Supplementary Materials for

Extensive Introgression In A Malaria Vector Species Complex Revealed By Phylogenomics

Michael C. Fontaine, James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang, Andrew B. Hall, Flaminia Catteruccia, Evdoxia Kakani, Sara N. Mitchell, Yi-Chieh Wu, Hilary A. Smith, R. Rebecca Love, Mara K. Lawniczak, Michel A. Slotman, Scott J. Emrich, Matthew W. Hahn*, Nora J. Besansky*

*correspondence to: mwh@indiana.edu (M.W.H.); nbesansk@nd.edu (N.J.B.)

This PDF file includes:

Supplementary Online Materials

Table of Contents

S1. Reference assembly alignments

Genome assemblies of six members of the *Anopheles gambiae* species complex and two outgroup *Pyretophorus* species (Table S1) were retrieved from VectorBase, www.vectorbase.org. Other relatively rare and narrowly distributed species in the complex have not been colonized (*Anopheles amharicus*, *Anopheles bwambae*) and were not sequenced due to the lack of suitable DNA template available at the time. In addition to the species sequenced as part of the *Anopheles* 16 genomes project (*52*), the genome assemblies of *An. gambiae* PEST (*19*), *An. gambiae* Pimperena S (*20*), and *An. coluzzii* (*20*) (formerly *An. gambiae* M molecular form) were employed. Before computing the multiple whole genome alignments, repetitive regions of the input genome assemblies were first masked to reduce the total number of potential genomic anchors formed by the many matches that occur among regions of repetitive DNA. Assemblies were analyzed using RepeatModeler (*53*) to build libraries of repetitive elements that were then combined and compared with known repeats from *An. gambiae* (from VectorBase). The combined library made up of repeats from all species was filtered to remove matches to known protein-coding repetitive sequences. Each genome assembly was subsequently masked with RepeatMasker (*54*).

A similar whole genome alignment strategy was employed to that used for other multi-species whole genome alignments such as the 12 *Drosophila* (*55*) and 29 mammal (*56*) genomes. Multiple whole genome alignments of the 9 *Anopheles* assemblies were built using the MULTIZ feature of the Threaded-Blockset Aligner (TBA) suite of tools (*57*). The progressive alignment approach of MULTIZ requires an input dendrogram of the expected relationships between the species so that the closest pairs are aligned first

followed by progressively stepping along the phylogeny to the most distant clades. While the position of the outgroup species was clear, the ordering of the species within the complex was initially unknown. Thus, simple ladder-ordered topologies were selected, and alternative orders made little difference to the overall results (data not shown). The topology arbitrarily chosen for the alignment employed for analyses was: (((((((AgamP3,AgamS1),AgamM1),AmerM1),AaraD1),AquaS1),AmelC1),AchrA1),AepiE1). The first step of the MULTIZ approach consists of running all-against-all pairwise LASTZ alignments, followed by a projection to ensure that each pair of species is "single-coverage", *i.e*. regions of both species may only be present once. Following the TBA alignment strategy, subsequent projection steps are then performed as guided by the species dendrogram to progressively combine the single-coverage pairwise alignments, and then the multiple alignments, until they encompass the complete dendrogram of all assemblies. Statistics of alignment coverage (Table S1) and density (Fig. S1) were computed using custom Perl scripts. The resulting TBA alignment (available in DRYAD, doi:10.5061/dryad.f4114) was the basis for all of our phylogenomic analyses based directly on the reference genome assemblies and the whole genome resequencing of single individuals per species from natural populations (S2.1).

For practical reasons, additional analyses using whole genome sequences from population samples of multiple individuals per species (S2.2) were based on a very similar but earlier whole genome alignment generated by ROAST (reference dependent multiple alignment tool). In this strategy the pairwise LASTZ alignments are projected to ensure that the reference species is "single-coverage" (unlike TBA where both species must be "single-coverage"), *i.e*. in any pairwise alignment, regions of the reference

species may only be present once. Subsequent projection steps are then performed as for TBA - guided by the species dendrogram to progressively combine alignments until they encompass the complete dendrogram.

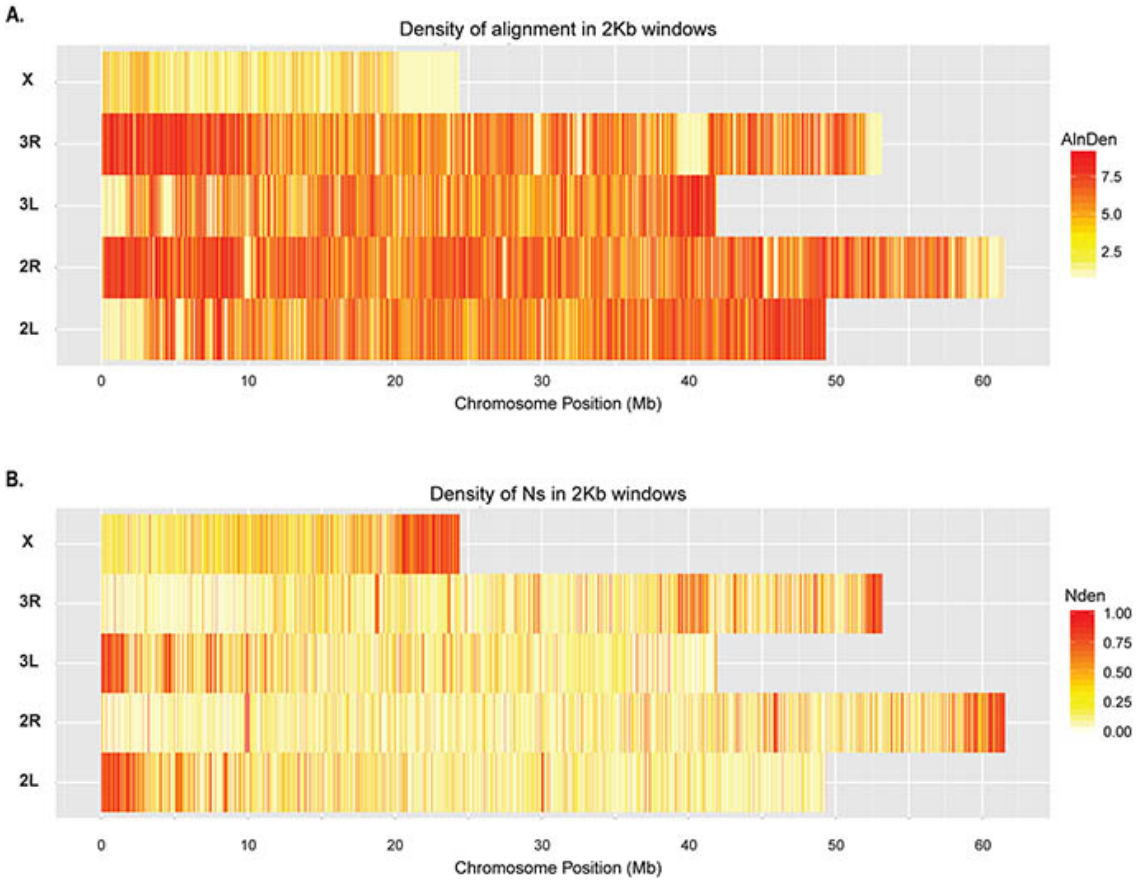


Fig. S1. Whole genome TBA alignment densities.
(A) Density of the whole genome alignment in 2 kb windows along chromosomes 2, 3, and X. Alignment density (AlnDen) ranges from 1 (not aligned to any other assembly) to 9 (aligned to all other assemblies). (B) Density of Ns (NDen, gaps in the assembly or masked regions) in 2 kb windows along chromosomes 2, 3, and X of *An. gambiae* PEST assembly. Density ranges from 0.00 (no Ns) to 1.00 (all Ns). Regions with the lowest alignment density (A) correspond to regions with the highest density of Ns (B).

Table S1. Whole genome TBA alignment statistics.
Statistics of the alignment of nine *Anopheles* genomes showing percentage aligned overall and percentage aligned to the *An. gambiae* PEST assembly (AgamP3) for all base-pairs and for non-gap and non-masked base-pairs.

Assembly	AgamP3	AgamS1	AgamM1	AaraD1	AquaS1	AmerM1	AmelC1	AchrA1	AepiE1
Assembly total (bp)	273,093,681	236,403,076	224,455,335	246,567,867	283,828,998	251,805,912	227,407,517	172,658,580	223,486,714
Aligned total (bp)	118,653,735	117,538,135	117,370,515	117,268,219	117,025,650	116,391,161	116,310,762	106,604,092	104,792,338
% total aligned	43.45	49.72	52.29	47.56	41.23	46.22	51.15	61.74	46.89
% AgamP3 total aligned	43.45	43.04	42.98	42.94	42.85	42.62	42.59	39.04	38.37
Gaps (bp)	20,654,948	8,362,861	14,926,268	35,124,750	74,862,329	33,613,976	20,677,584	2,671,395	20,854,535
Masked (bp)	55,247,274	41,205,745	33,114,661	30,887,257	30,062,727	36,451,174	26,733,654	11,802,245	21,157,887
Assembly non-N[§] (bp)	197,191,459	186,834,470	176,414,406	180,555,860	178,903,942	181,740,762	179,996,279	158,184,940	181,474,292
Aligned non-N[§] (bp)	118,819,805	117,399,459	117,280,089	117,158,705	116,901,945	116,280,288	116,156,111	106,510,444	104,704,559
% non-N[§] aligned	60.26	62.84	66.48	64.89	65.34	63.98	64.53	67.33	57.70
% AgamP3 non-N[§] aligned	60.26	59.54	59.48	59.41	59.28	58.97	58.91	54.01	53.10

[§]non-N: non-gap and non-masked base pairs; *An. gambiae*: AgamP3, AgamS1; *An. coluzzii*: AgamM1; *An. arabiensis*: AaraD1; *An. quadriannulatus*: AquaS1; *An. merus*: AmerM1; *An. melas*: AmelC1; *An. christyi*: AchrA1; *An. epiroticus*: AepiE1

S2. Whole genome re-sequencing from natural populations

S2.1. Single individual per species sequenced at high depth

One individual from each of six species of the *An. gambiae* complex (AGC) included in this study was sampled from field populations (Table S2) and sequenced at high depth (Table S3) for validation of the results obtained from the colony-based reference assemblies.

S2.1.1. Sample processing, sequencing

DNA extraction. Genomic DNA was extracted from whole individual female mosquitoes using a CTAB DNA extraction protocol (*58*). Species identification was ascertained from rDNA-based PCR diagnostic assays (*59-61*). For *An. gambiae* and *An. coluzzii* whose populations are polymorphic for the 2La inversion, the karyotype was determined molecularly using a PCR diagnostic assay (*62*).

Library construction and Sequencing. All samples were sequenced on an Illumina sequencing platform (HiSeq 2000 or 2500) with data production at two different sequencing centers, BGI at the University of California, Davis, or the Broad Institute of MIT and Harvard.

For samples sequenced at BGI (*An. gambiae, An. coluzzii,* and *An. arabiensis*), genomic fragment libraries were constructed in the laboratory of NJB, using the Illumina TruSeq® DNA Sample Preparation kit (Illumina) and 300 ng DNA per sample following the manufacturer's protocol. Final mean fragment library sizes were ~500 base-pairs (bp), corresponding to an insert size of ~ 340 bp. Paired-end 101 bp whole genome sequencing was performed using an Illumina HiSeq 2000 with one sample per lane.

For samples sequenced at the Broad Institute (*An. quadriannulatus, An. merus* and *An. melas*), genomic fragment libraries of 200 bp inserts were prepared. For each fragment library, 100 ng of genomic DNA was sheared to ~250 bp using a Covaris LE instrument and prepared for sequencing as previously described (*63*). Sequencing was performed with an Illumina HiSeq2500 platform with v3 chemistry and a 2 x 101 bp run configuration, and the indexed samples were pooled and sequenced across a total of four lanes. All sequencing datasets were processed through the Broad Institute's Picard sequencing analysis pipeline to demultiplex reads, generate standard sequencing metrics (*e.g*. read counts) and mark duplicate reads.

S2.1.2. Variant calling and conversion of genomic coordinates

The pipeline used for variant calling was based on the Broad Institute's best practices for GATK (*64*). Details on specific command line parameters for all steps can be found here: https://bitbucket.org/steelea/16genometoolkit/wiki/Pipeline

(a) Illumina Read Quality Control. Short reads from each individual sample were first analyzed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) to detect poor quality and overrepresented sequences in the raw sequencing data. If FastQC identified a bias at a given location in all of the reads, special care was taken to remove the bias by removing the lower quality regions or by hard-clipping all reads in the problematic region. Otherwise, Trimmomatic (*65*) was used to remove low quality regions along with any lingering Illumina adapters using the "IlluminaClip" option. Low quality regions were identified at the edges of the read if bases at the beginning or end had a quality less than 5. Subsequently, any 4-base sections of the reads that had an

average quality less than 15 were also removed. Only mated reads that passed these trimming processes and were at least 50 bases long were used for the remaining analysis, *i.e*., all singleton reads post-trimming were removed. (Example of the Trimmomatic parameters: `ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:50`)

(b) Read mapping. We used BWA 0.5.9-r16 (*66*) to map the paired-end reads for all species to a reference assembly (Fig. S2). Reads from *An. gambiae* and *An. coluzzii* were mapped to the *An. gambiae* PEST (AgamP3) reference assembly (*19, 67*). For the other species, reads were mapped both to the *An. gambiae* PEST (AgamP3) reference assembly and to the conspecific reference assembly produced in the framework of the 16 *Anopheles* genomes project (*18, 52*). The specific parameters used with the BWA *aln* algorithm (-o 1, -q 5, -l 32, -k 2) permitted only a single open gap, and a minimum base-quality of 5 for each base mapped. For performance reasons, the first 32 bases of a read were used as a seed with a maximum edit distance of two in the seed. These more stringent parameters minimize the number of gaps and mismatches that exist in the final alignment.

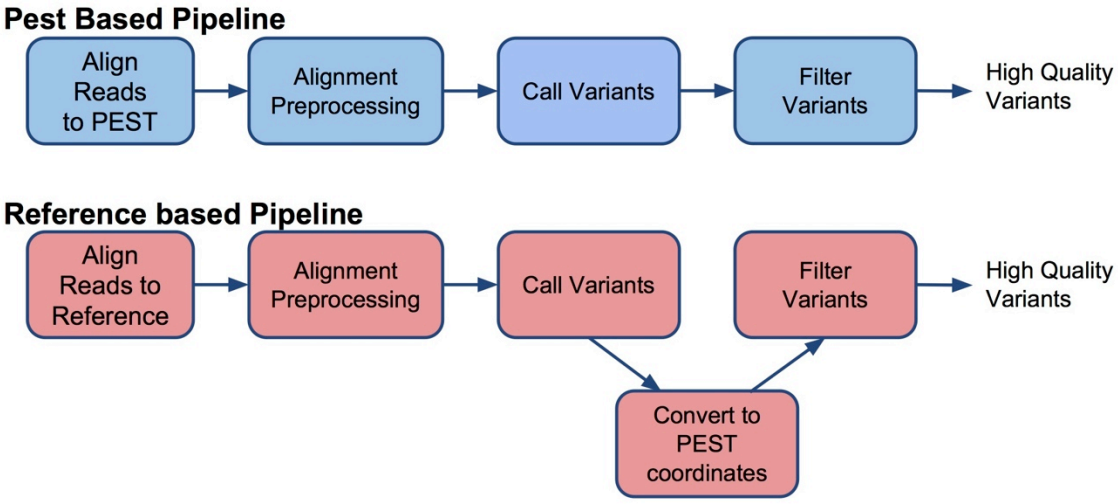


Fig. S2. Diagram of the two methods used for mapping reads and calling variants. The first method used the *An. gambiae* PEST genome as a reference to map reads and call variants. The alternative method used the conspecific reference assembly to map reads and call variants for sequences derived from species other than *An. gambiae* and *An. coluzzii*. To maintain consistency and allow interspecific comparisons, genomic coordinates of variants called using Method 2 were converted to PEST reference assembly coordinates prior to filtering (see S2.1.2f).

(c) Preprocessing. Further processing was performed on paired reads to ensure high quality alignments and compatibility with downstream tools. First, Picard Tools' CleanSam.jar (http://picard.sourceforge.net) was used to soft clip any reads where part of the read extended beyond the end of the sequence it was aligned to. Next, Picard Tools' SortSam.jar (http://picard.sourceforge.net) was used to sort reads based on alignment location, to facilitate compatibility with the Genome Analysis Toolkit (GATK). Third, Picard Tools' MarkDuplicates.jar (http://picard.sourceforge.net) was used to locate and mark duplicate molecules prior to SNP calling. Marked duplicates were retained in the dataset for consistency. Finally, reads were locally realigned with GATK's Indel Realignment tool (*64, 68*), which performs de-novo realignment in regions where a suspected indel exists. Combined, these additional preprocessing steps reduce the effects of indels on close-proximity SNPs.

(d) Variant Calling. For calling variants, we used both the UnifiedGenotyper and HaplotypeCaller tools that are packaged in GATK v2.8 and 3.1 (*64, 68*). For both variant callers the default parameters were used. Because these samples were of moderate coverage (Table S3), no downsampling was performed when calling variants. Additionally, calls were emitted for all sites (variant or otherwise) in the case where reads were mapped to the conspecific reference assembly to facilitate conversion from one coordinate system to another (see step f).

(e) Variant Annotation and filtering. To aid in quality assessment of variants, we added annotations (mapping quality, quality by depth, allelic balance, distance from a homopolymer run, and genotype quality) to the VCF file using GATK's VariantAnnotator. We applied hard filters on the variants to keep only those that were of

highest quality. The hard filters included a minimum depth of 10x, sites with quality score Q ≥ 30, a quality by depth QD ≥ 5, and allelic balance for heterozygote sites AB between 0.2 and 0.8.

(f) Coordinate Conversion. The second method of calling variants (Fig. S2) used the conspecific reference assemblies for sequences derived from species other than *An. gambiae* and *An. coluzzii*. To maintain consistency and to allow comparisons to Method 1, Method 2 biallelic SNPs were converted to *An. gambiae* PEST (AgamP3) coordinates prior to filtering. This was done using the whole genome alignments (S1), which provided conversion information, and base variant call format (VCF) files (2.1.2d). Specifically, the VCF file for each focal species was first filtered such that indels and multiallelic SNPs were removed for the sake of simplicity. Next, the specific alignment region of a scaffold *X* in a non-PEST genome to the PEST chromosome reference imposed a simple common coordinate system once indels were ignored. As a concrete example, if a position *a* on scaffold *X* aligns to a non-gap position *b* on arm *W* in PEST, this specific line of the VCF file (*a* on scaffold *X*) is converted to position *b* (arm *W*). We developed custom Java code (https://bitbucket.org/steelea/vcfmap) to implement this conversion. Although only biallelic SNPs were used as input, it is possible to have three alleles in the converted VCF if the PEST allele is different. By emitting all positions (see 2.1.2d) we were able to uncover invariant positions in the target species that differed from the reference (*i.e.*, fixed differences). See also SOM Text S3.1.

(g) Comparison of HaplotypeCaller and UnifiedGenotyper caller.

As recommended by GATK best practice (*64*), and unless stated otherwise, we used SNP calls from HaplotypeCaller (HC). However, we took the opportunity to compare SNP

calls from HC to those called from UnifiedGenotyper (UG), basing the comparison on SNPs that had a quality score Q ≥ 30, a map quality ≥ 30, and a sequencing depth (DP) ¼ (mode depth) ≥ DP ≤ 2X (mode depth).

The total number of high quality SNPs discovered (Fig. S3) was always higher from UG than HC. However, HC produced SNP calls with an overall higher genotype quality as measured by the genotype quality (GQ) and likelihood (PL) statistics (result not shown).

S2.1.3. Improvement of data quality when using a conspecific reference assembly instead of PEST for read mapping and SNP calling.

We investigated whether using the new reference assemblies of each focal species (other than *An. gambiae* or *coluzzii*) significantly improved the data quality relative to using a single reference (*An. gambiae* PEST, AgamP3) for all species irrespective of their genetic divergence from PEST.

Improvement in read mapping quality.

Short reads from *An. quadriannulatus*, *An. merus* and *An. melas* samples mapped considerably better to their own reference assemblies as compared to the *An. gambiae* PEST reference (Fig. S4 and Table S3). In contrast to those three species, short reads from *An. arabiensis* mapped reasonably well to the PEST reference (Fig. S4B), yet the coverage and the number of bases mapped improved if *An. arabiensis* reads were mapped to its own reference (Fig. S5A). Indeed, when reads from any species outside of the *An. gambiae* clade (*An. gambiae* and *An. coluzzii*) were mapped to the conspecific reference as opposed to PEST, the increase in average depth coverage was substantial: 17% in *An. arabiensis*, 16% in *An. merus,* 11% in *An. melas*, and 5% in *An. quadriannulatus*. The

proportion of sites with ≥ 10X coverage also increased measurably: by 4% in *An. arabiensis*, 3% in *An. quadriannulatus*, and 32% in *An. merus* and *An. melas* (Table S3). The mapping quality (MQ) also increased by a factor of 1.2 in *An. arabiensis* and *An. quadriannulatus*, and 1.5 in *An. merus* and *An. melas* (Table S3).

Improvement in SNP quality

The quality of the SNP calls also improved when using a species-specific reference, as reflected by the depth coverage (DP), mapping quality (MQ) and genotype quality (GQ) statistics (Fig. S5). Choice of reference had an especially marked impact on the results for the X chromosome.

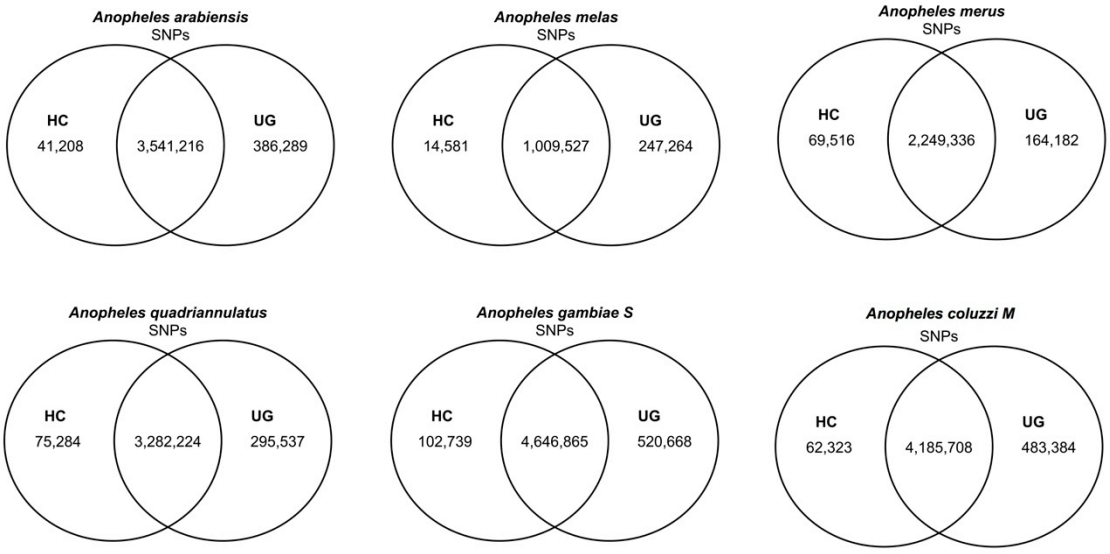


Fig. S3.
Venn diagram showing the total number of high quality SNPs discovered in individual high-depth genomic sequences from each species using the HaplotypeCaller (HC) or the UnifiedGenotyper (UG) algorithms.

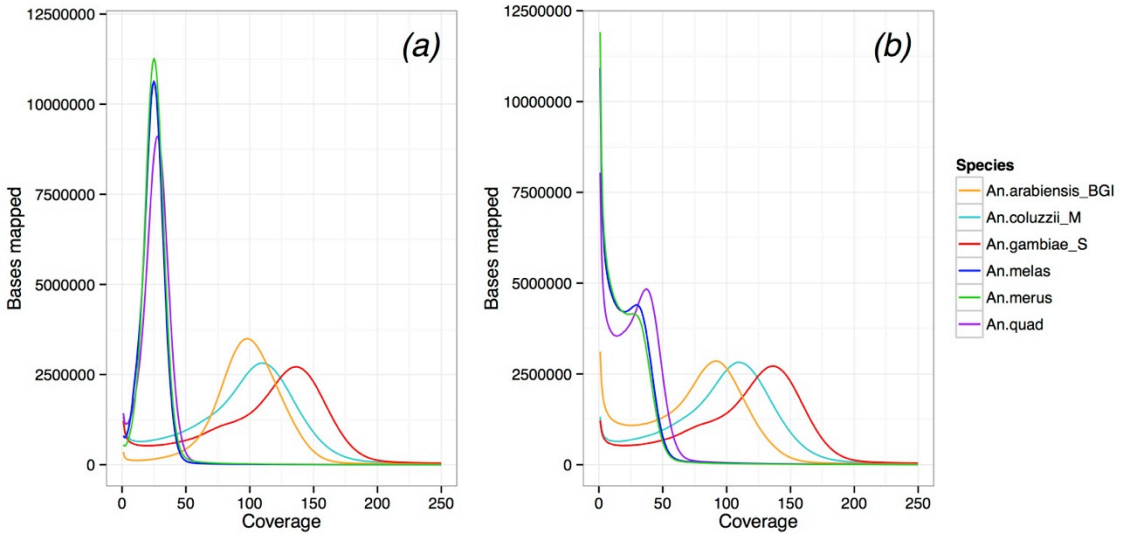


Fig. S4.
Bases mapped as a function of sequencing depth when short reads are mapped to **(A)** the species reference assemblies or **(B)** the *An. gambiae* PEST reference genome

Table S2.
Single individuals per species sequenced at high depth.

Species	ID	Country	Village	Coordinates	Year	2La karyotype	BioSample	BioProjects
An. gambiae	40.2	Burkina Faso	Pala	11°09'N,04°14'W	2012	a/+	SAMN02899205	PRJNA254046
An. coluzzii	C27.2	Burkina Faso	Bana	11°14'N,04°28'W	2012	a	SAMN02899195	PRJNA254046
An. arabiensis	4080	Burkina Faso	Monomtenga	12°06'N,01°17'W	2004	a	SAMN03083367	PRJNA262489
An. quad[*]	72	Zimbabwe	Chilongo, Chiredzi	21°03'S,31°40'E	1986	+	SAMN01760635	PRJNA177000
An. merus	Mpug686i	South Africa	Mpumalanga, Koomatipoort	25°26'S,31°57'E	1988	a	SAMN01760628	PRJNA176993
An. melas	CM1001067	Cameroon	Campo	02°22'N,0949'E	2010	+	SAMN01760621	PRJNA176986

[*] *An. quad : An. quadriannulatus.*

Table S3.
Statistics for read mapping to the conspecific reference or the An. gambiae PEST reference, calculated using QUALIMAP (*69*).

