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

Address: Department of Entomology, University of California Riverside. 900 University Ave. Riverside, CA 92521. Tel: +1 951 827 2626

Email address: caroline.fouet@ucr.edu; bwhite@ucr.edu

Abstract

Malaria vectors are exposed to intense selective pressures due to large-scale intervention programs that are underway in most African countries. One of the current priorities is therefore to clearly assess the adaptive potential of Anopheline populations, which is critical to understand and anticipate the response mosquitoes can elicit against such adaptive challenges. The development of genomic resources that will empower robust examinations of evolutionary changes in all vectors including currently understudied species is an inevitable step toward this goal. 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 samples 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. Estimates of F_{ST} and Linkage Disequilibrium (LD) across SNPs reveal a very low genetic differentiation throughout the genome and the absence of segregating LD blocks among populations, suggesting an overall lack of local adaptation. Our study provides the first investigation of the genetic structure and diversity in *An. moucheti* at the genomic scale. We conclude that, despite a weak genetic structure, this species has the potential to challenge current vector control measures and other rapid anthropogenic and environmental changes thanks to its great genetic diversity.

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

1. Introduction

Despite having a widely acknowledged epidemiological significance, most African malaria mosquitoes are so-called "neglected vectors" because the efforts devoted to their study and control are clearly insufficient. *Anopheles moucheti sensu lato* is one of the best examples. This mosquito vector is a group of three related species (*An. moucheti moucheti*, *An. moucheti nigeriensis*, and *An. moucheti bervoetsi*) distributed across the equatorial forest and distinguishable from each other by slight morphological differences (Kengne et al., 2007). The nominal species of the group, *An. moucheti moucheti* (hereafter *An. moucheti*), is a very efficient and anthropophilic vector especially in rural areas where the highest malaria burden due to *Plasmodium falciparum* infections are recorded (Antonio-Nkondjio et al., 2009, 2008, 2002). In such settings, abundant populations of *An. moucheti* breed year-round in slow moving streams and rivers and often outcompete other main malaria mosquitoes. Despite this epidemiological significance, the evolutionary history and the adaptive potential of this vector remain understudied. Early investigations of the genetic structure based on allozymes and microsatellites showed a significant genetic differentiation among samples from three different countries (Antonio-Nkondjio et al., 2008), but detected little divergence within populations from the same country (Antonio-Nkondjio et al., 2007, 2002). Precisely, very low levels of genetic differentiation were found between populations from Cameroon across eight microsatellite loci, suggesting extensive gene flow at such geographic scales, but detailed studies in other countries are still lacking to fully support this hypothesis. On the other hand, African anopheline populations are

62 increasingly exposed to strong selective pressures associated with insecticide-based
63 malaria control campaigns that have been recently intensified (World Health
64 Organization, 2013). Such pressures represent particularly efficient driving forces
65 that often contribute to the rapid diversification of vector populations in a few
66 decades (Clarkson et al., 2014; Kamdem et al., Unpublished data a; Norris et al.,
67 2015). As a result, a detailed characterization of the genomic architecture of all
68 vectors is important for a critical appraisal of the impacts of malaria control efforts.
69 In this framework, we set out to perform the first genome-wide investigation of
70 natural polymorphism in *An. moucheti*. One of our main goals was to know to what
71 extent assessing the genetic diversity could provide clues about the spatial
72 distribution and help predict the environmental resilience of this species. In
73 principle, evolutionary responses of species to human-induced or natural changes
74 rely largely on available heritable variation, which reflects the evolutionary
75 potential and adaptability to novel environments (Orr and Unckless, 2008).
76 Therefore, the screening of genome-wide variation is supposed to be a sensible
77 approach that may provide a generalized measure of evolutionary potential in
78 species like *An. moucheti* for which direct ecological, evolutionary or functional tests
79 are impossible (Harrisson et al., 2014).
80
81 Thanks to recent progresses in sequencing technology, high-resolution sequence
82 information can be generated for virtually any living organism. These technological
83 advances are extraordinary helpful for non-model species with limited genomic
84 resources like mosquitoes (Ellegren, 2014). However, at the exception of *Anopheles*

85 *gambiae* for which significant genomic studies have been carried out using high-
86 quality sequencing data (Fontaine et al., 2015; Kamdem et al., Unpublished data a;
87 O'Loughlin et al., 2014), the other African malaria vectors have yet to fully benefit
88 from the explosive growth of methods for assessing genetic variation at a fine scale.
89 These neglected vectors face a vicious cycle whereby the lack of basic genomic
90 resources that are critical to generate high-quality sequencing information and to
91 enable robust interpretations of natural polymorphisms greatly contributes to their
92 marginalization. One typical example is *An. moucheti*, which lacks all the vital
93 resources ranging from a laboratory strain, a reference genome assembly, and a
94 physical or linkage map.
95
96 To start filling this gap and to shed some light on the evolutionary history and
97 adaptive potential of this vector, we have performed a high-throughput sequencing
98 of reduced representation libraries in 98 wild-caught individuals from Cameroon
99 and identify thousands of RAD loci scattered throughout the genome. Using high-
100 quality Single Nucleotide Polymorphisms (SNPs) identified within these loci, we
101 have investigated the genetic structure of populations and scan genomes of our
102 samples to detect footprints of local adaptation and natural selection. We found that,
103 in our study zone, populations of *An. moucheti* are characterized by a great genetic
104 diversity and extensive gene flow. We argue that this vector is particularly adapted
105 to challenge the selective pressures imposed by vector controls and rapid
106 environmental modifications.
107

108 **2. Material and methods**

109 **2.1. Mosquito sampling and sequencing**

110 This study included two *An. moucheti* populations from the Cameroonian equatorial
111 forest. A total of 98 mosquitoes (97 adults and 1 larva) were collected in August and
112 November 2013 from Olama and Nyabessan, respectively (Table 1). The two
113 locations are separated by ~200 km (Fig. 1A) and are crossed respectively by the
114 Nyong and the Ntem rivers that provide the breeding sites for *An. moucheti* larvae.
115 Specimens were identified as *An. moucheti moucheti* using morphological
116 identification keys (Gillies and Coetzee, 1987; Gillies and De Meillon, 1968) and a
117 diagnostic PCR, which targets mutations on the ribosomal DNA (Kengne et al.,
118 2007). We extracted genomic DNA using the DNeasy Blood and Tissue kit (Qiagen)
119 for larvae and the Zymo Research MinPrep kit for adult mosquitoes. We used 10µl
120 (~50ng) of genomic DNA to prepare double-digest Restriction-site Associated DNA
121 libraries following a modified protocol of Peterson et al., 2012. *MluC1* and *NlaIII*
122 restriction enzymes were used to digest DNA of individual mosquitoes, yielding
123 RAD-tags of different sizes to which short unique DNA sequences (barcodes and
124 adaptors) were ligated to enable the identification of reads belonging to each
125 specimen. The digestion products were purified and pooled. DNA fragments of
126 around 400bp were selected and amplified via PCR. The distribution of fragment
127 sizes was checked on a BioAnalyzer (Agilent Technologies, Inc., USA) before
128 sequencing. The sequencing was performed on an Illumina HiSeq2000 platform
129 (Illumina Inc., USA) (Genomic Core Facility, University of California, Riverside) to
130 yield single-end reads of 101bp.

