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@4006a63 on January 27, 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.

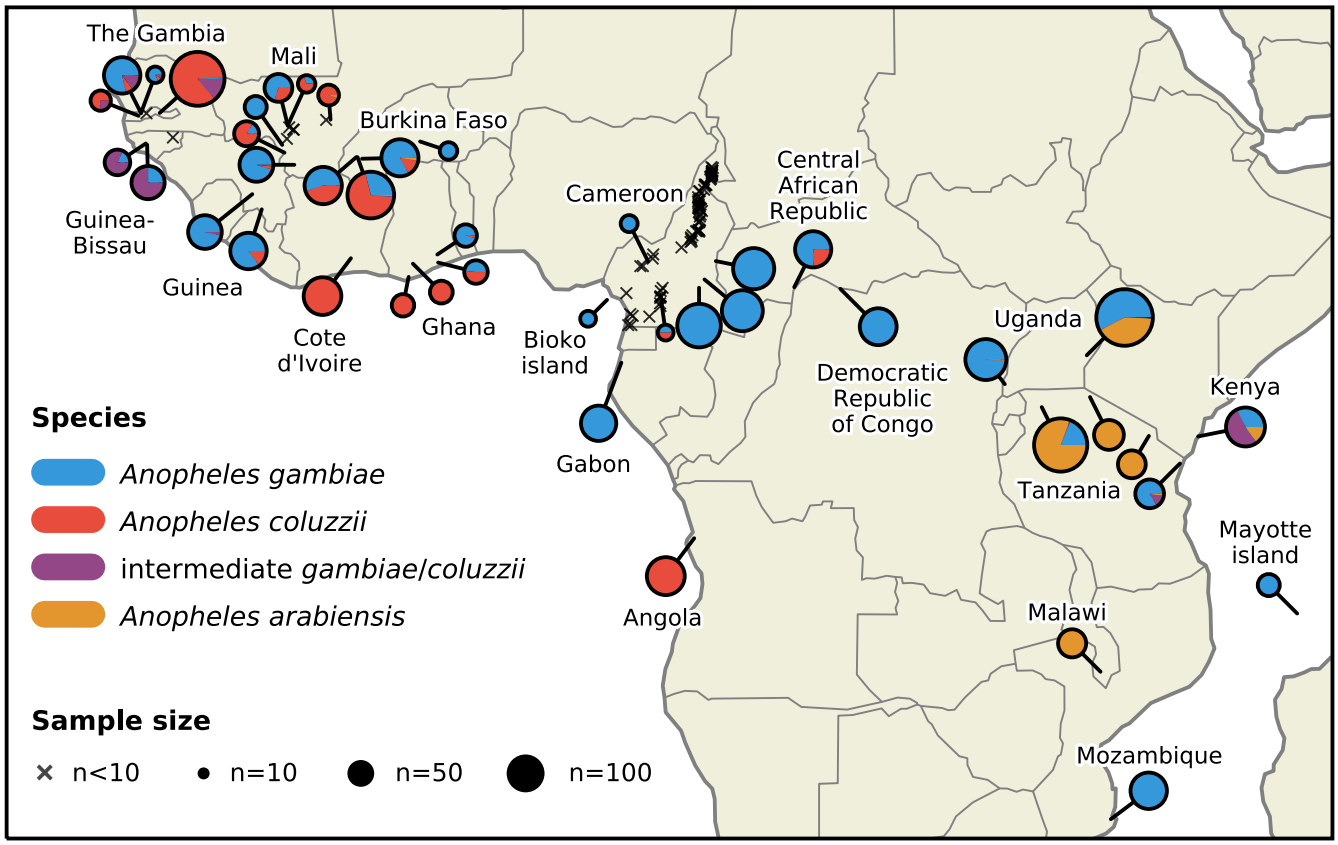

Figure 1: Sample Collection Map

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. (2013).

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*~ although and *A. arabiensis* have also been recorded in the province~.

Specimens were stored on silica gel and DNA extraction was performed by a phenol-chloroform protocol described in .

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.

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 1.

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 2.

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 (3).

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 2.

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 .

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. **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

2. **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

3. **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