Pf7: an open dataset of *Plasmodium falciparum* genome variation in 20,000 worldwide samples

Details of 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. Samtools fixmate v1.2 and Picard v2.6.0 MarkDuplicates were successively applied to the BAM files of each sample. GATK base quality score recalibration was applied using default parameters, using the variants from the Pf crosses 1.0 release as a set of known sites. All lanes from each sample were merged to create sample-level BAM files. The output of this stage was a set of 20,864 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 a minimum depth of 5 (`--minDepth 5`) to determine the genomic positions callable in each sample.

Variant discovery and genotyping

Genotypes were called from each sample-level bam independently using GATK HaplotypeCaller v4.1.4.0 with parameters `-contamination 0 -ERC GVCF` to produce a separate gVCF for each of the 20,864 samples. The gVCFs from all samples were combined in 2,342 10kbp intervals across the genome and inserted into 2,342 separate GenomicDBs using GATK GenomicsDBImport v4.1.4.0 with parameters `-L <10kbp interval> -ip 500`.

Each GenomicDB contained gVCF data for a non-overlapping 10kbp genomic interval, as well as the 500bp before and after the interval. Genotypes were jointly called across all samples from each GenomicDB in its 10kbp interval, as well as the 500bp before it using GATK GenotypeGVCFs v4.1.4.0 with parameters

```
--only-output-calls-starting-in-intervals
--use-new-qual-calculator -L <10kbp interval including 500bp
padding before interval> --annotation-group StandardAnnotation
--annotation-group AS_StandardAnnotation.
```

For both GATK HaplotypeCaller and GenotypeGVCFs, the default parameters were used to implicitly set max alternate alleles to 6, sample ploidy to 2, SNP heterozygosity to 1e-3, and indel heterozygosity to 1.25e-4.

BCFtools view v1.10.2 was used to excise redundant calls from the 500bp paddings to yield 2,342 VCFs, each containing genotypes for all samples for a 10kbp interval.

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 both SNPs and indels we used the following parameters: `-an QD, -an FS, -an SOR, -an MQRankSum, -an ReadPosRankSum --trust-all-polymorphic --maxGaussians 8`. We chose not to use the recommended DP annotation as coverage in many of our samples is highly variable. We chose not to use the recommended MQ annotation as when included the relationship with VQSLOD appeared to be bimodal with low values of MQ giving high values of VQSLOD. `ApplyRecalibration` was then run to assign each variant in the core region (as previously defined in the Pf crosses 1.0 release[4]) a quality score named VQSLOD. High values of VQSLOD indicate higher quality. To decide on a VQSLOD threshold for SNPs, we determined transition/transversion (Ti/Tv) ratios for each bin of VQSLOD scores. We noticed that Ti/Tv dropped off for VQSLOD scores below 2.0, but didn't increase appreciably for VQSLOD scores higher than this. As such we decided a threshold of 2.0 would be a good balance between sensitivity and specificity and filtered out all SNPs with VQSLOD < 2.0. We set the same threshold for indels. Variants from the non-core regions of the genome and the VQSLOD annotated variants from the core regions were merged together into the same VCF using GATK CombineVariants v3.8 with parameters `-genotypeMergeOptions PRIORITIZE -filteredRecordsMergeType KEEP_UNCONDITIONAL -mergeInfoWithMaxAC`

Variants in the VCFs were annotated using a number of different methods. Functional annotations were applied using snpEff[5] version 4.3, with gene annotations downloaded from GeneDB[6] February 2020 release at ftp://ftp.sanger.ac.uk/pub/genedb/releases/2020-02/Pfalciparum/Pfalciparum.gff.gz. This gff file was modified by replacing the mitochondria name Pf3D7_MIT_v3 with Pf_M76611 in

order to match the reference genome used for mapping. The following options were used with snpEff: `-no-downstream -no-upstream -onlyProtein`.

Genome regions were annotated using `bcftools` v1.10.2 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* gene 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.

