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@0fae711 on July 14, 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)

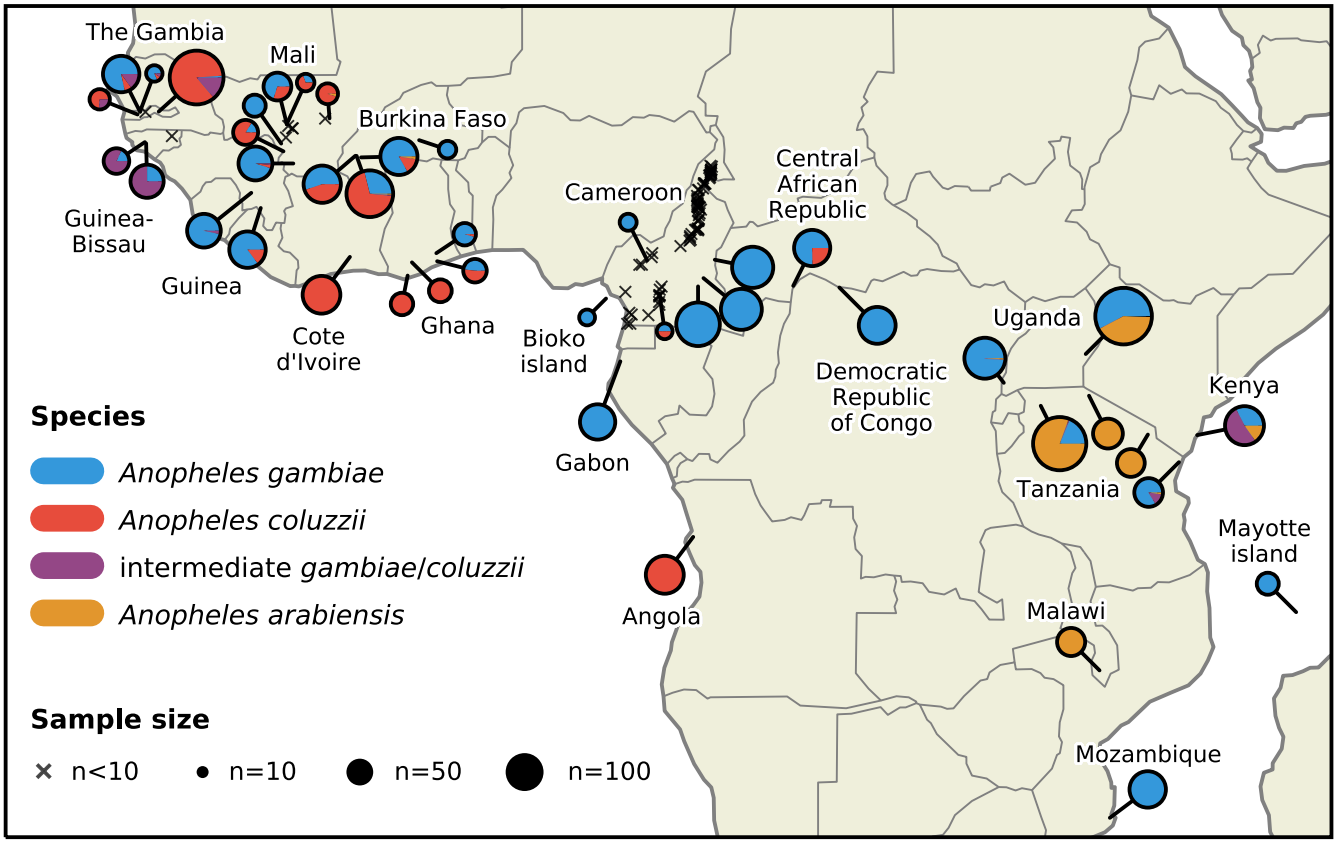

Figure 1: Sample Collection Map

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 in were excluded in favour of better quality samples, 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.

SNP filtering and quality

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

Features of specific regions of the Anopheles genome cause increases in calling errors in short-read technologies; these features include high divergence from the reference, high homology between regions, copy number variation, presence of transposable elements and others.

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

However, using the 15 available Anopheles pedigrees previously described, we were able to use 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, here we built a statistical model where cohort level genome annotations were used to predict the presence of mendelian error, becoming a binary classification problem.

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 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 included.

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

We used 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 50 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.

The error rate of the X chromosome was reduced by a median of XXX (table Y).

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

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

- Table A: Mendel errors per cross per chromosome. row indices: chromosome and raw/filtered column indexes: crosses + frac accessible. ie 10 rows, and 6 columns.

- Table B: comparison of 3 vs 2. row indices: as above column indices: MER, frac accessible, Youden, each for 2 and 3. ie 10 rows, and 6 columns.

Genome accessibility

SNP discovery

Species Assignment

Population Structure

Genetic Diversity within Populations

Insecticide Resistance

Gene Drive

References