Species	***An. gambiae "S"***	***An. coluzzii*** "M"	***An. arabiensis***	***An. arabiensis* (PEST)**	***An. quad****	***An. quad**** **(PEST)**	***An. merus***	***An. merus* (PEST)**	***An. melas***	***An. melas* (PEST)**
Reference size	273,093,681	273,093,681	246,567,867	273,093,681	283,828,998	273,093,681	251,805,912	273,093,681	227,407,517	273,093,681
# Reads	391,269,780	316,573,318	324,850,034	324,850,034	135,648,790	135,648,790	124,568,800	124,568,800	136,442,432	136,442,436
%, Reads mapped	83.78%	84.8%	69.16%	62.67%	65.39%	56.05%	70.09%	41.99%	57.78%	48.45%
Read length min/max/mean	100/100/100	100/100/100	100/100/100	100/100/100	50/101/99.0	50/101/99.0	50/101/98.9	50/101/98.9	50/101/98.9	50/101/98.9
Mean ± SD Coverage	118.34 ± 1,823.0	96.7 ± 691.0	90.2 ± 281.8	73.1 ± 357.0	30.1 ± 349.2	26.4 ± 149.8	33.7 ± 503.5	17.9 ± 217.8	33.8 ± 118.1	22.8 ± 386.2
% reference covered with ≥ 10X	87.8%	87.5%	84.7%	81.0%	69.7%	67.0%	84.8%	52.8%	88.9%	56.5%
Mean MQ	40.2	40.4	43.5	36.1	44.9	37.4	48.2	34.4	50.4	34.1
%GC	43.7%	43.5%	43.8%	44.4%	45.1%	46.2%	44.3%	46.5%	44.8%	46.7%
Insert sizes p25/ median/p75	316/340/359	180/334/359	278/334/390	267/330/388	117/152/197	113/151/198	121/155/201	112/152/199	121/158/206	108/153/202

**An. quad : An. quadriannulatus.*

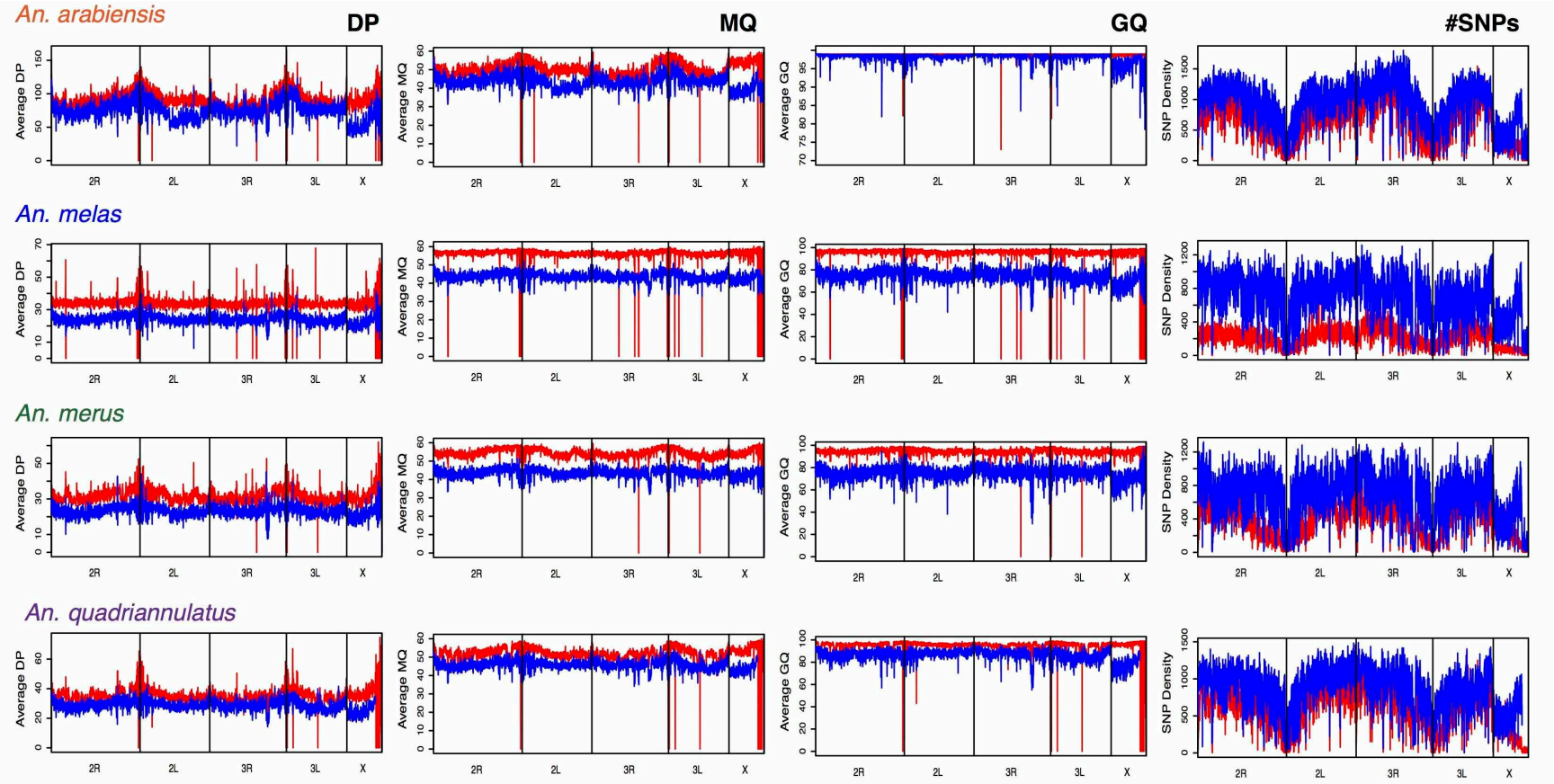


Fig. S5.
Distribution along each chromosome arm of metrics of SNP quality depending upon the reference genome (conspecific, red ; PEST, blue). Metrics shown are depth (DP), mapping quality (MQ), genotype quality (GQ) and number of SNPs calculated in non-overlapping 50kb windows.

S2.2. Multiple individuals per species sequenced at lower depth

We explored further the wide spectrum of diversity in each of the six sequenced species of the *Anopheles gambiae* complex by resequencing the whole genome of multiple individuals from each species at a lower depth of coverage.

S2.2.1 Sample collection and sequencing

Sampling and DNA extraction. Overall, 74 individual female mosquitoes were obtained from natural populations of six species of the *An. gambiae* complex (Fig. S6, Table S4): 26 *An. gambiae* (S-form), 12 *An. coluzzii* (M-form), 12 *An. arabiensis*, 10 *An. merus*, 4 *An. melas*, and 10 *An. quadriannulatus.* Genomic DNA was extracted from individual female mosquitoes using either a DNeasy Qiagen extraction kit or a CTAB DNA extraction protocol (*58*). Species identification and karyotypes for the 2La inversion of *An. gambiae* and *An. coluzzii* were ascertained as described in section S2.1.1.

Library construction and sequencing. Protocols followed those employed for the samples sequenced to higher depth (section S2.1), with data production at two different sequencing centers (all *An. gambiae* and *An. coluzzii* sequences were sequenced by BGI; the remainder were sequenced by the Broad). Barcoded mosquito libraries were indexed at 7-12 per each lane of Illumina HiSeq.

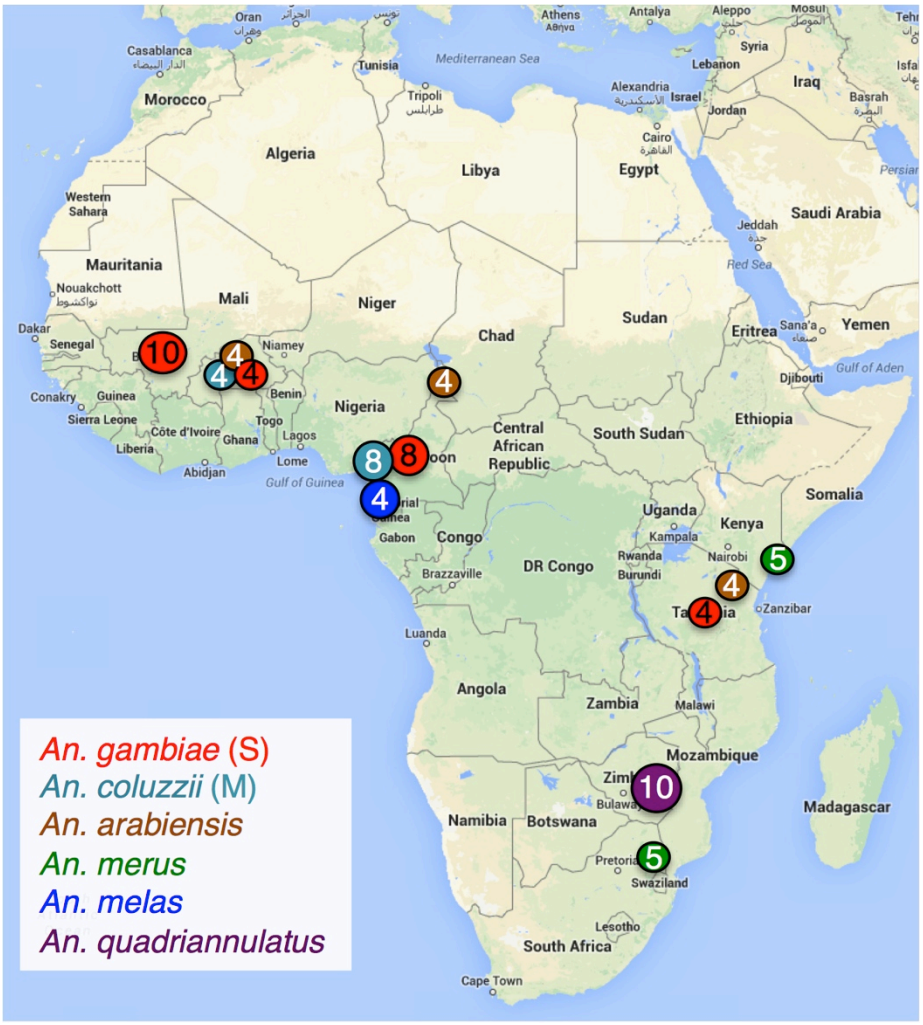


Fig. S6.
Approximate locations (circles) and sample size per locality (numbers within circles) for mosquitoes used for population sampling. Color of circles indicates species (see map inset). Table S4 provides more detailed information.

Table S4.
Metadata pertaining to the 74 population samples from six species, sequenced at low coverage.

Species	SampleID	SRA Identifier	BioProject	2La karyotype	Location Code	Village	Coord.	Year
An. arabiensis	BF0404094	SRP020595	PRJNA176974	2La/2La	BF	Monomtenga	12°06'N, 01°17'W	2004
An. arabiensis	BF0405003	SRP020603	PRJNA176971	2La/2La	BF	Monomtenga	12°06'N, 01°17'W	2004
An. arabiensis	BF0405012	SRP020600	PRJNA176975	2La/2La	BF	Monomtenga	12°06'N, 01°17'W	2004
An. arabiensis	BF0405744	SRP020521	PRJNA176976	2La/2La	BF	Monomtenga	12°06'N, 01°17'W	2004
An. arabiensis	CM0501012	SRP020594	PRJNA176968	2La/2La	CM	Moudawa	10°21'N, 14°11'E	2005
An. arabiensis	CM0501025	SRP020580	PRJNA176969	2La/2La	CM	Moudawa	10°21'N, 14°11'E	2005
An. arabiensis	CM0501026	SRP020597	PRJNA176970	2La/2La	CM	Moudawa	10°21'N, 14°11'E	2005
An. arabiensis	CM0501028	SRP020599	PRJNA176972	2La/2La	CM	Moudawa	10°21'N, 14°11'E	2005
An. arabiensis	TZ 71-158	SRP020524	PRJNA176978	2La/2La	TZ	Mabogini	03°24'S, 37°22'E	2009
An. arabiensis	TZ 71-163	SRP020601	PRJNA176973	2La/2La	TZ	Mabogini	03°24'S, 37°22'E	2009
An. arabiensis	TZ 71-199	SRP020529	PRJNA176977	2La/2La	TZ	Mabogini	03°24'S, 37°22'E	2009
An. arabiensis	TZ 71-211	SRP020519	PRJNA176979	2La/2La	TZ	Mabogini	03°24'S, 37°22'E	2009
An. coluzzii	44.2*	SAMN02899193	PRJNA254046	2La/2La	BF	Sourkoudiguan	11°14'N, 04°32'W	2012
An. coluzzii	A7.4	SAMN02899194	PRJNA254046	2La/2La	BF	Sourkoudiguan	11°14'N, 04°32'W	2012
An. coluzzii	C27.2†	SAMN02899195	PRJNA254046	2La/2La	BF	Bana	11°14'N, 04°28'W	2012
An. coluzzii	C27.3	SAMN02899196	PRJNA254046	2La/2La	BF	Bana	11°14'N, 04°28'W	2012
An. coluzzii	4631	SAMN02899197	PRJNA254046	2L+^{a}/2L+^{a}	CM	Ahala	03°48'N, 11°30'E	2005
An. coluzzii	4634	SAMN02899198	PRJNA254046	2L+^{a}/2L+^{a}	CM	Ahala	03°48'N, 11°30'E	2005
An. coluzzii	4691	SAMN02899199	PRJNA254046	2L+^{a}/2L+^{a}	CM	Nkolbisson	03°52'N, 11°27'E	2005
An. coluzzii	4697	SAMN02899200	PRJNA254046	2L+^{a}/2L+^{a}	CM	Nkolfoulou_II	03°55'N, 11°34'E	2005
An. coluzzii	5090	SAMN02899201	PRJNA254046	2L+^{a}/2L+^{a}	CM	Boussibelika	02°43'N, 09°52'E	2005
An. coluzzii	5107	SAMN02899202	PRJNA254046	2L+^{a}/2L+^{a}	CM	Boussibelika	02°43'N, 09°52'E	2005
An. coluzzii	5108	SAMN02899203	PRJNA254046	2L+^{a}/2L+^{a}	CM	Boussibelika	02°43'N, 09°52'E	2005
An. coluzzii	5113	SAMN02899204	PRJNA254046	2L+^{a}/2L+^{a}	CM	Boussibelika	02°43'N, 09°52'E	2005
An. gambiae	40.2†	SAMN02899205	PRJNA254046	2La/2L+^{a}	BF	Pala	11°09'N, 04°14'W	2012
An. gambiae	44.4	SAMN02899206	PRJNA254046	2La/2La	BF	Sourkoudiguan	11°14'N, 04°32'W	2012

An. gambiae	45.3	SAMN02899207	PRJNA254046	2La/2La	BF	Sourkoudiguan	11°14'N, 04°32'W	2012
An. gambiae	M20.7	SAMN02899208	PRJNA254046	2La/2L+[a]	BF	Pala	11°09'N, 04°14'W	2012
An. gambiae	4696	SAMN02899209	PRJNA254046	2La/2L+[a]	CM	Nkolfoulou_II	03°55'N, 11°34'E	2005
An. gambiae	4698	SAMN02899210	PRJNA254046	2L+[a]/2L+[a]	CM	Nkolfoulou_II	03°55'N, 11°34'E	2005
An. gambiae	4700	SAMN02899211	PRJNA254046	2L+[a]/2L+[a]	CM	Nkolfoulou_II	03°55'N, 11°34'E	2005
An. gambiae	4701	SAMN02899212	PRJNA254046	2L+[a]/2L+[a]	CM	Nkolfoulou_II	03°55'N, 11°34'E	2005
An. gambiae	5091	SAMN02899213	PRJNA254046	2L+[a]/2L+[a]	CM	Boussibelika	02°43'N, 09°52'E	2005
An. gambiae	5093	SAMN02899214	PRJNA254046	2L+[a]/2L+[a]	CM	Boussibelika	02°43'N, 09°52'E	2005
An. gambiae	5095	SAMN02899215	PRJNA254046	2L+[a]/2L+[a]	CM	Boussibelika	02°43'N, 09°52'E	2005
An. gambiae	5109	SAMN02899216	PRJNA254046	2L+[a]/2L+[a]	CM	Boussibelika	02°43'N, 09°52'E	2005
An. gambiae	KL0218	SAMN02899217	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0220	SAMN02899218	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0231	SAMN02899219	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0333	SAMN02899220	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0341	SAMN02899221	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0370	SAMN02899222	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0671	SAMN02899223	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0899	SAMN02899224	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0028*	SAMN02899225	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	KL0829*	SAMN02899226	PRJNA254046	2La/2La	MI	Kela	11°88'N, 8°45' W	2004
An. gambiae	TZ22*	SAMN02899227	PRJNA254046	2La/2L+[a]	TZ	Njage	08°12'S, 36°11'E	2008
An. gambiae	TZ102	SAMN02899228	PRJNA254046	2La/2L+[a]	TZ	Njage	08°12'S, 36°11'E	2008
An. gambiae	TZ65	SAMN02899229	PRJNA254046	2La/2L+[a]	TZ	Njage	08°12'S, 36°11'E	2008
An. gambiae	TZ67	SAMN02899230	PRJNA254046	2L+[a]/2L+[a]	TZ	Njage	08°12'S, 36°11'E	2008
An. melas	CM1001067†	SRP020530	PRJNA176986	2L+[a]/2L+[a]	CM	Campo	02°22'N, 09 49'E	2010
An. melas	CM1001069*	SRP020515	PRJNA176987	2L+[a]/2L+[a]	CM	Campo	02°22'N, 09 49'E	2010
An. melas	CM1001095	SRP020520	PRJNA176984	2L+[a]/2L+[a]	CM	Campo	02°22'N, 09 49'E	2010
An. melas	CM1002058	SRP020604	PRJNA176983	2L+[a]/2L+[a]	CM	Campo	02°22'N, 09 49'E	2010
An. merus	KN 00005	SRP020674	PRJNA176989	2La/2La	KN	Garithe village	?	2007-8

An. merus	KN 00006	SRP020665	PRJNA176996	2La/2La	KN	Garithe village	?	2007-8
An. merus	KN 00007	SRP020675	PRJNA176988	2La/2La	KN	Garithe village	?	2007-8
An. merus	KN 00035	SRP022553	PRJNA176994	2La/2La	KN	Garithe village	?	2007-8
An. merus	KN 00037	SRP020673	PRJNA176991	2La/2La	KN	Garithe village	?	2007-8
An. merus	Mpug 686g	SRP020531	PRJNA176995	2La/2La	SA	Mpumalanga, Koomatipoort	25°26'S, 31°57'E	1988
An. merus	Mpug 686h	SRP020577	PRJNA176990	2La/2La	SA	Mpumalanga, Koomatipoort	25°26'S, 31°57'E	1988
An. merus	Mpug 686i†	SRP020602	PRJNA176993	2La/2La	SA	Mpumalanga, Koomatipoort	25°26'S, 31°57'E	1988
An. merus	Mpug 686j	SRP020584	PRJNA176992	2La/2La	SA	Mpumalanga, Koomatipoort	25°26'S, 31°57'E	1988
An. merus	Mpug 803b	SRP020532	PRJNA176997	2La/2La	SA	Mpumalanga, Koomatipoort	25°26'S, 31°57'E	1988
An. quad.	24	SRP020525	PRJNA177007	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	41	SRP022554	PRJNA177003	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	42	SRP020583	PRJNA177002	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	47	SRP020582	PRJNA176998	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	51	SRP020579	PRJNA176999	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	53	SRP020518	PRJNA177004	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	71	SRP020576	PRJNA177001	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	72†	SRP020593	PRJNA177000	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	84	SRP020522	PRJNA177005	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. quad.	154	SRP020517	PRJNA177006	2L+^{a}/2L+^{a}	ZM	Chilongo,Chiredzi	21°03'S, 31°40'E	1986
An. christyi	ACHKN1017	SRP020566	PRJNA67213	2La/2La	KN	Kikuyu town	?	2010
An. epiroticus	MR4 _An.epiroticus_1	SRP028915	PRJNA191562	2La/2La	VN	Can Gio district, Ho Chi Minh City	?	2011

BF: Burkina Faso; CM: Cameroon; KN: Kenya; MI: Mali; SA: South Africa; TZ: Tanzania; VN: Vietnam; ZM: Zimbabwe. Coord.: GPS coordinates. *An. quad.: An. quadriannulatus.*

*Excluded from the nuclear data analyses because of very poor data quality.

†Sample also used for resequencing at high coverage.

S2.2.2. Read mapping and variant calling

Mapping, SNP calling, and QC filtering. Population samples were processed using steps *a* to *f* described in section S2.1.2. Each species was processed separately using its conspecific reference. Descriptive statistics of read mapping and sequencing depth are shown in Table S5. Intraspecific samples were combined for all steps following read mapping (*i.e.*, indel-realignment and variant calling). All sites (variant or not) were emitted across the genome using the HaplotypeCaller tool of GATK v.2.8 and 3.1 (*68, 70*). Genomic coordinates of the focal species reference assembly were then converted into a common PEST reference system as described (S2.1.2f) to allow interspecific analyses.

High quality (HQ) sites were retained for data analyses using hard filters, including a depth coverage (DP) between 4 and 30x (average per individual), a quality-by-depth ratio (QD) between 5.0 and 35.0, a mapping quality (MQ) $\geq$ 40.0, a probability of strand bias (FS) $\geq$ 60.0, and ReadPosRankSum $\leq$ -8.0. The boundaries for DP and QD were determined empirically based on their distribution defining a space where the average transition-transversion (Ti/Tv) ratio was stable (Fig. S7).

After hard filtering, the VCF files were merged using GATK (CombineVariants tool). We restricted our analyses only to biallelic SNPs (using GATK-SelectVariant), keeping only genotypes that had genotype quality (GQ) of at least 30, 4 reads to support the called genotype [vcftools v.1.0.12a (*71*), --minGQ 30 --minDP 4], and at least four diploid individuals called at that site (VCFtools v.0.1.12a, –mac 7).

After QC filtering, we excluded any sample with a proportion of missing data higher than 80%. This resulted in excluding 1 of 12 *An. coluzzii* samples, 3 of 26 *An.*

gambiae samples, and 1 of 4 *An. melas* samples. For analyses requiring an outgroup species, we also merged to the final VCF file two outgroups (*An. christyi* and *An. epiroticus*) taken from the haploid reference assemblies and converted to diploid homozygotes. The final data set (available in DRYAD, doi:10.5061/dryad.f4114) included 69 individual mosquitoes called for 7,531,874 SNPs across all species (Table S6). The proportion of missing data strongly correlated with the average individual sequencing depth (Fig. S8).

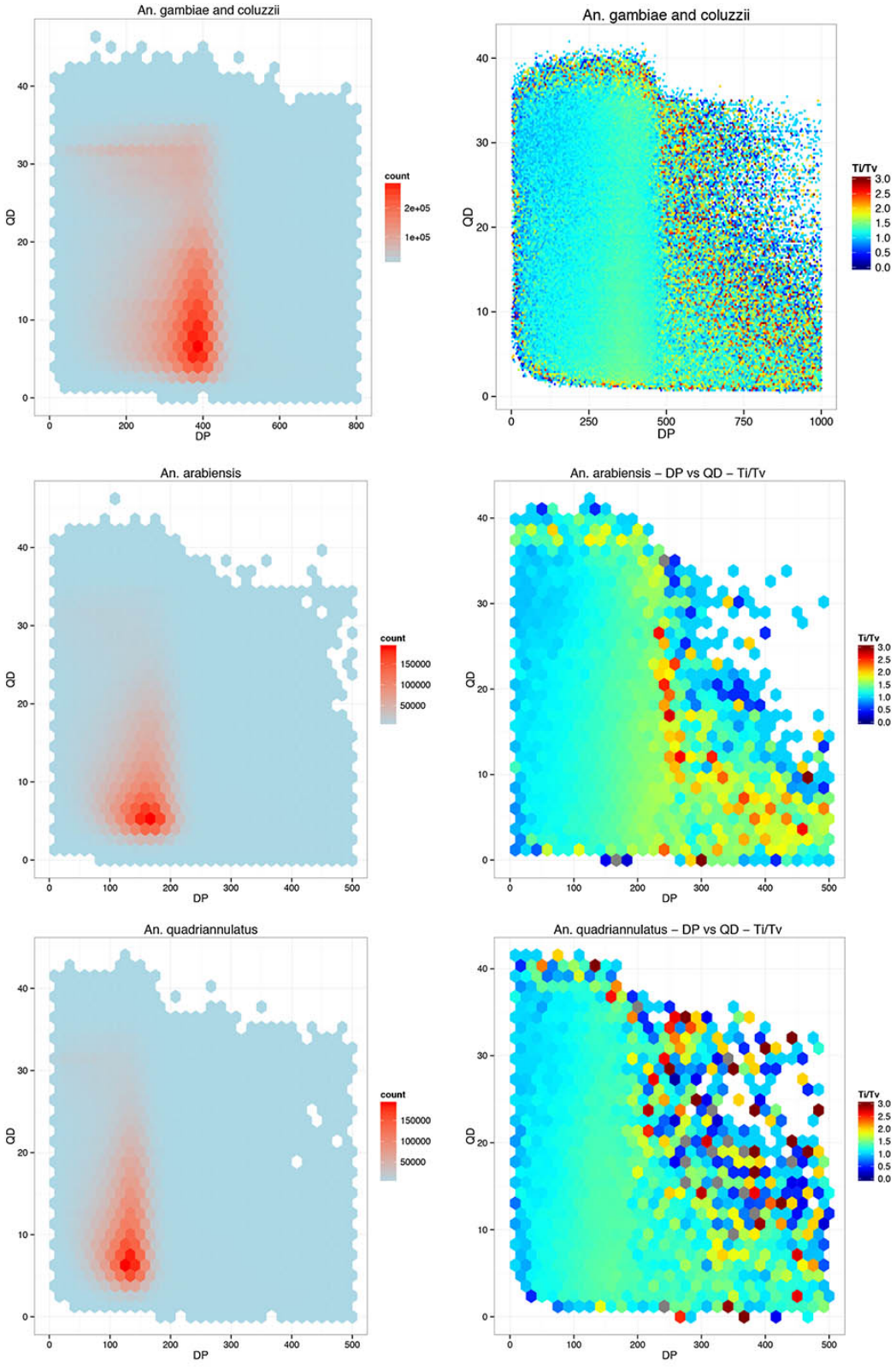
Table S5.
Descriptive statistics of the short-reads from intraspecific samples sequenced at low depth and mapped to their own reference assemblies. (*An. quad*., *An. quadriannulatus*).

Species		# Reads	Reads mapped	Reads paired	Reads MQ0	Total length	Bases mapped	Average Read length	Average quality	Average insert size	Average depth
An. arabiensis	mean	52,444,520	37,561,080	36,503,360	2,364,613	5,096,041,000	3,634,230,000	97	35.4	165	14.8
n = 12	sd	5,013,551	5,983,933	6,171,935	5,472,642	486,398,847	588,246,713	0	0.1	11	1.9
	min	44,861,610	29,969,840	29,042,310	648,210	4,362,147,000	2,899,938,000	96	35.2	136	12.0
	max	61,101,230	51,965,960	51,960,100	19,740,050	5,945,627,000	5,064,491,000	97	35.6	175	18.0
An. coluzzii	mean	32,161,160	21,232,950	20,591,250	2,812,346	3,216,116,000	2,123,295,000	100	36.5	465	7.5
n = 12	sd	7,897,420	4,398,670	4,300,782	569,943	789,742,015	439,866,954	0	0.2	19	1.7
	min	27,564,380	9,347,087	8,761,407	1,411,210	2,756,438,000	934,708,700	100	36.4	438	3.0
	max	56,608,680	26,719,940	25,731,240	3,584,974	5,660,868,000	2,671,994,000	100	36.9	509	10.0
An. gambiae	mean	35,019,060	26,249,950	25,541,400	3,662,446	3,472,143,000	2,601,507,000	99	36.3	361	9.5
n = 26	sd	6,111,850	6,250,016	6,102,560	874,187	606,962,448	628,332,996	1	0.2	114	2.7
	min	23,357,480	14,756,620	14,462,250	1,815,586	2,283,864,000	1,441,088,000	97	35.9	200	5.0
	max	48,480,770	38,237,250	37,127,080	5,560,724	4,848,077,000	3,823,725,000	100	36.6	487	15.0
An. merus	mean	50,230,360	32,355,590	31,257,710	2,407,006	4,890,652,000	3,133,435,000	97	35.6	170	11.6
n = 10	sd	6,683,652	5,280,179	5,155,522	351,060	651,421,770	511,869,616	0	0.1	3	2.4
	min	40,276,500	22,940,130	22,067,790	2,042,568	3,924,597,000	2,222,083,000	97	35.5	166	7.0
	max	57,941,170	39,146,720	37,949,300	3,058,085	5,645,641,000	3,793,025,000	97	35.6	175	15.0
An. quad.	mean	49,678,780	31,458,700	30,020,370	1,286,656	4,825,148,000	3,038,138,000	97	35.4	170	12.6
n = 10	sd	5,391,047	3,310,526	3,139,256	167,211	524,060,754	319,421,042	0	0.1	4	1.6
	min	39,562,880	25,492,420	24,366,600	968,265	3,844,544,000	2,464,603,000	97	35.4	162	10.0
	max	57,065,430	35,800,900	34,230,290	1,475,746	5,546,052,000	3,460,880,000	97	35.5	177	15.0
An. melas	mean	50,196,350	27,774,210	26,929,640	1,550,017	4,856,016,000	2,685,538,000	97	35.3	170	11.0
n = 4	sd	6,608,560	10,669,430	10,369,110	583,740	680,195,200	1,034,676,000	1	0.4	8	5.4
	min	40,901,620	11,797,350	11,406,550	675,427	3,894,107,000	1,136,199,000	95	34.7	159	3.0
	max	55,715,860	33,610,350	32,584,270	1,868,764	5,413,520,000	3,253,929,000	97	35.5	176	14.0

Table S6.
Descriptive statistics of population samples including the average depth (DP), number of complete genotypes (#SNP), number (#N) and proportion (%N) of missing genotypes relative to the total number of SNPs across the six species (7,531,874), and number of homozygous SNPs (Ho). Values are provided as per sample mean, median, SD, min, max and the total number of SNPs identified. *An. quad*., *An. quadriannulatus*; *An. chris*., *An. christyi*; *An. epir*., *An. epiroticus*.

