Genome variation and population structure in three African malaria vector species within the *Anopheles gambiae* complex

This manuscript (permalink) was automatically generated from malariagen/ag1000g-phase3-data-paper@a583aa8 on February 1, 2021.

Authors

- **The Anopheles gambiae 1000 Genomes Consortium**

Abstract

Population Sampling

The third and final phase of the Ag1000g project data resource contains wild-caught *Anopheles* mosquito genomes from Sub-Saharan Africa, collected from a total of 124 sites across 19 countries, 6 of which are novel.

Collections from Mali increase the density of coverage in West Africa, Central African Republic and Democratic Republic of Congo begin to fill the gap previously present in Central Africa while Malawi, Mozambique and Tanzania provide much more power to analyse East African malaria vectors, including *A. arabiensis* an important vector species not previously sequenced in the project.

Alongside sampling from natural populations, we include colony individuals from a number of laboratory crosses, comprising 11 crosses that were released as part of phase 2, and 4 additional pedigrees.

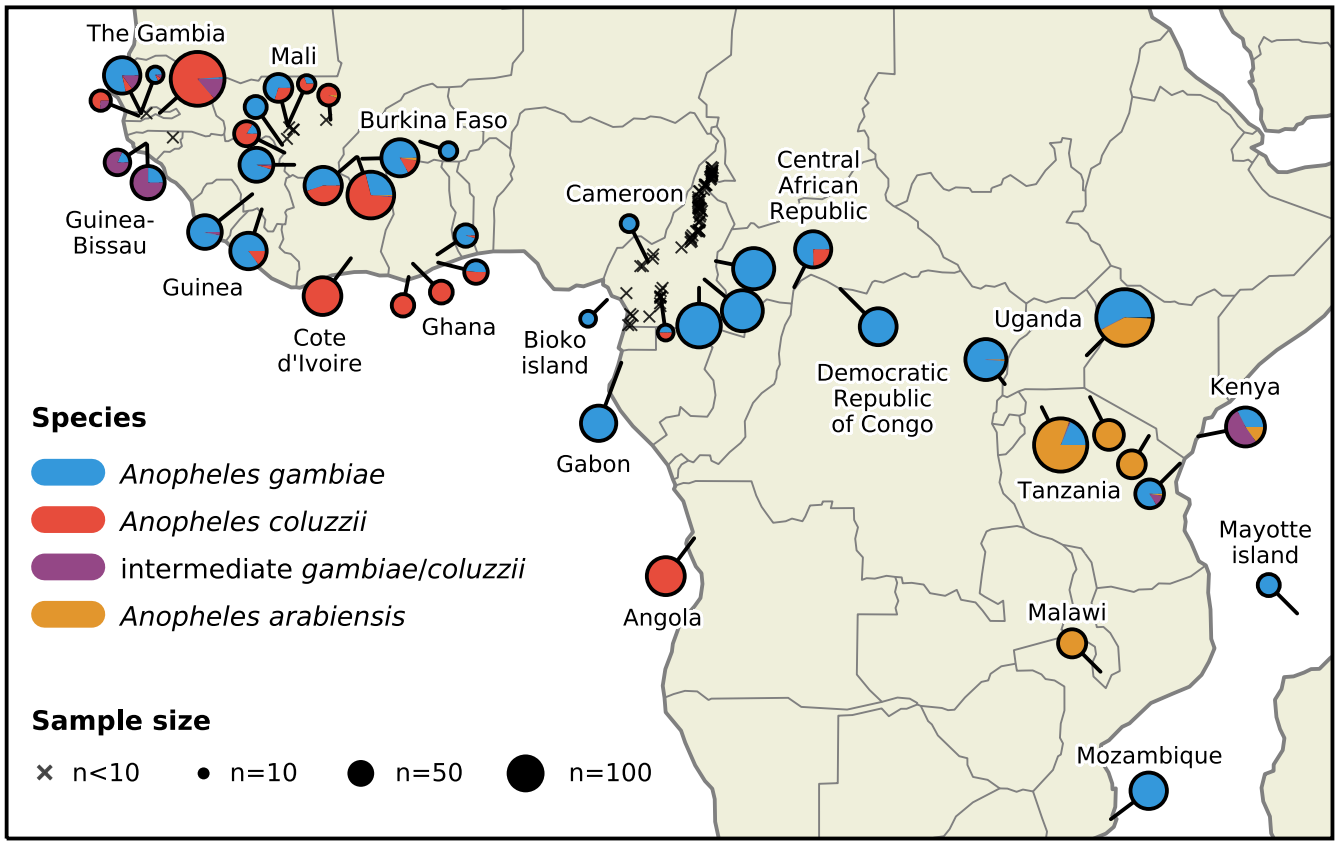

Figure 1: Sample Collection Map

Whole Genome Sequencing and Alignment

4,693 individual mosquito genomes were sequenced on either Illumina HiSeq2000 (n=3,130) or Illumina HiSeqX (n=1,563) to a target coverage of 30X.

Between machine types the median number of bases sequenced per sample was 9.76Gb and 10.33Gb respectively, representing a difference in yield (two-tailed mann-whitney U $p < 0.0001$).

These values correspond to a yield per reference base (vs AgamP4) of 35.76X and 37.82X.

91.9% of HiSeqX runs and 80.5% of HiSeq2000 runs met the target yield of 30X.

Reads were aligned to the AgamP4 reference and Single Nucleotide Polymophisms (SNPs) called using GATK UnifiedGenotyper.

All samples successfully completed the pipeline and entered the sample quality control (QC) process.

Sample QC

For wild-caught samples (n=XXXX), the QC process was composed of three stages, sequence quality assurance, replicate handling, and anomaly detection.

A total of 668 samples were removed where sequencing was of insufficient quality to accurately call genotypes across the whole genome.

Exclusions were due to poor coverage (n=410), potential contamination (n=229), and an ambiguous sex call (n=29).

Where technical replicates were available, we excluded 4 pairs (8 samples) with low genotype concordance.

Where pairs met the concordance threshold we excluded the lower quality sample (n=407).

Samples were also screened pairwise within submission sets for unexpected pairs, though none were detected.

The third stage used principal component analysis (PCA) to identify and exclude individual samples that were outliers based on available metadata.

A review process identified samples that could not be explained parsimoniously, and were therefore likely to be sample mix ups or instances of mislabelling.

28 samples were excluded as they respectively dominated the first principal components, indicating high divergence from all other samples and therefore likely members of other Anopheline species.

A further 82 samples were excluded as potential sample mix ups.

Following all sample QC steps, 3,483 wild-caught samples (74.2%) were retained from the original cohort for analysis.

This represents an additional 1,823 mosquitoes relative to the phase 2 release.

Due to a change in assessment of sample quality where technical replicates are available, the preferred replicate was changed for 172 mosquitoes between phase 2 and phase 3.

9 samples included in phase 2 are not present in this release (sup XXX).

The AG1000G-X submission set, made up of laboratory experimental crosses, was subject to a slightly different QC process.

Firstly an analysis based on rates of Mendelian error identified true fathers of crosses (where multiple males were introduced to cages), and validated provided pedigrees.

Of the 7XX samples provided we were able to validate 15 crosses to a high level of confidence, comprising 299 samples.

4 of these crosses are novel relative to phase 2.

These samples went through a modified sequence quality assurance process, where 1 sample was removed for insufficient coverage (methods).

The final data release therefore comprises 3,XXX samples, XXX from laboratory crosses, and YYY wild collected samples.

Coverage

%% TO DO %% (PLOTS DONE, but numbers needed).

