Genomic signatures of introgression at late stages of speciation in a malaria mosquito

Colince Kamdem1*, Caroline Fouet1, Stephanie Gamez1, Bradley J. White1,2*

1Department of Entomology, University of California, Riverside, CA 92521
2Center for Disease Vector Research, Institute for Integrative Genome Biology, University of California, Riverside, CA 92521

*Corresponding authors: colincek@ucr.edu; bwhite@ucr.edu

Key words: Introgression, speciation, linkage disequilibrium, divergent selection, Anopheles nili
Abstract

Hybridization plays a central role during the evolution of species boundaries, but the relative impact of gene flow on genomic divergence and vice versa remains largely unknown. The genome architecture of populations and emerging species exhibiting various levels of divergence along the speciation continuum should provide insights into the events that promote or prevent speciation. In this work, we have used a combination of population genomic approaches to examine the genomic signatures of hybridization between Anopheles nili sensu stricto and An. ovengensis, two malaria mosquitoes that have split ~3-Myr ago. Despite this substantial time since divergence, the two species hybridize extensively in nature, giving rise to a unique population of differentiated hybrids in contact zones. Using genomic clines and Bayesian models, we showed that signatures of introgression are widespread across the genome suggesting that recent hybridization between Anopheles nili sensu stricto and An. ovengensis involves multiple fitness traits and functional classes. Linkage Disequilibrium analyses allowed us to identified a block of 39 linked loci that segregated between hybrids and parental species and may harbour genes responsible for reproductive isolation. Our results demonstrate that genome-wide admixtures can persist in the face of species divergence over long periods of time during speciation due to increased gene flow at loci providing selective advantage.
Introduction

Most cases of speciation occur in a gradual manner so that emerging species continue to mate and exchange genes before the onset of complete reproductive isolation (Nosil 2012; Nadeau et al. 2013). As a result, secondary contacts between diverging lineages are pervasive in nature and can lead to several scenarios including extensive hybridization and the creation of a new hybrid species or genetic homogenization (James 2007; Ozerov et al. 2016). The role of hybridization either in the creation of biodiversity through speciation or in its reduction through genetic homogenization has been under increasing scrutiny over the last decades and has been recognized as one major force driving the evolution of plant and animal species (James 2007; Abbott et al. 2013; Mallet et al. 2015). However, although there is consistent evidence that hybridization can provide the raw material for evolutionary diversification (Grant 2015), examples of homoploid hybrid species (where the hybrid offspring has the same ploidy level as the two parental species) are rather rare in nature (James 2007; Abbott et al. 2013). Most often, hybrids have reduced fitness relative to parental populations and barely persist as a reproductively isolated unit in the wild.

It is particularly difficult to predict the outcome of secondary contacts between diverging lineages and specifically the rate of hybridization (Payseur & Rieseberg 2016). One main issue is that the level of divergence between species supposed to encourage or prevent hybridization remains obscure. Until a recent past, estimates of the population differentiation parameter $F_{ST}$ have been used to predict the rate of hybridization between incomplete species. Variations in $F_{ST}$ among groups of related species were often
interpreted as indicative of differences in rates of gene flow across populations, high $F_{ST}$ values being considered as evidence of limited gene flow. However, instances of hybridization and introgression between highly divergent taxa are accumulating, thereby indicating that estimates of population differentiation are particularly poor predictors of hybridization between two species (Nydam & Harrison 2011; Roux et al. 2013; Parchman et al. 2013; Martin et al. 2013; Canestrelli et al. 2014).

Likewise, genomic regions showing high differentiation among populations have been thought to harbor hybrid incompatibility or other reproductive isolation (RI) factors. In support of this view, genome scans in many couples of incipient species have identified genomic regions of extreme $F_{ST}$ values qualified as “speciation islands”, considered to be enriched in RI factors and resistant to gene flow (Turner et al. 2005; Hohenlohe et al. 2010; Lawniczak et al. 2010; Ellegren et al. 2012). Yet, not only cases where genes involved in RI have been explicitly identified in $F_{ST}$ outliers regions of the genome are extremely rare, but also extensive introgression has been found to sometimes coincide with regions of high differentiation. For example, (Parchman et al. 2013) used genomic clines and Bayesian models to study the correlation between regions of exceptional introgression and genetic differentiation manakins birds ($Manacus$). Contrary to expectations, they found that loci of strong introgression were relatively positively correlated to the genetic differentiation. The most plausible interpretation of this intriguing result is that divergent selection may promote introgression of select genes or genomic regions, resulting in differential patterns of introgression across the genome. A relative concordance between locus-specific
divergence and locus-specific measures of introgression has also been described in the
Lycaeides butterfly (Gompert et al. 2012, 2013) and in Drosophila melanogaster (Pool et al.
2012). Overall, the relationship between introgressive hybridization and genetic divergence
at the species or the genome level is more complex than previously thought. Whether
exceptionally differentiated regions of the genome harbor genes that can, paradoxically,
introgress easily is an empirical question that remains opened and should be answered by
studying closely related species at hybrid zones.

Mosquitoes of the genus Anopheles are ideal system in which to address the evolutionary
processes at hybrid zones due to the prevalence of adaptive speciation (Ndo et al. 2013;
Neafsey et al. 2015; Kamdem et al. 2016). In this paper, we focused on An. nili, a group of
African malaria vectors characterized by a reticulate evolution leading to complex
phylogenies that have been challenging to clarify (Kengne et al. 2003; Awono-Ambene et al.
2004, 2006; Ndo et al. 2010, 2013; Peery et al. 2011; Sharakhova et al. 2013). The group
harbour four known species identified by slight morphological differences: An. nili sensu
nili) is the most widespread across the continent while the three other species are more
patchily distributed primarily in the equatorial rainforest. An. nili and An. ovengensis are
important vectors of human malaria parasites (Antonio-Nkondjio et al. 2006). We sampled
populations across the ranges of the four species of the An. nili group in Cameroon and we
used a population resequencing approach to develop genome-wide SNP markers that we
genotyped in 145 individuals. Our first aim was to clarify the evolutionary relationships
between populations. We discovered new cryptic species within *An. ovengensis* and *An. nili* as well as a hybrid subgroup resulting from massive hybridization in sympatric areas. Finally, we took advantage of the recent implementation of genomic cline models that enables the investigation of footprints of introgression across genomes of non-model organisms without the need of a high-quality reference genome (Gompert & Alex 2011; Gompert & Buerkle 2012), to critically evaluate the interplays between the locus-specific divergent selection and introgression.
Materials and methods

Mosquito sample

We surveyed a total of 28 geographic locations that were representative of the main habitats of species of the An. nili group previously described in Cameroon (Fig 1A) (Awono-Ambene et al. 2004, 2006; Antonio-Nkondjio et al. 2009; Ndo et al. 2010, 2013). The different species were identified using reference morphological identification keys and a diagnostic PCR that discriminates the four currently known members of the An. nili group on the basis of a point mutation of the ribosomal DNA (Gillies & De Meillon 1968; Gillies & Coetzee 1987; Kengne et al. 2003).

