An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples

MalariaGEN *Plasmodium falciparum* Community Project

Details of the bioinformatics methods

Read mapping and coverage analysis

Reads mapping to the human reference genome were discarded before all analyses, and the remaining reads were mapped to the *P. falciparum* 3D7 v3 reference genome (ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/2016-07/Pfalciparum.genome.fasta.gz) using `bwa mem`[1] version 0.7.15 with -M parameter to mark shorter split hits as secondary.

For the steps requiring a known set of variants (base quality score recalibration and variant quality score recalibration) we used the PASS variants from the Pf crosses project (http://www.malariagen.net/data/pf-crosses-1.0).

BAM improvement steps were applied to the read mapping outputs before further analyses. The Picard (http://picard.sourceforge.net) tools `CleanSam`, `FixMateInformation` and `MarkDuplicates` were successively applied to the BAM files of each sample, using Picard version 2.6.0. GATK base quality score recalibration was applied using default parameters, and using the PASS variants from the Pf crosses 1.0 release as a set of known sites. All lanes from each library were merged to create library-level BAM files, and then all libraries for each samples were merged to create sample-level BAM files. The output of this stage was a set of 7,182 improved BAM files, one for each sample.

Standard alignment metrics were generated for each sample using the `stats` utility from samtools version 1.2[2]. We also used GATK's `CallableLoci`[3] with default parameters to determine the genomic positions callable in each sample.

Variant discovery and genotyping

We discovered potential SNPs and indels by running GATK's HaplotypeCaller[3] independently across each of the 7,182 sample-level BAM files. The following GATK parameters were used: `--emitRefConfidence GVCF --variant_index_type LINEAR --variant_index_parameter 128000 --max_alternate_alleles 6`

This resulted in the creation of 7,182 GVCF files. We merged these for each of the 16 reference sequences (14 chromosomes, 1 apicoplast and 1 mitochondria) using

GATK's `CombineGVCFs`. Each of the 16 reference sequences was then genotyped using GATK's `GenotypeGCVFs` with `--max_alternate_alleles 6`

Variant filtering and annotation

SNPs and indels were filtered separately. For each class of variant, filtering was done in two stages:

1) Each variant was assigned a quality score using GATK's Variant Quality Score Recalibration (VQSR). The tools `VariantRecalibrator` and `ApplyRecalibration` are used here.
2) Regions of the genome which we previously identified as being enriched for errors[4] are masked out.

`VariantRecalibrator` was run using the PASS variants from the Pf crosses 1.0 release as a training set with a prior of 15.0. For SNPs we used the following parameters: `-an QD -an FS -an SOR -an DP --maxGaussians 8 --MQCapForLogitJitterTransform 70`. For indels we used the following parameters: `-an QD -an DP -an SOR -an FS --maxGaussians 4 --MQCapForLogitJitterTransform 70`. `ApplyRecalibration` was then run to assign each variant a quality score named VQSLOD. High values of VQSLOD indicate higher quality. Variants (both SNPs and indels) with a VQSLOD score ≤ 0 were filtered out. As expected, the pass rate for SNPs was higher for coding variants (66%) than for non-coding variants (47%).

Variants in the VCFs were annotated using a number of different methods. Functional annotations were applied using snpEff[5] version 4.1, with gene annotations downloaded from GeneDB[6] at ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/2016-10/Pfalciparum.noseq.gff3.gz. The following options were used with snpEff: `-no-downstream -no-upstream -onlyProtein`.

Genome regions were annotated using `vcftools` and masked if they were outside the core genome. Variants in the apicoplast and mitochondrion were annotated *Apicoplast* and *Mitochondrion* respectively and masked by adding these annotations to the FILTER column. Subtelomeric regions in the 14 chromosomal sequences were identified by using the classification used in the Pf crosses 1.0 release[4]. Variants in these subtelomeric regions were annotated *SubtelomericHypervariable* or *SubtelomericRepeat* and masked by adding this annotation to the FILTER column. The internal *var* genes regions were annotated as *InternalHypervariable* and masked by adding this annotation to the FILTER column. The centromeres were annotated as *Centromere* and masked by adding this annotation to the FILTER column.

We removed 69 samples from lab studies to create the release VCF files which contain 7,113 samples. VCF files were converted to zarr format using `scikit-`

allel v 1.1.8 (https://github.com/cggh/scikit-allel) and subsequent analyses performed using the zarr files.