Summary of site coverage post QC exclusions.

- ie what frac of the genome is at 1X median
- what frac at 10X.
- What frac of exome
- what frac of X

At this point we do not mention arabiensis.

Species assignment and sex calling

The Anopheles gambiae complex is a crypic group of sibling species, with no single locus offering unambiguous resolution of species.

To identify species we looked beyond the conventional set of PCR based markers and applied a wider set of ancestry informative markers (AIMs).

Species were not assigned to samples from laboratory colony crosses due to inbreeding and high levels of genetic drift.

To distinguish *A. arabiensis* from *A. gambiae s.l* a set of novel markers was derived from data from the 16 genomes project (ref).

Using cut offs based on agreement with the established PCR marker, 368 individuals were classed as *A. arabiensis* and 2415 as *A. gambiae s.l*.

A single individual collected in Tororo, Uganda is classed as intermediate- given the majority (XX%) of AIM SNPs in the genome are heterozygous between the gambiae-like and arabiensis-like alleles, this individual is likely to be an F1 hybrid.

To resolve the *A. gambiae s.l* individuals as *A. gambiae* and *A. coluzzii* we applied 729 AIMs previously identified by Neafsey et al (ref), and used in previous analyses of Ag1000G data. (ref paper2, paper1).

Of the 2415 *A. gambiae s.l* individuals, 1571 were called as *A. gambiae s.s*, 675 as *A. coluzzii* and 169 as intermediate (ref collection map).

Many intermediate samples were sampled from the Western coast of West Africa (particularly The Gambia and Guinea Bissau), and given distinct populations of *A. gambiae s.l.* and *A. coluzzii* are also found in this region, this result highlights the complexity of species relationships here.

Additionally a number of intermediate samples were identified in coastal populations of East Africa, in Kilifi Kenya, and Muleba Tanzania.

%% TODO This analysis It is established that species barriers between members of the gambiae complex are porous, and numerous instances of introgression associated with selection have been observed in West Africa. (ref clarkson + li, others?).

We observe known introgression from gambiae to coluzzii of the kdr allele in West Africa.

In West African coluzzii populations, presence of gambiae-like alleles at this locus reach 95%.

However no introgression is obseved in Angola, or CAR.

%% TODO What about other loci

%% TODO Method to id these regions. Simply just plot frequency of gambiae allele in coluzzii samples? No clear introgression is observed between gambiae and arabiensis.

%% TODO ADD AIM FIGURES

SNP filtering and quality

Site filtering is necessary to ensure that reported variation is of high quality.

Features of specific regions of the Anopheles genome contribute to calling errors in short-read technologies; such features include regions of high divergence from the reference, high homology between regions, copy number variation, or the presence of transposable elements.

Where previously we have used manually curated cutoffs based on observed mendelian error rates to filter sites (ref phase1, phase2), here we built a statistical model where cohort level summary statistics were used to identify sites likely to contain genotyping errors.

Using the 15 available Anopheles pedigrees previously described, we used the presence of mendelian error at sites as a proxy for genotype discordance.

10 of the 15 crosses were used to train the model while 5 were held out for validation.

Each of the 5 pedigrees represent independent evaluation sets.

Before applying the site filters, the false discovery rate (FDR) of the 5 crosses over all autosomal sites ranged between 0.74% and 1.10% (table XXX).

The application of the site filters defines the accessible fraction of the autosomes at 72.58%, and the range of false discovery rates is 0.04% to 0.10%.

The median fold change of FDR was -3.71.

On the hemizygous X chromosome we used the more direct measure of heterozygote calls in males to ascertain mendelian error.

In the dataset are 220 *A. gambiae s.l* male samples, each of which represent an independent proxy for genotype discordance.

Pre-application of the site filters, subject to a Genotype Quality (GQ) threshold of 30, the median heterozygosity rate was 0.244%, and post filtering this drops to 0.023% (table XX).

The median fold change in error rate was -3.33, with 69.97% of the X chromosome passing site filters.

The new model based method represents a marked improvement over the site filters generated as part of phase 2; all 5 evaluation pedigrees showed a modest reduction in FDR, but the higher rate of accessibility in this release (72.58% vs 62.05%) resulted in an significant improvement in Youden score (Table XXX) across autosomes.

The X chromosome showed a similar pattern, the median heterozygosity rate in phase 2 is similar to the new site filters (0.028%), but the higher accessibility in the updated filter set (69.97% vs 62.46%) yields improved sensitivity.

As genomic features vary between species, different sets of site filters were generated to allow high quality analyses both within and between species.

The `gamb_colu` site filters were generated as above, and are appropriate for analyses that include *gambiae* and *coluzzii* samples only.

%% TODO Add accessibility of other site filters. The `arab` site filters were generated following application of the model to the summary statistics from arabiensis samples in the cohort (n=XXX), this set of site filters are appropriate when working with *A. arabiensis* samples only.

Finally, the `gamb_colu_arab` site filters allow analyses across all three species and are the intersection of the `gamb_colu` and the `arab` site filters.

Place holders for tables.

- Table A: Mendel errors per cross per autosome. row indices: chromosome and raw/filtered column indexes: crosses + frac accessible. ie 8 rows, and 6 columns.

- Table B: comparison of cross and X row indices: raw/filtered column indices: MER, frac accessible, Youden, each for 2 and 3. column indexes: crosses + frac accessible. ie 2 rows, and 6 columns.

```
---
caption: 'Result of heterozygote calls on male X chromosome'
alignment: LLLLLLLL
include: content/tables/mer_X.csv
csv-kwargs:
  dialect: unix
width: [0.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]
---
```

Genome accessibility

We define accessibility as the fraction of sites in a region passing the appropriate set of site filters.

Overall, 70% of the genome, and ??% of the exome are considered accessible in the `gamb_colu` set.

This is an improvement from phase 2, where XXX of the genome, and YYY of the exome was considered accessible.

As expected, accessiblity was generally lower around the centromeres, and in regions of heterochromatin (table ref).

One notable region of low accessibility spans 40-41Mbp of chromosome 3R, this corresponds to ??.

Accessibility of the `arab` site filters closely follows that of `gamb_colu`, with the exception of the X chromosome where we see substantially lower values.

This appears to be driven by high divergence between AgamP4 and our *A. arabiensis* samples, particularly around the Xag inversion at Q-Q Mbp, ref figure.

On the autosomes the divergence from the reference is comparable between *A.arabiensis* and *A. gambiael A. coluzzii* samples, suggesting a strong basis for comparison accross species.

The median divergence (Dxy) of 100kbp windows is XXX (5%/95% TTT/SSS) for gambiae/coluzzii and YYY (TTT/YYY) for arabiensis.

On the X chromosome these values are XXX (/) for gambiae/coluzzii and YYY (/) for arabiensis.

SNP discovery

Overall, we report XX,XXX,SSS single nucleotide polymophisms (SNPs) segregating in this cohort that pass filters, of which XX,XXX (%) are multiallelic.

12,223 SNPs are segregating in both species groups, while XXX are private to gambiae_/coluzzii_ and YYY to arabiensis [fig ref].

This phase of the study reports an additional XXX SNPs from phase 2.

In XXX gambiae and coluzzii individuals we report 12,222,222 SNPs (Q% multiallelic), corresponding to a SNP every 1.6 accessible bases.

In XXX arabiensis individuals we identify 10,000,000 SNPs (Q% multiallelic), a SNP every 2.5 accessible bases.

Population Structure

Genome wide patterns

%% TODO NOtes currently.

Re-introduce key idea of structure being different across the genome.

