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. Caputo¹, V. Pichler¹, E. Mancini¹*, M. Pombi¹, J. L. Vicente², J. Dinis³, K. Steen⁴, V. Petrarca⁵, A. Rodrigues³, J. Pinto², A. della Torre¹, D. Weetman¹

¹Istituto Pasteur-Fondazione Cenci-Bolognetti, Dipartimento di Sanità Pubblica e Malattie Infettive, Università “Sapienza”, Piazzale Aldo Moro 5, 00185, Rome, Italy; ²Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Rua daJunqueira,100, 1349-008 Lisboa, Portugal; ³Instituto Nacional de Saúde Pública, Ministério da Saúde Pública, Avenida Combatentes da Liberdade da Pátria, Apartado 861, 1004 Bissau Codex, Guinea Bissau, ⁴Vector Biology Department, Liverpool School of Tropical Medicine, Pembroke Pl, Liverpool, Merseyside L3 5QA, UK; ⁵Istituto 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

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/mec.13840

This article is protected by copyright. All rights reserved.
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-pericentromeric 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’-AATTTCACTCTACATCTACG-3’ and Cyp4G16rev: 5’-AAACATGTAAGGAGTAGTGG-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 MgCl₂, 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. Pericentromeric 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® genotyping 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.

This article is protected by copyright. All rights reserved.
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<sub>CO</sub>/IGS<sub>CO</sub>), An. gambiae (i.e. SINE<sub>GA</sub>/IGS<sub>GA</sub>) or nominally “F1-hybrids” (i.e. SINE<sub>CO/GA</sub>/IGS<sub>CO/GA</sub> 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<sub>GA</sub> specimens with IGS<sub>CO/GA</sub> genotypes (Safim: 16.7% in males and 11.89% in females; Mansoa: 6% and 6.52%; absent at Leibala). The discordant SINE<sub>GA</sub>-IGS<sub>CO</sub> 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 *TEP1*1 (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

This article is protected by copyright. All rights reserved.
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 its 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\textsuperscript{GA}-SINE\textsuperscript{CO}-IGS\textsuperscript{GA} males (which would be indicative of SINE insertion into an *An. gambiae* genotype) and a relatively high frequency of CYP\textsuperscript{CO}-SINE\textsuperscript{GA}-IGS\textsuperscript{CO} males. A gene conversion hot-spot in proximity of the SINE insertion, suggested by SINE\textsuperscript{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*

This article is protected by copyright. All rights reserved.
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


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.


This article is protected by copyright. All rights reserved.
This article is protected by copyright. All rights reserved.


**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**


**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.
This article is protected by copyright. All rights reserved.
**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 disequilibrium values are shown as D’ and r².

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](image-url)
Figure 2.

[Graph showing data for Guinea Bissau Co vs GAF 57 and Burkina Faso Co vs GAF 57 with axes labeled as follows:
- Guinea Bissau Co vs GAF 57: Y-axis range from -0.05 to 0.55, X-axis labeled as 2R, 2L, 3R, 3L, X.
- Burkina Faso Co vs GAF 57: Y-axis range from -0.05 to 0.55, X-axis labeled as 2R, 2L, 3R, 3L, X.]
Figure 3.

(a) [Diagram showing CO-BF, GA-BF, CO-GB, REC-GB, GA-GB with red and blue bars indicating data or results.]

(b) [Diagram showing similar categories as in (a).]
Figure 4.

Figure 5.

Recombinant haplotypes (N=62)

This article is protected by copyright. All rights reserved.