Evolutionary Applications

Open Access

Evolutionary Applications ISSN 1752-4571

ORIGINAL ARTICLE

Geographic population structure of the African malaria vector *Anopheles gambiae* suggests a role for the forest-savannah biome transition as a barrier to gene flow

Pinto J,[1] Egyir-Yawson A,[1,2] Vicente JL,[1] Gomes B,[1] Santolamazza F,[3] Moreno M,[4] Charlwood JD,[1,5] Simard F,[6] Elissa N,[7] Weetman D,[5] Donnelly MJ ,[5] Caccone A[8] and della Torre A[3]

1 Unidade de Parasitologia Médica, Centro de Malária e outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal
2 Biotechnology and Nuclear Agriculture Research Institute, Ghana Atomic Energy Commission, Legon, Ghana
3 Dipartimento di Sanità Pubblica e Malattie Infettive, Istituto Pasteur-Fondazione Cenci-Bolognetti, Università di Roma "La Sapienza", Rome, Italy
4 Division of Infectious Diseases, School of Medicine, University of California San Diego, La Jolla, CA, USA
5 Vector Group, Liverpool School of Tropical Medicine, Liverpool, UK
6 MIVEGEC (Maladies Infectieuses et Vecteurs: Ecologie, Genetique, Evolution et Contrôle), UMR IRD224-CNRS5290-UM1-UM2, Institut de Recherche pour le Développement, Montpellier, France
7 Unité d'Entomologie Médicale, Institut Pasteur de Madagascar, Antananarivo, Madagascar
8 Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

Keywords
Anopheles gambiae, geographic regions, microsatellites, molecular forms, population structure.

Correspondence
João Pinto, Unidade de Parasitologia Médica, Centro de Malária e outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Rua da Junqueira 100, Lisbon 1349-008, Portugal.
Tel.: + 351 213 652 666;
fax: + 351 213 632 105;
e-mail: jpinto@ihmt.unl.pt

Received: 28 August 2012
Accepted: 29 April 2013

doi:10.1111/eva.12075

Abstract

The primary Afrotropical malaria mosquito vector *Anopheles gambiae sensu stric*to has a complex population structure. In west Africa, this species is split into two molecular forms and displays local and regional variation in chromosomal arrangements and behaviors. To investigate patterns of macrogeographic population substructure, 25 *An. gambiae* samples from 12 African countries were genotyped at 13 microsatellite loci. This analysis detected the presence of additional population structuring, with the M-form being subdivided into distinct west, central, and southern African genetic clusters. These clusters are coincident with the central African rainforest belt and northern and southern savannah biomes, which suggests restrictions to gene flow associated with the transition between these biomes. By contrast, geographically patterned population substructure appears much weaker within the S-form.

Introduction

Many studies have attempted to identify genetic discontinuities between conspecific populations and to determine the factors that promote differentiation. This is a critical step for predicting the evolution of populations under different scenarios, including those that involve human-made environmental changes (Crispo et al. 2011). In medically important insects, the evolutionary relevance of these predictions gains a public health dimension, as they can be used to model the dispersal of genes of interest such as those related to insecticide resistance or refractoriness to infection by pathogens (Donnelly et al. 2002).

The nominal species of the *Anopheles gambiae* Giles complex (Diptera: Culicidae), *Anopheles gambiae sensu stricto* (hereafter termed '*An. gambiae*') is a primary vector of human malaria in Africa. It is widely distributed throughout sub-Saharan Africa in close association with humans. There is evidence that this species is undergoing a process of incipient speciation. The speciation process appears to be restricted to west Africa and involves sympatric populations. Initially, heterogeneities have been found in the

distribution of paracentric inversions at chromosome 2, which displayed strong heterokaryotype deficits. This led to the description of five chromosomal forms (cytoforms) each in Hardy–Weinberg equilibrium and characterized by distinct arrangements of inversions (Coluzzi et al. 1979, 2002). Following the early recognition of the five cytoforms, the species was tentatively split into two molecular forms, denoted M and S, identified by RFLP patterns in the X-linked ribosomal DNA (rDNA) intergenic spacer (IGS) (Favia et al. 1997; della Torre et al. 2001, 2002). The S-form has a continent-wide distribution, whereas the M-form appears to be confined to west Africa where it commonly occurs in sympatry with the S-form (della Torre et al. 2005). However, despite the extensive area of sympatry, MS hybrids are rarely seen (della Torre et al. 2005; Simard et al. 2009), with the exception of the extreme west of Africa (Caputo et al. 2008; Oliveira et al. 2008).

Initial genome-wide genotyping analyses revealed that differentiation between molecular forms was restricted to relatively small genomic regions located on the three chromosomes (Turner et al. 2005; White et al. 2010). More recently, however, whole-genome analyses based on next-generation sequencing and SNP microarrays have shown that M/S differentiation is more widespread across the genome than previously thought (Lawniczak et al. 2010; Neafsey et al. 2010; Weetman et al. 2010). Subsequently, the detection of genomic islands of divergence was found to be influenced by the degree of realized gene flow between the forms, which varies across west Africa (Weetman et al. 2012). As gene flow decreases, differentiation across the genome tends to increase and masks the initial divergent genomic regions. These findings point to a case of sympatric ecological speciation under divergent selection within *An. gambiae* (Diabaté et al. 2008; Costantini et al. 2009).

The phenotypic repercussions of the genetic divergence between molecular forms are still unresolved. Recent studies have shown that M-form larvae outcompete the S-form in the presence of predators, which may contribute to habitat segregation observed between forms (Diabaté et al. 2008; Gimonneau et al. 2010). M-form larvae prevail in areas with more permanent breeding sites (hence with higher predator pressure), whereas the S-form predominates in temporary rain-dependent breeding sites, perhaps due to a superior competitive ability where predation pressure is lower (Gimonneau et al. 2012). Genetic divergence between molecular forms may also impact both malaria transmission and vector control. A variant of the complement-like protein TEP1 with anti-parasitic activity was found to be fixed in M-form but absent in sympatric S-form populations of Mali and Burkina Faso (White et al. 2011). This was the first evidence of how subdivision within *An. gambiae* may affect vector competence. Another striking example comes from the contrasting differences in the frequency of knockdown resistance (*kdr*) mutations found between molecular forms. In spite of widespread sympatry between M- and S-forms, for a decade following their discovery in *An. gambiae* (Martinez-Torres et al. 1998), *kdr* mutations were found at high frequency in S-form populations but were rare in M-form (Santolamazza et al. 2008). Only recently, these mutations are becoming more common in M-form populations (Dabiré et al. 2009; Lynd et al. 2010).

In addition to the M- and S-forms partitioning, there is evidence for further population substructure within each of the molecular forms of *An. gambiae*. Microsatellite and AFLP analyses of S-form populations belonging to the SAVANNA and BAMAKO cytoforms revealed significant differentiation between these cytoforms in Mali (Taylor et al. 2001; Slotman et al. 2006). Similarly, Slotman et al. (2007) reported significant genetic differentiation between M-form populations of the FOREST and MOPTI cytoforms from Cameroon and Mali, respectively. These results led the authors to hypothesize that the M-form may actually consist of two partially isolated entities (Slotman et al. 2007; Lee et al. 2009).

With some exceptions (e.g. Lehmann et al. 2003; della Torre et al. 2005; Esnault et al. 2008; Choi and Townson 2012), the complex scenario of population subdivision within *An. gambiae* has been evidenced by studies that have often been based on a relatively limited geographic sampling coverage. Such local or regional sampling can be effective in detecting fine levels of population structure and revealing patterns of sympatric divergence but may mask other sources of substructuring, such as the presence of biogeographic or physical barriers to gene flow.

Here, we present the results of a microsatellite analysis of *An. gambiae* populations spanning the distribution of this species in the west of sub-Saharan Africa designed to assess geographic patterns of population structure within each molecular form.

Materials and Methods

Samples

Twenty-five *Anopheles gambiae* s.s. DNA samples were obtained from 24 collection sites in 12 African countries, between 1996 and 2006 (Fig. 1, see also Table S1 of the Supporting Information). These samples were collected mainly indoors by various adult sampling methods and identified to species by PCR (Scott et al. 1993), within the framework of previous entomological surveys. With the exception of a single site located in eastern Africa (Furvela, Mozambique), all sampling locations were in west Africa.