How does arabiensis fit into this? Are there regions of the genome where arabiensis ancestry is secondary?

PCA / UMAP

To highlight population structure we performed principal component analysis across all wild-caught samples in the dataset.

To avoid confounding of structure in genomic regions including paracentric inversions, extremely low diversity and regions under strong selection, we limited our analysis to euchromatic regions of chromosome 3L.

The most apparent signal in the dataset is PC1 clearly being driven by Arabiensis, with clear separation of arabiensis samples from gambiae/coluzzii.

The apparent hybrid sits between gambiae and arabiensis samples.

To view population structure within gambiae/coluzzii and arabiensis more independently, we performed subsequent PCA analysis Arabiensis and gambiae/coluzzii individuals separately.

Population structure between gambiae and coluzzii is significantly more complex.

Separately between species. What are the major findings?

- Arabiensis drives PC1.
- East Africa: Seems to be clear population structure between *gambiae* in KE and TZ.
- According to AIM analysis, a significant proportion of samples in these groups are classed as IM between gambiae. Certainly not coluzzii, but some kind of complex ancestry.
- Relevance to TENEGLRA
- West Africa- in far west Africa we see intermediate population. Not gambiae coluzzii, unlikely to be hybrids, but a related subspecies.

Interestingly seems to be stable in the presence of both gamb and colu. Although they sit close to col in the PCA they are distinct from coluzzii, given they are founf at the same site.

Genetic Diversity within sampling sites

Better to avoid use of population.

Using species groupings above, i.e. PCA clusters of samples not clearly gambcolu, but sympatric with them are classed as intermediate.

First look at diversity at a regional level within species. ie gambiae is more diverse in west than east africa. Central?

Coluzzii is similar within its range.

Arabiensis only found in EA, but do we see differences in diversity?

Justification of using wattersons theta.

THEN, we can start to speak about differences between species, within regions.

West African gambiae have higher diversity than coluzzii.

Then how do west african intermediate compare to these?

In east africa, we compare gambiae to arabiensis.

Insecticide Resistance

- kdr frequencies in different sampling groups
- we don't have CNVs... so? We can use markers?

Gene Drive

- repeat of phase 2 analysis.

Methods

Population Sampling

Summary

The Ag1000G project is coordinated by a consortium of partners from a range of different research institutions and countries. This includes consortium members who are carrying out independent research studies in malaria endemic regions, and who have contributed mosquito specimens or mosquito DNA samples collected in the course of their own research. The methods presented here describe the studies that have contributed samples to phase 3 of the Ag1000G project, including wild-caught samples from 19 African countries. This section also provides information about the collection locations and methods, the people involved in the studies, and references to any published articles providing further information about the studies. Throughout this document we use species nomenclature following Coetzee *et al.* [1]. Unless otherwise stated, the DNA extraction method used for the collections described below was Qiagen DNeasy Blood and Tissue Kit (Qiagen Science, MD, USA).

Angola

Sample sets

`AG1000G-AO`.

Study information

Adult mosquitoes were obtained by rearing larvae collected from breeding sites along the main roads connecting the municipalities of Kilamba-Kiaxi and Viana, Luanda province **(-8.821,13.291)**, in April/May 2009. These are peri-urban areas where malaria reaches hyperendemic levels. All specimens collected in the study area were typed as *A. coluzzii* [2] although *An. melas* and *A. arabiensis* have also been recorded in the province [3,4]. Specimens were stored on silica gel and DNA extraction was performed by a phenol-chloroform protocol described in [5].

Contributors

- Arlete D. Troco
  - Programa Nacional de Controle da Malária, Direcção Nacional de Saúde Pública, Ministério da Saúde, Luanda, Angola.
- João Pinto (jpinto@ihmt.unl.pt)
  - Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, UNL, Rua da Junqueira 100, 1349-008 Lisbon, Portugal.

Burkina Faso (1)

Sample sets

`AG1000G-BF-A`, `AG1000G-BF-B`

Study information

The Target Malaria project contributed samples from collections made in three villages separated by at most 30km: Bana **(11.233, -4.472)**, Souroukoudinga **(11.235, -4.535)** and Pala **(11.150, -4.235)**. These collections were made in July-August 2012, July and October 2014, and January, February and April 2015. The area is agricultural, with rice-growing areas near Bana and Souroukoudinga, and a large mango grove near Pala. Female mosquitoes were collected by human landing catch, pyrethrum spray collection or aspiration. Males were collected by swarm netting. Both *An. gambiae* and *An. coluzzii* [6] were collected. Specimens were stored in 80% ethanol and DNA was extracted using the DNeasy Tissue Kit (Qiagen) or using a simple CTAB method.

External resources

- **https://targetmalaria.org/**

Contributors

- Abdoulaye Diabaté
  - Institut de Recherche en Sciences de la Santé (IRSS), Bobo Dioulasso, B.P.545, Burkina Faso.
- Patric Stephane Epopa
  - Institut de Recherche en Sciences de la Santé (IRSS), Bobo Dioulasso, B.P.545, Burkina Faso.
- Franck Yao
  - Institut de Recherche en Sciences de la Santé (IRSS), Bobo Dioulasso, B.P.545, Burkina Faso.
- Samantha O'Loughlin (s.oloughlin@imperial.ac.uk)
  - Department of Life Sciences, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, UK.
- Austin Burt
  - Department of Life Sciences, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, UK.

Acknowledgements

We would like to thank the technicians of the Institute de Recherche en Sciences de la Santé/Target Malaria Burkina Faso, including Guel Hyacinthe, Diabate Brama, Ilboudo Seni, Kabre Rasmane, Diabate Noufou and Yeye Pascal, for their contributions to sample collections.

Burkina Faso (2)

Sample sets

`AG1000G-BF-C`

Study information

Samples were contributed from collections of indoor resting adults made by spray catch from Monomtenga in central Burkina Faso **(12.06, -1.17)**. These specimens were sorted morphologically to **An. gambiae s.l.**. Ovaries of half-gravid females were dissected and placed in numbered individual micro-tubes containing modified Carnoy's solution (1:3 glacial acetic acid: 100% ethanol). Carcasses were placed in correspondingly numbered micro-tubes over desiccant. Genomic DNA was isolated from individual mosquitoes using one of the following: DNeasy Extraction Kit (Qiagen, Valencia, CA), Puregene kit (Gentra Systems, Inc., Minneapolis, MN), DNAzol kit (Molecular Research Center, Inc., Cincinnati, OH.) or Easy-DNA kit (Invitrogen, Carlsbad, CA). **An. gambiae s.s.** and its molecular forms were identified using one of two rDNA-based PCR/RFLP assays, [2,7]. Ovaries from specimens of the desired species were subject to polytene chromosome analysis.

Contributors

- Carlo Costantini (carlo.costantini@ird.fr)
  - UMR MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France.
- N'Fale Sagnon
  - Centre National de Recherche et Formation sur le Paludisme (CNRFP), 01BP 2208 Ouagadougou 01, Burkina Faso.
- Nora J. Besansky (nbesansk@nd.edu)
  - Eck Institute for Global Health & Department of Biological Sciences, University of Notre Dame, IN 46556, USA.

Cameroon (1)

Sample sets

`AG1000G-CM-A`

Study information

Pyrethrum spray collections were conducted in three villages in Cameroon during September and October 2009. These villages comprise a transect from forest (village of Mayos: **(4.341, 13.558)**) to forest/savanna transition (village of Daiguene: **(4.777, 13.844)**) to savanna (villages of Gado-Badzere and Zembe-Borongo: **(5.747, 14.442)**) (8). All contributed specimens were **An. gambiae s.s.** [2]. A proportion of specimens were karyotyped via scoring of polytene chromosomes [9]. Specimens were stored on silica gel, and DNA was extracted using a simple CTAB protocol and run over Qiagen columns.