Library preparation, sequencing and SNP discovery

We created double-digest RAD (ddRAD) libraries using a modified version of the protocol described by Peterson et al, 2012 (Peterson et al. 2012). Genomic DNA of mosquitoes was extracted using the DNeasy Blood and Tissue kit (Qiagen) and the Zymo Research MinPrep kit on larvae and adult samples respectively. Approximately 50 ng (10µl) of DNA of each mosquito sample was digested simultaneously with MluCI and NlaIII restriction enzymes. Digested products were then ligated to adapter and barcode sequences to enable the unique identification of the individual associated with each sequencing read. The samples were pooled, purified, and 400-bp fragments selected. The resulting libraries were then amplified via PCR and purified. The distribution of library fragment-size was checked on a BioAnalyzer (Agilent Technologies, Inc., USA). The PCR products were quantified and diluted for sequencing on Illumina HiSeq2000 (Illumina Inc., USA) to yield single-end reads of 101 bp.
Bioinformatics pipeline

The *An. nili* Dinderesso draft genome assembly comprises 51,048 contigs, varying between 100 and 26775kb in length, to which short reads can be aligned and SNP called. However, members of the *An. nili* group found in Cameroon diverged ~0.2-6 million years ago (Ndo et al. 2013). Thus, alignments of our samples to *An. nili* Dinderesso reference genome might be subjected to an important bias associated with the inconsistent mapping of reads from highly divergent populations. To make sure that this potential reference sequence bias didn’t undermined our analyses, we compared our results across two SNP sets that were identified within RAD loci created using two distinct approaches: a *de novo* assembly and an assembly of reads aligned onto the reference genome. Recent studies using RAD sequencing on *Heliconius* butterflies showed that combining *de novo* assemblies and reference alignments provided a robust approach to perform rigorous test on introgression and phylogenetic relationships in distantly related species (The Heliconius Genome Consortium 2012). The *process_radtags* program of the Stacks v 1.35 pipeline was used to demultiplex and clean raw reads. Reads without the *NlaIII* restriction site and those bearing ambiguous barcode sequences or having low-quality score (average Phred score < 33) were discarded. Reads were trimmed to 96-bp by removing index and barcode sequences. We aligned the short reads to the *An. nili* Dinderesso draft genome assembly using Gsnap (Wu & Nacu 2010) with a maximum of five nucleotide mismatches allowed. The *ref_map.pl* and *denovo_map.pl* programs in Stacks were used to identify consensus RAD loci and to call SNPs within these loci across our populations using respectively the Gsnap-aligned SAM files or the individual fastq files as input. For both analyses, we set the minimum number of
reads required to form a stack to 3. In the denovo assembly, we allowed a maximum of three mismatches when creating loci in every individual (M parameter in denovo_map.pl) and two mismatches when building the catalogue of loci across individuals (n parameter). In reference-based assembly, we specified n = 2 in ref_map.pl to allow two mismatches during catalogue creation. We generated SNP files in different formats for further downstream analyses using the populations program of Stacks and PLINK v1.09 (Purcell et al. 2007).

**Population genetic structure**

We analyzed the genetic structure of An. nili s.l. populations using Principal Component Analysis (PCA) and Neighbor-Joining trees (NJ). We also examined ancestry proportions and admixtures between populations in ADMIXTURE v1.23 (Alexander et al. 2009) and STRUCTURE v2.3.4 (Pritchard et al. 2000). We performed these tests using filtered SNPs identified in RAD loci present in every population and in at least 50% of individuals by the populations program in Stacks. We used the R package adegenet (Jombart 2008) to implement the PCA. Neighbor-Joining trees were generated from matrixes of Euclidian distance computed from allele frequencies at genome-wide SNPs using the R package ape (Paradis et al. 2004). We ran ADMIXTURE with 10-fold cross-validation for values of k from 1 through 20. We analyzed patterns of ancestry from k ancestral populations in STRUCTURE, testing five replicates of k = 1-10. We used 200000 iterations and discarded the first 50000 iterations as burn-in for each STRUCTURE run. CLUMPP v1.1.2 (Jakobsson & Rosenberg 2007) was used to summarize assignment results across independent runs and DISTRACT v1.1 (Rosenberg 2004) to plot STRUCTURE results. To identify the optimal number of genetic clusters in our samples, we used simultaneously the lowest cross-
validation error in ADMIXTURE, the ad-hoc statistic deltaK (Evanno et al. 2005; Earl & VonHoldt 2012) and the Discriminant Analysis of Principal Component (DAPC) method in adegenet.

Population genomics analyses

To quantify the genetic differentiation between species and putative subgroups within species, we used a subset of 1000 high-quality SNPs to calculate the overall pairwise differentiation index $F_{ST}$ (Weir & Cockerham 1984) in Genodive v 1.06 (Meirmans & Van Tienderen 2004). Statistical significance was assessed with 10000 permutations. To further examine the genomic footprints of selection and introgression, we used ANGSD v 0.612 (Korneliussen et al. 2014) and ngsTools (Fumagalli et al. 2014) to derive locus-specific estimates of the nucleotide diversity (measured as $\theta_w$ and $\theta_\pi$), Tajima's $D$, absolute sequence divergence ($d_{sy}$) and $F_{ST}$ across 9622 sites using Gnap alignments without SNP calling. To identify statistical outliers of $F_{ST}$, we used the software LOSITAN which applies the coalescent simulation method (FDIST2) (Beaumont & Nichols 1996) to identify loci with exceptionally high $F_{ST}$ values relative to a “neutral” genome-wide $F_{ST}$ value expected under neutral evolution. We ran LOSITAN using the neutral $F_{ST}$ option, 50,000 simulations and a false discovery rate of 0.05%. Finally, we used the R package LDna (Kemppainen et al. 2015) for identifying clusters of linkage disequilibrium in our samples. This analysis was conducted primarily to assess the possible influence of inversion polymorphism on the genomic architecture of divergence and hybridization. Precisely, we wished to test if loci with great significance in the genetic divergence and/or introgression were independent loci scattered throughout the genome or clusters of linked loci encapsulated in low
recombination regions like chromosomal inversions. We used PLINK to calculate the
linkage disequilibrium (estimated as the $r^2$ correlation coefficient) between all pairs of SNPs
and we applied the graph-based method implemented in LDna to search for clusters of
strongly correlated SNPs.

