Appendix S1

MATERIALS AND METHODS

Sample collection: *A. coluzzii* and *A. gambiae* samples were collected from four sites in Mali: Selenkenyi (11.700°N, 8.2833°W), Kela (11.8868°N, 8.4474°W), Sidarebougou (11.4664°N, 5.7435°W), and Tissana (14.3612°N, 5.9131°W) (**Figure S1**). The distance between these sites ranged from 27 km (Selenkenyi to Kela) to 392 km (Selenkenyi to Tissana). Tissana is the driest site, with rainfall ranging from 505-694 mm/yr, followed by Kela (750-871 mm/yr), Sidarebougou (848-1106 mm/yr), and Selenkenyi (871-1240 mm/yr). *A. coluzzii* and *A. gambiae*, as well as *Anopheles arabiensis* (part of the *A. gambiae* sensu lato complex) are present in Selenkenyi and Kela; *A. coluzzii* and *A. gambiae* are present in Sidarebougou, and the Tissana population is mainly *A. coluzzii*, *A. arabiensis*, and extremely rare *A. gambiae* (**Figure S1**). Sidarebougou is located in a cotton-growing region with high agricultural insecticide use (1), while the other sites are in areas with moderate agricultural insecticide usage. These sites were chosen because we have multiple yearly collections allowing us to model selection coefficients.

Mosquitoes were collected during the rainy season (August-October) during the following years: Selenkenyi & Kela: 2002, 2004, 2006, 2009, 2010, 2011, 2012; Sidarebougou: 2002, 2009, 2011 and Tissana: 2006, 2011. Resting mosquitoes were collected by aspiration inside houses and stored in individual tubes in 80% ethanol.

DNA extraction: DNA from mosquito tissue was extracted using a Qiagen Biosprint (Valencia, CA) with the DNA Tissue protocol. Species identification PCRs (2, 3) were performed to distinguish *A. coluzzii* and *A. gambiae* from *A. arabiensis* and products were visualized on a QIAxcel instrument using a DNA Screening Kit cartridge (Qiagen, Valencia, CA).

SNP genotyping: Samples were analyzed with an iPLEX Gold multiplexed SNP genotyping array, using a Nanodispenser RS1000 and MassARRAY Analyzer Compact 96 (Sequenom, San Diego, CA) at the University of California - Davis Veterinary Genetics Lab. Samples were genotyped at five SNP loci: one in the X divergence island (28S rDNA intergenic sequence, also used to differentiate *A.*

gambiae and *A. coluzzii* (4)), one in the 2L divergence island (5), one in the 3L divergence island (5) and the L1014F SNP (TTA->TTT (6)) **(Table S1 and Figure S3)**. TyperAnalyzer v. 4.0.24.71 was used to design forward and reverse primers to amplify 80-120 bp fragments surrounding the SNPs, and internal extension primers (UEP) to capture SNP genotypes (see **Table S1** for primer sequences). Amplification and extension reactions were run with 2 μl DNA aliquots on a Sequenom MassARRAY Analyzer using manufacturer's protocol. SNP data was analyzed with TyperAnalyzer using the clustering algorithm, and manually checked for accuracy. Poorly amplified samples or poorly clustered samples were removed prior to further analysis. All karyotype and SNP data is publicly available in the PopI OpenProject AgKDR section (https://popi.ucdavis.edu/).

Statistical tests: Samples were classified into *A. coluzzii* and *A. gambiae* based on 28S IGS-540 SNP (Favia et al. 2001). *A. coluzzii* individuals have TT genotype for 28S IGS-540 SNP, *A. gambiae* individuals have CC and hybrid individuals have CT.

Standard molecular indices and Hardy-Weinberg equilibrium were calculated for *A. coluzzii* populations with Arlequin 3.1 (7). L1014F SNP frequencies were calculated in all groups by site and year (due to geographic proximity and similar L1014F frequency, samples from Selinkenyi and Kela were combined). Linkage disequilibrium (LD) between the 2L island SNP and L1014F was calculated for all *A. coluzzii* populations using the maximum likelihood method implemented in the *EMLD* program (8). Adjusted r^2 is used for further analysis. Linear regression was conducted to test the significance in decline of LD between the 2L SNP and L1014F using using Matplotlib (9).

