Research paper

Genetic analysis and population structure of the Anopheles gambiae complex from different ecological zones of Burkina Faso

Abdou Azaque Zouré, Grégoire Noël, Aboubacar Sombié, Zéphirin Somda, Athanase Badolo, Frédéric Francis

Keywords:
Anopheles gambiae s.l.
COI
rDNA
genetic structure
Burkina Faso

ARTICLE INFO

ABSTRACT

The Anopheles gambiae complex (Diptera: Culicidae) is the most important vector for malaria in Sub-Saharan Africa, besides other vectors such as Anopheles funestus. Malaria vector control should encompass specific identification, genetic diversity and population structure of An. gambiae to design vector control strategies.

The aim of this study was to determine the distribution of sibling species of the An. gambiae complex according to climatic regions related to cotton-growing or cotton-free areas by using polymerase chain-restriction fragment length polymorphism (PCR-RFLP). Then, variation in mitochondrial cytochrome c oxidase 1 (COI) was used to assess the genetic structure within and between populations from our selected ecological zones.

At the sibling species level, the following proportions were found across all samples (n = 180): An. coluzzii 65.56%, An. gambiae stricto sensu (s.s.) 21.11%, and An. arabiensis 3.33%. Hybrids between An. gambiae s.s. and An. coluzzii (7.78%) and hybrids between An. coluzzii and An. arabiensis (2.22%) were found. The phylogenetic tree and Integer Neighbour-Joining (IntNJ) haplotype network did not reveal any distinct genetic structure pattern related to climatic or agricultural conditions in Burkina Faso. The Fst (Wright’s F-statistic) values close to zero showed a free gene flow and no differentiation in An. gambiae complex populations. Furthermore, neutrality indices calculated by Tajima’s D, Fu and Li’s D, Fu and Li’s F, Fu’s F tests suggested an excess of rare mutations in the populations.

Overall, variation in the proportions of An. gambiae s.s., An. coluzzii and An. arabiensis was found according to climatic regions, but COI analysis did not evidence any population structuring of the An. gambiae complex. These scientific contributions can be used as a basis for further in-depth study of the genetic diversity of the An. gambiae complex for epidemiological risk assessment of malaria in Burkina Faso.

1. Introduction

In Sub-Saharan Africa, malaria is still a public health challenge (WHO, 2018). Early diagnosis of human infection by the parasite and medical therapy (prevention and treatment) constitute the basis of malaria control, while indoor residual spraying (IRS) and long-lasting insecticide nets (LLINs) are used for vector control (Diallo et al., 2015; WHO, 2018). Moreover, malaria control should encompass identification of the vector species or sub-species as well as their population structure (Besansky et al., 2003). To date, 485 described species belong to the Anopheles genus (Diptera: Culicidae) (Harbach and Kitching, 2016). An estimated 70 species are regularly associated with transmission of Plasmodium sp. to humans worldwide (Stevenson and Norris, 2017). Members of Anopheles gambiae sensu lato (s.l.) (named the Anopheles gambiae complex) are among the most important vectors for malaria transmission in Sub-Saharan Africa (Benson and Falade, 2013). The An. gambiae complex is also involved in the transmission of Bancroft filaria (Wuchereria bancrofti) and several arboviruses such as O’Nyong-Nyong virus, Tataguin virus, Chikungunya virus and Ilesha virus (Carnevale and Robert, 2009; Hegde et al., 2015). Before 1962, An. gambiae s.l. was considered to be a single species (Coetzee et al., 2000). But the evolution of molecular biology techniques caused An.
gambiae s.l. to break into seven sibling species, all morphologically indistinguishable (Borja-Cacho and Matthews, 2008). Among them, An. arabiensis Patton 1904, An. gambiense sensu stricto (s.s.) Giles 1902 (S-form or Savannah or Bamako form), and An. coluzzii Coetzee and Willkerson 2013 (M-form or Mopiti form) are the most anthropophagic vectors of malaria in Sub-Saharan Africa (Coetzee et al., 2013; Della Torre et al., 2001; Pates et al., 2014; Wiebe et al., 2017). Although hybridisation is possible in many sympatric areas, natural hybrids between these sibling species are very rare (Yawson et al., 2007; Pates et al., 2014). Speciation processes (Miles et al., 2017) and chromosomal inversion indicate genetic heterogeneity between An. arabiensis, An. gambiase s.s. and An. coluzzii in West Africa (Coetzee et al., 2000; Borja-Cacho and Matthews, 2008). Therefore, each of these sibling species (An. arabiensis, An. gambiase s.s. and An. coluzzii) has been clearly protected by effective reproductive isolation (Della Torre et al., 2001). This speciation process could also be explained by a combination of different factors, such as the intensity of rainfall or the type of agricultural system in any given area, as Anopheles mosquitoes depend on standing water for a breeding habitat (Bombly, 2012). In Burkina Faso (Africa), the differences in epidemiological facies of malaria follow the different rainfall levels in the three climatic regions, which are the Sahelian (short malaria transmission period of two to three months per year), Sudano-Sahelian (seasonal transmission four to six months per year) and Sudanian (perennial transmission) regions (Carnevale and Robert, 2009; Abiodun et al., 2016). One-third of the gross domestic product of Burkina Faso is possible from the agricultural economic sector, primarily food grains (sorghum, millet, maize, and rice). Because of the use of insecticides, farmland may be considered as a source of selection pressure on Anopheles mosquitoes, especially in cotton-growing areas (Jien et al., 2017) where cotton is the main cash crop (IFPRI, 2010). During the agricultural seasons of the last decade (2000), Burkina Faso was the leading cotton producer in West Africa with about 600,000 tons of seed cotton/year (Hauchart, 2007). Research suggests that certain fertilizer contents may influence mosquito egg-laying behaviour. The use of these fertilizers is combined with insecticide treatments, which have attractive properties for mosquitoes and are found in the waters where mosquitoes lay their eggs (Darriet and Corbel, 2008; Darriet et al., 2012). In sum, temperature, rainfall and agriculture could influence the species density and structure of the An. gambiae species complex (Minakawa and Sonye, 2002).