131 **2.2 SNP discovery and genotyping**

132 We used the bioinformatics pipeline Stacks v1.35 (Catchen et al., 2013) to process
133 Illumina short reads. The program *process_radtags* was first used to sort the reads
134 according to the barcodes and to trim all reads to 96bp in length by removing index
135 and barcode sequences from the ends of the reads. Reads with ambiguous barcodes,
136 those that did not contain the *NlaIII* recognition site and those with low-quality
137 scores (average Phred score < 33) were excluded. The program *ustacks* was then
138 utilized to perform a *de novo* assembly (i.e., the assembly of reads in "stacks"
139 enabling the creation of consensus RAD loci without prior alignment onto a
140 reference genome sequence) (Catchen et al., 2013, 2011) in each individual in our
141 populations. We allowed a maximum of 2 nucleotide mismatches between stacks (M
142 parameter in *ustacks*) and we required a minimum of three reads to create a stack
143 (m parameter in *ustacks*). Using the *cstacks* program, a catalogue of loci was built to
144 synchronize variations across all individuals in our populations. Finally, we utilized
145 the *populations* program to calculate population genetic parameters and output
146 SNPs in different formats. To avoid bias associated with less informative SNPs or
147 possible false positive SNPs (due to sequencing or pipeline errors), only RAD loci
148 scored in at least 70-75% of individuals were retained for further analyses.

149 **2.3. Population genomic analyses**

150 SNP files outputted by the *populations* program were used to assess the population
151 genetic structure with a Principal Component Analysis (PCA) and a Neighbor-Joining
152 (NJ) tree analysis using respectively the R packages *adegenet* and *ape* (Jombart,
153 2008; Paradis et al., 2004; R Development Core Team, 2008). We also explored

154 patterns of ancestry and admixture among individuals in ADMIXTURE v1.23
155 (Alexander et al., 2009) with 10-fold cross-validation for k assumed ancestral
156 populations (k= 1 through 6). The optimal number of clusters was confirmed using
157 the Discriminant Analysis of Principal Component (DAPC) method, which explores
158 the number of genetically distinct groups by running a k-means clustering
159 sequentially with increasing numbers and by comparing different clustering
160 solutions using Bayesian Information Criterion (BIC) (Jombart, 2008). We examined
161 the population genetic diversity, conformity to Hardy-Weinberg equilibrium and
162 demographic background using several statistics calculated with the *populations*
163 program. Precisely, to assess the global genetic diversity per population, we
164 calculated the overall nucleotide diversity (π) and the frequency of polymorphic
165 sites within population. To make inferences on the demographic history and to test
166 for departures from Hardy–Weinberg equilibrium, we used the allele frequency
167 spectrum and the Wright's inbreeding coefficient (F_{IS}). To quantify the geographic
168 and genetic differentiation between allopatric populations, we estimated the
169 genome-wide average F_{ST} (Weir and Cockerham, 1984) on 2000 randomly selected
170 SNPs in Genodive v1.06 (Meirmans and Van Tienderen, 2004). We also conducted
171 an hierarchical Analysis of Molecular Variance (AMOVA) (Excoffier et al., 1992) on
172 the same SNP set to quantify the effects of the geographic origin on the genetic
173 variance among individuals. The statistical significance of F_{ST} and AMOVA was
174 assessed with 10000 permutations. Finally, to have a detailed picture of the genomic
175 architecture of divergence, we inspected the genome-wide distribution of locus-
176 specific estimates of F_{ST}.

177 **2.4. Identification of segregating polymorphic chromosomal inversions**

178 In structured *Anopheles* populations whose ecological/genetic divergence is due to
179 polymorphic chromosomal inversions, high values of F_{ST} are expected between
180 divergent populations within inversion loci, a pattern consistent with local
181 adaptation of alternative karyotypes (Ayala and Coluzzi, 2005). This is the case for
182 most populations of the main African malaria vectors *An. funestus* and *An. gambiae*,
183 which depict multiple inversion clines in nature (Ayala et al., 2011; Fouet et al.,
184 Unpublished data; Kamdem et al., Unpublished data a; O'Loughlin et al., 2014). In
185 addition to scanning genomes of our individuals to identify outlier values of F_{ST} that
186 are indicators of selection and local adaptation, we also used Linkage Disequilibrium
187 (LD) analysis to search for the presence of LD blocks corresponding to putative
188 inversion polymorphisms. LD (the nonrandom association of alleles at different loci)
189 provides information about past events and is affected by local adaptation and
190 geographical structure, the demographic history, or the magnitude of selection and
191 recombination across the genome (Lewontin and Kojima, 1960). Notably, high LD is
192 expected in regions bearing inversions relative to the rest of the genome because
193 the neutral recombination rate is notoriously reduced within inversions
194 (Kirkpatrick and Barton, 2006). Thus, assessing genome-wide patterns of LD can
195 reveal clusters of strongly correlated SNPs (LD blocks) corresponding potentially to
196 chromosomal inversions. The R package LDna (Kemppainen et al., 2015) allows the
197 examination of the distinct LD network clusters within the genome of non-model
198 species without the need of a linkage map or reference genome. We have calculated
199 LD, estimated as the r^2 correlation coefficient between all pairs of SNPs, in PLINK

200 v1.09 (Purcell et al., 2007). To avoid spurious LD due to the strong correlation
201 between SNPs located on the same RAD locus, we randomly selected only one SNP
202 within each RAD locus resulting in a dataset of 2569 variants containing less than
203 15% missing data. LDna was then used to identify LD blocks whose population
204 genetic structure was examined with a PCA.
205

206 **3. Results**

207 **3.1. *De novo* assembly**

208 In total, 518,218 unique 96-bp RAD loci were identified from *de novo* assembly of
209 reads in 98 individuals. We retained 946 loci that were present in all sampled
210 populations and in at least 75% of individuals in every population, and we identified
211 3027 high-quality biallelic SNPs from these loci.