The VCFs were merged, yielding a separate VCF for each chromosome, mitochondria, and apicoplast using GATK GatherVcfs v4.1.4.0. VCF files were converted to zarr v2.4.0 format using scikit-allel v1.2.1 (https://github.com/cggh/scikit-allel). Subsequent analyses were performed using the zarr files.

Genetic distance

We calculate genetic distance between samples using biallelic coding 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 $f_1\left(1 - f_2\right) + f_2(1 - f_1)$. 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.

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. 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 set of 16,203 analysis samples after removing samples with unverified identity, mixed species, replicate and low coverage samples, and samples with excessive numbers of singleton SNPs.

We removed 120 samples for which the identity could not be verified, including samples that were identified as lab strains or continent mismatches in Pf6.

Then, to identify samples with significant DNA contamination from commonly used lab strains, we compared the mean genetic distance of each sample to data on lab strains 3D7, 7G8, GB4, HB3, Dd2 and IT. Where the mean distance was < 1 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 32 samples from the analysis set.

Afterwards, we removed 8 samples that were genetically similar to other samples from a different continent which could be due to mislabelling or unreported travel history; these are samples for which the majority of the eleven nearest neighbours are samples from a different continent.

We then aggregated and marked these 160 samples as “unverified identity” in the dataset.

We next removed 304 samples from the analysis set that were identified as containing mixed species, as this can affect genotyping of highly conserved loci (e.g. kelch13).

We then 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 1,298 samples. A further 2,893 samples with callability <50% were also removed.

Finally we removed 6 samples which had over 500 singleton SNP calls in the remaining sample set. The final analysis set contained 16,203 QC pass samples.

CNV genotypes at drug resistance loci

In our previous work, Pf6, tandem-duplication and DUP-TRP/INV-DUP breakpoint genotypes around mdr1 and plasmepsin2/3 were determined using two complementary forms of

evidence: 1) amplified copy-number segments, identified by applying a hidden Markov model to binned coverage (normalized to account for sample-specific total depth and GC-content bias), and 2) excesses in the proportion of read pairs with discordant orientations. Given the evidence in each sample, we called the breakpoint genotype with the highest support among a list of previously identified genotypes, using manual curation to resolve cases which were supported by only one form of evidence.

In this work, we focused on improving the methods for coverage-based evidence by utilizing the GATK GermlineCNVCaller (gCNV) pipeline. This effort was primarily motivated by the difficulties the previous hidden Markov model had in estimating copy number in samples that had undergone selective whole-genome amplification (sWGA), which typically yields heavily biased and noisy coverage profiles.

We adapted the gCNV pipeline to our use case of breakpoint genotyping by augmenting its capabilities as follows:

a) The gCNV pipeline performs inference on a probabilistic Bayesian model that jointly represents copy-number activity and sequencing systematics, given observed binned read-count data. Copy-number activity is represented by a hierarchical hidden Markov model, with global, binary parameters indicating bins with common copy-number activity and per-sample discrete parameters denoting copy-number state in each bin. In contrast, systematic sequencing bias and noise are represented by continuous parameters within a linear latent-factor model.

   An initial cohort of training samples can be used to learn the model for systematic sequencing bias and noise; this “denoising” model can then be applied to normalize and denoise subsequent samples. The implicit assumption is that the training samples are representative of subsequent samples. Thus, in cohorts with a high degree of heterogeneity, it becomes advantageous to divide samples into a small number of clusters, training and employing separate denoising models on each cluster. This allows each model to devote statistical power to modeling within-cluster sample-to-sample variation; in contrast, a single model trained on the entire analysis cohort would necessarily expend statistical power to model cluster-to-cluster variation.

