Anopheles gambiae 1000 Genomes Project (Ag1000G) phase 3

SNP calling methods

The *Anopheles gambiae* 1000 Genomes Consortium

January 14, 2021

Summary

This document contains a brief description of sequencing and variant calling methods for the Ag1000G phase 3 single nucleotide polymorphism (SNP) data release. This is a preliminary version of methods that will be released in final form as part of a future publication by the Ag1000G Consortium.

Contents

Whole-genome sequencing

All library preparation and sequencing was performed at the Wellcome Sanger Institute. Paired-end multiplex libraries were prepared using the manufacturer's protocol, with the exception that genomic DNA was fragmented using Covaris Adaptive Focused Acoustics rather than nebulization. Multiplexes comprised 12 tagged individual mosquitoes and three lanes of sequencing were generated for each multiplex to even out variations in yield between sequencing runs. Cluster generation and sequencing were undertaken according to the manufacturer's protocol for paired-end sequence reads with insert size in the range 100–200 bp. 4,693 individual mosquitoes were sequenced in total, of which 3,130 were sequenced using the Illumina HiSeq 2000 platform and 1,563 were sequenced using the Illumina HiSeq X platform. All individuals were sequenced to a target coverage of 30×. The HiSeq 2000 sequencing runs generated 100 bp paired-end reads, and the HiSeq X sequencing runs generated 150 bp paired-end reads.

Alignment and SNP calling

Reads were aligned to the AgamP4 reference genome using BWA version 0.7.15. Indel realignment was performed using GATK version 3.7-0 `RealignerTargetCreator` and `IndelRealigner`. Single nucleotide polymorphisms were called using GATK version 3.7-0 `UnifiedGenotyper`. Genotypes were called for each sample independently, in genotyping mode, given all possible alleles at all genomic sites where the reference base was not "N". Coverage was capped at 250× by random down-sampling. Complete specifications of the alignment[1] and genotyping[2] pipelines are available from the malariagen/pipelines GitHub repository[3]. Open source WDL implementations of the alignment[4] and genotyping[5] pipelines are also available from GitHub. Following successful completion of these pipelines, samples entered the sample quality control (QC) process.

Sample QC

The following subsections describe analyses performed to identify and exclude samples from the final dataset.

[1] `https://github.com/malariagen/pipelines/blob/v0.0.4/docs/specs/short-read-alignment-vector.md`

[2] `https://github.com/malariagen/pipelines/blob/v0.0.4/docs/specs/snp-genotyping-vector.md`

[3] `https://github.com/malariagen/pipelines`

[4] `https://github.com/malariagen/pipelines/tree/v0.0.4/pipelines/short-read-alignment-vector`

[5] `https://github.com/malariagen/pipelines/tree/v0.0.4/pipelines/SNP-genotyping-vector`

Coverage

For each sample, depth of coverage was computed at all genome positions. Samples were excluded if median coverage across all chromosomes was less than 10×, or if less than 50% of the reference genome was covered by at least 1×.

Cross-contamination

To identify samples affected by cross-contamination, we implemented the model for detecting contamination in NGS alignments described in Jun et al. (2012). Briefly, the method estimates the likelihood of the observed alternate and reference allele counts under different contamination fractions, given approximate population allele frequencies. Population allele frequencies were estimated from the Ag1000G phase 2 data release (The Anopheles gambiae 1000 Genomes Consortium 2020). The model computes a maximum likelihood value for a parameter α representing percentage contamination. Samples were excluded if α was 4.5% or greater.

Technical replicates

A number of samples were sequenced more than once within this project phase (technical replicates). To create a final dataset without any replicates suitable for population genetic analysis, we performed an analysis to confirm all technical replicates, and to choose the sample within each replicate with the best sequencing data. We computed pairwise genetic distance between all sample pairs within a submission set. The distance metric used was city block distance between genotype allele counts, to allow for handling of multiallelic SNPs. So, e.g., distance between genotypes of 0/1 and 0/1 is 0, distance between 0/0 and 0/1 is 2, distance between 0/1 and 1/2 is 2, distance between 0/0 and 1/1 is 4, etc. For each pair of samples, distance was averaged over all sites where both samples had a non-missing genotype call. Computations were initially carried out on a down-sampled set of 10 x 100,000 contiguous genomic sites, to be computationally feasible. Where a pair of samples fell beneath a conservative threshold of 0.012, the genetic distance was then recomputed across all genomic sites (i.e., without down-sampling). For each pair of samples that were expected to be technical replicates according to our metadata records, we excluded both members of the pair if genetic distance was above 0.006. Where an expected replicate pair had genetic distance below 0.006, we retained only one sample in the pair. We also identified and excluded both samples in any pair where genetic distance was below 0.006, but the samples were not expected to be replicates.

