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@0b62b1e on January 26, 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.

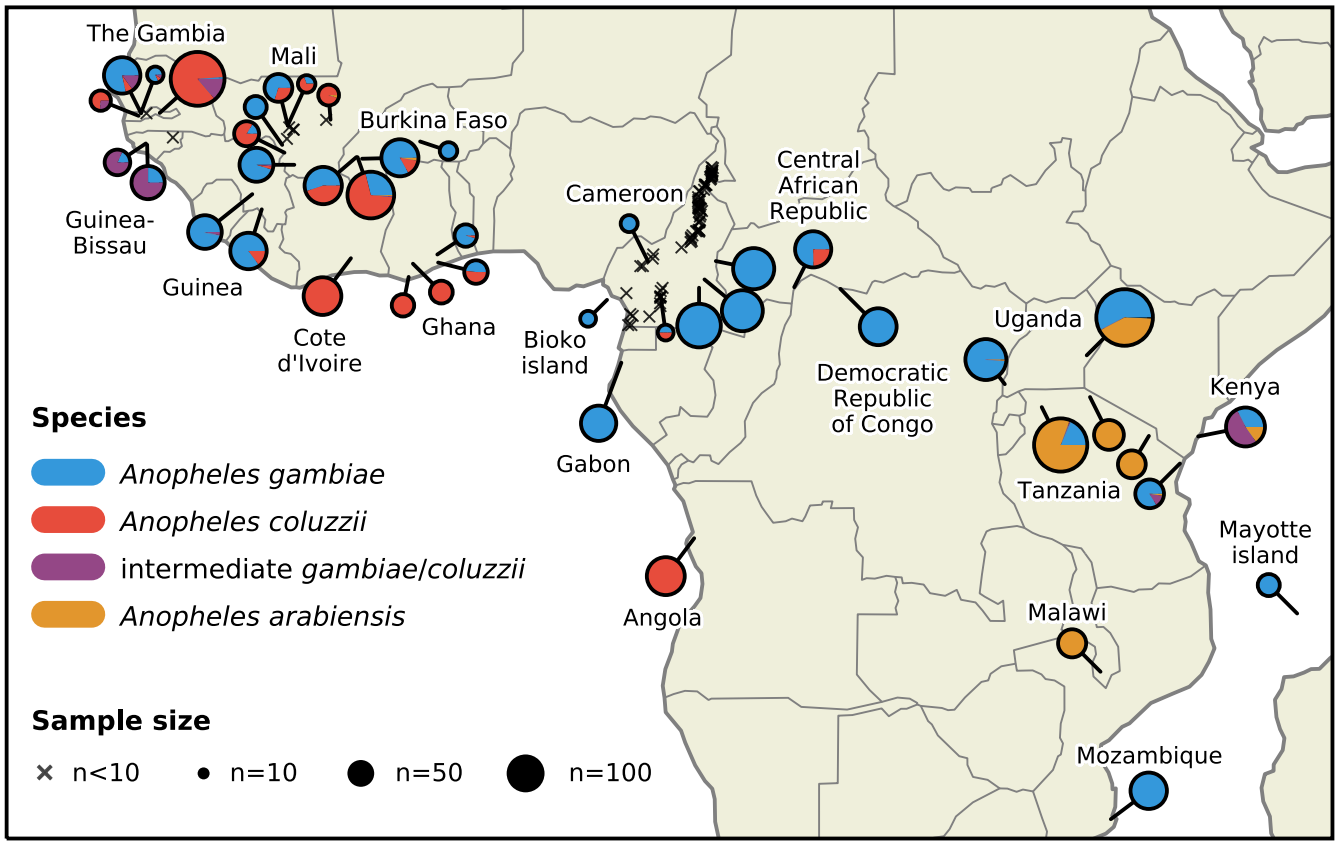

Figure 1: Sample Collection Map

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.

Population Sampling

Mosquitoes, from natural populations, were collected at 124 sites (unique latitude/longitudes) in 19 sub-Saharan African countries (Figure 1; Supplemental Table @@??).

