive (right panel).

551 **Extended Data Figure 4. ADMIXTURE analysis. a**, Ancestry proportions within individual mosquitoes for

552 ADMIXTURE models from $K=2$ to $K=10$ ancestral populations. Each vertical bar represents the proportion

553 of ancestry within a single individual, with colours corresponding to ancestral populations. These data

554 are the average of the major q-matrix clusters derived by CLUMPAK analysis. **b**, Violin plot of cross-

555 validation error for each of 100 replicates for each K.

556 **Extended Data Figure 5. Population structure and differentiation. a**, Principal components analysis of

557 the 765 wild-caught mosquitoes. **b**, Average allele frequency differentiation (F_{ST}) between pairs of

558 populations. The lower left triangle shows average F_{ST} between each population pair. The upper right

559 triangle shows the Z score for each F_{ST} value estimated via a block-jackknife procedure. CM*=Cameroon

560 savanna sampling site only. **c**, Allele sharing in doubleton (f_2) variants. The height of the coloured bars

561 represent the probability of sharing a doubleton allele between two populations. Heights are normalized

562 row-wise for each population.

563 **Extended Data Figure 6. Ancestry informative markers (AIMs).** Rows represent individual mosquitoes

564 (grouped by population) and columns represent SNPs (grouped by chromosome arm). Colours represent

565 species genotype. The column at the far left shows the species assignment according to the

566   conventional molecular test based on a single marker on the X chromosome, which was performed for

567   all individuals except Kenya (KE). The column at the far right shows the genotype for *kdr* variants in *Vgsc*

568   codon 995. Lines at the lower edge show the physical locations of the AIM SNPs.

569   **Extended Data Figure 7. Population size history. a**, Stairway Plot of inferred histories for each

570   population. The shaded area shows the 95% confidence interval from 199 bootstrap replicates. **b**,

571   Inferred histories from ∂a∂i three epoch models. The thick line shows the history with the highest

572   likelihood found by optimization; thin lines show 100 histories with the highest likelihoods from even

573   sampling of the model parameter space. **c**, Inferred histories from ∂a∂i 2-population models allowing for

574   migration. For each population pair, solutions from 5 optimization runs with the highest likelihoods are

575   shown, with the thick line showing the history with the highest likelihood. In all panels, time and N_e are

576   scaled assuming 11 generations per year and a mutation rate of $\mu=3.5\times10^{-9}$. Scaling of time and N_e is

577   proportional to $1/\mu$, e.g., if the true mutation rate is twice as high then estimates of time and N_e would
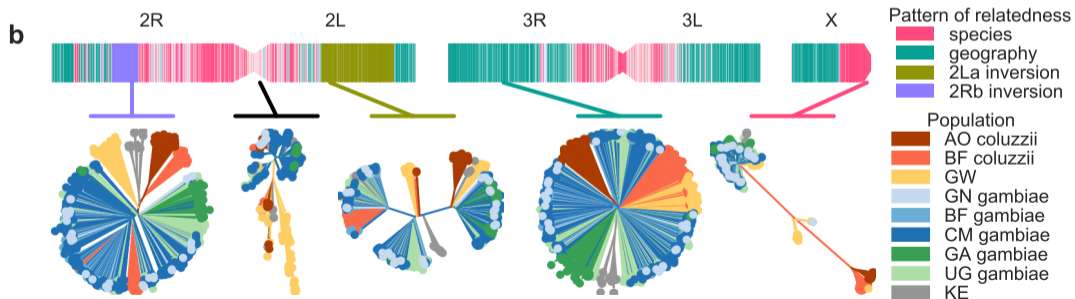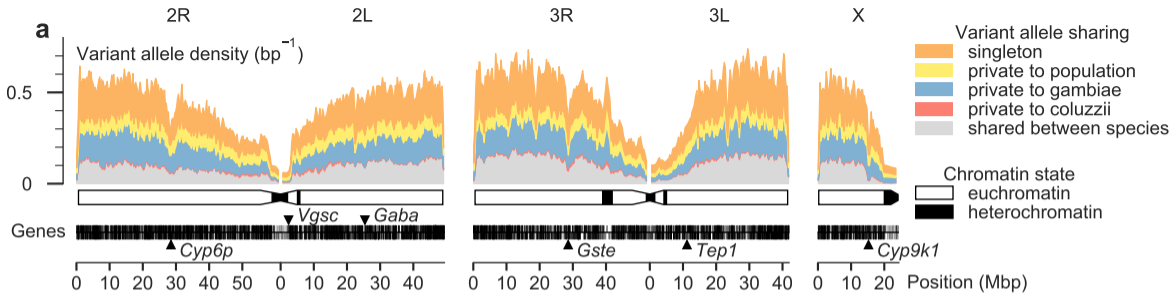
578   be halved.

