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@01b2ab0 on September 11, 2020.

Authors

- **The Anopheles gambiae 1000 Genomes Consortium**

Abstract

Population Sampling

DNA extracted from wild-caught *Anopheles* mosquitoes were submitted to the Ag1000G consortium in 23 sets by consortial partners. (chris' line)

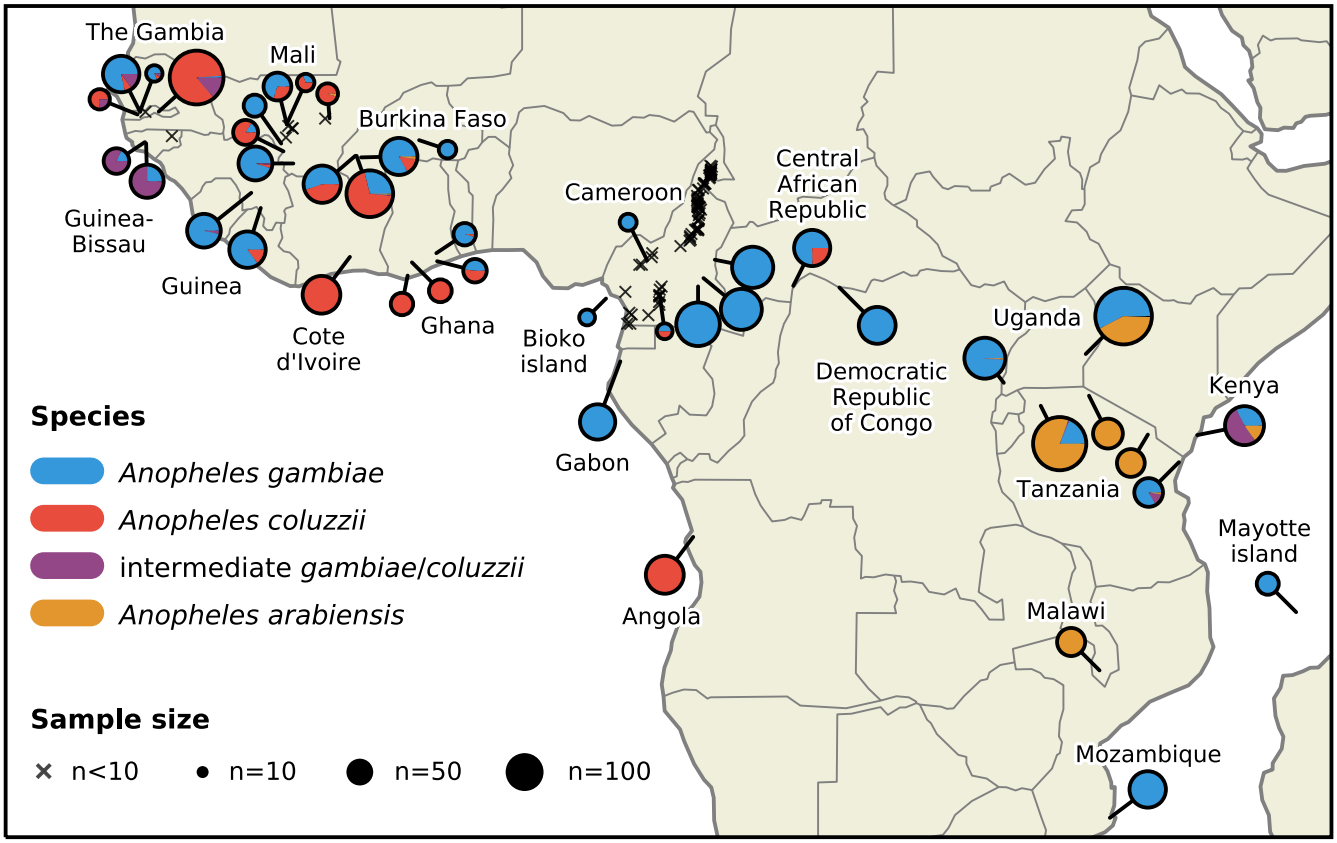

Figure 1: Sample Collection Map

Whole Genome Sequencing and Alignment

A total of 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 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 250.

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

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 the autosomal vs X coverage ratio not following the expected bimodal distribution (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, in total 407 samples were excluded, based on skewedness of the mean vs median.

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 the requirements of this stage 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.

Full details including exclusion thresholds are available in supplementary.

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.

Therefore, to assign species we looked beyond the conventional set of PCR based markers and applied a wider set of ancestry informative markers (AIMs).

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 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.

Owing to DNA availability, no second technology was available for direct benchmarking.

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.

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

5 of the 15 crosses were held out for validation, so performance could be evaluated against the previous site filtering scheme.

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 clear unambiguous decisions, and is similar in concept to the set of filters commonly used in non-model organism genomics.

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.

The 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.

The above definitions of PASS sites were retained, but independently over pedigrees, providing 5 distinct evaluation sets.

Before applying the site filters, the mendelian error rate of the 5 crosses over all autosomal sites ranged between XXX and XXX (table XXX).

The application of the site filters mask defines the accessible fraction of the genome at 70%, and reduces the mendelian error rate by a median factor of 10x on the autosomes.

In all 5 crosses the Youden score was substantially increased by a median factor of XXX.

On the hemizygous X chromosome we can use the more direct measure of heterozygote calls in males.

In the dataset are 220 gambiae/coluzzii male samples, each of which represent an independent proxy for genotype discordance.

TO FINALIZE? (Also some GQ threshold? when applied to the X chromosome).

Pre-application of the site filters, the median heterozygosity rate on X was 0.44%, and post filtering this drops to 0.12% (table XX).

The median fold change in error rate was -1.74, with 69.97% of the X chromosome passing site filters.

Directly comparing the numbers to the phase 2 site filters, we observe similar levels of mendelian error and X heterozygosity, however the updated site filters have a substantially higher sensitivity, yielding a higher Youden score over all crosses and chromosomes.

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	0.00476	0.00285	0.00402	0.00444	0.00522	0.01342	1.00000
post-filtering	220.00000	0.00165	0.00028	0.00078	0.00127	0.00207	0.00882	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. 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

Genetic Diversity within Populations

Insecticide Resistance

Gene Drive

References