Expected genotype calculations: Expected L1014F genotypes for F_1 hybrids were calculated assuming that *A. coluzzii* and *A. gambiae* hybridized at random, without regard to L1014F genotype. *A. coluzzii* was assumed to be 100% +/+ in our study sites in Mali (N=226 for pre-2006 +/+ abundance in *A. coluzzii* as shown in Table 1), while the distribution of *A. gambiae* L1014F genotypes was assumed from 2006 Selenkenyi data: 33.3% +/+, 40.7% +/r, and 25.9% r/r. F_1 hybrids would therefore be expected to be 53.7% +/+ and 46.3% +/r. L1014F frequencies for F_2 backcrosses were calculated using the above frequencies for F_1, *A. coluzzii* and *A. gambiae* populations. We assume hybridization occurred only in

2006 and provided the initial influx of resistant allele (r) and no further hybridization occurred in subsequent years based on our previous study (10). The lower bound of L1014F frequencies assume mating only occurred within *A. coluzzii* after 2006, providing resistant allele frequency of 0.045 (=7/(77*2)) from 2006 *A. coluzzii* (Table 1). The upper bound frequencies assume both F1 hybrids and *A. coluzzii* from 2006 contributed to gene pool in subsequent generations, providing resistant allele frequency of 0.074 (=14/(94*2)) from 2006 *A. coluzzii* (Table 1). For post-2006, genotype frequencies within *A. coluzzii* are calculated assuming random mating in the absence of selection. Deviation from expected values was calculated by a randomization goodness-of-fit test with 10,000 replications.

Selection coefficient calculation: Temporal changes in resistant allele frequency in *A. coluzzii* and *A. gambiae* in Selenakenyi/Kela and Sidarebougou were fitted to the recursive selection equation (11)

$$p_{t+1} = \frac{p_t^2(1+s) + p_t q_t(1+hs)}{1+s(p_t^2 + 2hp_t q_t)}$$

using the *leastsq* option in the *scipy.optimize* package in Python (http://scipy.org/). *A. coluzzii* from Tissana were not fitted, due to the low number (n=2) of time points available. To avoid overparametization for each population, h (dominance coefficient) was held constant at h = 0, 0.25, 0.5, 0.75, and 1.0, while p_0 (initial resistant allele frequency) and *s* (selection coefficient) were fitted to the data. For estimating *s*, we assumed one generation per month (11), or 12 generations per year. *A. gambiae* from Selenkenyi and *A. coluzzii* from both sites showed obvious selection for resistant allele, and values for *s* were similar across all values of h (standard deviation = 7-28% of the mean), while *A. gambiae* from Sidarebougou had relatively constant resistant allele frequencies (83-98%) and very low values for *s* (*s* = 0.014), such that variation in h had a greater impact on *s* (standard deviation = 81% of the mean). The best fit for all populations (h = 0.5) is reported in the results.

Genome sequencing: 33 *A. coluzzii* and *A. gambiae* from Kela/Selinkenyi were whole-genome sequenced: 12 *A. gambiae* (2012); 6 +/+, pre-2006 *A. coluzzii* (2002-2004); 8 r/+ and 5 r/r, post-2006 *A.*

coluzzii (2010-2012); and 2 post-2006 *A. coluzzii* individuals that had undergone recombination between the 2L island SNP and the L1014F locus (2012).

DNA concentration was quantified by a fluorescent assay for double-stranded DNA using a Qubit 2.0 fluorometer (Life Technologies). DNA was cleaned and concentrated with the DNA Clean and Concentrator kit (Zymo Research Corporation). Library preparations were made with the Nextera DNA Sample Preparation Kit (Illumina), using TruSeq dual indexing barcodes (Illumina), according to the manufacturer's protocol. We used 25-50 ng of input DNA for library construction. Libraries were size-selected with Agencourt AMPure XP beads (Beckman Coulter), according to manufacturer's instructions for Illumina Hi-Seq libraries. Final library preps were quantified and checked for insert size using a QIAxcel instrument (Qiagen, Valencia, CA) and Bioanalyzer 2100 (Agilent), and concentration was measured with a Qubit 2.0 fluorometer (Life Technologies). Barcoded libraries were pooled in equimolar amounts and sequenced with Illumina's HiSeq2500 platform with paired-end 100 bp reads, at the QB3 Vincent J Coates Genomics Sequencing Laboratory at UC Berkeley.

