Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*

Supplemental Material

The *Anopheles gambiae* 1000 Genomes Consortium

26th August 2020

Supplemental methods

Population sampling

This section provides details of the collection sites newly sampled in Ag1000G phase 2. Mosquitoes from natural populations were collected by a number of previous studies investigating the ecology and epidemiology of malaria vectors in different locations. Throughout, we use species nomenclature following Coetzee et al. (2013); prior to this, *An. gambiae* was known as *An. gambiae sensu stricto* (S form) and *An. coluzzii* was known as *An. gambiae sensu stricto* (M form). Unless otherwise stated, the DNA extraction method used for the collections described below was Qiagen DNeasy Blood and Tissue Kit (Qiagen Science, MD, USA). Information pertaining to the collection of samples released as part of Ag1000G phase 1 can be found in the Supplementary Information of The *Anopheles gambiae* 1000 Genomes Consortium (2017).

Côte d'Ivoire: Tiassalé (5.898, -4.823) is located in the evergreen forest zone of southern Côte d'Ivoire. The primary agricultural activity is rice cultivation in irrigated fields. High malaria transmission occurs during the rainy seasons, between May and November. Samples were collected as larvae from irrigated rice fields by dipping between May and September 2012. All larvae were reared to adults and females preserved over silica for DNA extraction. Specimens from this site were all *An. coluzzii*, determined by PCR assay (Santolamazza et al. 2008).

Bioko: Collections were performed during the rainy season in September, 2002 by overnight CDC light traps in Sacriba of Bioko island (3.7, 8.7). Specimens were stored dry on silica gel before DNA extraction. Specimens contributed from this site were *An. gambiae* females, genotype determined by two assays (Scott et al. 1993; Santolamazza et al. 2004). All specimens had the $2L^{+a}/2L^{+a}$ karyotype as determined by the molecular PCR diagnostics (White et al. 2007). These mosquitoes represent a population that inhabited Bioko Island before a comprehensive malaria control intervention initiated in February 2004 (Sharp et al. 2007). After the intervention *An. gambiae* was declining, and more recently almost only *An. coluzzii* can be found (Overgaard et al. 2012).

Mayotte: Samples were collected as larvae during March-April 2011 in temporary pools by dipping, in Bouyouni (-12.738, 45.143), M'Tsamboro Forest Reserve (-12.703, 45.081), Combani (-12.779, 45.143), Mtsanga Charifou (-12.991, 45.156), Karihani Lake forest reserve (-12.797, 45.122), Mont Benara (-12.857, 45.155) and Sada (-12.852, 45.104) in Mayotte island. Larvae were stored in 80% ethanol prior to DNA extraction. All specimens contributed to Ag1000G phase 2 were *An. gambiae* (Santolamazza et al. 2004) with the standard $2L^{+a}/2L^{+a}$ or inverted $2L^{a}/2L^{a}$ karyotype as determined by the molecular PCR diagnostics (White et al. 2007). The samples were identified as males or females by the sequencing read coverage of the X chromosome using LookSeq (Manske and Kwiatkowski 2009).

The Gambia: Indoor resting female mosquitoes were collected by pyrethrum spray catch from four hamlets around Njabakunda (-15.90, 13.55), North Bank Region, The Gambia between August and October 2011. The four hamlets were Maria Samba Nyado, Sare Illo Buya, Kerr Birom Kardo, and Kerr Sama Kuma; all are within 1 km of each other. This is an area of unusually high rates of apparent hybridization between *An. gambiae s.s.* and *An. coluzzii* (Caputo et al. 2008; Nwakanma et al. 2013). Njabakunda village is approximately 30 km to the west of Farafenni town and 4 km away from the Gambia River. The vegetation is a mix of open savannah woodland and farmland.

Ghana: Mosquitoes were collected from Twifo Praso (5.609, -1.549), a peri-urban community located in semi-deciduous forest in the Central Region of Ghana. It is an extensive

agricultural area characterised by small-scale vegetable growing and large-scale commercial farms such as oil palm and cocoa plantations. Mosquito samples were collected as larvae from puddles near farms between September and October, 2012. Madina (5.668, -0.219) is a suburb of Accra within the coastal savanna zone of Ghana. It is an urban community characterised by numerous vegetable-growing areas. The vegetation consists of mainly grassland interspersed with dense short thickets often less than 5 m high with a few trees. Specimens were sampled from puddles near roadsides and farms between October and December 2012. Takoradi (4.912, -1.774) is the capital city of Western Region of Ghana. It is an urban community located in the coastal savanna zone. Mosquito samples were collected from puddles near road construction and farms between August and September 2012. Koforidua (6.094, -0.261) is the capital city of Eastern Region of Ghana and is located in semi-deciduous forest. It is an urban community characterized by numerous small-scale vegetable farms. Samples were collected from puddles near road construction and farms between August and September 2012. Larvae from all collection sites were reared to adults and females preserved over silica for DNA extraction. Both *An. gambiae* and *An. coluzzii* were collected from these sites, determined by PCR assay (Santolamazza et al. 2008).

Guinea-Bissau: Mosquitoes were collected in October 2010 using indoor CDC light traps, in the village of Safim (11.957, -15.649), ca. 11 km north of Bissau city, the capital of the country. Malaria is hyperendemic in the region and transmitted by members of the *Anopheles gambiae* complex (Vicente et al. 2017). *An. arabiensis*, *An. melas*, *An. coluzzii* and *An. gambiae*, as well as apparent hybrids between the latter two species, are known to occur in the region (Gordicho et al. 2014; Vicente et al. 2017). Mosquitoes were preserved individually on 0.5ml micro-tubes filled with silica gel and cotton. DNA extraction was performed by a phenol-chloroform protocol (Donnelly et al. 1999).