   For our analysis, we implemented clustering by first using principal component analysis to reduce the dimensionality of coverage profiles (given by read counts in 500 bp bins naively normalized using per-sample and per-bin medians) to 20 dimensions, followed by fitting a Bayesian Gaussian mixture model to identify 6 clusters of samples. In each cluster, 300 samples were used to train the denoising model, which was then used to call the remainder of the samples in the cluster.

b) The gCNV pipeline is intended for discovery of copy-number variants and thus has limited ability to make use of prior information, such as known breakpoint

genotypes. We augmented the pipeline to perform copy-number and breakpoint genotyping by making use of intermediate results, namely the copy-number emission probabilities from the hierarchical hidden Markov model, to calculate conditional likelihoods for each genotype. Additional quantities, such as the log-odds ratio of the reference copy-number state to all other states in regions flanking the breakpoints (which can be used as an indicator of sample-level sequencing quality), were also calculated from the emission probabilities and used to annotate and filter genotype calls.

c) Finally, the gCNV model assumes germline copy-number states, and hence it cannot strictly model or call fractional copy-ratio states that may arise in samples with a mixture of populations (e.g., samples originating from coinfections). Nevertheless, fractional copy-ratio activity is still evident in the per-bin denoised copy ratios that are emitted by the pipeline; the model can still properly account for systematics, even if it cannot confidently infer integer copy-number states.

We thus implemented a straightforward generalization of a standard likelihood-ratio test (LRT) for changepoint detection[31] to generate conditional likelihoods and LRT statistics for each breakpoint genotype from the denoised copy ratios. The LRT essentially gives the relative likelihood of an alternative hypothesis---that a change in the distribution of the denoised copy ratios is induced by copy-number activity in the sample---to the null hypothesis---that the denoised copy ratios are uniformly distributed and the sample is copy neutral. For each sample and region, we heuristically combined the likelihood generated from the denoised copy ratios with that derived from the copy-number emission probabilities to derive a maximum likelihood estimate (MLE) of the breakpoint genotype. As above, we also calculated additional annotations, such as the MLE scale parameter for the denoised copy ratios (which is another indicator of sample-level sequencing quality), which were used in conjunction with the LRT statistic to filter genotype calls.

To mitigate the possible effect that unmodeled fractional copy-ratio activity might have on the training of the denoising model, we limited the selection of training samples to those with F_{ws} >= 0.95.

We were able to successfully call both tandem-duplication and DUP-TRP/INV-DUP breakpoint genotypes consistently within this framework, with only minor differences in how the different event types were treated. In particular, for tandem duplications, all copy-number states CN = 0, 1, …, 5 allocated by the gCNV model were considered; in contrast, only DUP-TRP-DUP copy-number states were considered along with the reference copy-number state for DUP-TRP/INV-DUP genotypes.

After MLE breakpoint genotypes were called using the above procedure, we performed an initial round of call filtering, using a combination of hard cuts on the various annotations generated from the copy-number emission probabilities and the denoised copy ratios. In particular, for each region of interest, we used the aforementioned log-odds ratio and MLE

scale parameter annotations to perform a quality-control cut to identify failing samples, followed by a cut on the LRT statistic to demarcate samples with duplication and reference genotypes. Guided by comparison with preliminary results generated using the Pf6 hidden Markov model, permissive thresholds were chosen for each of these cuts, resulting in a relatively sensitive initial call set. A final round of curation, based on manual inspection of the denoised copy ratios, was then applied to discard spurious duplication calls that passed the initial round of filtering. This round of curation was aided by sorting calls appropriately by the annotations, which allowed us to focus our attention on the most marginal calls.

Finally, this filtered gCNV call set was integrated with an analogous call set based on consideration of face-away read-pair evidence. Construction of this face-away read-pair evidence was accomplished using methods identical to those used in Pf6[9]. We determined numbers of read pairs around the sets of breakpoints identified in v6. We also calculated the proportion of read pairs around each set of breakpoints for which the reads are mapped in face-away orientation. If any set of breakpoints had fewer than 40 read pairs, we set the face-away call to missing. If the maximum proportion of reads mapping in face-away orientation around each set breakpoints was greater than 0 but below 2.5%, we also set the face-away call to missing. If the maximum proportion of reads mapping in face-away orientation around all breakpoints was zero we set the face-away call to non-duplicated (i.e. reference). Finally if any set of breakpoints has >= 2.5% read mapping in face-away orientation, we set the call to duplicated. If either the gCNV or face-away call set indicated a duplication call, the final call was a duplication. If either method resulted in a missing call, the call from the other method was used. If both methods failed to make a call, the final call was set as missing.