Tests of introgression

In addition to the interpretation of patterns of ancestry provided by clustering analyses in
STRUCTURE and ADMIXTURE, we conducted formal tests to infer the history of population
splits and admixtures. We used the three-population ($f_3$) and the four-population ($f_4$)
statistics, introduced in (Reich et al. 2009) as methods to estimate mixture proportions in
an admixed group, to test for introgression among species. The $f_3$ and $f_4$ statistics exploit the
idea that the genetic drift, defined as a function of allele frequency, should be uncorrelated
in unadmixed populations. As a result, correlations detected between allele frequencies of
three ($f_3$) or four ($f_4$) populations joined by an unrooted tree indicate episodes of gene flow.
Specifically, the $f_3$ ($X; Y, W$) tests for admixture between a test population X and two
reference populations Y and W. The expected value of $f_3$ is positive in case of no mixture and
negative if X is admixed with Y or W, or both. Similarly, the $f_4$ for an unrooted tree (A,B;C,D),
tests whether allele frequency differences between A and B are correlated with differences
between C and D. $f_4$ is equal to zero if there are no correlations and no admixtures across
branches of the tree. In contrast, $f_4$ is significantly positive if admixture occurs between A
and C, or between B and D, or both, and significantly negative if admixture occurs between
A and D, or between B and C, or both. The statistical significance of $f_3$ and $f_4$ values is
assessed using a Z-score: an $f_3$ or $f_4$ value divided by its standard deviation. We used a
threshold Z-score of 2.5 corresponding to a p-value of 0.05 as suggested in (Reich et al. 2009). The threepop and fourpop programs of the software TreeMix (Pickrell & Pritchard 2012) were used to calculate $f_3$ and $f_4$ statistics and Z-scores for all possible triples and quadruples of populations we described within An. nili s.l.

We next applied the graph-based method developed in TreeMix to determine the directionality and quantify the extent of gene flow among species. The TreeMix approach uses allele frequencies at genome-wide polymorphisms and a Gaussian approximation of the genetic drift among populations to first construct a Maximum Likelihood (ML) phylogeny connecting sampled populations by simple bifurcations. The model then compares the covariance structure modeled by this dendrogram and adds edges to the phylogeny to account for admixtures. We first conducted one TreeMix run without migration. We noted the percentage of explained variance of the models and visually inspected the residuals of covariance matrixes among populations. We next ran TreeMix 100 times without migration using different numbers of random seed and we built a ML consensus tree from the 100 trees in SumTrees (Sukumaran & Holder 2010) using the 95% majority rule. We finally added 1 and 4 migration edges to the ML trees and examined the changes in the percentage of explained variance and the residual fit. For all $f_3$, $f_4$ and TreeMix analyses, we used 4343 SNPs that were present in all populations and in at least 50% of individuals in every population.

**Genomic clines**

Loci of exceptional introgression compared with genome-wide average admixture may be important for local adaptation or the maintenance of species barriers. To identify sites with
the greatest contribution to introgression between *An. nili* and *An. ovengensis*, we used the Bayesian model of genomic clines as implemented in *bgc* (Gompert & Buerkle 2012). The model first estimates a genome-wide average of a hybrid index (*h*) that varies between zero and one for every potentially admixed individual, using allele frequencies of “pure” parental populations (Gompert & Alex 2011). The value zero corresponds to pure individuals of the alternative species (*An. ovengensis* in our case) and the value one to pure individuals of the reference species (*An. nili* in our case). Two genomic clines parameters α and β are then used to evaluate the deviation of individual loci in admixed individuals from the expected genome-wide hybrid index (*h*). The cline parameter α reflects the probability of *An. nili* ancestry relative to the base expectation (*h*), whereas the genomic cline parameter β denotes the rate of transition from low to high probability of *An. ovengensis* ancestry as a function of hybrid index (Gompert & Alex 2011). Deviations of loci from the genome-wide average of α and β were examined on the basis of departures from the 95% confidence envelope.

The “pure” parental populations we used in genomic clines analyses were apparently allopatric *An. nili* and *An. ovengensis* populations collected in Nyabessan and Nkoteng, respectively (Fig. 1A). The two locations are separated by ~420 km, which presumably reduces the rate of gene flow between *An. nili* and *An. ovengensis*. Hybrids were admixed individuals collected from the sympatric area in Mbébé (Fig. 1A). These individuals had almost equivalent ancestry proportions from *An. nili* and *An. ovengensis* as revealed by both STRUCTURE and ADMIXTURE clustering analyses. In *bgc*, we estimated genomic cline parameters using 9622 SNPs deriving from aligned RAD-loci in a dataset consisting of 11
admixed individuals, 8 *An. nili* parents and 15 *An. ovengensis* parents. We calculated average values of $\alpha$ and $\beta$ across five runs of *bgc*, each run including 50000 steps with samples from the posterior distribution recorded every 25th step following a 25000 step burn-in. We visually inspected the MCMC output to assess convergence to the stationary distribution.

To examine the relationship between the strength of selection and introgression, we tested whether loci with extreme or outlier genomic clines were enriched in genomic regions that were targets of selection. To do so, we studied the correlation between locus-specific cline parameters ($\alpha$ and $\beta$) and several divergence and diversity statistics ($F_{ST}$, $d_{sy}$, $\theta_{w}$, $\theta_{\pi}$, Tajima’s $D$) across 9622 SNPs in the genome. The strength of correlation was assessed using Pearson’s product moment correlation coefficient.
Results

SNP genotyping

We collected mosquitoes from four locations (Fig. 1A, Table S1) and sequenced 145 individuals belonging, according to morphological identifications and PCR, to two species (An. nili (n = 24) and An. ovengensis (n = 121)). We aligned reads from all 145 individuals to the reference genome and we assembled 197724 96-bp RAD loci that mapped to unique positions throughout the genome. We retained 408 loci present in all populations and in at least 50% of individuals in each population and identified 4343 high-quality biallelic markers within these loci. We also identified another set of 704408 unique loci by building consensus RAD loci de novo without aligning reads to the reference genome. We applied the same stringent filtration as with aligned reads to identify 3071 high-quality SNPs.

Genetic structure of populations

PCA and NJ trees showed that the genotype variation at 4343 genome-wide SNPs among the 145 sequenced individuals is best explained by more than two clusters, implying cryptic subdivisions among An. nili and An. ovengensis populations (Fig. 1C-D). The first three PCA axes and NJ trees clearly distinguished three subgroups within An. nili and two clusters in An. ovengensis. The five different clusters were associated with the sampling locations suggesting a strong correlation between the genetic structure and a local distinctness of populations. This marked geographic structure of An. nili and An. ovengensis populations can be explained by the ongoing adaptive divergence and ecological speciation within the two species. Importantly, both STRUCTURE and ADMIXTURE analyses revealed that, at k = 3, a subgroup containing 11 individuals corresponded to a cluster of hybrids with almost
half ancestry from *An. nili* and half from *An. ovengensis* (Fig. 1B). This result is surprising given the substantial time since divergence. We applied three different methods to identify the optimal number of genetic clusters in our samples. The DAPC suggested the presence of five clusters as indicated by PCA and NJ trees (Fig. 1E). However, the method of Evanno et al. indicated two probable ancestors while the distribution of the cross-validation error as a function of the number of putative populations in ADMIXTURE failed to unambiguously reveal an optimal number of genetic clusters (Fig. 1F,G). These conflicting results between methods are sometimes observed when the history of subdivisions and admixtures events is very complex as it is the case in *An. nili* s.l (Decker et al. 2014). The Evanno et al. method is overwhelmed by early divergence between *An. nili* and *An. ovengensis* while the results of DAPC and the ADMITURE cross-validation error reflect recent hierarchical population subdivisions within the two species. The inferred genetic structure was consistent when we used respectively 4343 and 3071 SNPs identified from reference-based and *de novo* assemblies (Fig. 1 and Fig. S1).