An accurate and predictive understanding of the geographic distribution of these Anopheles species should permit the identification of areas in which particular co-habituating species are potentially involved in further disease transmission (Levine et al., 2004). In Burkina Faso, most research studies have been carried out on An. gambiae s.l. (Dabiré et al., 2012) but not specifically according to ecological conditions. Therefore, the aim of this study was firstly to use molecular techniques to determine the distribution of the sibling species of the An. gambiae complex (An. arabiensis, An. gambiae s.s. and An. coluzzii in the present case) within and between populations according to their localisation in the Sahelian, Sudano-Sahelian and Sudanian climatic regions in cotton-growing or cotton-free areas. Secondly, a fragment of the COI gene was sequenced to distinguish the population structure according to agricultural and climatic features. The distribution of An. gambiae s.l. sibling species and its population structure could be an important monitoring tool for malaria control in Burkina Faso.

2. Materials and methods

2.1. Sample collection

Adult females of the An. gambiae s.l. complex were collected from human residential areas in different localities of Burkina Faso (Fig. 1) using a Prokopack aspirator (a semi-craft production) between August and October 2017. Malaria transmission generally reaches its seasonal peak at that time of year and corresponds to the population peak of An. gambiae s.l. in the three climate regions (INSD, 2015). The Sudanian climatic region (annual rainfall > 900 mm; average annual temperature of 31 °C) contributes to malaria transmission all year round. In the Sudano-Sahelian region (annual rainfall between 600 mm and 900 mm; average annual temperature of 33 °C) and the Sahelian region (annual rainfall < 600 mm; average annual temperature of 35 °C), malaria transmission is seasonal and lasts four to six months and two to three months, respectively (Kabore et al., 2017). The collection times ranged between 6:00–8:00 a.m. and 5:00–7:00 p.m. Mosquitoes were collected indoors and outdoors in six localities (two in each climatic region): Kongoussi (KG); Ouahigouya (OH); Sanguié (SG); Kombissiri (KB); Hounédé (HD), and Bobo (BB) (three villages in each collection site), including three sites in cotton-growing areas and three in cotton-free areas. Thirty individuals from the An. gambiae complex were collected per locality.

2.2. Morphological identification of mosquitoes

All samples (n = 180) were brought back to the Laboratoire d’Entomologie Fondamentale et Appliquée (Ouagadougou, Burkina Faso). All specimens were prepared and identified morphologically according to the identification keys of Gillies and De Meillon (1968) and Gillies and Coetzee (1987). These morphological keys can distinguish differences among groups of African Anopheles species but do not distinguish the sibling species. Specimens were washed in sterile phosphate-buffered saline (PBS 1 x, pH 7.8), and dissected into three parts (wings/legs, head/thorax, and midgut) preserved in ethanol 70%.

2.3. Molecular identification

All our dissected samples were brought to Gembloux Agro-Bio Tech, Functional & Evolutionary Entomology (Gembloux, Belgium) and stored at −80 °C until DNA extraction. First, the samples were washed twice in sterile distilled water to remove traces of ethanol and to avoid contamination. DNA was extracted from the mosquito samples (wings/legs) using a QIAamp DNA Micro Kit (QIAGEN, Valencia, CA) according to the manufacturer’s instructions. Briefly, the samples were crushed individually with plastic pestles in 1.5-ml microtubes with 180 μl of buffer ATL, and 20 μl of proteinase K were added. Mixtures were vortexed thoroughly for 15 s and incubated in a thermomixer (56 °C, 450 rpm) overnight. After processing and washing with the adequate product provided in the kit, the final DNA product was dissolved in 20 μl of elution buffer AE. DNA purity and quality were assessed using a spectrophotometer (Nanodrop®), and DNA was stored at −20 °C until PCR analysis.

A total of 180 mosquitoes of the An. gambiae complex was identified as being An. gambiae s.s., An. coluzzii or An. arabiensis by the PCR-RFLP method using the intergeneric spacer (IGS) of ribosomal DNA (rDNA) as the target gene (Fanello et al., 2002). We combined the protocols established by Scott et al. (1993) and Favia et al. (2001) with the enzyme Hha I and restriction site GGC C. Briefly, 25 ng of DNA were amplified in 25 μl of master mix containing 1× Buffer-MgCl2, 2.5 mM MgCl2, 0.2 μM of each dNTP, 12.5 ng of UN primer, 6.25 ng of GA primer, 18.75 ng of AR primer, 25 ng of QD primer, 12.5 ng of ME primer and 4 U of Platinum™ Taq DNA Polymerase (ThermoFisher Scientific™) (Table 1).

The PCR started with an initial step of ten minutes at 94 °C to activate the DNA polymerase, followed by 30 cycles each consisting of denaturation for 30 s at 94 °C, hybridisation for 30 s at 50 °C, extension for 30 s at 72 °C, and a final cycle for seven minutes at 72 °C. After amplification, 10 μl of PCR products were added directly to the 10-μl mix containing 5 U of Hha I (10 U/μl ThermoFisher Scientific™), 2 ng of BSA (10 ng/μl), and 1 x of RE Buffer C (10 x); digestion was performed at 37 °C for a minimum of three hours, following the manufacturer’s instructions. Amplified fragments were revealed and confirmed by electrophoresis on a 2% agarose gel and visualised under UV light. The
product sizes following the PCR and the fragment lengths of the sibling species following Hha I digestion are presented in Table 1.