For all samples we report the breakpoint which has the highest proportion of face-away reads. Note that in cases where there were some face-away reads but insufficient to make a face-away read call, we still report the breakpoint even if the final call is either reference (non-duplicated) or missing.

HRP2 and HRP3 deletion detection

In Pf6, deletions in *hrp2* and *hrp3*, genes which are located in subtelomeric regions of the genome with very high levels of natural variation, were identified by manual inspection of coverage profiles. For Pf7, we repeated this approach, but also sought to improve upon this approach by using the same breakpoint-genotyping framework introduced above, with the goal of having a consistent and unified treatment of both duplications and deletions.

However, the above framework implicitly assumes that the input set of possible breakpoint genotypes is complete; any novel genotypes present in the data that are not provided in this set will be incorrectly called as the closest available genotype. This especially presented an issue for hrp2 deletions; a wide variety of large deletions can occur in the subtelomeric region, but we were primarily interested in---and have thus only catalogued breakpoints for---homogeneous deletions that disrupt or delete the hrp2 gene itself. To filter out

deletions that are not of interest, we relied on annotations for summary statistics (namely, the minimum and the mean) of the denoised copy ratios of bins that overlap the gene; these statistics suitably separated homogeneous deletions that impact part or all of the gene from other types of deletions unlikely to be relevant for drug resistance.

As in the case of duplication genotyping, an initial round of permissive, annotation-based filtering was performed, with the goal of producing a relatively sensitive call set; this was again guided by comparison with a preliminary call set generated by the previous Pf6 method of manual inspection of coverage profiles. A final round of curation was then also applied to discard spurious deletion calls. In addition, the manual inspection approach identified two samples that appeared to have deletions in *hrp3* but were not called as deletions by the gCNV approach. For these two there was clear evidence of a deletion in the gCNV denoised copy ratios, but the deletion was not called by gCNV as the breakpoint was within the gene itself (Figure 1).

We created plots of sequence read coverage for the regions around *hrp2* and *hrp3* for all samples and these were used both as a cross check on the gCNV calls, and also for determining breakpoints of samples determined by gCNV to have deletions covering one or both the genes (Figure 1). Reads containing the telomeric repeat sequence GGGTTCA or GGGTTTA were highlighted on these plots, and where these appeared in positions where there was a sudden drop in coverage, the most frequent start position of the telomeric sequence in these reads was taken as the deletion breakpoint and the sample was classified as having a "Telomere healing" deletion type. All samples with *hrp2* deletions could be classified as telomere healing events and had precise breakpoints assigned.

In *hrp3* a large subset of samples appeared to have a drop in coverage at a position around 2.807 Mbp. This coincides with the end of a cluster of three ribosomal RNA genes (PF3D7_1371000, PF3D7_1371200 and PF3D7_1371300) spanning the region Pf3D7_13_v3:2800004-2807159. There is a similar cluster of ribosomal genes on chromosome 11 (PF3D7_1148600, PF3D7_1148620 and PF3D7_1148640) spanning the region Pf3D7_11_v3:1925995-1933138 with high levels of sequence identity to the cluster on chromosome 13. When we looked at coverage around this region of chromosome 11 in sample that appeared to have the deletion on chromosome 13 starting around 2.807 Mbp, we noticed that in all cases where the coverage variation across the genome was relatively low, there was an apparent doubling of coverage on chromosome beyond this cluster of genes. This suggests that there has been a recombination event between chromosomes 13 and 11 somewhere within these ribosomal RNA clusters, resulting in a new chromosome composed mostly of chromosome 13 sequence but with the right-hand sub-telomere composed of chromosome 11 sequence.