Genetic distance

We calculate genetic distance between samples using biallelic SNPs that pass filters. For each SNP in sample i we calculate the non-reference allele frequency f_i as the proportion of reads that carry the non-reference allele. For clonal samples, f_i should be either 0 (for homozygous reference allele calls) or 1 (for homozygous alternative allele calls). For samples containing mixtures of different strains, we should expect fractional values of f_i for heterozygous calls. f_i is set to 0 if there are < 2 or <5% alternative allele reads, and likewise to 1 if there are < 2 or <5% reference allele reads. We do not calculate f_i when there were less than 5 reads in total. Genetic distance between sample 1 and 2 is calculated as $f_1(1 - f_2) + f_2(1 - f_1)$. For each sample pair we calculate the mean genetic distance across all SNPs for which we have an estimate of f_i in each sample.

In addition to calculating genetic distance between all pairs of samples from the current data set, we also calculated the genetic distance between each sample and the lab strains 3D7, 7G8, GB4, HB3 and Dd2 from the Pf3k project

Species identification

We identified species using nucleotide sequence from reads mapping to six different loci in the mitochondrial genome, using custom java code (https://github.com/malariagen/GeneticReportCard). The loci were located within the *cox3* gene (PF3D7_MIT01400), as described in a previously published species detection method.[7] Alleles at various mitochondrial positions within the six loci were genotyped and used for classification as shown in Supplementary Table 14. A sample is assigned a species if it matches at least two of the six loci. At any given locus, the sample is considered a match to a species only if all the positions at that locus carry the matching allele.

Sample QC

We created a final analysis set of 5,970 samples after removing replicate, low coverage, expected mislabelled and mixed-species samples.

We calculated genome callability of each sample using GATK CallableLoci with a minimum depth of 5. Where we had multiple samples from the same individual, we removed samples with lower callability to leave a single sample for each individual in the final analysis set. This removed 487 samples. A further 591 samples with callability <50% were also removed.

We then compared mean genetic distance of each sample to data on lab strains 3D7, 7G8, GB4, HB3 and Dd2 from the Pf3k project (https://www.malariagen.net/projects/pf3k), in order to identify samples with significant DNA contamination from commonly used lab strains. For this analysis we restricted the SNP set to those that were discovered in both this dataset and in Pf3k. Where the mean distance was < 2.5 x 10^{-4} we assumed that the sample contained DNA from the matching lab strain and had been mislabelled or significantly contaminated, thus removing a further sixteen samples from the analysis set.

We also removed samples that were genetically similar to other samples from a different continent which could be due to mislabelling or unreported travel history. We first identified nine samples for which the majority of the eleven nearest neighbours were samples from a different continent. We subsequently removed one further sample which clustered together with samples from a different continent on visual inspection of a neighbour-joining tree (NJT) of all samples. NJTs were produced using the R package `ape`.

Finally, we removed 41 samples from the analysis set that were identified as containing mixed species, as this can affect genotyping of highly conserved loci (e.g. *kelch13*). The final analysis set contained 5,970 QC pass samples.

CNV genotypes at drug resistance loci

We used two complementary methods to determine tandem duplication genotypes around *mdr1*, *plasmepsin2/3* and *gch1*, namely a coverage-based method and a method based on position and orientation of reads near discovered duplication breakpoints. The outline algorithm is as follows:

1. Determine copy number at each locus using a coverage based hidden Markov model (HMM)
2. Manually determine breakpoints of identified duplications by manual inspection of reads
3. Automatically determine face-away read pairs around all sets of breakpoints
4. For each locus in each sample, initially set copy number to that determined by the HMM if <= 10 CNVs discovered in total, else consider undetermined
5. For the *plasmepsin2/3* locus in each sample:
    - If > 100 reads around breakpoints none of which are in face-away pairs, set copy number to 1, regardless of results from HMM
    - Else if > 100 reads, at least 2.5% of which are in face-away pairs and HMM has set copy number to 1 or undetermined, change copy number to 1.5
    - Else use copy number from HMM
6. For the *mdr1 and gch1* loci in each sample:
    - If > 100 reads, at least 2.5% of which are in face-away pairs and HMM has set copy number to 1 or undetermined, change copy number to 1.5
    - Else use copy number from HMM

7. For each locus in each sample, set the breakpoint to be that with the highest proportion of face-away reads

The rationale for and details of the above algorithm are given below.