Short-read Genome Sequence Mapping: We assessed the quality of our genome sequencing reads using the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptor sequences and poor quality sequence was trimmed from the raw Illumina fastq files using the *Trimmomatic* software, version 0.30 (12), using default options. Reads were aligned to the *A. gambiae* reference genome (AgamP3 version 3.x) using *Stampy* (version 1.0.22) (13), with *BWA* used as a pre-mapper to accelerate the mapping (14). We then used *MarkDuplicates* from *Picard* tools (15) to remove PCR duplicates and used the Genome Analysis Tool Kit (*GATK*) v1.7 (16) to realign reads around indels. The resulting sorted *bam* (Binary sequence Alignment/Map) files, which contain sequences for each read and its mapping position, were then used for later analysis.

SNP calling: In order to identify sequence variants that are over-represented (frequency of 0.67 or greater) in *A. coluzzii* or *A. gambiae*, we chose 3 individuals of each species and generated a multiple pileup file for those individuals using *SAMtools mpileup* (version: 0.1.19-44428cd) (17). Next, we produced a consensus sequence for *A. coluzzii* and *A. gambiae* using the two mpileup files and the *mpileup2cns* program from VarScan (18). Then, we identified all differences between the final consensus sequences between species. It should be noted that many but not all of the identified SNP locations are fixed between species. Therefore, these SNPs are appropriate for comparing relative trends in "*A. gambiae*-ness" between individuals and should not be used individually to genotype *A. coluzzii* and *A. gambiae*. This set of SNPs was used for calculating F_{ST} and "*A. gambiae* proportions" for Figure 2.

Genotype frequencies for the minor-effect SNPs that have co-introgressed within the *kdr* gene, identified in whole genome sequenced samples. These minor-effect SNPs are the synonymous/intronic [C/T] SNP at 2L: 2,417,678 (20) and N1575Y [A/T] at 2L: 2,429,745 (51). The genomic locations of these SNPs are shown in Figure S4.

Proportion *A. gambiae*: We used *SAMtools mpileup* (17) to genotype each individual at the *A. gambiae*/*coluzzii* differential SNP positions identified in our SNP calling step, excluding the individuals that were already used to define the SNPs. Using data from the variant call file (*VCF*) output, we calculated the proportion of *A. gambiae* alleles (limited to biallelic sites) in 100kb bins across the genome. For example, homozygous sites were counted as *A. coluzzii* (Proportion *A. gambiae* = 0) or *A. gambiae* (=1), depending on allele, while heterozygous sites were counted as half (Proportion *A. gambiae* =0.5) *A. coluzzii* and half *A. gambiae*. The trends in proportion *A. gambiae* were plotted using Matplotlib (9), with Gaussian smoothing.

F_{ST}, and Tajima's *D*: *SAMtools* and *BCFtools* were used to generate *mpileup* consensus files and call variants, against the PEST reference genome. *VCFtools* (19) was used to calculate F_{ST} in 100kb windows across chromosome 2 and Tajima's *D* in 5kb windows from 0-9 Mbp on chromosome 2L.

Table S1: SNPs used in the iPLEX Gold Assay, with genomic location, gene ID, primers, and primer concentrations.

SNP ID	location	gene	Primers	Variants	multiplex concentration
28SIGS-540	multiple locations in X centromere	rDNA IGS	F: TTGAGTGTAGCAAGGGATCG R: ACCAAGCTTCACCAGAGCAC UEP: GACCAAGATGGTTCGTT	[G/A]	1.0 µM 1.0 µM 7.0 µM
04679-157	2L: 209,534	*AGAP004679*	F: ATATCAAGGATATCACACG R: TCTGTTCGTCGTACCATCAG UEP: AAGGATATCACACGATTCGTTAA	[C/T]	1.0 µM 1.0 µM 9.3 µM
10313-052	3L; 296,897	*AGAP010313*	F: AAGAAGCTGTGGCGTGTTAC R: TAGGCTTGGATATTGTTCCT UEP: TGGATATTGTTCCTCGATAT	[C/T]	1.0 µM 1.0 µM 11.6 µM
L1014F	2L; 2,422,652	*AGAP04707*	F: CTTGGCCACTGTAGTGATAG R: TGTAAAAACGATCTTGGTCC UEP: GTTAATTTGCATTACTTACGAC	[A/T]	1.0 µM 1.0 µM 14.0 µM

