Received Date : 13-Aug-2015

Accepted Date : 01-Sep-2016

Article type : Original Article

The last bastion? X-chromosome genotyping of *Anopheles gambiae* species-pair males from a hybrid zone reveals complex recombination within the major candidate 'genomic island of speciation'

B. Caputo[1], V. Pichler[1], E. Mancini[1*], M. Pombi[1], J. L. Vicente[2], J. Dinis[3], K. Steen[4], V. Petrarca[5], A. Rodrigues[3], J. Pinto[2], A. della Torre[1], D. Weetman[4]

[1]Istituto Pasteur-Fondazione Cenci-Bolognetti, Dipartimento di Sanità Pubblica e Malattie Infettive, Università "Sapienza", Piazzale Aldo Moro 5, 00185, Rome, Italy; [2]Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Rua daJunqueira,100, 1349-008 Lisboa, Portugal; [3]Instituto Nacional de Saúde Pública, Ministério da Saúde Pública, Avenida Combatentes da Liberdade da Pátria, Apartado 861, 1004 Bissau Codex, Guinea Bissau, [4]Vector Biology Department, Liverpool School of Tropical Medicine, Pembroke Pl, Liverpool, Merseyside L3 5QA, UK; [5]Istituto Pasteur-Fondazione Cenci-Bolognetti, Dipartimento di Biologia e Biotecnologie Charles Darwin, Università "Sapienza", Piazzale Aldo Moro 5, 00185, Rome, Italy;
**current address: Dipartimento di Scienze, Università ROMA TRE, Viale Guglielmo Marconi 446, Rome, Italy*

Keywords: Sex chromosome, centromeres, hybridization, introgression, SNP, malaria vectors.

Corresponding author:

Beniamino Caputo
Istituto Pasteur-Fondazione Cenci-Bolognetti, Dipartimento di Sanità Pubblica e Malattie Infettive, Università "Sapienza", Piazzale Aldo Moro 5, 00185, Rome, Italy
beniamino.caputo@uniroma1.it

Running title: **Recombination in *A. gambiae* speciation island**

ABSTRACT

Speciation with gene flow may be aided by reduced recombination helping to build linkage between genes involved in the early stages of reproductive isolation. Reduced recombination on chromosome-X has been implicated in speciation within the *Anopheles gambiae* complex, species of which represent the major Afrotropical malaria vectors. The most recently diverged, morphologically-indistinguishable, species-pair, *An. gambiae* and *An. coluzzii*, ubiquitously display a 'genomic island of divergence' spanning over 4Mb from chromosome-X centromere, which represents a particularly promising candidate region for reproductive isolation genes, in addition to containing the diagnostic markers used to distinguish the species. Very low recombination makes the island intractable for experimental recombination studies, but an extreme hybrid zone in Guinea Bissau offers the opportunity for natural investigation of X-island recombination. SNP-analysis of chromosome-X hemizygous males revealed: (i) strong divergence in the X-island despite a lack of autosomal divergence; (ii) individuals with multiple-recombinant genotypes, including likely double crossovers and localized gene conversion; (iii) recombination-driven discontinuity both within and between the molecular species markers, suggesting that the utility of the diagnostics is undermined under high hybridization. The largely-, but incompletely-protected nature of the X-centromeric genomic island is consistent with a primary candidate area for accumulation of adaptive variants driving speciation with gene flow, whilst permitting some selective shuffling and removal of genetic variation.

INTRODUCTION

In species with heteromorphic sex chromosomes, the X (or Z) chromosomes exhibit a different mode of inheritance from autosomes, which may often act to increase rates of evolution (Charlesworth *et al.* 1987). Absence of recombination between X and Y chromosomes in the heterogametic sex reduces population recombination rate enhancing

the likelihood of extended selective sweeps (Nam *et al.* 2015), which may be expanded further by structural variations that reduce recombination (Feder *et al.* 2012). Though not necessarily an absolute pre-requisite for speciation with gene flow, some degree of physically-reduced recombination may help to sustain selection to maintain integrity of linked genes involved in reproductive isolation (Wu 2001; Wu & Ting 2004; Butlin 2005; Smadja & Butlin 2011) and can extend their influence by enhanced divergence hitchhiking (Via 2009). Examples of such physical features of the genome reducing recombination include paracentric inversion polymorphisms and expansive pericentromeric regions adjacent to heterochromatin. Whilst neither is unique to sex chromosomes, both are strongly implicated in speciation involving chromosome-X of members of the *Anopheles gambiae* complex (Coluzzi *et al.* 2002; Turner *et al.* 2005; Fontaine *et al.* 2015). Historically considered a single variable and opportunistic species, these mosquitoes are now recognized as a group of morphologically indistinguishable, but genetically discontinuous breeding units, that display considerable genetically-based ecological and behavioral differences and diverse roles as malaria vectors. Most *An. gambiae* complex species-pairs are characterized by intrinsic postzygotic reproductive barriers, in the form of sterile F1 males, with hybrid sterility factors mapping primarily to a 15Mb region of the chromosome-X, which exhibits fixed differences between some of the species pairs (Slotman *et al.* 2004, 2005). The same genomic region is characterized by large fixed chromosomal inversions, distinguishing some members of the complex and preventing recombination among species with different karyotypes (Coluzzi *et al.* 2002). Moreover, the same region, though comprising only about 8% of the genome, served to reveal the 'true' species tree within the complex, with autosomes providing a misleading signal as a result of extensive historical introgression (Fontaine *et al.* 2015), which continues at very low levels between some species (Weetman *et al.* 2014).

In contrast, the most recently-separated species of the complex, *An. gambiae* and *An. coluzzii* (Coetzee *et al.* 2013) - originally named as the *An. gambiae* S and M molecular forms (della Torre *et al.* 2001) - share the same chromosome-X karyotype, lack intrinsic

postzygotic isolating mechanisms (Diabaté *et al.* 2007) and are reproductively isolated by incompletely-understood pre-zygotic mechanisms, among which swarming segregation, ecological larval niche partitioning and selection against hybrids at the larval stage likely play important roles (Lehmann & Diabate 2008; Diabaté *et al.* 2009; Gimonneau *et al.* 2012). Although viable and fertile hybrids between the two species/molecular forms are easily obtained under laboratory conditions, hybrids in most wild populations are detected, albeit infrequently (della Torre *et al.* 2005). Coupled with detection of largely, but incompletely, differentiated "mosaic" genomes (Turner *et al.* 2005; Neafsey *et al.* 2010; Reidenbach *et al.* 2012; Clarkson *et al*. 2014), this has made the *Anopheles gambiae* species pair an attractive and influential model system for genomic studies of ecological speciation.

