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@0150c82 on September 17, 2020.

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 collected from a total of 124 sites across 19 countries in Sub-Saharan Africa.

Six novel countries have been added in phase 3; collections from Malawi 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.

In phase three, new quality control and filtering pipelines were developed so all 4,693 samples submitted across the three project phases were evaluated with these new methods.

High quality whole genome sequences from 2784 unique individuals passed all sample filters, 2532 females, 223 males and 29 samples with unknown sex (see Methods).

1823 of these wild-caught individuals are novel to phase 3.

181 samples from phase 2 did not pass the new quality thresholds are not included in this data release (@@ either - see Supplementary table OR explain 172 were lost via FILTER_second_rep_hi_skew <- then need to explain what this is?).

Alongside sampling from natural populations, 699 individuals (parents and offspring) from 15 lab crosses have been sequenced.

Five of these crosses are novel to phase 3, adding power to test and validate methods.

Parents of crosses were drawn from the @@ and @@... colonies.

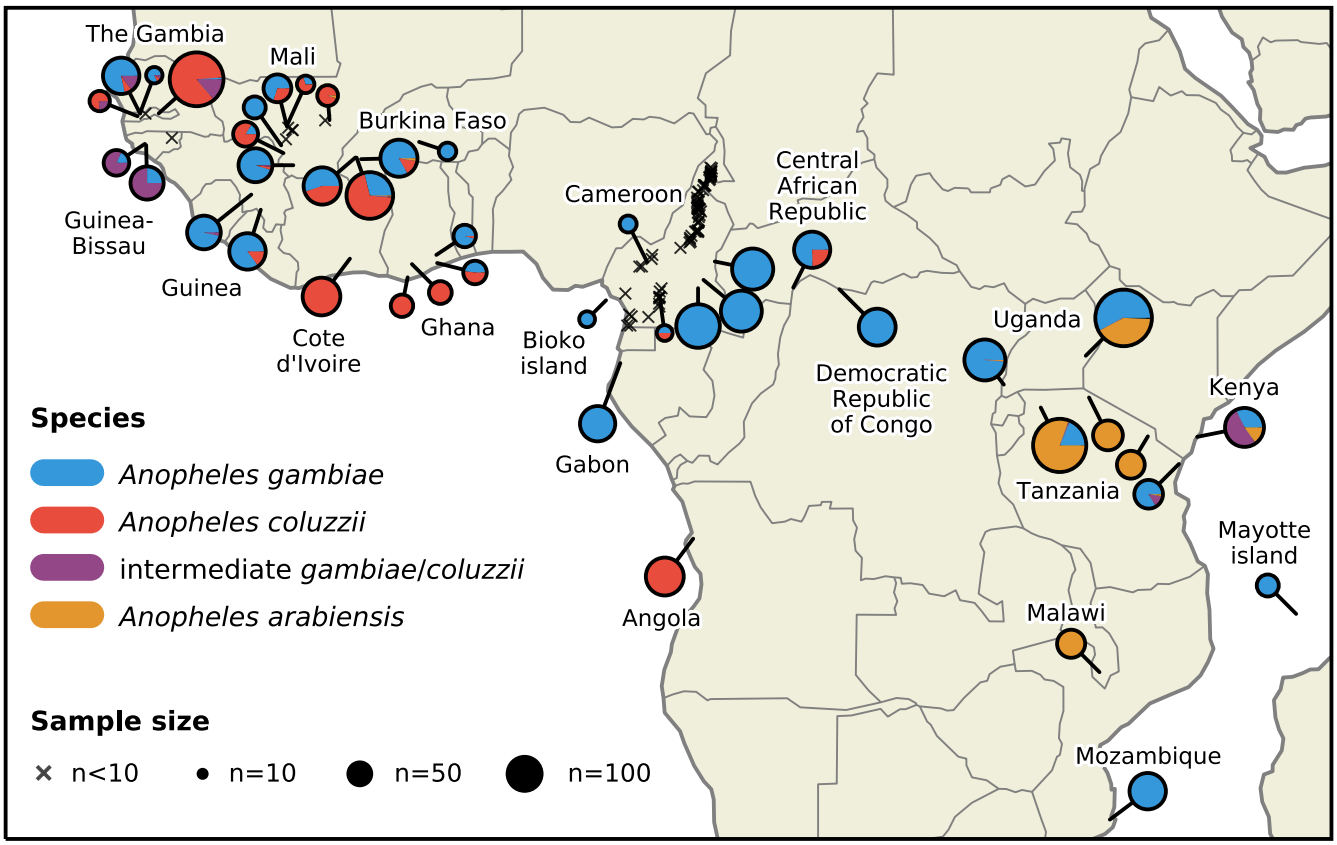

Figure 1: Sample Collection Map

Whole Genome Sequencing and Alignment

4,693 individual mosquitoes 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

The sample QC process was composed of three stages, sequence quality assurance, replicate handling, and anomaly detection.