212 **3.2. Population genetic structure**

213 First, we tested for the presence of cryptic genetic subdivision within *An. moucheti*
214 with PCA, NJ trees and the ADMIXTURE ancestry model. A NJ tree constructed from
215 a matrix of Euclidian distance using allele frequencies at 3027 genome-wide SNPs
216 showed a putative subdivision of *An. moucheti* populations in two genetic clusters
217 (Fig. S1A). The first three axes of PCA also revealed a number of outlier individuals
218 separated from a main cluster (Fig. S1B). However, when we ranked our sequenced
219 individuals based on the number of sequencing reads, we noticed that one of the
220 putative genetic clusters corresponded to a group of individuals having the lowest
221 sequencing coverage (Fig. S1 and Table S1). We excluded all these individuals and
222 reduced our dataset to 78 individuals. We conducted a new *de novo* assembly and
223 analyzed the relationship between the 78 remaining individuals at 6461 SNPs
224 present in at least 70% of individuals using PCA, NJ trees and ADMIXTURE. Both the
225 k-means clustering (DAPC) and the variation of the cross-validation error as a
226 function of the number of ancestral populations in ADMIXTURE revealed that the
227 polymorphism of *An. moucheti* resulted from only one ancestral population (k = 1)
228 (Fig. 1B and 1C). PCA and NJ depicted a homogeneous cluster comprising all 78

229 individuals providing additional evidence of the lack of genetic or geographic
230 structuring among populations (Fig. 1D and 1E). Unsurprisingly, the overall F_{ST} was
231 remarkably low between populations from the two sampling locations Olama and
232 Nyabessan (F_{ST} = 0.008, p < 0.005). Similarly, the distribution of F_{ST} values across
233 6461 SNPs showed a large dominance of very low F_{ST} values throughout the genome
234 (Fig. 2). The highest per locus F_{ST} was only 0.126, while 5006 of the 6461 loci
235 revealed F_{ST} near zero. The modest geographic differentiation was also well
236 illustrated by a hierarchical AMOVA, which showed that the genetic variance was
237 explained essentially by within-individual variations (99.7%). Finally, we found very
238 low overall Wright's inbreeding coefficient (F_{IS}= 0.0014, p < 0.005 in Nyabessan and
239 F_{IS} = 0.0025, p < 0.005 in Olama) (Table 2) suggesting that allelic frequencies within
240 both populations were in accordance with proportions expected under the Hardy-
241 Weinberg equilibrium.

242 **3.3. Genetic diversity and demographic history**

243 The estimates of the overall nucleotide diversity (π = 0.0020 and π = 0.0016,
244 respectively, in Olama and Nyabessan) (Table 2) were within the range of average
245 values found in other African *Anopheles* species using RADseq approaches (Fouet et
246 al., Unpublished data; Kamdem et al., Unpublished data (a, b); O'Loughlin et al.,
247 2014). Notorious demographic expansions have been described in natural
248 populations of this insect clade (Donnelly et al., 2001), and the values of π observed
249 in *An. moucheti* likely reflect the level of genetic diversity of a population with large
250 effective size. The great genetic diversity of *An. moucheti* was also illustrated by the
251 percentage of polymorphic sites. Of the 6461 variant sites, 89.60% were

252 polymorphic in Olama and 34.82% in Nyabessan (Table 2). The difference observed
253 between the two locations can be related to the sample size (n = 19 in Nyabessan
254 and n = 59 in Olama) or to demographic particularities that persists between the
255 two geographic sites despite a massive gene flow. To infer the demographic history
256 of *An. moucheti*, we examined the Allele Frequency Spectrum (AFS), summarized as
257 the distribution of the major allele in one population. This approach was a surrogate
258 to model-based methods that provide powerful examinations of the history of
259 genetic diversity by modeling the AFS at genome-wide SNP variants, but that
260 couldn't be implemented here due to the lack of a reference genome assembly. The
261 frequency distribution of the major allele p (Fig. 3) indicates that the majority of
262 polymorphic loci are highly frequent in Olama and Nyabessan as shown by the
263 predominance of SNPs at frequencies equal to 1. Ranges of allele frequencies are
264 also similar in both locations (between 0.47 and 1 in Olama and between 0.34 and 1
265 in Nyabessan). These frequency ranges are expected for old populations at
266 equilibrium capable of accumulating high amount of genetic diversity.

267 **3.4. Polymorphic chromosomal inversions and local adaptation**

268 When paracentric inversions are involved in local adaptation, high values of genetic
269 divergence are often observed within inversion loci in natural populations.
270 Cytogenetic analyses of the polytene chromosome of *An. moucheti* have identified
271 three polymorphic chromosomal inversions within samples collected from the sites
272 we have studied (Sharakhova et al., 2014). However, the weak overall population
273 structure and the very low F_{ST} values we have detected throughout the genome are
274 clear indicators of the absence of local adaptation. Interestingly, this finding also

275 suggests that none of the polymorphic inversions described previously is actually
276 segregating among our samples, as high values of F_{ST} are absent even within
277 inversion loci. We provided further support to this hypothesis by performing LD
278 analyses. First, we found a globally low LD in the *An. moucheti* genome (average
279 genome-wide r^2 = 0.0149) as expected in highly polymorphic populations with large
280 effective size. We next used LDna to cluster the LD values and to identify Single
281 Outlier Clusters (SOC) that can be associated with distinct or multiple evolutionary
282 phenomena in the *An. moucheti* history. We set the parameters to collect and screen
283 a high number of SOCs using 2569 highly filtered SNPs, which allowed us to identify
284 20 independent LD blocks in our samples (Fig. 4). In principle, when these blocks
285 are associated with important events in the evolutionary history of a species,
286 downstream analyses can reveal clear pattern reflecting the underlying process
287 (Kemppainen et al., 2015). This has been illustrated for example by studies
288 demonstrating that SNPs within SOCs generated by polymorphic inversions in
289 *Anopheles baimaii* clearly separate the three expected karyotypes (inverted
290 homozygotes, heterozygotes and standard homozygotes) (Kemppainen et al., 2015).
291 We conducted downstream analyses with a PCA using SNPs identified within the
292 SOCs. As shown in Fig S2, although individuals were occasionally spread along three
293 PCA axes, no distinct cluster could be identified from any of the 20 SOCs. These
294 results were consistent with the absence of segregating inversions and local
295 adaptation in our samples and corroborated low F_{ST} values observed throughout the
296 genome. Precisely, in our data, we couldn't identify polymorphic inversions whose
297 karyotype frequencies change between Olama and Nyabessan due to a differential

298 adaption between the two sites. Some of the different SOCs identified can be
299 associated with other processes that were not captured by our analytical approach;
300 others are probably methodological artifacts associated with the LDna pipeline
301 (Kemppainen et al., 2015).

302 **4. Discussion**

303 We have analyzed the genome-wide polymorphism and characterized some of the
304 baseline population genomic parameters in *An. moucheti*, an important malaria
305 vector in rural areas across the African rainforest. We found very little
306 differentiation among our samples, with most of the genetic variation distributed
307 within individuals. Although a more substantial sampling will be necessary to fully
308 dissect the population genetic structure of this species, our finding likely reflects the
309 current dynamic of *An. moucheti* populations in Cameroon. It is worth mentioning
310 that we have surveyed a total of 28 locations across the country (Fig 1A), some of
311 which were known from several past surveys to harbor *An. moucheti* populations
312 (Antonio-Nkondjio et al., 2013, 2009, 2008, 2006, 2002; Kengne et al., 2007), but we
313 confirmed the presence of the species in only 2 villages. Extant populations of *An.*
314 *moucheti* are distributed in patches of favorable habitats along river networks
315 where larval populations breed. Our results indicate that despite this apparent
316 fragmentation, connectivity and gene flow are high among population aggregates.
317 The weak population genetic structure of *An. moucheti* observed with genome-wide
318 markers corroborated results obtained with microsatellites and allozymes (Antonio-
319 Nkondjio et al., 2008, 2002). A survey of eight microsatellite loci revealed that the
320 highest F_{ST} among Cameroonian populations was as low as 0.003. Nevertheless, a
321 substantial differentiation was found between samples from different countries
322 consistent with an isolation-by-distance model (Antonio-Nkondjio et al., 2008). It is
323 clear that a deep sequencing of continental populations is necessary to further
324 clarify the status of these putative subpopulations. However, samples collected at