Anopheles gambiae/S-form and *An. coluzzii*-M-form were originally identified by single nucleotide polymorphisms in the intragenomic spacer region of the multi-copy ribosomal DNA deep in the pericentromeric region of chromosome- X (della Torre *et al.* 2001) and by a secondary diagnostic marker, the *An. coluzzii*-specific insertion of a transposable element of the SINE family in the same region (SINE-X; Santolamazza *et al.* 2008). Genomic studies showed that the two species segregate consistently at large pericentromeric 'genomic islands of divergence' on chromosome-X, as well as on the two autosomes (Turner *et al.* 2005; White *et al.* 2010), but multiple smaller regions of differentiation throughout the genome were also detected (Neafsey *et al.* 2010; Weetman *et al.* 2010; Reidenbach *et al.* 2012; Clarkson *et al.* 2014). These highly divergent genomic regions have been termed 'genomic islands of speciation' (Turner & Hahn 2010), implying that they contain a set of linked loci involved in the reproductive isolation process and are under divergent selection in the two species. However, serious doubts about the importance of these genomic regions for speciation have been raised based on the disputed reliability of the metrics used to identify them (Noor & Bennett 2009; Cruickshank & Hahn 2014). Moreover, recent homogenization of the putative autosomal genomic island on chromosomal arm 2L by adaptive introgression without any detectable impact on reproductive isolation was repeatedly observed in nature

(Lee *et al.* 2013; Clarkson *et al.* 2014; Norris *et al.* 2015). Yet recent evidence directly supports involvement of the X centromeric island in reproductive isolation. Aboagye-antwi *et al.* (2015) used a laboratory cross/back-cross protocol to dissociate an 8 Mb centromere-proximal section, containing the X-island of divergence from the rest of the chromosome-X and the autosomes. Regardless of the rest of their genome, backcross females almost exclusively preferred to mate with males possessing the same species-specific X-island, strongly implicating the region in assortative mating. Nevertheless, the huge size and lack of observable recombination within the X-island region not only makes pinpointing candidate genes difficult, but also potentially presents an evolutionary problem. Extremely low recombination across megabases could be both crucial for protection of co-adaptive variants from break-up, especially in earlier stages of divergence with gene flow, but also limit potential for adaptive shuffling of variants and removal of deleterious mutations, owing to inefficient background selection (Hill & Robertson 1966; Birky & Walsh 1988; Yeaman 2013). As a consequence, regions of low recombination might initially provide a selective advantage, but over time become progressively less so as linked deleterious variation accumulates.

Though recombination in the X-island appears far too low to be detected by experimentally crossing *An. gambiae* and *An. coluzzii* colonies (Aboagye-antwi *et al.* 2015), the key question of X-island recombination could be investigated in areas of high hybridization between the species-pair. Events of hybridization which repeatedly occur in their range of sympatry (Weetman *et al.* 2012; Lee *et al.* 2013) did not reveal any evidence of recombination in the X-island, even though they were shown to represent an important route for adaptive introgression in autosomal centromeric areas, at least under strong selection pressure, as in the case of the autosomal *kdr 1014F* insecticide resistance mutation (Clarkson *et al.* 2014; Norris *et al.* 2015). The temporarily stable hybrid zone (with a peak at >20% of hybrids in Guinea Bissau) observed in the extreme west of the species' sympatric range (Caputo *et al.* 2008, 2011; Oliveira *et al.* 2008; Niang *et al.* 2014) offer a unique

opportunity to investigate this phenomenon. Genomic studies on Guinea Bissau samples showed that only the centromere-proximal region of the chromosome-X retains the levels of very low intraspecific polymorphism and high interspecific divergence found elsewhere in the species range (Marsden et al., 2011; Weetman *et al.* 2012; Nwakanma *et al.* 2013). However, inconsistencies between species diagnostic markers (i.e. IGS-SNPs and SINE-X) in field females from this putative secondary contact zone are suggestive of recombination within the X-island (Caputo *et al.* 2011; Santolamazza *et al.* 2011). In this study we address the key question of whether recombination can impact the integrity of the primary candidate speciation island. By typing males from the extreme hybridization zone in Guinea Bissau, and thus overcoming the problems of extremely low recombination rates in the X-peri-centromeric area, and of ambiguous signals from previous genotyping of females. We typed markers throughout the genome, but with specific enrichment in the approximately 4 Mb chromosome-X island where hemizygosity permits unambiguous determination of haplotypes in males. Our results conclusively document evidence for recombination in the X-island, but the nature of the patterns observed suggests that gene conversion rather than meiotic crossing-over may be the more prevalent mechanism.

MATERIALS AND METHODS

Anopheles gambiae s.l. samples and genotyping.

Indoor resting mosquitoes were collected by manual aspiration (Coluzzi & Petrarca 1973), between October 10th and 30th 2010 from three villages situated along a west to east transect spanning from coastal to inland Guinea Bissau (Figure 1): Safim (11°57'24.8"N, 15°38'57.2"W), Mansoa (12°04'33.7"N, 15°19'16.3"W) and Leibala (12°16'18.3"N, 14°13'20.8"W). For comparison with mosquito samples from a typically-low hybridization area, males were also collected in October 2008 by indoor resting captures in two villages in Burkina Faso: Vallée du Kou (11°24' N, 4°25' W) and Soumousso (11°00' 46" N,

4°02'45"W), where *An. coluzzii* and *An. gambiae* prevail, respectively (Dabiré *et al.* 2008). Males and females were morphologically identified as belonging to the *An. gambiae* s.l. species complex based on Gillies & Coetzee (1987). DNA was extracted either from legs or wings of *An. gambiae* s.l. specimens using DNAzol® (Invitrogen).

Specimens were identified based on 'IMP-PCR' detection of two diagnostic SNPs in the IGS-rDNA region (Wilkins *et al.*, 2006; Santolamazza *et al.*, 2011) and on PCR-detection of the *An. coluzzii*-specific SINE insertion, approximately 1.5 Mb away from IGS (SINE-X; Santolamazza *et al.*, 2008). A subset of 29 IGS-genotypes were validated by sequencing (BMR s.r.l., Padua, Italy) a PCR-fragment including the two species-specific IGS-SNPs identified by the IMP-PCR using primers in Scott *et al.* (1993) (GenBank accession numbers KX828849 - KX828877; Table. S1). A subset of 15 SINE-X genotypes were validated by sequencing a PCR-fragment including regions flanking the SINE insertion where, importantly, two species-specific SNPs are also present (GenBank accession numbers KX828878 - KX828892; Table. S2; Santolamazza *et al.* 2008;).

We also developed a simple, novel PCR approach to detect an additional putatively *An. gambiae*-specific insertion polymorphism. The 57-bp insertion is located in the fourth intron of the *CYP4G16* gene (gene identifier AGAP001076; coordinates X: 22937392-22947129), located approximately 7kb from the SINE-X insertion site (Turner et al. 2005). Unlike the SINE and IGS markers, CYP is situated in a gene potentially involved directly in the process of differentiation and/or introgression between the two species and has been hypothesized to be involved in the process of steroidogenesis in *An. gambiae* (Pondeville *et al.* 2013) and in the regulation of female fertility and mating behavior and success (Baldini *et al.* 2013; Gabrieli *et al.* 2014), in addition to playing a role in insecticide resistance (Jones *et al.* 2013; Matowo *et al.* 2014), potentially via cuticular alteration (Qiu *et al.* 2012). The primers flanking the insertion (Cyp4G16fw: 5'-AATTTCACTCTACATCTACAG-3' and Cyp4G16rev: 5'-AAACATGTAAGGAAGTAGTGG-3') amplify fragments of 497 bp and 554 bp in *An. coluzzii*

and *An. gambiae*, respectively. PCR reactions were carried out in 15 µl reactions containing 3 pmol of each primer, 0.1 mM of each dNTP, 1.5 mM $MgCl_2$, 0.75 U Taq polymerase, and 1 µl of DNA template. Thermocycler conditions were 94°C for 10 min followed by 35 cycles of 94°C for 30 s, 54°C for 30 s and 72°C for 1 min, with a final elongation step of 72°C for 10 min. The species-specificity of the insertion was validated in samples collected from across the geographical range of the two species (Table S3). Genotyping results were confirmed by sequencing on 84 specimens (GenBank accession numbers KX828893 - KX828976).