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 AG1000G-X submission set, made up of laboratory experimental crosses, was exempted from this step due to familial similarity and high levels of inbreeding.

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 samples (74.2%) were retained from the original cohort for analysis.

Coverage

Summary of site coverage post QC exclusions.

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 identified by Neafsey et al (ref), and used by us in previous analyses (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, highlights the complexity of species relationships here.

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 genome 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 small reduction in FDR, but coupled with the lower rate of accessibility in phase 2 (62.05%) resulted in an significant improvement in Youden score (Table XXX) across autosomes.

The X chromosome showed a similar pattern, the heterozygosity rate 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.

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.

Result of heterozygote calls on male X chromosome

	count	mean	min	25%	50%	75%	max	fraction_accessible
pre-filtering	220.00000	0.00281	0.00165	0.00227	0.00244	0.00268	0.01687	1.00000
post-filtering	220.00000	0.00037	0.00016	0.00021	0.00023	0.00032	0.00853	0.69970

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. gambiae*/ *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

Genetic Diversity within Populations

Insecticide Resistance

Gene Drive

Whole Genome Sequencing and Alignment

4,693 individual mosquitoes were sequenced using the Illumina HiSeq2000 (n=3,130) and the Illumina HiSeqX (n=1,563) to a target coverage of 30X.

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

Indel realignment was performed using GATK `v3.7-0` RealignerTargetCreator and IndelRealigner.

Single nucleotide polymophisms were called against AgamP4 using GATK UnifiedGenotyper `v3.7-0`.

Sample genotypes were called independently, in genotyping mode, given all possible alleles at each site, allowing parallelisation over samples.

Coverage considered at individual sites was capped at 250X by random downsampling.

Full details of pipelines including all parameter settings are available in supplementary information.

Following successfull completion of the pipeline samples entered the sample quality control (QC) process.

Sample QC

The sample QC process was composed of three distinct stages: sequence quality assurance, replicate handling, and anomaly detection.

To meet the requirements of sequence quality assurance median coverage had to be at least 10X, and minumum 50% of the genome covered by at least 1X.

We also implemented the test for contamination in NGS alignments described in Jun et al (https://doi.org/10.1016/j.ajhg.2012.09.004).

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

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

The model computes a maximum likelihood value for a parameter representing percentage contamination (alpha).

Where this parameter was 4.5% or greater the sample was excluded.

We also made sex calls based on the modal coverage ratio X:3R.

3R was selected as representative of autosomal coverage as it is free from inversions and large regions of heterochromatin.

The modal coverage was used owing to concerns around the high skewdness of coverage distributions.

To mitigate this further, when computing the modal coverage we only considered sites where coverage was at least 2X.

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 sex call was not made and the sample dropped.

One of the submission sets from The Gambia, was composed entirely of Whole Genome Amplified (WGA) samples.

These received a sex call where possible but the decision was made not to exclude based on ratio, due to the inherent value of this submission.

The fact this set is entirely WGA was considered throughout the analysis of these samples.

The sample QC process also included assessment of technical replicates.

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

The metric used was Hamming between alleles, so a genotype of 0/1 records a value of 1 against the genotype 1/2, allowing straightforward handling of multiallelic SNPs.

Only sites where both samples had a genotype call were included, mean distance was calculated over a denominator of the number of assessed sites multiplied by 2 for diploidy.

Computations were initially carried out on a downsampled set of 10 x 100,000 contiguous sites genome wide to be computationally feasible.

This use of chunks was convenient to leverage the underlying storage of data in `zarr` format.

Where a pair of samples fell beneath a conservative threshold of 0.012, the true genetic distance (i.e. without downsampling) was computed across all sites.

For each pair of technical replicates, we excluded both members of the pair where genetic distance was above 0.006.

Where replicate pairs met the concordance threshold we excluded the lower quality sample.

Quality was determined based on the skewdness of mean to median, i.e. $1 - |mean_{cov}/median_{cov}|$.

The sample with the lower value was preferred as it suggests a more normal coverage distribution.

To identify unknown replicate pairs as a result of sample mix ups or mislabelling, we screened within submission sets for unexpected pairs, using the genetic distance cut off of 0.006 as above.