Genome accessibility

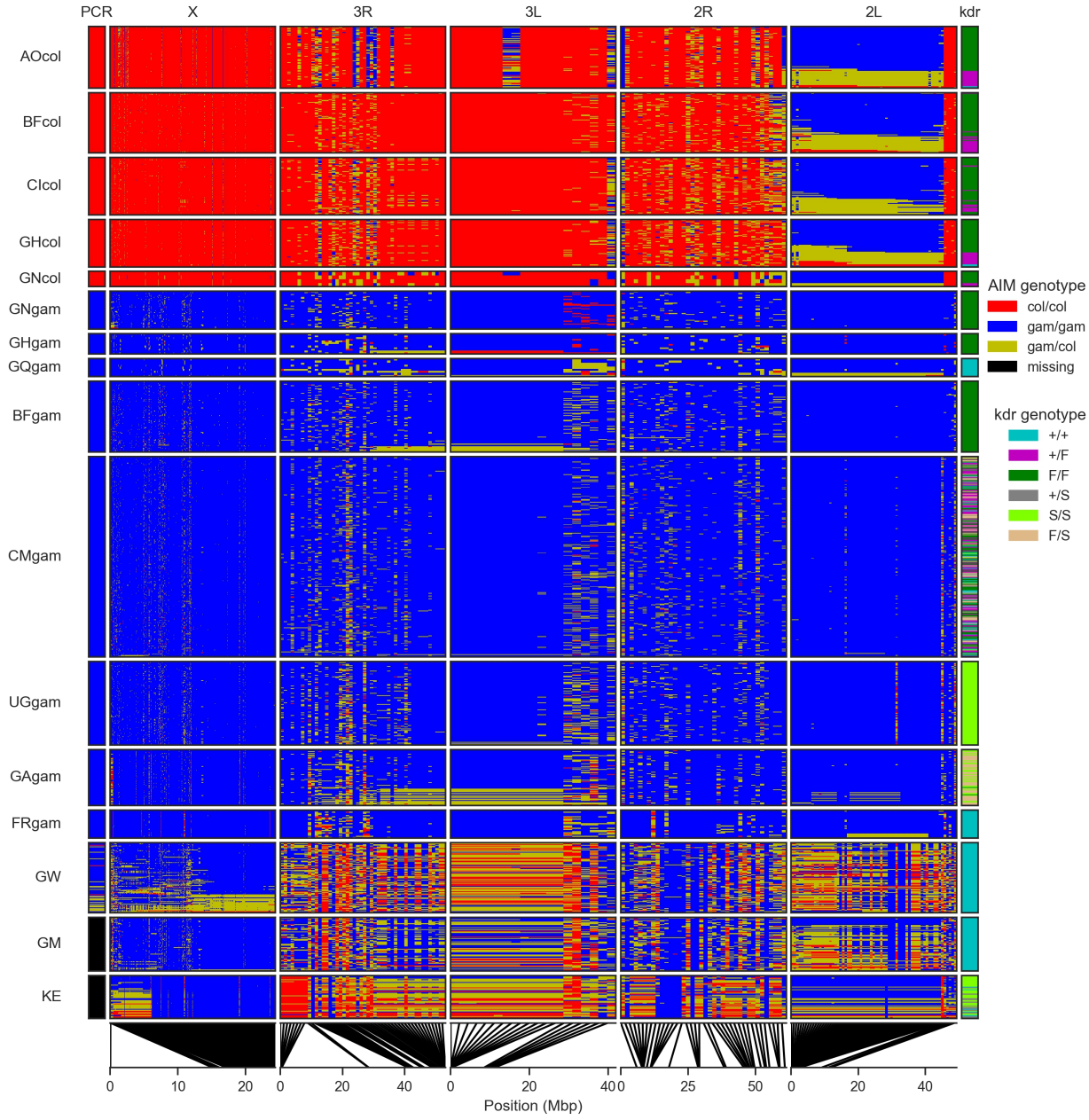
We performed additional analyses to verify that there was no significant bias towards one species or another given the use of a single reference genome AgamP3 (Holt et al. 2002) for alignment of reads from all individuals. We found that the genomes of *An. coluzzii* and *An. gambiae* individuals were similarly diverged from the reference genome (Supplemental Figure S9). The similarity in levels of divergence is likely to reflect the mixed ancestry of the PEST strain from which the reference genome was derived (Holt et al. 2002; Sharakhova et al. 2007). An exception to this was the pericentromeric region of the X chromosome, a known region of divergence between the two species (The *Anopheles gambiae* 1000 Genomes Consortium 2017) where the reference genome is closer to *An. coluzzii* than to *An. gambiae*. The similarity of this region to *An. coluzzii* may be due to artificial selection for the X-linked pink eye mutation in the reference strain (Holt et al. 2002), as this originated in the *An. coluzzii* parent it may have led to the removal of any *An. gambiae* ancestry in this region.