To study the population structure and to build a phylogenetic tree of our dataset related to the An. gambiae species complex, the COI gene or barcode was also amplified from our DNA extracts \((n = 180)\) using primers COI (Forward) and Cul-Rev (Reverse) (Table 1). The PCR reaction mixtures were as follows: a final volume of 25 μL containing 10 mM Tris-HCl (pH = 8.4), 150 mM KCl, 0.10% Triton X – 100, 0.2 μg/μl of bovine serum albumin (BSA), 1.5 mM MgCl₂, 0.2 mM of each dNTP, 30 ng of each primer, 2 U of Platinum™ Taq DNA Fig. 1. Map of the collection sites showing climatic regions and cotton-growing or cotton-free areas: Sahelian (Kongoussi; Ouahigouya), Sudano-Sahelian (Kombissiri; Sanguié) and Sudanian (Houndé; Bobo).

### Table 1

<table>
<thead>
<tr>
<th>Sequences of primers/Names of sibling species</th>
<th>Length of the fragments after Hha I digestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR primers for the intergenic space (IGS) of ribosomal DNA (rDNA) *</td>
<td></td>
</tr>
<tr>
<td>UN primer [5’-GTG TGC CCC TCT CTC GAT GT-3’]</td>
<td></td>
</tr>
<tr>
<td>GA primer [5’-CTG GTT TGG TCG GCA GGT TT-3’]</td>
<td></td>
</tr>
<tr>
<td>An. gambiae s.s. (S-form)</td>
<td>390 bp and 257 bp, 110 pb, 23 bp</td>
</tr>
<tr>
<td>An. coluzzii (M-form)</td>
<td>390 bp and 367 bp, 23 bp</td>
</tr>
<tr>
<td>AR primer [5’-AAG TGT CCT TCT CCA TCC TA-3’]</td>
<td>315 bp and 292 bp, 23 bp</td>
</tr>
<tr>
<td>An. arabiensis Patton</td>
<td></td>
</tr>
<tr>
<td>QD primer [5’-CAG ACC AAG ATG GCT ATT AT-3’]</td>
<td>153 bp and 124 bp, 6 bp, 23 bp</td>
</tr>
<tr>
<td>An. quadriannulatus (Theobald)</td>
<td></td>
</tr>
<tr>
<td>ME primer [5’-TGA CCA ACC CAC CTT TCC GA-3’]</td>
<td>464 bp and 435 bp, 6 bp, 23 bp</td>
</tr>
<tr>
<td>An. melas Theobald</td>
<td></td>
</tr>
<tr>
<td>An. merus J. Theobald</td>
<td>466 bp and 437 bp, 29 bp</td>
</tr>
<tr>
<td>PCR primers for the cytochrome oxidase subunit COI **</td>
<td></td>
</tr>
<tr>
<td>Forward-COI [5’-TTG ATT TTT TGG TCA TCC AGA AGT-3’]</td>
<td>877 bp</td>
</tr>
<tr>
<td>Reverse-Cul-Rev [5’-TAG AGC TTA AAT TCA TGG CAC TAA TC-3’]</td>
<td></td>
</tr>
</tbody>
</table>

bp: base pair; * and ** according to Fanello et al. (2002) and Oshaghi et al. (2006) respectively.
Polymerase (ThermoFisher Scientific™) and 25 ng of DNA sample (Oshaghi et al., 2006). The PCR profile was as follows: four minutes of initial denaturation at 94 °C, 32 cycles of one minute at 94 °C, one minute at 55 °C, two minutes at 72 °C, and a final extension step of seven minutes at 72 °C (Norris and Norris, 2015). A total of 10 μl of each PCR product was loaded for electrophoresis on a 1.5% agarose gel, stained with SYBR® Safe™ DNA, and visualised under UV light. The remainder of each successful PCR reaction was purified using a NucleoSpin® gel and PCR clean-up kit (Macherey-Nagel GmbH & Co. KG. Germany). After the purification step, COI amplicons were sequenced in both directions using Sanger sequencing at GATC Biotech (A Eur fins Genomics Company, Germany).

2.4. Molecular data processing and phylogenetic analysis

Barcode sequencing (Ratnasingham and Hebert, 2007) is commonly used for identifying Anopheles (Culicidae) species (Mohanty et al., 2009). The sequences obtained from Sanger sequencing were corrected manually using BioEdit v.7.2.6 software (Hall, 1999) and edited to acquire the consensus sequences. After BLASTn examination to control the quality of all sequences, each barcode sequence (n = 151) included a maximum of 765 bp. All COI sequences were available at GenBank under accession numbers MK681509-MK681659.

For phylogenetic reconstruction, we followed the guidelines of Harrison and Langdale (2006). The first step in the phylogenetic methodology relies on sequence alignment to highlight homologies in nucleotide positions. All COI sequences were aligned using ClustalX v.2.1 (Larkin et al., 2007) with default parameters and pairwise deletion for gap treatments.

Since nucleotide positions within a sequence do not evolve independently (e.g., unequal base frequencies or substitution rates) (Posada, 2003), the second step consisted in selecting the best evolutionary model that explained the empirical evolution of sequences (i.e., our aligned sequence dataset). Therefore, we searched for the best substitution model using PartitionFinder v.2.1.1. (Lanfear et al., 2016) with a linked branch length parameter, the AJCc model selection metrics and the greedy search algorithm (Lanfear et al., 2012). The transversional (TVM) model (Posada, 2003) was selected. The third step of phylogenetic reconstruction consisted in generating and analysing the phylogenetic tree with support values. We then conducted Bayesian Inference (BI) for the An. gambiae complex using BEAST2 v.2.5.2. environment (Bouckaert et al., 2014) to reconstruct an ultrametric tree. The yule model was set up as ‘Prior Tree’ with a Markov Chain Monte Carlo (MCMC) length of five million. We then summarised posterior probabilities using TreeAnnotator v.2.5.2. (included in BEAST v.2.5.2) with a burn-in of 10%. Phylogenetic trees were built using FigTree v.1.4.3. (Rambaut, 2017). For tree topologies, we considered posterior probabilities above 0.95 as the threshold for well-supported nodes (Wilcox et al., 2002).