325 lower spatial scales like ours are also very relevant as they can allow robust
326 inferences about ongoing selective processes that cannot be captured at continental
327 scale. Although RADseq samples only a small fraction of the genome and certain
328 signatures of selection are likely missing when reduce representation sequencing
329 approaches are used, it has been shown that such approaches can effectively capture
330 strong footprints of selection across genomes of *Anopheles* mosquitoes (Fouet et al.,
331 Unpublished data; Kamdem et al., Unpublished data a). We have found that
332 signatures of selection are rare in the genome of *An. moucheti* populations from the
333 Cameroonian rainforest. Populations remain largely undifferentiated throughout the
334 genome, with F_{ST} values near zero across the vast majority of variations suggesting
335 that no local adaptation is ongoing. This perception is further supported by the
336 absence of segregating linkage disequilibrium blocks between geographic locations.
337 The characterization of chromosomal inversions with cytogenetic methods can be
338 laborious and challenging (Kirkpatrick, 2010; Sharakhova et al., 2014). So far, three
339 paracentric polymorphic inversions have been discovered in *An. moucheti* in
340 Cameroon (Sharakhova et al., 2014). The ecological, behavioral or functional roles of
341 these inversion polymorphisms remain unknown. We have implemented a recently
342 designed method that uses Next Generation Sequencing and LD estimates to
343 indirectly identify paracentric inversions whose karyotype frequencies varies
344 among populations due to local adaptation (Kemppainen et al., 2015). Our LD
345 analyses revealed the presence of a few LD clusters that are however not associated
346 with inversions. On the other hand, the low overall LD observed across the genome
347 reflected the significant genetic polymorphism that seems to prevail within *An.*

348 *moucheti* populations. This polymorphism translates into exceptional levels of
349 overall genetic diversity and very high percentage of polymorphic sites that are in
350 the range of values observed in other mosquito species undergoing significant
351 demographic expansions (Donnelly et al., 2001; Fouet et al., Unpublished data;
352 Kamdem et al., Unpublished data a). The amount of neutral genetic diversity is often
353 viewed as a correlate of the adaptive potential of a species (Orr and Unckless, 2008).
354 Although the relationship is more complex in reality, estimates of neutral genetic
355 diversity are commonly used in conservation biology as an intuitive conceptual and
356 management framework to assess the genetic resilience of endangered species
357 (Bonin et al., 2007; Latta IV et al., 2010). Our population genomic analyses have
358 depicted *An. moucheti* as a species with a great genetic diversity and hence a
359 sustainable long-term adaptive resilience. Implications of our findings in malaria
360 epidemiology and control can be very significant. First, *An. moucheti* is essentially
361 endophilic and is particularly sensitive to the principal measures currently
362 employed to control malaria in Sub-Saharan Africa such as the massive use of
363 Insecticide Treated Nets (ITNs) and Indoor Residual insecticide Spraying (IRS). For
364 example, estimates of population effective size in one village in Equatorial Guinea
365 indicated that both mass distribution of ITNs and IRS campaigns resulted in a
366 decline of approximately 55% of *An. moucheti* (Athrey et al., 2012). However, the
367 great genetic diversity and the massive gene flow we observed within populations
368 could easily enable this vector to challenge population declines and recover from
369 shallow bottlenecks. Moreover, most insecticide resistance mechanisms found in
370 insects exploit standing genetic variation to rapidly respond to the evolutionary

371 challenge by increasing the frequency of existing variations rather than relying on
372 infrequent *de novo* mutations (Messer and Petrov, 2013). As a result, despite the
373 current sensitivity of *An. moucheti* to common insecticides, the significant amount of
374 standing genetic variation provides the species with a great potential to challenge
375 insecticide-based interventions and other types of human-induced stress.

376 **5. Conclusions**

377 Recent advances in sequencing allow sensitive genomic data to be generated for
378 virtually any species (Ellegren, 2014). However, the most important information we
379 can obtain from population resequencing approaches often depends on the
380 availability and the quality of genomic resources such as a well-annotated reference
381 genome. The reduced genome sequencing strategy (RADseq) offers a cost-effective
382 strategy that can be used to effectively study the genetic variation in a broad range
383 of species from yeast to plants, insects, etc., in the absence of a reference genome.
384 We have extended this approach to the study of the genetic structure of an
385 understudied mosquito species with a great epidemiological significance. We have
386 provided both significant baseline population genomic data and the methodological
387 validation of one approach that should motivate further studies on this species and
388 other understudied anopheline mosquitoes lacking genomic resources.

389

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. We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

397 **References**

398 Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of
399 ancestry in unrelated individuals. Genome Res. 19, 1655–1664.

400 Antonio-Nkondjio, C., Demanou, M., Etang, J., Bouchite, B., 2013. Impact of cyfluthrin
401 (Solfac EW050) impregnated bed nets on malaria transmission in the city of
402 Mbandjock : lessons for the nationwide distribution of long-lasting insecticidal
403 nets (LLINs) in Cameroon. Parasit. Vectors 6, 10. doi:10.1186/1756-3305-6-10

404 Antonio-Nkondjio, C., Kerah, C.H., Simard, F., Awono-Ambene, P., Chouaibou, M.,
405 Tchuinkam, T., Fontenille, D., 2006. Complexity of the malaria vectorial system
406 in Cameroon: contribution of secondary vectors to malaria transmission. J. Med.
407 Entomol. 43, 1215–1221.
408 doi:10.1603/0022-2585(2006)43[1215:COTMVS]2.0.CO;2

409 Antonio-Nkondjio, C., Ndo, C., Awono-Ambene, P., Ngassam, P., Fontenille, D., Simard,
410 F., 2007. Population genetic structure of the malaria vector Anopheles moucheti
411 in south Cameroon forest region. Acta Trop. 101, 61–68.
412 doi:10.1016/j.actatropica.2006.12.004

413 Antonio-Nkondjio, C., Ndo, C., Costantini, C., Awono-Ambene, P., Fontenille, D.,
414 Simard, F., 2009. Distribution and larval habitat characterization of Anopheles
415 moucheti, Anopheles nili, and other malaria vectors in river networks of
416 southern Cameroon. Acta Trop. 112, 270–276.
417 doi:10.1016/j.actatropica.2009.08.009

418 Antonio-Nkondjio, C., Ndo, C., Kengne, P., Mukwaya, L., Awono-Ambene, P.,
419 Fontenille, D., Simard, F., 2008. Population structure of the malaria vector