SNP annotation

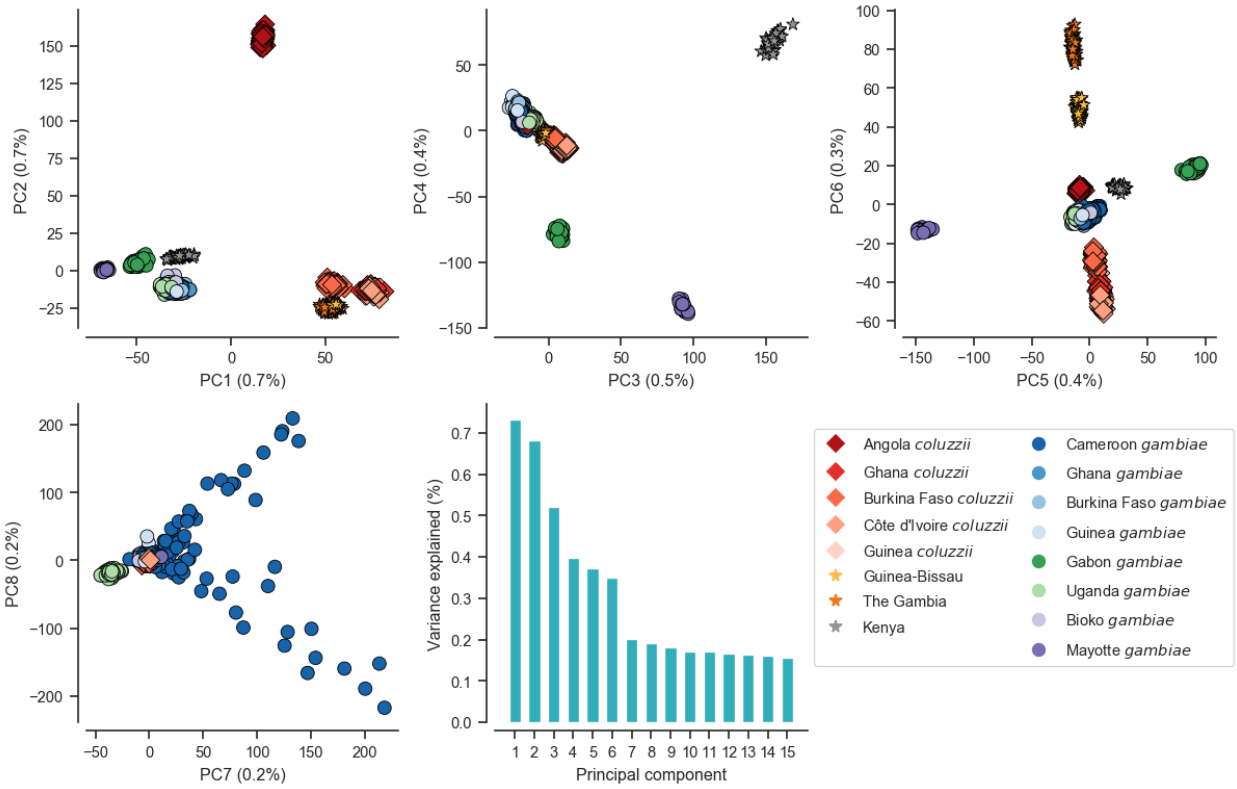
Of 105,486,698 SNPs reported in the raw callset, 57,837,885 passed all quality filters defined in the main Methods section. To produce an analysis-ready VCF file for each chromosome arm, we first removed all non-SNP variants. We then removed genotype calls for individuals excluded by the sample QC analysis, then removed any variants that were no longer variant after excluding individuals. We then added INFO annotations with genome accessibility metrics and added FILTER annotations per the criteria defined in

the main Methods section. Finally, we added INFO annotations with information about functional consequences of mutations using SnpEff version 4.1b (Cingolani et al. 2012). Further details of SNP filtering and annotation can be found in Supplementary Information of The *Anopheles gambiae* 1000 Genomes Consortium (2017).