For a subsample of males genotyped using the three markers above, DNA was extracted from the remainder of the carcasses using the DNeasy®-Blood and tissue kit (Qiagen) and whole genome amplified using the Genomiphi® kit (GE Healthcare). Genomes were then analyzed using two multilocus SNP-genotyping approaches. In the first approach samples were screened at 1,536 SNPs spread across all chromosomes using a custom Illumina SNP chip (see Weetman *et al.* (2014) for details and genomic locations of the SNPs). Standard Illumina laboratory protocols were followed to screen the chip on a Beadstation 500 GX, with established protocols for inclusion of SNPs, e.g. showed good clustering of genotypes, 80% call rate and no evidence for null alleles, and polymorphism in at least one comparison (Weetman *et al.* 2010, 2012, 2014). To target species-informative SNPs in a focal region proximate to the X-centromere (i.e. at AGAMP3 genomic positions exceeding 20 Mb in the assembled 24.4 Mb chromosome-X), a second genotyping approach was adopted. Peri-centromeric SNPs on chromosome-X were chosen based on fixed differences between *An. coluzzii* and *An. gambiae* identified in whole genome sequences from Ghana (Clarkson *et al.* 2014). Two Sequenom iPLEX® multiplexes were designed to genotype a total of 35 X-pericentromeric species-specific SNPs, yielding a density of 1 SNP per 0.12 Mb across the 4 Mb region targeted (Supporting Information 1). Out of the 35 SNPs scored, 31 were genotyped successfully (call rates above 80%). iPLEX® genoptyping was validated technically by inclusion of 3 SNPs in both multiplexes: for both that genotyped successfully in both multiplexes, results were fully concordant.

Statistical analyses

To evaluate correspondence between the IGS and SINE-X standard species-diagnostics and genome-wide differentiation based on Illumina genotyping, we applied individual-based clustering analysis to the multilocus genotypes of males from Safim (Guinea Bissau) and Burkina Faso using the Bayesian algorithm implemented by BAPS 5.4 (Corander & Marttinen 2006; Corander *et al.* 2008). Owing to the presence of multiple inversion polymorphisms on chromosome-2, clustering analysis was based on SNPs from chromosomes-X and -3, and was performed separately owing to hemizygosity of chromosome-X. Each run was repeated several times to check that the optimal clustering solution was obtained. Based on clustering results, admixture analyses were performed to permit the identification of significantly mixed individual genotypes, and the estimated proportions of their multilocus genotypes attributable to each cluster.

GENEPOP 4.2 (Rousset 2008) was used to compute F_{ST} values and to test differentiation at individual SNPs in pairwise comparisons. In order to highlight inter-specific differentiation and to compare results from Safim with those from Burkina Faso, we excluded from this analysis 31 specimens not assigned consistently to either *An. coluzzii* or *An. gambiae* by SINE-X and IGS diagnostic markers.

ARLEQUIN 3.5.1.2. was used to compute linkage disequilibrium (LD) measured as r^2 and D' between the two standard species diagnostic markers, SINE-X and IGS, and also the putative species diagnostic marker, CYP*,* genotyped in male specimens from Guinea Bissau, using default Markov chain settings to compute significance.

RESULTS

IGS and SINE-X diagnostic genotyping

IGS and SINE-X species diagnostic markers were genotyped in 264 *An. gambiae* s.l. males and 329 females collected from the three villages in Guinea Bissau, along with 33 males from Burkina Faso. IGS and SINE-X species diagnoses were entirely concordant for specimens from Burkina Faso, but not for those from Guinea Bissau (Figure 1). For clarity, specimens showing concordant IGS and SINE-X genotypes are hereafter referred to either as *An. coluzzii* (i.e. $SINE^{CO}/IGS^{CO}$), *An. gambiae* (i.e. $SINE^{GA}/IGS^{GA}$) or nominally "F1-hybrids" (i.e. $SINE^{CO/GA}/IGS^{CO/GA}$ females), while specimens showing discordant IGS/SINE-X genotypes are termed IGS/SINE-X recombinants.

Anopheles gambiae was almost the sole species found in the easternmost Guinean village (Leibala), but was found in sympatry with *An. coluzzii* in the two western sites, where higher frequencies of F1-hybrid females were also found (22.9% in Safim and 6.52% in Mansoa, compared to 1.8% in Leibala) (Figure 1). IGS/SINE-X recombinant genotypes were found in both males (18.6%) and females (10.9%) (Chi-sq=6.32 ; df=1, $p<0.05$; Table 1), although at a different frequency among the three villages (males: Chi-sq=25.158; df=2; $p<0.0001$; females: Chi-sq=9.44; df=2; $p<0.01$). These were mostly $SINE^{GA}$ specimens with $IGS^{CO/GA}$ genotypes (Safim: 16.7% in males and 11.89% in females; Mansoa: 6% and 6.52%; absent at Leibala). The discordant $SINE^{GA}$-IGS^{CO} genotype was found only in males (Safim 8.3%; Mansoa: 16.4%). Thus, discordant, recombinant genotypes within individuals were only present in the sample locations towards the coast of Guinea Bissau, where a mix of species and detection of hybrids were also highest.

SNP-genotyping

A total of 61 males from Guinea Bissau (Safim) and 33 males from Burkina Faso were Illumina-genotyped at 711 SNPs (43 on chromosome-X; 411 on chromosome-2; 257 on chromosome-3; Supporting Information 2). In Burkina Faso, high interspecific differentiation

was evident on each of the three chromosomes, whereas in Guinea Bissau, autosomal differentiation was very limited, and the only strongly divergent SNPs were on chromosome-X, primarily toward the centromere (Figure 2). Using either SNPs on chromosome-X or those on chromosome-3, two clusters matching *An. coluzzii* and *An. gambiae* in Burkina Faso were identified by BAPS analysis, with just one individual showing evidence of significant admixture on chromosome-3 (Figure 3). In contrast, clustering provided different pictures from different chromosomes in Guinea Bissau. Based on chromosome-X SNPs, two clusters were identified, one of which included all *An. coluzzii* (with a single exception) and 14 IGS/SINE-X recombinant specimens, while the other included all *An. gambiae* (again, with a single exception) and 17 IGS/SINE-X recombinants (Figure 3a). Yet, based on chromosome-3 SNPs (Figure 3b), in Guinea Bissau the cluster that corresponded to the *An. gambiae* cluster from Burkina Faso included all *An. gambiae,* all but one *An. coluzzii* and most IGS/SINE-X recombinant individuals. Therefore, all but one of the Guinea Bissau males that resembled Burkina Faso *An. coluzzii* based on chromosome-3 SNPs were actually identified as IGS/SINE-X recombinants.

Focal, higher resolution genotyping of the pericentromeric region of chromosome-X spanning approximately 4 Mb was successfully carried out for 91 of the 94 males genotyped using the Sequenom assay. Perfect linkage disequilibrium among all 31 SNPs was observed in Burkina Faso, where no IGS/SINE-X recombinants were identified (Figure 4). In the Guinea Bissau sample, however, three individuals showed evidence of recombination (each with multiple breakpoints) across the genotyping panel, despite the presence of 30 IGS/SINE-X recombinant males. Two of these 3 individuals showed signs of recombination immediately next to the SINE-X marker (Figure 4).

The lack of consistency between Sequenom- and IGS/SINE-X haplotypes is mostly due to two groups of males: 1) individuals characterized by either an *An. gambiae* (13/29) or an *An. coluzzii* (6/26) multilocus haplotype and showing mixed IGS arrays); and 2) individuals (8/29)

characterized by an *An. coluzzii* multilocus haplotype, but lacking the SINE insertion, which is specific to *An. coluzzii* in the rest of the species range. Sequences from the region flanking the SINE insertion (Santolamazza *et al*. 2008) in the latter specimens (group 2 above) showed the presence of two *An. gambiae*-specific SNPs (Table S2), supporting the hypothesis that this genomic region was acquired from *An. gambiae* by introgression, against alternative hypothesis that the latter specimens have never acquired or have lost secondarily the SINE-insertion, by some alternative mechanism.

