Supplemental text

Population genetic structures on collinear and inverted regions

667 *Anopheles gambiae* complex mosquito samples were collected at six sites in Cameroon (Fig. S1, coordinates listed in Table S1). They represent gradient changes in habitats (from forest, wet to dry savanna), which boast the highest diversity of polymorphic inversions within the species (previous collection data from PopI database: https://grassi2.ucdavis.edu/). Molecular identification showed that proportions of *An. arabiensis* increase from south to north. Only 6 *An. colluzzi* were found (previously known as the M form of *An. gambiae* [1]) in Mbe, while the rest were *An. gambiae* (previously known as the S form of *An. gambiae*). We selected 259 *An. gambiae* individuals (40-60 individuals with good DNA qualities per population except for Bankim, Table S1) and 8 *An. arabiensis*, prepared individually-barcoded double digest Radseq libraries [2], and generated two lanes of 100bp paired-end sequencing reads on Illumina HiSeq2000 platform.

After filtering for unmapped or ambiguously mapped reads and loci with low coverage per sample or low presence across samples, a total of 25,966 loci were mapped onto Chromosome 2, 3 and X. *An. gambiae* has a nucleotide diversity (π) of 0.01024 ± 4.0E-5, while π of *An. arabiensis* is 0.00888 ± 4.7E-5. One SNP per locus with at least 1000bp between them were selected for PCA analysis to detect population genetic structure of the species. When PCA were performed on all individuals in collinear regions, we found three clusters of individuals consistently across all chromosomes: 1) one big cluster of individuals from all populations; 2) one cluster of individuals from subset of Mbakaou (blue dots in Fig. S2); 3) *An. arabiensis* as a separate cluster (red dots in Fig. S2). When the latter two clusters of individuals were excluded, PCA showed no apparent geographic structures among any populations (Fig. S3) and DAPC analysis had highest support for one group (K = 1) of all individuals. It is intriguing to find a very distinctive population within molecularly-identified *An. gambiae* individuals in Mbakaou. However, since our goal is to identify inversion associated selected regions rather than population specific effects, we excluded these individuals for the following analysis. Contrary to collinear regions, when individuals are clustered by SNPs from 2La (2L: 20524058-42165532) or 2Rb (2R: 19023925-26758676) regions, three clusters of individuals can be seen from first PCs (explaining 20.3% and 11.9% of the total variance respectively, Fig. 1a,d). DAPC results supported K=3 as the most likely number of genetic clusters. When compared to molecular karyotyping results, the three clusters match nicely to inverted homokaryotypes (I/I hereafter), heterokaryotypes (I/S) and standard homokaryotypes (S/S) (Fig. 1c,f).

References

1. Coetzee M, Hunt RH, Wilkerson R, Della Torre A, Coulibaly MB, et al. (2013) Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. Zootaxa 3619: 246-274.

2. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one 7: e37135.

Supplemental Figure Legends

Figure S1. Sampling locations and species composition of *Anopheles gambiae* species complex. The area of each pie chart correspond to the sample size. Map color from blue to red stands for humid to dry areas.

Figure S2. Principle component analyses using SNPs in different collinear genomic regions. Color of the dots represent different populations. Red dots are individuals of *An. arabiensis*.

Figure S3. Principle component analyses of *An. gambiae* using SNPs in different collinear genomic regions. Color of the dots represent different populations.

Figure S4. Principle component analyses of *Anopheles gambiae* using SNPs from 2La and 2Rb. Left and Right panels are the result for 2La and 2Rb, respectively. Top panel is the result for PCA clustering of individuals from different populations. Middle panel finds the best number of clusters based on BIC scores. The bottom panel shows how divergent each cluster is from each other on the discriminant function space.

Figure S5. Comparison of detection power and false-discovery rates between two selection detection methods. Upper panel, 2La region; lower panel, 2Rb region. a), c), e), and g) are the results for the scenario in which selection occurs on the branch of *S*. b, d), f), and h) are the results for selection on the branch of *I*. The *x*-axis presents cases where the proportion of selected loci consists of 1%, 5%, and 10% out of all simulated loci (i.e., rest of the loci are simulated under neutral scenario). All selected loci are generated with $s = 0.001$. See methods for the details of how rates are calculated. The inset shows the false-positive rates for each method.

Figure S6. Comparison of detection power and false-discovery rates between two selection detection methods. Upper panel, 2La region; lower panel, 2Rb region. a), c), e), and g) are the results for the scenario in which selection occurs on the branch of *S*. b, d), f), and h) are the results for selection on the branch of *I*. The *x*-axis presents cases where the proportion of selected loci consists of 1%, 5%, and 10% out of all simulated loci (i.e., rest of the loci are simulated under neutral scenario). All selected loci are generated with $s = 0.0001$. See methods for the details of how rates are calculated. The inset shows the false-positive rates for each method.

Table S1. Collection sites, coordinates and sampling sizes of *Anopheles gambiae*.

Population	Latitude	Longitude	Sample size
MBE	7.78	13.55	89
NGOUNDERE	7.48	13.55	74
MEIGANGA	6.55	14.26	60
MBAKAOU	6.37	12.76	63
BANKIM	6.05	11.40	4
BAFOUSSAM	5.48	10.59	106

Table S2. Power of differentiating selection form drift using average summary statistics of loci in a 5kb segment with/without selected locus inside.

Inversion	True Scenario	Prediction			Correct Assignment % (Power)
		Neutral	Selection in I and Arab	selection in S	
2La	Neutral	936	31	33	93.6%
	Selection in I and Arab	65	927	8	92.7%
	selection in S	56	2	942	94.2%
2Rb	Neutral	981	9	10	98.1%
	Selection in I and Arab	72	927	1	92.7%
	selection in S	70	0	930	93.0%