95 of these sites are novel to Ag1000G phase 3, including 19 sites in six newly sampled countries, the remainder were previously sampled in phases 1 and 2 of the project ([**???**]; @[**???**] 2 doi when published).

New samples present in phase 3 comprised the following:

Burkina Faso

Two new submissions of collections from Burkina Faso are included here. The first added 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 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* (1) were collected.

Specimens were stored in 80% ethanol and DNA was extracted using the DNeasy Tissue Kit (Qiagen).

The second new submission from Burkina Faso added 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; 3).

Ovaries from specimens of the desired species were subject to polytene chromosome analysis.

Democratic Rebublic of the Congo

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.

Central African Republic

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

Cameroon

Two new submissions of samples from Cameroon are included in this phase.

In the first submission anopheline mosquitoes were taken from 64 locations covering a 1,500 km north-to-south transect that crossed all eco-geographical areas of Cameroon (4).

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 (>>Gillies and De Meillon 1968<<@[**???**] reference books with no ISBN?; 5).

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°C until they were molecularly processed. All half-gravid specimens collected in each village were identified to species and molecular forms using PCR-RFLP (2).

The second Cameroonian submission came 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* and *An. coluzzii* (2).

Population genomics studies indicated the presence of relatively differentiated subgroups within both species as well as clusters thriving in polluted breeding sites in large cities ([**???**]).

Specimens were stored on silica gel, and DNA was extracted using a Zymo research kit for adults and a Qiagen kit for larvae.

Mayotte

Phase 3 adds collections from three novel sites on the island of Mayotte.

Samples were collected as larvae during March-April 2011 in temporary pools by dipping in Tsounzou (-12.797, 45.185), Tsinkoura (-12.936, 45.138) and aerogare (-12.803, 45.283).

Larvae were stored in 80% ethanol prior to DNA extraction.

All specimens contributed were *An. gambiae* (3) with the standard $2L^{+a}/2L^{+a}$ or inverted $2L^{a}/2L^{a}$ karyotype as determined by the molecular PCR diagnostics (6).

The samples were identified as males or females by the sequencing read coverage of the X chromosome using LookSeq (7).

Gabon @[**???**] this submission "GA-B" included in the release?? I can't see it in the meta data, delete?

Anopheles sampling was carried out in three new locations in Gabon @[**???**]?.

Adult *An. gambiae* females were collected in Benguia (-1.633, 13.492) by human landing catches (National Research Ethics Committee of Gabon n°. 0031/2014/SG/CNE) in September 2015.

Benguia is a forest-savanna village with less than 300 inhabitants. Malaria is endemic across the year.

The specimens were stored immediately at -80°C.

An. coluzzii larvae were sampled in Libreville (0.390, 9.454) and Cocobeach (0.992, 9.576) by dipping in natural breeding sites in January 2016.

Libreville is the capital of the country and it is urban and polluted site, Cocobeach is a coastal village north of Libreville.

Across both malaria is endemic across the year.

The specimens were stored in alcohol at -20°C.

Total genomic DNA was extracted for all the mosquitoes using the CTAB protocol (8).

The Gambia

Two new submissions of samples from The Gambia are present in phase 3.

The first were collected along the Gambia River from the western coastal region of The Gambia, (Low River Area; Caputo et al. 2008), in August 2006.

An. gambiae and *An. coluzzii* specimens were identified to species following the PCR-RFLP protocol (2) using DNA extracted from the mosquito leg.

Only *An. coluzzii* specimens were collected from villages of Tankular (13.417, -16.033) and Kalataba (13.550, -15.617).

An. gambiae and *An. coluzzii* specimens were found in sympatry and collected from villages of Yallal Tankonjala (13.550, -15.700), Sare Samba Sowe (13.583, -15.900) and Hamdalai (13.567, -16.0167).

PCR-RFLP protocol also revealed the presence of mosquitoes with hybrid *An. gambiae*/*An. coluzzii* genotype in Yallal Tankonjala and Sare Samba Sowe.