**Genetic differentiation**

We estimated the level of population differentiation between the five genetic clusters identified in our samples using the overall pairwise $F_{ST}$ (Table 1). We found strong pairwise differentiation characterized by extreme $F_{ST}$ values, including between populations classified as belonging to the same species. The level of differentiation is even higher between some pairs of populations within the same species than between *An. nili* and *An. ovengensis*. The highest level of within-species genetic differentiation was recorded between “allopatric” populations of *An. ovengensis* collected from locations separated by
~350km ($F_{ST} = 0.896$, $p < 0.005$). These findings strongly suggest that, in addition to hybridization, local differentiation associated with late stages of adaptive divergence within species are overwhelming current taxonomic descriptions in *An. nili* s.l. populations. The results obtained with reference alignments perfectly mirrored those of de novo assemblies (Table 1, S2). Taken as a whole, the population genetic structure and divergence of *An. nili* and *An. ovengensis* depicts a radiating group involving a collection of species whose phylogenetic relationships are blurred by ongoing hybridizations. As a result, current taxonomic classifications based on morphological characters and point mutations on the ribosomal DNA cannot effectively describe the actual reproductive units and the vectorial capacity within the *An. nili* species group.

**Evidence for recent gene flow between *An. nili* and *An. ovengensis**

We used *TreeMix* to construct a tree connecting the five species and to effectively describe and visualize the mixture(s) event(s). The consensus of 100 ML *TreeMix* tree inferred from 4343 SNPs without migration resumes population-level relationships described by neighbor-joining analysis, PCA and clustering analyses (Fig. 2A). The long terminal branches leading to *An. nili* group 1, *An. nili* group 2 and *An. ovengensis* group 1 in the inferred tree reflect the signatures of strong bottlenecks in the history of these species. Interestingly, the residual fit from the ML model without migration suggested some correlations between species that were consistent with the admixture events inferred from ADMIXTURE and STRUCTURE analyses, particularly between the two species *An. nili* group 2 and *An. ovengensis* group 2 (Fig. 1B). Additional support to admixture was provided by the four-population test. Across all possible combinations of populations, there was a highly
significant allele frequency correlation between *An. nili* group 2 and *An. ovegensis* group 1 confirming that *An. nili* group 1 is the result of an admixture event between *An. nili* group 2 and *An. ovegensis* group 1 (Table 2 and S3). Nevertheless, five independent runs of TreeMix including one or two migrations edges captured none of these migration events. Also, the explained variance of the TreeMix ML model without migration edge was very high (99.8%) indicating that a bifurcation tree can also match the phylogenetic relationships among sampled populations. Finally, we found no significant $f_3$ statistics in all triplets of populations, but this test may be underpowered because of the low-level and the precocity of admixture or because of the complex demographic history of our populations (Decker et al. 2014). As we showed previously with the population structure and genetic divergence, the admixture events we described between *An. nili* and *An. ovegensis* also hold across assembly methods (de novo or reference alignments)(Fig. 2 and S2).

**Genomic signatures of divergent selection and differential introgression**

We retained 9622 SNP loci identified from a reference-based assembly that we used to examine the genomic footprints of divergent selection and introgression in a dataset containing 8 “pure” *An. nili* s.s. individuals, 15 “pure” *An. ovegensis* and 11 admixed individuals (see Materials and methods). The values of $F_{ST}$ among SNP loci between parent species *An. nili* and *An. ovegensis* were heterogeneous across the genome (Fig. 3). The empirical distribution of locus-specific $F_{ST}$ also revealed an unusually bimodal shape featuring $F_{ST}$ peaks centered on values around 0 and 1. The great majority of sites have low to moderate differentiation, but a substantial number of loci are extremely differentiated between the two species (Fig. 3A). We thinned our SNP set to 4003 using more stringent
filtration criteria and identified 42 statistical outliers of $F_{ST}$ in LOSITAN, which were all above the 99th percentile of the empirical distribution of $F_{ST}$. The bimodality of the genetic divergence between An. nili and An. ovengensis is also evident in the empirical distribution of $d_{xy}$ (Fig. 3B). As exemplified by some empirical cases like the introgressive hybridization observed between two closely related species of monkeyflowers (Mimulus) (Brandvain et al. 2014), the bimodal divergence is due to the coexistence between sites of abrupt differentiation reflecting the real level of divergence between the two species and sites of low divergence resulting from recent gene flow. We conducted Linkage Disequilibrium (LD) analyses to understand if highly differentiated sites are clustered into blocks of linked loci or dispersed throughout the genome. Particularly, polymorphic chromosomal inversions are important sources of genetic variation in Anopheles species and play a key role in local adaptation and speciation. High $F_{ST}$ values can aggregate in genomic regions containing chromosomal inversions polymorphisms that are under strong divergent selection and drive adaptive segregation as often observed in other African anopheline mosquitoes (Ayala et al. 2011; Cheng et al. 2012; Kamdem et al. 2016). Cytogenetic studies have detected no polymorphic chromosomal inversion among An. ovengensis samples. Two polymorphic inversions (2Rb, 2Rc) have been identified in An. nili, but samples from Cameroon bore only the 2Rc at a low frequency (Sharakhova et al. 2013). Rearranged regions of the genome are characterized by reduced recombination and increased LD relative to the genome-wide average. To effectively examine the extent of LD in the genomes of An. nili and An. ovengensis and to test for the effect of chromosomal inversion polymorphisms in adaptive evolution, we used the package LDna to identify clusters of LD.
To prevent spurious clusters due to LD between SNPs located on the same RAD locus, we used a dataset in which only one SNP was randomly selected within each RAD locus (1330 SNPs in total). We optimized LDna clustering parameters, which indicated the presence of 6 LD blocks (Single Outlier Clusters (SOCs)) among our samples (Fig. 4). These clusters represent signals of independent or compound events in the evolutionary history that left imprints on LD across the genome (Kemppainen et al. 2015). To further understand the role of the 6 LD clusters in the evolutionary history of An. nili and An. ovengensis, we conducted downstream analysis using PCA to examine the population structure of SNPs within each SOC. We found that one SOC (containing 39 SNPs) clearly discriminated a cluster encompassing all hybrid individuals from the two parent species (Fig. 4B). Two other SOCs (containing respectively 383 and 112 SNPs) consistently separated one parent species from a group comprising the other parent species and all hybrid individuals. The last three SOCs revealed no clear clustering patterns. None of the SOCs differentiated clusters that could be associated with the three alternative karyotypes (inverted homozygotes, heterozygotes and uninverted homozygotes) expected in case of polymorphic inversion. As suggested by cytogenetic observations, polymorphic inversions likely play only a moderate role in the genomic divergence in An. nili s.l, but a more intensive sampling is needed to make any definitive conclusion. Interestingly, the three most important blocks of LD found in the genomes of An. nili and An. ovengensis are instead associated with hybridization, which emphasizes the central role played by interspecific gene flow in the genomic architecture of adaptive speciation in the two species. Our results also suggest that loci resistant to introgression and genes responsible for reproductive isolation between An. nili and An.
*ovengensis* are in strong linkage disequilibrium. This finding is consistent with recent introgression whereby long-range haplotypes that are generated by recent gene flow between genetically distinct populations have not have sufficient time to be broken down by recombination (admixture-induced LD) (Martin *et al.* 2013). However, LDna estimates the LD irrespective of the physical linkage between markers and LD blocks can also result from a strong correlation between allelic frequencies of SNPs scattered throughout the genome. A detailed characterization of long-range haplotypes across genomes of *An. nili* and *An. ovengensis* will provide a more powerful examination of the genomic architecture of reproductive isolation between the two species.