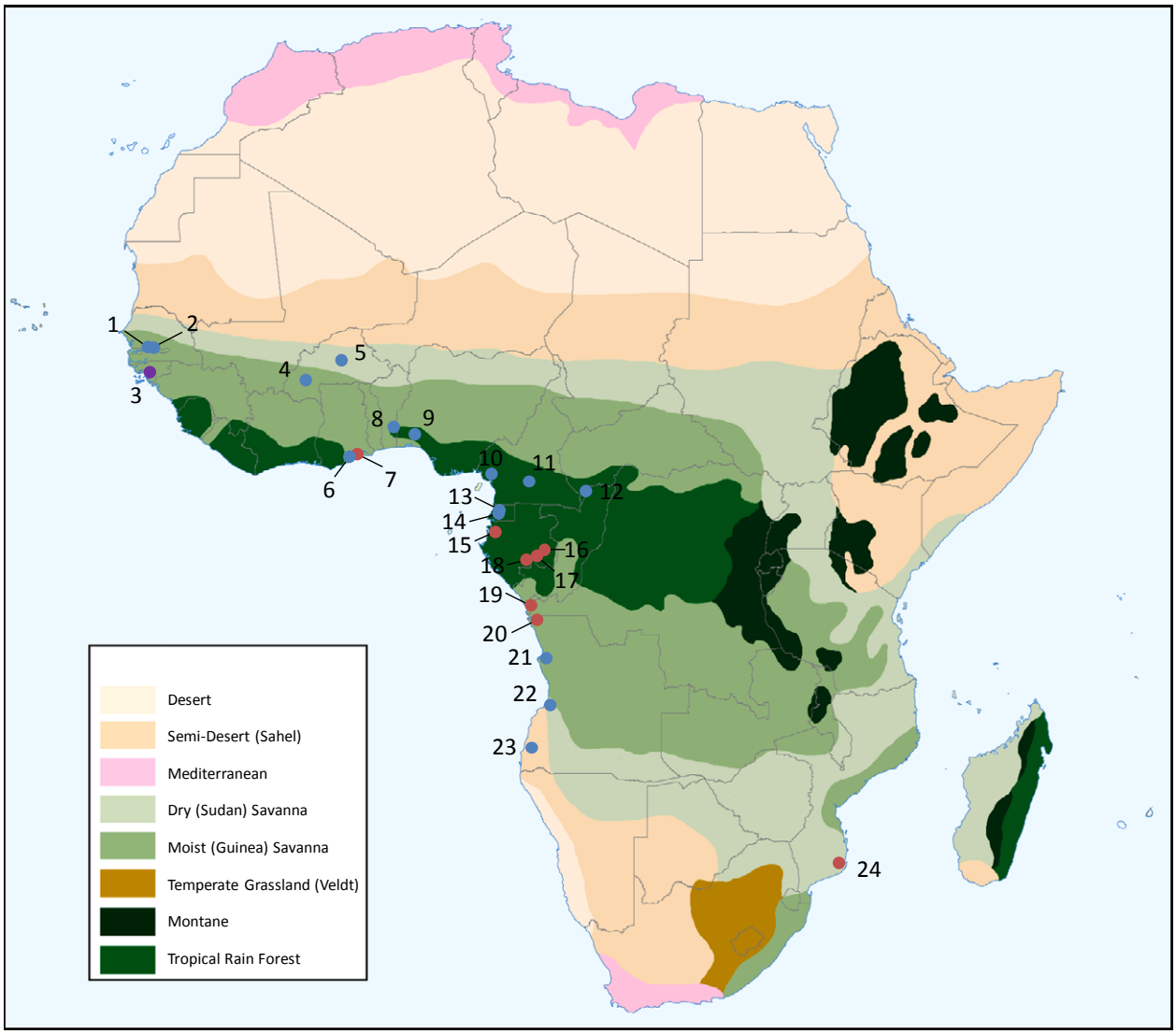

Figure 1 Map of Africa biomes (adapted from UNEP 2010) showing the location of the collection sites. Blue marks: M-form samples (identified by IGS-PCR); red marks: S-form samples; purple mark: locality with both M- and S-form samples. The Gambia: Wali Kunda (1), Maccarthy island (2); Guinea-Bissau: Bissau (3), Burkina Faso: Bobo-Dioulasso (4), Goundry (5); Ghana: Okyereko (6), Accra (7); Benin: Dassa (8); Nigeria: Kobape (9); Cameroon: Tiko (10), Simbok (11); Central African Republic (CAR): Bayanga (12); Equatorial Guinea: Ngonamanga (13), Bata (14); Gabon: Libreville (15), Benguia (16), Bakoumba (17), Dienga (18); Angola: Cabinda (19), Kikudo (20), Luanda (21), Cavaco (22), Namibe (23); Mozambique: Furvela (24). The dashed contour lines represent the approximate limits of the distribution of the S-form, and the dash-dotted contour line shows the limit of the distribution of the M-form, which is confined to west Africa.

The distribution of the west African sampling sites covered an overland distance of *ca.* 5700 km, from the Gambia to southern Angola.

Of the 25 samples analyzed, 16 were of the M-form and nine of the S-form according to the genotyping of the ribosomal DNA IGS marker (Favia et al. 1997; della Torre et al. 2001). The mean pair-wise distance among M-form sampling sites was 2031 km (median: 1810 km; SD: ± 1303 km) and 2129 km (median: 1977; SD: ± 1779 km) for S-form sampling sites. The mean pair-wise distance among S-form sites from west Africa (i.e. excluding the eastern African sample of Mozambique) was 1558 km (median: 797 km; SD: ± 1436 km). Although sympatric M- and S-forms are present in most west and central African sites, Bissau (Guinea-Bissau) was the only locality from which sympatric samples of both M- and S-forms were analyzed in this study. Information on sample size, year, type of collection, geographic coordinates, and biome type is given for each sample in Table S1 of the Supporting Information.

Microsatellite genotyping

Thirteen microsatellite loci were genotyped. All loci were located on chromosome 3 to avoid potential bias resulting from reduced recombination or selective pressures acting at chromosomal inversions (frequent in chromosome 2) or linkage with genomic regions of M/S divergence on chromosome X (Lanzaro et al. 1998; Turner et al. 2005). Each locus was amplified individually by PCR with fluorescently labeled primers using the protocols described by Donnelly et al. (1999). Details of the microsatellites genotyped can be found in Table S2 of the Supporting Information. Fragment analysis was performed by capillary electrophoresis on an automated sequencer (ABI®3730, Applied Biosystems, Foster City, CA, USA) at the Science Hill DNA Analysis Facility, Yale University. To control for variation in allele size scoring between capillary runs, the same positive controls, consisting of PCR products of two *An. gambiae* specimens from a laboratory colony, were used in all runs. One additional positive control (DNA template from a colony mosquito) and one negative control (no template) were also included to assess PCR quality. Allele sizes were scored from electropherograms using the software GENEMARKER® (SoftGenetics, State College, PA, USA).

Genetic data analysis

Genetic variation at each microsatellite locus was characterized by estimates of unbiased expected heterozygosity (Nei 1987) and allelic richness (El Mousadik and Petit 1996). The latter parameter was used to account for differences in sample sizes. Calculation of the estimates and comparisons among groups by permutation tests (1000 permutations) were performed using FSTAT v.2.9.3 (Goudet 1995). The same software was used to compute pair-wise estimates of the genetic differentiation parameter F_{ST} according to Weir and Cockerham (1984) and to assess their significance by permutation tests (1000 permutations). The number of shared alleles between groups was estimated in random subsamples of each group with size equal to the smallest group sample size. Exact tests against Hardy–Weinberg proportions and of linkage disequilibrium between pairs of loci were performed in GENEPOP v.4.1 (Raymond and Rousset 1995). Presence of null alleles at each locus and sample was tested using the procedure implemented by MICROCHECKER with a 99% confidence interval (Van Oosterhout et al. 2004). The coalescent-based simulation approach implemented in LOSITAN (Antao et al. 2008) was used to identify outlier microsatellites displaying unusually high or low F_{ST} values of by comparing observed F_{ST} estimates with values expected under neutrality (Beaumont and Nichols 1996). Runs were conducted under 'neutral mean F_{ST}' and stepwise or infinite alleles mutation models using 50 000 simulations over all loci. The significance threshold for outlier detection was set at ≥0.95 percentile of simulations.

Bayesian clustering methods were used to detect population subdivision without *a priori* assumptions on population boundaries. Two types of clustering methods, namely spatial and nonspatial, were employed based on whether geographic information was included as a prior in the analysis. Spatial models generally perform better in cases of low differentiation ($F_{ST} < 0.05$) among populations (Chen et al. 2007).

The nonspatial Bayesian clustering analysis method implemented in STRUCTURE 2.3.3 (Pritchard et al. 2000) was used to infer the number of genetic clusters (K) in the whole data set and within each molecular form separately. Analyses were carried out without prior information of sampling locations. A model with correlated allele frequencies within populations was assumed ($\lambda = 1$). The software was run with the option of admixture, allowing for some mixed ancestry within individuals, and the degree of admixture (α) was allowed to vary. For each value of K ($K = 1$–10), 10 independent runs were carried out with a burn-in period of 10 000 and 100 000 iterations. The ΔK statistic of Evanno et al. (2005) was calculated using STRUCTURE HARVESTER (Earl and vonHoldt 2012) to determine the most likely number of clusters. The information from the outputs of the 10 runs for each K was compiled by the greedy method implemented in CLUMPP (Jakobsson and Rosenberg 2007). Individual assignment to clusters was performed with a probability threshold (T_q) determined by the analysis of simulated parental and admixed individuals generated by HYBRIDLAB v1.0 (Nielsen et al. 2006). From the initial whole-sample STRUCTURE analysis, individuals showing a probability of membership $q_i > 0.90$ were selected to simulate 100 genotypes of each parental class and four hybrid classes (F1, F2, and backcrosses with each parental class). Simulated genotypes were analyzed by STRUCTURE under the same conditions as above. Following the example of Vähä and Primmer (2006), power and accuracy were calculated for four T_q values (0.70, 0.75, 0.80, and 0.90).