Collections of indoor daytime-resting half gravid mosquitoes were carried out mainly in human dwellings and, in few cases, in animal shelters.

Collections were carried out by pyrethroid and/or paper-cup mouth aspirators from 12 AM to sunset, and kept in vials with desiccant.

Ovaries were dissected, maintained into Carnoy fixative (three parts pure ethanol:one part glacial acetic acid) and stored at -20^{O}C before polytene chromosome preparations (9).

Chromosome scoring was carried out under a phase-contrast optical microscope.

Paracentric inversion karyotypes were scored according to the nomenclature and conventions of Coluzzi et al. ([**???**](79)90036-1) and Touré et al. (10).

The second new submission consists of adult mosquitoes collected at Wali Kunda in the rural, central river region of The Gambia (13.567, -14.917).

The area is 180 km from the sea, on the south bank of the River Gambia, in flat Sudan savannah with a small fishing village (and a research field station) as well as rice fields and swamplands.

The dominant *Anopheles* vector species is *An. coluzzii* (11).

Mosquitoes were captured using human landing collections both inside and outside huts for 19 days in October and November 2012. Mosquitoes were stored in RNAlater or dried over silica gel and stored at -20^{O}C.

Guinea and Mali

A novel sample submission to phase 3 included collections from both Guinea and Mali.

Mosquitoes were collected from four different study sites at the border with Mali and in Guinea Conakry. 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, we sampled in 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 in 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).

Further details on sampling and basic Southern Mali/Guinea vector biology are presented in Coulibaly et al. (doi:10.1186/s12936-016-1242-5).

Guinea Bissau

Two new Guinea Bissau collection sites are added here, both performed in October 2010 by indoor CDC light traps.

Samples were also collected from Ga-Mbana (12.052, -14.902) and Leibala (12.272, -14.222).

Ga-Mbana is a rural village located along a main road in central Guinea Bissau, while Leibala is a neighbourhood of the eastern town of Gabu.

The samples of Ga-Mbana comprised *An. coluzzii* whereas the samples of Leibala comprised *An. gambiae*, all being identified by IGS and SINEX markers as described in Vicente et al. (12).

The *kdr* pyrethroid target site resistance mutation L1014F occurs at high frequency in Leibala but at very low frequency in Ga-Mbana (12).

Malaria is meso-hyperendemic (13) 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 Donnelly et al. (14).

Kenya

New Kenyan specimens were obtained from villages located in Kilifi County near the Kenyan coast between 2000 and 2014.

All Anopheles mosquito sampling was conducted indoors using CDC light traps.

An. gambiae, *An. funestus*, *An. arabiensis* and *An. merus* were present at sampling locations (15).

Sporozoite rates for the area during previous studies were 1.47% (16).

Mali

Two further submissions containing collections from 11 sites in Mali are new for phase 3.

In the first, collections were made in four villages in the Koulikoro region; Tieneguebougou (12.810, -8.080) approximately 20 km north of Bamako, and Kababougou (12.890, -8.150), Ouassorola (12.900, -8.160), Sogolombougou (12.880, -8.140), approximately 30 km north of Bamako.

The collections were made in August 2014 by human landing catch and pyrethrum spray catch.

Both *An. gambiae* and *An. coluzzii* (2) were collected.

Specimens were stored in 80% ethanol.

In the second submission, collections of indoor resting adults were made by spray catch from seven villages in the southern part of Mali in August-September 2004: Banambani (12.800, -8.050), Bancoumana (12.200, -8.200), Douna (13.210, -5.900), Fanzana (13.200, -6.130), Kela (11.880, -8.450), Moribobougou (12.690, -7.870) and N'Gabakoro (12.680, -7.840).

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; 3).

Ovaries from specimens of the desired species were subject to polytene chromosome analysis.

Malawi

Specimens were obtained from villages within the catchment of the Majete Malaria Project, Chikhwawa District, Malawi (-15.933, 34.755) (17).

Mosquitoes were collected indoors and outdoors by Suna light trap in May 2015.

Chickhwawa District is an area with perennial and intense malaria transmission (18).