We initially called amplifications around *mdr1*, *plasmepsin2/3* and *gch1* in an interim data set containing a subset of the final sample set using a coverage-based hidden Markov model (HMM) as previously described[4]. Where we identified > 10 distinct putative CNVs in the core regions of chromosomes 5, 12 and 14, we conservatively set any HMM-based copy number call as 'undetermined', under the assumption that such a high density was biologically unlikely and more probably due to sequencing and/or mapping artefacts. Following this, we manually determined breakpoints of tandem duplications around *mdr1*, *gch1* and *plasmepsin2/3* by visual inspection of soft clipped reads and paired reads aligning either facing away from each other or aligned in the same orientation. This resulted in the identification of 27 pairs of tandem duplication breakpoints around *mdr1*, 9 pairs of tandem duplication breakpoints around *gch1*, and 3 pairs of tandem duplication breakpoints around *plasmepsin 2/3* (Supplementary Tables 4-6). In addition, we identified three sets of four breakpoints (1 set for *mdr1* and 2 sets for *gch1*), which are consistent with DUP-TRP/INV-DUP rearrangements[8]. To our knowledge, this is the first report of such events in *Plasmodium* parasites.

For each set of tandem duplications we identified regions of 600bp near the breakpoints (starting 100bp before the first breakpoint and 600bp before the second breakpoint), and identified pairs of reads where each read mapped starting within one of the two regions, and where the reads were oriented facing away from each other. Through manual inspection, we determined that where more than 2.5% of all reads mapping in such regions were in such face-away orientation, the amplification could be called reliably, regardless of whether it was called in the same sample by the coverage-based approach. Manual inspection of samples with >2.5% of reads mapping in face-away orientation, for which the HMM didn't detected a change in coverage, revealed that in most cases there appeared to be an increase that was less than a doubling, a possible indication that samples contain a mixture of clones within different duplication genotypes; such copy numbers were arbitrarily set to 1.5. Where samples had at least 1 read but less than 2.5% reads mapping in face-away orientation, or where there were < 100 reads mapping to the first 600bp region, we assumed that the breakpoint read evidence was inconclusive. In such cases we relied on the coverage-based HMM approach. From manual inspection, we found that we had found breakpoints around *plasmepsin2/3* for all of the duplications called reliably by the coverage HMM. For *mdr1* and *gch1*, we found many samples for which the coverage data looked conclusive for the presence of the duplication, but for which we could not find breakpoints reads. Based on these results, where we found at least 100 reads at all discovered breakpoint we took a different approach for *plasmepsin2/3* than we did for both *mdr1* and *gch1*. For *plasmepsin2/3*, we assumed that the lack of breakpoint-spanning read pairs signified the lack of *plasmepsin* duplications, and hence called the sample as single-copy. For *mdr1* and *gch1*, where we found no breakpoint-spanning read pairs, we deferred to the coverage HMM for amplification calls.

If we identified face-away reads from more than one pair of breakpoints, we considered the duplication to be due to the set of breakpoints with the highest proportion of face-away reads.

For all samples we report the breakpoint which has the highest proportion of face-away reads. Where a sample has less than 2.5% of reads in face-away pairs, and <= 10 total CNVs called by the HMM, we use the HMM call. For samples with < 2.5% of reads in face-away pairs and > 10 total CNVs, we give a missing duplication genotype call.

HRP2 and HRP3 deletion detection

We created plots of sequence read coverage for the regions around *hrp2* and *hrp3* for all samples. Three analysts independently classified each gene as "deleted" or "non-deleted" for each sample by manual inspection of these plots. In some cases the coverage profile suggested the sample might contain a mixture of deleted and non-deleted genotypes. In such cases samples were conservatively classified as non-deleted. The three sets of calls were then compared and combined in a final consensus classification in the small number of cases where there were discrepancies.

Population structure and characterisation

Neighbour-joining trees (NJTs) were produced using the `nj` implementation in the R package `ape`. Principal coordinate analysis (PCoA) was performed using `scikit-bio` v0.5.5. Based on these observations we grouped the samples into eight geographic regions: South America, West Africa, Central Africa, East Africa, South Asia, the western part of Southeast Asia, the eastern part of Southeast Asia and Oceania, with samples assigned to region based on the geographic location of the sampling site. Five samples from returning travellers were assigned to region based on the reported country of travel.

F_{WS} was calculated using custom python scripts using the method previously described[9]. Nucleotide diversity (π) was calculated in non-overlapping 25kbp genomic windows using the `mean_pairwise_difference` function in `scikit-allel` v1.1.9. We only considered coding biallelic SNPs to reduce the ascertainment bias caused by poor accessibility of non-coding regions. Note that these values are hard to interpret in absolute terms as this set of variants cannot be assumed to be evolutionary neutral and also only accounts for ~50% of the genome.