Species		**DP**	**#SNP**	**#N**	**%N**	**Ho**
An. arabiensis	**mean**	13.2	4,549,665	2,982,209	0.40	4,238,263
n=12	**median**	13.7	4,833,064	2,698,811	0.36	4,513,941
	sd	2.93	868,831	868,831	0.12	845,790
	min	6.38	2,143,699	2,164,973	0.29	1,940,605
	max	17.6	5,366,901	5,388,175	0.72	5,061,051
	Total		6,590,579			
An. coluzzii	**mean**	7.8	2,709,763	4,638,947	0.62	2,575,579
n=11	**median**	7.6	2,671,365	4,680,358	0.62	2,538,979
	sd	0.7	299,310	314,578	0.04	287,677
	min	7.1	2,372,321	4,095,079	0.54	2,250,125
	max	9.1	3,227,347	4,990,931	0.66	3,074,715
	Total		5,952,780			
An. gambiae	**mean**	9.2	3,055,908	4,252,458	0.56	2,900,338
n=23	**median**	8.6	3,004,926	4,323,433	0.57	2,854,780
	sd	2.0	625,168	659,627	0.09	612,047
	min	6.9	2,251,971	3,088,960	0.41	2,132,620
	max	13.3	4,186,380	5,115,458	0.68	4,032,878
	Total		6,304,438			
An. melas	**mean**	14.4	5,011,222	2,520,652	0.33	4,986,714
n=3	**median**	14.5	4,978,302	2,553,572	0.34	4,953,343
	sd	0.4	65,137	65,137	0.01	65,526
	min	14.0	4,969,116	2,445,626	0.32	4,944,591
	max	14.8	5,086,248	2,562,758	0.34	5,062,208
	Total		5,862,737			
An. merus	**mean**	11.3	4,650,290	2,881,584	0.38	4,468,566
n=10	**median**	10.9	4,634,788	2,897,086	0.38	4,455,079
	sd	2.1	730,837	730,837	0.10	716,496
	min	7.7	3,091,674	2,044,942	0.27	2,941,962
	max	14.2	5,486,932	4,440,200	0.59	5,287,031
	Total		6,853,032			

Species		DP	#SNP	#N	%N	Ho
An. quad.	**mean**	12.3	4,496,797	3,035,077	0.40	4,210,498
n=10	**median**	12.4	4,564,492	2,967,382	0.39	4,277,650
	sd	1.3	294,890	294,890	0.04	284,490
	min	10.1	3,939,870	2,666,208	0.35	3,670,941
	max	14.2	4,865,666	3,592,004	0.48	4,566,325
	Total		6,386,678			
An. chris.						
n=1		-	3,936,772	3,595,102	0.48	3,936,772
An. epir.						
n=1		-	3,865,595	3,666,279	0.49	3,865,595


An. gambiae and coluzzii
QD
DP
count
2e+05
1e+05
An. gambiae and coluzzii
QD
DP
Ti/Tv
3.0
2.5
2.0
1.5
1.0
0.5
0.0
An. arabiensis
QD
DP
count
150000
100000
50000
An. arabiensis – DP vs QD – Ti/Tv
QD
DP
Ti/Tv
An. quadriannulatus
QD
DP
count
150000
100000
50000
An. quadriannulatus – DP vs QD – Ti/Tv
QD
DP
Ti/Tv

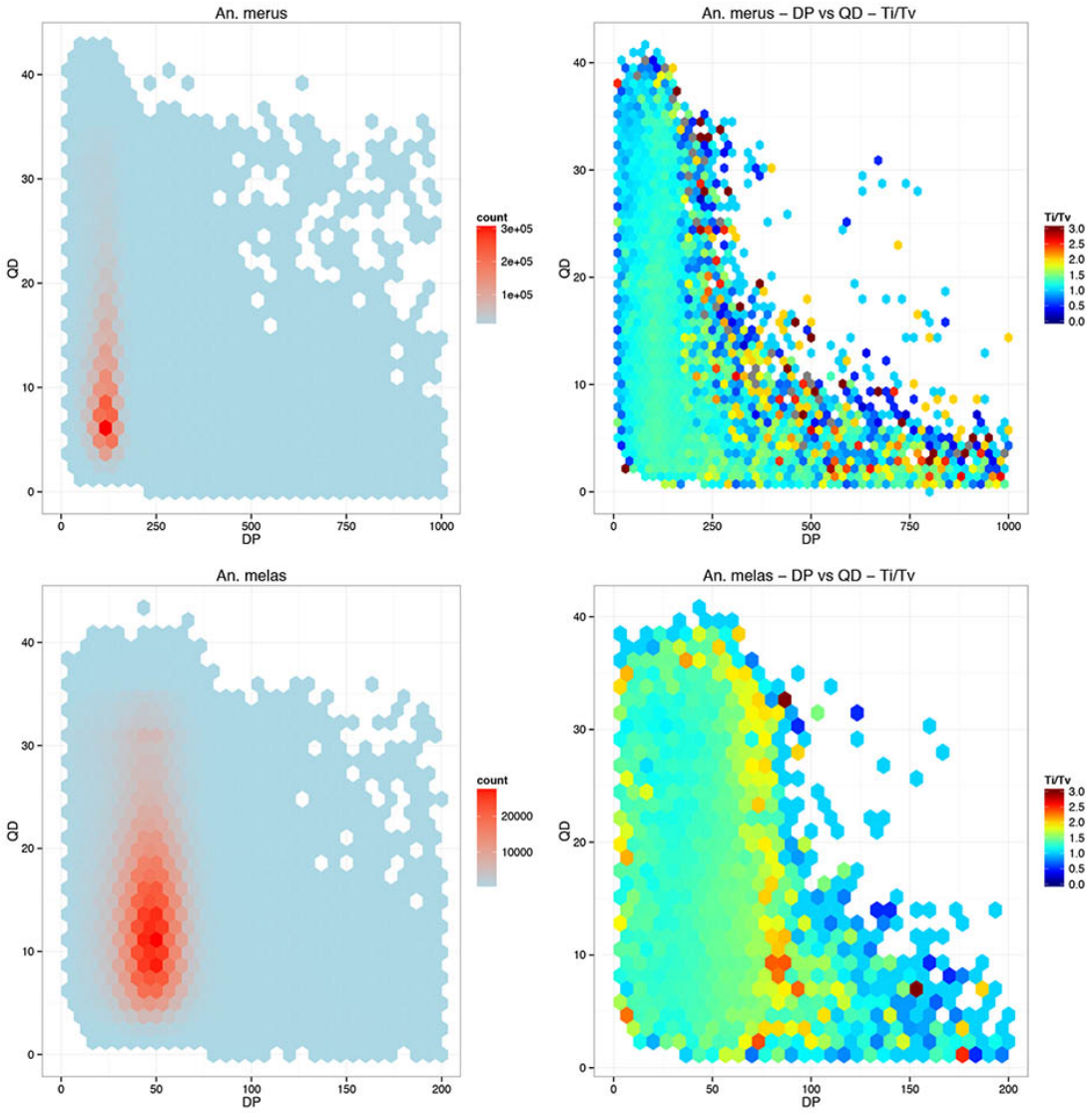


Fig. S7.
Two-dimensional histogram showing SNP counts (left panel) and Ti/Tv ratios (right panel) as a function of quality-by-depth ratio (QD) and the average depth (DP, summed over the individual in a focal species).

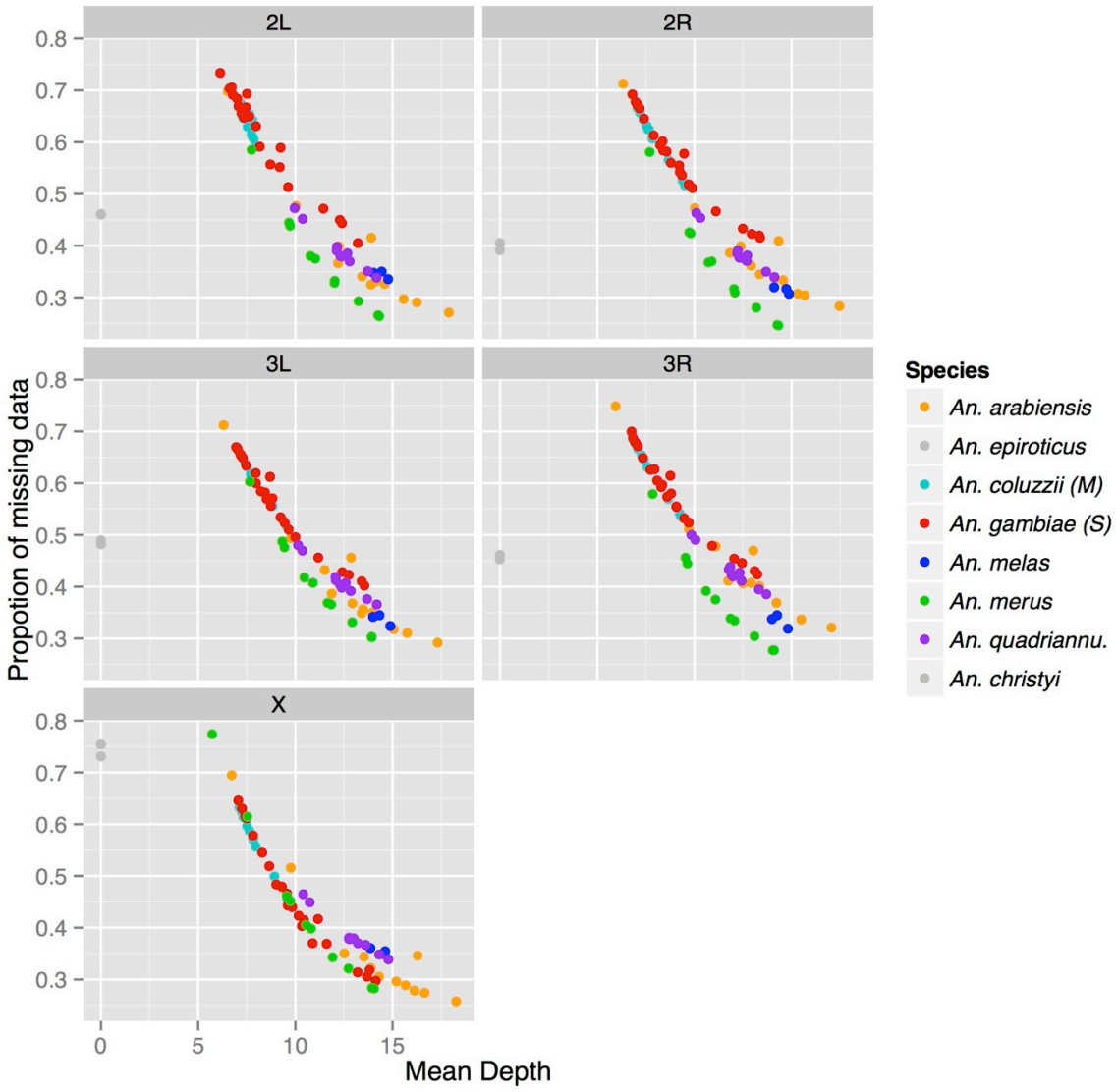


Fig. S8.
Proportion of missing data as a function of the sequencing depth for the 69 individual female mosquitoes (+2 outgroups) shown by chromosome arm.

S2.2.3 Polymorphism and divergence of nuclear genomes

Methodology. We analyzed intraspecific polymorphism and interspecific divergence using the SNP dataset derived from sequencing multiple individuals per species at lower coverage (see S2.2).

We used VCFtools v0.1.12a (*71*) to estimate the nucleotide diversity (π) within 50 kb non-overlapping windows along the genome, and the SNP allelic frequency in each of the six species of the *An. gambiae* complex included in our study. We plotted the two-dimensional allelic frequency spectrum between each pair of species using the *ipair* function of the IPDmisc v1.1.17 package in R v.3.0.2, using a $\log_2$ scale transformation. Linkage disequilibrium (LD), estimated as the r^2 correlation coefficient between pairs of SNPs that are at most 1 kb apart, was estimated in each species using PLINK v1.07 (*72*). We plotted the LD decay as a function of the distance between SNPs using *ggplot2* R package (*73*), fitting a GLM function. Differences in allelic frequencies between pairs of species was estimated using the weighted Weir and Cockerham's estimator of F_{ST} (*74*) calculated in 50 kb non-overlapping windows, using VCFtools.

Results. Descriptive statistics of the number of complete and missing genotypes in each species are shown in Table S6. The proportion of missing data was directly proportional to the sequencing depth (Fig. S8), which was highest in *An. coluzzii* and *An. gambiae,* and lowest in *An. merus* and *An. melas* (Table S6). Out of the total number of SNPs discovered (Table S6), only a fraction segregated in each species (from 1% in *An. melas* to 31% in *An. arabiensis* and 35% in *An. quadriannulatus*).

An. arabiensis and *quadriannulatus* had the highest genetic diversity both in terms of SNP density and heterozygosity (Fig. S9-S11). They are followed in order of decreasing

diversity by *An. gambiae, An. coluzzii* and *An. merus*. *An. melas* showed a markedly reduced level of genetic diversity (Fig. S9) and heterozygosity (Fig. S11). Reduced genetic diversity was also observed in the high depth samples (see section S2.1), and taken together, may suggest a severe bottleneck in the sampled population. This interpretation is consistent with the high level of linkage disequilibrium observed in *An. melas* compared to the other species, with the caveat that sample size is very small (Fig. S12). Another species displaying higher LD compared to the others was *An. coluzzii,* a result that is likely related to strong population structure, as observed previously within this taxon (*75*) and as shown in section S3.3.

With the caution that sample size is low in each species, differences in allelic frequencies, estimated by F_{ST} (Fig. S13), were always very high between species, except between the two very closely related species *An. gambiae* and *An. coluzzii*. The only chromosomal regions showing very high levels of differentiation between the latter pair were centromere-proximal on the autosomes and the X chromosome, consistent with previous observations (*76-78*). While F_{ST} values were very high on the autosomes for most species pairs, certain contrasts (*An. arabiensis* with *An. coluzzii*, *An. gambiae* and *An. quadriannulatus*; and *An. quadriannulatus* with *An. coluzzii* and *An. gambiae*) showed a slight reduction in F_{ST} values on the autosomes, but not on the X chromosome. This effect may be at least partially explained by the significant level of shared polymorphism related to introgression between *An. arabiensis*, *An. coluzzii* and *An. gambiae* (see S4).

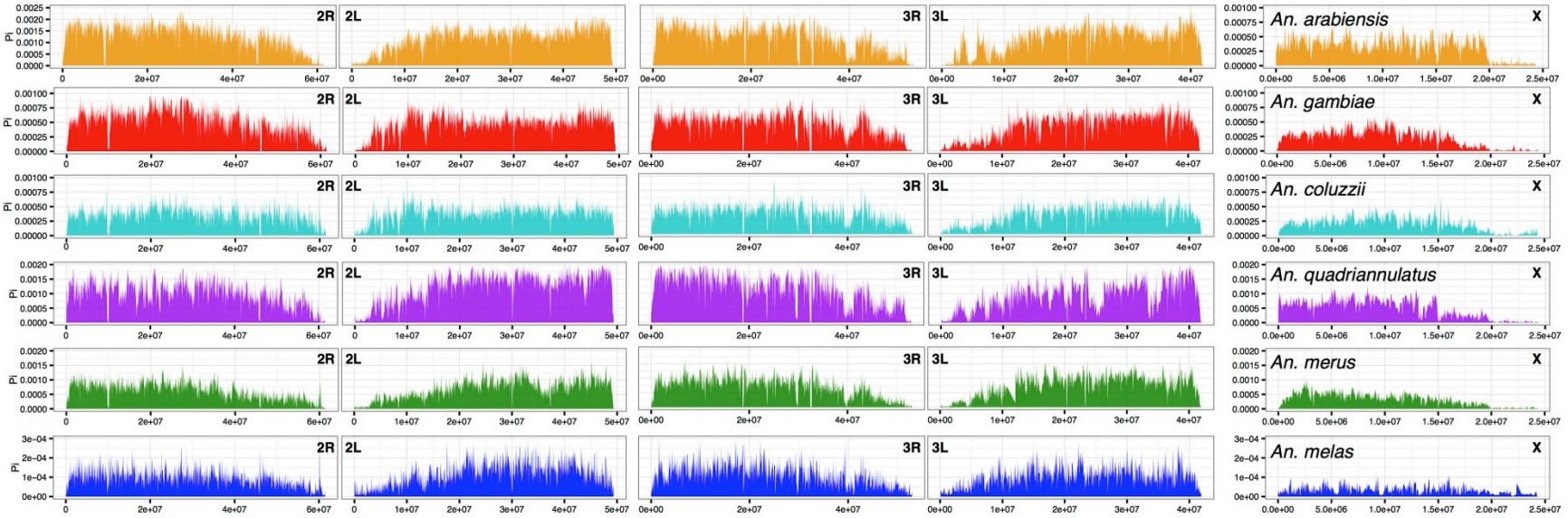


Fig. S9.

Nucleotide diversity (π) estimated within 50kb non-overlapping windows by chromosome arm for each studied species in the *An. gambiae* complex.

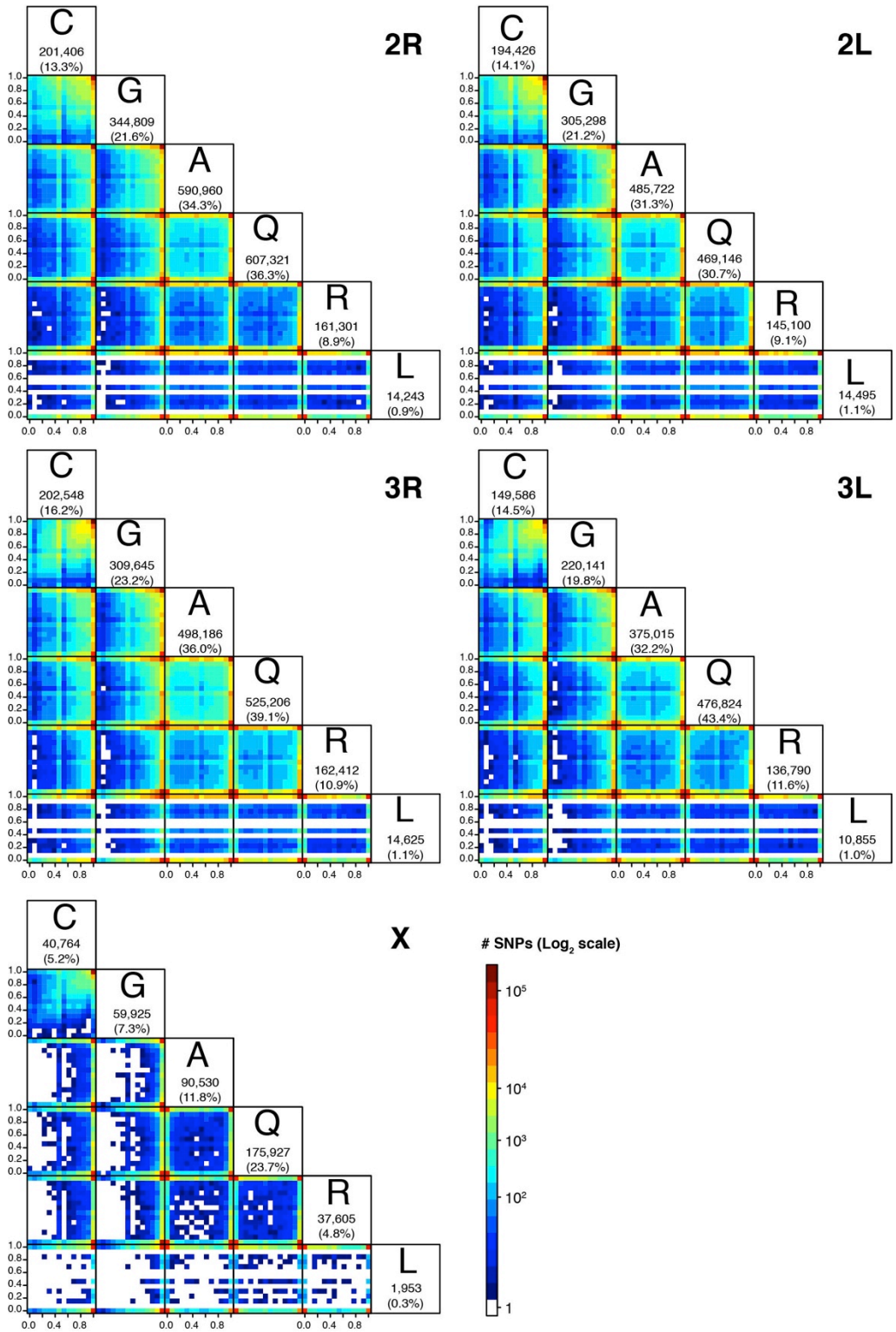


Fig. S10.

Pairwise comparisons of the allelic frequency spectrum (AFS) by chromosome arm. The number (and proportion) of segregating sites (with a minor allelic frequency ≥ 0.01) are shown in the box. A=*An. arabiensis,* C=*An. coluzzii,* G=*An. gambiae*, L=*An. melas*, Q=*An. quadriannulatus*, R=*An. merus.*

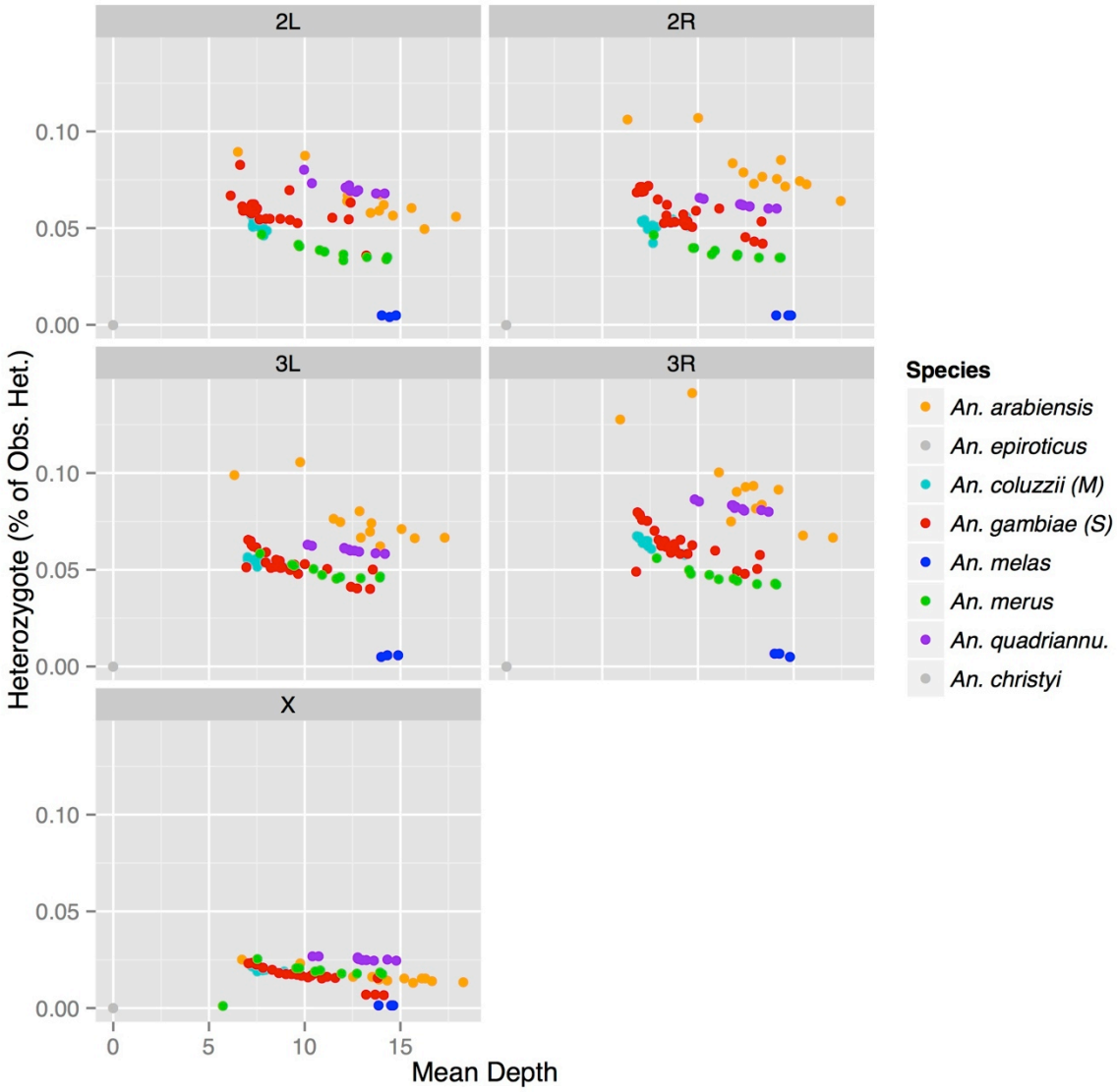


Fig. S11.
Proportion of heterozygous SNPs as a function of the mean sequencing depth by chromosome arm.

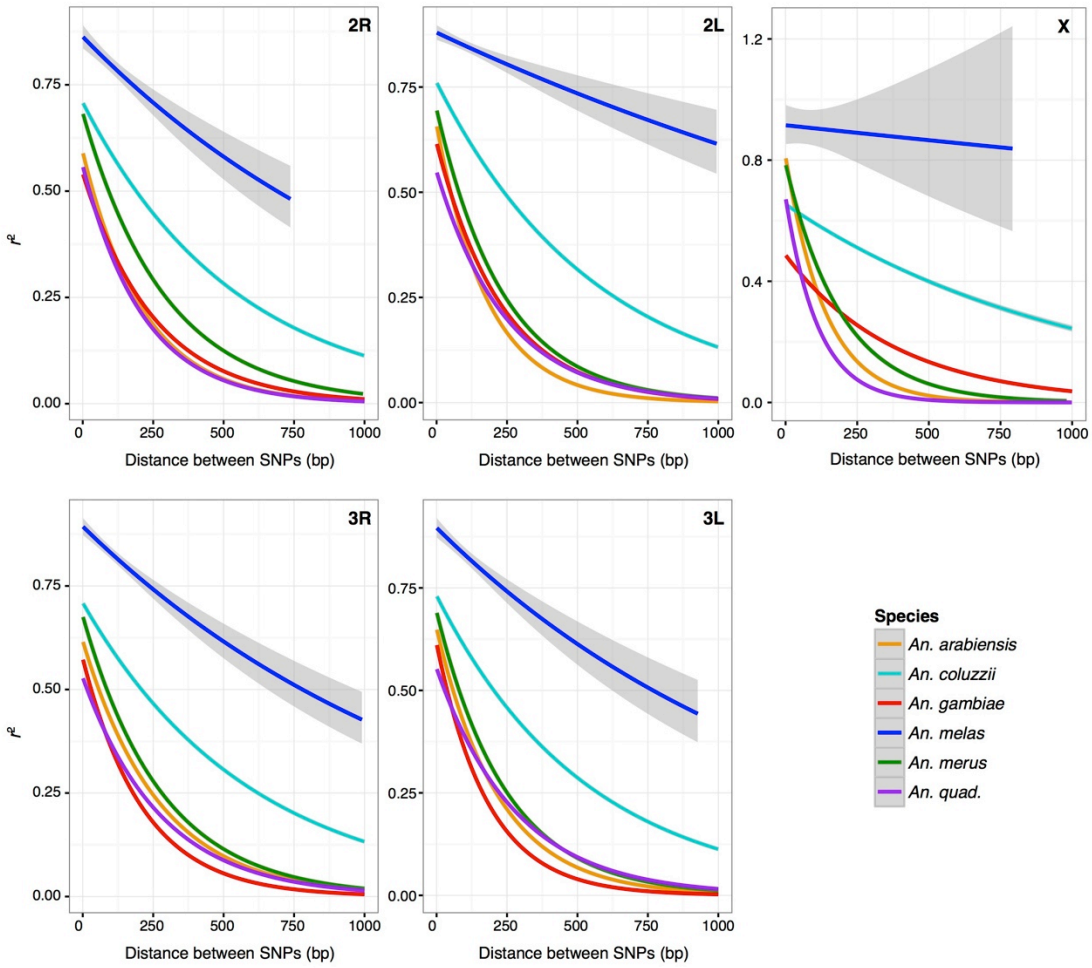


Fig. S12.
Linkage disequilibrium decay per chromosome arm in each species.