All specimens were *An. arabiensis* (2).

Additional details of vector population bionomics may be found in (17; 19).

Specimens were stored over silica and DNA was extracted using the Qiagen plate protocol.

Mozambique

Mosquito samples were collected in Furvela (-23.716, 35.299), Mozambique, by CDC light traps between December 2003 and April 2004.

Specimens were stored on silica gel and DNA was extracted according to Collins et al. (20).

Contributed specimens consisted of *An. gambiae* individuals identified according to Fanello et al. (2).

Furvela is a rural village located in Inhambane Province, where malaria is transmitted mainly by *An. gambiae* and *An. funestus* (21).

An. arabiensis and *An. merus* are also found at low frequency.

Sporozoite rates around 4% have been reported in *An. gambiae* from Furvela (21).

Tanzania

Tanzanian samples were collected from four distinct locations.

Moshi samples came from lower Mabogini (-3.400, 37.350), rice fields near lower Moshi on the southern slope of Mount Kilimanjaro, a region shown to have increasing resistance to pyrethroids (22).

Mosquitoes were collected as larvae, during the rice growing season in August-September 2012, raised to adults and females bioassayed in WHO tubes for one hour with 0.05% lambda cyhalothrin ([**???**]).

Alive and dead mosquitoes were preserved over silica.

In Tanzanian samples screened in Kabula et al. (23), Moshi was the most pyrethroid resistant population, they were found to be completely DDT susceptible, only in one out of 642 mosquitoes

assayed by Matowo et al. (22) was found to carry a *kdr* resistance mutation (Vgsc-995F).

Tarime collections took place in the village of Komaswa (-1.417, 34.183) about 410 km north west of Moshi, during August 2012.

Mosquito larvae were collected, raised to adults and females bioassayed with a range of insecticides in WHO tubes for one hour ([**???**]), finding almost complete multi-insecticide susceptibility: permethrin (100% mortality), lambda cyhalothrin (97%), fenitrothion (100%), DDT (100%) and bendiocarb (100%) (Nyka, T. unpublished data – Insecticide Resistance Monitoring Report 2012. NIMR Tanzania).

Mulheza samples were collected from Zeneti village (-5.217, 38.650), northeast Tanzania.

Malaria is intense and perennial with transmission peaking after the rainy season in May and June (23). Mosquitoes were sampled between November 2012 and May 2013.

Indoor resting collections were used to obtain live females for deltamethrin susceptibility testing and pyrethrum spray catches were used for mosquitoes that were collected for blood meal analysis.

Collections were conducted between 06:00 and 09:00 from randomly selected houses.

Live mosquitoes collected for susceptibility testing were provided with 10% glucose solution and transported to the field insectary.

Mosquitoes were sorted and morphologically identified to species, carcasses were stored individually over desiccant for laboratory processing. Muleba (1.750, 31.667), the final collection region, is in the North-western part of Tanzania.

The district is known to be a malaria epidemic prone area with unstable transmission of varying seasonality.

The highest peak of malaria transmission is usually reached between May-July and November-January, which results from proceeding rain seasons.

There have been malaria vector control efforts since 2007 when indoor residual spraying using Lambdacyhalothrin was introduced.

Insecticide resistance in this district is coupled with high frequency of *kdr* pyrethroid target site mutations in the *An. gambiae* s.s population (23; 24).

Sampling was conducted over six months, which include both dry and rainy season and covers 6 villages selected to represent all major ecological systems in the district.

Uganda

In Uganda, a single new site is sampled in phase 3.

In Kihihi subcounty, Kanungu District (-0.751, 29.701), resting mosquitoes were collected during October and November 2012.

Kihihi is located in an upland area with seasonal malaria transmission (25).

All specimens were *An. gambiae* (2).

Additional details of vector population bionomics may be found in ([**???**]).

Specimens were stored in 80% ethanol and DNA was extracted using the Qiagen plate protocol.

Details of natural population samples previously described in phases 1 and 2 can be found in the Supplementary Information of The Anopheles gambiae 1000 Genomes Consortium ([**???**]; @[**???**] 2 doi when published - respectively).

