Extensive genetic diversity among populations of the malaria mosquito *Anopheles moucheti* revealed by population genomics

Caroline Fouet\(^1\)*, Colince Kamdem\(^1\), Stephanie Gamez\(^1\), Bradley J. White\(^{1,2}\)*

\(^1\)Department of Entomology, University of California, Riverside, CA 92521

\(^2\)Center for Disease Vector Research, Institute for Integrative Genome Biology, University of California, Riverside, CA 92521

*Corresponding authors: caroline.fouet@ucr.edu; bwhite@ucr.edu

**Key words:** *Anopheles moucheti*, population genomics, RADseq, *de novo* assembly
Abstract

Background: Recent intensive efforts to control malaria in African countries expose vector populations to additional adaptive challenges. The main malaria mosquitoes of the continent display an array of adaptive strategies to cope with such challenges. The development of genomic resources will empower genetic studies that are crucial to understand the evolutionary history and adaptive potential of these vectors.

Methodology/Principal findings: Here we constructed double-digest Restriction Associated DNA (ddRAD) libraries and generated 6461 Single Nucleotide Polymorphisms (SNPs) that we used to explore the population structure and demographic history of wild-caught Anopheles moucheti from Cameroon. The genome-wide distribution of allelic frequencies among sampled populations best fitted that of an old population at equilibrium, characterized by a weak genetic structure and extensive genetic diversity, presumably due to a large long term effective population size. In contrast to other important African malaria vectors, polymorphic chromosomal inversions play little role in the genome architecture and evolutionary adaptation of An. moucheti.

Conclusions/Significance: Our study provides the first investigation of the genetic structure and diversity in An. moucheti at the genomic scale. Despite a weak genetic structure and absence of adaptive divergence, the adaptive potential of this mosquito remains significant owing to a great diversity and standing genetic variation that can be used to face current vector control measures and other rapid anthropogenic and environmental changes.
Background

*An. moucheti sensu lato* is a group of three related mosquito species (*An. moucheti moucheti*, *An. moucheti nigeriensis*, and *An. moucheti bervoetsi*) distributed across forested areas in Sub-Saharan Africa and distinguishable from each other by slight morphological differences [1]. The nominal species *An. moucheti moucheti* (hereafter *An. moucheti*) is the most widespread and one of the leading malaria vectors in the equatorial region [2–4]. The species breeds in slow moving streams and rivers close to villages where it contributes to high malaria transmission and often outcompete other main malaria vectors species. Despite this epidemiological significance, the evolutionary history and the evolvability of this mosquito remain understudied. Early investigations of the genetic structure based on allozymes and microsatellites detected no significant differentiation among *An. moucheti* populations over substantial geographic areas. This finding was particularly intriguing given the widespread distribution of *An. moucheti* in the evergreen forest areas of Central Africa. Moreover, African anopheline mosquito populations are increasingly exposed to intense selective pressure associated with the use of insecticides and insecticide-impregnated materials associated with recent expansive malaria control campaigns [5]. Such pressures are strong driving forces that often contribute to rapid diversification of vector populations at the scales of a few decades [6–8]. As a result, a more detailed characterization of the genomic architecture of *An. moucheti* will be crucial to: (1) understand the genetic bases allowing populations of this species to adapt to such a wide geographic area and (2)
predict the evolutionary potential of this vector in face of selective constraints imposed by rapidly changing environments.

Thanks to recent progresses in sequencing technology, high-resolution sequence information can be generated for virtually any living organism. However, while significant progresses have been made in genomic studies of other African anopheline mosquitoes [9], An. moucheti has lagged behind mainly because of the absence of genomic resources. So far, this species crucially lacks a reference genome assembly, a physical or linkage map and laboratory strains that are essential to generate high-quality sequencing information and to enable robust interpretations of natural polymorphisms.