Here we have used another innovative approach (genomic cline) to further understand the relationship between divergence, selection and differential introgression at the genomic level. First, consistent with STRUCTURE and ADMIXTURE results, the empirical distribution of hybrid index among admixed individuals (average hybrid index ~0.3) shows a slight predominance of *An. ovengensis* ancestry (Fig. 5A, Fig. 1B). Using estimates of genomic cline parameters, we noted that introgression was very heterogeneous across loci (Fig. 5B). Indeed, the level of introgression at 1297 SNP loci differed significantly from the genome-wide average (outliers) and we detected an excess *An. ovengensis* ancestry for 2506 loci (lower bound of 95% CI for $\alpha > 0$) and excess *An. nili* ancestry for 2635 loci (upper bound of 95% CI for $\alpha < 0$). Estimates of the genomic cline parameter $\alpha$ range from -5 to 5.25 while our values of $\beta$ are low overall ($\beta$ varies from -0.13 to 0.09). Moreover, we identified no loci with significantly elevated estimates of genomic cline rate. The $\beta$ parameter assesses the rate of transition from one ancestry to the other and thereby scores the steepness of the
genomic cline at each locus. Extreme values of $\beta$ are expected when there is population structure in the hybrid zone, selection against hybrids or gene flow among parent species (Parchman et al. 2013). Our hybrid population is highly differentiated ($F_{ST} > 0.8$) from both parent species and is well adapted in hybrid zone, which certainly explain the low $\beta$ values observed. One key aspect in the Bayesian implementation of genomic clines is the relationship between $\alpha$ and $\beta$, which is crucial to understand the rate of transition of sites of exceptional introgression from one side to the other of the genome. As we have shown previously with $F_{ST}$ and $d_{xy}$, recent introgression in divergent genomes of An. nili and An. ovengensis has resulted in bimodal genomic divergence featuring two blocks of sites with extreme divergence values. This pattern presumes that most variant sites will have very steep genomic clines because the transition from one ancestry to the other is very abrupt. In agreement with this prediction, the scatterplot of $\beta$ as a function of $\alpha$ indicates the presence of two blocks of SNPs with either high probability of An. nili or of An. ovengensis ancestry (Fig. 5B). The genomes of both species are split into two compartments of ancestry due to recent introgression resulting in steep genomic clines at hybrid zone (Fig. 5B). To better understand the biological significance of these outlier loci of introgression in admixed individuals, we assessed the correlation between $\alpha$ and $\beta$ and locus-specific estimates of divergence and selection parameters. We found a detectable negative correlation between $F_{ST}$ and values of $\alpha$ ($r = -0.15$, $p < 0.005$). Among the 32 LOSITAN $F_{ST}$ outliers, 13 loci had extreme $\alpha$ estimates, and 11 were outlier of $\beta$, but the average values of both $\alpha$ and $\beta$ were not significantly different between $F_{ST}$ outliers and the 9622 genome-wide SNP loci (Mann-Whitney U-test, $P < 0.001$). Erroneous correlation between $F_{ST}$ and
exceptional introgression can be inferred when the $F_{ST}$ between parental populations is $< 0.1$ (Parchman et al. 2013). The high overall $F_{ST}$ ($\sim 0.8$) between An. nili and An. ovengensis minimizes such errors in our study. We also detected a negative correlation between $\alpha$ and $d_{xy}$ ($r = -0.13$, $p < 0.005$), which confirms previous results with $F_{ST}$, suggesting that sites that diverge strongly resist to differential introgression because they likely contain reproductive isolation factors. The negative correlation between $d_{xy}$ and estimates of locus-specific admixture indicates that perhaps gene flow has had sufficient time to reduce sequence divergence at admixed loci despite the relatively recent hybridization process. We next correlated cline parameters ($\alpha$ and $\beta$) to the estimates of locus-specific diversity ($\theta_w$ and $\theta_\pi$) and the allele frequency spectrum (Tajima’s $D$). The results provided another clear illustration of the steep genomic clines described previously with alpha and beta values. Notably, there is a strong positive correlation between $\theta_w$ and $\alpha$ ($r = 0.53$, $p < 0.005$) in An. ovengensis and a negative correlation of the same magnitude with the diversity of An. nili ($r = -0.49$, $p < 0.005$), which translate the segregation of genomes of admixed individuals in two sources of ancestry that are at opposite ends of the genomic cline. Interestingly, the high correlation between alpha and diversity indicate that sites that are favored by introgression are not constrained by divergent selection that should have likely resulted in depression in nucleotide diversity. In contrast to what has been shown in manakin birds (Parchman et al. 2013) and in Lycaeides butterflies (Gompert et al. 2012, 2013), there is no evidence of increased introgression at loci under divergent selection between An. nili and An. ovengensis.
Discussion