Colony Crosses

15 crosses were contributed to Ag1000G phase 3.

The crosses were generated using parents from eight different colonies: G3(MRA-112); Kisumu(MRA-762); Pimperena (canonical representative of *An. gambiae* species; MRA-861); Ghana (recent colony of *An. coluzzii* from Okyereko, southern Ghana (26)); Mali-NIH (canonical representative of *An. coluzzii* species; Niono, MRA-860); (P)Akron (Benin, MRA-913); Nagongera (Tororo, Uganda); and Tiassalé (southern Cote d'Ivoire (26)).

The labels, e.g. "29-2", are identifiers used for each of the crosses within the project and have no special meaning.

Full full details of cross production, sequencing and quality control is described in the Supplementary Information of The Anopheles gambiae 1000 Genomes Consortium ([**???**]).

The only difference between the production of the four crosses that are novel to phase 3: B4, K2, K4 and K6, and those from earlier phases of the project, is that multiple males and multiple females were placed together for mating, then all male and all egg laying female mosquitoes were sequenced.

This necessitated matching up potential fathers of crosses with the correct mother and offspring.

For each cross for which the father was in doubt, the list of potential parental pairs was computed.

For each of these pairs, for each chromosome, the Mendelian error was computed for every sample of the progeny and the median value (along samples) was plotted for every computation.

If one pair yielded median Mendelian errors significantly lower for every chromosome than any other pair (except X which was consistent no matter the parents), that pair was chosen as the parsimonious parents.

In four crosses (B4, K2, K4 and K6) parental pairs could be clearly identified and these crosses could, therefore, be included in the phase 3 data release.

Two of the novel crosses, K4 and K6, were found to be fathered by the same male, AC0398.

Methods

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 27.

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 28.

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 and , we used publicly available data from the 16 genomes project .

Whole genome SNP calls for 12 and 38 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 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 where a fraction $>$0.6 of alleles were arabiensis-like.

To resolve the non-individuals into and , we applied the AIMs previously used in .

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

Individuals were called as where this fraction was $<$0.12 and 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 .

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

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

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 Ag1000G phase 3. % Each cross comprises two parents and up to 20 progeny, and thus it is possible to identify 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 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 and/or

All the 15 crosses involved and/or parents, and none of the crosses involved , so we used the crosses to first develop site filters suitable for use with and/or . % Hereafter we refer to these filters as the ” 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 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 and 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 an error indicator. % 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. % Performance of the decision tree model was better than the hand-crafted site filters created during the previous project phase (Ag1000G phase 2), with lower Mendelian error rates, and a larger number of sites passing the filter. % Full performance metrics will be reported in a future publication.

Site filters for use with

To generate site filters for use with , we recomputed site summary statistics using only wild-caught individuals, then applied the decision tree model described above. % These filters, which we refer to as the ” site filters, are appropriate when working with 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 and the site filters. % We refer to these filters as the ” 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 .

For further information about the Ag1000G phase 3 SNP data release, please visit .

If you have any questions regarding the data release, please start a new discussion at .

References

1. **:(unav)**
   Elien E Wilkins, Paul I Howell, Mark Q Benedict
   Malaria Journal (2006) https://doi.org/dbtfkr
   DOI: 10.1186/1475-2875-5-125 · PMID: 17177993 · PMCID: PMC1769388

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. **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**
   ALESSANDRA DELLA TORRE, ADALGISA CACCONE, FEDERICA SANTOLAMAZZA
   The American Journal of Tropical Medicine and Hygiene (2004-06-01) https://doi.org/ghcqgn
   DOI: 10.4269/ajtmh.2004.70.604