To start filling this gap, we have conducted the first screening of genetic variation across the genome of wild populations of An. moucheti. We have performed a high-throughput sequencing of reduced representation libraries in 87 wild-caught individuals from Cameroon and used a de novo assembly to identify thousands of RAD loci scattered throughout the genome. Using high-quality Single Nucleotide Polymorphisms (SNPs) identified within these loci, we have investigated the genetic relatedness and population history of our samples. We found that populations of An. moucheti are characterized by extensive gene flow and a great genetic diversity. This vector appears particularly adapted to challenge the selective pressures imposed by vector controls and rapid human-induced environmental modifications.
Methods

Mosquito sampling and sequencing

This study included two An. moucheti populations from the Cameroonian equatorial rainforest. A total of 98 mosquitoes (97 adults and 1 larva) were collected in August and November 2013 Olama and Nyabessan, respectively (Table 1). The two locations are separated by ~200 km (Fig. 1A) and are crossed respectively by the Nyong and the Ntem rivers that provide the breeding sites for An. moucheti larvae. Specimens were identified as An. moucheti moucheti using morphological identification keys and a diagnostic PCR [1,10,11]. We extracted genomic DNA of specimens using either the DNeasy Blood and Tissue kit (Qiagen) or the Zymo Research MinPrep kit. We used 10ul (~50ng) of genomic DNA to prepare double-digest Restriction-site Associated DNA libraries following a modified protocol of Peterson et al.,2012 [12]. MluC1 and NlaIII restriction enzymes were used to digest DNA of individual mosquitoes yielding RAD-tags of different sizes to which short unique DNA sequences (barcodes) were ligated to enable the identification of reads belonging to each specimen. The digestion products were purified and pooled. DNA fragments with size ~400bp were selected and amplified via PCR. The distribution of fragment sizes was checked on a BioAnalyzer (Agilent Technologies, Inc., USA) before sequencing. The sequencing was performed on an Illumina HiSeq2000 platform (Illumina Inc., USA) (Genomic Core Facility, University of California, Riverside) to yield single-end reads of 101 bp.

De novo assembly and SNP discovery
We used the bioinformatics pipeline Stacks v1.35 [13] to process Illumina short reads. The program `process_radtags` was first used to sort the reads according to the barcodes and to trim all reads to 96-bp in length by removing index and barcode sequences from the ends of the reads. Reads with ambiguous barcodes, those that did not contain the NlaIII recognition site and those with low-quality scores (average Phred score < 33) were excluded. The program `ustacks` was then utilized for a de novo assembly enabling the identification of consensus RAD loci in each individual in our populations. We allowed a maximum of 2 nucleotide mismatches between stacks (M parameter in `ustacks`) and we required a minimum of three reads to create a stack (m parameter in `ustacks`). Using the `cstacks` program, a catalogue of loci was built to synchronize variations across all individuals in our populations. Finally, we utilized the `populations` program to calculate population genetic parameters and output SNPs in different formats. To avoid bias associated with less informative SNPs or possible false positive SNPs (due to sequencing or pipeline errors), only RAD loci scored in at least 70-80% of individuals were retained for further analyses.

**Population genomic analyses**

SNP files outputted by the `populations` program were used to assess the population genetic structure with a Principal Component Analysis (PCA) and a Neighbor-joining tree analysis using respectively the R packages `adegenet` and `ape` [14,15]. We also explored patterns of ancestry and admixture in *An. moucheti* individuals in ADMIXTURE v1.23 [16] with 10-fold cross-validation for k assumed ancestral populations (k= 1 through 6). The optimal number of clusters was confirmed using
the Discriminant Analysis of Principal Component (DAPC) method which explores the number of genetically distinct groups by running a k-means clustering sequentially with increasing numbers and by comparing different clustering solutions using Bayesian Information Criterion (BIC) [14]. We examined the population genetic diversity, conformity to Hardy-Weinberg equilibrium and demographic background using several statistics calculated with the populations program. Precisely, to assess the global genetic diversity per population, we calculated the overall nucleotide diversity ($\pi$) and the frequency of polymorphic sites within population. To make inferences on the demographic history and to test for departures from Hardy–Weinberg equilibrium, we utilized the Allele frequency spectrum and the Wright’s inbreeding coefficient ($F_{st}$). To quantify the geographic and genetic differentiation between allopatric populations, we estimated the genome-wide average $F_{ST}$ [17] on 2000 randomly selected SNPs in Genodive v1.06 [18]. We also conducted an hierarchical Analysis of Molecular Variance (AMOVA) [19] on the same SNP set to quantify the effects of the geographic origin on the genetic variance among individuals. The statistical significance of $F_{ST}$ and AMOVA was assessed with 10000 permutations. Finally, to examine the genomic architecture divergence, we inspected the genome-wide distribution of locus-specific estimates of $F_{ST}$.