2.5. Population structure visualisation and statistical analysis

Regardless of the subspecies, we defined six populations of the An. gambiae complex corresponding to sampling localities based on the combination of agricultural pressure and climatic regions. All our population structure analyses used COI as a molecular marker. Population diversity indices such as number of polymorphic (segregating) sites (S), haplotype number (h), haplotype diversity (Hd), nucleotide diversity (π) and average number of pairwise nucleotide differences within a population (K) were estimated using DnaSP v.6.12.01 software (Librado and Rozas, 2009). The haplotype number (h) refers to a nucleotide sequence that may be common to several individuals but differs from other haplotypes in at least one nucleotide substitution. Haplotype diversity (Hd) defines the probability for two randomly selected genes in a sample to be identical (Nei, 1987). Nucleotide diversity (π) measures the average nucleotide divergence between all sequence pairs in a given sample with the probability that two randomly selected genes are different by a nucleotide (Tajima, 1983). The neutrality indices of Tajima’s D, Fu and Li’s D* and F*, and Fu’s Fs were also calculated in each population with DnaSP v.6.12.01. Tajima’s D was estimated by comparing the differences between the number of segregating sites (S) and the average number of nucleotide differences (K). Fu and Li’s D* and F* statistics compare estimates of theta based on mutations in the internal and external branches of a genealogy. Fu’s Fs were used to assess the structure of the haplotype frequency distribution. In addition, the average number of pairwise nucleotide differences (Ksy) and nucleotide substitutions per site (Dxy) between populations were also calculated using DnaSP v.6.12.01. Pairwise genetic differences were estimated for all populations by calculating Wright’s F-statistic (Fst) and carrying out analyses of molecular variance (AMOVA) and gene flow (Nm) in Arlequin 3.5.1.2 software (Excoffier et al., 2005). According to Wright (1969), the Fst metric estimates the extent of gene flow between various genotypes (Raymond and Rousset, 2013). AMOVA was tested to detect population (i.e. locality) segregation with a Euclidian distance matrix derived from the combined climatic region and agricultural pressure data. We implemented the non-parametric permutation approach (n = 999) to test if our populations were significantly structured.

Isolation by distance was calculated by examining the correlation between genetic distance Fst/(1-Fst) and geographical distance in log (km) of the samples. Null hypotheses assumed that the regression slope of genetic distance over geographical distance was null.

For the analysis and visualisation of An. gambiae s.l. haplotypes, we used the representation of reticulate networks. This method encompasses reticulation events such as evolutionary histories of the different haplotypes. The edges provide the evolutionary history of the lineages such as speciation events, and the nodes correspond to supposed ancestors (Huson and Bryant, 2006). Therefore, our final aligned barcode dataset (in FASTA format) was transformed in NEXUS format (Maddison et al., 1997), and then POPART v.1.7. (Leigh and Bryant, 2015) was used to build haplotype networks for climatic region and agricultural areas with the IntNJ method and a reticulation tolerance parameter (α) of 0.5. This approach of haplotype network construction is more suitable for a dataset with low genetic divergence, as expected in a species complex.

3. Results

3.1. Identification and distribution of the Anopheles gambiae complex

We identified 180 Anopheles specimens using the PCR-RFLP method. An. coluzzii was the dominant species with 65.56% occurrence among all samples, followed by An. gambiae s.s. (21.11%) and An. arabiensis (3.33%). These three species of the An. gambiae complex were found in all climatic regions as well as in cotton-growing and cotton-free areas. Hybrids between An. gambiae s.s. and An. coluzzii (7.78%) and between An. coluzzii and An. arabiensis (2.22%) were also found (Table 2). An. coluzzii, An. arabiensis and their hybrid sequences were present in all climatic regions. Across the three climatic regions, chi-square goodness of fit tests highlighted a significant dominance of An. coluzzii (P < .006; ddl = 2; χ² = 10.403) in the Sahelian region, followed by the Sudano-Sahelian and Sudanian regions. An. gambiae s.s. (P < .001; ddl = 2; χ² = 16.769) was dominant in the Sudano-Sahelian region, followed by the Sudanian region, and hybrid specimens between An. gambiae s.s. and An. coluzzii (P < .015; ddl = 2; χ² = 8.400) were dominant in the Sudanian region, followed by the Sudano-Sahelian region. However, this difference in species distribution among climatic regions was not observed for the An. arabiensis population. No significant difference was observed in the distribution of An. gambiae populations between cotton-free and cotton-growing areas.
3.2. Phylogenetic reconstruction

Our phylogenetic analyses based on COI (Fig. 2) were poorly supported (except for some nodes) and revealed no significant clusters according to population origin. The topology of the unrooted COI phylogeny tree was not consistent with classical molecular taxonomy for *An. gambiae* s.s., *An. coluzzii* or *An. arabiensis*. Therefore, no genetic structure was observed in these three species’ populations based on COI as a single locus. Moreover, no population cluster was related to any climatic region or agricultural use in Burkina Faso. In supplementary materials A1, neighbour-joining (NJ) trees are used as a standard phylogenetic approach in DNA barcoding (Saitou and Nei, 1987) to visualise genetic distances and pairwise p-distances (i.e. proportions of variable sites) (Tamura et al., 2004).