420 Anopheles moucheti in the equatorial forest region of Africa. Malar. J. 7, 120.
421 doi:10.1186/1475-2875-7-120
422 Antonio-Nkondjio, C., Simard, F., Cohuet, A., Fontenille, D., 2002. Morphological
423 variability in the malaria vector, Anopheles moucheti, is not indicative of
424 speciation: evidences from sympatric south Cameroon populations. Infect.
425 Genet. Evol. 2, 69–72. doi:10.1016/S1567-1348(02)00084-9
426 Athrey, G., Hodges, T.K., Reddy, M.R., Overgaard, H.J., Matias, A., Ridl, F.C.,
427 Kleinschmidt, I., Caccone, A., Slotman, M. a, 2012. The effective population size
428 of malaria mosquitoes: large impact of vector control. PLoS Genet. 8, e1003097.
429 doi:10.1371/journal.pgen.1003097
430 Ayala, D., Fontaine, M.C., Cohuet, A., Fontenille, D., Vitalis, R., Simard, F., 2011.
431 Chromosomal inversions, natural selection and adaptation in the malaria vector
432 Anopheles funestus. Mol. Biol. Evol. 28, 745–758. doi:10.1093/molbev/msq248
433 Ayala, F.J., Coluzzi, M., 2005. Chromosome speciation: humans, Drosophila, and
434 mosquitoes. Proc. Natl. Acad. Sci. U. S. A. 102 Suppl, 6535–42.
435 doi:10.1073/pnas.0501847102
436 Bonin, A., Nicole, F., Pompanon, F., Miaud, C., Taberlet, P., 2007. Population adaptive
437 index: A new method to help measure intraspecific genetic diversity and
438 prioritize populations for conservation. Conserv. Biol. 21, 697–708.
439 doi:10.1111/j.1523-1739.2007.00685.x
440 Catchen, J., Hohenlohe, P. a., Bassham, S., Amores, A., Cresko, W.., 2013. Stacks: An
441 analysis tool set for population genomics. Mol. Ecol. 22, 3124–3140.
442 doi:10.1111/mec.12354

443 Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J.H., 2011. Stacks:
444 building and genotyping Loci de novo from short-read sequences. G3
445 (Bethesda). 1, 171–82. doi:10.1534/g3.111.000240
446 Clarkson, C.S., Weetman, D., Essandoh, J., Yawson, A.E., Maslen, G., Manske, M., Field,
447 S.G., Webster, M., Antão, T., MacInnis, B., Kwiatkowski, D., Donnelly, M.J., 2014.
448 Adaptive introgression between Anopheles sibling species eliminates a major
449 genomic island but not reproductive isolation. Nat. Commun. 5, 4248.
450 doi:10.1038/ncomms5248
451 Donnelly, M.J., Licht, M.C., Lehmann, T., 2001. Evidence for recent population
452 expansion in the evolutionary history of the malaria vectors Anopheles
453 arabiensis and Anopheles gambiae. Mol. Biol. Evol. 18, 1353–1364.
454 doi:10.1093/oxfordjournals.molbev.a003919
455 Ellegren, H., 2014. Genome sequencing and population genomics in non-model
456 organisms. Trends Ecol. Evol. 29, 51–63. doi:10.1016/j.tree.2013.09.008
457 Excoffier, L., Smouse, P.E., Quattro, J.M., 1992. Analysis of molecular variance
458 inferred from metric distances among DNA haplotypes: Application to human
459 mitochondrial DNA restriction data. Genetics 131, 479–491.
460 doi:10.1007/s00424-009-0730-7
461 Fontaine, M.C., Pease, J.B., Steele, a., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I. V.,
462 Jiang, X., Hall, a. B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.
463 a., Love, R.R., Lawniczak, M.K., Slotman, M. a., Emrich, S.J., Hahn, M.W., Besansky,
464 N.J., 2015. Extensive introgression in a malaria vector species complex revealed
465 by phylogenomics. Science (80). 347, 1258524–1258524.

466 doi:10.1126/science.1258524
467 Fouet, C., Kamdem, C., White, B.J., 2016. Chromosomal inversions facilitate
468 chromosome-scale evolution in Anopheles funestus. bioRxiv.
469 Gillies, M.T., Coetzee, M., 1987. A supplement to the Anophelinae of Africa south of
470 the Sahara. The South African Institute for Medical Research, Johannesburg.
471 Gillies, M.T., De Meillon, B., 1968. The Anophelinae of Africa South of the Sahara,
472 Second Edi. ed. Publications of the South African Institute for Medical Research,
473 Johannesburg.
474 Harrisson, K. a., Pavlova, A., Telonis-Scott, M., Sunnucks, P., 2014. Using genomics to
475 characterize evolutionary potential for conservation of wild populations. Evol.
476 Appl. n/a-n/a. doi:10.1111/eva.12149
477 Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic
478 markers. Bioinformatics 24, 1403–1405.
479 Kamdem, C., Fouet, C., Gamez, S., White, B.J., 2016a. Pollutants and insecticides drive
480 local adaptation in African malaria mosquitoes. bioRxiv.
481 Kamdem, C., Fouet, C., Gamez, S., White, B.J., 2016b. Genomic signatures of
482 introgression at late stages of speciation in a malaria mosquito. bioRxiv.
483 Kemppainen, P., Knight, C.G., Sarma, D.K., Hlaing, T., Prakash, A., Maung Maung, Y.N.,
484 Somboon, P., Mahanta, J., Walton, C., 2015. Linkage disequilibrium network
485 analysis (LDna) gives a global view of chromosomal inversions, local adaptation
486 and geographic structure. Mol. Ecol. Resour. n/a-n/a. doi:10.1111/1755-
487 0998.12369
488 Kengne, P., Antonio-Nkondjio, C., Awono-Ambene, H.P., Simard, F., Awolola, T.S.,

489 Fontenille, D., 2007. Molecular differentiation of three closely related members
490 of the mosquito species complex, Anopheles moucheti, by mitochondrial and
491 ribosomal DNA polymorphism. Med. Vet. Entomol. 21, 177–182.
492 doi:10.1111/j.1365-2915.2007.00681.x
493 Kirkpatrick, M., 2010. How and why chromosome inversions evolve. PLoS Biol. 8.
494 doi:10.1371/journal.pbio.1000501
495 Kirkpatrick, M., Barton, N., 2006. Chromosome inversions, local adaptation and
496 speciation. Genetics 173, 419–434. doi:10.1534/genetics.105.047985
497 Latta IV, L.C., Fisk, D.L., Knapp, R.A., Pfrender, M.E., 2010. Genetic resilience of
498 Daphnia populations following experimental removal of introduced fish.
499 Conserv. Genet. 11, 1737–1745. doi:10.1007/s10592-010-0067-y
500 Lewontin, R.C., Kojima, K., 1960. The evolutionary dynamics of complex
501 polymorphisms. Evolution (N. Y). doi:10.2307/2405995
502 Meirmans, P., Van Tienderen, P., 2004. GENOTYPE and GENODIVE: two programs for
503 the analysis of genetic diversity of asexual organisms. Mol. Ecol. Notes 4, 792–
504 794.
505 Messer, P.W., Petrov, D., 2013. Population genomics of rapid adaptation by soft
506 selective sweeps. Trends Ecol. Evol. 28, 659–669.
507 doi:10.1016/j.tree.2013.08.003
508 Norris, L.C., Main, B.J., Lee, Y., Collier, T.C., Fofana, A., Cornel, A.J., Lanzaro, G.C., 2015.
509 Adaptive introgression in an African malaria mosquito coincident with the
510 increased usage of insecticide-treated bed nets. Proc. Natl. Acad. Sci.
511 201418892. doi:10.1073/pnas.1418892112