Spatial genetic clustering analysis was conducted with the whole data set and with M- and S-form data sets using the software TESS v.2.3. (François et al. 2006; Chen et al. 2007). This method implements a Bayesian clustering algorithm that integrates genetic and spatial information to ascertain population structure without *a priori* population information, by inferring the most likely maximum number of clusters. As geographic coordinates were available only for each collection site, individual coordinates for each specimen were randomly generated within a circle with 10-km radius around the coordinate of each site. The 10-km radius was chosen based on previous observations on anopheline maximal flight distances that seem to vary

around 9–12 km (Kaufmann and Briegel 2004). The two admixed models available in TESS, CAR and BYM (Chen et al. 2007; Durand et al. 2009), were used in the analysis. Ten independent runs were carried out with a burn-in period of 100 000 iterations and 100 000 replications for each value of K_{max} (K = 2–10). The Deviance Information Criterion (DIC) was used to select the admixture model that performed better and to infer the number of clusters. The maximum number of clusters was selected from DIC versus K_{max} plots as the lowest value at which the DIC curve reached a plateau. The estimated individual membership probabilities of the ten runs of the optimal K_{max} were averaged using the greedy algorithm in CLUMPP to correct for discrepancies between runs.

Principal coordinates analysis (PCoA) was used to visualize patterns of genetic differentiation among samples in a two-dimensional plot. Calculations were performed in GENALEX 6.41 (Peakall and Smouse 2006) using the standardized covariance method for the distance matrix conversion.

Isolation by distance was tested by the linear regression between logarithmic geographic distances and linearized 1/(1-F_{ST}) values (Rousset 1997). Pair-wise overland distances between sites were estimated using the metric tool available in Google® Earth. The software GENALEX was used to assess the significance of the correlation by Mantel tests (1000 permutations).

Whenever multiple tests were performed, the nominal significance level (α = 0.05) was adjusted by the sequential Bonferroni procedure (Holm 1979).

Results

A total of 967 *An. gambiae* were analyzed. Estimates of genetic diversity are shown in Table S3 of the Supporting Information. Mean allele richness (R_s) of the microsatellite loci varied from 5.4 (AG3H242) to 12.7 (AG3H128) and expected heterozygosity (H_e) from 0.575 (AG3H577) to 0.894 (AG3H128). There were 48 significant departures from Hardy–Weinberg proportions of 325 tests performed. These were associated with positive F_{IS} values indicating heterozygote deficits. Loci AG3H88, AG3H127, and AG3H750 comprised 39 (81.3%) of the 48 significant tests, suggesting that departures from Hardy–Weinberg expectations were locus-specific. Presence of null alleles was detected by MICROCHECKER in 44 of the 48 (92.7%) significant heterozygote deficits (Table S3, Supporting Information). There were 35 significant linkage disequilibrium (LD) tests of 1950 performed, of which 24 (68.6%) were observed in the sample from Cabinda, Angola (sample 19, Fig. 1) and 6 (17.1%) in Kobape, Nigeria (sample 9). Of the 13 loci analyzed for signatures of selection using LOSITAN, only AG3H127 showed a significant signal of positive selection in both mutation models (Fig. S1, Supporting Information). Two additional loci, AG3H758 and AG3H93, displayed marginally significant signals of selection and only under the IAM or SMM mutation models, respectively.

The results of the Bayesian clustering analysis implemented in STRUCTURE are shown in Fig. 2. Graphical representations of Evanno's ΔK can be seen in Fig. S2 of the Supporting Information. When all samples were analyzed together, the optimum number of clusters was K = 2. This partitioning generally corresponded to the M (cluster 1) and S (cluster 2) molecular form composition of the samples and it was independent of geographic location. However, samples from west African sites (i.e. samples 1–9 in Fig. 2, K = 2) displayed more inconsistencies between the form determined by the IGS marker and the respective genetic background when compared to samples from central and southern Africa. In west African samples, the average probability of assignment to cluster 1 for M-form specimens was 0.515 and 0.636 for S-form assignment to cluster 2. When individuals were assigned to each cluster based on a $T_q \geq 0.8$, as determined by the analysis of simulated data (see Table S4, Supporting Information), there were only 8.7% (25 of 289) M-form individuals assigned to cluster 1 and 33.3% (23 of 69) S-form individuals to cluster 2. The proportion of individuals with admixed ancestry (i.e. $0.20 < T_q < 0.80$) was 83.7% and 65.2% for M- and S-form, respectively. In contrast, the average probabilities of assignment for M- and S-form individuals from central and southern African sites were 0.831 and 0.847, respectively. The proportion of consistent assignments was also much higher: 73.3% (225 of 307) in the M-form and 75.8% (229 of 302) in the S-form. The second highest ΔK value corresponded to K = 3. Here, M-form populations were subdivided into two genetic clusters (Fig. 2, K = 3): cluster 1 contained mainly individuals from the samples collected in west Africa (samples 1–9) and also from Bayanga, CAR (sample 12); cluster 2 included the remaining samples from central Africa (samples 10, 11, 13, and 14) and Angola (samples 21–23). These results did not differ qualitatively when analyses were repeated excluding the three loci that revealed most heterozygote deficits indicating that locus-specific Hardy–Weinberg deficits had little impact in the analysis (Fig. S3, Supporting Information).

STRUCTURE was also performed within each molecular form separately. When M-form samples were analyzed, a third subdivision was evident (Fig. 2, M-form). West African samples were again grouped in a single cluster (cluster 1, samples 1-9, 12), but there was a separation between central African samples (cluster 2, samples 10-11, 13-14) and the southernmost samples from Angola (cluster 3, samples 21-23). The only exception to this geographic partitioning was the sample from Bayanga (sample 12). This sample has a central African location, but individuals displayed a