LD decay (r^2) was calculated using `rogers_huff_r` function in `scikit-allel` v1.1.9.

To calculate mean F_{ST} between populations we used Hudson's method as implemented in `scikit-allel` v1.2.0.

Allele frequencies and F_{ST}

Allele frequencies stratified by geographic regions and sampling sites were calculated using the genotype calls produced by GATK run in diploid mode (see above; GT field in the VCF), with heterozygous calls contributing a non-reference allele frequency of 0.5. We only calculated frequencies at sites for which we had at least 25 QC pass samples. F_{ST} was calculated between all 8 regions, and also between all sites with at least 25 QC pass samples within the regions WAF, EAF, SAS, WSEA, ESEA and OCE (no site in SAM and only one site in CAF had >= 25 samples). F_{ST} between different locations for individual SNPs was calculated using Weir and Cockerham's estimate implemented in `scikit-allel` v1.1.8.

SNP genotypes at drug resistance mutations and samples classification

We initially extracted genotypes at loci implicated in drug resistance from the VCF files (GT fields). At some loci that we could not derive amino acid changes directly from the VCF files because a) the codon contains multiple variable positions, b) some positions within the codon have multi-allelic variants, or, c) as is the case for *crt*, some key mutations are resolved as a combination of two short indels rather than multiple SNPs. We developed a custom `python` script to call amino acids at selected loci by first determining the reference amino acids and then, for each sample, applying all variations using the GT field of the VCF file. We also created calls for amino acids 72-76 in *crt*, and for amino acids 349-726 in *kelch13*. Where a locus included multiple heterozygous variants, we used the PID and PGT VCF fields to phase the variants where possible. For *crt* 72-76 we output 5-amino acid haplotypes, whereas for *kelch13* 349-726, we output all non-synonymous changes seen. For all other loci we output the single amino acid.

The amino acid and copy number calls generated were used to classify all samples into different types of drug resistance. Our methods of classification were heuristic and based on the available data and current knowledge of the molecular mechanisms. Each type of resistance was considered to be either present, absent or unknown for a given sample. The procedure used to map genetic markers to inferred resistance status classification is described in the details for each drug in the accompanying data release (https://www.malariagen.net/resource/26).

Global and local gene differentiation score

We defined the global differentiation score for a gene as $1 - \frac{\log_{10} N}{\log_{10} \max (N)}$, where N is the rank (using the mean rank where there are ties) of the non-synonymous SNP with the highest global F_{ST} value within that gene. To define the local differentiation score, we first calculated for each region containing multiple sites (WAF, EAF, SAS, WSEA, ESEA and OCE) F_{ST} for each SNP between sites within that region. For each gene, we then calculated the rank (using the mean rank where there are ties) of the highest F_{ST} non-synonymous SNP within that gene for each of the six regions. We defined the local differentiation score for each gene using the second highest of these six ranks (N), to ensure that the gene was highly ranked in at least two populations, i.e. to minimise the chance of artefactually ranked a gene highly due to a single variant in a single population. The final local differentiation score was normalised to ensure that the range of possible scores was between 0 and 1, local differentiation score was defined as $1 - \frac{\log_{10} N}{\log_{10} \max (N)}$.

In order to ensure genes weren't highly ranked due to LD with a nearby gene, we also calculated, for each gene, the distance in bp to the nearest gene on the same chromosome with a higher local differentiation score. In Supplementary Table 8 we excluded genes where this distance was < 50,000bp.

References

1 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–60.

2 Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–9.

3 Depristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43**: 491–501.

4 Miles A, Iqbal Z, Vauterin P, *et al.* Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Res* 2016; **26**: 1288–99.

5 Cingolani P, Platts A, Wang LL, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*; **6**: 80–92.

6 Logan-Klumpler FJ, De Silva N, Boehme U, *et al.* GeneDB--an annotation database for pathogens. *Nucleic Acids Res* 2012; **40**: D98-108.

7 Echeverry DF, Deason NA, Davidson J, *et al.* Human malaria diagnosis using a single-step direct-PCR based on the Plasmodium cytochrome oxidase III gene. *Malar J* 2016; **15**: 128.

8 Carvalho CMB, Ramocki MB, Pehlivan D, *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet* 2011; **43**: 1074–81.

9 Manske M, Miotto O, Campino S, *et al.* Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature* 2012; **487**: 375–9.