512 O'Loughlin, S.M., Magesa, S., Mbogo, C., Mosha, F., Midega, J., Lomas, S., Burt, A., 2014.
513 Genomic Analyses of Three Malaria Vectors Reveals Extensive Shared
514 Polymorphism but Contrasting Population Histories. Mol. Biol. Evol. 1–14.
515 doi:10.1093/molbev/msu040
516 Orr, H.A., Unckless, R.L., 2008. Population extinction and the genetics of adaptation.
517 Am. Nat. 172, 160–9. doi:10.1086/589460
518 Paradis, E., Claude, J., Strimmer, K., 2004. Analyses of Phylogenetics and Evolution in
519 R language. Bioinformatics 20, 289–290.
520 Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double Digest
521 RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in
522 Model and Non-Model Species. PLoS One 7, e37135.
523 doi:10.1371/journal.pone.0037135
524 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J.,
525 Sklar, P., de Bakker, P., Daly, M., Sham, P., 2007. PLINK: a toolset for whole-
526 genome association and population-based linkage analysis. Am. J. Hum. Genet.
527 81.
528 Sharakhova, M. V., Antonio-Nkondjio, C., Xia, a., Ndo, C., Awono-Ambene, P., Simard,
529 F., Sharakhov, I. V., 2014. Polymorphic chromosomal inversions in Anopheles
530 moucheti, a major malaria vector in Central Africa. Med. Vet. Entomol. 28, 337-
531 340. doi:10.1111/mve.12037
532 Team, R.D.C., 2008. R: A language and environment for statistical computing. R
533 Foundation for Statistical Computing, Vienna, Austria.
534 Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of

population structure. Evolution (N. Y). 38, 1358–1370.

World Health Organization, 2013. World malaria report 2013. World Health WHO/HTM/GM, 238. doi:ISBN 978 92 4 1564403

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.

545 **Tables**

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

Sampling locations	Geographic coordinates	Sampling methods			Total
		HLC-OUT	HLC-IN	LC	
Nyabessan	2°24'00"N, 10°24'00"E	21	15	1	37
Olama	3°26'00"N, 11°17'00"E	30	31	0	61
Total					98

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

547

548 **Table 2**: Population genomic parameters based on 6461 variant sites reflecting the

549 genetic diversity and conformity to Hardy-Weingberg equilibrium.

	Number of individuals*	Sites	Observed heterozygosity	F_{IS}	Nucleotide diversity (π)	% polymorphic sites
Variant positions						
Olama	54.12	6 461	0.0445	0.0631	0.0505	89.60
Nyabessan	16.78	6 461	0.0334	0.0372	0.0402	34.82
All positions						
Olama	54.47	165 975	0.0017	0.0025	0.0020	3.49
Nyabessan	16.86	165 975	0.0013	0.0014	0.0016	1.36

* Mean number of individuals per locus in this population (as estimated by Stacks v 1.35)

550

551

552 **Figure legends**

553 **Figure 1**: Relationship between *An. moucheti* individuals from Olama and
554 Nyabessan. (A) Map of the study site showing both the locations surveyed (small
555 black dots) and the two villages (large red and blue squares) where *An. moucheti*
556 samples were collected. (B) and (C) Plots of the ADMIXTURE cross-validation error
557 and the Bayesian Information Criterion (BIC) (DAPC) as a function of the number of
558 genetic clusters indicating that k = 1. The lowest BIC and CV error indicate the
559 suggested number of clusters. (D) and (E) Absence of genetic structure within
560 populations illustrated by neighbor-joining and PCA. The percentage of variance
561 explained by each PCA axis is indicated.

562 **Figure 2**: Frequency distribution of F_{ST} between Olama and Nyabessan across 6461
563 SNP loci and plot of these F_{ST} values along arbitrary positions in the genome.

564 **Figure 3**: Allele Frequency Spectrum for 6461 SNP loci in Nyabessan and Olama
565 populations. The x-axis presents the frequency of the major allele and the y-axis the
566 frequency distribution of loci in each class of the major allele frequency.

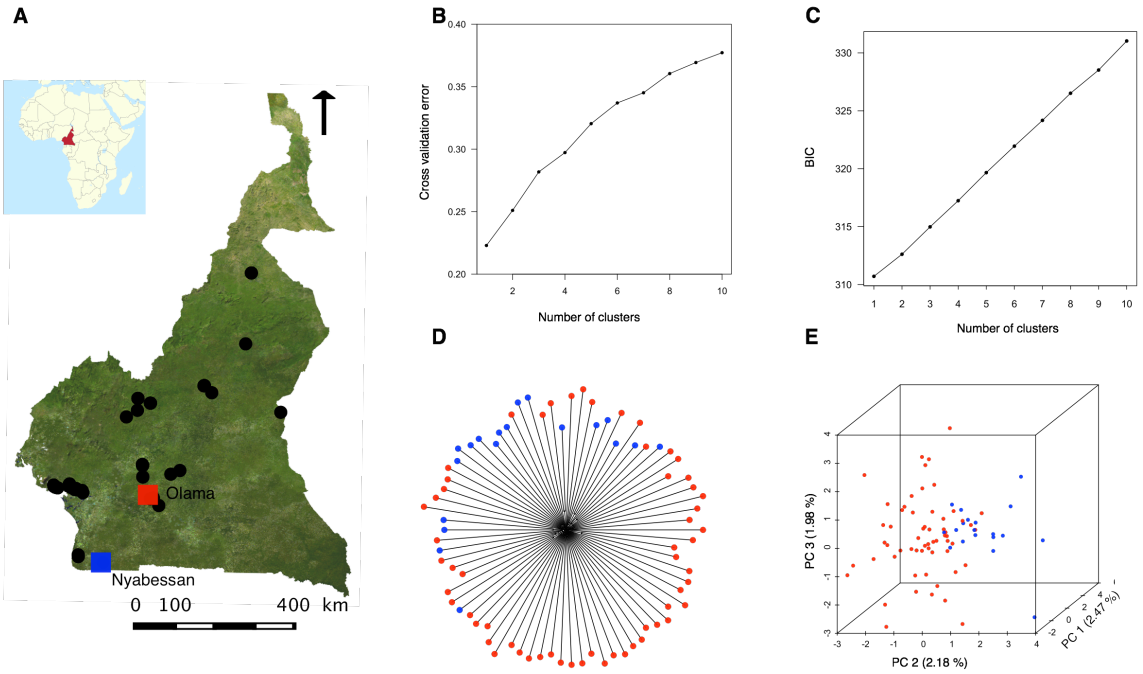
567 **Figure 4**: LDna analyses on 2569 SNPs showing the presence of 20 Single Outlier
568 Clusters (SOCs) of linkage disequilibrium in *An. moucheti*. The graph presents the
569 results obtained with values of the two parameters: φ (which controls when clusters
570 are defined as outliers) and $|E|_{\min}$, the minimum number of edges required for a LD
571 cluster to be considered as an outlier, indicated on top. LD thresholds are shown on
572 the x-axis.