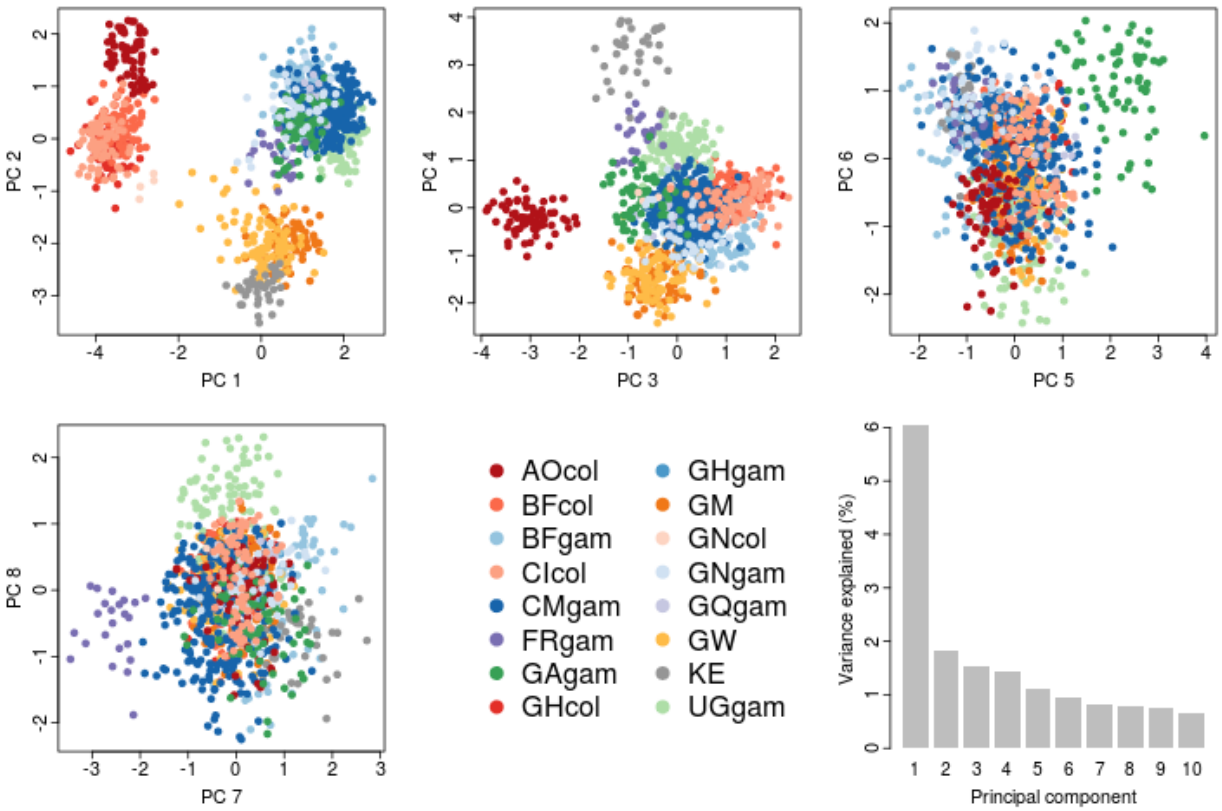
Supplemental figures



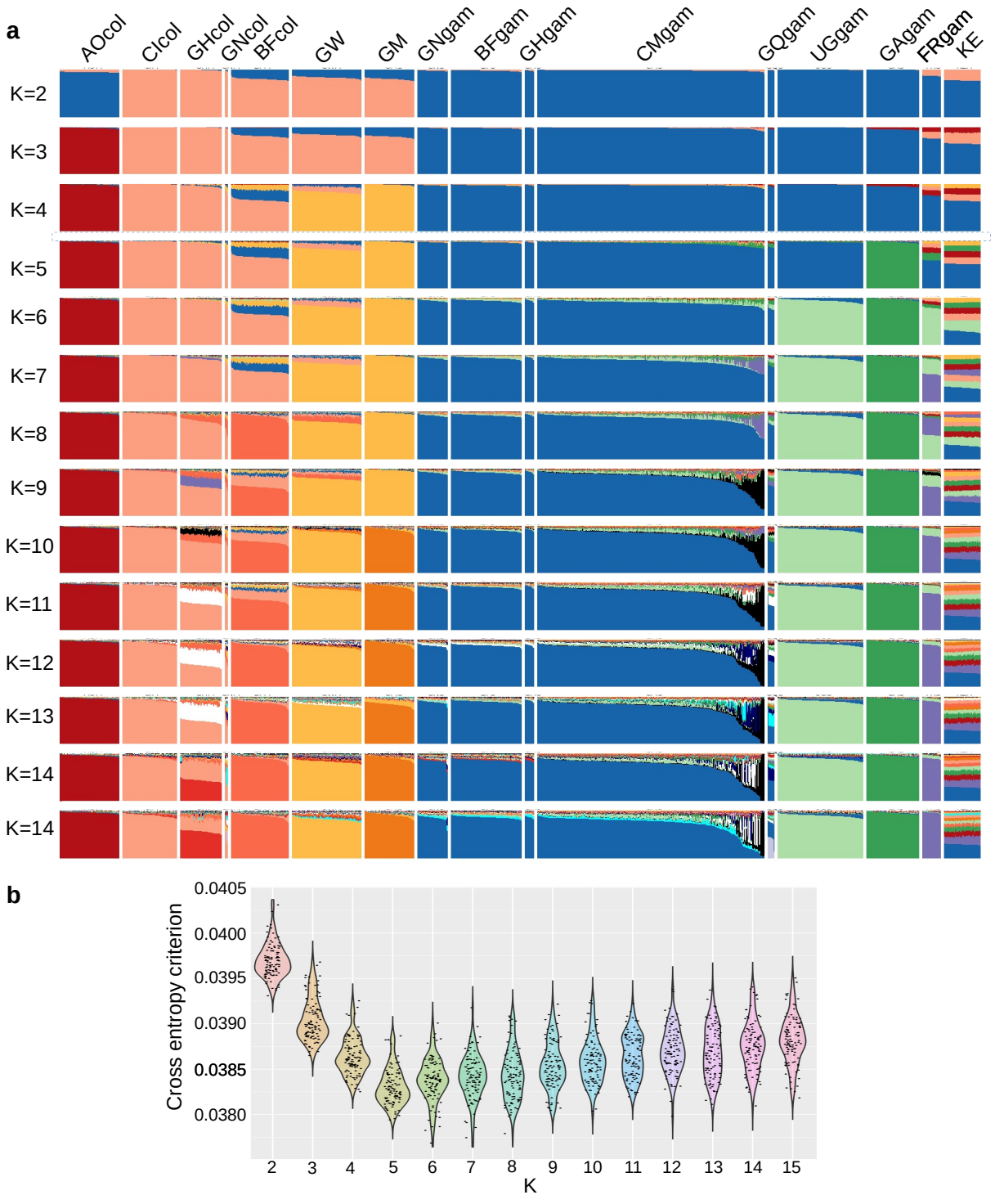
Supplemental Figure S1. Ancestry informative markers (AIM). Rows represent individual mosquitoes (grouped by population) and columns represent SNPs (grouped by chromosome arm). Colours represent species genotype. The column at the far left ("PCR") shows the species assignment according to the conventional molecular test based on a single marker on the X chromosome, which was performed for all populations except The Gambia (GM) and Kenya (KE). The column at the far right shows the genotype for *kdr* variants in *Vgsc* codon 995. Lines at the lower edge show the physical locations of the AIM SNPs.