3.3. Diversity indices and nucleotide polymorphism

Among 151 sequences, haplotype and nucleotide diversity indices were highest in the Houndé population and lowest in the Bobo population (Sudanian region). However, the *An. gambiae* complex had a high haplotypic diversity (HD) (0.95252 for all 151 sequences) combined with low nucleotide diversity (Table 6). The Fst values between Kongoussi and Ouahigouya (between Ouahigouya and Bobo) to 0.0263 (between Sanguié and Bobo) were highest in the Houndé population and lowest in the Bobo population (Table 5). The average number of nucleotide substitutions varied from 3.4331 (Ouahigouya and Sanguié) to 4.91244 (Kombissiri and Kongoussi). The Fst values between populations varied from 3.33% (Ouahigouya and Sanguié) to 0.00737 (Kombissiri and Kongoussi) (Table 5). Pairwise genetic distance (Fst) between populations varied from −0.01366 (between Ouahigouya and Bobo) to 0.0263 (between Sanguié and Ouahigouya) (Table 6). The Fst values between Kongoussi and Ouahigouya, Kombissiri and Kongoussi, Bobo and Kongoussi, Ouahigouya and Bobo, and between Sanguié and Houndé were negative, indicating no differentiation in the COI sequence (Lehmann et al., 1996). For all the other populations taken pairwise (e.g. Sanguié and Ouahigouya, Sanguié and Koudougou), the Fst value ranged from 0.00052 to 0.0263, indicating that these populations were differentiated by low gene flow (Table 6). In addition, isolation by distance was not significantly correlated with the geographical distance between populations (Pearson correlation $R = 0.376, P < .167$) (Supplementary materials A2). Furthermore, both IntNJ haplotype networks (Fig. 3) showed no distinct genetic structure pattern according to climatic or agricultural conditions. Haplotype 1 and haplotype 2 clustered 22 (14.6%) and 18 (11.9%) barcode sequences from all three climatic zones and the two agricultural pressures, respectively. Most *An. gambiae* complex haplotypes encompassed only one barcode sequence.

3.4. Analysis of molecular variance (AMOVA)

AMOVA results showed that 99.27% of the diversity was distributed within populations and 0.59% among populations, but the value was non-significant (Table 7). No significant difference in variance (F statistics) was observed, pointing to the absence of a genetic structure in the population. The low $F_{st}$ value (0.00723, $P = .26$) of variance within populations indicated non-significant genetic variability, suggesting that the sampled populations were unstructured (Table 7).

4. Discussion

4.1. Distribution of *An. arabiensis*, *An. gambiae* s.s. and *An. coluzzii*

In Sub-Saharan Africa, *An. arabiensis*, *An. coluzzii* and *An. gambiae* s.s. are sympatric, with different occurrences in Burkina Faso (Costantini et al., 2009) according to climatic conditions (Coetzee et al., 2000). Among the seven species of the *An. gambiae* complex, our PCR-RFLP analysis from indoor and outdoor collections revealed three species: *An. gambiae* s.s. (21.1%), *An. coluzzii* (65.5%), and *An. arabiensis* (3.3%). In contrast, another study in Burkina Faso found *An. gambiae* in a higher proportion (51.7%) than *An. coluzzii* (21.6%) and *An. arabiensis* (26.7%) (Costantini et al., 2009). We found a smaller population of *An. gambiae*, *An. arabiensis* as compared to *An. gambiae* s.s. and *An. coluzzii*. Our results corroborate the findings of a previous study suggesting that *An. gambiae* s.s dominates in the humid Sudanian region, while *An. arabiensis* is better represented in the Sahel steppes (Coetzee et al., 2000). But urbanisation in West Africa and Burkina Faso is linked to the distribution of *An. arabiensis*, so its distribution does not necessarily follow ecoclimatic regions (Markianos et al., 2016). In general, *An. gambiae* s.s. and *An. coluzzii* occupy wetlands (forest, wet savannah) but can be found in arid Sahelian regions bordering the Sahara, from Mauritania to Chad, Senegal, Mali, Niger and Nigeria. These two species are absent in the horn of Africa, southern Arabia and most of the Indian Ocean islands except in Madagascar, Comoros and Mauritius (Coetzee et al., 2000; Simard et al., 2009). Previous studies on the *An. gambiae* complex found *An. gambiae* s.s. and *An. coluzzii* to be more common in West Africa than in Central and Eastern Africa, while *An. arabiensis* had a wider range including the arid areas south of the Sahara and the horn of Africa (Kyalo et al., 2017). Furthermore, *An. gambiae* s.s. was found mostly dominant in savannah (Sudanian) regions, while *An. coluzzii* was typical of the Niger River delta (Simard et al., 2009) and predominant in shrublands (Sahelian region) (Lee et al., 2009), as they are in Burkina Faso and Mali. *An. coluzzii* has a decisive adaptive advantage in the dry season, while *An. gambiae* s.s. distribution is linked to its advantage in the rainy season (Bombly, 2012).

*An. coluzzii* is dominant in southwestern Burkina Faso, while *An. gambiae* s.s. occurs in rice-growing areas in a relatively high proportion, with a maximum peak at 51% in October (end of the rainy season) (Dabiré et al., 2008). *An. coluzzii* is mainly found in the Sudano-Sahelian region and is highly predominant (93%) in the irrigated rice fields.
(caption on next page)
area, while *An. gambiae* s.s. is found predominantly in the Sudanese zone (Namountougou et al., 2012). Migration models of polymorphism suggest that expansions of *An. gambiae* s.s. and *An. coluzzii* are related to agricultural pressure (Crawford and Lazzaro, 2010) while *An. arabiensis* has a high potential by responding to selective pressures from environmental changes and vector control efforts (Ishimura et al., 2015).