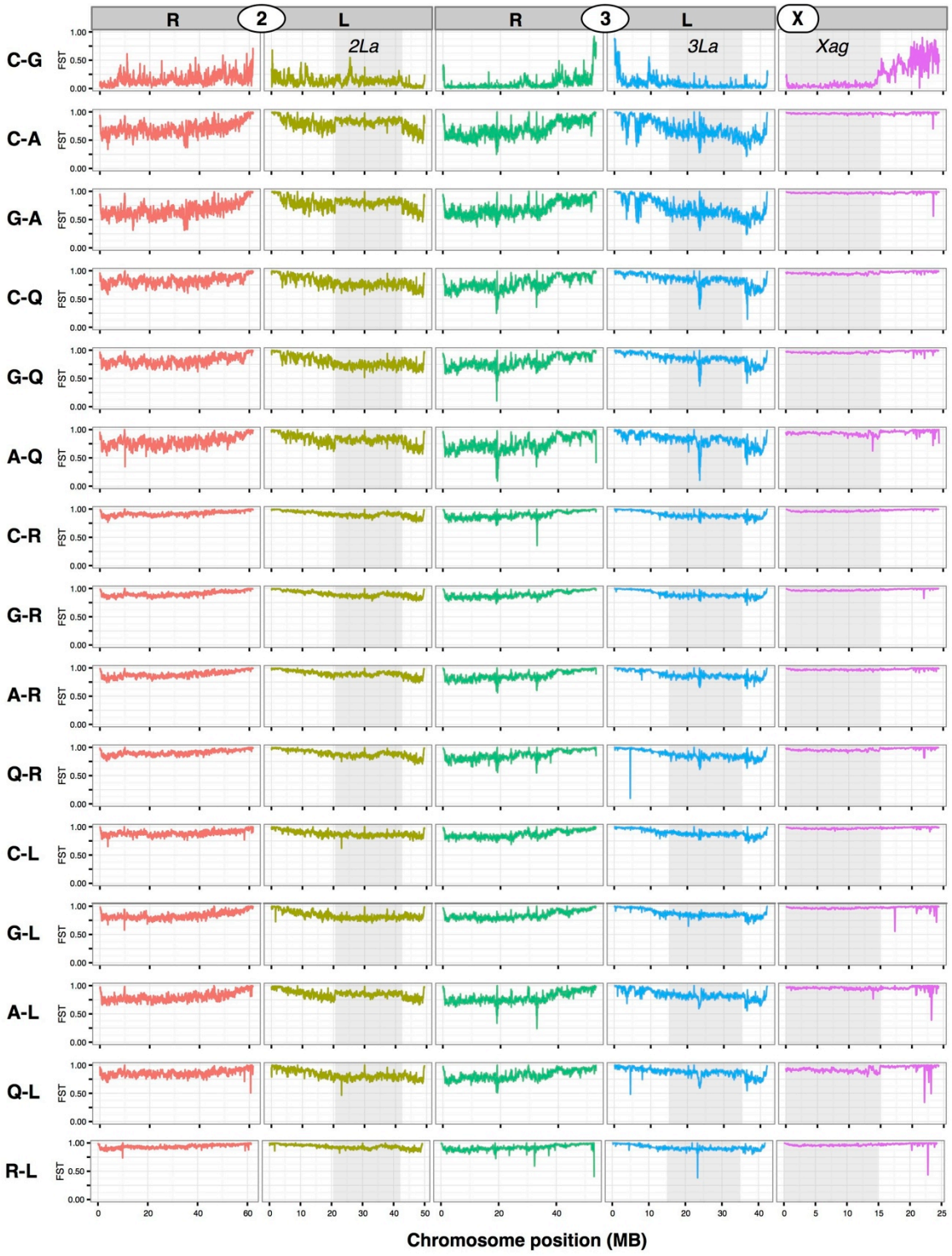


Fig. S13.
Differences in allelic frequency expressed as F_{ST} between each species in non-overlapping 50 kb windows by chromosome arm. A=*An. arabiensis,* C=*An. coluzzii,* G=*An. gambiae*, L=*An. melas*, Q=*An. quadriannulatus*, R=*An. merus.*

S2.2.4. Mitochondrial genome assembly

Methods. We used the whole genome resequencing (WGS) data from multiple individuals per species (section S3) to assemble the mitochondrial genome (mtDNA) of each of the six studied species in the *An. gambiae* complex. We used the previously published *An. gambiae s.s.* mitochondrial genome (*79*) as a reference (length: 15,363 base-pairs, bps). Major steps of the pipeline (Fig. S14) were adapted from (*80*). For each sample, we captured reads from the mitochondrial genome by mapping the trimmed reads using the BWA '*mem'* algorithm to the previously published *An. gambiae* mitochondrial genome as well as a modified assembly with the ends joined and the middle split (Table S7). This second round of mapping is intended to increase the number of captured reads at the extremes of the assemblies and take advantage of the circularity of mtDNA (*80*). Reads that mapped to each assembly were extracted and randomly downsampled to 400x coverage. These downsampled sets of 400x coverage were replicated five times for both assemblies. This protocol reduces the probability of inadvertent capture of nuclear-integrated mitochondrial DNA. Contigs were constructed for each sample using Hapsembler (http://compbio.cs.toronto.edu/hapsembler), thus creating 10 assemblies per sample. Using Nucmer (*81*), the 10 assemblies were aligned to the original reference assembly to order and orient the contigs, from which a consensus sequence was formed representing the final mitochondrial assembly for the given sample. We removed the highly variable non-coding control region from this consensus.

Results. We were able to reconstruct the mtDNA genome of 75 samples including 73 individuals from each of the six ingroup species and the two outgroups, *An. christyi* and *An. epiroticus* (Table S7). The mtDNA alignment is available in DRYAD

(doi:10.5061/dryad.f4114). Out of the 14,843 bp (14,833 without gaps), 1341 sites were polymorphic, 722 were singletons and 619 were polymorphisms observed more than once, defining 72 haplotypes (Table S8). The mtDNA nucleotide diversity (estimated with π and θ_W, Table S8) was highest in *An. gambiae* and *An*. *arabiensis* and lowest in *An. merus* and *An. melas*. Measures of genetic diversity were computed using DNASP v5 (*82*).

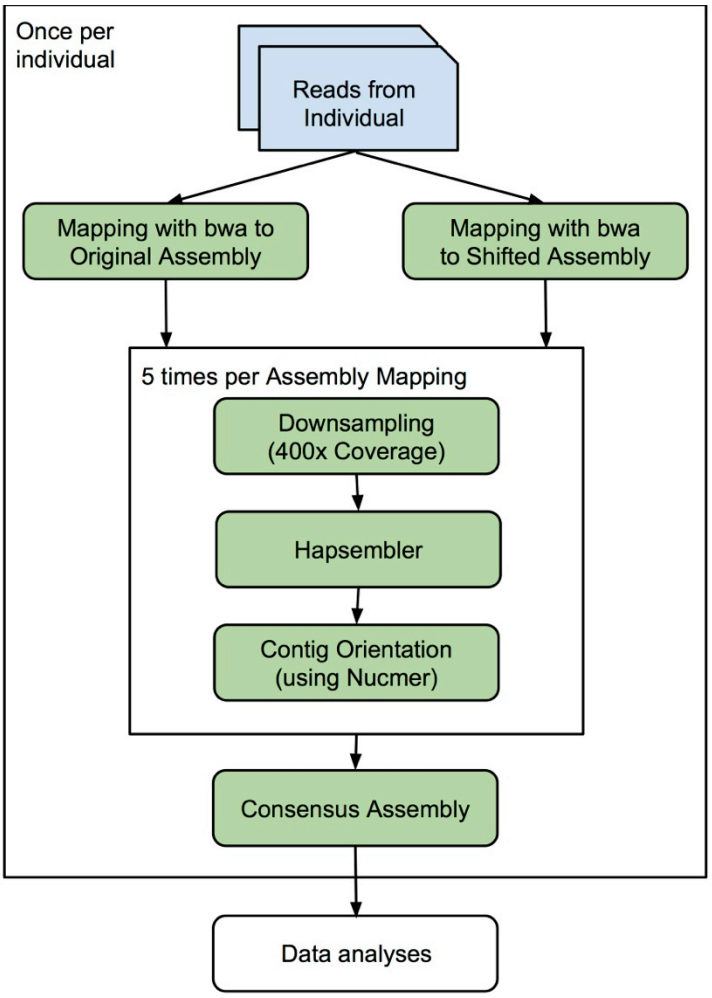


Fig. S14.
Pipeline for mitochondrial genome assembly.

Table S7.

Descriptive statistics of the reads mapped to the *An. gambiae* mitochondrial genome in each of six *Anopheles gambiae* complex species and two outgroups.

Statistics	Species	*AARAB*	*ACOL*	*AGAMB*	*AMELA*	*AMERU*	*AQUAD*	*ACHRI*	*AEPI*
N		12	12	26	4	10	9	1	1
Read Mapped to Ref.	**mean**	**131988.8**	**167845.2**	**141895.5**	**102731.5**	**213786.0**	**238871.2**	**214423.0**	**156267.0**
	SD	72034.2	146588.2	126094.6	62624.2	108416.0	100658.7	-	-
	median	132715.0	99697.0	108674.0	118619.0	193144.0	223917.0	-	-
	min	42429.0	47041.0	2303.0	19589.0	44043.0	104985.0	-	-
	max	303213.0	460969.0	481115.0	154099.0	379115.0	382065.0	-	-
Depth	**mean**	**859.1**	**1092.5**	**923.6**	**668.7**	**1391.6**	**1554.8**	**1395.7**	**1017.2**
	SD	468.9	954.2	820.8	407.6	705.7	655.2	-	-
	median	863.9	648.9	707.4	772.1	1257.2	1457.5	-	-
	min	276.2	306.2	15.0	127.5	286.7	683.4	-	-
	max	1973.7	3000.5	3131.6	1003.1	2467.7	2486.9	-	-
Read Mapped to modified Ref.	**mean**	**131112.3**	**166112.7**	**139904.7**	**101801.0**	**212242.6**	**236945.9**	**215667.0**	**153857.0**
	SD	71759.7	145184.5	124053.5	62101.3	107420.1	99918.1	-	-
	median	131745.0	98553.0	107636.0	117496.0	192355.0	221555.0	-	-
	min	42135.0	46653.0	2317.0	19459.0	43851.0	104191.0	-	-
	max	301827.0	457717.0	473975.0	152753.0	376531.0	379287.0	-	-
Depth modified Ref.	**mean**	**853.4**	**1081.3**	**910.7**	**662.6**	**1381.5**	**1542.3**	**1403.8**	**1001.5**
	SD	467.1	945.0	807.5	404.2	699.2	650.4	-	-
	median	857.5	641.5	700.6	764.8	1252.1	1442.1	-	-
	min	274.3	303.7	15.1	126.7	285.4	678.2	-	-
	max	1964.6	2979.3	3085.2	994.3	2450.9	2468.8	-	-
Sequence length	**mean**	**15362.0**	**15362.0**	**15361.9**	**15362.0**	**15362.0**	**15362.0**	**14862.0**	**14877.0**
	SD	0.0	0.0	0.6	0.0	0.0	0.0	-	-
	median	15362.0	15362.0	15362.0	15362.0	15362.0	15362.0	-	-
	min	15362.0	15362.0	15359.0	15362.0	15362.0	15362.0	-	-
	max	15362.0	15362.0	15362.0	15362.0	15362.0	15362.0	-	-
Non-N's sites	**mean**	**15362.0**	**15361.3**	**15361.3**	**15361.5**	**15362.0**	**15362.0**	**14858.0**	**14873.0**
	SD	0.0	0.9	0.9	0.6	0.0	0.0	-	-
	median	15362.0	15361.5	15362.0	15361.5	15362.0	15362.0	-	-
	min	15362.0	15360.0	15359.0	15361.0	15362.0	15362.0	-	-
	max	15362.0	15362.0	15362.0	15362.0	15362.0	15362.0	-	-
Non-N's w/o CR (14845 bps)	**mean**	**14843.0**	**14842.8**	**14843.0**	**14842.5**	**14843.0**	**14843.0**	**14839.0**	**14839.0**
	SD	0.0	0.4	0.2	0.6	0.0	0.0	-	-
	median	14843.0	14843.0	14843.0	14842.5	14843.0	14843.0	-	-
	min	14843.0	14842.0	14842.0	14842.0	14843.0	14843.0	-	-
	max	14843.0	14843.0	14843.0	14843.0	14843.0	14843.0	-	-

AARAB: *An. arabiensis;* ACOL: *An. coluzzii* (M); AGAMB: *An. gambiae* (S); AMELA: *An. melas;* AMERU: *An. merus;* AQUAD: *An. quadriannulatus;* ACHRI: *An. christyi;* AEPI: *An. epiroticus.*

Table S8.
Mitochondrial genome DNA polymorphism.

Statistics	*All + 2 outgroups*	*AARAB*	*ACOL*	*AGAMB*	*AQUAD*	*AMELA*	*AMERU*
N	75	12	12	26	9	4	10
S	1341	234	153	414	162	88	133
h	72	11	12	24	9	4	10
Hd	0.99	0.99	1.00	0.99	1.00	1.00	1.00
π	0.00822	0.0037	0.0034	0.0039	0.0035	0.0031	0.0025
θ_W	0.0185	0.0052	0.0034	0.0075	0.0040	0.0032	0.0032
K	121.88	54.6	50.3	58.3	51.7	46.5	36.6
D	–	-1.46	-0.07	-1.90	-0.69	-0.33	-1.13
***P-value* on D**	–	0.10	0.10	0.05	0.10	0.10	0.10
D*	–	-1.70	-0.31	-2.31	-0.74	-0.33	-1.24
P-value* on D	–	0.1	0.1	.1>P>.05	0.1	0.1	0.1
F*	–	-1.86	-0.03	-2.57	-0.82	-0.35	-1.37
P-value* on F	–	0.1	0.1	.1>P>.05	0.1	0.1	0.1

AARAB: *An. arabiensis;* ACOL: *An. coluzzii* (M); AGAMB: *An. gambiae* (S); AMELA: *An. melas;* AMERU: *An. merus*, AQUAD: *An. quadriannulatus;* ACHRI: *An: christyi;* AEPI: *An, epiroticus. Pi (per site); Theta (per site).*

S3. Phylogenomic analysis of the *An. gambiae* complex

S3.1. Whole genome alignments of single field-collected mosquitoes of each species.

Our objective was to compare phylogenies reconstructed from alignments of the reference assemblies (S1), which were largely derived from colonized anopheline mosquitoes, to phylogenies reconstructed from single field caught mosquitoes sequenced at a high depth (S2.1). Realizing this objective required whole genome alignments based on sequencing reads from the field samples, rather than the original alignments based on reference assemblies. To this end, we generated haploid consensus sequences from the diploid samples, using variants called with the GATK UnifiedGenotyper (*68*). For every species, each site of its conspecific reference assembly was replaced with the majority allele in the corresponding diploid field-caught specimen, as determined by the filtered VCF file (see section 2.1.2e). This was achieved by leveraging positional information stored in association with the reference genome alignments (section S1). Encoded in the Multiple Alignment Format (MAF) (*83*) file is a scaffold and a starting position for every alignment block of a focal reference genome to the PEST reference. We leveraged this positional information to replace sequences in the original reference assembly with the "haploidified" consensus sequence from wild specimens. In the instance of a deletion or missing data in the field specimen, the allele in the updated MAF file was made into an "N." All insertions relative to PEST were ignored to simplify analysis in PEST coordinate space. In addition to 6 species of the *An. gambiae* complex, *An. epiroticus* and *An. christyi* were also added to the alignments as outgroups. As a point of reference given its wide use, we included *An. gambiae* PEST in the same alignment with the haploid consensus sequences from wild specimens. The final alignments for the reference

assemblies and the high-depth field samples are available in DRYAD (doi:10.5061/dryad.f4114). Descriptive statistics in terms of nucleotide diversity and number of sites aligned in 50 kb windows are shown in Figure S15.

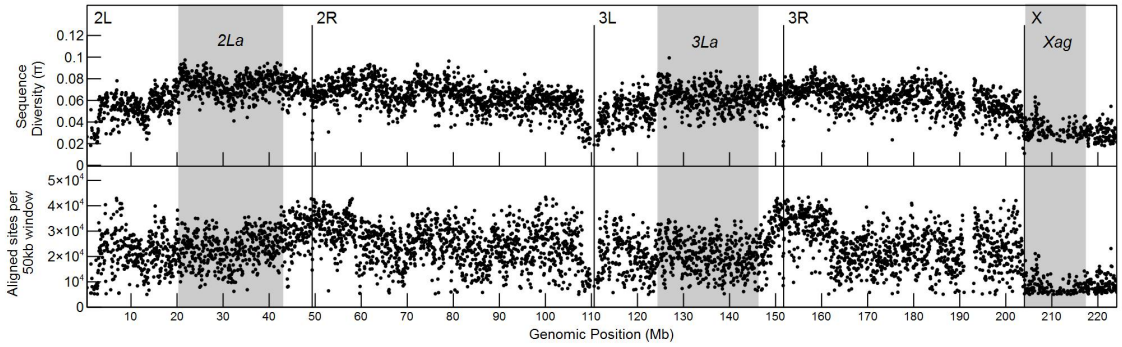


Fig. S15.
Chromosomal spatial plot for non-overlapping 50 kb genomic alignment windows for the high-depth genomic sequences determined from individual *An. arabiensis*, *An. christyi*, *An. coluzzii*, *An. gambiae*, *An. melas*, *An. merus*, and *An. quadriannulatus*. The two plots show the sequence diversity (π) for all taxa and the number of full-depth aligned sites per 50kb window.

S3.2 Window-based phylogenies and identification of the species branching order

As expected for a clade with large population sizes, rapid divergence, and apparently widespread introgression, there is a high degree of discordance among the observed gene (locus) trees. The conclusions below are based on analysis of phylogenies inferred from 50 kb non-overlapping windows from the high-depth whole-genome alignment of the taxa (A=*An. arabiensis,* C=*An. coluzzii,* G=*An. gambiae*, L=*An. melas*, Q=*An. quadriannulatus*, R=*An. merus,* O=outgroup=*An. christyi*). We will also refer to *An. gambiae* and *An. coluzzii* as the "gambiae group," and *An. arabiensis*, *An. melas*, and *An. quadriannulatus*, as the "melas group." Taking into account the relative abundance (Table S9), spatial distributions (Fig. S16), and relative divergence times (main text, Fig. 1c; Fig. 3; Fig. 5) of the various phylogenies inferred from 50 kb genomic regions we can infer the following model of the species phylogeny and its features.

Gene tree reconstruction. Gene trees (which do not require genic regions) were inferred from 50 kb non-overlapping windows of the alignment, using RAxML (v.7.2.8) with GTRGAMMA model and rapid bootstrapping for 100 replicates and specifying *An. christyi* as the outgroup: "`-m GTRGAMMA -f a -# 100 -o Achr`" (*84, 85*). Only sites that were fully covered by sequence (*i.e.*, without gaps in any aligned species) were considered in any 50 kb window. For a 50 kb window to be included in the analysis, at least 10% (5 kb) of full-coverage sites were required. No correlation was found between sequence divergence and coverage, indicating that gaps in the coverage did not affect the results. Several other window sizes were also tested, with no substantive change in the results. All processing of MAF alignments into window alignments, automation of

RAxML, and postprocessing and counting of trees was handled by custom Python scripts. Chromosomal plots of gene tree distributions were generated using Veusz (v. 1.17.1 http://home.gna.org/veusz/).

The true species tree and phylogenies on the X chromosome. The phylogeny (O,((C,G),(R,(L,(A,Q))))) is likely the true species phylogeny (tree vii in Fig. 2 and Fig. S16). It is the most common topology inferred in the Xag region (45.6%), which has the oldest inferred divergence times in the whole genome (except ancient polymorphic inversion regions, see below). The second- and third-most common phylogenies in the Xag region (trees viii and ix in Fig. 2 and Fig. S16) are (O,((L,(A,Q)),(R,(C,G)))) and (O,(R,((C,G),(L,(A,Q))))). These are simply rearrangements due to ILS of the majority X tree, due to the short time in between speciation events separating the melas group, gambiae group, and *An. merus*. Together these three topologies make up 94.5% of the Xag region, and 63.9% of the X chromosome as a whole.

The proximal region of the X chromosome (15-24Mb) appears to have phylogenies similar to the autosomes, and 40% of inferred trees from the region are the most common autosomal trees (trees i and ii). This is in agreement with previous hypotheses that the X proximal region may be introgressed between *An. arabiensis* and the ancestor of *An. gambiae* + *An. coluzzii* (*14*).

Introgressed trees on the autosomes. The overall most common phylogeny genome-wide is (O,((L,R),(Q,(A,(C,G))))) inferred from 39.7% of total genomic windows and 41.8% of autosomal windows (tree i, Fig. 2 and Fig. S21). The two ILS rearrangements of the

melas group, gambiae group, and *An. merus* are (O,(R,(L,(Q,(A,(C,G)))))) and (O,(L,(R,(Q,(A,(C,G)))))), which make up 11% and 1% of the total trees, respectively (trees ii and iii). Under the proposed species phylogeny from the X, we conclude these autosomal phylogenies are the result of introgression between *An. arabiensis* and *An. gambiae* + *An. coluzzii*. This is especially apparent in the 2La region (see next section). As *An. gambiae* + *An. coluzzii* are placed on the tree within the melas group, rather than *An. arabiensis* being placed with *An. gambiae* + *An. coluzzii*, we conclude that the majority of autosomal introgression was from *An. arabiensis* into *An. gambiae* + *An. coluzzii* (see also S4).

Introgression and the 2La inversion region. The 2La inversion region (2L coordinates: 20.5Mb – 42.1 Mb) contains phylogenies not seen anywhere else in the genome due to two factors:

1. The apparent genetic sequence distance between taxa with the 2La or 2L+^{a} haplotypes exceeds that inferred from the species phylogeny on the X, indicating that these two haplotypes diverged prior to the root of the complex.

2. The field-collected *An. gambiae s.s.* (G) sample (S2.1) is heterokaryotypic 2La/2L+^{a}, so this region contains a mix of alleles from both haplotypes. This makes the placement of G in the phylogeny highly variable.

The three most common phylogenies in this region—(O,((R,(A,C)),(L,(G,Q)))), (O,((L,R),(Q,(G,(A,C))))) and (O,((A,C),(R,(L,(G,Q))))))—together comprise 48.5% of trees inferred in the 2La region. These are three ILS variants of trees all showing a strong relationship of *arabiensis* and *coluzzii* (both homokaryotypic for 2La) due to strong

recent introgression of the 2La haplotype from the ancestor of *An. gambiae* (G) + *An. coluzzii* (C) into *arabiensis* (A). The presence of this introgression is inferred by a sharp reduction of sequence divergence between A and C in this region. The direction of introgression is inferred both by a partial reduction of G-A divergence and considering the model of inversion gain and loss on the species phylogeny (S5). Since Q and L both have the $2L+^{a}$ haplotype, we infer that the 2La haplotype was introgressed from the gambiae group into A.

An. merus-quadriannulatus introgression, 3La inversion region, and 3R enrichment. Genome wide, 24.8% of trees place Q and R as sister taxa, including 26.3% of autosomal regions and 1.6% (n=4) regions on the X chromosome. From this evidence we infer that R and Q may also be introgressing on the autosomes (Fig. 4). This inference is further supported by their strongly overlapping current geographical ranges and strong local enrichment of trees that place R and Q in a sister taxon relationship within the 3La region (see also Fig. S19). The 3La inversion region (3L coordinates ~15Mb – 35Mb) shows a strong enrichment of phylogenies that place R and Q as sister taxa (94.7%, compared to 24.8% overall). Chromosome 3R also shows an enrichment of Q-R sister trees at the proximal end (coordinates ~ 5-15Mb).

Determination of an introgressed phylogeny using divergence times: When introgression occurs between different species, the transfer of alleles causes the inferred time of divergence to be based on the time of introgression rather than the species divergence time. In a rooted three-taxon phylogeny there are two divergence times: the

earlier time when the first taxon diverges from the remaining sister pair (T_1) and the time when the paired taxa diverge (T_2, Fig. S16C,D; see also Fig. 3A). We use these times to determine which tree topologies represent those affected by introgression.

As the *gambiae* complex species have low divergence, we can measure the divergence times using simple pairwise distance measures. The vast majority of phylogenetically informative sites are biallelic, so allelic patterns can be represented as combinations of ancestral alleles (`A`) and derived alleles (`B`). For example, given the topology "((*P*1,*P*2),*P*3),*O*" we can represent the allelic patterns in the order $P_1P_2P_3O$ (e.g. `BBAA`), where *O* the outgroup always defines the ancestral allele (`A`). The counts of each pattern (e.g. $n_{\mathtt{BBAA}}$) can be compared to the total number of sites in the sampled region (*N)* to compute the divergence times. For the example tree, T_2 is the divergence of P_1 and P_2, and can be calculated as:

$$T_2 = \frac{1}{N}\left(\frac{n_{\mathrm{ABAA}} + n_{\mathrm{BAAA}}}{2}\right) \quad (1)$$

T_1 is calculated as T_2 plus the length of the branch ancestral to P_1 and P_2:

$$T_1 = \frac{1}{N}\left(\frac{n_{\mathrm{ABAA}} + n_{\mathrm{BAAA}}}{2} + n_{\mathrm{BBAA}}\right) \quad (2)$$

When the non-paired taxon (P_3) is the source of the introgression (i.e. the donor species), T_2 is predicted to be lower, since the observed T_2 will instead be the time of introgression rather than the true T_2 divergence time (Fig. S16C-D). When one of the two paired taxa (P_1 or P_2) is the source of introgression, both T_1 and T_2 will be lowered, since: (1) the true T_1 will not be observed, (2) the true T_2 will instead be the observed T_1, and (3) the time of introgression will be the observed T_2. In the case of bidirectional introgression we also expect that the observed T_1 and T_2 will be lower than their true

values. Therefore, when deciding between two incompatible alternative majority phylogenies, the topology with the lowest average divergence times is more likely to have experienced introgression, leaving the true species tree with older divergence times.