Population outliers

We used principal component analysis (PCA) to identify and exclude individual samples that were population outliers. SNPs were down-sampled to use 100,000 segregating non-singleton sites from chromosomes 3R and 3L, to avoid regions complicated by known introgression loci or paracentric inversions. PCA was computed using scikit-allel version 1.2.0. We iteratively identified and excluded any individual samples that were outliers

along a single principal component. We then identified and excluded any individual samples or small sample groups that clustered together with other samples in a way that was not plausible given metadata regarding their collection location.

Colony crosses

Samples in the `AG1000G-X` sample set were parents and progeny from colony crosses and were subject to a slightly different set of QC steps. For each cross, we performed an analysis of Mendelian inheritance and consistency to confirm the true parents and the validity of the cross. Not all crosses were able to be successfully resolved, and samples that were not in a resolved cross were excluded. From the samples originally submitted in the AG1000G-X sample set, 297 samples from 15 crosses were retained for release. We did not include the colony crosses in the population outlier analysis due to their relatedness.

Sex calling

We called the gender of all samples based on the modal coverage ratio between the X chromosome and the autosomal chromosome arm 3R. 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 sample was excluded. One of the sample sets from The Gambia, AG1000G-GM-B, included whole-genome amplified (WGA) samples which displayed some skew in their coverage ratios, which meant that sex could not be called via the same process. These samples received a sex call where possible, but no samples were excluded based on uncertain sex call.

Species assignment

We assigned a species to each individual that passed sample QC using their genomic data, via two independent methods: ancestry-informative markers (AIMs) and principal components analysis (PCA).

Species calling via ancestry-informative markers

To derive AIMs between *An. arabiensis* and *An. gambiae*, we used publicly available data from the *Anopheles* 16 genomes project (Neafsey et al. 2014). Whole genome SNP calls for 12 *An. arabiensis* and 38 *An. gambiae* individuals were used. Alleles were mapped onto the same alternate allele space, and allele frequencies were computed for both species. 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. We randomly down-sampled these SNPs to a set of 50,000 AIMs, then computed the fraction of alleles at these SNPs that were arabiensis-like for each

individual in the Ag1000G phase 3 cohort. Given the relatively small number of *An. arabiensis* samples in the 16 genomes project, it was clear that a significant proportion of putative AIMs were not likely to be truly informative across the broader sampling in Ag1000G. Individuals in Ag1000G were classed as *An. arabiensis* where a fraction >0.6 of alleles were arabiensis-like.

To resolve the non-*An. arabiensis* individuals into *An. gambiae* and *An. coluzzii*, we applied the AIMs previously used in The Anopheles gambiae 1000 Genomes Consortium (2020). For each individual, we computed the fraction of coluzzii-like alleles at these AIMs. Individuals were called as *An. gambiae* where this fraction was <0.12 and *An. coluzzii* where this fraction was >0.9, with individuals in between classed as intermediate.

Species calling via principal components analysis

To provide a complementary view of species assignments, we also used the results of the principal components analysis of Chromosome 3 computed during the outlier analysis described above. Based on a comparison with the AIM species calls, it was apparent that the first two principal components could be used to assign species. Individuals where PC1 > 150 were called as *An. arabiensis*. Individuals where PC1 < 0 and PC2 > -7 were called as *An. gambiae*. Individuals where PC1 < 0 and PC2 < -24 were called as *An. coluzzii*. All other individuals were called as intermediate. The results of the PCA and AIM species calls were highly concordant in most sample sets, except for the Far West (Guinea-Bissau, The Gambiae) and Far East (Kenya, Tanzania). Further investigation is required to resolve the species status of these individuals.

Site filtering

We developed filters that identify genomic sites where SNP calling and genotyping is likely to be less reliable in one or more mosquito species. To guide the design and calibration of the site filters, we made use of the 15 colony crosses included in Ag1000G phase 3. Each cross comprises two parents and up to 20 progeny, and thus it is possible to identify sites where genotypes in one or more progeny are not consistent with Mendelian inheritance (Mendelian errors). A small number of Mendelian errors may be due to *de novo* mutation, but the vast majority of Mendelian errors are likely to be due to errors in sequencing, alignment or SNP calling. The general approach we took was to use Mendelian consistency to identify sets of positive and negative training sites, then used these to train a machine learning model that classified all genomic sites as either PASS or FAIL.