Hybridization at late stages of speciation

We have described a complex case of speciation and hybridization in a group of related *Anopheles* mosquitoes endemic to Sub-Saharan Africa. Both incomplete speciation and pervasive hybridization are leading to a very dynamic pattern of genetic structure between and within species. We first analysed the population genetic structure and revealed cryptic subdivisions between the two malaria vectors *An. nili* and *An. ovengensis*. The exact number of demes or cryptic subgroups within each species in our samples was difficult to determine. We found conflicting results between estimates from three different genetic clustering methods. However, three different methods suggested the existence of 2 to 5 clusters in our sample. *An. nili* is subdivided into three clusters while ongoing adaptive speciation in *An. ovengensis* results so far in two cryptic species. The extremely high genetic differentiation between populations indicates that all these cryptic subgroups are almost complete species. Our results therefore suggest that *An. nili* and *An. ovengensis*, contrary to the current taxonomy, represent probably two different complexes of cryptic species. The geographic origin of samples explains a great part of the genetic variance among individuals consistent with strong local differentiation and adaptive speciation. Significant population structure has been described among *An. nili* populations from Cameroon with 8 microsatellite loci (Ndo *et al.* 2013). Ndo *et al.* 2013 also found $F_{ST}$ values as high as 0.48 between two populations from the forest area. Using genome-wide SNPs, we have identified new subdivisions and revealed that the population genetic structure of *An. nili* is more complex. Our work highlights the strength of Next Generation Sequencing (NGS)
approaches and the necessity of fine-scale genomic examinations in the resolution of intricate patterns of ancestry in this group of mosquito. By contrast, the ongoing speciation we have observed within *An. ovengensis* has never been described in the past, and this species has been sometimes considered as a sibling of *An. nili* (Kengne et al. 2003; Awono-Ambene et al. 2004, 2006; Ndo et al. 2013). Nevertheless, more recent studies have started to challenge the assumed relatedness between the two species (Ndo et al. 2013; Sharakhova et al. 2013). Precisely, analyses of polytene chromosomes revealed high karyotypic divergence of *An. ovengensis* from *An. nili* (Sharakhova et al. 2013) and estimates of time since divergence indicated that the two species split from one another 3 to 6-Myr ago (Ndo et al. 2013). Intriguingly, our work provides multiple lines of evidence supporting the existence of extensive ongoing gene flow between the two species despite this strong divergence. First, clustering methods and Bayesian implementation of genomic clines identified individuals with almost half ancestry from both species and formal tests of population admixture corroborated these ongoing admixture events. Second, perhaps the most compelling evidence for hybridization is the presence of admixture-induced linkage disequilibrium (ALD) characterized by blocks of linked SNPs that discriminate hybrid populations from parental species. In general, linkage disequilibrium (LD) or the correlation between allele frequencies of different loci across the genome, which can have multiple origins including selection, genetic drift, or population structure, is normally eroded by recombination in the course of time. ALD is caused by associations between nearby loci co-inherited on an intact chromosomal block from one of the ancestral mixing populations (Loh et al. 2013). Signatures of ALD are frequent in genomes of recently
admixed populations for which recombination has not yet broken the large introgressed chromosomal segments into smaller portions. ALD is a well-known feature in evolutionary history of humans and estimates of long-range LD have been proposed as an approach to measure the extent and the timing of admixture events that have shaped the genetic polymorphism across genomes of extant populations (Loh et al. 2013). In insects, signatures of ALD have been found for example in the Heliconius butterfly genome, especially around loci involved in mimicry of color patterns that circulate between species (Martin et al. 2013). An. nili provides another rare example of ALD across the genome of an insect species. Further studies using a more comprehensive genomic sequencing and a reference genome of better quality will help us to improve our knowledge of functional and sequential characteristics of admixed LD blocks found in the An. nili genome.

Although the concept of speciation-with-gene-flow has become the dominant paradigm in speciation studies, we remain ignorant about the conditions that prevent or motivate gene flow between divergent lineages before the onset of complete reproductive isolation. Moreover, even the notion of “complete reproductive isolation” is now challenged because an increasing number of examples from diverse taxa showed rampant gene exchange across strong reproductive barriers, sometimes between established species that diverged several million years ago (Nydam & Harrison 2011; Roux et al. 2013; Parchman et al. 2013; Martin et al. 2013; Canestrelli et al. 2014). These studies and ours suggest that processes underlying hybridization and introgression in the presence of clear genetic differentiation will be best addressed within speciation continuum rather than across couples of occasionally mating species. A continuum of speciation featuring a collection of taxa
occupying a gradient of genetic/ecological divergence provides ideal conditions where the
relation between divergence and hybridization can be inferred. Contrary to what we
initially thought, the An. nili group does not represent a speciation continuum but instead a
collection of complete species sharing a common ancestry. Moreover, despite extensive
efforts, we couldn’t sample populations of the two remaining species of the group: An.
somalicus and An. carnevalei. Therefore, patterns of admixtures and ancestry among
populations of this group of mosquitoes are probably more complex than what we have
shown. Nevertheless, instead of a comprehensive description of splits and admixtures in An.
nili s.l., we have focused our efforts on the examination of the genomic architecture of
divergence and introgression between An. nili and An. ovengensis. The results we discuss
below are among the rare cases that address the genomic signatures of gene flow and the
relationship between divergence, selection and introgression at late stages of the speciation
process.

Genomic architecture of adaptive introgression

Owing to the increasing availability of high-throughput sequencing information, genomes of
multiple taxa have been scanned and analyzed in comparative frameworks to search for
genomic signatures of divergence and speciation. The prevailing idea behind these
approaches is that regions of extreme differentiation between incipient or complete species
contain factors that maintain reproductive isolation (RI) among their populations. In
reality, the genomic distribution of highly differentiated regions has been more contentious
(Nosil & Feder 2012, 2013). Overall, two models are well documented: the “speciation
island” model whereby RI loci are thought to be caught up in in a few regions of the genome
where outliers of genetic differentiation cluster (e.g. (Andrew & Rieseberg 2013)) and the
“heterogeneous” distribution model, which posits that genomic divergence is instead lead
by numerous small genomic regions scattered throughout all chromosomes (e.g.
(Lawniczak et al. 2010; Roesti et al. 2012)). A shift in this paradigm is envisioned as
evidence is accumulating indicating that highly differentiated loci can paradoxically
coincide with regions of elevated introgression between two species (Gompert et al. 2012,
2013; Pool et al. 2012; Parchman et al. 2013). The genomic distribution of the genetic
divergence between An. ovengensis and An. nili shows that the compound effects of strong
divergence and recent introgression generates a bimodal pattern of divergence which
assigns most sites into two main categories: a majority of low divergence sites and a small
cluster of high divergence loci with $F_{ST}$ values centered around 1. In most of the widespread
Anopheles species, signatures of high divergence can be found in large chromosomal
segments corresponding to rearranged regions of chromosomes where recombination is
rare (Neafsey et al. 2015). This is the case for example in the 2La inversion locus, which
depicts signatures of strong divergent selection along a latitudinal cline (Cheng et al. 2012).
In agreement with cytogenetic studies, which found roughly no polymorphic inversions
among An. nili samples from Cameroon, we observed no tendency for high $F_{ST}$ or $d_{xy}$ to
cluster within regions that can be assimilated to chromosomal rearrangements. In contrast
to what has been recently demonstrated in manakin birds (Parchman et al. 2013) and in
Lycaeides butterflies (Gompert et al. 2012, 2013), there is a negative correlation between
the introgression parameter $\alpha$ and genetic divergence (estimated both as $F_{ST}$ an $d_{xy}$)
between An. nili and An. ovengensis. As expected, gene flow is favored across neutral loci
and those that provide selective advantage, but are presumably not under strong divergent selection. Moreover, the relationship between $\alpha$ and $\beta$ suggests a prevalence of steep clines in genomes of admixed individuals at contact zones. This also translates into a positive correlation between $\alpha$ and the nucleotide diversity in \textit{An. ovengensis} and a negative correlation of $\alpha$ to the diversity of \textit{An. nili}. In theory, SNPs with steep cline are hypothesized to be near genes involved in reproductive isolation and as such are possibly under selection in hybrids (Janoušek \textit{et al.} 2015). Therefore, the abundance of steep clines provides another clear illustration of the mosaic genome characterized by coexistence between high divergence and consistent gene flow observed in the hybrid species.