Table S9.
Counts of bootstrapped tree topologies from non-overlapping 50 kb genomic windows per chromosomal arm and for inversion regions (italicized). Top tree for each chromosome/region is bolded. (A=*An. arabiensis,* C=*An. coluzzii,* G=*An. gambiae*, L=*An. melas*, Q=*An. quadriannulatus*, R=*An. merus,* O=outgroup=*An. christyi*)

Topology	**2L**	***2La***	**2R**	**3L**	***3La***	**3R**	**X**	***Xag***	**Total**
Total	**918**	**437**	**1165**	**771**	**419**	**965**	**244**	**164**	**4063**
(O,((L,R),(Q,(A,(C,G)))))	**291**	7	**609**	192	13	**507**	16	0	**1615**
(O,(R,(L,(Q,(A,(C,G))))))	79	2	201	61	1	112	16	0	469
(O,((A,(C,G)),(L,(R,Q))))	18	2	58	**207**	**178**	94	1	0	378
(O,((R,Q),(L,(A,(C,G)))))	26	1	118	98	67	115	0	0	357
(O,(L,((R,Q),(A,(C,G)))))	2	0	1	112	110	15	0	0	130
(O,((A,(C,G)),(Q,(L,R))))	26	1	31	15	1	45	6	0	123
(O,((L,R),(Q,(G,(A,C)))))	77	76	4	1	0	1	0	0	83
(O,((R,(A,C)),(L,(G,Q))))	82	**82**	0	0	0	0	0	0	82
(O,((C,G),(R,(L,(A,Q)))))	0	0	0	0	0	0	**76**	**75**	76
(O,((L,R),(Q,(C,(A,G)))))	10	0	48	9	1	4	0	0	71
(O,((A,C),(R,(L,(G,Q)))))	54	54	0	0	0	0	0	0	54
(O,((L,(A,Q)),(R,(C,G))))	0	0	0	0	0	0	53	53	53
(O,(L,(R,(Q,(A,(C,G))))))	6	0	20	4	1	13	1	0	44
(O,((R,Q),(L,(G,(A,C)))))	29	29	0	3	2	1	0	0	33
(O,((R,Q),(L,(C,(A,G)))))	5	0	15	9	7	1	0	0	30
(O,((C,(A,G)),(L,(R,Q))))	0	0	9	19	17	2	0	0	30
(O,(R,(Q,(L,(A,(C,G))))))	5	0	17	2	0	4	0	0	28
(O,(R,((C,G),(L,(A,Q)))))	0	0	0	0	0	0	27	27	27
(O,(R,(L,(Q,(G,(A,C))))))	26	26	0	0	0	0	0	0	26
(O,((L,R),((A,Q),(C,G))))	2	0	3	1	0	12	7	0	25
(O,((G,(A,C)),(R,(L,Q))))	24	24	0	0	0	0	0	0	24
(O,((L,Q),(R,(G,(A,C)))))	24	23	0	0	0	0	0	0	24
(O,((G,(A,C)),(L,(R,Q))))	20	20	0	1	1	1	0	0	22
(O,((L,R),(A,(Q,(C,G)))))	4	0	1	7	1	7	2	0	21
(O,((A,(C,G)),(R,(L,Q))))	7	1	4	4	0	4	0	0	19
(O,((R,(A,C)),(G,(L,Q))))	18	18	0	0	0	0	0	0	18
(O,(R,(L,(Q,(C,(A,G))))))	1	0	13	1	1	2	0	0	17
(O,(L,((R,Q),(C,(A,G)))))	0	0	1	15	15	0	0	0	16
(O,(R,((L,Q),(G,(A,C)))))	15	15	0	0	0	0	0	0	15
(O,((L,R),((A,C),(G,Q))))	11	11	0	0	0	0	0	0	11
(O,((C,G),(A,(Q,(L,R)))))	1	0	1	1	1	3	4	0	10
(O,((G,(A,C)),(Q,(L,R))))	9	9	0	0	0	0	0	0	9
(O,((L,Q),(R,(A,(C,G)))))	4	0	0	0	0	4	0	0	8
(O,(R,(L,(A,(Q,(C,G))))))	2	0	0	3	0	1	1	0	7

Table S9 (continued):

Topology	2L	*2La*	2R	3L	*3La*	3R	X	*Xag*	Total
(O,((A,C),((G,Q),(L,R))))	6	6	0	0	0	0	0	0	6
(O,(R,((A,C),(L,(G,Q)))))	6	6	0	0	0	0	0	0	6
(O,(R,(L,((A,Q),(C,G)))))	1	0	0	1	0	1	3	0	6
(O,(R,((L,Q),(A,(C,G)))))	0	0	3	0	0	1	2	0	6
(O,((A,C),(R,(G,(L,Q)))))	4	4	0	0	0	0	0	0	4
(O,(R,(L,((A,C),(G,Q)))))	4	4	0	0	0	0	0	0	4
(O,(R,(Q,(L,(G,(A,C))))))	4	4	0	0	0	0	0	0	4
(O,(Q,(R,(L,(A,(C,G))))))	1	0	1	1	0	1	0	0	4
(O,((C,G),(A,(L,(R,Q)))))	0	0	1	0	0	1	2	0	4
(O,(L,((R,Q),(G,(A,C)))))	2	2	0	0	0	1	0	0	3
(O,((C,(A,G)),(Q,(L,R))))	0	0	3	0	0	0	0	0	3
(O,(Q,((L,R),(A,(C,G)))))	0	0	1	0	0	2	0	0	3
(O,((A,Q),((C,G),(L,R))))	0	0	0	0	0	1	2	0	3
(O,((Q,(A,L)),(R,(C,G))))	0	0	0	0	0	0	3	1	3
(O,((C,G),((A,Q),(L,R))))	0	0	0	0	0	0	3	0	3
(O,((A,C),(G,(R,(L,Q)))))	2	2	0	0	0	0	0	0	2
(O,((G,Q),(L,(R,(A,C)))))	2	2	0	0	0	0	0	0	2
(O,((A,(L,R)),(Q,(C,G))))	1	0	0	0	0	1	0	0	2
(O,(R,(Q,(L,(C,(A,G))))))	0	0	2	0	0	0	0	0	2
(O,((C,G),((A,L),(R,Q))))	0	0	0	0	0	1	1	0	2
(O,(R,(Q,((A,L),(C,G)))))	0	0	0	0	0	1	1	0	2
(O,(Q,(A,(L,(R,(C,G))))))	0	0	0	0	0	0	2	2	2
(O,(R,((C,G),(Q,(A,L)))))	0	0	0	0	0	0	2	1	2
(O,((C,G),(R,(Q,(A,L)))))	0	0	0	0	0	0	2	0	2
(O,(A,((C,G),(Q,(L,R)))))	0	0	0	0	0	0	2	0	2
(O,((G,L),((A,C),(R,Q))))	1	1	0	0	0	0	0	0	1
(O,((L,(A,C)),(R,(G,Q))))	1	1	0	0	0	0	0	0	1
(O,((L,Q),(G,(R,(A,C)))))	1	1	0	0	0	0	0	0	1
(O,((R,(A,C)),(Q,(G,L))))	1	1	0	0	0	0	0	0	1
(O,(L,(R,(Q,(G,(A,C))))))	1	1	0	0	0	0	0	0	1
(O,(L,(Q,(R,(G,(A,C))))))	1	1	0	0	0	0	0	0	1
(O,(A,(Q,((C,G),(L,R)))))	1	0	0	0	0	0	0	0	1
(O,(L,(R,(G,(C,(A,Q))))))	0	0	0	1	1	0	0	0	1
(O,(L,(R,(G,(Q,(A,C))))))	0	0	0	1	1	0	0	0	1
(O,(L,(Q,((A,R),(C,G)))))	0	0	0	1	0	0	0	0	1
(O,(Q,(L,(R,(C,(A,G))))))	0	0	0	1	0	0	0	0	1
(O,((A,(R,Q)),(L,(C,G))))	0	0	0	0	0	1	0	0	1
(O,((A,L),((C,G),(R,Q))))	0	0	0	0	0	1	0	0	1

(O,((L,R),(C,(Q,(A,G)))))	0	0	0	0	0	1	0	0	1

Table S9 (continued):

Topology	2L	*2La*	2R	3L	*3La*	3R	X	*Xag*	Total
(O,(L,(A,((C,G),(R,Q)))))	0	0	0	0	0	1	0	0	1
(O,(L,(R,(A,(C,(G,Q))))))	0	0	0	0	0	1	0	0	1
(O,(Q,((C,G),(A,(L,R)))))	0	0	0	0	0	1	0	0	1
(O,((A,Q),(L,(R,(C,G)))))	0	0	0	0	0	0	1	1	1
(O,((C,G),(L,(Q,(A,R)))))	0	0	0	0	0	0	1	1	1
(O,((R,(A,Q)),(L,(C,G))))	0	0	0	0	0	0	1	1	1
(O,(L,((A,Q),(R,(C,G)))))	0	0	0	0	0	0	1	1	1
(O,(R,((C,G),(A,(L,Q)))))	0	0	0	0	0	0	1	1	1
(O,((C,G),(L,(R,(A,Q)))))	0	0	0	0	0	0	1	0	1
(O,((C,G),(Q,(A,(L,R)))))	0	0	0	0	0	0	1	0	1
(O,(L,(R,((A,Q),(C,G)))))	0	0	0	0	0	0	1	0	1
(O,(R,((A,L),(Q,(C,G)))))	0	0	0	0	0	0	1	0	1

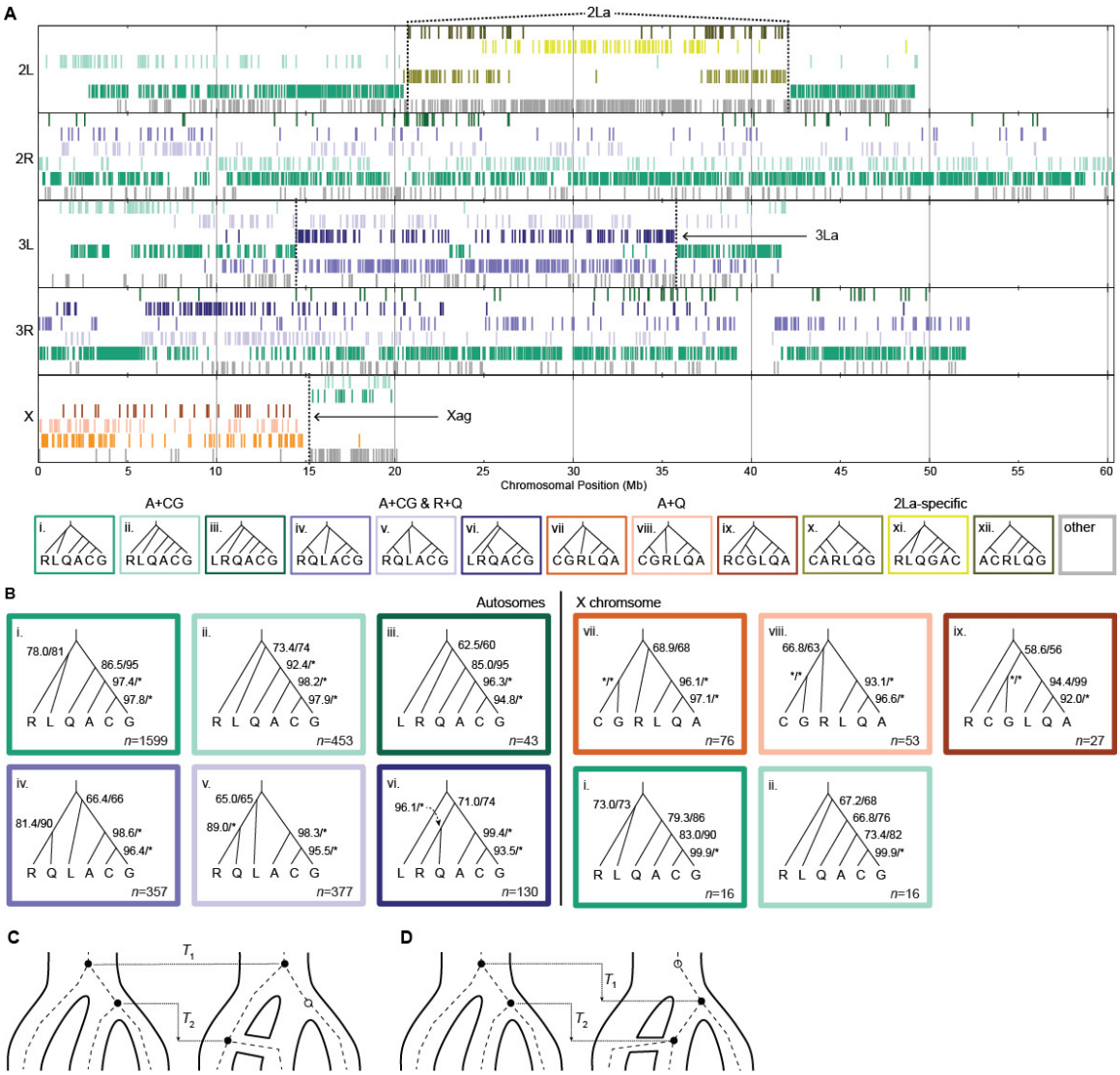

A
2La
2L
2R
3L
3La
3R
X
Xag
0
5
10
15
20
25
30
35
40
45
50
55
60
Chromosomal Position (Mb)
A+CG
A+CG & R+Q
A+Q
2La-specific
i. RLQACG
ii. RLQACG
iii. LRQACG
iv. RQLACG
v. RQLACG
vi. LRQACG
vii CGRLQA
viii. CGRLQA
ix. RCGLQA
x. CARLQG
xi. RLQGAC
xii. ACRLQG
other
B
Autosomes
X chromsome
i. 78.0/81 86.5/95 97.4/* 97.8/* R L Q A C G n=1599
ii. 73.4/74 92.4/* 98.2/* 97.9/* R L Q A C G n=453
iii. 62.5/60 85.0/95 96.3/* 94.8/* L R Q A C G n=43
vii. 68.9/68 */* 96.1/* 97.1/* C G R L Q A n=76
viii. 66.8/63 */* 93.1/* 96.6/* C G R L Q A n=53
ix. 58.6/56 */* 94.4/99 92.0/* R C G L Q A n=27
iv. 81.4/90 66.4/66 98.6/* 96.4/* R Q L A C G n=357
v. 89.0/* 65.0/65 98.3/* 95.5/* R Q L A C G n=377
vi. 96.1/* 71.0/74 99.4/* 93.5/* L R Q A C G n=130
i. 73.0/73 79.3/86 83.0/90 99.9/* R L Q A C G n=16
ii. 67.2/68 66.8/76 73.4/82 99.9/* R L Q A C G n=16
C
T_1
T_2
α
D
T_1
T_2
ϕ

Fig. S16. (A) Spatial chromosomal distribution of the most common rooted topologies (i-xii) for phylogenies inferred from 50 kb non-overlapping genomic regions for *An. gambiae* complex species (A=*An. arabiensis,* C=*An. coluzzii,* G=*An. gambiae*, L=*An. melas*, Q=*An. quadriannulatus*, R=*An. merus,* outgroup=*An. christyi*). The trees fall into four general categories (1) A+GC trees where A has introgressed with the *gambiae* group, (2) A+GC & R+Q trees where A-GC introgression and R-Q introgression have both occurred, (3) X-linked non-introgressed trees where A+Q are sister taxa, and (4) 2La-specific trees where sorting of the ancient 2La/+ haplotypes has occurred. **(B)** The mean/median bootstrap support for internal nodes among 50kb trees for each topology on either the autosomes (left) or X chromosome (right) demonstrate three features of the phylogenies generated from the autosomes and the X. First, regardless of topology the most poorly supported node is the first divergence among the *gambiae* group (C+G), *An. merus* (R), and the group *An. melas*, *quadriannulatus*, and *arabiensis* (L+Q+A). All trees show strong support for C and G as sister taxa (as expected), and autosomal trees show strong support for A as sister to C+G. The few autosome-like trees that appear on the X (right bottom row) have generally weaker support at all nodes (except C+G) than the X-majority trees. **(C)** For any rooted phylogeny of three taxa there are two divergence times, labeled T_1 and T_2. Introgression will generally tend to lower the apparent divergence time of T_2 when the taxon that is not in the sister pair (i.e. the first to diverge) is the source of introgression. Instead of inferring the true divergence time, the observed divergence time between these taxa instead will be the time of introgression. **(D)** When one of the sister taxa is the source of introgression, both T_1 and T_2 will be lowered, as: (1) the true T_1 will not be observed, (2) the true T_2 will instead be the observed T_1, and (3) the time of introgression will be the observed T_2.

S3.3. Molecular phylogeny reconstruction by chromosome arm, chromosomal inversions, and across the entire genome

Phylogenetic tree reconstruction. After trimming the reference genome alignments as well as the genome alignments based on single field collected specimens with Trimal (*86*) and excluding sites with gaps in more than 60% of the sequences, maximum likelihood phylogenies were reconstructed for both alignments with RAxML v8.0.14 (*84*), using a GTR-GAMMA substitution model. Node support was assessed using 500 rapid bootstrap resampling replicates (*85*). Phylogenetic reconstructions were conducted on the sequence alignments for the whole genome, for individual chromosomal arms, and for inversions 2La (2L coordinates: 20.5Mb – 42.1 Mb) and 3La (3L coordinates ~15Mb – 35Mb) (Fig. S17-S19). We included both *An. christyi* and *An. epiroticus* and used the most distant outgroup, *An. epiroticus,* to root the trees.

Nuclear and mtDNA phylogenies for population samples of six species. We assessed the genetic divergence between each species and the substructure within each species by reconstructing distance-based Neighbour Joining trees (fig. S20-S21) based on the nuclear biallelic SNP dataset obtained from multiple individual mosquitoes per species from natural populations (section S2.2). We used ADEGENET v1.4-2 (*87, 88*) to store the data in R v3.0.2 and computed a genotype-based Euclidian distance matrix between individuals. We drew a distance–based Neighbour Joining (NJ) tree rooted with *An. christyi* and evaluated node support using 100 bootstrap replicates (*89*). These two latter steps were performed in R using the APE v3.1-2 R-package (*90*).

For mtDNA, a maximum likelihood phylogenetic (Fig. S22) tree based on 75 mitochondrial genomes (section S2.2.4) was constructed using RAxML (parameters -m GTRGAMMA -# 1000 -T 16 –f a –x 12345 –p 12345, for deducing the best tree and T 16 -# 1000 –f b -m GTRGAMMA, for calculating bootstrap values using the fast-bootstrap method) (*84, 85*). *An. christyi* was used as an outgroup to root the tree.

Comparison of reference assembly-based and field sample-based phylogenies. Phylogenetic trees obtained from the reference assemblies and single wild specimens all display very high branch support and similar topologies when considering the sequence alignments for the whole genome, chromosome X, and arms 2R and 3R (Fig. S17-S18). However, conflicting topologies were observed on arms 2L and 3L, consistent with different karyotype combinations involving the 2La and 3La inversion systems influencing the inferred phylogenetic relationships (see below). Based on the reference assemblies, the 2L topology (Fig. S17C) is identical to that of the other autosomes: the *An. gambiae* AgamS1 reference (2La/+^{a}) is sister to the clade containing the *An. arabiensis* AaraD1 reference (2La/a) and the *An. coluzzii* AgamM1 reference (2La/a). Based on the field-collected samples, the 2L topology (Fig. S18C) shows the *An. gambiae* field sample (also 2La/+^{a}) clustering with PEST (2L+^{a}/+^{a}). Phylogenies based exclusively on the 2La inversion in the reference (Fig. 3.5a) and field-based alignments (Fig. S19B) indicate that tree topology is driven by inversion status, and that different genomic sampling of the two alleles in the heterokaryotypic samples explains the differences in topologies for *An. gambiae*.

The differences in topology on chromosome 3L between the reference- and field-based alignments (Fig. S17E and S19E) involve the relationship between *An. quadriannulatus* and *An*. *merus*, and also are influenced by an inversion, 3La, that has introgressed between these two species (see below, and main text). Phylogenies based only on the 3La inversion in the reference (Fig. S19C) and field-based alignments (Fig. S19D) show that the two species cluster together, in disagreement with expectation based on the species branching order.

Confirmation of phylogenetic relationships with population samples of each species. The NJ tree based on multiple individual genomic sequences sampled from natural populations of each species (Fig. S20) confirmed and extended the single-genome results discussed above. Importantly, these trees revealed monophyly of intraspecific samples, even for species that were sampled from multiple, distant geographic locations. The only exception was the relationship between *An. coluzzii* and *An. gambiae*, which was paraphyletic, although *An. coluzzii* samples formed a cohesive clade inside of *An. gambiae* with respect to sequences on chromosome 3 and 2R, and *An. gambiae* samples formed a cohesive clade inside of *An. coluzzii* based on X chromosome sequences. Genetic subdivision was notable in both *An. merus* (between Kenya and South Africa) and *An. coluzzii* (between Cameroon and Burkina Faso).

The 2La and 3La inversions. The relationship between *An. coluzzii* and *An. gambiae* is particularly notable based on sequences from chromosome 2L, as neither taxon forms a single cohesive clade. This pattern is associated with the individual karyotype status of

the 2La inversion. A NJ tree based solely on the 2La inversion (Fig. S21A) reveals that individual *An. arabiensis* samples cluster within a group composed of the *An. gambiae* and *An. coluzzii* that are 2La/a homokaryotypes. By contrast, those *An. gambiae* and *An. coluzzii* samples that are $2L+^{a}/+^{a}$ homokaryotypes cluster in another group, and heterokaryotypic ($2La/2L+^{a}$) samples form a third group.

A NJ tree focusing on the 3La inversion region (Fig. S21B) shows that *An. quadriannulatus* samples cluster with *An. merus* samples, counter to the expected species relationships. This underlines the strong introgression signal we have observed from *An. merus* into *An. quadriannulatus* (see S4).

Mitochondrial phylogenies. Phylogenies based on the mitochondrial genome (Fig. S22) confirmed and extended previous observations of extensive mtDNA intermingling and shared haplotypes between *An. gambiae* and *An. arabiensis* even across 7000 km of the African continent (*91, 92*). Our data reveal a complete lack of *An. arabiensis*-specific mtDNA haplotypes, despite the fact that we sampled 12 individuals from eastern and western African populations, suggesting that the mtDNA genome of *An. arabiensis* has been completely replaced by the mtDNA genome from *An. gambiae* (or *An. coluzzii*). This result further stresses the very important level of introgression between *An. arabiensis* and *An. gambiae-An. coluzzii.* It is noteworthy that although we detected nuclear introgression between *An. merus* and *An. quadriannulatus*, their mtDNA sequences are monophyletic, as are the mtDNA sequences of all other species sampled. Another peculiar observation revealed by the mtDNA phylogeny is the clustering of *An. merus* and *An. melas* sequences into the same clade, although each species remains

monophyletic. This might suggest that some mitochondrial exchange occurred in the past, despite the present allopatric distribution of these species. However, we did not detect any evidence of nuclear introgression (see S4).

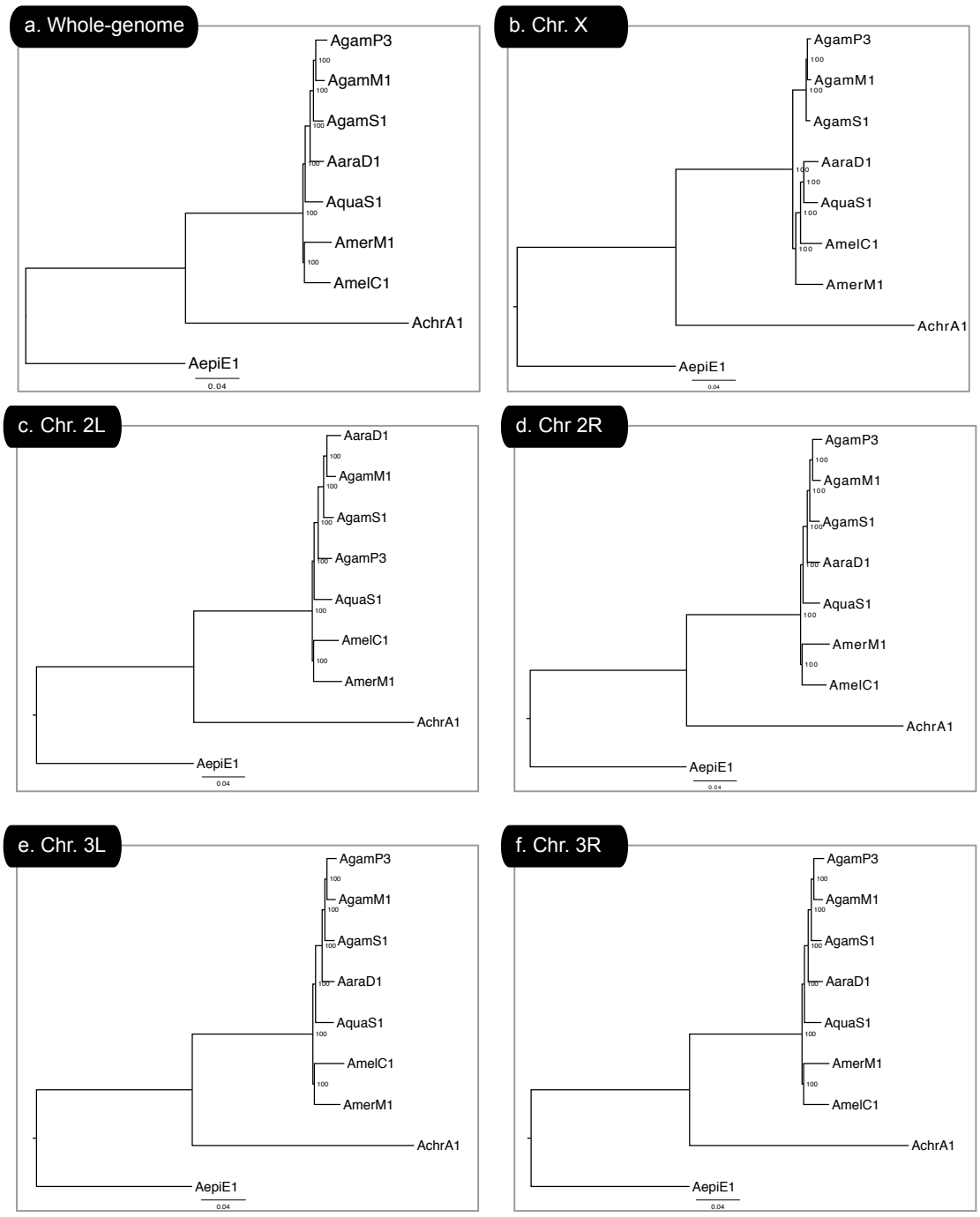


Fig. S17.
Maximum likelihood phylogenies based on the reference assemblies using TBA sequence alignments across the whole genome and by chromosomal arm.

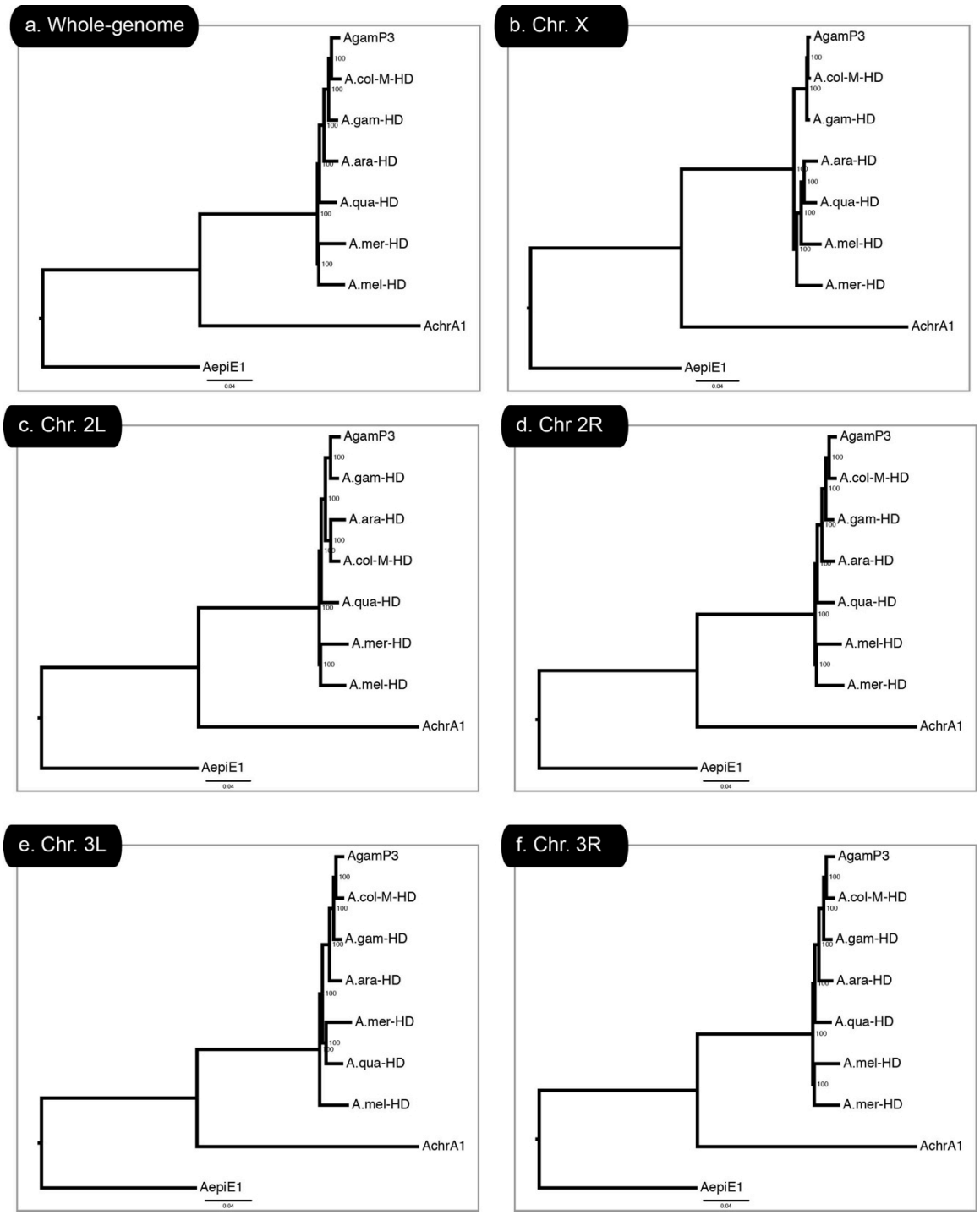


Fig. S18.
Maximum likelihood phylogenies of individual field-collected samples and the *An. gambiae* PEST reference (AgamP3) based on the whole genome or chromosomal arm TBA alignments. *An. gambiae*, A.gam-HD; *An. coluzzii*, A.col-M-HD; *An. arabiensis*, A.ara-HD; *An. quadriannulatus*, A.qua-HD; *An. merus*, A.mer-HD; *A. melas*, A.mel-HD.