CYP marker genotyping

To investigate fine-scaled recombination patterns further at an independently-genotyped locus, presence/absence of the putatively species diagnostic insertion in the *CYP4G16* gene (termed CYP) was scored in 150 *An. gambiae* and 119 *An. coluzzii* females from 6 African countries and in females and males from Guinea Bissau typed previously using IGS and SINE-X. Outside of Guinea Bissau, CYP is fixed in *An. gambiae* and absent in *An. coluzzii*, yielding perfect 3-locus LD with IGS and SINE-X (Table S3) and supporting the validity of CYP as a species diagnostic. In Guinea Bissau, however, the three markers were not in LD in 15% of females and 23% of males (Table S4). CYP, SINE-X and IGS haplotype data from the latter Guinean males showed that (Figure 5): i) the most frequent class of recombinants arose from $CYP^{GA}/SINE^{GA}$ males carrying mixed IGS arrays (35.5%), while $CYP^{CO}/SINE^{CO}$ males carrying mixed IGS arrays were far less common (4.8%); ii) LD appears to be stronger between CYP and IGS (r^2=0.69, even though more than 1Mb apart on the chromosome), than between CYP and SINE-X (r^2=0.48, located only about 7kb from one other); iii) more than 25% of recombinant haplotypes were $CYP^{CO}/SINE^{GA}/IGS^{CO}$, while "opposite" $CYP^{GA}/SINE^{CO}/IGS^{GA}$ haplotypes were never observed.

DISCUSSION

Recently diverged species of the *An. gambiae* complex represent a very good model to study inter-specific gene-flow (Turner & Hahn, 2010; Fontaine *et al.* 2015). Conditions under which interspecific hybridization is favored (at least temporarily) have been repeatedly observed within the complex most notably for adaptive introgression of the autosomal *kdr 1014F* insecticide resistance mutation without any detectable long term impact on reproductive isolation (Lee *et al.* 2013; Clarkson *et al.* 2014; Norris *et al.* 2015). In the extreme west of the *An. gambiae* and *An. coluzzii* range, however, the partial breakdown of reproductive barriers and/or of selective pressures against hybrids seems to have allowed the two species to mate with each other and/or with hybrids far more freely and viably than elsewhere. This has resulted in widespread autosomal mixture (Weetman *et al.* 2012; Nwakanma *et al.* 2013) including introgression of potentially adaptive variants, such as the immune-resistant *TEP1r1* (Mancini *et al.* 2015). In the hybrid zone introgression may also dramatically increase genetic variability in autosomal centromeric regions of typically low diversity, as detected by genotyping of intron-1 of the para Voltage-Gated Sodium Channel gene on chromosome arm 2L (Santolamazza *et al.* 2015).

In the present study, by genotyping *An. coluzzii* and *An. gambiae* (chromosome-X hemizygous) males from the extreme hybrid zone in coastal Guinea Bissau using different molecular markers, we detected a complete loss of autosomal differentiation, including at the two islands of divergence near the autosomal centromeres. Moreover, we identified a complex pattern of recombination within the largest and only universally-observed genomic island in the low-recombination centromeric region of chromosome-X (Weetman *et al.* 2012; Nwakanma *et al.* 2013) due to recombination among rDNA repeats and to what appears most likely to be gene conversion.

Recombination in the rDNA multi-copy DNA region

Recombination among rDNA repeats was definitively demonstrated to occur in the hybrid zone by virtue of the presence of copies of both species-specific IGS-alleles in hemizygotic males characterized by otherwise "pure" SNP-genotypes, confirming previous suggestions based on hybrid female data (Caputo *et al.* 2011; Santolamazza *et al.* 2011). No mixed IGS arrays were observed in males from the inland population (Leibala) where hybrid females are rare (Vicente et al., Submitted), suggesting that recombination within rDNA is indeed associated with inter-specific hybridization, and thus may be primarily localized to the coastal region. Multicopy rDNA arrays are homogenized rapidly by the process of concerted evolution and mixed arrays though very occasionally detected, e.g. in 11 *Drosophila* species (Stage & Eickbush 2007), are unlikely to persist at high frequencies in a natural population (Eickbush & Eickbush 2007. With relatively frequent detection here, our results are consistent with a process of high interspecific hybridization and introgression in the secondary contact zone in coastal Guinea Bissau ongoing, as highlighted by other studies in the *Anopheles gambiae* species pair (Marsden *et al.* 2011; Weetman *et al.* 2012) and also in a grasshopper *Podesma pedestris* hybrid zone (Keller *et al.* 2008).

Recombination: meiotic crossovers or gene conversion?

The presence in the genotyped sample of seven males characterized by a pure *An. coluzzii* multilocus SNP-genotype but lacking the species-specific SINE-X insertion suggests that recombination does not occur homogeneously within the single-copy DNA of the 4 Mb X-centromeric island. Detection of *An. gambiae*-specific SNPs in the regions flanking the SINE-X insertion (Santolamazza *et al.* 2008) in these males supports the hypothesis that the SINE-X insertion was lost due to introgression from *An. gambiae*, rather than either being polymorphic in the *An. coluzzii* population, or being excised once inserted (a mechanism not known to occur, Shedlock & Okada 2000). Surprisingly, multiple instances of discordance were also observed between the SINE-X and CYP markers, despite their close proximity. Given the improbability of double-recombinants at such a fine scale, this suggests a

prevalence of gene conversion in the region neighboring SINE-X insertion site. Gene conversion typically involves small (<2kb) genomic regions and breaks the expected relationship between LD and distance (Talbert & Henikoff, 2010). Gene conversion may actually be a hallmark of pericentromeric regions and a key, but – given it's small scale - easily overlooked, mechanism of genetic exchange (Korunes & Noor, 2016). In *Drosophila melanogaster* the ratio of crossovers to diversity is positive, but that between the gene conversion/crossover rate and diversity is strongly negative (Comeron *et al.* 2012), i.e. in regions of low diversity such as toward centromeres, gene conversion becomes increasingly important (Bhakta *et al.* 2015). Owing to the relatively limited number of males for which multilocus data are available it is difficult to conclude as to whether crossover-associated or non-crossover gene conversion may be more prevalent. However, the observation that two out of the three individuals for which evidence for recombination was detected in the multilocus data exhibited a neighboring discordant SINE-X genotype suggests that crossover-associated gene conversion might have occurred.

Gene conversion may also readily lead to asymmetry in genetic exchange. Genotyping of the *An. gambiae*-specific insertion in *CYP4G16*-Intron4 in a larger male sample (N=264) than that genotyped by multilocus SNPs revealed a strong asymmetry in frequency of the observed recombinant haplotypes (Figure 5), with a lack of CYP^{GA}-$SINE^{CO}$-IGS^{GA} males (which would be indicative of SINE insertion into an *An. gambiae* genotype) and a relatively high frequency of CYP^{CO}-$SINE^{GA}$-IGS^{CO} males. A gene conversion hot-spot in proximity of the SINE insertion, suggested by $SINE^{GA}$ in males otherwise characterized by an *An. coluzzii* X-centromere genotype, could be aided by the large number of transposable elements and repeats close to SINE-insertions (Feschotte & Pritham 2007; Oliver & Greene 2009). This hypothesis of a hotspot is further reinforced by the different frequencies of recombinant individuals revealed by the two sets of data (i.e. 5% among multilocus SNPs compared to 12% between CYP and SINE).