4. **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

5. **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

6. **MOLECULAR KARYOTYPING OF THE 2LA INVERSION IN ANOPHELES GAMBIAE**
   NORA J. BESANSKY, OLGA GRUSHKO, ALESSANDRA DELLA TORRE, IGOR SHARAKHOV, CECILE BRENGUES, JONATHAN K. KAYONDO, FEDERICA SANTOLAMAZZA, FREDERIC SIMARD, MARCO POMBI, MAMADOU COULIBALY, ... KARINE MOULINE
   The American Journal of Tropical Medicine and Hygiene (2007-02-01) https://doi.org/ghcqgq
   DOI: 10.4269/ajtmh.2007.76.334

7. **LookSeq: A browser-based viewer for deep sequencing data**
   H. M. Manske, D. P. Kwiatkowski
   Genome Research (2009-08-13) https://doi.org/b5hmbh
   DOI: 10.1101/gr.093443.109 · PMID: 19679872 · PMCID: PMC2775587

8. **POPULATION STRUCTURE OF ANOPHELES ARABIENSIS ON LA RÉUNION ISLAND, INDIAN OCEAN**
   FRÉDÉRIC SIMARD, RICHARD HUNT, DIDIER FONTENILLE, ISABELLE MORLAIS, ROMAIN GIROD
   The American Journal of Tropical Medicine and Hygiene (2005-12-01) https://doi.org/ghcqgp
   DOI: 10.4269/ajtmh.2005.73.1077

9. **Anopheles gambiae complex along The Gambia river, with particular reference to the molecular forms of An. gambiae s.s**
Beniamino Caputo, Davis Nwakanma, Musa Jawara, Majidah Adiamoh, Ibrahima Dia, Lassana Konate, Vincenzo Petrarca, David J Conway, Alessandra della Torre
Malaria Journal (2008) https://doi.org/b97c6c
DOI: 10.1186/1475-2875-7-182 · PMID: 18803885 · PMCID: PMC2569043

10. **The distribution and inversion polymorphism of chromosomally recognized taxa of the Anopheles gambiae complex in Mali, West Africa.**
YT Touré, V Petrarca, SF Traoré, A Coulibaly, HM Maiga, O Sankaré, M Sow, MA Di Deco, M Coluzzi
Parassitologia (1998-12) https://www.ncbi.nlm.nih.gov/pubmed/10645562
PMID: 10645562

11. **Does insecticide resistance contribute to heterogeneities in malaria transmission in The Gambia?**
Kevin Ochieng' Opondo, David Weetman, Musa Jawara, Mathurin Diatta, Amfaal Fofana, Florence Crombe, Julia Mwesigwa, Umberto D'Alessandro, Martin James Donnelly
Malaria Journal (2016-03-15) https://doi.org/ghcqgh
DOI: 10.1186/s12936-016-1203-z · PMID: 26980461 · PMCID: PMC4793517

12. **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

13. **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

14. **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

15. **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

16. **Wind direction and proximity to larval sites determines malaria risk in Kilifi District in Kenya**
Janet T. Midega, Dave L. Smith, Ally Olotu, Joseph M. Mwangangi, Joseph G. Nzovu, Juliana Wambua, George Nyangweso, Charles M. Mbogo, George K. Christophides, Kevin Marsh, Philip Bejon
Nature Communications (2012-02-14) https://doi.org/ghcqgd
DOI: 10.1038/ncomms1672 · PMID: 22334077 · PMCID: PMC3292715

17. **Assessment of the effect of larval source management and house improvement on malaria transmission when added to standard malaria control strategies in southern Malawi: study protocol for a cluster-randomised controlled trial**
Robert S. McCann, Henk van den Berg, Peter J. Diggle, Michèle van Vugt, Dianne J. Terlouw, Kamija S. Phiri, Aurelio Di Pasquale, Nicolas Maire, Steven Gowelo, Monicah M. Mburu, … Willem Takken
BMC Infectious Diseases (2017-09-22) https://doi.org/ggr5g7
DOI: 10.1186/s12879-017-2749-2 · PMID: 28938876 · PMCID: PMC5610449

18. **Mapping Malaria Transmission Intensity in Malawi, 2000–2010**
Adam Bennett, Lawrence Kazembe, Don P. Mathanga, Damaris Kinyoki, Doreen Ali, Robert W. Snow, Abdisalan M. Noor
The American Journal of Tropical Medicine and Hygiene (2013-11-06) https://doi.org/f5h9zz
DOI: 10.4269/ajtmh.13-0028 · PMID: 24062477 · PMCID: PMC3820324