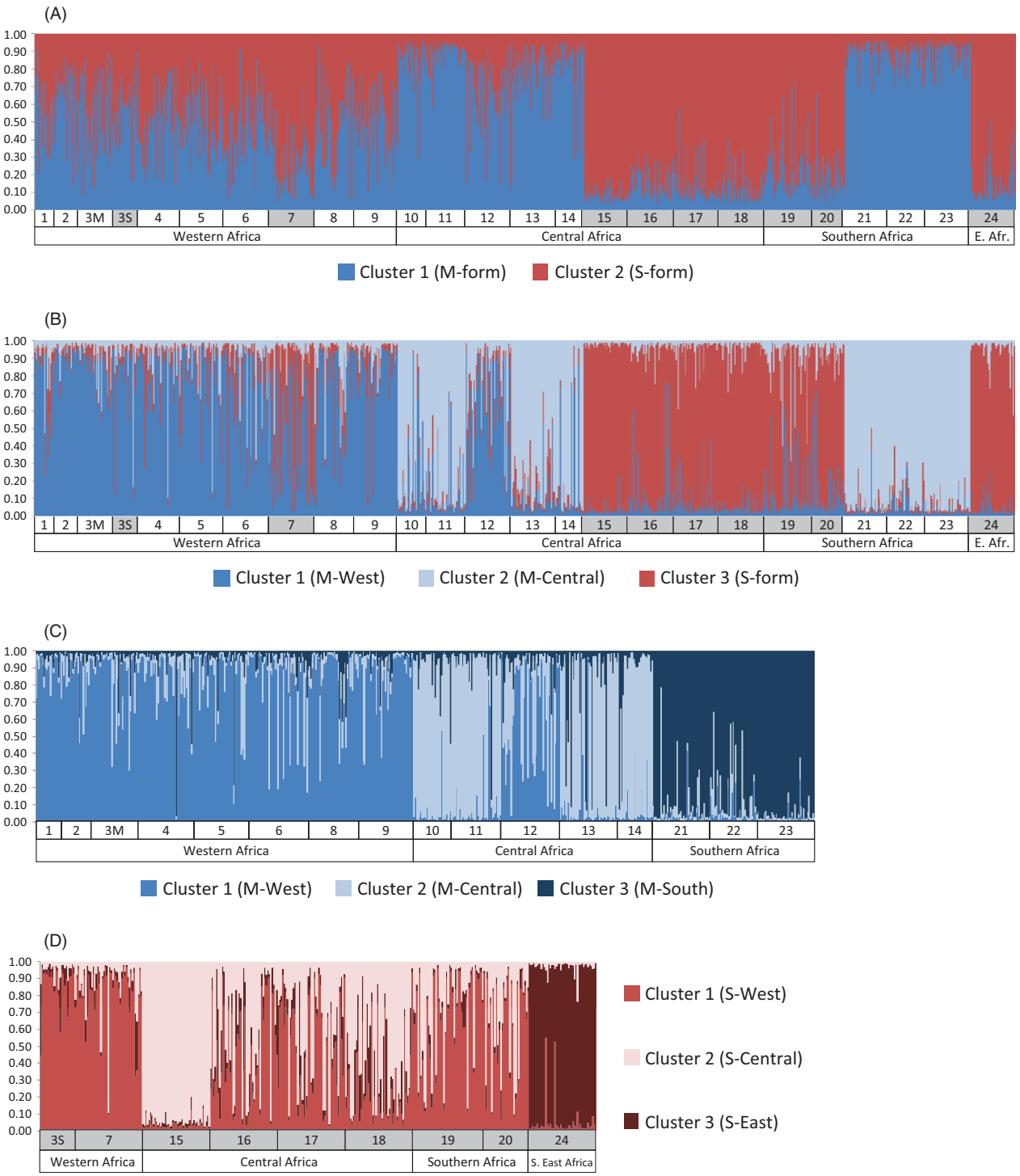

Figure 2 Bayesian clustering analysis implemented by STRUCTURE (Pritchard et al. 2000). Localities are numbered according to Fig. 1 in a northwest–southeast direction along the *X*-axis (see also Table S1 of Supporting Information). White boxes indicate M-form and gray boxes indicate S-form samples as determined by the IGS marker. *Y*-axis: probability of ancestry to each cluster. In the graphs, each column corresponds to the multilocus genotype of a single individual partitioned into colors representing the probability of assignment to each cluster. (A) analysis performed with all samples ($N = 967$), $K = 2$; (B) analysis performed with all samples ($N = 967$), $K = 3$; (C) analysis performed with M-form samples only ($N = 596$), $K = 3$; (D) analysis performed with S-form samples only ($N = 371$), $K = 3$.

higher probability of assignment to cluster 1 (mean = 0.633) compared to cluster 2 (mean = 0.289). For $T_q \geq 0.80$, 46.7% of the 45 individuals analyzed were assigned to the west African cluster 1 and only 2 individuals (4.4%) were assigned to the central African cluster 2.

Subdivision among S-form populations was also observed when STRUCTURE analysis was performed with these samples only (Fig. 2, S-form). The two west African samples (Bissau, 3S and Accra, 7) were grouped into a west African cluster (cluster 1). In central Africa, a second cluster was detected (cluster 2). This genetic background predominates in the sample from Libreville, Gabon (sample 15) and gradually intergrades southwards with cluster 1. The proportion of individuals assigned to cluster 2 ($T_q \geq 0.80$) decreased from 97.8% in Libreville (sample 15) to 28.9% (Dienga, 18), 17.8% (Benguia, 16), 13.3% (Bakoumba, 17), 8.5% (Cabinda, 19), and 6.7% in the southernmost Kikudo (sample 20). Finally, a third cluster comprised specimens from the southeast African sample of Furvela (sample 24), in Mozambique.

The geographic structuring of M-form populations was also evident in the principal coordinates analysis (Fig. 3). The distribution of the M-form samples in the plot reflects their geographic grouping into west, central, and southern clusters. The S-form samples were clearly separated from the M-form with the single exception of Bissau (3S, Fig. 3). The separation between west and central African S-form samples was less pronounced than in the M-form. The S-form sample of Furvela, Mozambique (sample 24), was placed as an outlier of the S-form group, in agreement with the results obtained by STRUCTURE.

The results of the spatially explicit analysis conducted in TESS were very similar for the two admixture models used.

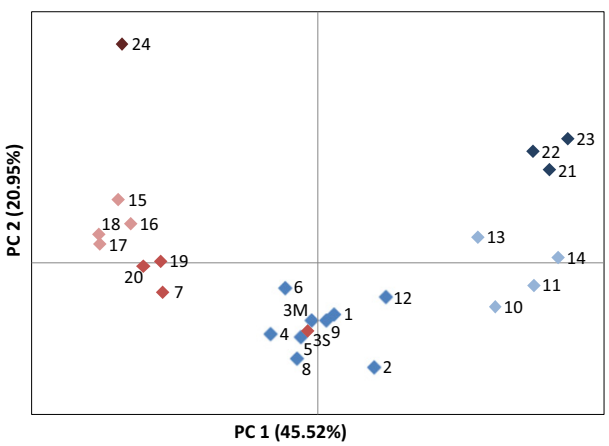

Figure 3 Principal coordinates analysis of the 25 *An. gambiae* samples. Each mark represents a sample numbered according to Fig. 1. Marks are colored according to the within-form genetic clusters revealed by STRUCTURE (Fig. 2). Blue: M-west, light blue: M-central, dark blue: M-south; red: S-west, light red: S-central, dark red: S-east.

The CAR model gave, however, less-dispersed DIC values between runs, so that only the results for this model are presented (Fig. S4, Supporting information). When both M- and S-form samples were analyzed together, an optimal $K_{max} = 6$ was obtained (Fig. 4, A). There were three major clusters that consisted in the partitioning of the M-form into west, central, and south clusters, thus confirming the results of the nonspatial analyses. In the S-form, however, only the east African sample of Furvela, Mozambique (sample 24), formed a distinct cluster, whereas the remaining S-form samples from west and central Africa grouped together. There was one additional minor cluster in which the highest individual probability of membership was only 0.38, for a specimen from Bata (sample 14). The spatial analysis of M- and S-form samples alone did not disclose any additional substructuring. For the M-form, a $K_{max} = 4$ was obtained confirming west, central, and southern clusters (Fig. 4, B and C). A fourth minor cluster comprised again the same single individual from Bata, Equatorial Guinea (sample 14) with a probability of membership $q_i = 0.731$. The assignment of this individual into a minor cluster was also consistent in the TESS analyses performed with 10 loci (i.e. excluding the three loci with greatest heterozygote deficits; Fig. S3, C and D). This consistency led us to re-analyze the IGS molecular identification (Scott et al. 1993) of this and the other specimens of this locality. This revealed the presence of two misidentified individuals. One was found to be *Anopheles melas*, another sibling species of the *An. gambiae* complex, and corresponded to the individual assigned to the minor cluster. The other gave a banding pattern consistent with a hybrid between *An. melas* and *An. gambiae s.s.* This individual was assigned to the M-west cluster with $q_i = 0.621$. Removing these two individuals had little influence on the estimates of pair-wise genetic differentiation between this locality and the others (Table S6). For the S-form, an optimal $K_{max} = 3$ also confirmed the separation of the east African sample of Furvela, Mozambique (sample 24), but did not disclose any subdivision between central and west African samples (Fig. 4, D and E). There was one additional minor cluster represented by five specimens, four from Bissau (sample 3S) and one from Accra (sample 7).

Significant positive slopes were obtained for all the regressions of (F_{ST}/1-F_{ST}) with logarithmic distance (Table S5, Supporting Information). The proportion of the variation explained by the regression (r^2) was generally low, particularly when both M- and S-form were analyzed together (all samples, Table S5, Supporting Information). The largest r^2 value was recorded for the regression involving S-form samples (0.43). When the most distant S-form sample of Furvela, Mozambique (sample 24), was removed, the regression remained significant but with a lower r^2 (0.28). Plots of the regression of linearized F_{ST} and logarithmic