Site filters for use with *An. gambiae* and/or *An. coluzzii*

All the 15 crosses involved *An. gambiae* and/or *An. coluzzii* parents, and none of the crosses involved *An. arabiensis*, so we used the crosses to first develop site filters suitable for use with *An. gambiae* and/or *An. coluzzii*. Hereafter we refer to these filters as the "`gamb_colu`" site filters. Five of the 15 crosses were held out for validation, so performance

could be evaluated objectively. Sites were assigned to the positive training set where all genotypes across all 10 crosses were called, and no Mendelian errors were observed. Sites were assigned to negative training set where one or more Mendelian errors were observed in any cross. All other sites were not considered eligible for inclusion in model training. A balanced training set was then generated containing 100,000 autosomal sites from each of the positive and negative training sets.

The inputs to the machine learning model were a set of per-site summary statistics computed from the sequence read alignments and SNP genotypes across all wild-caught *An. gambiae* and *An. coluzzii* individuals. These input summary statistics are described further in the appendix. Male individuals were excluded from the summary statistic calculations, so that the model could also be applied without modification to the X chromosome. We used these summary statistics, together with the positive and negative training sites, to train a decision tree model. We initially trained a set of trees with different hyperparameter values, exploring the depth of trees, and the number of samples allowed at a terminal node. Each of these trees was evaluated on an unbalanced set of sites randomly sampled from the whole genome (2% of all sites, without replacement). Leaves of these trees contained different proportions of positive and negative training sites, and 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 set of hyperparameter values. The best performing hyperparameter set based on AUROC was selected as the final model, and the leaf classification cutoff used was optimised based on the Youden statistic. The resulting model was a decision tree of depth 8, where leaves were assigned to PASS where > 0.533 of training data in that leaf were positive training sites. All sites in the genome were then assigned to PASS or FAIL via this model.

The 5 remaining cross pedigrees were used to perform a final evaluation of the approach. For each of these crosses, we computed the Mendelian error rate (fraction of variants with one or more Mendelian errors among progeny) before and after applying the site filters, to provide five independent evaluation results. We also evaluated performance on the X chromosome using heterozygote calls in males as an error indicator. The fraction of variants with a heterozygous genotype call in or more males was computed before and after applying site filters. Male error rates were estimated from genotype calls with a minimum Genotype Quality (GQ) value of 30. Performance of the decision tree model was better than the hand-crafted site filters created during the previous project phase (Ag1000G phase 2), with lower Mendelian error rates, and a larger number of sites passing the filter. Full performance metrics will be reported in a future publication.

Site filters for use with *An. arabiensis*

To generate site filters for use with *An. arabiensis*, we recomputed site summary statistics using only wild-caught *An. arabiensis* individuals, then applied the decision tree model described above. These filters, which we refer to as the "`arab`" site filters, are appropriate when working with *An. arabiensis* samples only.

Site filters for joint analyses of all three species

We created site filters suitable for joint analysis of individuals from all three species by taking the intersection of the `gamb_colu` and the `arab` site filters. We refer to these filters as the "`gamb_colu_arab`" site filters.

Acknowledgments

We would like to thank the staff of the Wellcome Sanger Institute Sample Logistics, Sequencing and Informatics facilities for their contributions to the production of this data release. We would like to thank the members of the Data Engineering team of the Broad Institute of Harvard and MIT for their work on open source implementations of the alignment and SNP calling pipelines used in Ag1000G phase 3.

Further information

For further information about the Ag1000G project, please visit `https://www.malariagen.net/ag1000g`. For further information about the Ag1000G phase 3 SNP data release, please visit `www.malariagen.net/data/ag1000g-phase3-snp`. If you have any questions regarding the data release, please start a new discussion at `https://github.com/malariagen/vector-public-data/discussions`.

Appendix

Summary statistics used as input to site filter modelling

Name	Description
No Coverage	Number of samples with no coverage whatsoever at the given position
Low Coverage	Number of samples with low depth of coverage at the given position (less than half modal coverage for whole chromosome)
Low Coverage GC normed	Number of samples with low depth of coverage at the given position (less than half modal coverage for whole chromosome)
High Coverage	Number of samples with high depth of coverage at the given position (more than twice modal coverage for the whole chromosome)
High Coverage GC normed	Number of samples with high depth of coverage at the given position (more than twice modal coverage for the whole chromosome)
Mean GQ	Mean of GQ
Median GQ	Median of GQ
Variance GQ	Variance of GQ
Low GQ 30	Number of samples with a GQ of <30
Low GQ 10	Number of samples with a GQ <10
Low MQ	Number of samples with an MQ (RMS) <30
Median MQ	Median of mapping quality (RMS)
Mean MQ	Mean of MQ
Variance MQ	Variance of MQ
Error Fraction	Fraction of samples with an allele not called in the genotype
Allele Balance Het	Product of binomial probability of read counts at het calls