nature portfolio

Corresponding author(s): Alessandra della Torre

Last updated by author(s): Aug 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Full genomes of 1,190 mosquito females collected at 35 sites from 7 different countries in West Africa provided by MalariaGEN. Anopheles gambiae 1000 Genomes Project Phase-3 data resource. Details of the samples and associateciated references are provided in the Agl000G partner studies page (https://malariagen.github.io/vector-data/studies-ag1000g.html).
Data analysis	Python 3.10; ADMIXTURE 1.3.0; scikit-allel 1.3.7; TreeMix; dadi; coding notebooks and custom scripts with their references are reported in the paper' github: https://github.com/randomxsk8/caputo-et-al-2 024

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Accession codes far each individuai analyzed are reported on the Supplementary Data 1

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not Applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not Applicable
Population characteristics	Not Applicable
Recruitment	Not Applicable
Ethics oversight	Not Applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The aim of this work was to understand the actual nature of 'far-west' (e.g. FW) populations focusing on their relationships with west-central An. coluzzii and An. gambiae populations. We took advantage of high-throughput whole genome sequence data of 1,190 adult females made available by the Ag1000G Project project (https://www.malariagen.net/project/ag1000g/). The Ag1000G Project data resource includes high-confidence single-nucleotide polymorphism (SNP) calls at 57 million variable sites, genome-wide copy number variation (CNV) calls, and haplotypes phased at biallelic SNPs. We performed population genomics analyses in order to exploit: i) genetic variation among west african populations; ii) nucleotide diversity and divergence in FW; iii) patterns of genomic divergence and admixture in FW; iv) demographic history in FW; v) putative patterns involved in the target-site and metabolic insecticide resistance.
Research sample	The sample data set utilised in the present study includes the genome sequences and associated metadata of 1,190 mosquito females collected at 35 sites from 7 different countries in West Africa released by the Ag1000G Project. Sequence data utilised to explore the genetic variation and structure are available from the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home). ENA accession numbers for the specific samples and sequencing runs used in this study are provided in the Supplementary Table 1. Further information on availability of these data is available from the MalariaGEN website at https://malariagen.github.io/vector-data/ag3/ag3.0.html
Sampling strategy	Only individuals from sampling locations with sample sizes >5 are included in the analysed dataset to provide the best representative sample size per location. SNPs on the three chromosomes (e.g. 2-3-X) were filtered by the "gamb_colu" site filters as defined by the Ag1000G Project. Different SNPs set have been generated: i) for all chromosomes a variant set was generated by selecting SNPs with a minor allele frequency ≥1% and randomly downsampling to 100,000 SNPs per chromosomal arm, then a linkage disequilibrium was applied by excluding variants SNPs above an r2 threshold of 0.01 in moving windows of 500 SNPs with a step size of 250 SNPs; ii) for chromosome-3, a second variant set was generated by including 34,707,855 SNPs located in the euchromatic region (3R: 1-24 Mbp; 3L: 15-41 Mbp), which is recognized to provide the most coherent view of population structure due to the absence of polymorphic inversions; iii) in order to carry on the analysis of genomic diversity and divergence across the whole genome, SNPs were chosen by selecting a total of 191,422,813 biallelic SNPs on all chromosomes.
Data collection	Specimens were collected in the frame of the Ag1000G Consortium and other researches on the An. gambiae complex. Detailed descriptions of sampling methods and studied included are reported on the following link: https://malariagen.github.io/vector-data/studies-ag1000g.html. Whole-genome sequencing was conducted at the Wellcome Sanger Institute, using both Illumina HiSeq 2000 and HiSeq X platforms, targeting a coverage of 30× per individual. Library preparation involved paired-end multiplexing with 100–200 bp insert sizes. Alignment to the reference genome (AgamP4) and SNP calling were performed using BWA and GATK, respectively. SNP calling included indel realignment and coverage was capped at 250×. The pipelines were developed by the MalariaGEN Resource Centre and Broad Institute Data Engineering Team, with open-source implementations available on GitHub. Additional details regarding thresholds and detailed information regarding read alignment, SNP calling, CNV calling and haplotype phasing are detailed on the Ag1000G website (https://malariagen.github.io/vector-data/ag3/methods.html).
Timing and spatial scale	We investigated on the West Sub-Saharan African region were the two species An. gambiae and An. coluzzii are both present. Individuals were collected by the Ag1000G Consortium between 2010 and 2014 with some individuals collected in Mali sampled in

	2004.
Data exclusions	Analyses were performed considering the biallelic position along the genome reported on the our sample populations (PCA, ADMIXTURE, TreeMix, F3, Patterson's D, SFS, dadi, genome-wide FST and diversity statistics such as Dxy, Tajima's D, pi, Watterson's Theta. Some analyses were carried out using specific regions of the genome such as: i) regions associated to chromosomal inversion 2Rd (33.57-41.36 Mbp) and 2La (20.52-42.16 Mbp) - whose frequencies characterize the so called BISSAU chromosomal form (PCA); ii) euchromatic region of chromosome-3 (located at 3R-arm = 1-24 Mbp and 3L-arm = 15-41 Mbp), which is recognized to provide the most coherent view of population structure due to the absence of polymorphic inversions (PCA, ADMIXTURE, TreeMix); iii) centromeric chromosome-X region, which is known as the CO and GA speciation island, implicated in the species divergence (PCA).
Reproducibility	Code provided at the github repository (https://github.com/randomxsk8/caputo-et-al-2024) is available and reproducible.
Randomization	In order to facilitate the representation of genomic regions and utilize population genomics tools which can be CPU-intensive such as PCA, ADMIXTURE, and Treemix analysis, we performed for each chromosome (e.g. 2-3-X) a downsampling to 100k by selecting random SNPs and then we selected those SNPs for analysis using the ld pruning methodology with an r2 threshold of 0.01 in moving windows of 500 SNPs with a step size of 250 SNPs.
Blinding	NA

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	No laboratory animals were used in the work
Wild animals	The study focuses on Afrotropical mosquitoes of the Anopheles gambiae complex
Reporting on sex	Not Applicable
Field-collected samples	Mosquito specimens were collected in the frame of the Ag1000G Consortium. Detailed descriptions of sampling methods are reported in the following link: https://malariagen.github.io/vector-data/studies-ag1000g.html
Ethics oversight	Not Applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	Not Applicable
Novel plant genotypes	Not Applicable
Authentication	Not Applicable