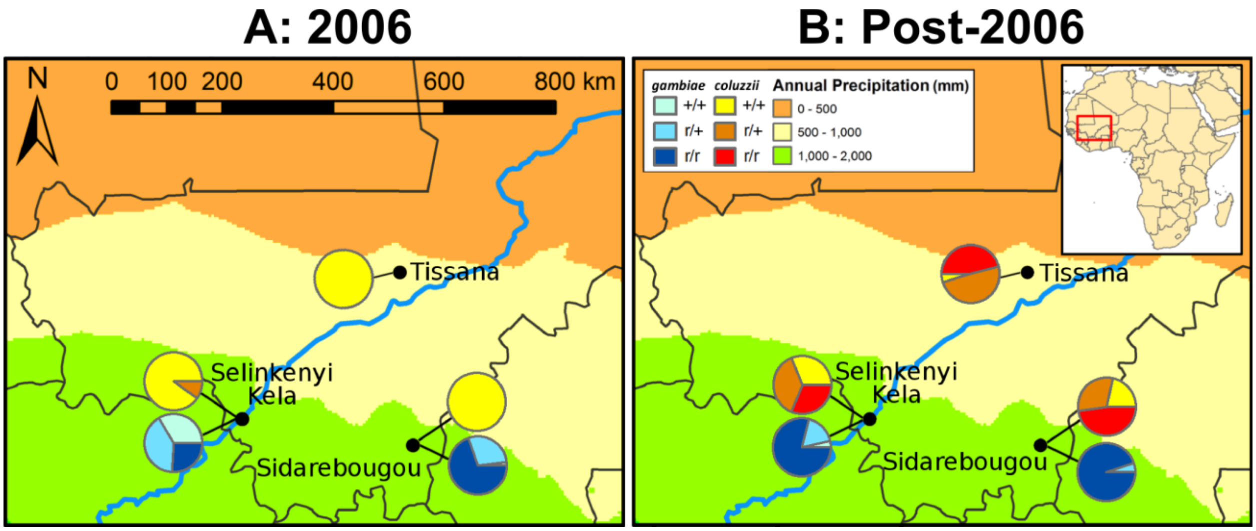
Table S2: linkage disequilibrium (LD) between the 2L SNP and L1014F, within *A. coluzzii* populations. EMLD was used to calculate D’, r^2 and p-values using Fisher’s exact test.

year	site	N	D'	r^2	p-value
2006	Tissana	40		Monomorphic	
2006	Kela	27	1	0.4909	0.03108911
2006	Selenkenyi	49	0.6279	0.226	<0.00001
2009	Sidarebougou	47	1	0.9583	<0.00001
2009	Kela	72	0.903	0.8155	<0.00001
2010	Kela	60	1	0.8415	<0.00001
2010	Selenkenyi	59	0.926	0.8269	<0.00001
2011	Tissana	70	0.8258	0.659	<0.00001
2011	Sidarebougou	22	0.8333	0.463	<0.00001
2011	Selinkenyi	38	1	0.7133	<0.00001
2012	Selenkenyi	69	0.9222	0.6487	<0.00001

Table S3: Estimated selection coefficient and initial resistant allele frequency for L1014F in *A. coluzzii* and *A. gambiae* in Selenkenyi/Kela and Sidarebougou. p_0 = initial allele frequency, h = dominance coefficient, s = selection coefficient, N = number of timepoints sampled, R^2 = correlation coefficient, RMSE = reduced mean square error. Parameters were estimated using h = 0, 0.25, 0.5, 0.75, and the best fit (h = 0.5) was reported.