We did not attempt to identify unknown replicate pairs in the AG1000G-X submission set, made up of laboratory experimental crosses, due to familial similarity and high levels of inbreeding.

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.

Using the PCA implementation in `scikit-allel v2.1.0` we downsampled to 100,000 segregating non-singleton sites from chromosomes 3R and 3L, to avoid regions complicated by known introgression loci or paracentric inversions.

Multiallelic sites were included as dummy rows by melting the data structure.

A careful review process identified: a) samples that dominated single principal components, either individually or in very small numbers.

This suggests an individual belonging to another Anopheline species, or some inherent problem with the sample.

Or b) samples that clustered with other samples inconsistently with metadata.

This was a subjective assessment, but bore in mind the given collection location, time, and PCR species assignment (where available) of the sample.

Multiple geographical sites submitted by the same partner were also considered, where sample mix ups formed the most parsimonious explanations of incongruent clustering.

Species assignment and sex calling

Ancestry informative markers were used to assign species in our cohort.

To derive markers informative between *A. arabiensis* and *A. gambiae s.l.*, we used publicly available data from the 16 genomes project (ref).

Whole genome SNP calls called against the AgamP3 reference for 12 *A. arabiensis* and 38 *A. gambiae s.l* individuals were available.

Alleles were mapped onto the same alternate allele space, the frequencies of which were computed across both groups.

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.

The full AIM set of positions and alleles are available as part of the phase 3 data release.

Called genotypes for each individual in the dataset were cross referenced against a random subset of 50,000 ancestry informative marker alleles, genotype alleles were accordingly classified as gambiae-like or arabiensis-like.

AIM fractions were cross referenced against PCR results available from a subset of inviduals.

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.

Therefore, classification requirements are less rigourous than other sets of validated markers; individuals were classed as *A. arabiensis* where a fraction >0.6 of alleles were arabiensis-like (n=368), and as *A. gambiae s.l* where this value was < 0.03 (n=2415).

Species were not assigned to samples from the AG1000G-X submission due to inbreeding and high levels of genetic drift.

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

To resove the *A. gambiae s.l* individuals into *A. gambiae* and *A. coluzzii* we applied the 729 AIMs previously identified by Neafsey et al (ref).

?? Need to find code that creates gambiae from neafsey set

Cutoffs were made at <0.12 (gambiae) and >0.9 (coluzzii), with individuals between classed as intermediate.

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

SNP filtering and quality

Site filtering ensures that reported variation is of high quality.

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 using a decision tree model, and are appropriate for analyses that include *gambiae* and *coluzzii* samples only.

Inputs to the decision tree model are summary statistics from the set of SNP calls and genotype alignments.

The `arab` site filters were generated following application of the resulting model to the summary statistics from arabiensis samples in the cohort, 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.

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

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 predict the presence or absence of mendelian error, becoming a binary classification problem.

Summary statistics used as input to the model are presented in table (ref).

Pedigrees included *A. gambiae* and *A. coluzzii* mosquitoes only, and summary statistics to build the initial site filters model came from females of these species (n=2415).

Males were excluded, so that the model could also be applied without modification to the X chromosome.

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

Sites were defined as PASS where all genotypes across all 10 remaining crosses were called, and no mendelian inconsistencies were observed.

Sites were defined as FAIL where a mendelian inconsistency was observed in any pedigree.

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

A balanced training set was generated from the remaining 10 crosses containing XXX autosomal(?) sites.

We applied a decision tree, as it provides decisions with clear explanations, and is similar in concept to the set of hard thresholds commonly used in SNP calling in non-model organisms.

A set of trees with different parameter settings were learned, exploring the depth of trees and the number of samples allowed at a terminal node.

Parameter settings were evaluated on an unbalanced evaluation set, consisting of XXX sites randomly from sampled from the whole genome.

Leaves of the trained models contain different proportions of PASS sites, 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 parameter set.

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

The resulting model was a decision tree of depth 8, with a maximum of XX (CHECK NUMBERS) terminal nodes, where leaves were assigned to PASS where > 0.533 of training data in that leaf were PASS.

All sites in the genome were then assigned to PASS or FAIL given the model inputs.

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

Above definitions of PASS sites were retained, but independently within pedigrees, providing 5 distinct evaluation sets.

To evaluate performance on the hemizygous X chromosome we use the more direct measure of heterozygote calls in males.

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

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

A non heterozygote call is labelled PASS, and a heterozygote call FAIL.

Error rates were computed for all crosses, over all chromosomes before and after application of site filters.

We report the false positive rate (FPR), (i.e proportion of the genome considered accessible), and the Youden statistic (ratio of sensitivity to specificity).

References