Species diagnosis and studies of introgression

A practical implication of these results is a limited cross-prediction possible between diagnostic markers and multilocus genotypes. *Anopheles gambiae* and *An. coluzzii* are currently defined by the IGS and (secondarily) the SINE-X diagnostic markers. Barring occasional methodological issues, these markers are concordant and highly reliable across the species range (Santolamazza *et al.* 2011). However the markers themselves are of no known adaptive value with respect to divergence between the species, and therefore their utility relies upon LD with key species-defining functional genetic variants, which remain to be discovered, but appear highly likely to be located within the same genomic region (Aboagye-antwi *et al.* 2015). Whilst genomewide differentiation may often be a correlate of variation on chromosome-X, comparative genomic work has shown that this is not true among the other species of the *An. gambiae* complex (Fontaine *et al.* 2015). Although the IGS and SINE-X species diagnostics are good representatives of the X peri-centromeric region of *An. coluzzii* and *An. gambiae* in the majority of their range, a SNP multiplex assay such as that applied here would be more appropriate to represent the majority of the X peri-centromeric region and to study adaptive introgression. This is likely to apply not only to the hybrid zone of Guinea Bissau but also to neighboring countries in the "far-west" region where interspecific hybridization has been reported, although typically at lower rates than in Guinea Bissau (Caputo *et al.* 2008; Nwakanma *et al.* 2013; Niang *et al.* 2014).

Asymmetric introgression from *An. coluzzii* to *An. gambiae* has been hypothesized in Guinea Bissau to explain far greater similarity of IGS-diagnosed hybrids to *An. gambiae* than to *An. coluzzii* in females genotyped at approximately 50 SNP markers from mixed genomic locations (Marsden *et al.* 2011). Sequence data from female X-chromosomes in hybrids was also found to show a comparable pattern of inequality between species (Nwakanma *et al.* 2013), and both studies were interpreted as evidence for directional backcrossing of hybrids to *An. gambiae*, rather than to *An. coluzzii*. From our chromosome-3 SNP-genotype data, all but one of the males from Guinea Bissau clustered with Burkina Faso *An. gambiae*

regardless of species, while chromosome-X SNPs exhibited species-specific clustering. This might appear suggestive of biased gene flow leading to autosomal replacement in *An. coluzzii*. However, limited autosomal differentiation, coupled with hemizygosity, and evidently low recombination among the SNPs used for clustering on chromosome-X, mean that hybrids and backcrosses could not be identified from the data in the same way as for females and so initial definitions of 'pure' species may have been inaccurate. Indeed, perspectives on asymmetry of introgression may need to be revised by identifying the species based on aggregate multilocus analysis of the X peri-centromeric region. However, apparent loss of the SINE insertion in some *An. coluzzii* chromosomes, yet a complete lack of insertion observed in *An. gambiae,* could be consistent with asymmetric introgression from the latter, as suggested by autosomal data.

Conclusion

The secondary contact zone between the two principal and recently diverged Afro-tropical malaria vectors provides a "natural" laboratory to study in real time the genomic effects of hybridization and introgression. Genomic variation in the chromosome-X of hemizygotic males unveiled a complex pattern of recombination, which we hypothesize most likely reflects a combination of relatively infrequent crossover and more frequent, but localized, gene conversion operating within the major candidate 'genomic island of speciation'. Our discovery of the existence of X-centromere recombinants involving multiple breakpoints suggests that the potential exists for re-assortment of variants to avoid progressive long-term fitness loss via a Hill-Robertson effect (Hill & Robertson 1966; Birky & Walsh 1988; Yeaman 2013), which could enhance the likelihood of key involvement of this chromosomal area in adaptive divergence (Wu 2001; Wu & Ting 2004). The importance of gene conversion in both the break-up of LD and asymmetric introgression has received relatively limited attention to date, but warrants further investigation (Korunes & Noor 2016).

Moreover, these results have practical implications by disclosure of very limited cross-prediction between species-specific diagnostic markers and multilocus genotypes. While this is relevant from the evolutionary biology perspective, it remains to be addressed if malaria control programs in regions of high hybridization should invest in more expensive sophisticated technologies for this purpose. Decisions need to be made in light of potential differences between species (and hybrids) in vectorial capacity or in responses to vector control.

Acknowledgements

The authors wish to thank the people of Guinea Bissau for their hospitality and goodwill during mosquito collections. We thank Ana Alfiveric and Daniel Carr of the Institute of Translational Medicine, University of Liverpool, for help with design and scoring of Sequenom assays.

References

Aboagye-antwi F, Alhafez N, Weedall GD, Brothwood J, *et al.* (2015) Experimental swap of *Anopheles gambiae* ' s assortative mating preferences demonstrates key role of X-chromosome divergence island inincipient sympatric speciation. *Plos Genetics* **11**, 1–19.

Baldini F, Gabrieli P, South A et al. (2013) The Interaction between a Sexually Transferred Steroid Hormone and a Female Protein Regulates Oogenesis in the Malaria Mosquito *Anopheles gambiae*. PLoS Biology, **11**. e1001695

Bhakta MS, Jones VA, Vallejos CE (2015) Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris* L. *PloS ONE*, **10**, e0116822.

Birky CW, Walsh JB (1988) Effects of linkage on rates of molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 6414–8.

Butlin RK (2005) Recombination and speciation. *Molecular Ecology*, **14**, 2621–2635.

Caputo B, Nwakanma D, Jawara M *et al.* (2008) *Anopheles gambiae* complex along The Gambia river, with particular reference to the molecular forms of *An. gambiae* s.s. *Malaria Journal*, **7**, 182.

Caputo B, Santolamazza F, Vicente JL *et al.* (2011) The “far-west” of *Anopheles gambiae* molecular forms. *PLoS ONE*, **6**, 1–7.

Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. *The American Naturalist*, **130**, pp. 113–146.

Clarkson CS, Weetman D, Essandoh J *et al.* (2014) Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. *Nature Communications*, **5**, 4248.

Coetzee M, Hunt RH, Wilkerson R *et al.* (2013) *Anopheles coluzzii* and *Anopheles* amharicus, new members of the *Anopheles gambiae* complex. *Zootaxa*, **3619**, 246–274.

Coluzzi M, Petrarca V (1973) Aspirator with paper cup for collecting mosquitoes and other insects. *Mosquito News*, **33**, 249–250.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science*, **298**, 1415–1418.

Comeron JM, Ratnappan R, Bailin S (2012) The Many Landscapes of Recombination in *Drosophila melanogaster*. *PLoS Genetics*, **8**, 33–35. e1002905

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, **15**, 2833–2843.

Corander J, Marttinen P, Sirén J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

della Torre A, Fanello C, Akogbeto M *et al.* (2001) Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West AfricAn. *Insect Molecular Biology*, **10**, 9–18.

della Torre A, Tu Z, Petrarca V (2005) On the distribution and genetic differentiation of *Anopheles gambiae* s.s. molecular forms. *Insect Biochemistry and Molecular Biology*, **35**, 755–769.

Dabiré KR, Diabaté A, Djogbenou L *et al.* (2008) Dynamics of multiple insecticide resistance in the malaria vector *Anopheles gambiae* in a rice growing area in South-Western Burkina Faso. *Malaria Journal*, **7**, 188.

Diabaté A, Dabire RK, Millogo N, Lehmann T (2007) Evaluating the effect of postmating isolation between molecular forms of *Anopheles gambiae* (Diptera : Culicidae). *Journal of Medical Entomology*, **44**, 60–64.

Diabaté A, Dao A, Yaro AS *et al.* (2009) Spatial swarm segregation and reproductive isolation between the molecular forms of *Anopheles gambiae*. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* **276**, 4215–4222.

Eickbush TH, Eickbush DG (2007) Finely orchestrated movements: Evolution of the ribosomal RNA genes. *Genetics*, **175**, 477–485.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–50.

Feschotte C, Pritham EJ (2007) DNA Transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, **41**, 331–369.