573

574

575 **Figures**

576 **Figure 1**



577

Figure 2

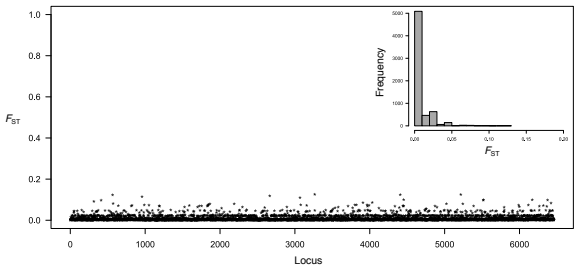

1.0
0.8
0.6
0.4
0.2
0.0
F_{ST}
0
1000
2000
3000
4000
5000
6000
Locus
Frequency
F_{ST}

Figure 3

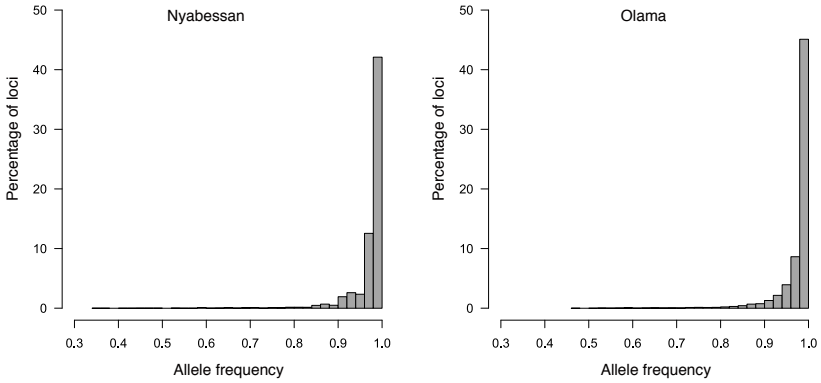
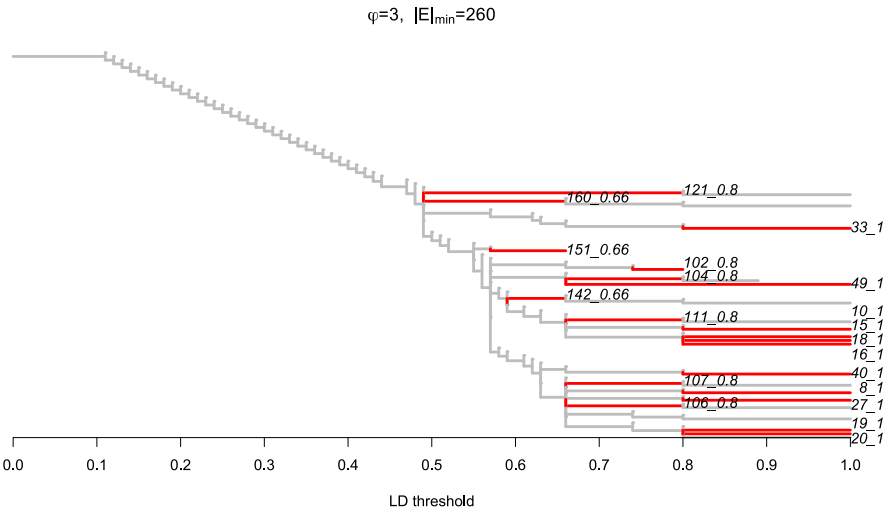

Nyabessan
Olama
Percentage of loci
Allele frequency
0
10
20
30
40
50
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4


φ=3, |E|min=260
160_0.66
121_0.8
33_1
151_0.66
102_0.8
104_0.8
49_1
142_0.66
10_1
111_0.8
15_1
18_1
16_1
40_1
107_0.8
8_1
106_0.8
27_1
19_1
20_1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
LD threshold

587 **Supplemental Material**

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

592 **Figure S2**: PCA indicating the population genetic structure inferred from SNPs
593 within the 20 Single Outlier Clusters (SOCs) of linkage disequilibrium identified in
594 *An. moucheti* (red: Olama; blue: Nyabessan).

595

Figure S1:

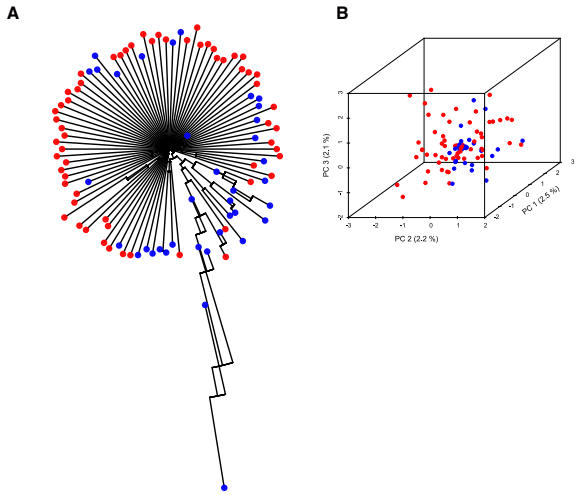
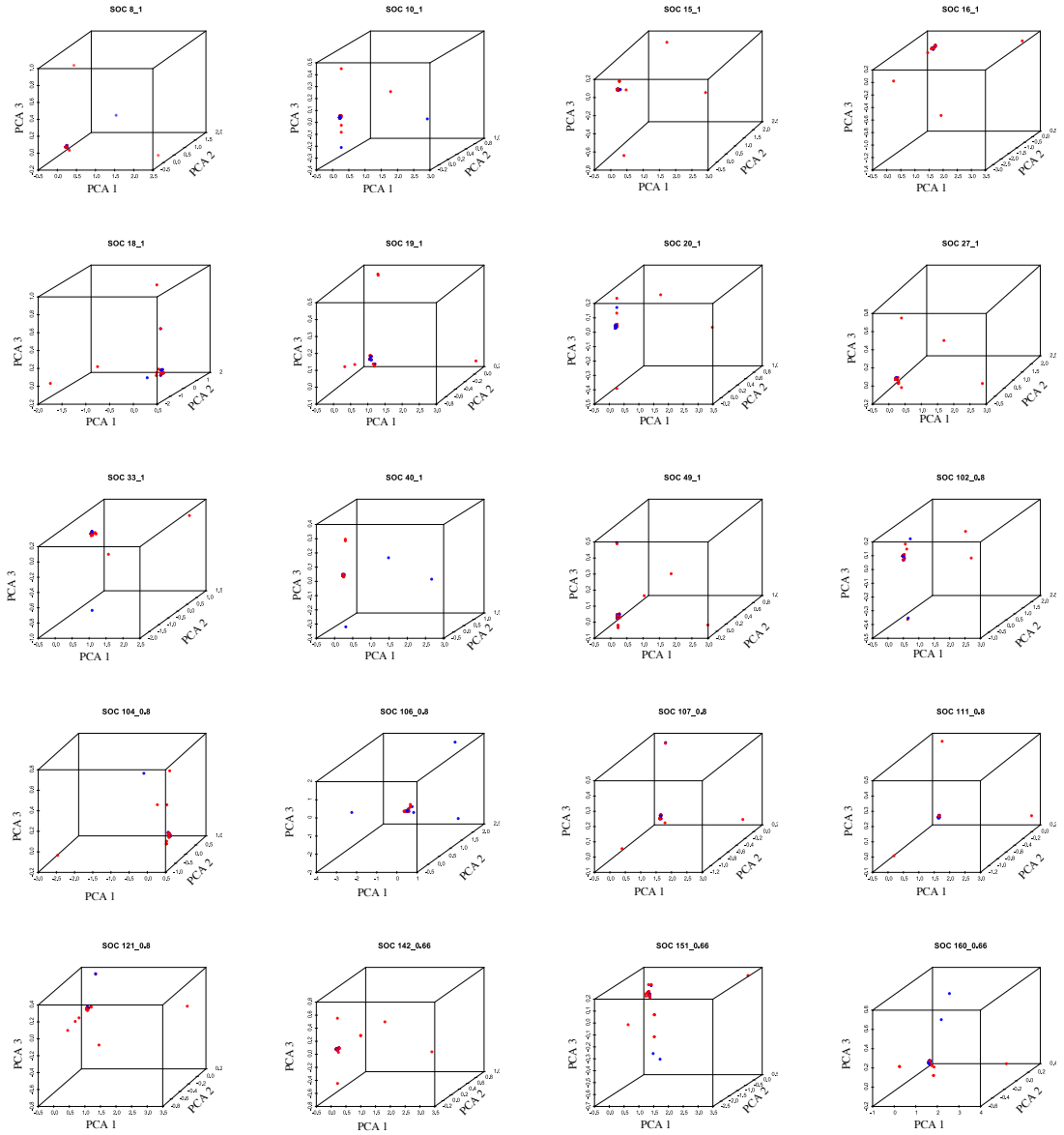