Related publications

- NF Lobo et al. 2010. Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. Malaria Journal. 9:293. (8).

Contributors

- Carlo Costantini (carlo.costantini@ird.fr)
  - UMR MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France.
- Kyanne R. Rohatgi
  - Eck Institute for Global Health & Department of Biological Sciences, University of Notre Dame, IN 46556, USA.
- Nora J. Besansky (nbesansk@nd.edu)
  - Eck Institute for Global Health & Department of Biological Sciences, University of Notre Dame, IN 46556, USA.

Cameroon (2)

Sample sets

`AG1000G-CM-B`

Study information

These samples were collected as part of a study which took place in Cameroon in Central Africa. The country is commonly referred to as "miniature Africa", owing to the diversity of its climate, topography, landscape, and bio-ecological settings: arid savannas in the north gradually turn into rain

forest in the south, along with highland areas, contribute to increase diversity of ecological settings. Anopheline mosquitoes were collected in 2005 from 64 locations covering a 1,500 km north-to-south transect that crossed all eco-geographical areas of Cameroon [10] Mosquito collection involved spraying aerosols of pyrethroid insecticides inside human dwellings, dead mosquitoes were retrieved from white sheets that were laid on the floor. Anopheline mosquitoes were identified using morphological identification keys [11,12]. Ovaries from half-gravid An. gambiae s.l. females were dissected and stored in Carnoy's fixative solution (absolute ethanol:glacial acetic acid 3:1) for cytogenetic analyses. Carcasses were stored individually in tubes containing a desiccant and kept at -20^{o} C until they were processed for molecular analysis. All half-gravid specimens collected in each village were identified to species and molecular forms using PCR-RFLP [2].

Related publications

- Simard, Frédéric, et al. "Ecological Niche Partitioning between Anopheles Gambiae Molecular Forms in Cameroon: The Ecological Side of Speciation." BMC Ecology, vol. 9, no. 1, 2009, p. 17, 10.1186/1472-6785-9-17. [10].

Contributors

- Frédéric Simard (frederic.simard@ird.fr)
  - UMR MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France.
- Diego Ayala (diego.ayala@ird.fr)
  - UMR MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France.
- Nora J. Besansky (nbesansk@nd.edu)
  - Eck Institute for Global Health & Department of Biological Sciences, University of Notre Dame, IN 46556, USA.
- Carlo Costantini (carlo.costantini@ird.fr)
  - UMR MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France.

Cameroon (3)

Sample sets

`AG1000G-CM-C`

Study information

Samples were contributed from pyrethrum spray collections, larval sampling and human landing catches conducted in twenty locations during October 2013.

These villages are scattered throughout the country and reflect a gradient of human-dominated environments, for example, forest (Manda: **(5.726, 10.868)** and Campo: **(2.367, 9.817)**; **forest/savanna transition (Tibati:** (6.469, 12.629)**); savanna (Lagdo:** (9.049, 13.656)); suburban area (Nkolondom: **(3.972, 11.516)**) and urban areas (Douala: **(4.055, 9.721)** and Yaoundé: **(3.880, 11.506)**.

Contributed specimens were *An. gambiae* or *An.coluzzii* [2].

Population genomics studies indicated the presence of relatively dierentiated subgroups within both species as well as clusters thriving in polluted breeding sites in large cities [13].

Specimens were stored on silica gel.

DNA was extracted using a Zymo research kit for adults, and a Qiagen kit for larvae.

Related publications

- C Kamdem et al. 2017. Pollutants and Insecticides Drive Local Adaptation in African Malaria Mosquitoes. Molecular Biology and Evolution. 34: 1261-1275. [13].

Contributors

- Colince Kamdem
  - Laboratoire de Recherche sur le Paludisme, Organisation de Coordination pour la lutte contre les Endémies en Afrique Centrale (OCEAC), B.P. 288, Yaoundé, Cameroon.
- Caroline Fouet
  - Department of Entomology, University of California, Riverside, CA, USA.
- Bradley J. White (bradwhite@verily.com)
  - Verily Life Sciences, South San Francisco, CA 94080, USA.

Central African Republic

Sample sets

`AG1000G-CF`

Study information

Collections were carried out in Bangui **(4.367, 18.583)**, during December 1993, by indoor resting aspiration or pyrethrum spray catch.

Contributors

- Alessandra della Torre (alessandra.dellatorre@uniroma1.it)
  - Istituto Pasteur Italia Fondazione Cenci Bolognetti, Dipartimento di Sanita Pubblica e Malattie Infettive, Università di Roma SAPIENZA, Rome, Italy.

Bioko Island, Equatorial Guinea

Sample sets

`AG1000G-GQ`

Study information

Collections were performed during the rainy season in September 2002 by overnight CDC light traps in Sacriba of Bioko island **(3.7, 8.7)**.

Specimens were stored dry on silica gel before DNA extraction.

Specimens contributed from this site were *An. gambiae* females, genotype determined by two assays [14].

All specimens had the $2L^{+a}/2L^{+a}$ karyotype as determined by the molecular PCR diagnostics [15].

These mosquitoes represent a population that inhabited Bioko Island before a comprehensive malaria control intervention initiated in February 2004 [16].

After the intervention *An. gambiae* was declining, and more recently almost only *An. coluzzii* can be found [17].

Contributors

- Jorge Cano
  - London School of Hygiene and Tropical Medicine, Keppel Street, Bloomsbury, London WC1E 7HT, UK.
- Maryam Kamali
  - Department of Medical Entomology and Parasitology, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran.
  - Department of Entomology, Virginia Tech, Blacksburg, Virginia 24061, USA.
- Igor Sharakhov (igor@vt.edu)
  - Department of Entomology, Virginia Tech, Blacksburg, Virginia 24061, USA.
  - Department of Genetics and Cell Biology, Tomsk State University, Tomsk 634050, Russia.

Côte d'Ivoire

Sample sets

`AG1000G-CI`

Study information

Samples were collected in Tiassale **(5.898, -4.823)**, located in the evergreen forest zone of southern Côte d'Ivoire. The primary agricultural activity is rice cultivation in irrigated fields. High malaria transmission occurs during the rainy seasons, between May and November. Samples were collected as larvae from irrigated rice fields by dipping between May and September 2012. All larvae were reared to adults and females preserved over silica for DNA extraction. Specimens from this site were all *An. coluzzii* , determined by PCR assay [7].

Related publications

- Grau-Bové, Xavier, et al. "Resistance to Pirimiphos-Methyl in West African Anopheles Is Spreading via Duplication and Introgression of the Ace1 Locus." PLOS Genetics, vol. 17, no. 1, 21 Jan. 2021, p. e1009253, 10.1371/journal.pgen.1009253. Accessed 1 Feb. 2021. [18].

Contributors

- David Weetman (david.weetman@lstmed.ac.uk)
  - Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, United Kingdom.
- Edi Constant
  - Centre Suisse de Recherches Scientifiques. Yopougon, Abidjan - 01 BP 1303 Abidjan, Côte d'Ivoire.