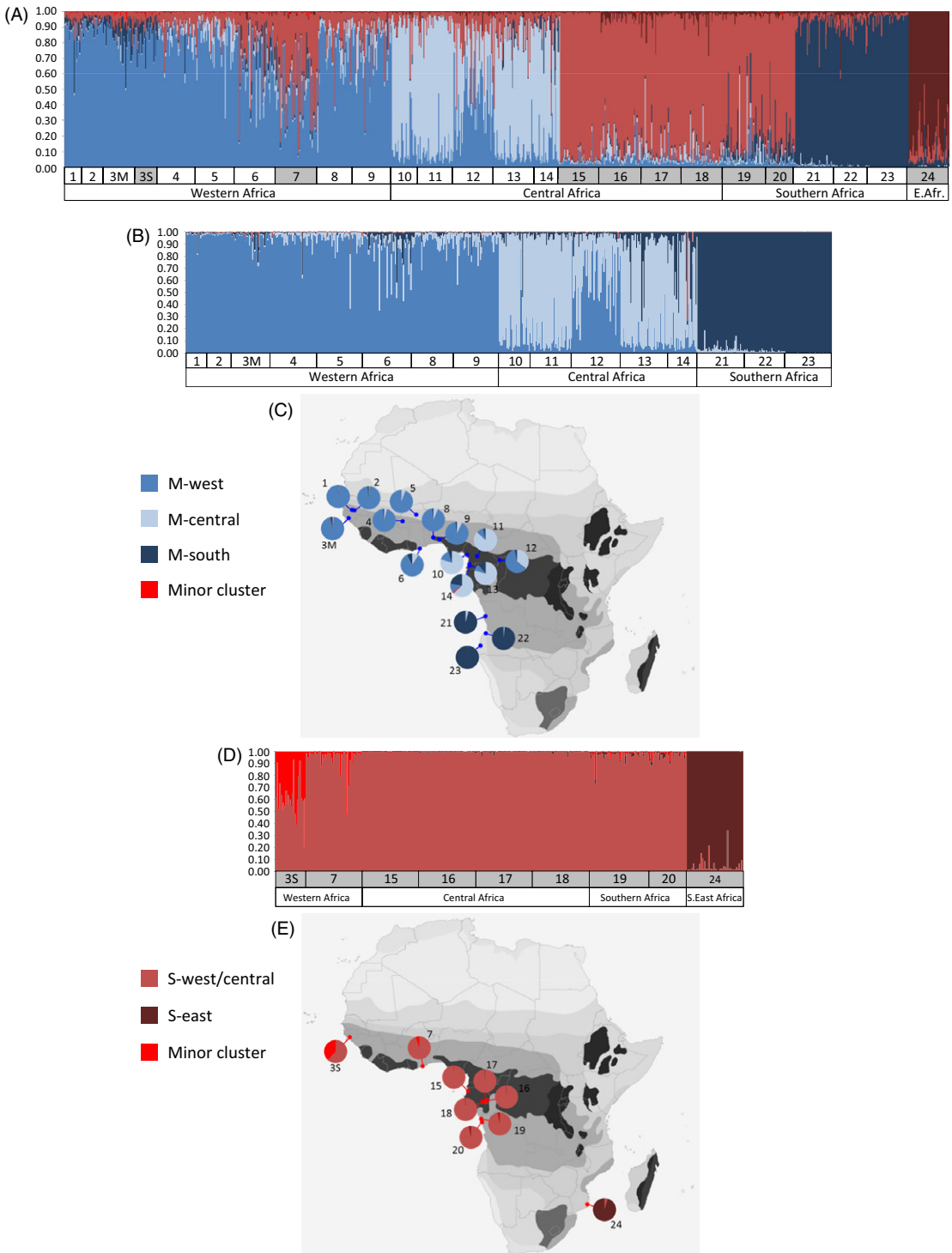

Figure 4 Individual assignment plots and maps showing mean membership probabilities to each cluster at each locality, obtained by TESS (Chen et al. 2007). The bar plots depict individual assignment probabilities averaged for the ten runs using CLUMPP (Jakobsson and Rosenberg 2007). The maps show pie charts of the average probability of membership to each cluster for each locality. Samples are numbered according to Fig. 1 and Table S1 (Supporting Information). (A) analysis performed with all samples (i.e. both M- and S-form), with clusters colored according to the labels of the following bar plots; (B and C) analysis performed with M-form samples only; (D and E) analysis performed with S-form samples only.

distance for M- and S-forms are shown in Fig. 5. For the M-form, comparisons between sampling sites within the same genetic cluster (obtained by STRUCTURE) had in general lower F_{ST} than comparisons involving sites from distinct genetic clusters (Fig. 5, A). The mean of pair-wise F_{ST} estimates between samples within each cluster varied between 0.015 and 0.022, corresponding to comparisons between collection sites 18–3658 km apart (Table S6, Supporting Information). The mean of pair-wise F_{ST} between samples from different clusters ranged from 0.030 to 0.042 and involved comparisons with distances between 541 km and 5317 km. This pattern was not so evident in the S-form, where differentiation appears to reflect less cluster ancestry and depend more on geographic distance (Fig. 5, B).

Estimates of genetic diversity for each genetic cluster within the M-form and for the S-form are shown in Table 1. There was a trend for a south–north increase in diversity within the M-form. Estimates of R_s and H_e were lowest in the M-south cluster, intermediate in the M-central, and highest in the M-west cluster. These differences were significant for both parameters (permutation tests, R_s: $P = 0.001$; H_e: $P = 0.027$). The S-form displayed R_s and H_e values similar to those of the M-central cluster. The average number of shared alleles among clusters was higher between the S-form and M-west clusters than between any other comparison (Table 1). These two clusters also had the lowest pair-wise F_{ST} estimate (0.023) with the highest (0.075) being observed between S-form and M-south (Table 1).

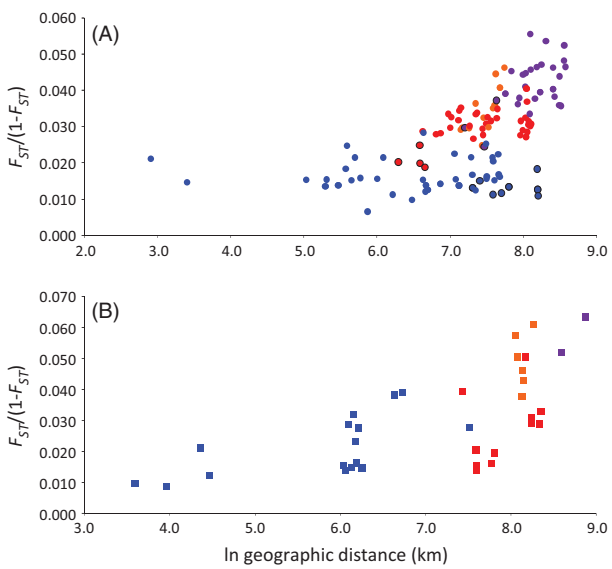

Figure 5 Plots of the regression between $F_{ST}/(1-F_{ST})$ and logarithmic geographic distance. (A) M-form. Blue circles: comparisons between localities for which the majority of individuals were assigned to the same genetic cluster (i.e. M-west, M-central, and M-south). Red circles: comparisons between M-west and M-central localities. Orange circles: comparisons between M-central and M-south localities. Purple: comparisons between M-west and M-south localities. Circles with a black line correspond to comparisons involving the locality of Bayanga (CAR), which was considered as representative of the M-west cluster. (B) S-form. Blue squares: comparisons between localities belonging to the same genetic cluster (i.e. S-west, S-central, and S-east/Mozambique). Red squares: comparisons between S-west and S-central localities. Orange squares: comparisons between S-central localities and S-east/Mozambique. Purple squares: comparisons between S-west localities and S-east/Mozambique.

Table 1. Estimates of genetic diversity, pair-wise F_{ST}, and proportions of shared alleles among S-form and M-form clusters.

	M-west	M-central	M-south	S-form
R_s	8.8 (0.3)	7.5 (0.2)	6.3 (0.4)	7.4 (1.0)
H_e	0.806 (0.016)	0.773 (0.013)	0.720 (0.027)	0.743 (0.051)
M-west	–	0.030	0.048	0.023
M-central	11.5	–	0.035	0.061
M-south	9.4	8.7	–	0.075
S-form	13.0	10.8	8.9	–

H_e, mean over loci expected heterozygosity; R_s, mean over loci allele richness; in parenthesis: standard deviation of mean; above diagonal: F_{ST} estimates (all significant, $P < 0.001$); below diagonal: mean over loci number of shared alleles estimated in randomly selected subsamples of each group with samples size equal to the lowest sample size (M-south, $N = 124$).

Discussion

The macrogeographic scale microsatellite analysis on *An. gambiae* presented here revealed a significant association between genetic differentiation and geographic distance. This pattern of isolation by distance was not an unexpected result given the relatively low active dispersal capacity of this mosquito (<13 km, Kaufmann and Briegel 2004) and also agrees with a previous study covering similar geographic ranges (Lehmann et al. 2003). However, isolation by distance appears not to be the only factor shaping the genetic structure of this species. Two additional sources of variation were disclosed: the well-known subdivision of the species into the M and S molecular forms and the split of the M-form into three geographic clusters corresponding to west, central, and southern African populations.

Subdivision corresponding to the two molecular forms was revealed by both Bayesian clustering analyses and was also confirmed by PCoA. This pattern was detected using molecular markers located outside the previously described genomic regions of divergence (Turner et al. 2005; White et al. 2010). It was also independent of the geographic location. At $K = 2$ of the STRUCTURE analysis, all M-form samples clustered together regardless of being from west, central, or southern Africa. Likewise, the S-form samples

from west and central Africa also clustered with the East African sample of Mozambique. The single exception was the clustering of the majority of the S-form individuals from Bissau in the M-form cluster, which reflects the high levels of inter-form hybridization and asymmetric introgression previously described for this region (Oliveira et al. 2008; Caputo et al. 2011; Marsden et al. 2011). The introgression of more M-form genes into the S-form detected in these reports agrees with the position of the S-form sample from Bissau in the west M-form cluster of the PCoA conducted in this study.

The degree of inter-form differentiation appeared to be higher in central and southern African samples than in west African ones, judging by the individual probabilities of assignment to M- and S-form clusters obtained in STRUCTURE at $K = 2$. An explanation for this observation could be the nearly monotypic composition of *An. gambiae* in some of the collection sites. This is the case for S-form samples of Mozambique, Gabon and northern Angola and also for the M-form samples of Angola (Pinto et al. 2002; Calzetta et al. 2008). However, this hypothesis is less probable for the sites sampled in Cameroon and Equatorial Guinea, in which both forms have been found in sympatry at minimum relative frequencies of *ca.* 10:90 (Moreno et al. 2007; Ridl et al. 2008; Simard et al. 2009; Weetman et al. 2010; Kamdem et al. 2012).