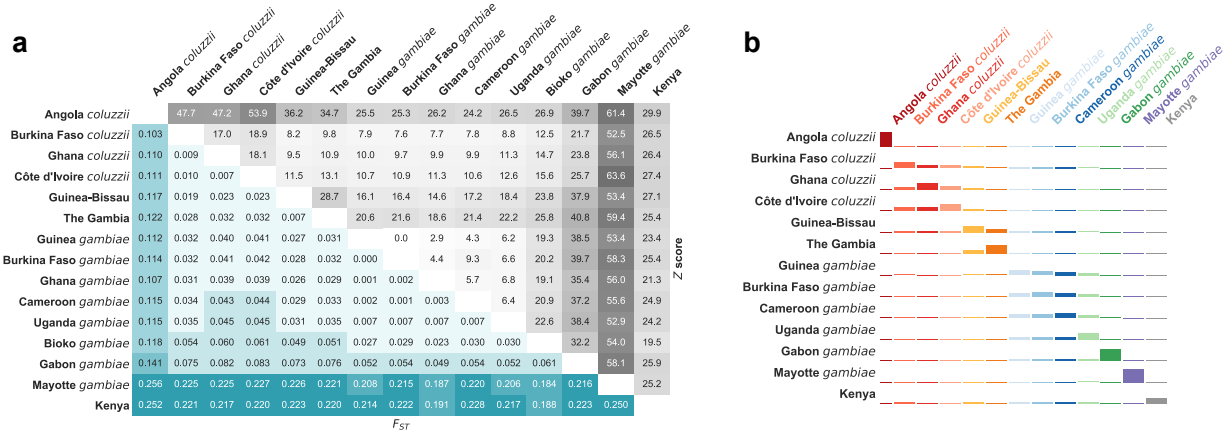
Supplemental Figure S2. Principal component analysis of the 1,142 wild-caught mosquitoes using biallelic SNPs from euchromatic regions of Chromosome 3. Scatter plots show relationships of principle components 1-8 where each marker represents an individual mosquito. Marker shape and colour denotes population. The bar chart shows the percentage of variance explained by each principal component.



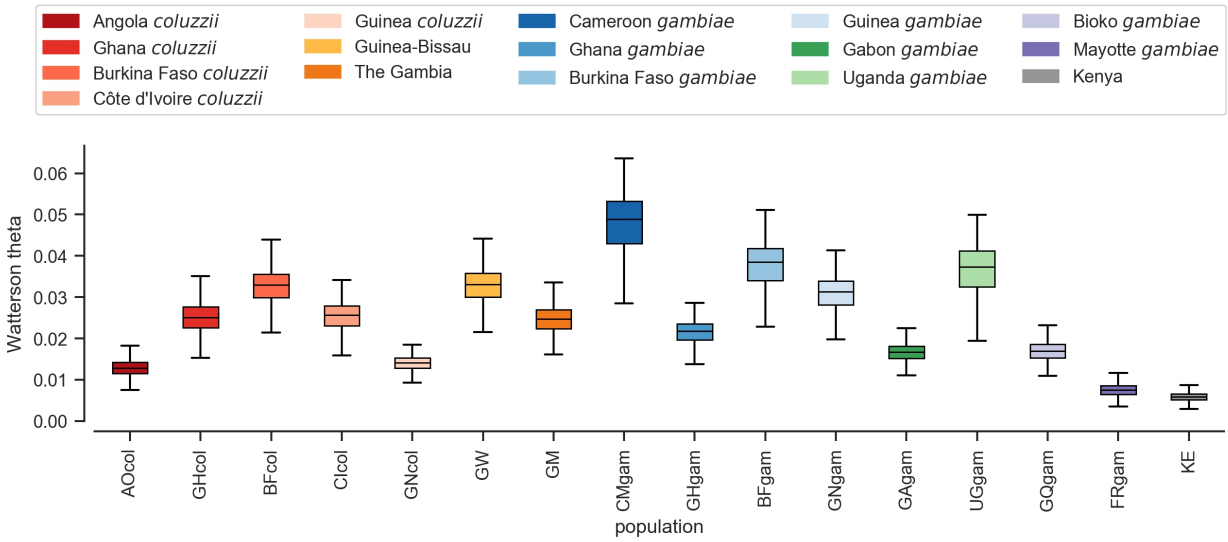
Supplemental Figure S3. Principal component analysis of the 1,142 wild-caught mosquitoes using copy number variant calls. Bar chart shows the percentage of variance explained by each component.