Ghana

Sample sets

`AG1000G-GH`

Study information

Samples were collected in Twifo Praso (5.609,-1.549), a peri-urban community located in semi-deciduous forest in the Central Region of Ghana. It is an extensive agricultural area characterised by small-scale (vegetable growing) and large-scale commercial farms such as oil palm and cocoa plantations. Mosquito samples were collected as larvae from puddles near farms between September and October, 2012. Madina (5.668,-0.219) is suburb of Accra within a coastal savanna zone of Ghana. It is an urban community characterised by myriad vegetable-growing areas. The vegetation consists of mainly grassland interspersed with dense short thickets often less than 5m high with a few trees. Specimens were sampled from puddles near roadsides and farms between October and December 2012. Takoradi (4.912,-1.774) is the capital city of the Western Region of Ghana. It is an urban community located in the coastal savanna zone. Mosquito samples were collected from puddles near road construction and farms between August and September 2012. Koforidua (6.094,-0.261) is the capital city of the Eastern Region of Southern Ghana and is located in semi-deciduous forest. It is an urban community characterized by numerous small-scale vegetable farms. Samples were collected from puddles near road construction sites and farms between August and September 2012. Larvae from all collection sites were reared to adults and females preserved over silica for DNA extraction. Both *An. gambiae* and *An. coluzzii* were collected from these sites, determined by PCR assay [7].

Related publications

- Essandoh, John, et al. “Acetylcholinesterase (Ace-1) Target Site Mutation 119S Is Strongly Diagnostic of Carbamate and Organophosphate Resistance in Anopheles Gambiae S.s. And Anopheles Coluzzii across Southern Ghana.” Malaria Journal, vol. 12, no. 1, 2013, p. 404, 10.1186/1475-2875-12-404. [19].

Contributors

- David iWeetman (david.weetman@lstmed.ac.uk)
  - Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, United Kingdom.
- John Essandoh
  - Department of Wildlife and Entomology, University of Cape Coast, Cape Coast, Ghana.

Democratic Republic of the Congo

Sample sets

`AG1000G-CD`

Study information

Samples were collected from Gbadolite (4.283,21.017), a town located in the far north of the Democratic Republic of Congo (DRC) near the border with the Central African Republic, surrounded by forest. In common with much of DRC, malaria transmission rates are high, and the samples are *An. gambiae s.s.* , which is the dominant vector. Samples were collected as larvae from temporary pools within and around the town by dipping in early August 2015. All larvae were reared to adults and females preserved over silica for DNA extraction using Qiagen DNAEasy kits.

Contributors

- David Weetman (david.weetman@lstmed.ac.uk)
  - Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool, United Kingdom.

Gabon

Sample sets

`AG1000G-GA-A`

Study information

Mosquitoes were collected by landing catches in the capital city Libreville **(0.384,9.455)** in December 2000 [**???**], an urban and polluted site. Malaria is endemic throughout the year. Specimens were stored in alcohol at -20^{O}C. Co-occurrence of both *kdr* resistance alleles and absence of wild-type susceptible alleles have been reported in this population [**???**]. *An coluzzii* and *An. melas* are also present in the region but at frequencies <1% [20]. Specimens were stored on silica gel and DNA extraction was performed by a phenol-chloroform protocol described in [21]

Contributors

- Nohal Elissa
  - Institut Pasteur de Madagascar, Avaradoha, BP 1274, 101 Antananarivo, Madagascar.
- João Pinto (jpinto@ihmt.unl.pt)
  - Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, UNL, Rua da Junqueira 100, 1349-008 Lisbon, Portugal.

Guinea Bissau

Sample sets

`AG1000G-GW`

Study information

Guinea Bissau samples were collected from three sites in October 2010 by indoor CDC light traps. Safim **(11.957,-15.649)** and Antula **(11.891,-15.582)**, from a south-western coastal region, characterised mainly by mixed flooded forests and croplands. Leibala is a neighbourhood of the eastern town of Gabu where shrubland and open deciduous forest predominate **(12.272,-14.222)**. According to PCR-RFLP of the IGS [2] and SINEX [7] all samples were identified as *An. gambiae*. The *kdr* pyrethroid target site resistance mutation L995F occurs at high frequency in Leibala but at very low frequency in the western coastal region [22] Malaria is meso-hyperendemic [23] and sporozoite rates are below 1% in the region. Specimens were stored on silica gel and DNA extraction was performed by a phenol-chloroform protocol described in [21].

Related publications

- Vicente, José L., et al. “Massive Introgression Drives Species Radiation at the Range Limit of Anopheles Gambiae.” Scientific Reports, vol. 7, no. 1, 18 Apr. 2017, 10.1038/srep46451. [22]

Contributors

- Amabélia Rodrigues
  - Instituto Nacional de Saaúde Paública, Ministaério da Saaúde Paública, Bissau, Guinaé-Bissau
- João Dinis
  - Instituto Nacional de Saaúde Paública, Ministaério da Saaúde Paública, Bissau, Guinaé-Bissau
- Marco Pombi

    - Istituto Pasteur Italia-Fondazione Cenci Bolognetti, Dipartimento di Sanita Pubblica e Malattie Infettive, Università di Roma SAPIENZA, Rome, Italy.
- Beniamino Caputo (beniamino.caputo@uniroma1.it)
    - Istituto Pasteur Italia-Fondazione Cenci Bolognetti, Dipartimento di Sanita Pubblica e Malattie Infettive, Università di Roma SAPIENZA, Rome, Italy.
- João Pinto (jpinto@ihmt.unl.pt)
    - Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, UNL, Rua da Junqueira 100, 1349-008 Lisbon, Portugal.

Guinea and Mali (1)

Sample sets

`AG1000G-GN-A`, `AG1000G-GN-B`

Study information

Collections were made from four different study sites around the border between Guinea and Mali. From Mali; Takan **(11.47,-8.33)** and Toumani Oulena **(10.83,-7.81)** are both small villages in the Yanfolila district of southern Mali and represent the Sudanian savannah ecological zone. Takan is arid savannah, while Toumani Oulena is humid savannah. In Guinea Conakry, mosquitoes were sampled from Koraboh **(9.28,-10.03)**, a small village in the Kissidougou district in the Faranah region representing a semi-forest site with intermediate ecology, a mix of savannah and forest, and in Koundara **(8.48,-9.53)**, a small village in the Macenta district in the Nzerekore region representing deep forest ecology. All reported collections occurred in October and November 2012. At each site, mosquitoes were collected using three different methods: human-landing capture, indoor manual aspirator or pyrethroid spray catch, and larval capture - where the first and second instar larvae were raised to adult in a field insectary under standard insectary conditions prior to DNA isolation from the adults, and the third and fourth instar larvae were preserved directly for DNA isolation, without rearing in the insectary.

The two distinct methods of larval collection were used to control for possible genetic bias inherent in lab rearing of captured larvae. Across sites, all types of larval sites were sampled, including both temporary and permanent sites. Human-landing captures were performed both inside dwellings and outside (>10 m from dwelling) at night between 18:00 and 06:30. The indoor aspirator or spray catches were done in the morning between 06:00 and 12:00. Adult specimens or third and fourth instar larvae were preserved immediately in 80% ethanol until later DNA extraction. First and second instar larvae were raised to adults in nearby field insectaries and upon emergence were preserved in 80% ethanol. DNA was extracted from mosquitoes using DNAzol by the provided protocol (Invitrogen, CA, USA).