There is evidence that inter-form gene flow and introgression varies across the *An. gambiae* distribution range. In the central African region, the degree of inter-form divergence appears to be highest and coincident with no reported MS hybrids (della Torre et al. 2005; Simard et al. 2009), although there is evidence for at least sporadic recent gene flow (Etang et al. 2009; Weetman et al. 2012). In contrast, the isolation between forms seems to be less marked in west Africa. Here, MS hybrid rates have been found to vary greatly, from ~1% (della Torre et al. 2005; Costantini et al. 2009) to over 20% (Caputo et al. 2008; Oliveira et al. 2008). High levels of inter-form hybridization and a pattern of asymmetric introgression have been described in Guinea-Bissau (Oliveira et al. 2008; Caputo et al. 2011; Marsden et al. 2011). Low inter-form differentiation was also reported in a previous microsatellite analysis of samples from different ecological zones in Ghana (Yawson et al. 2007). These results contrast with the high levels of inter-form differentiation revealed by genome-wide SNP analyses in *An. gambiae* from Ghana (Weetman et al. 2010) and also from Mali (Neafsey et al. 2010). This discrepancy might be influenced by the propensity of microsatellites to underestimate genetic differentiation as a result of allelic homoplasy (Estoup et al. 2002). However, in the SNP analyses of M- and S-forms from Ghana, differentiation was markedly heterogeneous and far lower on chromosome-3 than on chromosome-2 and chromosome X (Weetman et al. 2012). Thus, differences might also be explained by variation in the genomic location of markers. Comparative genome-wide SNP analysis of samples from central and west African regions displaying varying levels of hybridization showed that the degree of genomic divergence was dependent on the amount of realized gene flow between forms (Weetman et al. 2012). Altogether, these results point to a considerable variation in the degree of isolation between molecular forms throughout the species range. This variation may be a consequence of an intricate assemblage of factors such as local or regional differences in the stage or history of the speciation process, the occurrence of secondary contact zones, and differences in the ecological trade-offs of hybridization (Caputo et al. 2011; Marsden et al. 2011).

Additional partitioning into three distinct geographic M-form clusters corresponding to west, central, and southern African populations was revealed by both spatial and nonspatial Bayesian clustering analyses and also by PCoA. This subdivision appears to be coincident with the transition from a rainforest biome to northern and southern savannah biomes, respectively. The genetic discontinuity imposed by the forest-savannah transition is not complete, as evidenced by the maintenance of a significant isolation-by-distance signal across all M-form samples and also by the presence of a locality (Bayanga, CAR) displaying a higher proportion of an M-west genetic background in spite of its rainforest location. Bayesian clustering methods may overestimate genetic structure by generating spurious clusters when applied to populations displaying isolation by distance (Frantz et al. 2009; Schwartz and Mckelvey 2009). However, pair-wise F_{ST} values within clusters were generally lower than those involving comparisons between clusters, even when the distances between sampling sites of the same genetic cluster were similar to those between sampling sites of different clusters (Fig. 5, A). This suggests that differentiation within the M-form is not only dependent on geographic distance but that restrictions to gene flow may also be present. The intermediate F_{ST} values of the sample of Bayanga in the plot of Fig. 5 (A) are also consistent with a higher admixture between two distinct genetic clusters in this particular locality. Also, levels of differentiation do not seem to have been influenced by temporal differences between samples. Nonsignificant F_{ST} values were obtained between samples from the Gambia and Guinea-Bissau, located *ca.* 200 km apart, in spite of a 7-year interval between these collections.

Initial evidence of a separation between west and central African M-form *An. gambiae* populations emerged from two previous microsatellite-based studies. In the only microsatellite-based continent-wide study carried out before the present one, the grouping of Senegal and Ghana samples apart from central African ones was observed in a

F_{ST}-based neighbor-joining population tree (Lehmann et al. 2003). Moreover, a high degree of genetic differentiation was found between M-form populations from a savannah area in Mali and those from a forested area in Cameroon, suggesting that M-form may not be a single entity (Slotman et al. 2007).

Population subdivision associated with forest-savannah transitions is not uncommon. A similar scenario was recently described within the *An. gambiae* sibling species *Anopheles melas* Theobald, in which genetically distinct west and central/southern African clusters were detected with a degree of divergence comparable to that observed among other species of the *An. gambiae* complex (Deitz et al. 2012). Significant differentiation between rainforest populations and one southern savannah population of *Anopheles nili* (Theobald) in central Africa also suggested a role of the evergreen forest as a barrier to gene flow in this vector species (Ndo et al. 2010). A recent study has also shown the occurrence of a cryptic central African group genetically distinct from west African populations within the tsetse fly *Glossina palpalis palpalis* Robineau-Desvoidy (Dyer et al. 2009). Altogether, these results are consistent with a role of the transition between rain forest and savannah biomes as a barrier to gene flow in insect species.

Recent studies have shown that central African M-form populations are becoming more adapted to densely urbanized areas where they explore polluted breeding sites of anthropogenic nature (Simard et al. 2009; Kamdem et al. 2012). In contrast, west African M-form populations appear more closely associated with irrigated agricultural areas, occupying more permanent breeding sites such as rice fields and irrigation reservoirs (Gimonneau et al. 2012). Local adaptation to different ecological niches coupled with the effect of isolation by distance and restrictions to mosquito active dispersal imposed by the rainforest environment could explain the observed patterns of population subdivision within the M-form.

Another factor that may have contributed to the differentiation between west and central African M-form clusters could be a higher degree of genetic introgression between M- and S-forms in west Africa. This effect is suggested by the highest mean number of shared alleles and lowest pairwise F_{ST} estimate between the M-west cluster and the S-form, in line with a hypothesis of highest introgression between these clusters. Introgression may also explain the higher levels of genetic diversity of the M-west cluster as measured by estimates of H_e and R_s. The data suggest that substantial MS inter-form introgression is a less-probable cause for the differentiation between central and southern M-form samples, because evidence of inter-form gene flow (i.e. admixture in the $K = 2$ analysis of STRUCTURE) was much lower in these samples. However, sequence analysis of an X-linked locus revealed that the majority of M-form individuals in Angola had a 16-bp insertion that was fixed in the S-form but absent in M-form individuals from west and central Africa (Choi and Townson 2012), a finding that suggests inter-form introgression has occurred in this geographic region.

The results obtained for the S-form did not conclusively show a genetic discontinuity at the transition between rainforest and savannah. A central African S-form cluster was detected by STRUCTURE analysis but appears to be mostly represented by a single sample (Libreville). Rainforest samples of Gabon also appeared more closely related in the PCoA. However, S-form samples from savannah biomes in west Africa (Ghana) and Angola were grouped together in the PCoA and into a single cluster in STRUCTURE. The intergradation between S-form clusters observed southwards of Libreville in the STRUCTURE analysis also suggests gradual differentiation, in line with an expectation of isolation by distance. Moreover, the results of the spatial genetic analysis conducted by TESS for the S-form did not show a clustering of central African samples within the rainforest belt. Instead, the two major clusters corresponded to the separation of the East African sample of Mozambique from central and west African samples. However, it should be noted that in spite of the continent-wide distribution of the S-form, the number of samples available for this study was quite limited, particularly in west Africa. Moreover, central African samples were also concentrated within a relatively small area separated by a maximum distance of <500 km. This restricted sampling could have influenced the results, especially for the spatial cluster analysis as the accuracy of these methods tends to increase with the inclusion of more spatial points (Guillot et al. 2009). Thus, while our data suggest that isolation by distance may be the predominant force in genetic structuring of the S-form, greater geographic coverage would be required to confirm if a pattern of population subdivision associated with the forest-savanna transition also occurs in this form. The third minor cluster detected included only five specimens, four of which were collected in Bissau. While this minor cluster may represent an artifact of the analysis, as the effective number of clusters may be lower than K_{max} (Durand et al. 2009), it may also represent admixed individuals between M- and S-forms, given the high levels of hybridization reported for this locality (Oliveira et al. 2008; Caputo et al. 2011; Marsden et al. 2011). In fact, this particular S-form sample from Bissau grouped together with the west African M-form samples in the PCoA and was not distinguishable from the M-form in both spatial and nonspatial Bayesian analyses performed with all samples. The differences found between spatial and nonspatial clustering models in the S-form highlight the importance of adding a spatial component into the analysis especially in cases where isolation by distance is

likely to influence the patterns of population differentiation. When all samples were analyzed together by TESS, the optimal *K* obtained reflected both the M/S subdivision and the geographic partitioning within each form.