Figure S2:


SOC 8_1
SOC 10_1
SOC 15_1
SOC 16_1
SOC 18_1
SOC 19_1
SOC 20_1
SOC 27_1
SOC 33_1
SOC 40_1
SOC 49_1
SOC 102_0.8
SOC 104_0.8
SOC 106_0.8
SOC 107_0.8
SOC 111_0.8
SOC 121_0.8
SOC 142_0.66
SOC 151_0.66
SOC 160_0.66
PCA 1
PCA 2
PCA 3

603 **Table S1:** Distribution of the number of reads among sequenced individuals.
604 Individuals below the dashed line were excluded from analysis.

Mosquito ID	Total number of reads	Ambiguous/low quality reads	Retained reads	Sampling location
43	5190686	140706	5044739	Olama
850	4211705	136872	4066857	Nyabessan
27	2888406	77609	2800516	Olama
97	2773623	56956	2714480	Olama
94	2729590	61848	2660187	Olama
26	1915405	45856	1866121	Olama
82	1910651	50411	1854108	Olama
37	1764370	51133	1706506	Olama
95	1720323	37181	1678873	Olama
93	1450516	31318	1406481	Olama
25	1334643	45439	1285047	Olama
79	1263586	35615	1222166	Olama
80	1259353	42228	1202777	Olama
98	957698	20322	934945	Olama
49	874718	20033	852633	Olama
36	867344	21542	842175	Olama
99	853444	17206	834357	Olama
84	836989	18434	816949	Olama
76	850380	26481	814198	Olama
31	782599	18231	759965	Olama
86	761099	17235	741314	Olama
81	735291	13915	716347	Olama
1007	732960	19269	710564	Nyabessan
48	725668	17270	705049	Olama
743	703141	15098	686817	Nyabessan
46	639840	16191	621897	Olama
42	634781	16180	617726	Olama
92	593819	12260	579932	Olama
34	575175	11846	561789	Olama
90	579585	17480	556841	Olama
868	572514	16379	553831	Nyabessan
47	569810	12524	551627	Olama
28	557485	12363	534835	Olama
30	552365	13813	534263	Olama
38	541076	13190	525801	Olama
L_627	531803	12874	516440	Nyabessan
32	520421	11244	507425	Olama
78	510457	11112	493998	Olama
19	511063	15054	493676	Olama

87	507445	10896	492038	Olama
85	498180	10747	485508	Olama
792	478211	10753	465355	Nyabessan
35	467060	9616	455391	Olama
33	463964	10001	453130	Olama
60	465276	11304	452748	Olama
77	428787	10568	415155	Olama
45	391371	8565	380806	Olama
61	372074	10772	360550	Olama
851	363537	10079	352761	Nyabessan
40	346713	6421	339007	Olama
89	350001	12049	331665	Olama
15	341218	8157	331365	Olama
91	337294	12007	320481	Olama
44	323246	7406	313719	Olama
731	315382	10965	302496	Nyabessan
100	311056	5531	301968	Olama
729	295394	6781	287687	Nyabessan
18	290682	8206	281904	Olama
39	225981	6085	217620	Olama
73	220793	6020	212919	Olama
70	189567	4906	183785	Olama
756	188521	4639	182974	Nyabessan
75	181946	6799	173364	Olama
730	178968	5081	171828	Nyabessan
758	177709	5115	170782	Nyabessan
732	161697	4019	149839	Nyabessan
742	149159	3223	144972	Nyabessan
793	132552	4456	126135	Nyabessan
14	109314	3398	104900	Olama
96	99548	2293	95813	Olama
741	94635	2614	90106	Nyabessan
724	88884	3176	83605	Nyabessan
16	85486	2665	81748	Olama
88	83193	3414	78289	Olama
777	80457	2346	77695	Nyabessan
41	80127	2363	75238	Olama
72	76135	2366	73512	Olama
723	71212	2266	67258	Nyabessan
17	67037	2188	64443	Olama
795	65448	1778	62623	Nyabessan
869	57496	2654	54157	Nyabessan
760	45620	1610	43127	Nyabessan

727	43892	1402	40694	Nyabessan
71	37279	1233	35811	Olama
794	36669	1359	34439	Nyabessan
754	35476	1137	33712	Nyabessan
726	35274	1149	33230	Nyabessan
757	37431	1289	32648	Nyabessan
761	29783	947	28368	Nyabessan
833	27881	800	26582	Nyabessan
776	24245	783	22874	Nyabessan
755	24690	1012	22274	Nyabessan
753	19802	1301	17379	Nyabessan
728	17829	713	16750	Nyabessan
725	16069	706	14676	Nyabessan
810	14088	610	12963	Nyabessan
762	12307	568	11575	Nyabessan
763	11562	901	10374	Nyabessan

605