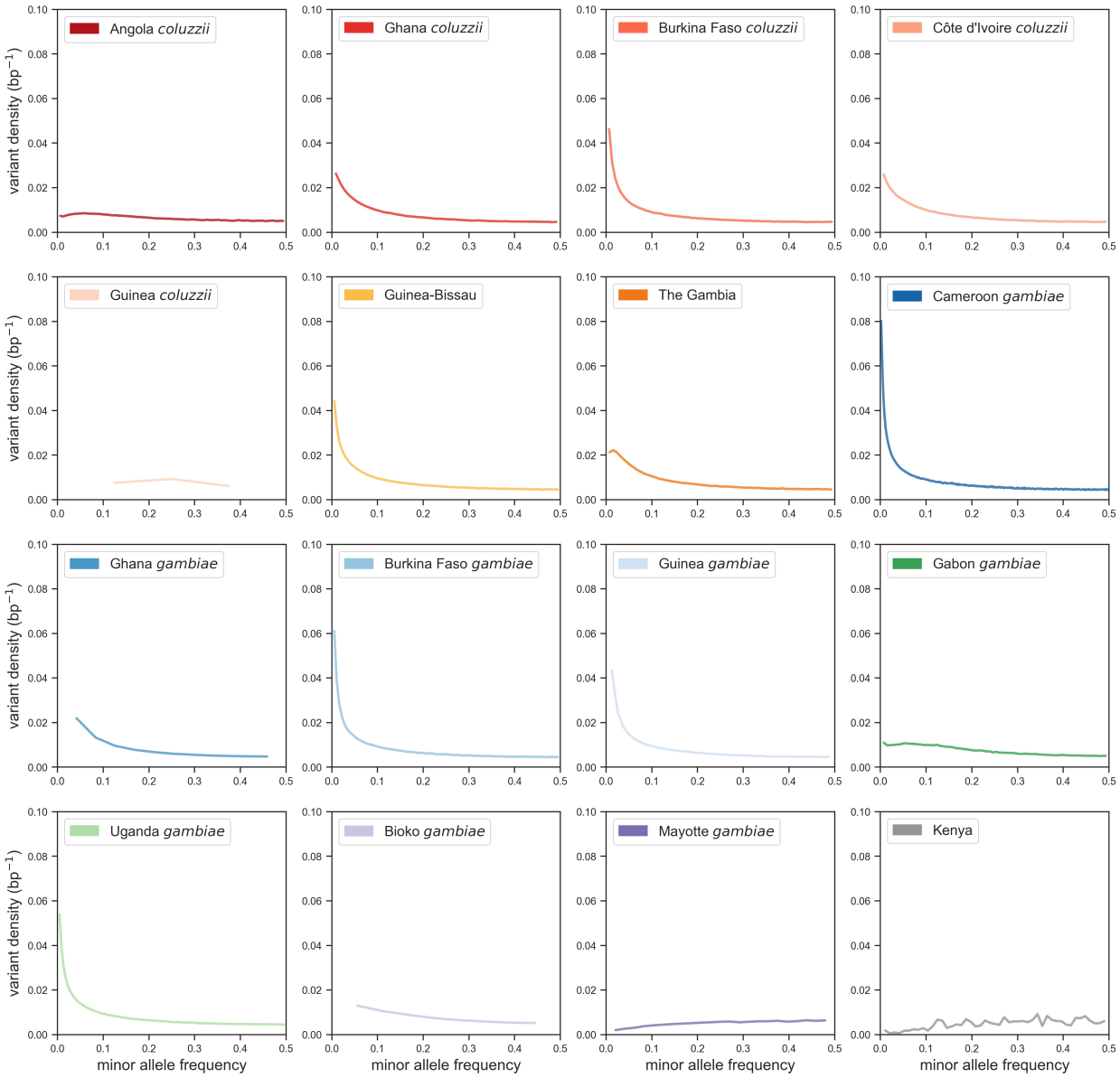
Supplemental Figure S4. Analysis of population structure and admixture. **(a)** Each row shows results of modelling ancestry in sampled individuals assuming a given number K of ancestral populations (Frichot and François 2015). Within each row, individual mosquitoes are represented as vertical bars, grouped according to sampling location and species, and coloured according to the proportion of the genome inherited from each ancestral population. **(b)** Cross-entropy criterion values obtained for each value of K ancestral populations, where lower values imply a better fit of the model to the data.



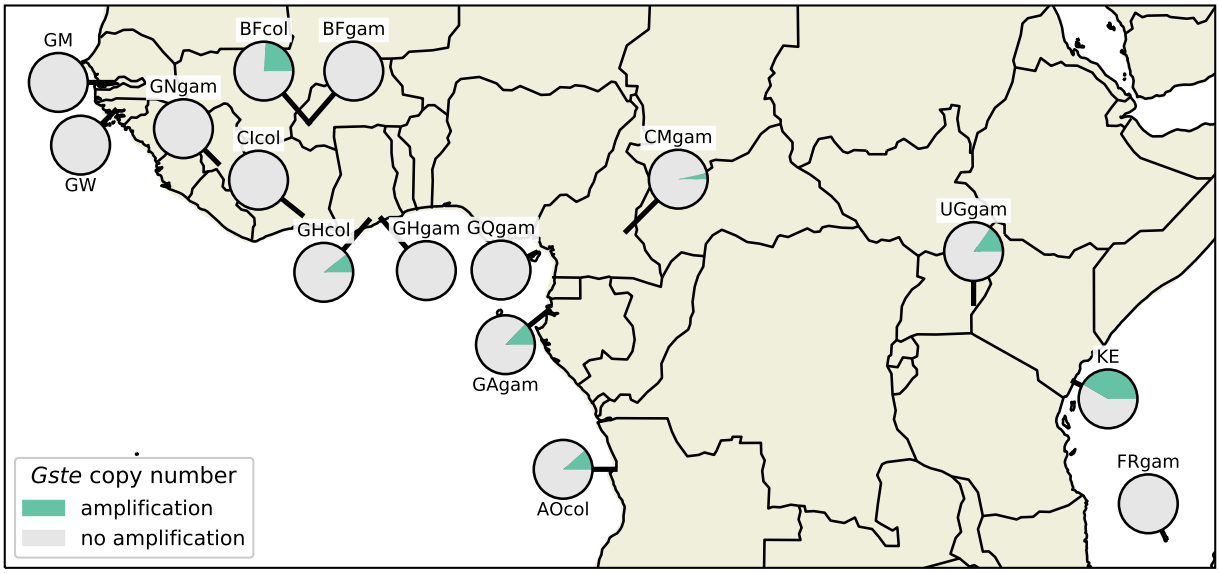
Supplemental Figure S5. Genetic differentiation between populations, computed using using biallelic SNPs from euchromatic regions of Chromosome 3. **(a)** Average allele frequency differentiation (F_{ST}) between pairs of populations. The bottom left triangle shows average F_{ST} values between each population pair. The top right triangle shows the Z score for each F_{ST} value estimated via a block-jackknife procedure. **(b)** Allele sharing in doubleton (f_2) variants. For each population, we identified the set of doubletons with at least one allele originating from an individual in that population. We then computed the fraction of those doubletons shared with each other population and the fraction shared only within itself. The height of the coloured bars represent the probability of sharing a doubleton allele between or within populations. Heights are normalized row-wise for each population so that the sum of coloured bars in each row equals 1.