Fontaine MC, Pease JB, Steele A *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**, 1258524–1258524.

Gabrieli P, Kakani EG, Mitchell SN et al. (2014) Sexual transfer of the steroid hormone 20E induces the postmating switch in *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 16353–8.

Gillies MT, Coetzee M (1987) A Supplement to the Anophelinae of Africa South of the SaharAn. *Publications of the South African Institute for medical research*, **55**.

Gimonneau G, Pombi M, Dabiré RK *et al.* (2012) Behavioural responses of *Anopheles gambiae* sensu stricto M and S molecular form larvae to an aquatic predator in Burkina Faso. *Parasites & Vectors*, **5**, 65.

Hill WG, Robertson A, (1966) The effect of linkage on limits to artificial selection. *Genetical Research*, **8**, 269–294.

Jones CM, Haji K a, Khatib BO et al. (2013) The dynamics of pyrethroid resistance in *Anopheles arabiensis* from Zanzibar and an assessment of the underlying genetic basis. *Parasites & Vectors*, **6**, 343.

Keller I, Veltsos P, Nichols RA (2008) The frequency of rDNA variants within individuals provides evidence of population history and gene flow across a grasshopper hybrid zone. *Evolution*, **62**, 833–844.

Korunes KL, Noor MAF (2016) Gene Conversion and Linkage: Effects on Genome Evolution and Speciation. *Molecular Ecology*, doi: 10.1111/mec.13736

Lee Y, Marsden CD, Norris LC *et al.* (2013) Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 19854–9.

Lehmann T, Diabate A (2008) The molecular forms of *Anopheles gambiae*: A phenotypic perspective. *Infection, Genetics and Evolution*, **8**, 737–746.

Mancini E, Spinaci MI, Gordicho V *et al.* (2015) Adaptive potential of hybridization among malaria vectors: introgression at the immune locus TEP1 between *Anopheles coluzzii* and AN. *gambiae* in “Far-West” AfricAn. *Plos ONE*, **10**, e0127804.

Marsden CD, Lee Y, Nieman CC *et al.* (2011) Asymmetric introgression between the M and S forms of the malaria vector, *Anopheles gambiae*, maintains divergence despite extensive hybridization. *Molecular Ecology*, **20**, 4983–4994.

Matowo J, Jones CM, Kabula B et al. (2014) Genetic basis of pyrethroid resistance in a population of *Anopheles arabiensis*, the primary malaria vector in Lower Moshi, north-eastern Tanzania. *Parasites & Vectors*, **7**, 274.

Nam K, Munch K, Hobolth A *et al.* (2015) Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proceedings of the National Academy of Sciences*, **112**, 201419306.

Neafsey DE, Lawniczak MKN, Park DJ *et al.* (2010) SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science*, **330**, 514–517.

Niang EHA, Konaté L, Diallo M, Faye O, Dia I (2014) Reproductive isolation among sympatric molecular forms of *An. gambiae* from inland areas of south-eastern Senegal. *PloS ONE*, **9**, e104622.

Noor M a F, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

Norris LC, Main BJ, Lee Y *et al.* (2015) Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proceedings of the National Academy of Sciences*, **112**.

Nwakanma DC, Neafsey DE, Jawara M *et al.* (2013) Breakdown in the process of incipient speciation in *Anopheles gambiae*. *Genetics*, **193**, 1221–1231.

Oliveira E, Salgueiro P, Palsson K *et al.* (2008) High levels of hybridization between molecular forms of *Anopheles gambiae* from Guinea Bissau. *Journal of Medical Entomology*, **45**, 1057–1063.

Oliver KR, Greene WK (2009) Transposable elements: Powerful facilitators of evolution. *BioEssays*, **31**, 703–714.

Pondeville E, David JP, Guittard E et al. (2013) Microarray and RNAi analysis of P450s in *Anopheles gambiae* male and female steroidogenic tissues: CYP307A1 is required for ecdysteroid synthesis. *PLoS ONE*, **8**. e79861

Qiu Y, Tittiger C, Wicker-Thomas C et al. (2012) An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 14858–14863.

Reidenbach KR, Neafsey DE, Costantini C *et al.* (2012) Patterns of genomic differentiation between ecologically differentiated M and S forms of *Anopheles gambiae* in West and Central AfricAn. *Genome Biology and Evolution*, **4**, 1202–1212.

Rousset F (2008) GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Santolamazza F, Caputo B, Calzetta M *et al.* (2011) Comparative analyses reveal discrepancies among results of commonly used methods for *Anopheles gambiae*molecular form identification. *Malaria Journal*, **10**, 215.

Santolamazza F, Caputo B, Nwakanma DC *et al.* (2015) Remarkable diversity of intron-1 of the para voltage-gated sodium channel gene in an *Anopheles gambiae*/*Anopheles*

coluzzii hybrid zone. *Malaria Journal*, **14**, 1–10.

Santolamazza F, Mancini E, Simard F *et al.* (2008) Insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms. *Malaria Journal*, **7**, 163.

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

Shedlock AM, Okada N (2000) SINE insertions: powerful tools for molecular systematics. *Bioessays*, **22**, 148–160.

Slotman M, della Torre A, Calzetta M, Powell JR (2005) Differential introgression of chromsomal regions between *Anopheles gambiae* and *An. arabiensis*. *American Journal of Tropical Medicine and Hygiene*, **73**, 326–335.

Slotman M, della Torre AN., Powell JR (2004) The genetics of inviability and male sterility in hybrids between *Anopheles gambiae* and *An. arabiensis*. *Genetics*, **167**, 275–287.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123–5140.

Stage DE, Eickbush TH (2007) Sequence variation within the rRNA gene loci of 12 *Drosophila* species. Genome research, 17, 1888–1897.

Talbert PB, Henikoff S (2010) Centromeres Convert but Don't Cross. *PLoS Biology*, **8**, e1000326.

Turner TL, Hahn MW (2010) Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, **19**, 848–50.

Turner TL, Hahn MW, Nuzhdin SV. (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, 1572–1578.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9939–9946.

Weetman D, Steen K, Rippon EJ *et al.* (2014) Contemporary gene flow between wild *An. gambiae* s.s. and *An. arabiensis*. *Parasites & Vectors*, **7**, 345.

Weetman D, Wilding CS, Steen K *et al.* (2010) Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: Major variants identified in a low-linkage disequilbrium genome. *PLoS ONE*, **5**.

Weetman D, Wilding CS, Steen K, Pinto J, Donnelly MJ (2012) Gene flow-dependent genomic divergence between *Anopheles gambiae* M and S forms. *Molecular Biology and Evolution*, **29**, 279–291.

White BJ, Cheng C, Simard F, Costantini C, Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology*, **19**, 925–939.

Wilkins EE, Howell PI, Benedict MQ (2006) IMP PCR primers detect single nucleotide polymorphisms for *Anopheles gambiae* species identification, Mopti and Savanna rDNA types, and resistance to dieldrin in *Anopheles arabiensis*. *Malaria Journal*, **5**, 125.

Wilkins EE, Howell PI, Benedict MQ (2007) X and Y chromosome inheritance and mixtures of rDNA intergenic spacer regions in *Anopheles gambiae*. *Insect Molecular Biology*, **16**, 735–741.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Wu C-I, Ting C-T (2004) Genes and speciation. *Nature Reviews. Genetics*, **5**, 114–122.

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E1743–51.