A further set of 21 samples had a sudden drop off in coverage at around 2.835 Mbp but no evidence of any telomeric sequences in the reads around this position. All of these samples had duplications of the gene *mdr1* on chromosome 5. Close inspection of these 21 samples revealed that they all had reads which were soft-clipped after Pf3D7_13_v3:2835612. In

many cases the soft-clipped part of the read had a supplementary alignment on Pf3D7_05_v3, and the mate often mapped to a similar region. There is an $(AT)_{14}$ repeat at Pf3D7_13_v3:2835587-2835612 and at $(AT)_9$ repeat at Pf3D7_05_v3:979192-979209, and manual inspection of coverage plots suggests the duplication on chromosome 5 ends around this position. At the 5' end of the duplication region in most samples we see soft-clipped reads where the soft-clipped portion contains telomeric repeat sequences. As such, it appears that there has been a recombination between chromosome 13 and an inverted part of the centre of chromosome 5, including the gene *mdr1*. Furthermore, the resulting chromosome appears to be viable as there is evidence of telomeric repeat sequence at the end of the resultant hybrid chromosome.

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` v0.5.6. Based on these observations we grouped the samples into eight geographic regions: South America, west Africa, central Africa, east Africa, northeast Africa, the western part of South Asia, the eastern part of South Asia, the western part of Southeast Asia, the eastern part of Southeast Asia and Oceanian island of New Guinea, 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 for QC pass samples 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 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.

Drug resistance gene haplotypes, amino acid calls and samples classification

We created full gene amino acid haplotypes for the genes *crt*, *dhfr* and *dhps*. We also determined the amino acid calls at *crt* amino acids 76 and 72-76, dhfr amino acids 51, 59, 108 and 164, and dhps amino acids 437, 540, 581 and 613. We determined amino acid changes in *kelch13* 349-726. To do all of these we wrote custom python code to apply genotypes from the GT field of the VCF to the reference sequence, and then translate the nucleotide sequence into a) amino acid sequence and b) lists of non-synonymous variants. We used all SNPs that passed filtering, and also two short indels in crt at positions Pf3D7_07_v3:403618 (1bp insertion) and Pf3D7_07_v3:403622 (1bp deletion). The combination of these two indels results in a haplotype that is three differences from the 3D7

reference sequence, but HaplotypeCaller resolves this haplotype as two indels rather than three SNPs. For multi-exon genes such as *crt* we concatenated nucleotide sequence from each exon before translating to amino acid haplotype. Where there was one or more heterozygous genotype call we created separate nucleotide sequence for each allele and then translated each into separate amino acid haplotypes. We used the AD field of the VCF to assign the allele with the greater read count as the first allele. Where there was more than one heterozygous call we used the PID and PGT VCF fields to phase the variants where possible. If any variant had a missing genotype call for any variant, we called the haplotype as missing, but still reported any non-synonymous variant seen at other variant within the haplotype. This is important because, taking the example of amino acids 349-726 in *kelch13*, if we see a non-synonymous mutation such as C580Y we can declare the sample as resistance to artemisinin, even if genotypes at other positions are missing.

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/34).

eba175 allelic type calling

We identified kmers of length 19 that are unique to the two major haplotypes C and F in EBA-175. F and C calls are made simply by first identifying samples that have kmers present that are unique to the C and F haplotypes. A normalisation of the F kmer using a neighbouring interval containing roughly the same number of reads is used as the F fragment breakpoint is often deleted. Samples with kmers present in both haplotypes are identified as mixed regardless of fraction, and samples without kmers from either F or C are identified as no calls. Proportion of C allele was calculated as number of samples with C allele / number of samples with C or F allele (i.e. samples with mixed or no calls were not considered in the calculation.