We found new hybrids between *An. gambiae* s.s. and *An. coluzzii* (7.78%) and between *An. arabiensis* and *An. coluzzii* (2.22%). Although hybridisation has been shown to be possible in many sympatric areas, natural hybrids between these three species of the *An. gambiae* complex are very rare (Pates et al., 2014; Pombi et al., 2017). A study in Uganda reported occasional hybridisation between *An. arabiensis*, *An. gambiae* s.s. and *An. coluzzii* in nature, showing that hybrid forms between these three species may occur (Weetman et al., 2014). In West Africa, hybridisation between *An. gambiae* s.s. and *An. coluzzii* is extremely low for the most part, with hybrid frequencies below 1%, comparable to hybrids between *An. gambiae* and *An. arabiensis* (Della Torre et al., 2005). Hybrids of all sympatric members of the *An. gambiae* complex have been encountered in nature with frequencies ranging from 0.02% to 0.76% (Besansky et al., 2003). All hybrid females are fertile, and some even show higher fertility than the female parent (Besansky et al., 2003). However, in our study, the frequency of hybrids between *An. gambiae* s.s. and *An. coluzzii* was 7.7%, and the frequency of hybrids between *An. arabiensis* and *An. coluzzii* was 2.22%. An earlier study in south-eastern Burkina Faso found no hybrid between *An. arabiensis* and *An. gambiae* s.s., but a 6% hybrid frequency between *An. coluzzii* and *An. gambiae* s.s. (Robert et al., 1989). In contrast, GOUNDRY, a recent cryptic subgroup of *An. gambiae* s.l. found in Burkina Faso (Crawford et al., 2016), displays a high rate of hybridisation (36%) (Riehle et al., 2011). GOUNDRY is a hybrid between *An. coluzzii* and *An. gambiae* s.s. and is more closely related to *An. coluzzii* (Crawford et al., 2016).

### 4.2. Phylogenetic analysis

Phylogenetic analysis as a tool for taxonomic studies is useful in mosquito systematics when the sequences of the genome and molecular markers of the *An. gambiae* complex are available (Mohanty et al., 2009). The *Anopheles* genus has an impact on human evolution (i.e. sickle cell anaemia as a mode of resistance to malarial protozoa), but little work has been done to understand the evolution and phylogenetic relationships of these mosquitoes (Harbach, 2013). The topologies of the phylogenetic trees we obtained from the COI gene did not allow for a reasonable classification of *An. gambiae* s.s., *An. coluzzii* and *An. arabiensis*, as compared to other molecular techniques using the IGS of rDNA. Phylogenetic relationships between these sub-species are not fully delineated, especially when based on morphology or the

### Table 3

<table>
<thead>
<tr>
<th>Population code (Number of DNA Sequences)</th>
<th>Number of polymorphic (segregating) sites (S)</th>
<th>Number of haplotypes (Hd)</th>
<th>Haplotypes diversity</th>
<th>Average number of differences (K)</th>
<th>Nucleotide diversity (π)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG (29)</td>
<td>34</td>
<td>21</td>
<td>0.96305</td>
<td>4.24384</td>
<td>0.00620</td>
</tr>
<tr>
<td>OH (29)</td>
<td>21</td>
<td>18</td>
<td>0.95567</td>
<td>3.51232</td>
<td>0.00513</td>
</tr>
<tr>
<td>SG (25)</td>
<td>28</td>
<td>16</td>
<td>0.92667</td>
<td>3.17333</td>
<td>0.00464</td>
</tr>
<tr>
<td>KB (31)</td>
<td>58</td>
<td>19</td>
<td>0.94194</td>
<td>5.88602</td>
<td>0.00861</td>
</tr>
<tr>
<td>HD (28)</td>
<td>28</td>
<td>24</td>
<td>0.98942</td>
<td>3.82804</td>
<td>0.00560</td>
</tr>
<tr>
<td>BB (18)</td>
<td>22</td>
<td>13</td>
<td>0.90196</td>
<td>3.61438</td>
<td>0.00528</td>
</tr>
<tr>
<td>Total (151)</td>
<td>102</td>
<td>74</td>
<td>0.95252</td>
<td>4.13310</td>
<td>0.00604</td>
</tr>
</tbody>
</table>

Each sample collection site had a unique population code: Kongoussi (KG); Ouahigouya (OH); Sanguï (SG); Kombissiri (KB); Houndé (HD) and Bobo (BB). a: collection site cotton-growing and b: collection site cotton-free.

### Table 4

<table>
<thead>
<tr>
<th>Population code</th>
<th>Fu and Li's D*</th>
<th>Fu and Li's F*</th>
<th>Fu's Fs</th>
<th>Tajima's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>-2.73374&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-2.95469&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-14.297</td>
<td>-2.18264&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
<tr>
<td>OH</td>
<td>-0.80365</td>
<td>-1.13087</td>
<td>-9.725</td>
<td>-1.30076</td>
</tr>
<tr>
<td>SG</td>
<td>-2.89057&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-3.04071&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-10.123</td>
<td>-1.93802&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
<tr>
<td>KB</td>
<td>-3.02537&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-3.27671&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-8.765</td>
<td>-2.26179&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>HD</td>
<td>-2.44943&lt;sup&gt;***&lt;/sup&gt;</td>
<td>-2.63967&lt;sup&gt;***&lt;/sup&gt;</td>
<td>-22.564</td>
<td>-1.80675&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
<tr>
<td>BB</td>
<td>-2.68499&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-3.03067&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-7.697</td>
<td>-1.98661&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total</td>
<td>-4.42854&lt;sup&gt;***&lt;/sup&gt;</td>
<td>-4.27649&lt;sup&gt;***&lt;/sup&gt;</td>
<td>-91.382</td>
<td>-2.52979&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Statistical significance: *, P < .05; **, P < .02; ***, P < .01. a: collection site cotton-growing and b: collection site cotton-free.