Identification of polymorphic chromosomal inversions

The neutral recombination rate is notoriously reduced in genomic regions bearing chromosomal inversions, which results in elevated linkage disequilibrium (LD) in those regions relative to the rest of the genome. Thus, assessing genome-wide
patterns of LD can reveal clusters of highly correlated SNPs (LD blocks) corresponding potentially to chromosomal inversions. The R package LDna [20] allow the examination of the distinct LD network clusters within the genome of nonmodel species without the need of a linkage map or reference genome. We calculated LD, estimated as the $r^2$ correlation coefficient between all pairs of SNPs, in PLINK v1.09 [21]. To avoid spurious LD due to the strong correlation between SNPs located on the same RAD locus, we randomly selected only one SNP within each RAD locus resulting in a dataset of 1056 variants containing less than 10% missing data. LDna was then used to identify clusters of highly correlated SNPs that were interpreted in light of the population demographic history, the population genetic structure and previous cytological studies of *An. moucheti*. 
Results

De novo assembly

In total, 518,218 unique 96-bp RAD loci were identified from de novo assembly of reads in 98 individuals. We retained 946 loci that were present in all sampled populations and in at least 75% of individuals in every population, and we identified 3027 high-quality biallelic SNPs from these loci.

Population genetic structure

First, we tested for the presence of cryptic genetic subdivision within An. moucheti with PCA, NJ trees and the ADMIXTURE ancestry model. A NJ tree constructed from a matrix of Euclidian distance using allele frequencies at 3027 genome-wide SNPs showed a putative subdivision of An. moucheti populations in two genetic clusters (Fig. S1). The first three axes of PCA also revealed a number of outlier individuals separated from a main cluster (Fig. S1). However, when we ranked our sequenced individuals based on the number of sequencing reads, we noticed that one of the putative genetic clusters corresponded to a group of individuals having the lowest sequencing coverage. We excluded all these individuals and reduced our dataset to 78 individuals. We conducted a de novo assembly and analyzed the relationship between the 78 remaining individuals at 6461 genome-wide SNPs using PCA, NJ trees and ADMIXTURE. Both the k-means clustering (DAPC) and the variation of the cross-validation error as a function of the number of ancestral populations in ADMIXTURE revealed that the polymorphism of An. moucheti resulted from only one ancestral population (k = 1) (Fig. 1E,F). PCA and NJ depicted a homogeneous cluster comprising all 78 individuals providing additional evidence of the lack of genetic or
geographic structuring among populations. Unsurprisingly, the overall $F_{ST}$ was remarkably low between populations from the two sampling locations Olama and Nyabessan ($F_{ST} = 0.008, p < 0.005$). Similarly, the distribution of $F_{ST}$ values across the 6461 SNPs showed a large dominance of very low $F_{ST}$ values (Fig. 2). The highest per locus $F_{ST}$ was only 0.126, while 5006 of the 6461 loci revealed $F_{ST}$ near zero. The modest geographic differentiation was also well illustrated by a hierarchical AMOVA, which showed that the genetic variance was explained essentially by within-individual variations (99.7%). Finally, we found very low overall Wright's inbreeding coefficient ($F_{IS} = 0.0014$ in Nyabessan and $F_{IS} = 0.0025$ in Olama) (Table 2) suggesting that allelic frequencies within both populations were in accordance with proportions expected under the Hardy-Weinberg equilibrium.