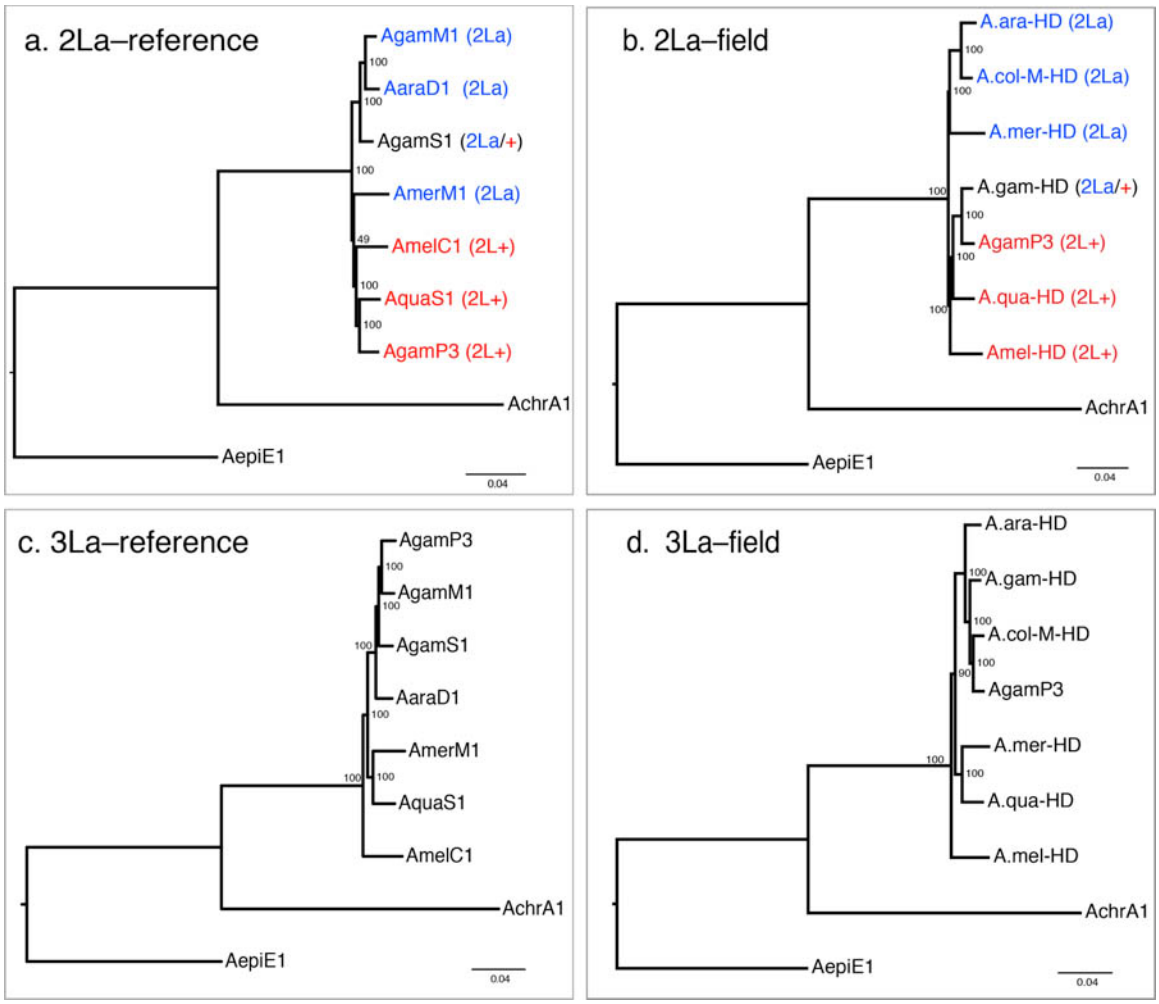


Fig. S19.
Maximum likelihood phylogenies for the 2La and 3La inversions based on the reference assemblies (a and c) and on the individual field-collected samples (b and d). *An. gambiae*, A.gam; *An. coluzzii*, A.col; *An. arabiensis*, A.ara; *An. quadriannulatus*, A.qua; *An. merus*, A.mer; *An. melas*, A.mel. 2La and 2L+ refer to 2La/2La and $2L+^{a}/2L+^{a}$ homokaryotypes for the 2La inversion, respectively. The heterokaryotype $2La/2L+^{a}$ is indicated by 2La/+.

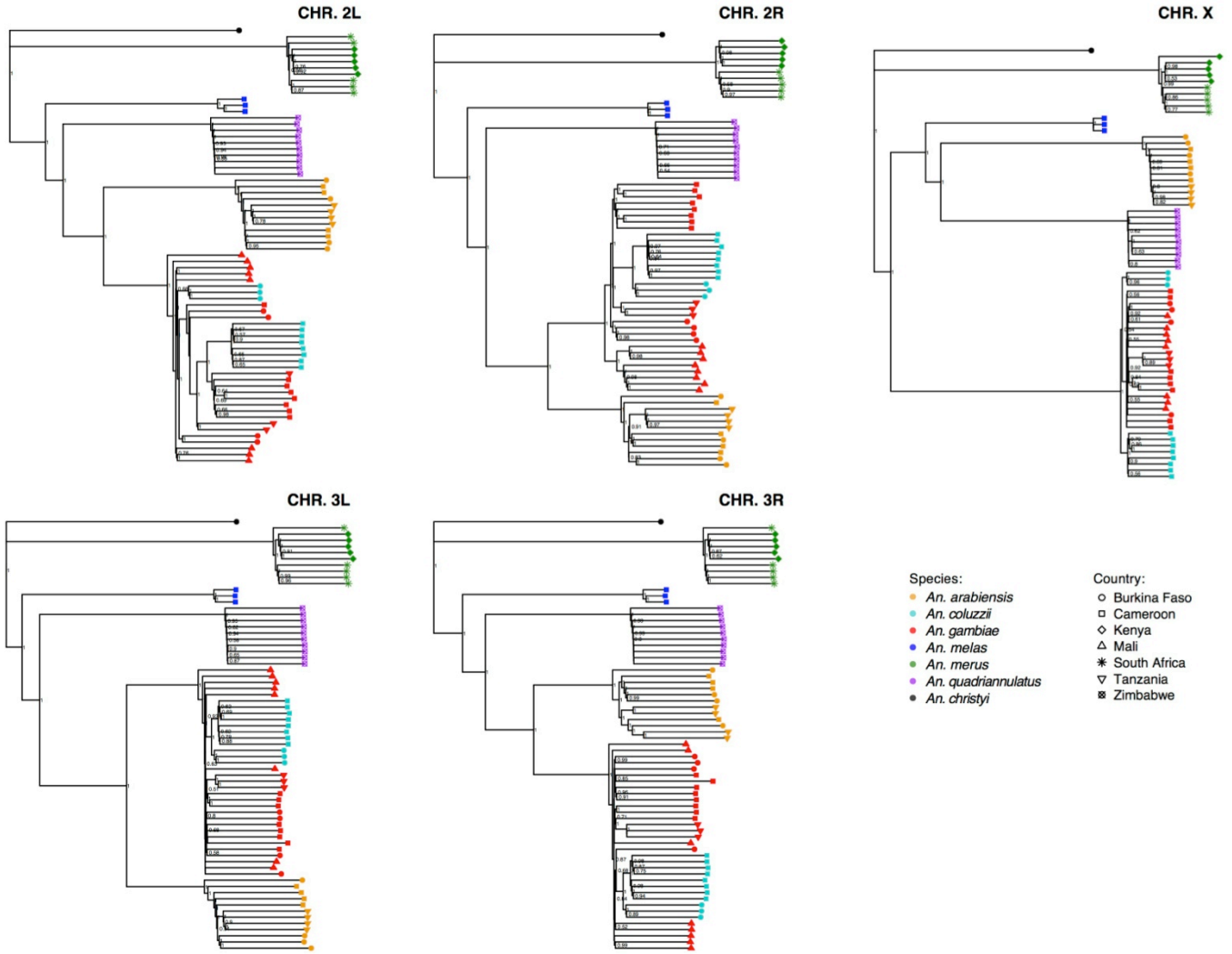


Fig. S20.
NJ tree displaying the Euclidian distance between individuals from population samples of each species, calculated from the SNP data of each chromosomal arm. Each species is represented by a distinct symbol color, and the symbol shape indicates the sampling location. Node support was evaluated by 100 bootstrap resampling replicates; only values greater than 0.5 are shown.

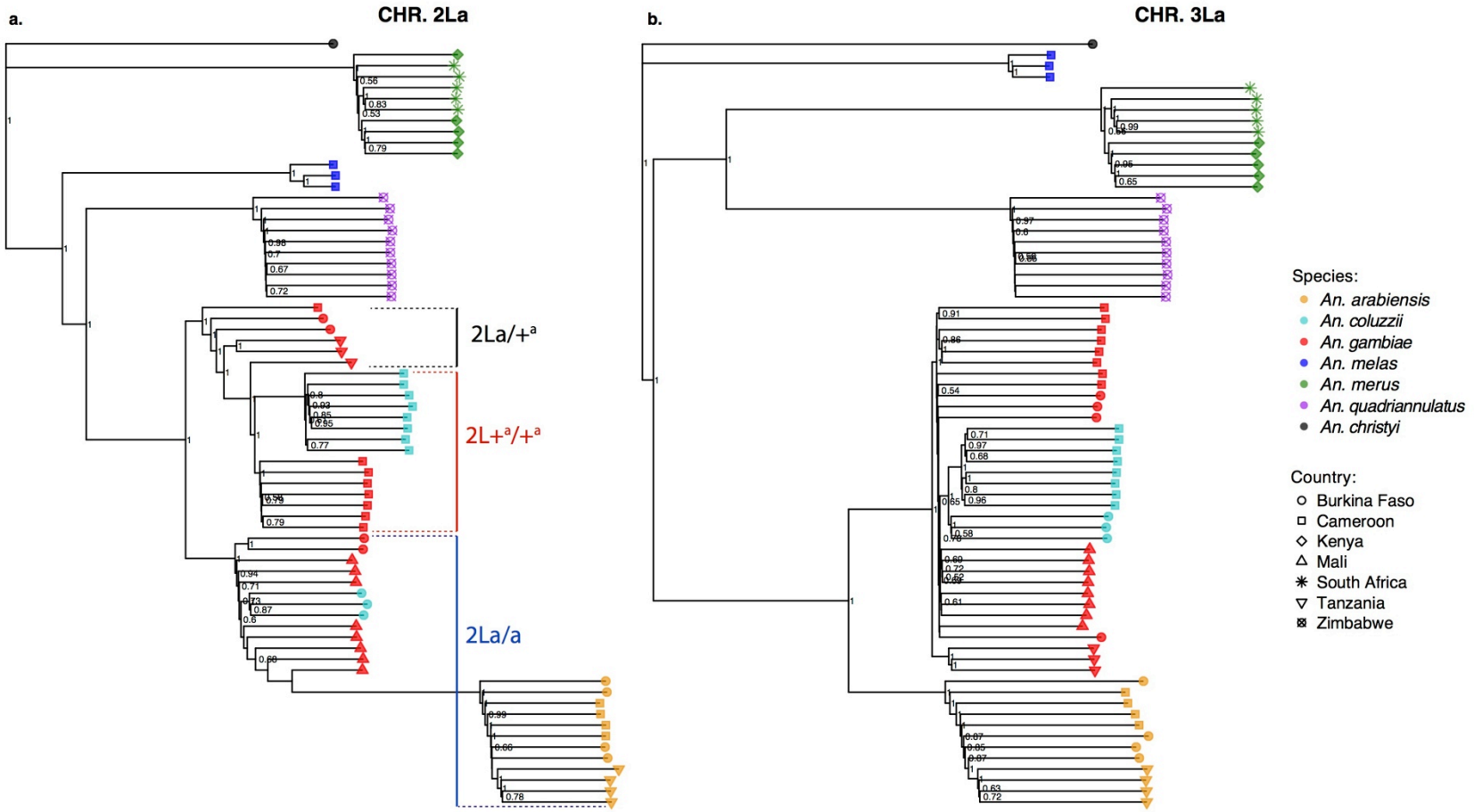


Fig. S21.
NJ trees displaying the Euclidian distance between individuals from population samples of each species, calculated from the SNP data for the SNPs found in the **(A)** 2La and **(B)** 3La inversions. Each species is represented by a distinct symbol color, and the symbol shape indicates the sampling location. Node support was evaluated by 100 bootstrap resampling replicates; only values greater than 0.5 are shown. The tree **(A)** shows that each *An. gambiae*, *An. coluzzii*, and *An. arabiensis* cluster according to their karyotype status for the 2La inversion.

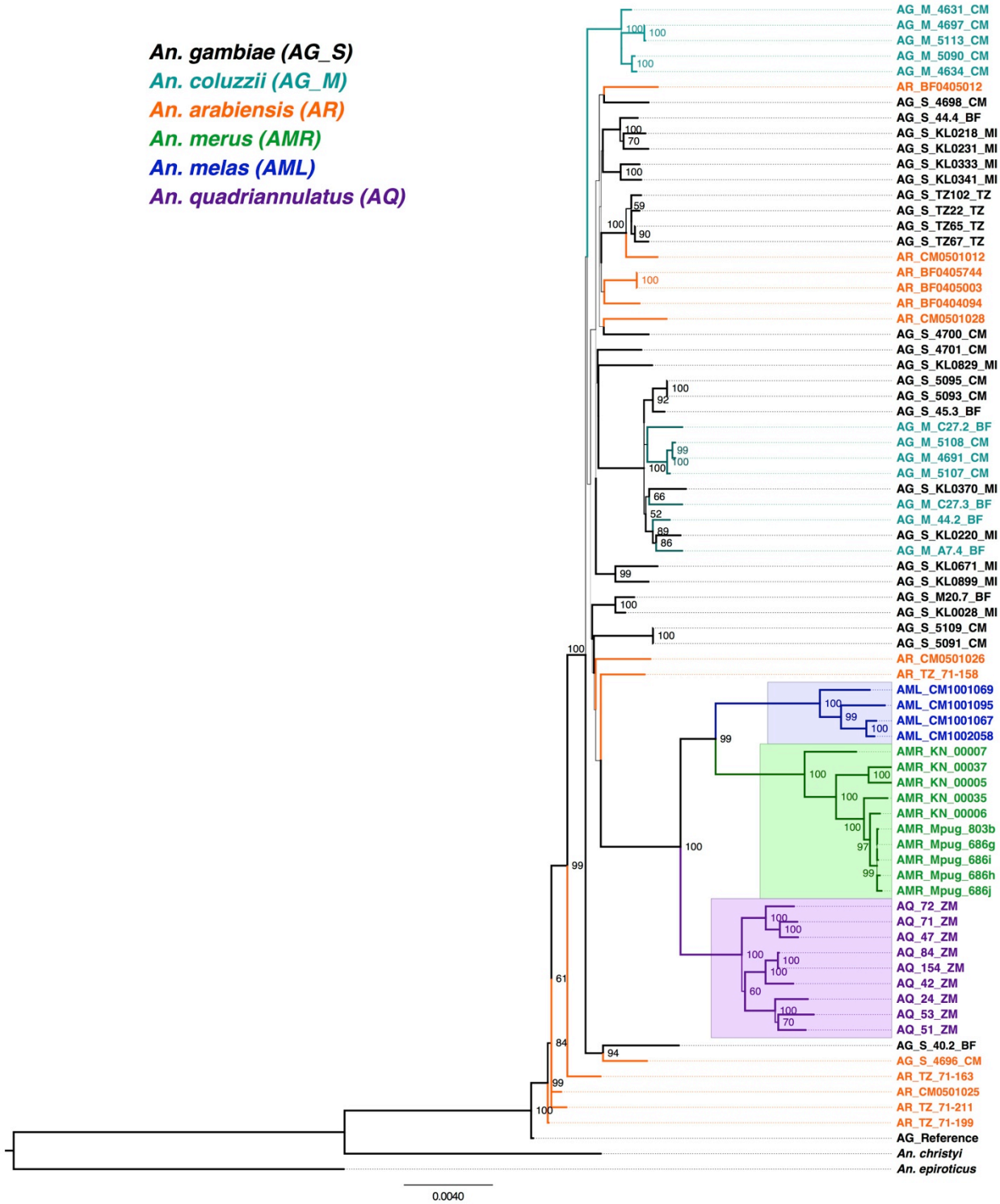


Fig. S22.
Maximum Likelihood phylogeny of the mitochondrial genomes of field-collected samples, estimated with RAxML. Each species is represented by a different color. Labels are color-coded by species and indicate sampling location (BF : Burkina Faso, CM : Cameroon, KN : Kenya, MI : Mali, Mpug : South Africa, TZ : Tanzania, ZM : Zimbabwe). Bootstrap support (%) is indicated at the nodes when it exceeded 50%.

S3.4. Higher molecular evolutionary rate of the X chromosome versus the autosomes

To avoid the potentially confounding effect on inferences of evolutionary rate caused by differential introgression of autosomes versus X chromosomes within the *An. gambiae* complex, we selected a species pair for which introgression is unlikely. *An. gambiae* PEST was chosen to represent the *An. gambiae* complex, as its assembly and annotation are the most complete, and we compared this species to the more distant Pyretophorus outgroup species, *An. epiroticus*. Single copy orthologs were identified using Orthodb (*93*) (http://orthodb.org). Protein sequence alignments were generated first using MUSCLE (*94*), and then used to inform CDS alignments with the codon-aware PAL2NAL alignment program (*95*). Ambiguous regions were removed from the CDS alignments using TrimAL (*86*), with the gap tolerance parameter set to 0.8 to exclude alignment columns with missing data in 20% or more of the sequences in the orthogroup. Sequences were removed from the CDS alignments if more than 40% of their length was gap characters following multiple alignment. PAML v4.7 (*96*) was used to calculate dN and dS values for each aligned ortholog pair (runmode = -2, codeml model=0, NSsites=0, ncatG=1).
To explore whether rates of evolution are different for autosomes vs the X chromosome, PAML estimated values of dN and dS were tested using the Wilcoxon rank sum test in R (version 3.0.3). The X chromosome shows significantly higher rates of both dN and dS than the autosomes ($p < 10^{-15}$ for each; Autosome mean dN=0.07, X chromosome mean dN=0.09; Autosome mean dS = 0.80, X chromosome mean dS= 1.05) as shown in Fig. S23.

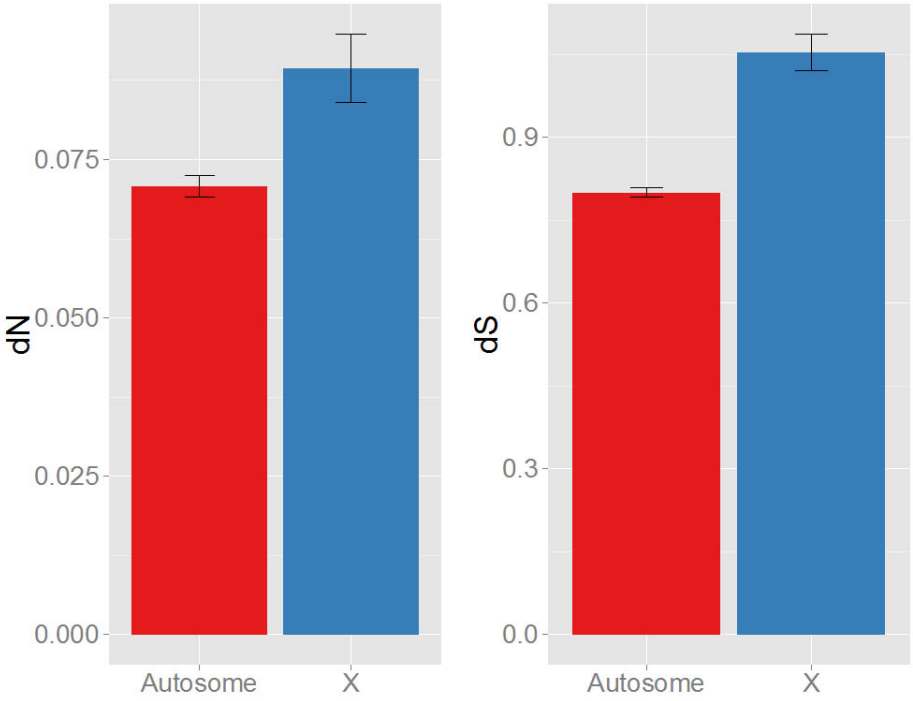


Fig. S23.
Autosome versus X chromosome rates of evolution. The mean rates of amino acid (dN) and silent site (dS) evolution between *An. gambiae* and *An. epiroticus* are plotted with 95% confidence interval bars. Rates of evolution of both site types are significantly faster on the X than on the autosomes (Wilcoxon rank sum test *p-value* < 10^{-15} for both site types).

S4. Formal tests of introgression between species

S4.1. Chromosomal patterns of introgression from *D* and D_{FOIL} statistics for field samples.

Disentangling incomplete lineage sorting from gene flow. In clades of closely related taxa, discordant genealogies due to incomplete lineage sorting (ILS) can complicate the detection of introgression. When four taxa are considered (three in-group and one out-group), the *D*-statistic (a.k.a. the "ABBA/BABA test") has been proposed to distinguish between ILS and introgression by testing for an imbalance in the relative frequency of the two discordant gene trees (*1, 4*). Given a major species phylogeny, ILS should produce the two minor topologies with equal frequency (*97-99*). However, introgression causes an imbalance toward a closer relationship between the two taxa exchanging alleles. Therefore, a statistically significant imbalance toward one discordant topology (indicated by the allele pattern `ABBA` or `BABA`) indicates that introgression has occurred.

Given site pattern counts (e.g., n_{ABBA}) of each site type in a given region of the sequence alignment, we can calculate the *D*-statistic as:

$$D = \frac{n_{\mathrm{ABBA}} - n_{\mathrm{BABA}}}{n_{\mathrm{ABBA}} + n_{\mathrm{BABA}}} \tag{3}$$

Therefore, the four-taxon *D*-statistic is a measure of inequality in the prevalence of site patterns that support the two possible discordant gene tree topologies. Note that this approach cannot detect introgression between sister species because such gene flow does not produce discordant topologies. It also does not tell us the direction of introgression; for this, we need to apply an extension of the *D*-statistic for five-taxon trees.

Pease and Hahn (*30*) proposed a five-taxon test to distinguish ILS from gene flow (the D_{FOIL} statistics). Unlike the standard *D*-statistic, their D_{FOIL} statistics can determine the direction of any detected introgression in a symmetric phylogeny (i.e., two sets of paired in-groups and a single out-group). D_{FOIL} consists of a system of four *D*-statistics to distinguish among the 16

possible introgressions in a symmetric five-taxon phylogeny. The four in-group taxa are labeled P_1-P_4, with the four D_{FOIL} statistics corresponding to: D_{FO} ("first"=P_1/P_3 *vs.* "outer" =P_1/P_4), D_{IL} ("inner"=P_2/P_3 *vs.* "last"=P_2/P_4), D_{FI} ("first" *vs.* "inner"), and D_{OL} ("outer" *vs.* "last"). Given counts of the site types corresponding to the presence of biallelic ancestral (A)/ derived (B) states, these statistics are defined as:

$$D_{FO} = \frac{\left(n_{\mathrm{BABAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{ABABA}} + n_{\mathrm{AAABA}}\right) - \left(n_{\mathrm{BAABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{ABBAA}} + n_{\mathrm{AABAA}}\right)}{\left(n_{\mathrm{BABAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{ABABA}} + n_{\mathrm{AAABA}}\right) + \left(n_{\mathrm{BAABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{ABBAA}} + n_{\mathrm{AABAA}}\right)} \quad (4)$$

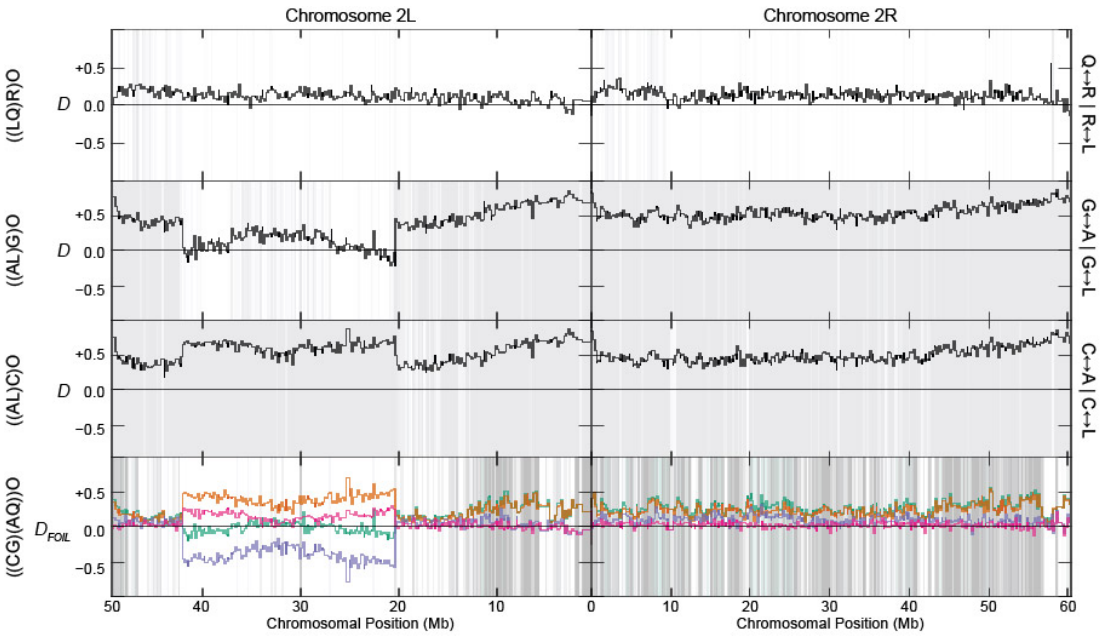
$$D_{IL} = \frac{\left(n_{\mathrm{ABBAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{BAABA}} + n_{\mathrm{AAABA}}\right) - \left(n_{\mathrm{ABABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{BABAA}} + n_{\mathrm{AABAA}}\right)}{\left(n_{\mathrm{ABBAA}} + n_{\mathrm{BBBAA}} + n_{\mathrm{BAABA}} + n_{\mathrm{AAABA}}\right) + \left(n_{\mathrm{ABABA}} + n_{\mathrm{BBABA}} + n_{\mathrm{BABAA}} + n_{\mathrm{AABAA}}\right)} \quad (5)$$

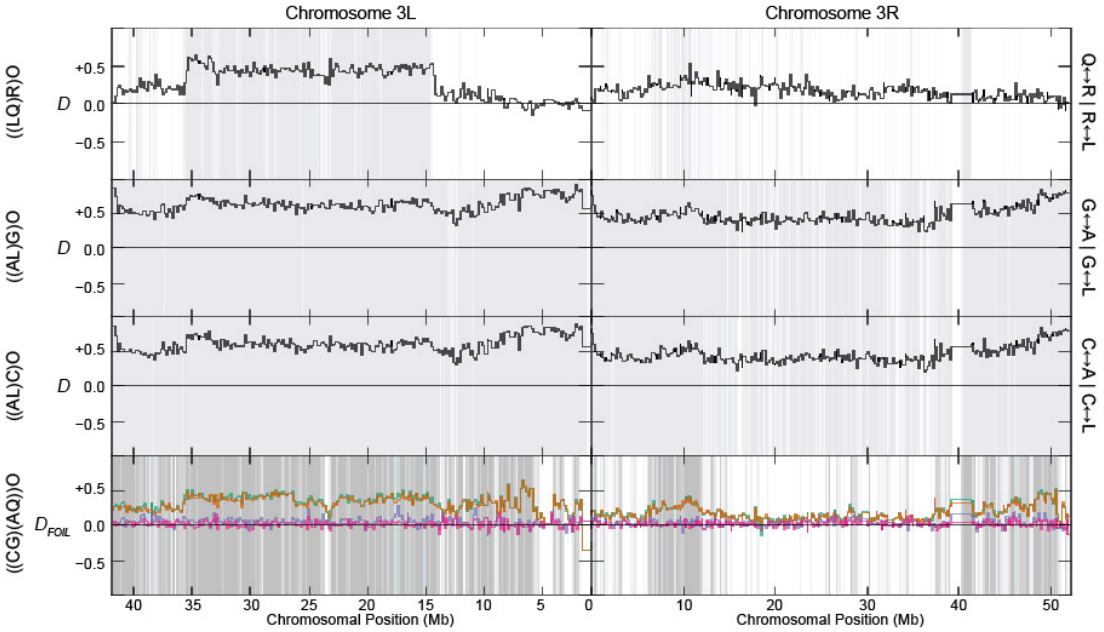
$$D_{FI} = \frac{\left(n_{\mathrm{BABAA}} + n_{\mathrm{BABBA}} + n_{\mathrm{ABABA}} + n_{\mathrm{ABAAA}}\right) - \left(n_{\mathrm{ABBAA}} + n_{\mathrm{ABBBA}} + n_{\mathrm{BAABA}} + n_{\mathrm{BAAAA}}\right)}{\left(n_{\mathrm{BABAA}} + n_{\mathrm{BABBA}} + n_{\mathrm{ABABA}} + n_{\mathrm{ABAAA}}\right) + \left(n_{\mathrm{ABBAA}} + n_{\mathrm{ABBBA}} + n_{\mathrm{BAABA}} + n_{\mathrm{BAAAA}}\right)} \quad (6)$$

$$D_{OL} = \frac{\left(n_{\mathrm{BAABA}} + n_{\mathrm{BABBA}} + n_{\mathrm{ABBAA}} + n_{\mathrm{ABAAA}}\right) - \left(n_{\mathrm{ABABA}} + n_{\mathrm{ABBBA}} + n_{\mathrm{BABAA}} + n_{\mathrm{BAAAA}}\right)}{\left(n_{\mathrm{BAABA}} + n_{\mathrm{BABBA}} + n_{\mathrm{ABBAA}} + n_{\mathrm{ABAAA}}\right) + \left(n_{\mathrm{ABABA}} + n_{\mathrm{ABBBA}} + n_{\mathrm{BABAA}} + n_{\mathrm{BAAAA}}\right)} \quad (7)$$

Testing for introgression. D/D_{FOIL} tests and plots were calculated for 50 kb windows using the DFOIL program (https://www.bitbucket.org/jbpease/dfoil). Windows with fewer than 5000 sites were excluded. Windows with less than a count of 10 for any of the *D*-statistic or D_{FOIL} components (i.e. site counts) were also excluded. Significance was determined by the binomial exact test as described in Pease and Hahn (*30*). The *D*-statistics and D_{FOIL} components were averaged over three consecutive windows for the plots. Biallelic site counts were calculated using custom Python scripts.