site	**population**	**p_0**	**h**	**s**	**N**	**R^2**	**RMSE**	**reduced chi-squared**
Selenkenyi/Kela	*A. coluzzii*	1.70E-06	0.5	0.13159	8	0.95570	0.06396	4.09E-03
Selenkenyi/Kela	*A. gambiae*	8.81E-03	0.5	0.06478	6	0.76304	0.13122	1.72E-02
Sidarebougou	*A. coluzzii*	1.12E-06	0.5	0.14355	5	0.99997	0.00166	2.79E-06
Sidarebougou	*A. gambiae*	8.00E-01	0.5	0.01430	5	-0.06539	0.06343	4.02E-03

Figure S1: Map of field site locations in Mali, including annual precipitation categories. **A)** *A. coluzzii* and *A. gambiae* populations in 2006, **B)** populations in 2009-2012. Due to geographic proximity, Selenkenyi and Kela are shown together. Pie graphs indicate *A. gambiae* in shades of blue, *A. coluzzii* in yellow/orange/red. Darker segments on pie charts indicate proportion resistant homozygotes (r/r), medium shades indicate heterozygotes (+/r), and light colors indicate susceptible homozygotes (+/+).



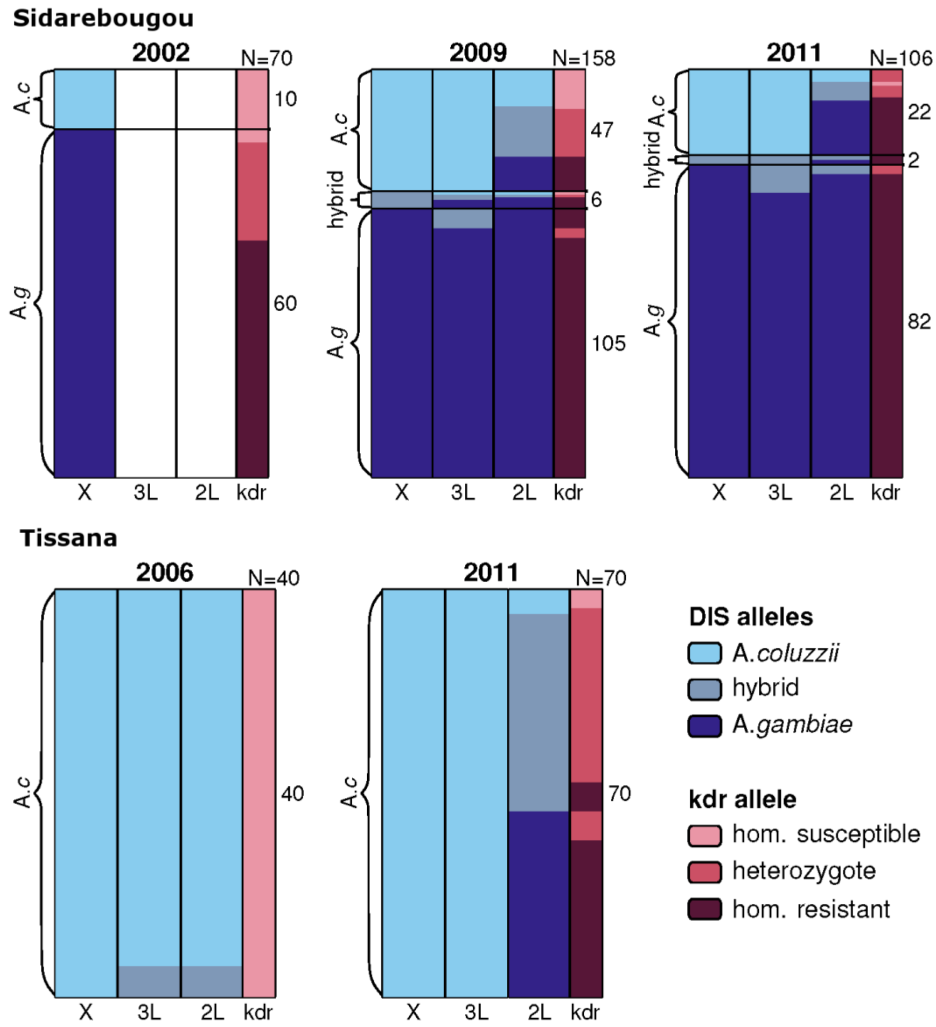
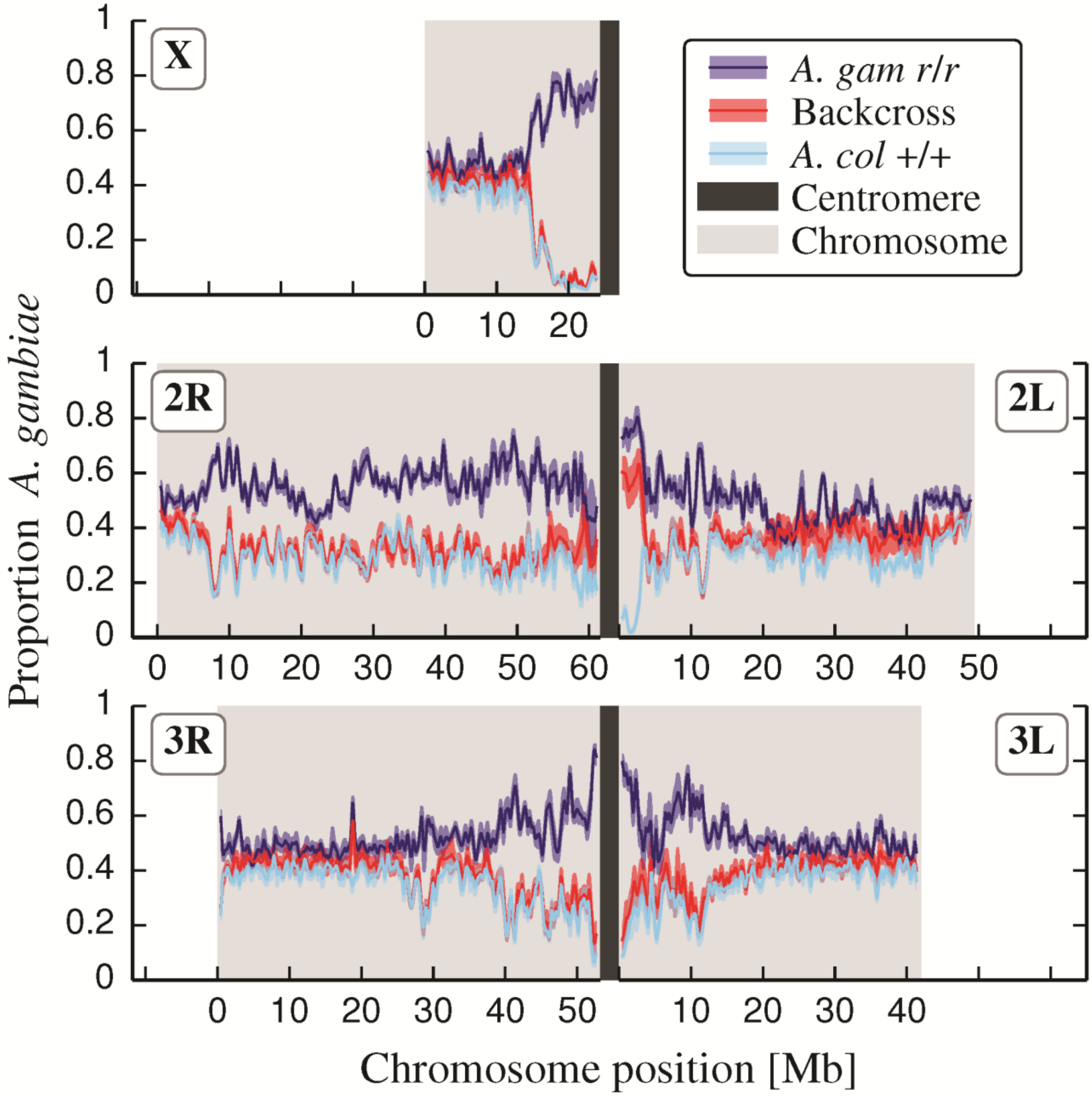


Figure S2: Heat maps of three divergence island SNPs and the L1014F SNP for Sidarebougou and Tissana pre- and post-introgression. Columns represent SNPs (X divergence island, 3L island, 2L island, L1014F/*kdr*-*w*), individual mosquitoes are represented by colored horizontal lines, with individuals stacked vertically. Light blue = homozygous for *A. coluzzii*-associated alleles, dark blue = homozygous for *A. gambiae*-associated alleles, grey = heterozygous, white = missing data, dark red = r/r, medium red = +/r and pink = +/+. Samples that are heterozygous (grey) across the X, 2L, and 3L SNPs are assumed to be F_1 hybrids. Population assignments (*A. coluzzii*, hybrid, and *A. gambiae*) are indicated by brackets on the left of each heat map.