**Genetic diversity and demographic history**

The estimates of the overall nucleotide diversity ($\pi = 0.0020$ and $\pi = 0.0016$, respectively, in Olama and Nyabessan) (Table 2) were within the range of average values found in other African *Anopheles* species using RAD-seq approaches [8,22–24]. Notorious demographic expansions have been described in natural populations of this insect clade [25], and the values of $\pi$ observed in *An. moucheti* likely reflect the level genetic diversity of a population with large effective size. The great genetic diversity of *An. moucheti* was also illustrated by the percentage of polymorphic sites. Of the 6461 variant sites, 89.60% were polymorphic in Olama and 34.82% in Nyabessan (Table 2). The difference observed between the two locations can be related to the sample size ($n = 19$ in Nyabessan and $n = 59$ in Olama) or to demographic particularities that persists between the two geographic sites despite a
massive gene flow. To infer the demographic history of *An. moucheti*, we examined the Allele Frequency Spectrum (AFS), summarized as the distribution of the major allele in one population. This approach represented a substitute to model-based methods that provide powerful examinations of the history of genetic diversity by modeling the AFS at genome-wide SNP variants, but that couldn’t be implemented here due to the lack of reference genome assembly. The frequency distribution of the major allele \( p \) (Fig. 3) indicates that the majority of loci that are polymorphic between Olama and Nyabessan are fixed within each population as shown by the predominance of SNPs at frequencies equal to 1. Ranges of allele frequencies are similar in Olama and Nyabessan (between 0.47 and 1 in Olama and between 0.34 and 1 in Nyabessan). These frequency ranges are expected for old populations at equilibrium capable of accumulating high amount of genetic diversity.

**Polymorphic chromosomal inversions**

LD analyses at 1056 highly filtered SNPs revealed a globally low LD in the *An. moucheti* genome (average genome-wide \( r^2 = 0.0149 \)) as expected in highly polymorphic populations with large effective size. We used LDna to cluster the LD values and to identify Single Outlier Clusters (SOC) that can be associated with distinct or multiple evolutionary phenomena in the *An. moucheti* history. Results showed the presence of two independent LD blocks in our samples (Fig. 4). Cytogenetic analyses have identified three polymorphic chromosomal inversions in Cameroon samples [26] and the two LD clusters could well correspond to linked SNPs within inversions. Studies on *Anopheles baimaii* [20] have shown that when SOCs are generated by polymorphic inversions, SNPs within the SOCs can clearly
separate the three expected karyotypes (inverted homozygotes, heterozygotes and uninvlected homozygotes). We conducted downstream analyses with a PCA using respectively 30 and 34 SNPs identified within the two SOCs. As shown on Fig, although individuals were spread along three PCA axes, no distinct cluster could be identified for the first SOC. likewise, except three outlier individuals, the second SOC formed a single genetic cluster (Fig. 4B,C). The two LD clusters were apparently not associated with polymorphic inversions, but caution must be used when interpreting this result as evidence of absence of inversions in our samples because other LDna analyses conducted on An. funestus and An. gambiae which contain tens of polymorphic inversions did not systematically identified the segregating karyotypes of key adaptive inversions (results not shown). However, based on our findings and previous cytogenetic studies, we can reasonably think that inversion polymorphisms are less important in the genome architecture An. moucheti compared with that of other important African malaria vectors such as An. gambiae and An. funestus whose populations often segregate along clines of multiple polymorphic inversions in nature.
**Discussion**