A substantial body of evidence indicates that genomic material coming from related species can confer an advantage to populations (adaptive introgression). Adaptive introgression of one or two loci has been widely studied over the last two decades and excellent examples have been described. The most prominent cases include the transfer of genes involved in mimicry of color patterns in \textit{Heliconius} butterflies (Consortium 2012), the circulation of resistant alleles of insecticides in mosquitoes (Clarkson \textit{et al.} 2014; Norris \textit{et al.} 2015) and rodenticide in mice (Song \textit{et al.} 2011). However, although convincing signatures of adaptive introgression around one gene or a few linked loci can now be described in an increasing number of species, the extent and the magnitude of introgressive hybridization across the genome remained unknown. Further, due in part to the fact that most species in which the concomitance of introgressive hybridization and high divergence has been observed are nonmodel species with little genomic resources, the knowledge of genomic characteristics of adaptive introgression in these species remains relatively modest. Meanwhile, a
consistent pattern has started to emerge from the few cases that have been studies with
substantial genomic details. In general, genomic regions exhibiting non-random
introgression are widely dispersed across the genome, rather than co-localized in a few
discrete blocks. This heterogeneity of genome-wide introgression patterns has been
observed for example between manakin birds (*Manacus candei* and *M. vitellinus*)
(Parchman *et al.* 2013) in different mice subspecies (Liu *et al.* 2015), and among
mosquitoes of the *Anopheles gambiae* species complex (Fontaine *et al.* 2014; Kamdem *et al.*
2016). Our results also show that selective introgression can be widespread across the
genome of two highly divergent species. However, in contrast to most of the reference cited,
LD analyses in *An. nili* and *An. ovengensis* have revealed a LD cluster separating hybrids
from the two parental species, which suggests that at least some of the recently
introgression loci consist of relatively large chromosome segments that have yet to be
further characterized sequentially and functionally with a high-quality reference genome.
In addition to the lack of knowledge about the genomic architecture, the functional and
phenotypic aspects of introgressive hybridization between established species remain
obsures. In mice for example, introgression is associated with a polarization of GO terms,
regions of elevated introgression exhibiting a disproportionate number of genes involved in
signal transduction and olfactory receptor genes (Janoušek *et al.* 2015). Hybrids between
*An. nili* and *An. ovengensis* were collected from an area of the equatorial forest whose
environmental features were not apparently very divergent from those of the locations
where the parent species were sampled (Fig. 1A). As a result, it is hard to pinpoint
environmental gradients and the life history traits that can be considered as the main drivers of speciation and introgression among species of *An. nili* and *An. ovengensis.*
Conclusions and implications

Although hybridization has been recognized as one of the major forces that affect the evolution of living species, the detailed study of its fundamental and applied implications has been hampered by methodological limitations. Advances in high-throughput DNA sequencing and statistical genomics are revolutionizing experimental and conceptual approaches, allowing a very sensitive examination of the heterogeneity of hybridization across species and genomes. We have used a combination of tests tailored to infer patterns of introgression across genomes of nonmodel species. Although our results still need to be replicated in other contact zones across the distribution range of An. nili in Africa, they highlight the complex relationships between divergence, selection and introgression during the split of taxa. Our work has methodological, conceptual and applied implications. Most genome scans assume a negative correlation between genetic divergence and introgression. It has been suggested that that the opposite is possible (Parchman et al. 2013; Gompert et al. 2013), but our findings do no support this hypothesis. Climate change and anthropogenic disturbance are contributing to expand geographic ranges of mosquito species worldwide thereby increasing contact between previously isolated species that are capable of exchanging gene flow. This interspecific gene flow in mosquitoes often leads to the spread of insecticide resistance alleles and other epidemiologically significant genes. Our work provided a methodological validation of a cost-effective population genomic approach that can be applied to investigate the bases of introgressive hybridization in other mosquito species.
Acknowledgements

Funding for this project was provided by the University of California Riverside and NIH grants 1R01AI113248 and 1R21AI115271 to BJW.
References


Genetics Resources, 4.


Molecular Ecology, n/a–n/a.


Purcell S, Neale B, Todd-Brown K et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.


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

**Table 1:** Pairwise $F_{ST}$ between *An. nili* and *An. ovengensis* populations. $p < 0.005$ for all values.

<table>
<thead>
<tr>
<th></th>
<th><em>An. nili</em> group 1</th>
<th><em>An. nili</em> group 2</th>
<th><em>An. nili</em> group 3</th>
<th><em>An. ovengensis</em> group 1</th>
<th><em>An. ovengensis</em> group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>An. nili</em> group 1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>An. nili</em> group 2</td>
<td>0.794</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>An. nili</em> group 3</td>
<td>0.838</td>
<td>0.863</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>An. ovengensis</em> group 1</td>
<td>0.655</td>
<td>0.834</td>
<td>0.873</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><em>An. ovengensis</em> group 2</td>
<td>0.857</td>
<td>0.863</td>
<td>0.858</td>
<td>0.896</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 2: Results of the most significant $f_{4}$ tests for gene flow.

<table>
<thead>
<tr>
<th>Test</th>
<th>$f_{4}$ ± std err</th>
<th>Z-score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{4}(\text{Oveng 1, Nili 3 ; Nili 2, Nili 1})$</td>
<td>-0.03506 ± 0.00287</td>
<td>-12.20</td>
<td>&lt; 0.00001</td>
</tr>
<tr>
<td>$f_{4}(\text{Oveng 1, Nili 3 ; Nili 2, Oveng 2})$</td>
<td>0.02241 ± 0.00263</td>
<td>8.52</td>
<td>&lt; 0.00001</td>
</tr>
<tr>
<td>$f_{4}(\text{Oveng 1, Oveng 2 ; Nili 3, Nili 1})$</td>
<td>-0.05775 ± 0.00355</td>
<td>-16.25</td>
<td>&lt; 0.00001</td>
</tr>
<tr>
<td>$f_{4}(\text{Oveng 1, Oveng 2 ; Nili 2, Nili 1})$</td>
<td>-0.03718 ± 0.00292</td>
<td>-12.72</td>
<td>&lt; 0.00001</td>
</tr>
<tr>
<td>$f_{4}(\text{Nili 3, Nili 1 ; Nili 2, Oveng 2})$</td>
<td>-0.02272 ± 0.00263</td>
<td>-8.63</td>
<td>&lt; 0.00001</td>
</tr>
</tbody>
</table>

Oveng 1: *An. ovengensis* group 1; Oveng 2: *An. ovengensis* group 2; Nili 1: *An. nili* group 1; Nili 2: *An. nili* group 2; Nili 3: *An. nili* group 3
Table 3: Pearson’s correlation coefficient assessing the relationship between cline parameter ($\alpha$ and $\beta$) and locus-specific estimates of five population genomic parameters (pairwise genetic divergence ($F_{ST}$ and $d_{xy}$), nucleotide diversity ($\theta_w$ and $\theta_\pi$) and allele frequency spectrum (Tajima’s $D$)).