### Table 5

<table>
<thead>
<tr>
<th>Population code</th>
<th>Average number of nucleotide differences (Kxy) (above the diagonal)</th>
<th>Nucleotide divergence (Dxy) (below the diagonal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>3.86801&lt;sup&gt;***&lt;/sup&gt;</td>
<td>3.73379&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
<tr>
<td>OH</td>
<td>0.00565&lt;sup&gt;***&lt;/sup&gt;</td>
<td>3.4331&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
<tr>
<td>SG</td>
<td>0.00546&lt;sup&gt;***&lt;/sup&gt;</td>
<td>0.00502&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
<tr>
<td>KB</td>
<td>0.00737&lt;sup&gt;***&lt;/sup&gt;</td>
<td>0.00689&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
<tr>
<td>HD</td>
<td>0.00591&lt;sup&gt;***&lt;/sup&gt;</td>
<td>0.00514&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
<tr>
<td>BB</td>
<td>0.00572&lt;sup&gt;***&lt;/sup&gt;</td>
<td>0.00514&lt;sup&gt;***&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Statistical significance: *, P < .05; **, P < .02; ***, P < .01. a: collection site cotton-growing and b: collection site cotton-free.

mitogenome (Harbach, 2004; Hanemaaijer et al., 2019). The considerable sequence similarity, common genetic elements and common ancestral molecular polymorphisms of the *An. gambiae* complex make it difficult to determine the direction of its evolution (Besansky et al., 1994; Mohanty et al., 2009). The mitochondrial gene COI is commonly sequenced for *Anopheles* phylogeny (Ratnasingham and Hebert, 2007; Mohanty et al., 2009). In Nigeria, Matthews et al. (2007) used the COI gene to assess *Anopheles* population genetic structures and found a trend towards decreasing genetic diversity in *An. arabiensis* from the northern savannah to the southern rainforest (Matthews et al., 2007). Phylogenetic analysis is used as a tool for dynamic population studies because the genome sequences and molecular markers of the *An. gambiae* complex are now available (Mohanty et al., 2009). Therefore, gene flows and population structures are built based on allelic distribution and dispersal of genes of interest, such as the mitochondrial gene COI and the internal transcribed spacer 2 (ITS2) rDNA region (Sum et al., 2014; Norris and Norris, 2015). Their spatial structure analysis is therefore necessary to ensure the success of selective vector
control campaigns and to predict epidemiological risks (Stevenson and Norris, 2017).

The COI is a mitochondrial protein-coding gene that was used to infer the phylogeny of Anopheles (Sallum et al., 2002). This gene encompasses the barcode region that was used for identifying sibling and cryptic species of the genus Anopheles (Wang et al., 2017; Mohanty et al., 2009). However, the phylogenetic relationships among An. gambiae s.s., An. coluzzii and An. arabiensis are not fully delineated although their mitochondrial DNA (mtDNA) genome sequences are available (Hanemaaijer et al., 2019). In fact, An. gambiae s.s., An. coluzzii and An. arabiensis are a monophyletic taxon (Benson and Falade, 2013; Kamali et al., 2014), even though the subdivision of An. coluzzii into distinct West, Central, and Southern African genetic clusters suggests restrictions to gene flow (Pinto et al., 2013).

4.3. Population genetic structure

Geographical distance between populations may lead to genetic differentiation; therefore, we considered each climatic region or cotton-growing area as a basic reproductive unit. We observed non-significant intraspecific nucleotide variation and nucleotide diversity (\(\pi\)) in An. gambiae complex populations for the mitochondrial (COI) gene sequences. Our Fst values varied from \(-0.01366\) to \(0.0263\) across all populations; Fst values close to zero show a free gene flow and the absence of differentiation in the An. gambiae complex populations. Furthermore, our neutrality indices calculated using Tajima’s D, Fu and Li’s D*; Fu and Li’s F*, Fu’s Fs tests were significantly negative in all populations except Ouahigouya. A significantly negative value (\(P < .05\)) of D and Fs is expected in cases of population expansion, purifying selection or selective scanning (Tajima, 1983).

AMOVA indicated that a high proportion of the total genetic variance was due to variation within populations (99.27%), but the value was non-significant. These results indicate that the distribution of genetic variance between the sampling sites of each climate region did not take significant genetic differentiation between Anopheles gambiae populations into account (Sum et al., 2014). These observations suggest interspecies gene flow among An. coluzzii, An. gambiae s.s. and An. arabiensis populations. Further gene analysis by Fontaine et al. (2016) showed that An. gambiae s.s. was more closely related to An. arabiensis.

There was no isolation by distance, and the haplotype network showed no distinct genetic structure. Although the locations sampled in each climatic region were separated by a maximum of 90 km, these results reveal that there are no barriers to gene flow for mosquitoes from Burkina Faso. This high level of gene flow can be explained either by the wide flight range of the mosquitoes (up to seven km in a lifetime) or by the extensive human-associated migration throughout Burkina Faso (Donnelly et al., 2002; Chen et al., 2004). Our results are similar to those found in Mali, where interspecies gene flow was low among villages (Taylor et al., 2001), and they are consistent with the majority of previous studies on An. gambiae in similar regions of Africa (Lehmann et al., 2003; Prugnolle et al., 2008). In western Kenya, low genetic differentiation was found between island and mainland populations.

The authors explained that wind or human-assisted dispersal probably caused a large extent of the gene flow between the populations (Chen et al., 2004).