We have analyzed genome-wide polymorphism and characterized some of the baseline population genomic parameters in *An. moucheti*, an important malaria vector in the African rainforest. We found very little differentiation among our samples, with most of the genetic variation distributed within individuals. Although a more substantial sampling will be necessary to fully appreciate the population genetic structure of this species, our sampling likely reflects the current dynamic of *An. moucheti* populations in Cameroon. It is worth mentioning that we have surveyed a total of 28 locations across the country, some of which were known from several past surveys to harbor *An. moucheti* populations [1–4,27,28], but we confirmed the presence of the species in only 2 villages. Extant populations of *An. moucheti* are distributed in patches of favorable habitats along river networks where larval populations breed. Our results indicate that despite this apparent fragmentation, connectivity and gene flow are high among population aggregates. The weak population genetic structure of *An. moucheti* observed with genome-wide markers corroborated results obtained with microsatellites and allozymes [2,3]. A survey of eight microsatellite loci revealed that the highest $F_{ST}$ among Cameroonian populations was as low as 0.003. Nevertheless, a substantial differentiation was found between samples from different countries consistent with an isolation-by-distance model [3]. It is clear that a deep sequencing of continental populations is necessary to further clarify the status of these putative subpopulations that seem to be endemic to different countries. But, samples collected at lower spatial scales like ours are also relevant as they can allow robust inferences about ongoing selective
processes. We have found that signatures of selection are very rare in the genome of *An. moucheti* populations from the Cameroonian rainforest. Populations remain largely undifferentiated throughout the genome, with $F_{ST}$ values near zero across the vast majority of segregating sites. Interestingly, polymorphic chromosomal inversions that are the prime source of adaptive polymorphisms in many other important malaria vectors seem to be less vital in the evolution of *An. moucheti*. The characterization of chromosomal inversions with cytogenetic methods can be laborious, particularly with regard to the identification of new inversions [26,29]. We have implemented a recently designed method that used Next Generation Sequencing and LD estimates to identify paracentric inversions. Although the method holds great promise in identifying inversions in genomes of nonmodel species like *An. moucheti*, care should be taken when interpreting the results. So far, only three paracentric polymorphic inversions have been discovered in *An. moucheti* thanks to cytogenetics [26]. The ecological, behavioral or functional roles of these inversion polymorphisms remain unknown. In keeping with cytogenetic observations, LD analyses revealed the presence of only a few LD clusters that could not be unambiguously associated with inversions. On the other hand, the low overall LD observed across the genome reflected the significant genetic polymorphism that seems to prevail in *An. moucheti* populations. This polymorphism translates into exceptional levels of overall genetic diversity and very high percentage of polymorphic sites that are in the range of values found in other mosquito species undergoing a significant demographic expansion [8,24,25]. One of our main goals was to know if the genetic diversity of *An. moucheti* can provide cues about its
extensive spatial distribution and can help to predict its environmental resilience. In principle, evolutionary responses of species to human-induced or natural changes rely largely on available heritable variation, which reflects the evolutionary potential and adaptability to novel environments [30]. Therefore, the screening of genome-wide variation is supposed to be a sensible approach that may provide a generalized measure of evolutionary potential in species like *An. moucheti* for which direct ecological, evolutionary or functional tests are impossible [31]. Our population genomic analyses have depicted *An. moucheti* as a species with a great genetic diversity and consequently a sustainable long-term adaptive resilience.

Implications of our findings in malaria epidemiology and control can be very significant. First, *An. moucheti* is essentially endophilic and is particularly sensitive to the principal measures currently employed to control malaria in Sub-Saharan Africa such as the massive use of Insecticide Treated Nets (ITNs) and Indoor Residual Insecticide Spraying (IRS). For example, estimates of population effective size in one village in Equatorial Guinea indicated that both mass distribution of ITNs and IRS campaigns resulted in a decline of approximately 55% of *An. moucheti* [32].

However, the great genetic diversity and the massive gene flow we observed within populations could easily enable *An. moucheti* to challenge population declines and recover from shallow bottlenecks. Moreover, most insecticide resistance mechanisms found in insects exploit standing genetic variation that allow to rapidly respond to the evolutionary challenge by increasing the frequency of existing variations rather than that of *de novo* mutations [33]. As a result, despite the current sensitivity of *An. moucheti* to common insecticides, the significant amount of
standing genetic variation provides the species with a great potential to challenge insecticides and other types of human-induced stress.

Recent advances in sequencing allow sensitive genomic data to be generated for virtually any species [30]. However, the most important information we can obtain from population resequencing approaches often depends on the availability and the quality of genomic resources such as a well-annotated reference genome. Nevertheless, the reduced genome sequencing strategy (RAD-seq) offers a cost-effective strategy that can been used to effectively study the genetic variation in a broad range of species from yeast to plants, insects, etc., in the absence of reference genome. We have extended this approach to the study of the genetic structure of an understudied epidemiologically important mosquito species. We have provided both significant baseline population genomic data and the methodological validation of one approach that should motivate further studies on this species and other understudied anopheline mosquitoes lacking genomic resources.
Acknowledgements

Funding for this project was provided by the University of California Riverside and NIH grants 1R01AI113248 and 1R21AI115271 to BJW. We thank populations and authorities of the locations surveyed for their kind collaboration.
References


Author contributions

Conceived and designed the experiments: CF CK BJW. Performed the experiments: CF CK SG BJW. Analyzed the data: CF CK BJW. Wrote the paper: CF CK BJW.