Details of bioinformatics methods figures

Figure 1. Example of HRP deletion diagnostic plot for sample not called as duplication by gCNV. The panels show coverage in 300 bp windows across regions of Pf3D7_08_v3 (top), Pf3D7_13_v3 (centre) and Pf3D7_11_v3 (bottom). Below the top two panels, locations of reads containing 1, 2 and 3 repeats of the telomeric sequences GGGTTCA or GGGTTTA are shown in purple. Below this, the location of reads with at least 2 telomeric sequences are shown together with the position of the start of the telomeric sequence (i.e. the breakpoint) and the number of such reads. In this example there are 64 reads containing at least 2 telomeric repeat sequences where the sequence begins at position Pf3D7_13_v3:2841024 (which is the assumed position of the *hrp3* deletion breakpoint in this sample. The pink vertical lines show the positions of the genes *hrp2* and *hrp3*, and in this case it is apparent that the breakpoint of the *hrp3* deletion is within the *hrp3* gene itself.

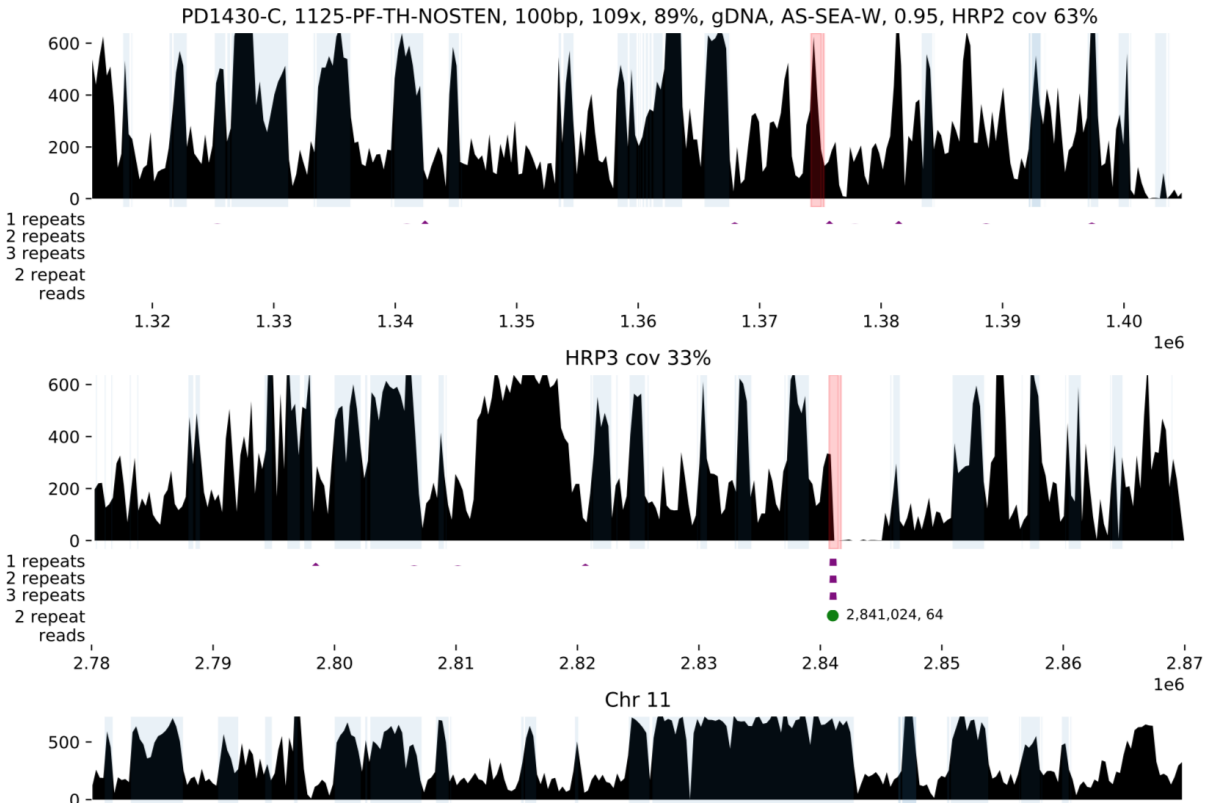
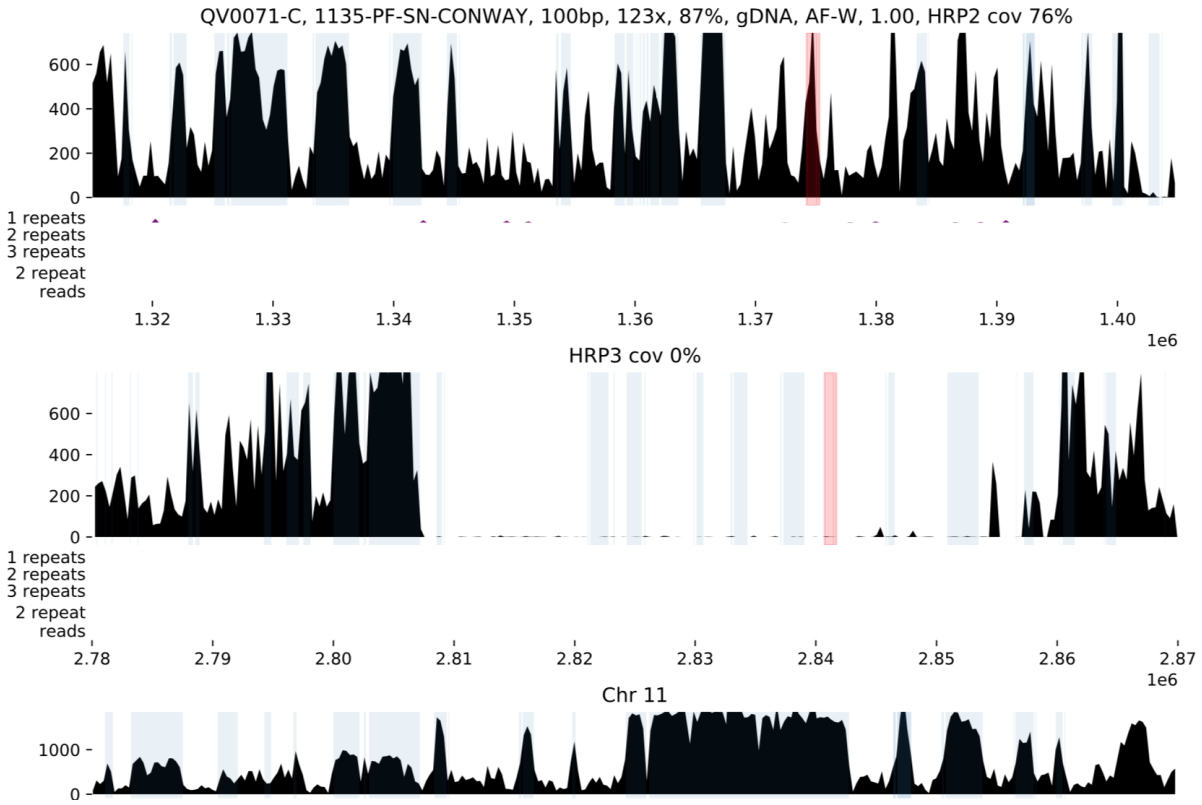


Figure 2. Example of HRP deletion diagnostic plot for sample with chromosome 11 recombination. The panels show coverage in 300 bp windows across regions of Pf3D7_08_v3 (top), Pf3D7_13_v3 (centre) and Pf3D7_11_v3 (bottom). For this sample there is a clear drop off in coverage around 2.807Mbp on chromosome 13. It can also be seen that the coverage at the right-hand end of the bottom plot is typically about twice that a the left hand. These plots suggest that there has been a recombination event between chromosomes 13 and 11.



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.