Related publications

Coulibaly, Boubacar, et al. “Malaria Vector Populations across Ecological Zones in Guinea Conakry and Mali, West Africa.” Malaria Journal, vol. 15, no. 1, 8 Apr. 2016, 10.1186/s12936-016-1242-5. [9]

Contributors

- Boubacar Coulibaly
    - Malaria Research and Training Centre (MRTC), Faculty of Medicine and Dentistry, University of Mali, BP: E 423 Bamako, Mali
- Kenneth D. Vernick (kenneth.vernick@pasteur.fr)

    - Unit for Genetics and Genomics of Insect Vectors, Institut Pasteur, 75015, Paris, France.
- Michelle M. Riehle
    - Department of Microbiology and Immunology, Medical College of Wisconsin, Milwaukee, WI 53226, USA.

Whole Genome Sequencing

All library preparation and sequencing was performed at the Wellcome Sanger Institute.

Paired-end multiplex libraries were prepared using the manufacturer's protocol, with the exception that genomic DNA was fragmented using Covaris Adaptive Focused Acoustics rather than nebulization.

Multiplexes comprised 12 tagged individual mosquitoes and three lanes of sequencing were generated for each multiplex to even out variations in yield between sequencing runs.

Cluster generation and sequencing were undertaken according to the manufacturer's protocol for paired-end sequence reads with insert size in the range 100-200 bp.

4,693 individual mosquitoes were sequenced in total, of which 3,130 were sequenced using the Illumina HiSeq 2000 platform and 1,563 were sequenced using the Illumina HiSeq X platform.

All individuals were sequenced to a target coverage of 30X.

HiSeq 2000 sequencing runs generated 100 bp paired-end reads, while HiSeq X sequencing runs generated 150 bp paired-end reads.

Alignment and SNP calling

Reads were aligned to the AgamP4 reference genome using `bwa` version 0.7.15.

Indel realignment was performed using GATK version 3.7-0 `RealignerTargetCreator` and `IndelRealigner`.

Single nucleotide polymorphisms were called using GATK version 3.7-0 `UnifiedGenotyper`.

Genotypes were called for each sample independently, in genotyping mode, given all possible alleles at all genomic sites where the reference base was not `N`.

Coverage was capped at 250X by random down-sampling.

Complete specifications of the **alignment** and **genotyping** pipelines are available from the **malariagen/pipelines** GitHub repository.

Open source WDL implementations of the **alignment** and **genotyping** pipelines are also available from GitHub.

Following successful completion of these pipelines, samples entered the sample quality control (QC) process.

Sample QC

The following subsections describe analyses performed to identify and exclude samples from the final dataset.

Coverage

For each sample, depth of coverage was computed at all genome positions.

Samples were excluded if median coverage across all chromosomes was less than 10X, or if less than 50% of the reference genome was covered by at least 1X.

Cross-contamination

To identify samples affected by cross-contamination, we implemented the model for detecting contamination in NGS alignments described in 24.

Briefly, the method estimates the likelihood of the observed alternate and reference allele counts under different contamination fractions, given approximate population allele frequencies.

Population allele frequencies were estimated from the Ag1000G phase 2 data release 25.

The model computes a maximum likelihood value for a parameter α representing percentage contamination.

Samples were excluded if α was 4.5% or greater.

Technical replicates

A number of samples were sequenced more than once within this project phase (technical replicates).

To create a final dataset without any replicates suitable for population genetic analysis, we performed an analysis to confirm all technical replicates, and to choose the sample within each replicate with the best sequencing data.

We computed pairwise genetic distance between all sample pairs within a submission set.

The distance metric used was city block distance between genotype allele counts, to allow for handling of multiallelic SNPs.

So, *e.g.*, distance between genotypes of `0/1` and `0/1` is 0, distance between `0/0` and `0/1` is 2, distance between `0/1` and `1/2` is 2, distance between `0/0` and `1/1` is 4, etc.

For each pair of samples, distance was averaged over all sites where both samples had a non-missing genotype call.

Computations were initially carried out on a down-sampled set of 10 x 100,000 contiguous genomic sites, to be computationally feasible.

Where a pair of samples fell beneath a conservative threshold of 0.012, the genetic distance was then recomputed across all genomic sites (*i.e.*, without down-sampling).

For each pair of samples that were expected to be technical replicates according to metadata records, we excluded both members of the pair if genetic distance was above 0.006.

Where an expected replicate pair had genetic distance below 0.006, we retained only one sample in the pair.

We also identified and excluded both samples in any pair where genetic distance was below 0.006, where samples were not expected to be replicates.

Population outliers

We used principal component analysis (PCA) to identify and exclude individual samples that were population outliers.

SNPs were down-sampled to use 100,000 segregating non-singleton sites from chromosomes 3R and 3L, to avoid regions complicated by known introgression loci or paracentric inversions.

PCA was computed using `scikit-allel` version 1.2.0.

We iteratively identified and excluded any individual samples that were outliers along a single principal component.

We then identified and excluded any individual samples or small sample groups that clustered together with other samples in a way that was not plausible given metadata regarding their collection location.

Colony crosses

Samples in the `AG1000G-X` sample set were parents and progeny from colony crosses and were subject to a slightly different QC process.

For each cross, we performed an analysis of Mendelian inheritance and consistency to confirm the true parents and the validity of the cross.

Not all crosses were able to be successfully resolved, and samples that were not in a resolved cross were excluded.

From the samples originally submitted in the `AG1000G-X` sample set, 297 samples from 15 crosses were retained for release.

We did not include the colony crosses in the population outlier analysis due to their relatedness.

Sex calling

We called the sex of all samples based on the modal coverage ratio between the X chromosome and the autosomal chromosome arm 3R.

The sample was classed as male where the coverage ratio was between 0.4-0.6, and female between 0.8-1.2.

Where the ratio was outside these limits, the sample was excluded.

One of the sample sets from The Gambia, `AG1000G-GM-B`, included whole-genome amplified (WGA) samples which displayed some skew in their coverage ratios, which meant that sex could not be called via the same process.

These samples received a sex call where possible, but no samples were excluded based on uncertain sex call.

Species assignment

We assigned a species to each individual that passed sample QC using their genomic data, via two independent methods: ancestry-informative markers (AIMs) and principal components analysis (PCA).

Species calling via ancestry-informative markers

To derive AIMs between *A. arabiensis* and *A. gambiae*, we used publicly available data from the *Anopheles* 16 genomes project (26).

Whole genome SNP calls for 12 *A. arabiensis* and 38 *A. gambiae* individuals were used.

Alleles were mapped onto the same alternate allele space, and allele frequencies were computed for both species.

Sites that were multiallelic in either group were excluded, as well as sites where any genotypes were missing.

565,329 SNPs were identified as potentially informative, where no shared alleles were present between groups.

These were spread throughout the genome, but were concentrated on the X chromosome (63.2%), particularly around the Xag inversion.

We randomly down-sampled these SNPs to a set of 50,000 AIMs, then computed the fraction of alleles at these SNPs that were arabiensis-like for each individual in the Ag1000G phase 3 cohort.

Given the relatively small number of *A. arabiensis* samples in the 16 genomes project, it was clear that a significant proportion of putative AIMs were not likely to be truly informative across the broader sampling in Ag1000G.

Individuals in Ag1000G were classed as *A. arabiensis* where a fraction >0.8 of alleles were arabiensis-like.

To resolve the non-*A. arabiensis* individuals into *A. gambiae* and *A. coluzzii*, we applied the AIMs previously used in 25.

For each individual, we computed the fraction of coluzzii-like alleles at these AIMs.

Individuals were called as *A. gambiae* where this fraction was <0.12 and *A. coluzzii* where this fraction was >0.9, with individuals in between classed as intermediate.

Species calling via principal components analysis

To provide a complementary view of species assignments, we also used the results of the principal components analysis of Chromosome 3 computed during the outlier analysis described above.