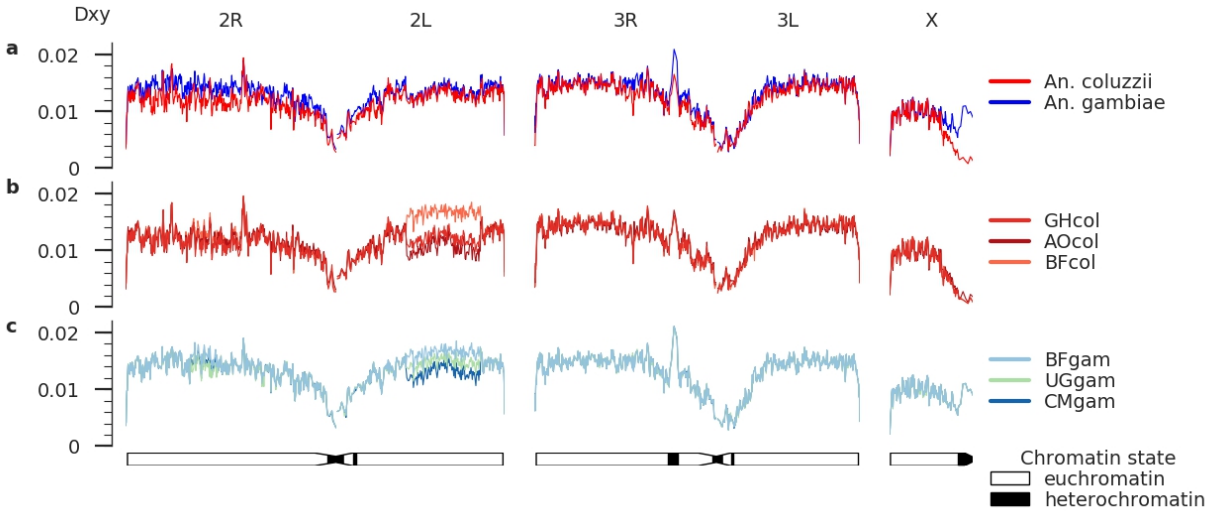
Supplemental Figure S6. Watterson's theta (θ_W), the density of segregating sites, calculated in non-overlapping 20 kbp genomic windows using SNPs from euchromatic regions of Chromosome 3.



Supplemental Figure S7. SNP density. Plots depict the distribution of allele frequencies (site frequency spectrum) for each population, scaled such that a population with constant size over time is expected to have a constant SNP density over all allele frequencies.



Supplemental Figure S8. Prevalence of copy number amplifications at the *Gste* locus. Each pie shows the frequency of individuals from a given population carrying an amplification spanning at least one gene in the *Gste* gene cluster. The Guinea *An. coluzzii* population is omitted due to small sample size.



Supplemental Figure S9. Divergence from the AgamP3 reference genome, calculated as *Dxy*, is largely similar for *An. coluzzii* and *An. gambiae*, with the exception of the centromere of the X chromosome (a). Comparing three populations of *An. coluzzii* (b) or *An. gambiae* (c) highlights the strong effect of the 2La chromosomal inversion on the accumulation of genetic variation.

Supplemental tables

Supplemental Table S1. Ag1000G phase 2 sampling locations.