The apparent shallow differentiation between west and southern African S-form samples is consistent with previous studies pointing to an overall shallow population differentiation within this form (Lehmann et al. 1999, 2003). These studies have detected a single major subdivision of S-form populations in east Africa associated with gene-flow restrictions imposed by the rift valley. In contrast with the M-form, whose distribution is limited to the occidental side of Africa, the relatively continuous distribution of the S-form throughout the sub-Saharan continent may provide a connection between west and southern S-form populations through the intermediate central African region west to the Rift Valley. On the other hand, the heterogeneous haplotype distribution of genes conferring knockdown insecticide resistance is consistent with a possible partitioning between rainforest and savannah S-form populations (Pinto et al. 2007; Lynd et al. 2010). While differences in insecticide selection pressure are likely to be the major force shaping the distribution of *kdr* haplotypes, forest/savannah restrictions to gene flow may also contribute to the observed heterogeneities.

The results obtained in this study show that in addition to the M- and S-forms partitioning and to the existence of local or regional genetic variants (Coluzzi et al. 1979; Riehle et al. 2011), population subdivision may occur at a macrogeographic scale in *An. gambiae*, at least within the M-form. This trend appears to be associated with the transition between forest and savannah biomes and appears to be evident both northwards and southwards from the central African rainforest belt. This complexity is of importance to the management of malaria vector control programs. A genetic discontinuity between savannah and forest biomes is likely to influence dispersal and distribution of genes of practical importance to malaria epidemiology and control, such as genes associated with insecticide resistance or with vector competence.

Acknowledgements

We are grateful to the colleagues that performed the mosquito collections in the localities included in this study. This work received financial support by the UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR, A50239). BG was funded by a PhD fellowship of Fundação para a Ciência e Tecnologia/MCTES/FEDER Portugal (SFRH/BD/36410/2007).

Data archiving statement

Data for this study available on Dryad: doi:10.5061/dryad.201rm.

Literature cited

Antao, T., A. Lopes, R. J. Lopes, A. Beja-Pereira, and G. Luikart 2008. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinformatics **9**:323.

Beaumont, M. A., and R. A. Nichols 1996. Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society B: Biological Sciences **263**:1619–1626.

Calzetta, M., F. Santolamazza, G. C. Carrara, P. J. Cani, F. Fortes, M. A. Di Deco, A. della Torre et al. 2008. Distribution and chromosomal characterization of the *Anopheles gambiae* complex in Angola. American Journal of Tropical Medicine and Hygiene **78**:169–175.

Caputo, B., D. Nwakanma, M. Jawara, M. Adiamoh, I. Dia, L. Konate, V. Petrarca 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., F. Santolamazza, J. L. Vicente, D. C. Nwakanma, M. Jawara, K. Palsson, T. Jaenson et al. 2011. The "far-west" of *Anopheles gambiae* molecular forms. PLoS ONE **6**:e16415.

Chen, C., E. Durand, F. Forbes, and O. Francois 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Molecular Ecology Notes **7**:747–756.

Choi, K. S., and H. Townson 2012. Evidence for X-linked introgression between molecular forms of *Anopheles gambiae* from Angola. Medical and Veterinary Entomology **26**:218–227.

Coluzzi, M., A. Sabatini, V. Petrarca, and M. A. Di Deco 1979. Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. Transactions of the Royal Society of Tropical Medicine and Hygiene **73**:483–497.

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

Costantini, C., D. Ayala, W. M. Guelbeogo, M. Pombi, C. Y. Some, I. H. Bassole, K. Ose et al. 2009. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. BMC Ecology **9**:16.

Crispo, E., J. S. Moore, J. A. Lee-Yaw, S. M. Gray, and B. C. Haller 2011. Broken barriers: human-induced changes to gene flow and introgression in animals: an examination of the ways in which humans increase genetic exchange among populations and species and the consequences for biodiversity. BioEssays **33**:508–518.

Dabiré, K. R., A. Diabaté, M. Namountougou, K. H. Toé, A. Ouari, P. Kengne, C. Bass et al. 2009. Distribution of pyrethroid and DDT resistance and the L1014F *kdr* mutation in *Anopheles gambiae* s.l. from Burkina Faso West Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene **103**:1113–1120.

Deitz, K. C., G. Athrey, M. R. Reddy, H. J. Overgaard, A. Matias, J. Musa, A. della Torre et al. 2012. Genetic isolation within the malaria mosquito *Anopheles melas*. Molecular Ecology **21**:4498–4513.

Diabaté, A., R. K. Dabiré, K. Heidenberger, J. Crawford, W. O. Lamp, L. E. Culler, and T. Lehmann 2008. Evidence for divergent selection

between the molecular forms of *Anopheles gambiae*: role of predation. BMC Evolutionary Biology **8**:5.

Donnelly, M. J., N. Cuamba, J. D. Charlwood, F. H. Collins, and H. Townson 1999. Population structure in the malaria vector, *Anopheles arabiensis* Patton, in East Africa. Heredity **83**:408–417.

Donnelly, M. J., F. Simard, and T. Lehmann 2002. Evolutionary studies of malaria vectors. Trends in Parasitology **18**:75–80.

Durand, E., F. Jay, O. E. Gaggiotti, and O. Francois 2009. Spatial inference of admixture proportions and secondary contact zones. Molecular Biology and Evolution **26**:1963–1973.

Dyer, N. A., A. Furtado, J. Cano, F. Ferreira, M. O. Afonso, N. Ndong-Mabale, P. Ndong-Asumu et al. 2009. Evidence for a discrete evolutionary lineage within Equatorial Guinea suggests that the tsetse fly *Glossina palpalis palpalis* exists as a species complex. Molecular Ecology **18**:3268–3282.

Earl, D. A., and B. M. vonHoldt 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources **4**:359–361.

El Mousadik, A., and R. J. Petit 1996. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* L. Skeels] endemic to Morocco. Theoretical and Applied Genetics **92**:832–839.

Esnault, C., M. Boulesteix, J. B. Duchemin, A. A. Koffi, F. Chandre, R. Dabiré, V. Robert et al. 2008. High genetic differentiation between the M and S molecular forms of *Anopheles gambiae* in Africa. PLoS ONE **3**:e1968.

Estoup, A., P. Jarne, and J.-M. Cornuet 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Molecular Ecology **11**:1591–1604.

Etang, J., J. L. Vicente, P. Nwane, M. Chouaibou, I. Morlais, V. E. do Rosario, F. Simard et al. 2009. Polymorphism of intron-1 in the voltage-gated sodium channel gene of *Anopheles gambiae* s.s. populations from Cameroon with emphasis on insecticide knockdown resistance mutations. Molecular Ecology **18**:3076–3086.

Evanno, G., J. Goudet, and S. Regnaut 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology **14**:2611–2620.

Favia, G., A. della Torre, M. Bagayoko, A. Lanfrancotti, N. Sagnon, Y. T. Touré, and M. Coluzzi 1997. Molecular identification of sympatric chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive isolation. Insect Molecular Biology **6**:377–383.

François, O., S. Ancelet, and G. Guillot 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics **174**:805–816.

Frantz, A. C., S. Cellina, A. Krier, L. Schley, and T. Burke 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? Journal of Applied Ecology **46**:493–505.

Gimonneau, G., J. Bouyer, S. Morand, N. J. Besansky, A. Diabate, and F. Simard 2010. A behavioral mechanism underlying ecological divergence in the malaria mosquito *Anopheles gambiae*. Behavioral Ecology **21**:1087–1092.

Gimonneau, G., M. Pombi, M. Choisy, S. Morand, R. K. Dabiré, and F. Simard 2012. Larval habitat segregation between the molecular forms of the mosquito *Anopheles gambiae* in a rice field area of Burkina Faso, West Africa. Medical and Veterinary Entomology **26**:9–17.

Goudet, J. 1995. FSTAT version 1.2: a computer software to calculate *F*-statistics. Journal of Heredity **86**:485–486.

Guillot, G., R. Leblois, A. Coulon, and A. C. Frantz 2009. Statistical methods in spatial genetics. Molecular Ecology **18**:4734–4756.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**:65–70.

Jakobsson, M., and N. A. Rosenberg 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics **23**:1801–1806.

Kamdem, C., B. Tene Fossog, F. Simard, J. Etouna, C. Ndo, P. Kengne, P. Boussès et al. 2012. Anthropogenic habitat disturbance and ecological divergence between incipient species of the malaria mosquito *Anopheles gambiae*. PLoS ONE **7**:e39453.

Kaufmann, C., and H. Briegel 2004. Flight performance of the malaria vectors *Anopheles gambiae* and *Anopheles atroparvus*. Journal of Vector Ecology **29**:140–153.

Lanzaro, G. C., Y. T. Touré, J. Carnahan, L. Zheng, G. Dolo, S. Traoré, V. Petrarca et al. 1998. Complexities in the genetic structure of *Anopheles gambiae* populations in west Africa as revealed by microsatellite DNA analysis. Proceedings of the National Academy of Sciences of the United States of America **95**:14260–14265.

Lawniczak, M. K., S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson, B. White, S. Redmond et al. 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science **330**:512–514.

Lee, Y., A. J. Cornel, C. R. Meneses, A. Fofana, A. G. Andrianarivo, R. D. McAbee, E. Fondjo et al. 2009. Ecological and genetic relationships of the Forest-M form among chromosomal and molecular forms of the malaria vector *Anopheles gambiae* sensu stricto. Malaria Journal **8**:75.