Based on a comparison with the AIM species calls, it was apparent that the first two principal components could be used to assign species.

Individuals where PC1 > 150 were called as *A. arabiensis*.

Individuals where PC1 < 0 and PC2 > -7 were called as *A. gambiae*.

Individuals where PC1 < 0 and PC2 < -24 were called as *A. coluzzii*.

All other individuals were called as intermediate.

The results of the PCA and AIM species calls were highly concordant in most sample sets, except for the Far West (Guinea-Bissau, The Gambiae) and Far East (Kenya, Tanzania).

Further investigation is required to resolve the species status of these individuals.

Site filtering

We developed filters that identify genomic sites where SNP calling and genotyping is likely to be less reliable in one or more mosquito species.

To guide the design and calibration of the site filters, we made use of the 15 colony crosses included in this release.

Each cross comprises two parents and up to 20 progeny, allowing identification of sites where genotypes in one or more progeny are not consistent with Mendelian inheritance (Mendelian errors).

A small number of Mendelian errors may be due to *de novo* mutation, but the vast majority of Mendelian errors are likely to be due to errors in sequencing, alignment or SNP calling.

The general approach we took was to use Mendelian consistency to identify sets of positive and negative training sites, then used these to train a machine learning model that classified all genomic sites as either PASS or FAIL.

Site filters for use with *A. gambiae* and/or *A. coluzzii*

All the 15 crosses involved *A. gambiae* and/or *A. coluzzii* parents, while none of the crosses involved *A. arabiensis*.

We therefore used the crosses to first develop site filters suitable for use with *A. gambiae* and/or *A. coluzzii*.

Hereafter we refer to these filters as the `gamb\_colu` site filters.

Five of the 15 crosses were held out for validation, so performance could be evaluated objectively.

Sites were assigned to the positive training set where all genotypes across all 10 training crosses were called, and no Mendelian errors were observed.

Sites were assigned to negative training set where one or more Mendelian errors were observed in any cross.

All other sites were not considered eligible for inclusion in model training.

A balanced training set was then generated containing 100,000 autosomal sites from each of the positive and negative training sets.

The inputs to the machine learning model were a set of per-site summary statistics computed from the sequence read alignments and SNP genotypes across all wild-caught *A. gambiae* and *A. coluzzii* individuals.

These input summary statistics are described further in the appendix.

Male individuals were excluded from the summary statistic calculations, so that the model could also be applied without modification to the X chromosome.

We used these summary statistics, together with the positive and negative training sites, to train a decision tree model.

We initially trained a set of trees with different hyperparameter values, exploring the depth of trees, and the number of samples allowed at a terminal node.

Each of these trees was evaluated on an unbalanced set of sites randomly sampled from the whole genome (2% of all sites, without replacement).

Leaves of these trees contained different proportions of positive and negative training sites, and by increasing the cutoff for these proportions required to label a leaf as PASS, we were able to compute the area under the receiver operating curve (AUROC) for each set of hyperparameter values.

The best performing hyperparameter set based on AUROC was selected as the final model, and the leaf classification cutoff used was optimised based on the Youden statistic.

The resulting model was a decision tree of depth 8, where leaves were assigned to PASS where > 0.533 of training data in that leaf were positive training sites.

All sites in the genome were then assigned to PASS or FAIL via this model.

The 5 remaining cross pedigrees were used to perform a final evaluation of the approach.

For each of these crosses, we computed the Mendelian error rate (fraction of variants with one or more Mendelian errors among progeny) before and after applying the site filters, to provide five independent evaluation results.

We also evaluated performance on the X chromosome using heterozygote calls in males as indicator of error rates.

The fraction of variants with a heterozygous genotype call in or more males was computed before and after applying site filters.

Male error rates were estimated from genotype calls with a minimum Genotype Quality (GQ) value of 30.

Site filters for use with *A. arabiensis*

To generate site filters for use with *A. arabiensis*, we recomputed site summary statistics using only wild-caught *A. arabiensis* individuals, then applied the decision tree model described above.

These filters, which we refer to as the `arab` site filters, are appropriate when working with *A. arabiensis* samples only.

Site filters for joint analyses of all three species

We created site filters suitable for joint analysis of individuals from all three species by taking the intersection of the `gamb\_colu` and the `arab` site filters.

We refer to these filters as the `gamb\_colu\_arab` site filters.

Acknowledgments

We would like to thank the staff of the Wellcome Sanger Institute Sample Logistics, Sequencing and Informatics facilities for their contributions to the production of this data release.

We would like to thank the members of the Data Engineering team of the Broad Institute of Harvard and MIT for their work on open source implementations of the alignment and SNP calling pipelines used in Ag1000G phase 3.

Further information

For further information about the Ag1000G project, please visit **https://www.malariagen.net/ag1000g**.

For further information about the Ag1000G phase 3 SNP data release, please visit **www.malariagen.net/data/ag1000g-phase3-snp**.

If you have any questions regarding the data release, please start a new discussion at **https://github.com/malariagen/vector-public-data/discussions**.

References

1. **Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex.**
   Maureen Coetzee, Richard H Hunt, Richard Wilkerson, Alessandra Della Torre, Mamadou B Coulibaly, Nora J Besansky
   Zootaxa (2013) https://www.ncbi.nlm.nih.gov/pubmed/26131476
   PMID: 26131476

2. **Simultaneous identification of species and molecular forms of the *Anopheles gambiae* complex by PCR-RFLP**
   C. Fanello, F. Santolamazza, A. Della Torre
   Medical and Veterinary Entomology (2002-12) https://doi.org/ds5pmz
   DOI: 10.1046/j.1365-2915.2002.00393.x · PMID: 12510902

3. **:(unav)**
   Nelson Cuamba, Kwang Choi, Harold Townson
   Malaria Journal (2006) https://doi.org/cd3n27
   DOI: 10.1186/1475-2875-5-2 · PMID: 16420701 · PMCID: PMC1363361

4. **Distribution and Chromosomal Characterization of the Anopheles gambiae Complex in Angola**
   Pedro J. Cani, Maria Calzetta, Maria Angela Di Deco, Federica Santolamazza, Alessandra della Torre, Gian Carlo Carrara, Vincenzo Petrarca, Filomeno Fortes
   The American Journal of Tropical Medicine and Hygiene (2008-01-01) https://doi.org/ghv47t
   DOI: 10.4269/ajtmh.2008.78.169

5. **Population structure in the malaria vector, Anopheles arabiensis Patton, in East Africa**
   MJ Donnelly, N Cuamba, JD Charlwood, FH Collins, H Townson
   Heredity (1999-10-01) https://doi.org/bg4xmm
   DOI: 10.1038/sj.hdy.6885930 · PMID: 10583542

6. **IMP PCR primers detect single nucleotide polymorphisms for Anopheles gambiae species identification, Mopti and Savanna rDNA types, and resistance to dieldrin in Anopheles arabiensis.**
   Elien E Wilkins, Paul I Howell, Mark Q Benedict
   Malaria journal (2006-12-19) https://www.ncbi.nlm.nih.gov/pubmed/17177993
   DOI: 10.1186/1475-2875-5-125 · PMID: 17177993 · PMCID: PMC1769388

7. **Short report: A new polymerase chain reaction-restriction fragment length polymorphism method to identify Anopheles arabiensis from An. gambiae and its two molecular forms from degraded DNA templates or museum samples.**
   Federica Santolamazza, Alessandra Della Torre, Adalgisa Caccone
   The American journal of tropical medicine and hygiene (2004-06)
   https://www.ncbi.nlm.nih.gov/pubmed/15210999
   PMID: 15210999