Figure S3. Proportion *A. gambiae* ancestry (Gaussian-smoothing with 100kb windows and a degree of 5) for *A. gambiae* (purple), pre-introgression *A. coluzzii* (blue), introgressed *A. coluzzii* (red). Line width indicates standard error among samples. Dark Gray lines indicate centromeres. Light gray area indicates the chromosome boundaries.



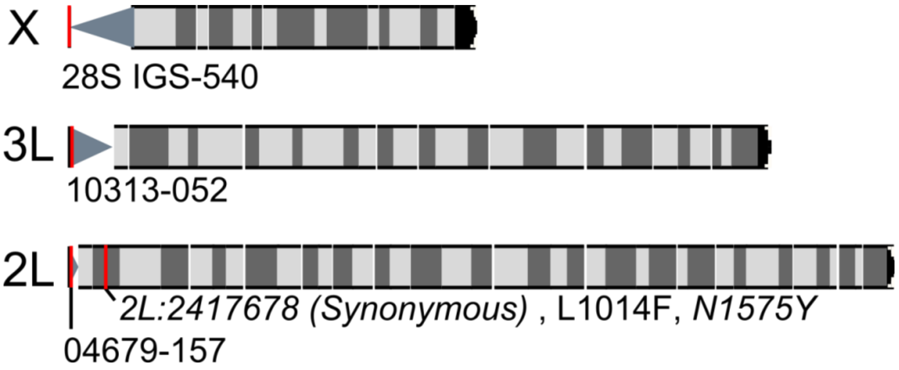


Figure S4. Genomic location of SNPs used in the manuscript. A synonymous SNP at the 2,417,678 bp of the 2L chromosomal arm and a non-synonymous mutation, N1575Y are within the *kdr* gene and genotypes were retrieved from genome sequence data. Other SNP genotypes were determined using iPLEX Gold Assay.

SUPPLEMENTAL REFERENCES

1. Reimer L, *et al.* (2008) Relationship between kdr mutation and resistance to pyrethroid and DDT insecticides in natural populations of *Anopheles gambiae*. *J Med Entomol* 45(2):260-266.
2. Scott JA, Brogdon WG, & Collins FH (1993) Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. *Am J Trop Med Hyg* 49(4):520-529.
3. Favia G, *et al.* (1997) Molecular identification of sympatric chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive isolation. *Insect Mol Biol* 6(4):377-383.
4. Favia G, Lanfrancotti A, Spanos L, Siden-Kiamos I, & Louis C (2001) Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae* s.s. *Insect Mol Biol* 10(1):19-23.
5. White BJ, Cheng C, Simard F, Constantini C, & Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular ecology* 19:925-939.
6. Martinez-Torres D, *et al.* (1998) Molecular characterization of pyrethroid knockdown resistance (kdr) in the major malaria vector *Anopheles gambiae* s.s. *Insect Mol Biol* 7(2):179-184.
7. Excoffier L, Laval G, & Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1:47-50.
8. Huang Q, Shete S, Swartz M, & Amos CI (2005) Examining the effect of linkage disequilibrium on multipoint linkage analysis. *BMC genetics* 6 Suppl 1:S83.
9. Hunter JD (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9(3).
10. Lee Y, *et al.* (2013) Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci U S A* 110(49):19854-19859.
11. Barbosa S, Black WCt, & Hastings I (2011) Challenges in estimating insecticide selection pressures from mosquito field data. *PLoS neglected tropical diseases* 5(11):e1387.
12. Lohse M, *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res* 40(Web Server issue):W622-627.

13. Lunter G & Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* 21(6):936-939.

14. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.

15. Anonymous (Picard Tools.

16. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.

17. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.

18. Badolo A*, et al.* (2012) Three years of insecticide resistance monitoring in Anopheles gambiae in Burkina Faso: resistance on the rise? *Malar J* 11:232.

19. Danecek P*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156-2158.