Data accessibility

Table S1, S2, S3 and S4 are included in the Supplementary Material.
Information on the design of the two Sequenom iPLEX multiplexes used to genotype a total of 35 X-pericentromeric species-specific SNPs and results obtained for Whole-Genome-SNP-genotyping using the Illumina-platform are uploaded in the Dryad repository as Supporting Information1 (SNPs genotyped by Sequenom-assay) and Supporting Information 2 (Illumina-SNP-genotyping results); Dryad DOI: http://dx.doi.org/10.5061/dryad.7ck7r .
Sequences are released in GenBank with the following accession numbers:
IGS sequences for males (N=29) from Guinea Bissau: GenBank accession numbers KX828849 - KX828877; SINE sequences for males (N=15) from Guinea Bissau: GenBank accession numbers KX828878 - KX828892; CYP sequences for females (N=84) from 6 African countries: GenBank accession numbers KX828893 - KX828976.

Authors' contributions

B.C., J.P., A.d.T., D.W. conceived and designed the experiments; B.C., Ve.P., E.M., M.P., J.L.V., J.D., K.S., A.R., J.P., D.W., performed and analysed experiments; B.C., Ve.P., Vi.P., J.P., A.d.T., and D.W. wrote the paper.

Table and Figure captions

Table 1. Frequencies of IGS- and SINE-X-genotypes of *Anopheles gambiae* and *An.coluzzii* collected in Guinea Bissau (N=593). CO=*An. coluzzii*-specific allele; GA=*An. gambiae*-specific allele. Discordant diagnostic results between the two genotyping approaches are in bold.

		% SINE^{CO}	% SINE$^{CO/GA}$	% SINE^{GA}	TOTAL
MALES	**IGS^{CO}**	19.7	-	**8.0**	73
	IGS$^{CO/GA}$	**1.5**	-	**9.1**	28
	IGS^{GA}	-	-	61.7	163
	TOTAL	56	-	208	264
FEMALES	**IGS^{CO}**	10.9	**1.2**	-	40
	IGS$^{CO/GA}$	**0.6**	17	**9.1**	88
	IGS^{GA}	-	-	61.1	201
	TOTAL	38	60	231	329

Figure 1. Distribution of SINE-X- and IGS- genotypes in *Anopheles gambiae* and *An. coluzzii* males and females collected in Guinea Bissau, and in males from Burkina Faso. CO=*An. coluzzii*-specific allele; GA=*An. gambiae*-specific allele.

Figure 2. Genomic differentiation between *Anopheles gambiae* and *An. coluzzii* males collected in Guinea Bissau and Burkina Faso. F*ST* values were computed for polymorphic SNPs: N=43 on chromosome-X; N=411 on chromosome-2; N=257 on chromosome-3. Vertical bars indicate approximate centromere and telomere positions. CO=*An. coluzzii*; GA=*An. gambiae*.Light blue = SNPs on chromosome arm 2R, orange = SNPs on chromosome arm 2L, light green = SNPs on chromosome arm 3R, purple = SNPs on chromosome arm 3L, red = SNPs on X-chromosome.

Figure 3. Admixture results, produced by individual-based (BAPS) clustering analyses, based on male samples from Guinea Bissau (GB, N=61) and Burkina Faso (BF, N=33). Analysis was based on 43 SNPs on chromosome-X (a), and 257 SNPs on chromosome-3 (b). red = *An. coluzzii;* blue =*An. gambiae*. Letters below the graph show sample identification by standard diagnostic PCRs (CO = *An.coluzzii*, GA = *An.gambiae*, REC = recombinant individuals i.e. specimens showing discordant IGS/SINE-X genotypes.

Figure 4. SNP-genotyping through the pericentromeric region of chromosome-X in males from Guinea Bissau and Burkina Faso. Results from the 31 multiplexed SNPs are shown left to right ordered by (AGAMP4) position on chromosome-X; results from IGS- and SINE-X- markers are shown at the left side with physical positions indicated by grey arrows on the top and species-ID reported as defined in the main text. Each genotyped individual is represented by a horizontal line; black arrows indicate individuals identified as recombinants by the Sequenom-SNP-genotyping; red=An. coluzzii typical alleles; blue = An. gambiae typical alleles; yellow = mixed IGS genotypes (i.e. IGS$^{CO/GA}$); grey = missing genotypes.

Figure 5. Recombinant haplotypes among the three X-centromeric species-diagnostic markers (CYP, SINE-X, IGS) in males from Guinea Bissau (in blue= allele typical for *An. gambiae*, in red= allele typical for *An.coluzzii*). Haplotype frequencies are shown with sample sizes in parentheses. Approximate marker positions on chromosome-X centromere: *CYP4G16*: X: 22.937.392-22.947.129; *SINE200 X6.1*: X:22.951.445–22.951.671; IGS arrays start from position X: 23.490.000.

SUPPLEMENTARY MATERIAL

Table S1. **Comparison among genotypes obtained for 29 Guinea Bissau males** by PCR using the CYP-, SINE-X- and IGS-approach and by sequencing a fragment of the IGS-region. CO=*An.coluzzii*-specific allele; GA= *An.gambiae*-specific allele

Table S2. **Comparison among genotypes obtained for 15 Guinea Bissau males** by PCR using the CYP-, SINE-X- and IGS-approach and by sequencing the region including the SINE-insertion. CO=*An.coluzzii*-specific allele; GA= *An.gambiae*-specific allele.
* indicate specimens analyzed also by the Sequenom-approach.

Table S3. **CYP-genotyping results of female mosquitoes** (N=269), collected in different African countries. Species – ID was identified by the IGS-approach (Wilkins *et al.* 2006) and confirmed by the SINE-X –approach (Santolamazza *et al.* 2008). CO=*An.coluzzii;* GA= *An.gambiae;* CO/GA= Hybrid result. Linkage disequlibrium values are shown as D' and r^2.

Table S4. **Numbers (and frequencies) of *An. gambiae, An.coluzzii* and recombinant males (left, N=264) and females (right, N=329)** collected at three sites in Guinea Bissau and genotyped with the IGS-approach (Wilkins et al. 2006), the SINE-X-approach (Santolamazza et al. 2008) and the CYP-approach. Concordant results in boldface. CO=*An.coluzzii*-specific allele; GA= *An.gambiae*-specific allele.

Figure 1.

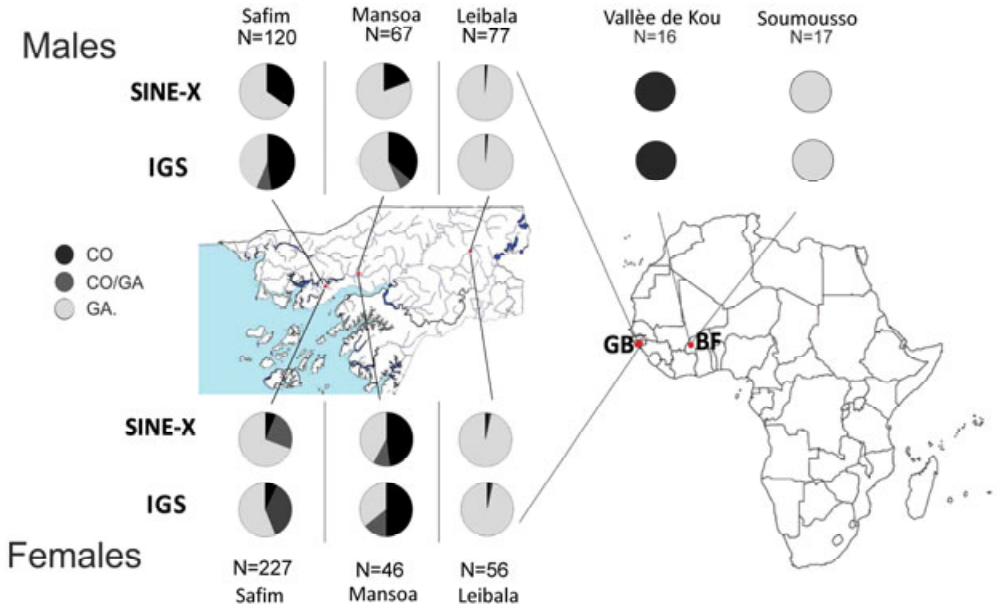

Figure 2.