We applied the *D*-statistic to two different sets of three taxa and an outgroup (O; Fig. S24): *An. quadriannulatus* (Q), *An. melas* (L), and *An. merus* (R) (top row); *An. melas* (L), *An. arabiensis* (A), and *An. gambiae* (G) (second row); and *An. melas* (L), *An. arabiensis* (A), and *An. coluzzii* (C) (third row). We applied the D_{FOIL} statistics to the tree ((CG)(AQ))O (bottom row).


Chromosome 2L
Chromosome 2R
((LQ)R)O
((AL)G)O
((AL)C)O
((CG)(AQ))O
D
D_{FOIL}
+0.5
0.0
−0.5
Q↔R | R↔L
G↔A | G↔L
C↔A | C↔L
50 40 30 20 10 0
0 10 20 30 40 50 60
Chromosomal Position (Mb)


Chromosome 3L
Chromosome 3R
((LQ)R)O
((AL)G)O
((AL)C)O
((CG)(AQ))O
D
D_{FOIL}
+0.5
0.0
−0.5
Q↔R | R↔L
G↔A | G↔L
C↔A | C↔L
40 35 30 25 20 15 10 5 0
0 10 20 30 40 50
Chromosomal Position (Mb)

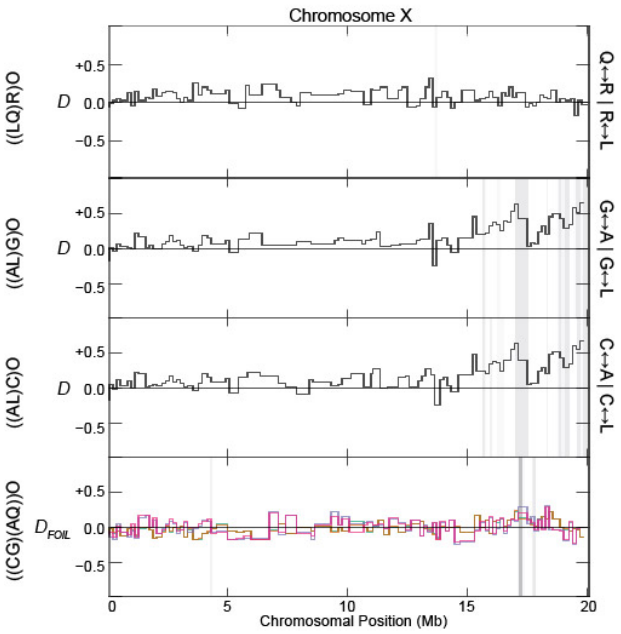

Chromosome X
((LQ)R)O
((AL)G)O
((AL)C)O
((CG)(AQ))O
D
D_{FOIL}
+0.5
0.0
−0.5
Q↔R | R↔L
G↔A | G↔L
C↔A | C↔L
0 5 10 15 20
Chromosomal Position (Mb)

Fig. S24.
Chromosomal plots of the *D*-statistic (black line) show the variation in signal of introgression across 50 kb non-overlapping genomic windows on all five chromosomal arms. Grey shading indicates regions of significant introgression ($P < 2\times 10^{-4}$, binomial exact test). For the group ((L,Q),R,)O (top row), generally $D>0$, indicating low levels of autosome-wide introgression between *An. quadriannulatus* (Q) and *An. merus* (R), when compared to *An. melas* (L) with outgroup (O) *An. christyi*. Introgression appears particularly strong or recent in the 3La inversion region (~15–35Mb) and on 3R (~5–15Mb). For the groups ((AL)G)O and ((AL)C)O (second and third rows), *An. arabiensis* (A) shows significant introgression with both *An. gambiae* (G) and *An. coluzzii* (C). Note also that *D* differs between these two in the 2La region (~20–41Mb) due to the difference in karyotype between G and C. The D_{FOIL} test (bottom row) infers the taxa involved in and direction of introgression from the combined signature of four tests: D_{FO} (green), D_{IL} (orange), D_{FI} (blue), and D_{OL} (magenta). The D_{FOIL} tests for the group ((CG)(AQ))O indicates that nearly all introgression detected is between A and G+C (grey shading). Therefore, introgression between A and the gambiae group is inferred to have occurred either prior to the split of G and C, or with approximately equal frequency with both species after their split. Again, the pattern in 2La is exceptional in these plots due to the difference in karyotype.

S4.2 Geographic pattern of introgression

For the species between which introgression was detected based on analyses of the reference assemblies and single field-collected individuals, we tested whether geographic variation in introgression occurred in a consistent way across population samples. The rationale underlying this test is that if the signal of introgression can be detected in all the sampled populations, this could indicate relatively old introgression in one population that spread to others, or pervasive introgression in multiple populations. In contrast, an introgression signal only detected in some populations of the focal species could reveal more recent or geographically restricted events. This is similar to the rationale followed to demonstrate that Neanderthal introgression into human populations only occurred in non-African populations and could be consistent with a single episode of admixture from Neanderthals into the ancestors of all non-Africans when the two groups coexisted in the Middle East 50–80 Kyr (*1, 100*).

To conduct this analysis, we used the SNP dataset based on multiple individual mosquito genomes sampled in the field from various locations for each of the 6 species (Fig. S6). We used the four-taxon *D*-statistic (described in section S4.1) extended to multiple individuals per group as implemented in ADMIXTOOLS v.1.1 (*101*). As described in Patterson *et al.* (*101*) and Wall *et al.* (*100*), diploid genomes can be subject to this test by using the frequencies (instead of the count) of the derived allele in each group considered. Let W, X, Y, Z be four taxa with a phylogeny ((W,X),Y),Z), with W and X being the focal species, Y the tested species potentially introgressing with W or X, and Z the outgroup used to polarize variants as ancestral (A) or derived (B). Only those sites with configurations ABBA and BABA are used, where the order is W, X, Y, Z. The requirement that two copies of both the derived and the ancestral alleles be present greatly reduces the effect of sequencing error (*4*).

When only a single sequence from each population is available,

$$D = \frac{n_{\mathrm{BABA}} - n_{\mathrm{ABBA}}}{n_{\mathrm{ABBA}} + n_{\mathrm{BABA}}} \tag{8}$$

where n_{ABBA} and n_{BABA} are the numbers of sites with each of the two configurations. Note that, compared to Eq. 3 in supplementary text S4.1, n_{BABA} and n_{ABBA} were switched in the numerator to align with the ADMIXTOOLS implementation of the *D*-statistic. Following Patterson *et al.* (*101*), when diploid sequences from each taxon W, X, Y, Z are available, the numerator for the SNP *i* (Num_i) and denominator (Den_i) equal:

$$Num_i = P(BABA) - P(ABBA) = (w' - x')(y' - z') \quad (9)$$

$$Den_i = P(BABA) + P(ABBA) = (w' + x' - 2w'x')(y' + z' - 2y'z') \quad (10)$$

where w', x', y', z' are variant population allele frequencies in the taxon W, X, Y, Z for a SNP i. Using the sample allelic frequency, the *D*-statistic D(W,X;Y,Z) can be defined as

$$D = \frac{\sum_i \hat{N}um_i}{\sum_i \hat{D}en_i} \quad (11)$$

ADMIXTOOLS computes a standard error for *D* using the weighted block jackknife (*101*). The number of standard errors by which *D* departs from zero reflects the Z-score, which is approximately normally distributed and thus yields a formal test of introgression. An absolute Z-score value greater than 3 is considered as significant evidence of introgression.

The triplet of species (W, X, Y) considered in this analysis use the following code: A=*An. arabiensis,* C=*An. coluzzii,* G=*An. gambiae*, L=*An. melas*, Q=*An. quadriannulatus*, R=*An. merus*. We used two outgroup species to polarize the variants: *An. christyi* and *An. epiroticus*. Under the ADMIXTOOLS implementation of the *D*-statistic, $D = 0$ means no introgression, $D > 0$ and abs(Z) > 3 means introgression between W–Y, $D < 0$ and abs(Z) > 3 means introgression between X–Z.

Validation of interspecific introgression. We conducted the analysis first at the species level, considering all individuals available in each species in order to validate the findings based on the analyses of one single genome per species from the reference assemblies and from the single genome field samples (see S4.1). This analysis confirmed the previous findings: the

introgression signal was highly significant between A–G in the QAG test and A–C in the QAC test for all the chromosomes except the X and 2L (Fig. S25). The absence of an introgression signal on 2L results from the fact that the samples in G and C are polymorphic for the 2La and 2L+^{a} inversion (see below and section S3.2 and S.3; Fig. S21). Significant Z-scores were also found between Q–G (in the QLG test) and Q–C (in the QLC test), but these signals are side-effects of the massive introgression between A–C and A–G as shown using the D_{FOIL} test applied to single genome analyses (see S4.1). Finally, significant introgression was also found between Q–R in QLR and QAR tests, in agreement with previous results. The signal was especially strong on chromosome 3L, as a result of the introgression of the ~22 Mb 3La inversion from R to Q (polarity implied by the topology of the 3La inversion, Fig. S21). All the other significant Q–R *D* values were very close to 0 with Z-scores close to the threshold of 3 (Fig. S25). This test also confirmed that the X chromosome did not display evidence of introgression in any of the comparisons.

Geographic variation in introgression. As distinct populations were sampled in A, G, C, and R for different parts of their distribution (Fig. S6), we tested whether the signal of introgression differed according to the set of populations considered in the test (Table S10).

The two populations of R from Kenya and Mozambique did not display any distinct pattern of introgression when considered as focal groups (W,X) with Q as Y, or when considering them in turn as the Y test taxon (lines 1 to 3, Table S10).

We then tested introgression between C and G populations (lines 4–21, Table S10). The most important signal of introgression was observed on chromosome 2L, relating to the inversion karyotype composition in each population, with populations (W or X) sharing the same inversion status as Y implicated in introgression. Introgression was also detected between all C populations and G_{TZ} (*An. gambiae* sampled from Tanzania; lines 4–8). The signal was present on 2R, 3L and 3R arms, and was especially strong on the X chromosomes. Some introgression was also detectable between local populations (lines 9 – 11).

When testing whether G or C were introgressing with A, all tests (lines 15–21) showed significant introgression on all autosomes except 2L (due to the 2La inversion). We further refined this analysis by testing whether a specific population of G had stronger evidence of introgression with A (lines 22–30), and found this to be the case for G_{TZ} and any A populations, even when considering the two A populations from West Africa. We tested whether a particular A population could exchange more derived alleles with G (or C), but all the *D* values (lines 31 to 40) were very close to 0 and the Z scores were under or near the threshold limits of 3. These results may suggest that recent introgression between A–G occurred in East-Africa (G_{TZ}), and that the other G populations sampled in West Africa displays a more ancient signal of introgression. However a more intensive sampling is required to make any definitive conclusions.

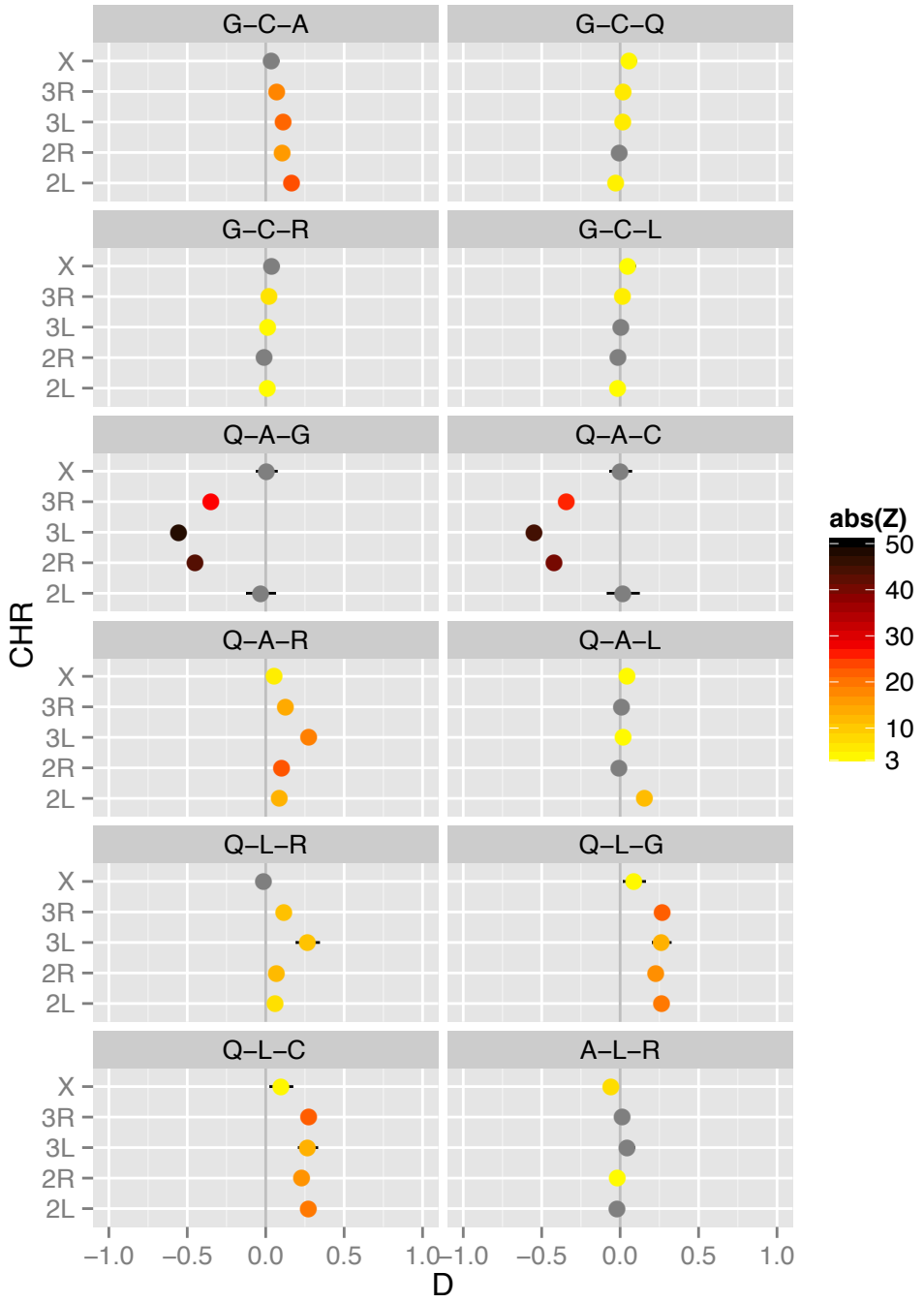


Fig. S25.
Introgression *D*-statistic computed per chromosome at the species level, considering all the individuals within each species. The Z-score, computed by block-jackknife, provides the statistical significance. An absolute Z value greater than 3 is considered significant (grey points indicate non-significant Z-values). The whiskers show *D* ± 3SE. The triplet of species considered is indicated by the code above each plot (W, X, Y), where *An. gambiae* = G, *An. coluzzii* = C, *An. arabiensis* = A, *An. quadriannulatus* = Q, *An. merus* = R, and *An. melas* = L. *D*=0, no introgression; *D*>0 and abs(Z) >3, introgression between W–Y; *D*<0 and abs(Z)>3, introgression between X–Z.

Table S10.
Introgression *D*-statistic test for triplets of populations within- and between-species. The Z-score provides the statistical significance level with absolute values greater than 3 being statistically significant (in bold). *D*>0 and Z>3 = W–Y; *D*<0 and Z<-3 = X–Y. Bold values have a *D*-statistic greater than 0.1 or lower than 0.1 indicating a substantial level of introgression. The 2La karyotype was determined by molecular diagnostic assay (S2.1.1).

	Groups (((WX)Y)O)						2L		2R		3L		3R		X	
	W	**W-2La**	**X**	**X-2La**	**Y**	**Y-2La**	**D**	**Z**	**D**	**Z**	**D**	**Z**	**D**	**Z**	**D**	**Z**
1	R_{KN}	+	R_{MZ}	+	Q_{ZM}	+	-0.01	-2.11	0.00	0.90	0.01	1.55	0.00	-0.49	0.01	0.84
2	Q_{ZM}	+	A_{TZ}	a	R_{KN}	+	**0.09**	**13.64**	**0.11**	**24.22**	**0.28**	**18.53**	**0.13**	**14.77**	0.06	5.35
3	Q_{ZM}	+	A_{TZ}	a	R_{MZ}	+	**0.09**	**13.50**	**0.11**	**24.23**	**0.28**	**18.87**	**0.13**	**14.74**	0.06	5.50
4	G_{MI}	a	G_{TZ}	+	C_{BF}	a	**0.17**	**6.87**	-0.08	-8.59	-0.01	-1.61	-0.01	-1.07	**-0.15**	**-7.80**
5	G_{BF}	a/+	G_{TZ}	+	C_{BF}	a	0.09	5.16	-0.04	-4.97	-0.02	-3.52	-0.02	-2.33	**-0.17**	**-8.48**
6	G_{CM}	+	G_{TZ}	+	C_{CM}	+	0.03	2.85	-0.06	-9.22	-0.05	-7.10	-0.05	-6.64	**-0.21**	**-11.30**
7	G_{BF}	a/+	G_{TZ}	+	C_{CM}	+	**-0.20**	**-13.36**	**-0.11**	**-8.59**	-0.02	-2.91	-0.02	-2.58	**-0.17**	**-8.06**
8	G_{CM}	+	G_{TZ}	+	C_{BF}	a	**-0.14**	**-11.70**	-0.07	-11.82	-0.06	-8.65	-0.05	-7.49	**-0.22**	**-12.54**
9	G_{BF}	a/+	G_{CM}	+	C_{BF}	a	**0.20**	**9.19**	0.03	8.09	0.03	7.18	0.04	11.33	0.05	3.75
10	G_{BF}	a/+	G_{CM}	+	C_{CM}	+	**-0.20**	**-9.78**	-0.04	-3.20	0.02	6.43	0.03	7.31	0.04	3.70
11	C_{BF}	a	C_{CM}	+	G_{BF}	a/+	**0.10**	**9.22**	**0.07**	**6.41**	0.01	1.68	**0.01**	**3.13**	0.03	1.67
12	C_{BF}	a	C_{CM}	+	G_{CM}	+	**-0.26**	**-8.69**	-0.02	-2.04	-0.01	-1.11	0.00	0.82	0.02	1.05
13	C_{BF}	a	C_{CM}	+	G_{TZ}	+	**-0.15**	**-7.67**	-0.01	-1.32	0.01	1.15	0.01	2.90	0.03	1.41
14	G_{BF}	a/+	C_{BF}	a	A_{BF}	a	-0.06	-2.96	**0.11**	**17.17**	0.08	11.43	0.03	7.16	-0.02	-0.91
15	G_{CM}	+	C_{CM}	+	A_{CM}	a	0.08	12.50	**0.12**	**16.96**	0.09	14.07	0.05	11.56	0.05	2.82
16	G_{TZ}	+	C_{BF}	a	A_{BF}	a	**-0.11**	**-3.65**	**0.15**	**13.50**	**0.17**	**18.00**	**0.13**	**11.37**	-0.01	-0.27
17	G_{TZ}	+	C_{CM}	+	A_{CM}	a	**0.27**	**22.91**	**0.23**	**28.95**	**0.22**	**23.43**	**0.17**	**16.29**	0.10	4.50
18	G_{TZ}	+	C_{BF}	a	A_{TZ}	a	**-0.11**	**-3.60**	**0.16**	**14.85**	**0.17**	**17.33**	**0.13**	**11.66**	-0.01	-0.16
19	G_{TZ}	+	C_{CM}	+	A_{TZ}	a	**0.27**	**23.37**	**0.23**	**28.29**	**0.22**	**23.33**	**0.18**	**17.14**	0.10	4.44
20	G_{BF}	a/+	C_{CM}	+	A_{BF}	a	**0.28**	**16.57**	**0.17**	**21.88**	**0.12**	**19.93**	**0.08**	**14.59**	0.08	4.36
21	G_{CM}	+	C_{BF}	a	A_{CM}	a	**-0.22**	**-7.32**	0.05	5.44	0.05	8.37	0.01	3.23	-0.04	-1.79

Table S10 (Continued).

	Groups (((WX)Y)O)						2L		2R		3L		3R		X	
	W	**W-2La**	**X**	**X-2La**	**Y**	**Y-2La**	**D**	**Z**	**D**	**Z**	**D**	**Z**	**D**	**Z**	**D**	**Z**
22	G_{MI}	a	G_{TZ}	+	A_{BF}	a	**0.22**	**7.39**	**-0.12**	**-13.07**	**-0.10**	**-12.23**	-0.09	-9.25	0.02	0.94
23	G_{BF}	a/+	G_{TZ}	+	A_{BF}	a	0.09	5.38	-0.03	-2.57	**-0.09**	**-10.92**	-0.09	-8.67	0.02	0.60
24	G_{CM}	+	G_{TZ}	+	A_{CM}	a	**-0.19**	**-15.58**	**-0.09**	**-14.42**	**-0.12**	**-16.22**	**-0.11**	**-10.94**	-0.03	-1.15
25	G_{TZ}	+	G_{CM}	+	A_{TZ}	a	**0.19**	**15.85**	0.09	13.91	**0.12**	**15.98**	**0.11**	**11.14**	0.03	1.42
26	G_{TZ}	+	G_{BF}	a/+	A_{TZ}	a	-0.09	-5.15	0.04	3.66	**0.10**	**11.59**	**0.10**	**9.19**	-0.02	-0.60
27	G_{TZ}	+	G_{MI}	a	A_{TZ}	a	**-0.22**	**-7.29**	**0.12**	**13.35**	**0.11**	**13.13**	**0.10**	**9.88**	-0.02	-1.09
28	G_{BF}	a/+	G_{CM}	+	A_{BF}	a	**0.24**	**11.86**	0.06	7.27	0.03	6.77	0.02	5.89	0.02	1.13
29	G_{BF}	a/+	G_{CM}	+	A_{CM}	a	**0.23**	**11.80**	0.07	7.16	0.03	7.25	0.02	6.02	0.02	0.95
30	G_{BF}	a/+	G_{CM}	+	A_{TZ}	a	**0.24**	**11.85**	0.05	7.87	0.03	7.70	0.02	5.29	0.02	1.18
31	A_{BF}	a	A_{CM}	a	G_{BF}	a/+	0.00	0.76	-0.01	-3.72	-0.01	-1.37	-0.01	-1.48	0.04	2.99
32	A_{BF}	a	A_{CM}	a	C_{BF}	a	0.00	0.76	-0.01	-1.77	0.00	-0.36	0.00	-0.42	0.04	3.12
33	A_{BF}	a	A_{CM}	a	G_{CM}	+	0.01	1.74	0.00	-1.47	0.00	-0.68	0.00	-0.40	0.04	2.98
34	A_{BF}	a	A_{CM}	a	C_{CM}	+	0.01	1.94	0.00	-0.28	0.00	-0.42	0.00	-0.95	0.04	2.70
35	A_{CM}	a	A_{TZ}	a	G_{CM}	+	0.01	3.27	0.03	3.62	0.01	0.99	0.01	3.08	0.04	2.66
36	A_{CM}	**a**	A_{TZ}	**a**	C_{CM}	**+**	0.01	3.04	0.03	3.62	0.01	0.94	0.01	2.90	0.03	2.26
37	A_{BF}	**a**	A_{TZ}	**a**	G_{BF}	**a/+**	0.01	3.19	0.03	4.43	0.00	0.22	0.01	2.26	0.06	4.88
38	A_{BF}	**a**	A_{TZ}	**a**	C_{BF}	**a**	0.01	2.23	0.03	4.50	0.00	0.63	0.01	1.84	0.06	4.41
39	A_{BF}	**a**	A_{TZ}	**a**	G_{TZ}	**+**	0.01	3.12	0.02	2.75	-0.01	-0.77	0.00	-0.07	0.06	4.86
40	A_{CM}	**a**	A_{TZ}	**a**	G_{TZ}	**+**	0.01	2.43	0.03	3.47	0.00	0.76	0.01	1.31	0.03	2.27
						# SNP	1,744,882		1,941,790		1,329,351		1,626,846		888,901	
						# Block	99		123		84		107		49	

S5. Chromosomal inversion phylogeny of the *An. gambiae* complex

Based on the banding pattern of polytene chromosomes, 10 fixed inversions in the *An. gambiae* complex have been observed (*31*), of which five are on the X chromosome. Based on these five fixed X inversions, the *An. gambiae* complex can be divided into three groups: 1) *An. merus* and *An. gambiae*, which share the compound Xag inversion; 2) *An. quadriannulatus*, *An. bwambae* and *An. melas*, which carry the arbitrary standard X arrangement ; and 3) *An. arabiensis,* which carries the compound Xbcd inversion. We aimed to A) identify the genomic coordinates for breakpoints of fixed inversions, using the newly available *Anopheles* genome assemblies (*18, 52*); B) determine the ancestral and derived arrangements for fixed chromosomal inversions in the *An. gambiae* complex based on comparisons to outgroup species; C) reconstruct phylogenetic relationships within the *An. gambiae* complex using fixed inversions as markers; D) estimate the divergence time in the *An. gambiae* complex using rates of X chromosome evolution estimated for the genus (*18*).

S5.1. Genomic coordinates for breakpoints of fixed inversions

Ortholog information was retrieved from OrthoDB (*93*). The gene IDs of *An. gambiae* were retrieved based on GFF3 annotation from VectorBase (http://www.vectorbase.org/). The *Anopheles* species ortholog groups IDs for all genes were identified using OrthoDB. Within the same species, any two or more genes that share the same ortholog group ID were removed for further analysis, thus one-to-one gene ortholog pairs between all species were identified. The genomic coordinates within scaffolds for one-to-one orthologs for each species were then obtained from the GFF3 annotation file downloaded from VectorBase.

Using the relationships between genes and their position on scaffolds, tab-format files for the R program genoPlotR (*102*) were generated. The relationships between these genes and their order on scaffolds were visualized in genoPlotR. Visualizing the results and comparing them with the cytogenetic (*31*) and physical (*103*) maps identified breakpoints of fixed inversions. To

narrow down the breakpoint positions, the sequences between the neighboring genes of breakpoint regions were extracted using BEDtools "fastaFromBed" (*104*) and compared between species. To assemble a chromosome from scaffolds, the cytogenetic (*31*) and physical (*103*) maps of *An. gambiae* were used as a guide to concatenate long scaffolds in a species-specific order.