579   **Extended Data Figure 8. Identity by descent (IBD) and recent effective population size history. a**,

580   Patterns of IBD sharing within populations. Each marker represents a pair of individuals. **b**, The

581   distribution of IBD tract lengths within populations. **c**, Recent population size history for the Kenyan

582   population inferred by IBDN_e. **d**, Comparison of the IBD tract length distribution between Kenya and four

583   simulated demographic scenarios. **e**, Population size histories inferred by IBDN_e (red dashed lines) from

584   data generated by simulations (black line shows the simulated population size history). **f**, Comparison of

585   patterns of IBD sharing generated by simulations (black contour lines) with Kenyan data (filled blue

586   contours). See Supplementary Text 8.4 for details of simulations.

587   **Extended Data Figure 9. Genome scans for signatures of recent selection. a**, Haplotype diversity. Each

588   track plots the H12 statistic in non-overlapping windows over the genome. A value of 1 indicates low

589   haplotype diversity within a window, expected if one or two haplotypes have risen to high frequency

590   due to recent selection. A value of 0 indicates high haplotype diversity, expected in neutral regions. **b**,

591   XP-EHH scans. For each population comparison (e.g., BF *gambiae* versus BF *coluzzii*), positive scores

592   indicate longer haplotypes and therefore recent selection in the first population (e.g., BF *gambiae*), and

593   negative scores indicate selection in the second population (e.g., BF *coluzzii*).

594   **Extended Data Figure 10. Haplotype structure at metabolic insecticide resistance loci.** Plot components

595   are as described for Fig. 4. For both loci, SNPs shown in the lower panel are all either non-synonymous

596   or splice site variants, and are associated with one or more haplotypes under selection. **a**, Haplotype

597   clustering using 1,375 SNPs within the region 3R:28,591,663-28,602,280 spanning 8 genes (*Gste1-*

598   *Gste8*). **b**, Haplotype clustering using 1,844 SNPs within the region 2R:28,491,415-28,502,910 spanning 5

599   genes (*Cyp6p1-Cyp6p5*).

600

a

Variant allele density (bp⁻¹)

2R 2L 3R 3L X

Variant allele sharing
- singleton
- private to population
- private to gambiae
- private to coluzzii
- shared between species

Chromatin state
- euchromatin
- heterochromatin

Genes

▲ Cyp6p   ▼ Vgsc   ▼ Gaba   ▲ Gste   ▲ Tep1   ▲ Cyp9k1

Position (Mbp)

b

2R 2L 3R 3L X

Pattern of relatedness
- species
- geography
- 2La inversion
- 2Rb inversion

Population
- AO coluzzii
- BF coluzzii
- GW
- GN gambiae
- BF gambiae
- CM gambiae
- GA gambiae
- UG gambiae
- KE

Labels across the top: AO coluzzii, BF coluzzii, GW, GN gambiae, BF gambiae, CM gambiae, UG gambiae, GA gambiae, KE

Map annotations:
- 0.018 (2.1)
- 0.031 (1.2)
- 0.001 (18.5)
- 0.007 (4.5)
- 0.026 (1.3)
- 0.000 (>20)
- 0.053 (2.0)
- 0.213 (0.2)
- 0.139 (0.7)

a

Time (kya)

Population size (scaled in units of $4\mu N_e$)

N_e (millions)

Time (scaled in units of $2\mu T \times 10^{-5}$)

- AO coluzzii
- BF coluzzii
- GW
- GN gambiae
- BF gambiae
- CM gambiae
- GA gambiae
- UG gambiae
- KE

b

ROH count

$F_{ROH(100\ kbp)}$