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@bfb83a0 on July 23, 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)

Table 1. Population sampling table. Each row represents a unique latitude/longitude where 10 or more mosquitoes were collected (pie charts in **Figure 1**). Sex and species counts for each site are given. "arabiensis x gambiae" gives counts for hybrids between *An. arabiensis* and either *An. gambiae* or *coluzzii*. Sampling on the island of Mayotte consists of seven sites with low numbers at each, these were combined and latitudes/longitudes averaged due to the small size of the island. For collection data on individual samples see **Supplementary Table 1**.

country	site	latitude	longitude	counts	female	male	sex unknown	arabiensis	arabiensis x gambiae	coluzzii	gambiae	gambiae x coluzzii
Angola	Luanda	-8.884	13.302	81	77	4	0	0	0	81	0	0
Bioko	Bioko	3.7	8.7	10	10	0	0	0	0	0	10	0
Burkina Faso	Bana	11.233	-4.472	128	81	47	0	1	0	89	37	1
Burkina Faso	Pala	11.15	-4.235	77	67	10	0	2	0	11	64	0
Burkina Faso	Sourouko udinga	11.235	-4.535	78	77	1	0	0	0	35	43	0
Burkina Faso	Monomtenga	12.06	-1.17	13	13	0	0	0	0	0	13	0
Cameroon	Manda	5.726	10.868	12	12	0	0	0	0	0	12	0
Cameroon	Gado Badzere	5.747	14.442	97	82	15	0	0	0	0	97	0
Cameroon	Daiguene	4.777	13.844	96	81	15	0	0	0	0	96	0
Cameroon	Nkolondom	3.972	11.516	10	10	0	0	0	0	5	5	0
Cameroon	Mayos	4.341	13.558	110	95	15	0	0	0	0	110	0
Central African Republic	Bangui	4.367	18.583	73	73	0	0	0	0	18	55	0
Cote d'Ivoire	Tiassale	5.898	-4.823	80	80	0	0	0	0	80	0	0
Democratic Republic of Congo	Gbadolite	4.283	21.017	76	44	32	0	0	0	0	76	0
Gabon	Libreville	0.384	9.455	69	69	0	0	0	0	0	69	0
Ghana	Takoradi	4.912	-1.774	24	24	0	0	0	0	24	0	0
Ghana	Twifo Praso	5.609	-1.549	25	25	0	0	0	0	25	0	0
Ghana	Madina	5.668	-0.219	27	27	0	0	0	0	14	13	0
Ghana	Koforidua	6.094	-0.261	24	24	0	0	0	0	1	23	0
Guinea	Koraboh	9.28	-10.03	62	50	12	0	0	0	0	60	2
Guinea	Koundara	8.48	-9.53	74	62	12	0	0	0	11	63	0
Guinea-Bissau	Antula	11.891	-15.582	60	60	0	0	0	0	0	15	45
Guinea-Bissau	Safim	11.957	-15.649	33	33	0	0	0	0	0	6	27
Kenya	Kilifi	-3.511	39.909	86	75	11	0	13	0	0	28	45
Malawi	Chikhwawa	-15.933	34.755	41	41	0	0	41	0	0	0	0
Mali	Toumani Oulena	10.83	-7.81	63	52	11	0	0	0	2	60	1
Mali	Kela	11.88	-8.45	23	23	0	0	0	0	0	23	0

country	site	latitude	longitude	counts	female	male	sex unknown	arabiensis	arabiensis x gambiae	coluzzii	gambiae	gambiae x coluzzii
Mali	Ouassorola	12.9	-8.16	13	13	0	0	0	0	9	4	0
Mali	Kababougou	12.89	-8.15	40	32	8	0	0	0	12	28	0
Mali	Douna	13.21	-5.9	20	20	0	0	1	0	19	0	0
Mali	Takan	11.47	-8.33	31	25	6	0	0	0	26	5	0
Mayotte	Mayotte	-12.857	45.137	23	11	12	0	0	0	0	23	0
Mozambique	Furvela	-23.716	35.299	74	74	0	0	0	0	0	74	0
Tanzania	Tarime	-1.431	34.199	47	47	0	0	47	0	0	0	0
Tanzania	Muleba	-1.962	31.651	170	160	10	0	137	0	0	32	1
Tanzania	Moshi	-3.482	37.308	40	39	1	0	40	0	0	0	0
Tanzania	Muheza	-4.94	38.948	43	43	0	0	1	0	0	36	6
The Gambia	Tankular	13.417	-16.033	19	1	0	18	0	0	14	0	5
The Gambia	Njabakunda	13.55	-15.9	74	74	0	0	0	0	5	58	11
The Gambia	Wali Kunda	13.567	-14.917	174	174	0	0	0	0	148	2	24
The Gambia	Sare Samba Sowe	13.583	-15.9	11	1	0	10	0	0	1	9	1
Uganda	Nagongera	0.77	34.026	194	194	0	0	81	1	0	112	0
Uganda	Kihihi	-0.751	29.701	96	96	0	0	1	0	0	95	0

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.

Genomic features cause unavoidable 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.

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 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 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 eligible to be included in model training.

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.

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

Rather than mendelian errors, on the hemizygous chromosome we can use the more direct measure of heterozygote calls in males.

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

Pre-application of the site filters, the median heterzygosity 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.

? (Also some measure of GQ? when applied to the X chromosome).

Direct comparison to the phase 2 site filters is favourable; we observe similar levels of mendelian error, but with substantially higher sensitivity, yielding a higher Youden score over all crosses and chromosomes.

- 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

SNP discovery

Species Assignment

Population Structure

Genetic Diversity within Populations

Insecticide Resistance

Gene Drive

References