Collection						Sample size		
Country	Location	Site	Year	Latitude	Longitude	Total	Female	Male
Angola	Luanda		2009	-8.821	13.291	78	78	0
Burkina Faso	Bana		2012	11.233	-4.472	60	40	20
	Pala		2012	11.150	-4.235	56	48	8
	Souroukoudinga		2012	11.235	-4.535	51	51	0
Cameroon	Daiguene		2009	4.777	13.844	96	81	15
	Gado Badzere		2009	5.747	14.442	73	58	15
	Mayos		2009	4.341	13.558	105	91	14
	Zembe Borongo		2009	5.747	14.442	23	23	0
Cote d'Ivoire	Tiassale		2012	5.898	-4.823	71	71	0
Equatorial Guinea	Bioko		2002	3.700	8.700	9	9	0
France	Mayotte	Bouyouni	2011	-12.738	45.142	1	1	0
		Combani	2011	-12.779	45.143	5	2	3
		Karihani Lake	2011	-12.797	45.122	3	3	0
		Mont Benara	2011	-12.857	45.155	2	1	1
		Mtsamboro Forest Reserve	2011	-12.703	45.081	1	1	0
		Mtsanga Charifou	2011	-12.991	45.156	8	3	5
		Sada	2011	-12.852	45.104	4	1	3
Gabon	Libreville		2000	0.384	9.455	69	69	0
Gambia, The	Njabakunda	Kerr Birom Kardo	2011	13.550	-15.900	19	19	0
		Kerr Sama Kuma	2011	13.550	-15.900	8	8	0
		Maria Samba Nyado	2011	13.550	-15.900	18	18	0
		Sare Illo Buya	2011	13.550	-15.900	20	20	0
Ghana	Koforidua		2012	6.094	-0.261	1	1	0
	Madina		2012	5.668	-0.219	24	24	0
	Takoradi		2012	4.912	-1.774	20	20	0
	Twifo Praso		2012	5.609	-1.549	22	22	0
Guinea	Koraboh		2012	9.250	-9.917	22	22	0
	Koundara		2012	8.500	-9.417	22	22	0
Guinea-Bissau	Antula		2010	11.891	-15.582	58	58	0
	Safim		2010	11.957	-15.649	33	33	0
Kenya	Kilifi	Junju	2012	-3.862	39.745	16	16	0
		Mbogolo	2012	-3.635	39.858	32	32	0
Uganda	Tororo	Nagongera	2012	0.770	34.026	112	112	0

Supplemental Table S2. Colony crosses.

Cross ID	Mother Colony	Father Colony	N progeny
18-5	Ghana	Kisumu/G3	20
29-2	Ghana	Kisumu	20
36-9	Ghana	Mali	20
37-3	Kisumu	Pimperena	20
42-4	Mali	Kisumu/Ghana	14
45-1	Mali	Kisumu	20
46-9	Pimperena	Mali	20
47-6	Mali	Kisumu	20
73-2	Akron	Ghana	19
78-2	Mali	Kisumu/Ghana	19
80-2	Kisumu	Akron	20

References

Caputo B, Nwakanma D, Jawara M, Adiamoh M, Dia I, Konate L, Petrarca V, Conway DJ and Torre A della. 2008. *Anopheles gambiae* complex along The Gambia river, with particular reference to the molecular forms of *An. gambiae* ss. *Malaria Journal*. **7**: 182.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. **6**: 80–92.

Coetzee M, Hunt RH, Wilkerson R, Della Torre A, Coulibaly MB and Besansky NJ. 2013. *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa*. **3619**: 246–274.

Donnelly M, Cuamba N, Charlwood J, Collins F and Townson H. 1999. Population structure in the malaria vector, *Anopheles arabiensis* Patton, in East Africa. *Heredity*. **83**: 408.

Frichot E and François O. 2015. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*.

Gordicho V, Vicente JL, Sousa CA, Caputo B, Pombi M, Dinis J, Seixas G, Palsson K, Weetman D, Rodrigues A et al. 2014. First report of an exophilic *Anopheles arabiensis* population in Bissau City, Guinea-Bissau: recent introduction or sampling bias? *Malaria Journal*. **13**: 423.

Holt RA et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*. **298**: 129–149.

Manske HM and Kwiatkowski DP. 2009. LookSeq: a browser-based viewer for deep sequencing data. *Genome Research*. **19**: 2125–2132.

Nwakanma DC, Neafsey DE, Jawara M, Adiamoh M, Lund E, Rodrigues A, Loua KM, Konate L, Sy N, Dia I et al. 2013. Breakdown in the process of incipient speciation in Anopheles gambiae. *Genetics*. **193**: 1221–1231.

Overgaard HJ, Reddy VP, Abaga S, Matias A, Reddy MR, Kulkarni V, Schwabe C, Segura L, Kleinschmidt I and Slotman MA. 2012. Malaria transmission after five years of vector control on Bioko Island, Equatorial Guinea. *Parasites & Vectors*. **5**: 253.

Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z and Torre A della. 2008. Insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms. *Malaria Journal*. **7**: 163.

Santolamazza F, Torre A della and Caccone A. 2004. A new polymerase chain reaction-restriction fragment length polymorphism method to identify *Anopheles arabiensis* from *An. gambiae* and its two molecular forms from degraded DNA templates or museum samples. *The American Journal of Tropical Medicine and Hygiene*. **70**: 604–606.

Scott JA, Brogdon WG and Collins FH. 1993. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *The American Journal of Tropical Medicine and Hygiene*. **49**: 520–529.

Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, Bruggner RV, Birney E and Collins FH. 2007. Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biology*. **8**: R5.

Sharp BL, Ridl FC, Govender D, Kuklinski J and Kleinschmidt I. 2007. Malaria vector control by indoor residual insecticide spraying on the tropical island of Bioko, Equatorial Guinea. *Malaria Journal*. **6**: 52.

The *Anopheles gambiae* 1000 Genomes Consortium. 2017. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature*. **552**: 96.

Vicente JL, Clarkson CS, Caputo B, Gomes B, Pombi M, Sousa CA, Antao T, Dinis J, Bottà G, Mancini E et al. 2017. Massive introgression drives species radiation at the range limit of *Anopheles gambiae*. *Scientific Reports*. **7**: 46451.

White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, Mouline K, Brengues C, Guelbeogo W, Coulibaly M, Kayondo JK et al. 2007. Molecular karyotyping of the

2La inversion in *Anopheles gambiae*. *The American Journal of Tropical Medicine and Hygiene*. **76**: 334–339.