8. **Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion**
   Neil F Lobo, Djibril M Sangaré, Allison A Regier, Kyanne R Reidenbach, David A Bretz, Maria V Sharakhova, Scott J Emrich, Sekou F Traore, Carlo Costantini, Nora J Besansky, Frank H Collins

Malaria Journal (2010-10-25) https://doi.org/c8td4x
DOI: 10.1186/1475-2875-9-293 · PMID: 20974007 · PMCID: PMC2988034

9. **Malaria vector populations across ecological zones in Guinea Conakry and Mali, West Africa**
Boubacar Coulibaly, Raymond Kone, Mamadou S. Barry, Becky Emerson, Mamadou B. Coulibaly, Oumou Niare, Abdoul H. Beavogui, Sekou F. Traore, Kenneth D. Vernick, Michelle M. Riehle
Malaria Journal (2016-04-08) https://doi.org/f8gzkb
DOI: 10.1186/s12936-016-1242-5 · PMID: 27059057 · PMCID: PMC4826509

10. **Ecological niche partitioning between Anopheles gambiae molecular forms in Cameroon: the ecological side of speciation**
Frédéric Simard, Diego Ayala, Guy Kamdem, Marco Pombi, Joachim Etouna, Kenji Ose, Jean-Marie Fotsing, Didier Fontenille, Nora J Besansky, Carlo Costantini
BMC Ecology (2009) https://doi.org/bd8bz5
DOI: 10.1186/1472-6785-9-17 · PMID: 19460146 · PMCID: PMC2698860

11. **The Anophelinae of Africa south of the Sahara. Suppl: Afrotropical region**
M. T. Gillies, Botha de Meillon
Publications of the South African Institute for Medical Research (1987)
ISBN: 9780620103213

12. **The anophelinae of Africa south of the Sahara (Ethiopian zoogeographical region)**
M. T. Gillies, Botha de Meillon

13. **Pollutants and Insecticides Drive Local Adaptation in African Malaria Mosquitoes.**
Colince Kamdem, Caroline Fouet, Stephanie Gamez, Bradley J White
Molecular biology and evolution (2017-05-01) https://www.ncbi.nlm.nih.gov/pubmed/28204524
DOI: 10.1093/molbev/msx087 · PMID: 28204524 · PMCID: PMC5400387

14. **Identification of Single Specimens of the Anopheles Gambiae Complex by the Polymerase Chain Reaction**
Julie A. Scott, William G. Brogdon, Frank H. Collins
The American Journal of Tropical Medicine and Hygiene (1993-10-01) https://doi.org/ghcqgk
DOI: 10.4269/ajtmh.1993.49.520 · PMID: 8214283

15. **Molecular karyotyping of the 2La inversion in Anopheles gambiae.**
Bradley J White, Federica Santolamazza, Luna Kamau, Marco Pombi, Olga Grushko, Karine Mouline, Cecile Brengues, Wamdaogo Guelbeogo, Mamadou Coulibaly, Jonathan K Kayondo, ... Nora J Besansky
The American journal of tropical medicine and hygiene (2007-02)
https://www.ncbi.nlm.nih.gov/pubmed/17297045
PMID: 17297045

16. **Malaria vector control by indoor residual insecticide spraying on the tropical island of Bioko, Equatorial Guinea**
Brian L Sharp, Frances C Ridl, Dayanandan Govender, Jaime Kuklinski, Immo Kleinschmidt
Malaria Journal (2007) https://doi.org/czzjf5
DOI: 10.1186/1475-2875-6-52 · PMID: 17474975 · PMCID: PMC1868751

17. **Light traps fail to estimate reliable malaria mosquito biting rates on Bioko Island, Equatorial Guinea**
Hans J Overgaard, Solve Sæbø, Michael R Reddy, Vamsi P Reddy, Simon Abaga, Abrahan Matias, Michel A Slotman

Malaria Journal (2012-02-24) https://doi.org/ghwxxm
DOI: 10.1186/1475-2875-11-56 · PMID: 22364588 · PMCID: PMC3384454

18. **Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the Ace1 locus**
Xavier Grau-Bové, Eric Lucas, Dimitra Pipini, Emily Rippon, Arjèn E. van 't Hof, Edi Constant, Samuel Dadzie, Alexander Egyir-Yawson, John Essandoh, Joseph Chabi, … The Anopheles gambiae 1000 Genomes Consortium
PLOS Genetics (2021-01-21) https://doi.org/ghwzd9
DOI: 10.1371/journal.pgen.1009253 · PMID: 33476334

19. **Acetylcholinesterase (Ace-1) target site mutation 119S is strongly diagnostic of carbamate and organophosphate resistance in Anopheles gambiae s.s. and Anopheles coluzzii across southern Ghana**
John Essandoh, Alexander E Yawson, David Weetman
Malaria Journal (2013) https://doi.org/gbfcqt
DOI: 10.1186/1475-2875-12-404 · PMID: 24206629 · PMCID: PMC3842805

20. **Malaria transmission in Libreville: results of a one year survey**
Jean-Romain Mourou, Thierry Coffinet, Fanny Jarjaval, Christelle Cotteaux, Eve Pradines, Lydie Godefroy, Maryvonne Kombila, Frédéric Pagès
Malaria Journal (2012) https://doi.org/gdkzp5
DOI: 10.1186/1475-2875-11-40 · PMID: 22321336 · PMCID: PMC3310827

21. **Population structure in the malaria vector, Anopheles arabiensis patton, in East Africa.**
MJ Donnelly, N Cuamba, JD Charlwood, FH Collins, H Townson
Heredity (1999-10) https://www.ncbi.nlm.nih.gov/pubmed/10583542
DOI: 10.1038/sj.hdy.6885930 · PMID: 10583542

22. **Massive introgression drives species radiation at the range limit of Anopheles gambiae**
José L. Vicente, Christopher S. Clarkson, Beniamino Caputo, Bruno Gomes, Marco Pombi, Carla A. Sousa, Tiago Antao, João Dinis, Giordano Bottà, Emiliano Mancini, … João Pinto
Scientific Reports (2017-04-18) https://doi.org/f93m36
DOI: 10.1038/srep46451 · PMID: 28417969 · PMCID: PMC5394460

23. **TRANSMISSION OF MIXED PLASMODIUM SPECIES AND PLASMODIUM FALCIPARUM GENOTYPES**
VIRGÍLIO E. DO ROSÁRIO, KATINKA PÅLSSON, THOMAS G. T. JAENSON, GEORGES SNOUNOU, JOÃO PINTO, ANA PAULA AREZ
The American Journal of Tropical Medicine and Hygiene (2003-02-01) https://doi.org/ghcqgm
DOI: 10.4269/ajtmh.2003.68.2.0680161

24. **Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data**
Goo Jun, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michael Boehnke, Hyun Min Kang
The American Journal of Human Genetics (2012-11) https://doi.org/f4cf5g
DOI: 10.1016/j.ajhg.2012.09.004 · PMID: 23103226 · PMCID: PMC3487130

25. **Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii***
The Anopheles gambiae 1000 Genomes Consortium

Genome Research (2020-10) https://doi.org/ghvn76
DOI: 10.1101/gr.262790.120 · PMID: 32989001 · PMCID: PMC7605271

26. **Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes**
Daniel E. Neafsey, Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, Bruno Arcà, Peter Arensburger, Gleb Artemov, … Nora J. Besansky
Science (2015-01-02) https://doi.org/gdkzt5
DOI: 10.1126/science.1258522 · PMID: 25554792 · PMCID: PMC4380271