Competing interests

The authors declare that they have no competing interests.
Tables

Table 1: Information on *An. moucheti* samples included in this study.

<table>
<thead>
<tr>
<th>Sampling locations</th>
<th>Geographic coordinates</th>
<th>Sampling methods</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HLC-OUT</td>
<td>HLC-IN</td>
</tr>
<tr>
<td>Nyabessan</td>
<td>2°24'00&quot;N, 10°24'00&quot;E</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>Olama</td>
<td>3°26'00&quot;N, 11°17'00&quot;E</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HLC-OUT, human landing catches performed outdoor; HLC-IN, human landing catches performed indoor; LC, larval collection

Table 2: Population genomic parameters based on 6461 variant sites reflecting the genetic diversity and conformity to Hardy-Weingberg equilibrium.

<table>
<thead>
<tr>
<th>Variant positions</th>
<th>Sites</th>
<th>Observed heterozygosity</th>
<th>$F_{IS}$</th>
<th>$\pi$</th>
<th>% polymorphic sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olama</td>
<td>6461</td>
<td>0.0445</td>
<td>0.0631</td>
<td>0.0505</td>
<td>89.60</td>
</tr>
<tr>
<td>Nyabessan</td>
<td>6461</td>
<td>0.0334</td>
<td>0.0372</td>
<td>0.0402</td>
<td>34.82</td>
</tr>
<tr>
<td>All positions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olama</td>
<td>165 975</td>
<td>0.0017</td>
<td>0.0025</td>
<td>0.0020</td>
<td>3.49</td>
</tr>
<tr>
<td>Nyabessan</td>
<td>165 975</td>
<td>0.0013</td>
<td>0.0014</td>
<td>0.0016</td>
<td>1.36</td>
</tr>
</tbody>
</table>
**Figures**

**Figure 1:** Relationship between *An. moucheti* individuals from Olama and Nyabessan. (A) Map of the study site showing both the locations surveyed (small black dots) and the two villages (large red and blue squares) where *An. moucheti* samples were collected. (B) and (C) Absence of genetic structure within populations illustrated by neighbor-joining and PCA. The percentage of variance explained by each PCA axis is indicated. (D) and (E) Plots of the ADMIXTURE cross-validation error and the Bayesian Information Criterion (BIC) (DAPC) as a function of the number of genetic clusters indicating that k = 1. The lowest BIC and CV error indicate the suggested number of clusters.
**Figure 2:** Frequency distribution of $F_{ST}$ between Olama and Nyabessan across 6461 SNP loci and plot of these $F_{ST}$ values along arbitrary positions in the genome.
**Figure 3:** Allele Frequency Spectrum for 6461 SNP loci in Nyabessan and Olama populations. The x-axis presents the frequency of the major allele and the y-axis the frequency distribution of loci in each class of the major allele frequency.
**Figure 4**: LDna analyses on 1056 SNPs showing the presence of two Single Outlier Clusters (SOCs) of linkage disequilibrium in *An. moucheti*. (A) Graph presenting the results obtained with values of the two parameters: $\varphi$ (which controls when clusters are defined as outliers) and $|E|_{\text{min}}$, the minimum number of edges required for a LD cluster to be considered as an outlier, indicated on top. LD thresholds are shown on the x-axis. (B) and (C) PCA indicating the population genetic structure inferred from SNPs within the two SOCs (red: Olama; blue: Nyabessan).
Supplemental Material

Figure S1: Selection of individuals included in final analyses based on the average per individual sequencing coverage. (A) and (B) Neighbor-joining tree and PCA indicating spurious population structure due to individuals with low sequencing coverage in Olama (red) and Nyabessan (blue).