19. **Entomological indices of malaria transmission in Chikhwawa district, Southern Malawi**
Themba Mzilahowa, Ian M Hastings, Malcolm E Molyneux, Philip J McCall
Malaria Journal (2012-11-21) https://doi.org/gbcb3h
DOI: 10.1186/1475-2875-11-380 · PMID: 23171123 · PMCID: PMC3536595

20. **A Ribosomal RNA Gene Probe Differentiates Member Species of the Anopheles gambiae Complex**
Frank H. Collins, Nora J. Besansky, M. Alina Mendez, Melissa O. Rasmussen, Victoria Finnerty, Philip C. Mehaffey
The American Journal of Tropical Medicine and Hygiene (1987-07-01) https://doi.org/ghcqgj
DOI: 10.4269/ajtmh.1987.37.37 · PMID: 2886070

21. **Analysis of the sporozoite ELISA for estimating infection rates in Mozambican anophelines**
J. D. CHARLWOOD, E. V. E. TOMÁS, N. CUAMBA, J. PINTO
Medical and Veterinary Entomology (2015-03) https://doi.org/f62ckd
DOI: 10.1111/mve.12084 · PMID: 25088021

22. **Genetic basis of pyrethroid resistance in a population of Anopheles arabiensis, the primary malaria vector in Lower Moshi, north-eastern Tanzania**
Johnson Matowo, Christopher M Jones, Bilali Kabula, Hilary Ranson, Keith Steen, Franklin Mosha, Mark Rowland, David Weetman
Parasites & Vectors (2014) https://doi.org/ghcqgg
DOI: 10.1186/1756-3305-7-274 · PMID: 24946780 · PMCID: PMC4082164

23. **Susceptibility status of malaria vectors to insecticides commonly used for malaria control in Tanzania**
Bilali Kabula, Patrick Tungu, Johnson Matowo, Jovin Kitau, Clement Mweya, Basiliana Emidi, Denis Masue, Calvin Sindato, Robert Malima, Jubilate Minja, … William Kisinza
Tropical Medicine & International Health (2012-06) https://doi.org/f3zsbt
DOI: 10.1111/j.1365-3156.2012.02986.x · PMID: 22519840

24. **High level of resistance in the mosquito Anopheles gambiae to pyrethroid insecticides and reduced susceptibility to bendiocarb in north-western Tanzania**
Natacha Protopopoff, Johnson Matowo, Robert Malima, Reginald Kavishe, Robert Kaaya, Alexandra Wright, Philippa A West, Immo Kleinschmidt, William Kisinza, Franklin W Mosha, Mark Rowland
Malaria Journal (2013-05-02) https://doi.org/ghcqgf
DOI: 10.1186/1475-2875-12-149 · PMID: 23638757 · PMCID: PMC3655935

25. **Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae s.l. using three sampling methods in three sites in Uganda**
Maxwell Kilama, David L Smith, Robert Hutchinson, Ruth Kigozi, Adoke Yeka, Geoff Lavoy, Moses R Kamya, Sarah G Staedke, Martin J Donnelly, Chris Drakeley, … Steve W Lindsay
Malaria Journal (2014-03-21) https://doi.org/gdkzsz
DOI: 10.1186/1475-2875-13-111 · PMID: 24656206 · PMCID: PMC4001112

26. **CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito Anopheles gambiae**
Constant V. Edi, Luc Djogbénou, Adam M. Jenkins, Kimberly Regna, Marc A. T. Muskavitch, Rodolphe Poupardin, Christopher M. Jones, John Essandoh, Guillaume K. Kétoh, Mark J. I. Paine, … David Weetman
PLoS Genetics (2014-03-20) https://doi.org/f56k77
DOI: 10.1371/journal.pgen.1004236 · PMID: 24651294 · PMCID: PMC3961184

27. **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

28. **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