<table>
<thead>
<tr>
<th></th>
<th>$F_{ST}$</th>
<th>$d_{xy}$</th>
<th>$\theta_w$</th>
<th>$\theta_\pi$</th>
<th>TD</th>
<th>$\theta_w$</th>
<th>$\theta_\pi$</th>
<th>TD</th>
<th>$\theta_w$</th>
<th>$\theta_\pi$</th>
<th>TD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>-0.151</td>
<td>-0.130</td>
<td>0.057</td>
<td>0.052</td>
<td>0.016</td>
<td>*</td>
<td>-0.498</td>
<td>-0.414</td>
<td>0.088</td>
<td>0.537</td>
<td>0.432</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.060</td>
<td>0.048</td>
<td>-0.016</td>
<td>*</td>
<td>-0.017</td>
<td>*</td>
<td>-0.011</td>
<td>*</td>
<td>0.230</td>
<td>0.188</td>
<td>-0.047</td>
</tr>
</tbody>
</table>

* not significant (p>0.005)

TD: Tajima's $D$
Figures

Figure 1: Population genetic structure of An. nili sensu lato inferred from 4343 SNPs identified with a reference-based assembly. (A) Map showing the sampling locations and relative frequencies of the two An. nili and An. ovengensis. (B) ADMIXTURE plots with k from 2 through 8. (C) and (D) neighbor-joining tree and PCA. Each PCA axis is labeled with the percentage of variance explained. (E), (F) and (G) Identification of the optimal number of genetic clusters using the delta k method of Evanno et al, DAPC, and 10-fold cross-validation in ADMIXTURE. The lowest BIC and CV error and the highest delta k indicate the most probable number of genetic clusters.
**Figure 2.** TreeMix Maximum Likelihood (ML) trees depicting the signals of gene flow between *An. nili* and *An. ovengensis*. ML tree and residual fit from the ML model inferred with (A) no migration edge, (B) a single migration edge and (C) two migration edges. The small arrow on each indicates the directionality of gene flow migration edge and the color of the edge reflect the intensity of admixture. Heat colors depict the residual covariance between each pair of populations. Darker colors indicate populations more closely related to each other than expected under a bifurcating maximum likelihood tree, suggestive of gene flow.
Figure 3: Frequency distribution of $F_{ST}$ (A) and $d_{xy}$ (B) based on 9622 variant sit...
Figure 4: Results of Linkage disequilibrium analyses in LDna. (A) LDna graph suggesting the presence of 6 LD clusters (Single Outlier Clusters (SOCs)) based on 1330 SNPs in a dataset containing *An. nili*, *An. ovengensis*, and hydrids of the two species. Values of the two parameters: \( \varphi \) (which controls when clusters are defined as outliers) and \( |E|_{\min} \), the minimum number of edges required for a LD cluster to be considered as an outlier, are indicated on top of the graph. Corresponding LD thresholds are shown on the x-axis. (B) Population genetic structure of the six SOCs identified.
A

$q=3$, $|E_m|=160$

B

SOC 31_0.96 (383 SNPs)

SOC 32_0.96 (39 SNPs)

SOC 68_0.84 (112 SNPs)

SOC 10_1 (42 SNPs)

SOC 77_0.82 (34 SNPs)

SOC 81_0.8 (30 SNPs)

A. nili

An. ovengensis

Hybrids
Figure 5: Genomic cline analysis. (A) Frequency distribution of the hybrid index in admixed individuals (hybrid index of pure An. nili = 1.0 and pure An. ovengensis = 0.0). (B) Prevalence of steep genomic clines illustrated by a scatterplot between the cline parameters $\alpha$ and $\beta$. 

![Graph A](image1.png)  

![Graph B](image2.png)
Supplemental Material

Table S1: Information on An. nili sensu lato mosquitoes 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>Ebebda</td>
<td>4°20'00&quot;N, 11°17'00&quot;E</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Nkoteng</td>
<td>4°31'00&quot;N, 12°02'00&quot;E</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Nyabessan</td>
<td>2°24'00&quot;N, 10°24'00&quot;E</td>
<td>63</td>
<td>44</td>
</tr>
<tr>
<td>Mbéhé</td>
<td>4°10'00&quot;N, 11°04'00&quot;E</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>76</td>
<td>47</td>
</tr>
</tbody>
</table>

HLC-OUT, human landing catches performed outdoor; HLC-IN, human landing catches performed indoor; LC, larval collection.
**Table S2**: Pairwise $F_{ST}$ estimated from a *de novo* assembly. $p < 0.005$ for all values.

<table>
<thead>
<tr>
<th>$F_{ST}$</th>
<th><em>An. nili</em> group 1</th>
<th><em>An. nili</em> group 2</th>
<th><em>An. nili</em> group 3</th>
<th><em>An. ovengensis</em> group 1</th>
<th><em>An. ovengensis</em> group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>An. nili</em> group 1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>An. nili</em> group 2</td>
<td>0.791</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>An. nili</em> group 3</td>
<td>0.844</td>
<td>0.862</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>An. ovengensis</em> group 1</td>
<td>0.705</td>
<td>0.861</td>
<td>0.902</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><em>An. ovengensis</em> group 2</td>
<td>0.854</td>
<td>0.862</td>
<td>0.867</td>
<td>0.907</td>
<td>-</td>
</tr>
</tbody>
</table>
**Table S3:** Results of the most significant $f_4$ tests for gene flow (*de novo* assembly).

<table>
<thead>
<tr>
<th>Test</th>
<th>$f_4 \pm \text{std err}$</th>
<th>Z-score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_3$(Oveng 1, Nili 3; Nili 2, Nili 1)</td>
<td>-0.03692 ± 0.00360</td>
<td>-10.25</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>$f_3$(Oveng 1, Nili 3; Nili 2, Oveng 2)</td>
<td>0.02787 ± 0.00351</td>
<td>7.94</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>$f_3$(Oveng 1, Oveng 2; Nili 3, Nili 1)</td>
<td>-0.06610 ± 0.00451</td>
<td>-14.64</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>$f_3$(Oveng 1, Oveng 2; Nili 2, Nili 1)</td>
<td>-0.04053 ± 0.00368</td>
<td>-11.01</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>$f_3$(Nili 3, Nili 1; Nili 2, Oveng 2)</td>
<td>-0.02776 ± 0.00349</td>
<td>-7.96</td>
<td>&lt;0.00001</td>
</tr>
</tbody>
</table>

Oveng 1: *An. ovengensis* group 1; Oveng 2: *An. ovengensis* group 2; Nili 1: *An. nili* group 1; Nili 2: *An. nili* group 2; Nili 3: *An. nili* group 3
**Figure S1:** Population genetic structure of An. nili sensu lato inferred from 3071 SNPs identified with a *de novo* assembly. (A) ADMIXTURE plots with k from 2 through 8. (B) and (C) neighbor-joining tree and PCA. (D), (E) and (F) Identification of the optimal number of genetic clusters using the delta k method of Evanno et al, DAPC, and a 10-fold cross-validation in ADMIXTURE.
**Figure S2:** *TreeMix* Maximum Likelihood (ML) trees estimated from 3071 SNPs identified with a *de novo* assembly. ML tree and residual fit from the ML model inferred with (A) no migration, (B) a single migration and (C) two migration edges. See Fig. 2. in the main text for additional description.