Lehmann, T., W. A. Hawley, H. Grebert, M. Danga, F. Atieli, and F. H. Collins 1999. The rift valley complex as a barrier to gene flow for *Anopheles gambiae* in Kenya. Journal of Heredity **90**:613–621.

Lehmann, T., M. Licht, N. Elissa, B. T. Maega, J. M. Chimumbwa, F. T. Watsenga, C. S. Wondji et al. 2003. Population Structure of *Anopheles gambiae* in Africa. Journal of Heredity **94**:133–147.

Lynd, A., D. Weetman, S. Barbosa, A. Egyir Yawson, S. Mitchell, J. Pinto, I. Hastings et al. 2010. Field, genetic, and modeling approaches show strong positive selection acting upon an insecticide resistance mutation in *Anopheles gambiae* s.s. Molecular Biology and Evolution **27**:1117–1125.

Marsden, C. D., Y. Lee, C. C. Nieman, M. R. Sanford, J. Dinis, C. Martins, A. Rodrigues 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.

Martinez-Torres, D., F. Chandre, M. S. Williamson, F. Darriet, J. B. Bergé, A. L. Devonshire, P. Guillet et al. 1998. Molecular characterization of pyrethroid knockdown resistance *kdr* in the major malaria vector *Anopheles gambiae* s.s. Insect Molecular Biology **7**:179–184.

Moreno, M., P. Salgueiro, J. L. Vicente, J. Cano, P. J. Berzosa, A. de Lucio, F. Simard et al. 2007. Genetic population structure of *Anopheles gambiae* in Equatorial Guinea. Malaria Journal **6**:137.

Ndo, C., C. Antonio-Nkondjio, A. Cohuet, D. Ayala, P. Kengne, I. Morlais, P. H. Awono-Ambene et al. 2010. Population genetic structure of the malaria vector *Anopheles nili* in sub-Saharan Africa. Malaria Journal **9**:161.

Neafsey, D. E., M. K. Lawniczak, D. J. Park, S. N. Redmond, M. B. Coulibaly, S. F. Traoré, N. Sagnon et al. 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. Science **330**:514–517.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nielsen, E. E., L. A. Bach, and P. Kotlicki 2006. Hybridlab version 1.0: a programme for generating simulated hybrids from population samples. Molecular Ecology Notes **6**:971–973.

Oliveira, E., P. Salgueiro, K. Palsson, J. L. Vicente, A. P. Arez, T. G. Jaenson, A. Caccone et al. 2008. High levels of hybridization between molecular forms of *Anopheles gambiae* from Guinea Bissau. Journal of Medical Entomology **45**:1057–1063.

Peakall, R., and P. E. Smouse 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes **6**:288–295.

Pinto, J., M. J. Donnelly, C. A. Sousa, V. Gil, C. Ferreira, N. Elissa, V. E. do Rosário et al. 2002. Genetic structure of *Anopheles gambiae* Diptera: Culicidae in São Tomé and Príncipe West Africa: implications for malaria control. Molecular Ecology **11**:2183–2187.

Pinto, J., A. Lynd, J. L. Vicente, F. Santolamazza, N. P. Randle, G. Gentile, M. Moreno et al. 2007. Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector *Anopheles gambiae*. PLoS ONE **2**:e1243.

Pritchard, J. K., M. Stephens, and P. Donnelly 2000. Inference of population structure using multilocus genotype data. Genetics **155**:945–959.

Raymond, M., and F. Rousset 1995. GENEPOP version 1.2: population genetics software for exact tests and ecumenicism. Journal of Heredity **86**:248–249.

Ridl, F. C., C. Bass, M. Torrez, D. Govender, V. Ramdeen, L. Yellot, A. E. Edu et al. 2008. A pre-intervention study of malaria vector abundance in Rio Muni, Equatorial Guinea: their role in malaria transmission and the incidence of insecticide resistance alleles. Malaria Journal **7**:194.

Riehle, M. M., W. M. Guelbeogo, A. Gneme, K. Eiglmeier, I. Holm, E. Bischoff, T. Garnier et al. 2011. A cryptic subgroup of *Anopheles gambiae* is highly susceptible to human malaria parasites. Science **331**:596–598.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics **145**:1219–1228.

Santolamazza, F., M. Calzetta, J. Etang, E. Barrese, I. Dia, A. Caccone, M. J. Donnelly et al. 2008. Distribution of knock-down resistance mutations in *Anopheles gambiae* molecular forms in west and west-central Africa. Malaria Journal **7**:74.

Schwartz, M. K., and K. S. Mckelvey 2009. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conservation Genetics **10**:441–452.

Scott, J. A., W. G. Brogdon, and F. H. Collins. 1993. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. American Journal of Tropical Medicine and Hygiene **49**:520–529.

Simard, F., D. Ayala, G. C. Kamdem, M. Pombi, J. Etouna, K. Ose, J. M. Fotsing et al. 2009. Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. BMC Ecology **9**:17.

Slotman, M. A., M. M. Mendez, A. D. Torre, G. Dolo, Y. T. Touré, and A. Caccone 2006. Genetic differentiation between the BAMAKO and SAVANNA chromosomal forms of *Anopheles gambiae* as indicated by amplified fragment length polymorphism analysis. American Journal of Tropical Medicine and Hygiene **74**:641–648.

Slotman, M. A., F. Tripet, A. J. Cornel, C. R. Meneses, Y. Lee, L. J. Reimer, T. C. Thiemann et al. 2007. Evidence for subdivision within the M molecular form of *Anopheles gambiae*. Molecular Ecology **16**:639–649.

Taylor, C., Y. T. Touré, J. Carnahan, D. E. Norris, G. Dolo, S. F. Traoré, F. E. Edillo et al. 2001. Gene flow among populations of the malaria vector, *Anopheles gambiae*, in Mali, West Africa. Genetics **157**:743–750.

della Torre, A., C. Fanello, M. Akogbeto, J. Dossou-yovo, G. Favia, V. Petrarca, and M. Coluzzi 2001. Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. Insect Molecular Biology **10**:9–18.

della Torre, A., C. Costantini, N. J. Besansky, A. Caccone, V. Petrarca, J. R. Powell, and M. Coluzzi 2002. Speciation within *Anopheles gambiae*-the glass is half full. Science **298**:115–117.

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

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biology **3**:e285.

UNEP 2010. Africa water atlas. United Nations Environment Programme. Progress Press Limited, Malta. http://na.unep.net/atlas/africaWater/book.php (accessed on January 2013).

Vähä, J. P., and C. R. Primmer 2006. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Molecular Ecology **15**:63–72.

Van Oosterhout, C., W. F. Hutchinson, D. P. M. Wills, and P. Shipley 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes **4**:535–538.

Weetman, D., C. S. Wilding, K. Steen, J. C. Morgan, F. Simard, and M. J. Donnelly 2010. Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilbrium genome. PLoS ONE **5**:e13140.

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

Weir, B. S., and C. C. Cockerham 1984. Estimating F-statistics for the analysis of population structure. Evolution **38**:1358–1370.

White, B. J., C. Cheng, F. Simard, C. Costantini, and N. J. Besansky 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. Molecular Ecology **19**:925–939.

White, B. J., M. K. Lawniczak, C. Cheng, M. B. Coulibaly, M. D. Wilson, N. Sagnon, C. Costantini et al. 2011. Adaptive divergence between incipient species of *Anopheles gambiae* increases resistance to *Plasmodium*. Proceedings of the National Academy of Sciences of the United States of America **108**:244–249.

Yawson, A. E., D. Weetman, M. D. Wilson, and M. J. Donnelly 2007. Ecological zones rather than molecular forms predict genetic differentiation in the malaria vector *Anopheles gambiae* s.s. in Ghana. Genetics **175**:751–761.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Geographic information, year of collection and molecular form composition of the samples analyzed.

Table S2. Microsatellite loci genotyped.

Table S3. Microsatellite genetic diversity estimates according to collection site.

Table S4. Power and accuracy of the Bayesian clustering analysis implemented by STRUCTURE (Pritchard et al. 2000) to detect M and S form simulated individuals (N = 100 for each form).

Table S5. Isolation by distance model regressions of $F_{ST}/(1\text{-}F_{ST})$ on logarithm of distance.

Table S6. Pair-wise estimates of F_{ST} (below diagonal) and geographic distance (above diagonal, in kilometres) between sampling sites.

Figure S1. Plots between genetic differentiation and expected heterozygosity to detect candidate microsatellite loci under selection according to the method implemented in LOSITANT (Antao et al. 2008).

Figure S2. Graphics of Evanno's ΔK for the different Bayesian clustering analyses implemented by STRUCTURE.

Figure S3. Bayesian clustering analysis implemented by STRUCTURE (Pritchard et al. 2000) and spatially explicit analyses implemented by TESS (Chen et al. 2007), performed with 10 microsatellite loci.

Figure S4. Plots of DIC values (Y-axis) against K_{max} (X-axis) obtained for TESS analysis (Chen et al. 2007) under two admixture models (CAR and BYM).

