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@fa5b0fb on July 17, 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)

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

table demo

Africa centroids

name_long	pop_est	gdp_md_est	lastcensus	Longitude	Latitude
Angola	12799293	110300	1970	17.53736768	-12.29336054
Burundi	8988091	3102	2008	29.87512156	-3.35939666
Benin	8791832	12830	2002	2.32785254	9.6417597
Burkina Faso	15746232	17820	2006	-1.75456601	12.26953846
Botswana	1990876	27060	2011	23.79853368	-22.18403213
Bioko	334463		2015	8.749618	3.616311
Central African Republic	4511488	3198	2003	20.46826831	6.56823297
Cote d'Ivoire	20617068	33850	1998	-5.5692157	7.6284262
Cameroon	18879301	42750	2005	12.73964156	5.69109849
Democratic Republic of Congo	68692542	20640	1984	23.64396107	-2.87746289
Republic of Congo	4012809	15350	2007	15.21965762	-0.83787463
Comoros	752438	751.2	2003	43.68253968	-11.87783444
Cape Verde	429474	1626	2010	-23.9598882	15.95523324
Djibouti	516055	1885	2009	42.5606754	11.74871806
Algeria	34178188	232900	2008	2.61732301	28.15893849
Egypt	83082869	443700	2006	29.86190099	26.49593311
Eritrea	5647168	3945	1984	38.84617011	15.36186618
Ethiopia	85237338	68770	2007	39.60080098	8.62278679
Gabon	1514993	21110	2003	11.7886287	-0.58660025
Ghana	23832495	34200	2010	-1.21676566	6.85345644
Guinea	10057975	10600	1996	-10.94066612	10.43621593
The Gambia	1782893	2272	2003	-15.39601295	13.44965244
Guinea-Bissau	1533964	904.2	2009	-14.94972445	12.04744948
Equatorial Guinea	650702	14060	2002	10.34137924	1.70555135
Kenya	39002772	61510	2009	37.79593973	0.59988022
Liberia	3441790	1526	2008	-9.32207573	6.45278492
Libya	6310434	88830	2006	18.00866169	27.03094495
Lesotho	2130819	3293	2006	28.22723131	-29.58003188
Morocco	34859364	136600	2004	-8.45615795	29.83762955
Madagascar	20653556	20130	1993	46.70473674	-19.37189587
Mali	12666987	14590	2009	-3.54269065	17.34581581
Mayotte	270372	3550	2019	45.156544	-12.796385
Mozambique	21669278	18940	2007	35.53367543	-17.27381643
Mauritania	3129486	6308	2000	-10.34779815	20.25736706
Malawi	14268711	11810	2008	34.28935599	-13.21808088
Namibia	2108665	13250	2001	17.20963567	-22.13032568
Niger	15306252	10040	2001	9.38545882	17.41912493
Nigeria	149229090	335400	2006	8.08943895	9.59411452
Rwanda	10473282	9706	2002	29.91988515	-1.99033832
Western Sahara	-99	-99	-99	-12.21982755	24.22956739
Sudan	25946220	88080	2008	29.94046812	15.99035669
South Sudan	10625176	13227	2008	30.24790002	7.30877945

name_long	pop_est	gdp_md_est	lastcensus	Longitude	Latitude
Senegal	13711597	21980	2002	-14.4734924	14.36624173
Sierra Leone	6440053	4285	2004	-11.79271247	8.56329593
Somaliland	3500000	12250	-99	46.25198395	9.73345496
Somalia	9832017	5524	1987	45.70714487	4.75062876
SaoTome and Principe	212679	276.5	2001	6.72429658	0.44391445
Swaziland	1123913	5702	2007	31.4819369	-26.55843045
Chad	10329208	15860	2009	18.64492513	15.33333758
Togo	6019877	5118	2010	0.96232845	8.52531356
Tunisia	10486339	81710	2004	9.55288359	34.11956246
Tanzania	41048532	54250	2002	34.81309981	-6.27565408
Uganda	32369558	39380	2002	32.36907971	1.27469299
South Africa	49052489	491000	2001	25.08390093	-29.00034095
Zambia	11862740	17500	2010	27.77475946	-13.45824152
Zimbabwe	12619600	9323	2002	29.8514412	-19.00420419

References