Chromosomal positions of the 10 known fixed inversions were taken from the *An. gambiae* cytogenetic map (*31*). The approximate coordinates of the breakpoints in the PEST AgamP3 genome assembly were identified using a physical map of *An. gambiae* polytene chromosomes (*103*). These coordinates were used as a guide to find precise coordinates of fixed inversion breakpoints in genomic scaffolds of species from the *An. gambiae* complex with reference assemblies. Several breakpoints could not be found in some highly fragmented genome assemblies: one of two breakpoints for Xag in *An. quadriannulatus*, all Xag breakpoints in *An. merus*, the 2Rp breakpoints in *An. merus,* and the 2Rm and 3La breakpoints in *An. melas.* There are two fixed inversions between *An. gambiae* and *An. quadriannulatus* on the X chromosome (Xa and Xg), three fixed inversions between the *An. quadriannulatus* and *An. arabiensis* X chromosome (Xb, Xc, and Xd), and five fixed inversions between the *An. gambiae* and *An. arabiensis* X chromosome (Xa, Xg, Xb, Xc, and Xd) (*31*). All five inversions were found when we aligned five concatenated *An. arabiensis* scaffolds to the *An. gambiae* X chromosome assembly (Fig. S26).

We identified 42 genomic scaffolds in *An. quadriannulatus* that belong to the X chromosome based on synteny. By comparing genome assemblies of *An. gambiae* and *An. quadriannulatus*, we found only one ~2 Mb-long scaffold (KB667689) in *An. quadriannulatus* that has the Xag breakpoints. We demonstrated that cytologically identical distal Xa and Xg breakpoints have different genomic positions, they are separated by ~329 kb in *An. quadriannulatus,* by ~211 kb in *An. gambiae,* and by~153 kb in *An. arabiensis.* Therefore, we found no evidence for breakpoint reuse of distal Xa and Xg breakpoints in *An.*

quadriannulatus and *An. arabiensis.* However, we do see evidence for breakpoint reuse of breakpoints Xb and Xd in *An. arabiensis* (Fig. S26). The genomic localization of breakpoints for the fixed inversions on the X chromosome and autosomes is provided in S11.

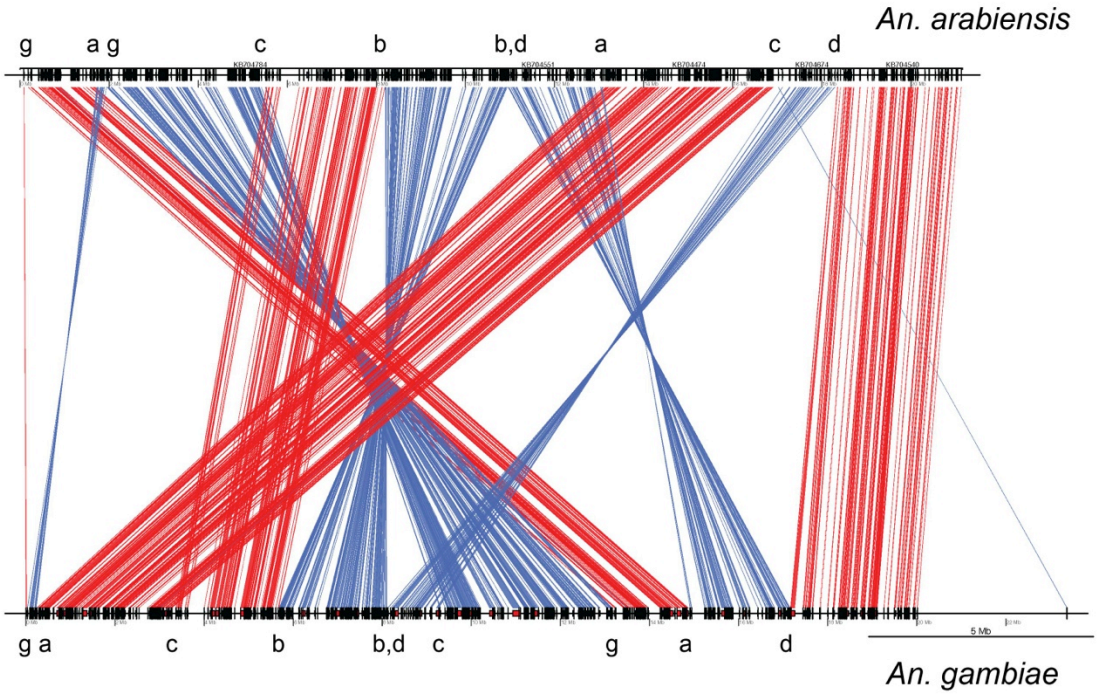


Fig. S26.
Concatenated X chromosome scaffolds of *An. arabiensis* (top) aligned to the *An. gambiae* X chromosome (bottom). The breakpoints are shown with small letters. The chromosomes are oriented with telomeres on the left and centromeres on the right.

Table S11.
Coordinates of breakpoints of 10 fixed inversions in the *An. gambiae* genome assembly.

Bp*	Flanking gene 1 id	Gene 1 coord A	Gene 1 coord B	Flanking gene 2 id	Gene 2 coord A	Gene 2 coord B	Bp* coord A	Bp* coord B
Xg	AGAP000002	582	16387	AGAP000005	32382	38843	21645	27311
Xa	AGAP000017	212901	233685	AGAP000018	245976	251810	238265	240688
Xc	AGAP000201	3279532	3280615	AGAP000203	3364642	3371951	3342563	3355861
Xb	AGAP000319	5662879	5676527	AGAP000320	5679568	5695844	5676527	5679568
Xbd	AGAP000469	8098317	8102973	AGAP000470	8105206	8110152	8102973	8105206
Xc	AGAP000519	9215504	9266532	AGAP000520	9291162	9297115	9281318	9288838
Xg	AGAP000724	13154935	13156918	AGAP000726	13162864	13169057	13160765	13162864
Xa	AGAP000809	14838271	14839995	AGAP000813	14912127	14913229	14839995	14912127
Xd	AGAP000891	16768557	16772936	AGAP000892	16792766	16794816	No data	No data
2Ro	AGAP001760	9483432	9485151	AGAP001762	9488303	9493744	9485167	9486712
2Rp	AGAP013533	13138062	13145420	AGAP001984	13150830	13154837	No data	No data
2Ro	AGAP002933	29835569	29838048	AGAP002935	29839372	29840708	29838366	29839163
2Rp	AGAP003327	35998660	36020065	AGAP003328	36027561	36028502	No data	No data
2La	AGAP005778	20521765	20523605	AGAP005779	20528560	20529407	20524058	20528089
2La	AGAP007068	42163507	42164602	AGAP007069	42165842	42176356	42165182	42165532

*breakpoint

S5.2. Ancestral and derived genome arrangements

To reconstruct the ancestral karyotype of the *An. gambiae* complex, the gene orders at the breakpoints were analyzed using species outside the *An. gambiae* complex. If the gene order at the breakpoint is the same as in outgroups, then this gene order was considered ancestral in the *An. gambiae* complex. In order to confirm these results at the nucleotide level, the sequences at breakpoints were retrieved and subjected to BLASTn analysis against the outgroup species with a $1e^{-5}$ e-value cut-off. We determined gene orders at the breakpoints of the fixed inversions in 12 outgroup species: *An. albimanus, An. atroparvus, An. christyi, An. culicifacies, An. dirus, An. epiroticus, An. farauti, An. funestus, An. maculatus, An. minimus, An. sinensis,* and *An. stephensi*. Table S11 demonstrates the gene orders at the breakpoints in the X chromosome inversions. The data show that all arrangements in the *An. gambiae* X chromosome are ancestral. The Xag chromosome arrangement is shared by *An. gambiae* and *An. merus* (*31*), but the *An. merus* assembly is too fragmented to find breakpoints on the X chromosome. No supporting information was found for *An. arabiensis* having ancestral X chromosome arrangements. We found that the X+ arrangement typical to *An. quadriannulatus* and the Xbcd arrangements found only in *An. arabiensis* are derived. Table S11 shows the gene orders at the breakpoints in the autosomal inversions in ingroup and outgroup species. We confirm the previous finding that the 2Ro, 2La, and $2R+^{p}$ are ancestral autosomal arrangements, while the 2Rp and $2L+^{a}$ are derived arrangements (*105*). We also found that the $2R+^{b}$ arrangement of the polymorphic inversions 2Rb/+ is ancestral. We conclude that the genome arrangements of *An. gambiae* and *An. merus* closely resemble the ancestral karyotype.

S5.3. Rooted inversion phylogeny of the *An. gambiae* complex

Our task was to generate a tree using ancestral chromosomal arrangements (Xa, Xg, 2Ro, $2R+^{p}$, 2La) and derived chromosomal arrangements (Xb, Xc, Xd, $X+^{ag}$, $2R+^{o}$, 2Rp, $2L+^{a}$). Because the genome assembly for *An. melas* is too fragmented, we could not identify

breakpoints for 2Rm and 3La. Therefore, we did not include *An. melas* in the phylogeny. The calculation of inversion distances among the included species of the *An. gambiae* complex and the outgroup species was performed using the *Multiple Genome Rearrangement (*MGR) program available at http://grimm.ucsd.edu/MGR/. The signed option of the MGR program was used. This program implements an algorithm that uses a parsimony approach, *i.e*. it minimizes the sum of the rearrangements over all the edges of the phylogenetic tree (*106*). To create an inversion phylogenetic tree, numbers were assigned to represent each conserved synteny block in ingroup and outgroup species using our breakpoint data.

Figure S27A shows one result, a parsimony tree that allows no independent fixations of inversions in different lineages. This tree does not require but allows introgression of 2La from *An. gambiae* to *An. arabiensis.* Figure S27B shows an alternative result, a parsimony tree that requires independent fixations of $2R+^{o}$ in two lineages and introgression of 2La from *An. gambiae* to *An. arabiensis*. This tree topology is identical to that of the species tree inferred from the X chromosome (Fig. 1B, main text), and it assumes an ancient inversion polymorphism of the 2La and 2Ro inversions that likely predate speciation in the complex. While the $2R+^{o}$ inversion became fixed in *An. gambiae* and other species of the complex, the 2La inversion still remains polymorphic in *An. gambiae*.

These inversion trees consider *An. merus* and *An. gambiae* as species descended from the earliest branching events, in agreement with the species tree inferred from the X chromosome. The ambiguity of placing either species as sister to the *An. quadriannulatus* + *An. arabiensis* clade is consistent with the detection of incomplete lineage sorting (ILS) at the basal node of the species tree (main text). The majority trees supported by autosomal genomic sequences strongly conflict with the inversion phylogeny. To reconcile these contrasting topologies, it is necessary to postulate a long-term ancestral polymorphism of 2Ro/+, 2La/+, and Xag/+, as well as independent fixations of $2R+^{o}$, $2L+^{a}$, Xag, and $X+^{ag}$ in at least two lineages. Only two polymorphic inversions have been found on the X chromosome in the *An. gambiae* complex (*31*). A comprehensive survey of inversion polymorphisms detected in thousands of *An. gambiae* population samples spanning multiple decades and geographic locations in Africa detected at least 82 polymorphic inversions on autosomes and none on the X chromosome *(107)*, suggesting that inversion polymorphisms on the X chromosome are relatively rare. More important, placing *An. gambiae* and *An. arabiensis* as sister taxa as the majority autosomal topologies predict is problematic because it implies that Xbcd originated directly from Xag. In reality, the Xbcd inversion of *An. arabiensis* and the Xag inversion of *An. gambiae* and *An. merus* are complex rearrangements that differ by 5 overlapping inversions; Xbcd can only arise from $X+^{ag}$, not directly from Xag. In conclusion, our independent phylogenetic approach rejects the majority autosomal topology and produces an inversion phylogeny consistent with the species phylogeny determined from X chromosome sequences.

S5.4. Dating the initial radiation of the *An. gambiae* complex

We calculated the rate of the X chromosome rearrangement in the genus *Anopheles* using chromosome-based genome assemblies for *An. stephensi, An. funestus, An. atroparvus*, and *An. albimanus* (*18*). The rate of X chromosome rearrangement was consistent among different species pairs ranging only between 0.120 and 0.130 breaks/Mb/Myr. The mean rate of X

chromosome evolution in genus *Anopheles* was 0.126±0.004 breaks/Mb/Myr (*18*). The standard deviation was calculated using four different species pairs. There are five fixed inversions between X chromosomes of *An. gambiae* and *An. arabiensis* (*31*). The length of the X chromosome genome assembly in *An. arabiensis* is 21.162 Mb. Assuming a constant rate of X chromosome rearrangement and two breaks per inversion, we calculated the split between the branches leading to *An. gambiae* and *An. arabiensis* using the following formula: Divergence (Myr) = (breaks/Mb)÷(rearrangement rate × 2). We multiplied the rearrangement rate by 2 in order to account for both lineages. Accordingly, divergence between the *An. gambiae* and *An. arabiensis* lineages occurred about 1.88±0.05 Myr ago, an estimate very close to the approximate divergence time of 1.85±0.47 Myr inferred independently from sequence divergence on the X chromosome (Fig. 1C, main text).

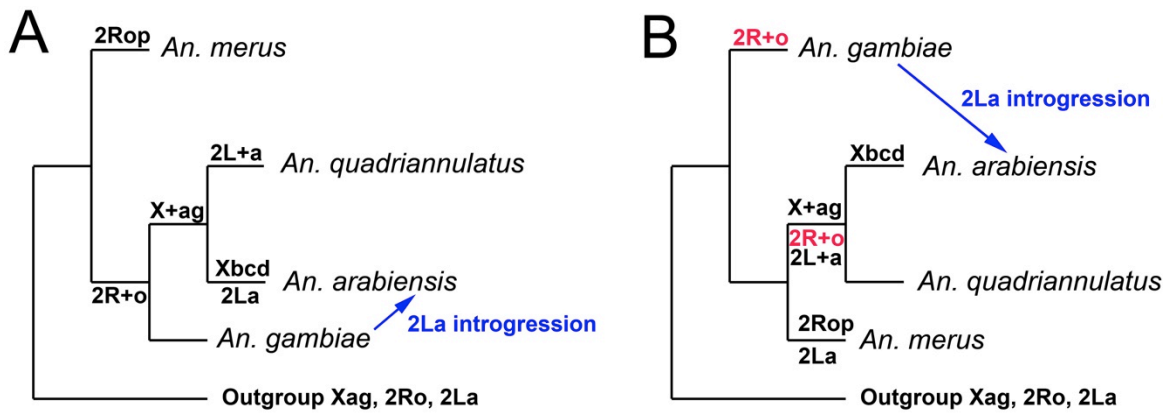


Fig. S27.
The rooted chromosomal phylogeny of the *An. gambiae* complex, based on fixed chromosomal inversions whose ancestral-derived relationships could be inferred from available genome assemblies of ingroup and outgroup species. *An. coluzzii* is not shown, as it does not differ from An. gambiae by any fixed inversions. **(A)** A parsimony tree that allows no independent fixations of inversions in different lineages but allows introgression of 2La from *An. gambiae* to *An. arabiensis*. **(B)** A tree that requires independent fixations of 2R+^{o} in two lineages and introgression of 2La from *An. gambiae* to *An. arabiensis*.

S6. Functional analysis of differentially introgressed regions

S6.1. Ecdysteroid quantification in *An. gambiae* and *An. arabiensis*.

Ecdysteroid titers in the male accessory glands (MAGs) of *An. gambiae* and *An. arabiensis* males were determined by an ACE Competitive Enzyme Immunoassay (EIA), using 20-hydroxyecdysone (20E) conjugated to acetylcholinesterase as a tracer and 20E EIA antiserum (Cayman Chemical). A standard curve was prepared from 625 pg of 20E (Sigma-Aldrich) in EIA buffer, with a series of seven 2-fold dilutions. MAGs were dissected from 4-day old virgin male mosquitoes and total ecdysteroids were extracted in 30 µl of methanol, re-dissolved in 50 µl of EIA buffer and loaded on a 96-well plate pre-coated with mouse anti-rabbit IgG (Cayman Chemical). Plates were incubated with tracer (50 µl) and antiserum (50 µl) for 18 h at 4°C and then developed with Ellman's Reagent (200 µl) for 90-120 min. Absorbance was measured on an ELISA plate reader at 412 nm. All samples were assayed in duplicate. The results are expressed as mean values ± SEM of 10 independent samples containing MAGs from 2 males each. Differences in mean amount of ecdysteroids in the MAGs were determined by 1-way ANOVA and the significance of pairwise comparisons was calculated using Tukey's HSD in Prism 6.0.

S6.2 Functional Enrichment analyses of (non-) introgressed genes

Functional enrichment analyses of genes in a target region of interest were conducted using DAVID (*108*), which clusters terms with similar functional categories. Here we report only significantly enriched clusters (enrichment score >1.3). Additionally, a specific analysis of Gene Ontology (GO) terms was conducted with the R software *goseq* (*109*), adapted from its original purpose (to test for GO enrichment in genes differentially expressed) to test for enrichment in (non-) introgressed gene lists. We also compared GO terms of genes in the target region to GO terms of genes in a random subset of the genome (excluding the target genes). The mean GO content of 15 random subsets, each containing an equal number of genes to the target region, was

compared to the mean GO content of the target region. To test for over-representation of genes implicated in the mosquito immune response, the number of immune-related genes from an updated version of ImmunoDB (*110*) was compared with chi-square tests of the target region versus genes outside the target region.

S6.2.1. *An. merus* and *An. quadriannulatus* introgression

Functional enrichment analyses were based on genes in regions detected as introgressed between this species pair based on D_{FOIL} statistics; the background used as reference was defined as the set of genes across the genome with sufficient data to pass our filters and allow for testing using D_{FOIL} (S4.1). We performed analyses on all introgressed genes across the genome, and additionally on genes in the 3La inversion (in PEST coordinates, from position 14,452,080 to 35,641,019 corresponding to AGAP010981 to AGAP011962).

The results of the genome-wide analysis and the 3La analysis of introgressed genes are presented in Tables S12-S15. Overall, functional enrichment was similar in both data sets. Functions related to immune/stress defense, apoptosis, oxidation-reduction, ion transport and polyamine synthesis were over-represented among the introgressed genes. Significantly more immune-related genes were observed (46 genes) than expected (29) in 3La (chi-square test, d.f.=1, P=0.002) versus all regions of the genome outside this inversion.

S6.2.2. Autosomal genes resistant to introgression between *An. arabiensis* and *An. gambiae*

Because most of the autosomes showed evidence of introgression between *An. gambiae* and *An. arabiensis*, it was of interest to examine the set of autosomal genes that resisted introgression, as these would be candidate genes potentially involved in reproductive isolation. For this purpose, we considered "non-introgressed" autosomal genes to be those whose gene trees support the species topology of ((*An. arabiensis*, *An. quadriannulatus*)*An. gambiae*) with >80% RAxML bootstrap support for the clade. To identify this set of genes, we selected a total of 7

Anopheles species for phylogenetic analysis: *An. gambiae* (PEST), *An. merus*, *An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. christyi*, and *An. epiroticus*. In addition, we identified single-copy gene families using OrthoDB (*93*); these included 5782 strict single-copy gene families (one gene in each of seven species) and 1130 relaxed single-copy gene families (one species was allowed to have more than one gene in the gene family, and the longest one was selected for further analysis). After eliminating the set of genes on the X chromosome, for each remaining gene, the peptide sequences were aligned using MUSCLE (*94*). The peptide alignment was back-aligned to a CDS alignment using Pal2Nal (*95*), and alignments were cleaned with TrimAl (*86*). RAxML (*84, 85*) was then run on the cleaned CDS alignment using the GTRGAMMA model of evolution and 100 fast bootstraps, and gene trees were re-rooted using *An. epiroticus* as the outgroup.

Of the 6319 autosomal orthologs whose gene trees were analyzed, the 485 that supported the species topology ((*An. arabiensis*, *An. quadriannulatus*)*An. gambiae*) were considered "non-introgressed" and submitted to functional enrichment analysis. DAVID revealed two clusters with an enrichment score >1.3, but the second cluster (zinc finger, FYVE-type) only contained four genes. The first cluster was annotated as 3',5'-cyclic-nucleotide phosphodiesterase (enrichment score of 2.28, 7 genes). Cyclic-nucleotide phosphodiesterases also appeared as the top over-represented GO terms from *goseq* analysis (Table S16). Phosphodiesterases again emerged as significantly over-represented when comparing GO terms in the target list with random lists of equal size; "phosphoric diester hydrolase activity" and "3'5'-cyclic-nucleotide phosphodiesterase activity" were ranked as the 9th and 18th most over-represented terms, respectively.

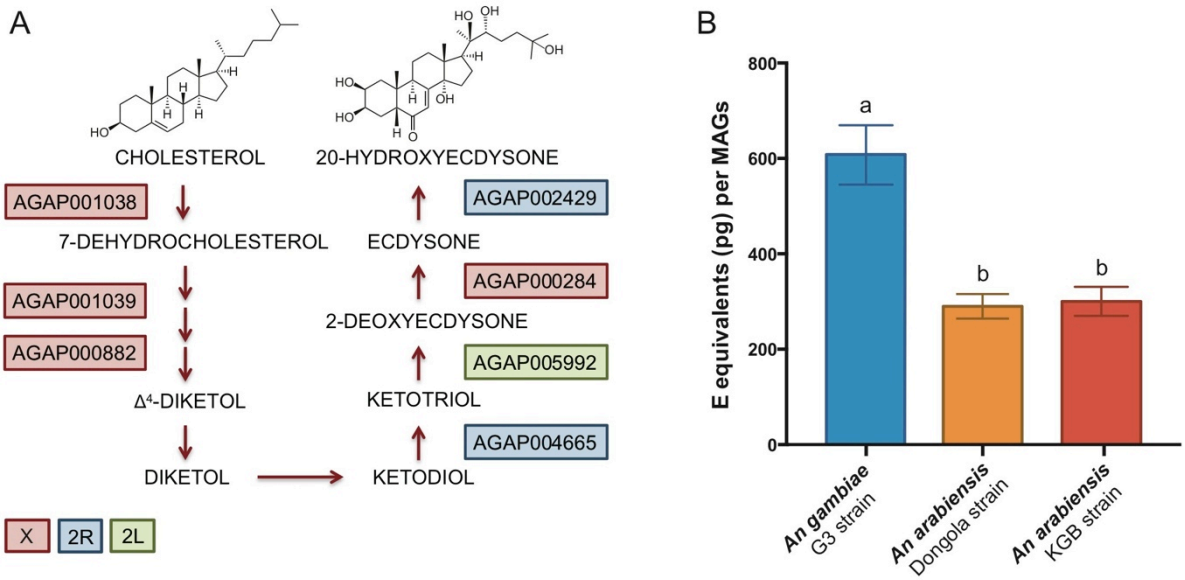


Fig. S28.
(A) Diagram of insect ecdysteroid synthesis pathway. *An. gambiae* genes corresponding to the seven enzymes currently known to be implicated in the production of 20-hydroxyecdysone (20E) are indicated along with their chromosomal location. Remarkably, in *An. gambiae* there is an overrepresentation of X chromosome genes in the pathway, with four out of seven 20E genes present on this chromosome (genes highlighted in pink). Interestingly, AGAP000284, the penultimate enzyme in the pathway, is present in the region (Xag) resistant to introgression between *An. gambiae* and *An. arabiensis*. **(B)** Comparison of 20E titers between *An. gambiae* and *An. arabiensis* males. Hormone levels were measured from male accessory glands (MAGs) dissected from males of *An. gambiae s.s.* and two different strains of *An. arabiensis*. Male *An. arabiensis* from both strains have significantly lower levels of ecdysone in their MAGs compared to *An. gambiae* as indicated by letters (ANOVA $F_{2,27}=17.92$, $P < 0.0001$, pairwise comparison Tukey's HSD).

Table S12.
DAVID annotation clusters with enrichment scores >1.3 for all genes introgressed between *An. merus* and *An. quadriannulatus*.

Cluster	Enrichment score	Number of genes	David cluster name
1	2.92	96	peptidase
2	2.6	18	digestion
3	1.68	15	ornithine decarboxylase
4	1.68	13	SET domain
5	1.68	24	leucine-rich repeat
6	1.61	11	nucleotide transport and
7	1.61	47	chitin binding
8	1.42	9	alcohol dehydrogenase
9	1.4	55	ion binding and transport

Table S13.
The top 20 GO terms significantly over-represented among genes introgressed between *An. merus* and *An. quadriannulatus.* Number of occurrences is given for each GO term in the target introgressed and background (total) gene lists.

GO term name	Over represented *P*-value	Introgressed	Total (Introgressed + Non-introgressed)
catalytic activity	0.0005	149	694
cytoskeleton	0.0005	29	148
dynein complex	0.0005	3	12
enzyme regulator activity	0.0005	30	168
focal adhesion	0.0005	2	2
hydrolase activity	0.0005	111	542
motor activity	0.0005	6	17
myosin complex	0.0005	5	16
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced ascorbate as one donor, and incorporation of one atom of oxygen	0.0005	4	4
polyamine biosynthetic process	0.0005	5	6
potassium ion transmembrane transport	0.0005	9	28
potassium ion transport	0.0005	8	24
protein binding	0.0005	246	1632
regulation of GTPase activity	0.0005	7	44
regulation of ion transmembrane transport	0.0005	9	24
voltage-gated ion channel activity	0.0005	7	12
voltage-gated potassium channel activity	0.0005	4	12
voltage-gated potassium channel complex	0.0005	4	7
digestion	0.0010	7	11
homophilic cell adhesion	0.0010	7	33

Table S14.
DAVID annotation clusters with enrichment scores >1.3 based on genes introgressed between *An. merus* and *An. quadriannulatus* in the 3La inversion.

Cluster	Enrichment score	Number of genes	David cluster name
1	3.9	14	SET domain
2	1.56	15	ornithine decarboxylase
3	1.54	11	monooxygenase
4	1.51	14	aldo/keto reductase
5	1.48	13	leucine-rich repeat
6	1.47	57	serine-type peptidase activity
7	1.45	8	fibrinogen
8	1.33	8	caspase

Table S15.
The top 20 GO terms significantly (P<0.05) over-represented in 3La genes introgressed between *An. merus* and *An. quadriannulatus*. Number of occurrences is given for each GO term in the target introgressed and background (total) gene sets.

GO term name	Over-represented *P*-value	Introgressed	Total (Introgressed + Non-introgressed)
catalytic activity	0.0005	77	749
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced ascorbate as one donor, and incorporation of one atom of oxygen	0.0005	4	6
polyamine biosynthetic process	0.0005	4	6
mitosis	0.0015	6	18
protein binding	0.0015	151	1740
hydrolase activity	0.0035	58	579
phosphatase activity	0.0035	14	92
cell cycle	0.0045	10	58
lipid metabolic process	0.0050	22	176
copper ion binding	0.0055	6	22
dephosphorylation	0.0055	12	76
organelle	0.0055	122	1421
COPII vesicle coating	0.0060	2	2
motor activity	0.0070	5	19
condensin complex	0.0110	2	3
alkaline phosphatase activity	0.0125	3	7
chitin binding	0.0125	14	104
chitin metabolic process	0.0125	14	104
serine-type endopeptidase activity	0.0125	36	341
apoptotic process	0.0145	6	26

Table S16.
GO terms significantly ($P<0.05$) over-represented among autosomal genes resistant to introgression between *An. gambiae* and *An. arabiensis*.

GO term name	Over-represented *P*-value	Number of GO in orthologs not introgressed	Number of GO in all autosomal orthologs
phosphoric diester hydrolase activity	0.0005	7	14
3',5'-cyclic-nucleotide phosphodiesterase activity	0.0010	5	8
neurotransmitter:sodium symporter activity	0.0050	4	9
symporter activity	0.0050	4	9
neurotransmitter transport	0.0065	4	10
carbohydrate metabolic process	0.0075	18	126
monovalent inorganic cation transport	0.0100	5	17
iron-sulfur cluster binding	0.0115	6	25
2 iron, 2 sulfur cluster binding	0.0145	4	13
translational termination	0.0150	2	3
mitochondrial inner membrane	0.0160	3	8
Mo-molybdopterin cofactor biosynthetic process	0.0170	2	3
sodium ion transmembrane transport	0.0170	6	27
negative regulation of catalytic activity	0.0240	4	15
Rab GTPase activator activity	0.0245	5	22
generation of precursor metabolites and energy	0.0250	6	31
cation transport	0.0260	8	45
sensory perception of taste	0.0360	5	25
regulation of endopeptidase activity	0.0450	3	10