In the Nyanza province of Kenya, An. gambiae populations have been found only very slightly differentiated genetically across collection sites (two populations with a maximum distance of about 68 km) (Prugnolle et al., 2008). In a malaria endemic region of southern Tanzania, high levels of genetic differentiation and genetic divergence among An. arabiensis populations have been observed, as opposed to An. gambiae s.s. which displayed no differentiation. However, it appeared that genetic differentiation was not attributed to physical barriers or distance but possibly to ecological diversification (Ng’Habi et al., 2011). In Cameroon (Central Africa), both An. coluzzii and An. gambiae s.s. have been found to occur in sympathy, with low geographic differentiation despite a minimum geographical distance of 35 km between collection sites (Wondji et al., 2002). All these authors worked on microsatellites, while we worked on the COI gene sequence (Wondji et al., 2002; Prugnolle et al., 2008; Borja-Cacho and Matthews, 2010; Ng’Habi et al., 2011). Furthermore, these previous studies showed low genetic differentiation between some high-transmission sites in Western and Eastern Africa although they are separated by thousands of kilometres with a large local population of An. gambiae in these areas (Prugnolle et al., 2008). Reidenbach et al. (2012) concluded that An. coluzzii, An. gambiae s.s. and An. arabiensis are evolving collectively on independent evolutionary trajectories through the analysis of 400,000 single nucleotide polymorphisms (SNPs) across the genomes of paired population samples from Mali, Burkina Faso and Cameroon (Coetzee et al., 2013). Based on population genomic evidence from whole genome sequencing and SNP genotyping, An. coluzzii and An. gambiae s.s. are cohesive and exclusive taxonomic groups across their shared range. This may contribute substantially to the lack of gene flow (Coetzee et al., 2013). Lastly, many studies on gene flow across Africa suggest that the An. gambiae complex exchanges individuals at a rate sufficient in magnitude to prevent them from diverging genetically. There is no ‘isolation by distance’, hence, no relationship between the levels of genetic divergence (Fst) and geographic distance (Lanzaro and Triplet, 2009). However, Miles et al. (2017) recently sequenced the genomes of 765 specimens of An. gambiae and An. coluzzii sampled from 15 locations across Africa, and identified over 50 million SNPs. Contrary to our study, their data revealed high levels of genetic diversity, a complex population structure, and patterns of gene flow, with evidence of ancient expansions, recent bottlenecks, and local variation in the effective population (Miles et al., 2017).

The limitations of this study are that nuclear markers of phylogeny such as nuclear ribosomal internal transcribed spacer 2 (ITS2) could be used instead of the COI marker to determine the extent of population structure and gene flow. Additionally, the present study concerns only one country (Burkina Faso). It will gain in robustness by extending sampling to several malaria-affected countries or localities in Sub-Saharan Africa.

5. Conclusions

This study was carried out in Burkina Faso to assess the taxonomic
Fig. 3. Haplotype networks (n = 74) of COI from the *Anopheles gambiae* complex in Burkina Faso. The haplotype networks show genetic distance between two sequences. Each coloured circle corresponds to haplotype; the circle size displays the number of the sequence, and the colour displays the origin where the sequence was generated: (A) according to climatic region of Burkina Faso (i.e. Sahelian, Sudanian, Sudan-Sahelian) and (B) according to agriculture (i.e. cotton-growing and cotton-free areas). Each black circle corresponds to missing haplotypes. Each edge between two haplotypes corresponds to single changes between sequences, which could be increased according to the number of straps on the edge.

Table 7

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Df</th>
<th>Sum of squares</th>
<th>Variance components</th>
<th>Percentage of variation</th>
<th>F-statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>2</td>
<td>8,092</td>
<td>0.004 Va</td>
<td>0.140</td>
<td>F_C = 0.001</td>
<td>0.34</td>
</tr>
<tr>
<td>Among populations within groups</td>
<td>3</td>
<td>11,408</td>
<td>0.019 Vb</td>
<td>0.590</td>
<td>F_S = 0.006</td>
<td>0.12</td>
</tr>
<tr>
<td>Within populations</td>
<td>145</td>
<td>481,096</td>
<td>3.317 Vc</td>
<td>99.270</td>
<td>F_T = 0.007</td>
<td>0.26</td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>500,596</td>
<td>3.342</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FCT, FSC and FST are the F-statistics. F_C denotes genetic differentiation among groups; F_S denotes genetic differentiation among populations within groups; F_T denotes genetic differentiation between populations among groups. Groups (three) refer to collected localities: Kongoussi (KG) and Ouahigouya (OH); Sangué (SG) and Kombissiri (KB); Houndé (HD) and Bobo (BB). Va, Vb, Vc are the associate variance components.

and genetic structure of the *Anopheles gambiae* complex, a major vector of malaria. Our genomic approaches revealed the distribution of *An. gambiae* s.s., *An. coluzzii* and *An. arabiensis* populations according to selected agro-climatic areas. Using the COI gene, no genetic differentiation was evidenced in the *An. gambiae* complex in Burkina Faso. However, this scientific contribution can not only be used as a basis for in-depth genetic diversity studies of the *An. gambiae* complex in Burkina Faso, but also extended to neighbouring countries to further refine the model and control of malaria transmission.

Availability of data and materials

Data generated or analysed during this study is included in this published article and its supplementary information files. All datasets are available from the corresponding author on reasonable request.

Authors' contributions

AAZ, AB and FF conceived and designed the study. AAZ, AS, ZS and AB supervised the data and sample collection. AAZ performed practical work and wrote the manuscript. AAZ, GN, AB and FF analysed and interpreted the results. GN performed haplotype networks analysis and phylogenetic reconstruction. AB, GN and FF revised the manuscript. All authors were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Ethics statement and consent to participate

This study was approved by the National Health Ethic Committee in Burkina Faso (reference number 2017-9-143 on 12 September 2017). For mosquito collection in residential areas, written and informed consent was obtained from homeowners in each location.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgements

We are sincerely grateful to all the inhabitants of the mosquito collection villages for participating in this study. We acknowledge Nicolas Poncelet who helped in PCR analysis. We also thank the mosquito collectors (Martin Tientiga et al.) and Mr. Dramé, who dissected the mosquitoes, for their valuable contribution to the successful completion of the study. We also acknowledge Dr. Sylvie Cornelline from MiVEGEC (Maladies Infectieuses et Vecteurs, Ecologie, Génétique, Évolution et Contrôle) and Institut de Recherche pour le Développement (IRD) for graciously giving us the KIMUSU strain (insecticide-sensitive) of *Anopheles gambiae* as a positive control in PCR in our study. Finally, we sincerely thank Prof. Kevin J. Emerson for his helpful suggestions and also the second reviewer.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104261.

References


Crawford, J.E., Lazzaro, B.P., 2010. The demographic histories of the M and S molecular


Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses - a re-