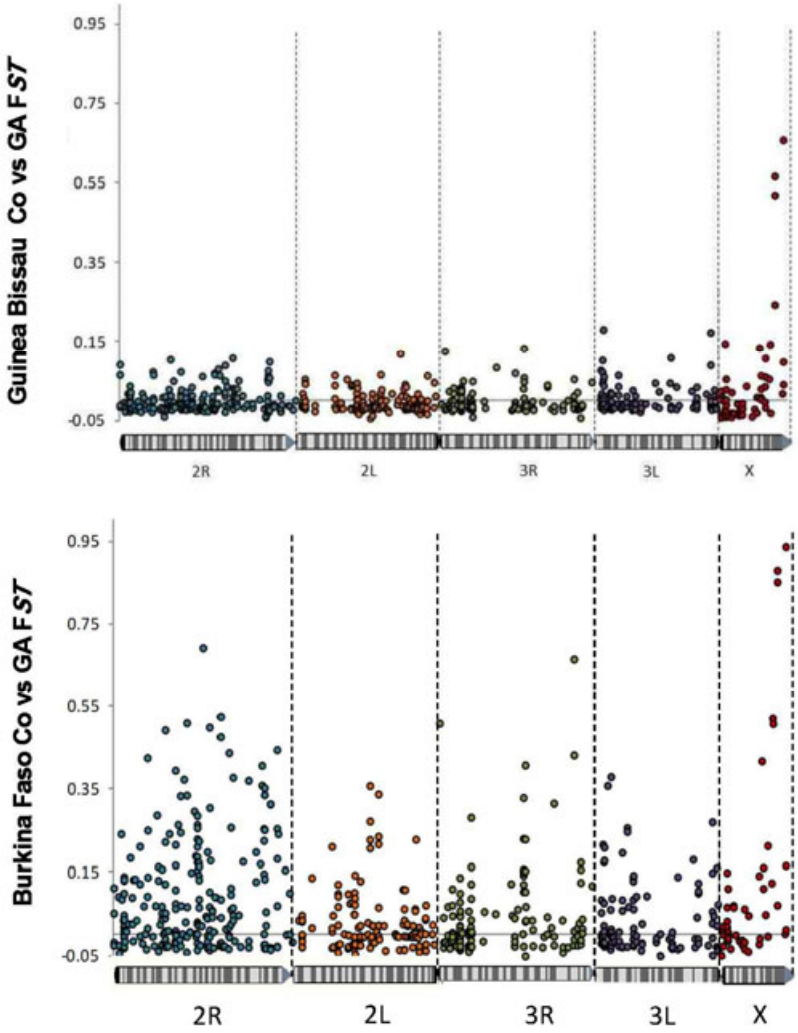

Figure 3.

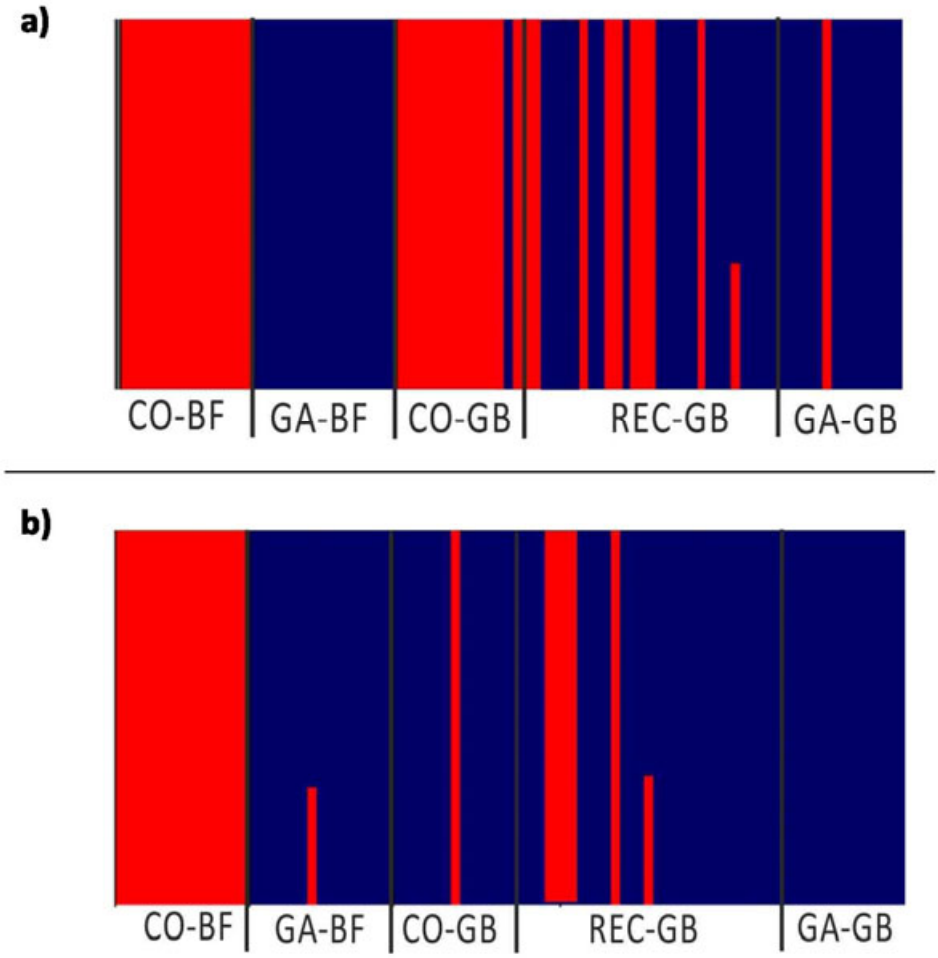

Figure 4.

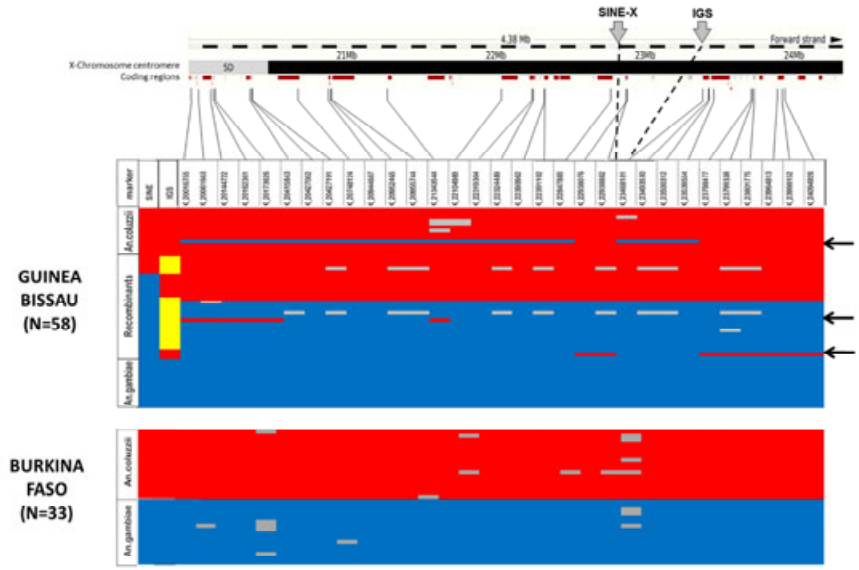

Figure 5.