Deep sequencing of Plasmodium falciparum genetic crosses: a resource for the study of genome variation and meiotic recombination

Supplementary information

Table of Contents

Index of Figures

Index of Tables

1. Whole genome sequencing

@@TODO

Note that typically in high throughput sequencing studies of humans or other higher eukaryotes multiple sequencing runs will be obtained for each sample, then data from each run (lane) are combined to increase coverage. However in this study a single sequencing run was sufficient to obtain ~100X coverage of the *P. falciparum* genome, so only a single sequencing run was obtained for each sample. Samples that represented biological replicates (DNA derived from the same clone but obtained from different cultures) were treated separately, with separate DNA library preparation and sequencing runs. Thus in this study there is always a one-to-one mapping from sample (biological replicate) to sequence run.

For convenience throughout this document we use a three-part identifier for each sample, e.g., “3D7/PG0051-C/ERR019061”, where the first part identifies the clone (e.g., “3D7”), the second part is our internal lab identifier for the sample (i.e., biological replicate, e.g., “PG0051-C”), and the third part is the accession for the sequencing run at the ENA (e.g., “ERR019061”). The second and third parts are redundant, because as mentioned above there is a one-to-one mapping from sample to sequencing run, however we include both for transparency. The data files available from the FTP site and the web application use the same identifier system for consistency.

2. Sequence alignment and genome region classification

Sequence reads from each sample were aligned to the 3D7 version 3 reference genome using BWA (Li & Durbin, 2009) version 0.6.1-r104 with the following parameter settings:

```
bwa aln -n 0.01 -k 4
bwa sampe
```

We found that the custom parameters to the `aln` command served to slightly increase the sensitivity and improve consistency of the alignment in regions with clusters of SNPs, such as the polymorphisms found at the chloroquine resistance locus (Fidock et al., 2000), however the vast majority of alignments are identical under the custom and default settings (data not shown). We recommend the custom settings for alignment of *P. falciparum* short sequence reads where possible, however the increased sensitivity does increase the runtime required by approximately an order of magnitude over the default settings, and therefore the default settings are the only practical option for large numbers of samples.

Various metrics were then calculated from the alignments of each sample. These metrics were computed per genome position based on the pileup of aligned reads, using the program `pysamstats`[1]. Metrics calculated include the total depth of coverage, percentage of reads aligned in a proper pair (i.e., in correct orientation and reasonable distance apart, as defined by the aligner), average mapping quality and percentage of reads aligned ambiguously (mapping quality zero).

Alignment metrics for each of the parental samples were then plotted for each chromosome, alongside other metrics derived from the reference genome sequence, including the %GC content in a 300bp window and the non-uniqueness score (defined as the smallest k-mer size at which all k-mers overlapping a given position are unique within the genome; a high score for this metric is bad, in the sense that it indicates low uniqueness). An example plot for sample HB3/PG0052-C/ERR019054 and chromosome 4 is shown in Figure S1. The alignments themselves were also

[1] https://github.com/alimanfoo/pysamstats

visualised using the LookSeq web application (Manske & Kwiatkowski, 2009), which can be viewed via the web application at @@URL.

From these visualisations a clear, qualitative distinction could be seen between regions of the genome with consistent coverage across all parent samples, and regions with significant alignment issues in one or more parents. To capture these large-scale qualitative differences we defined the following heuristic scheme for classifying genome regions:

- **Core –** Regions with near-continuous coverage in all samples, with a high percentage of reads mapping in a proper pair and a low proportion of reads aligned ambiguously.
- **Subtelomeric Hypervariable** – Gene-containing regions towards the sub-telomere of a chromosome, with patchy and/or highly variable coverage in one or more samples and/or a low proportion of reads aligned in a proper pair and/or a high percentage of ambiguous alignments and/or a high percentage of aligned bases mismatching the reference.
- **Internal Hypervariable –** Gene-containing regions towards the centromere of a chromosome, with patchy and/or highly variable coverage in one or more samples and/or a low proportion of reads aligned in a proper pair and/or a high percentage of ambiguous alignments and/or a high percentage of aligned bases mismatching the reference.
- **Subtelomeric Repeat –** Gene-free regions with repetitive sequence at the end of a chromosome, typically with highly variable coverage and a high percentage of ambiguous alignments.
- **Centromere –** Centromere as given in the GeneDB genome annotation.

Within each chromosome we defined boundaries for these regions by eye from the visualisations described above. Figure S2 shows a map of the genome regions defined, and Figure S3 gives a summary of alignment statistics for each parental clone by region class. At least 99.6% of core genome positions were covered in all parents, and at least 98.8% of the core genome was covered by unambiguously mapped reads.

Our definition of the core genome is subjective, and more sophisticated methods could be devised to partition the genome into regions with different alignment characteristics. However the contrast between these different regions of the genome is very striking, and we believe the definitions given here capture the major qualitative features in a useful way.

The genome region classification can be browsed alongside coverage, mapping quality and other metrics via the web application at the following URL:

http://www.malariagen.net/apps/pf-crosses/#genome

A BED file defining the region boundaries can be downloaded from the FTP site:

@@TODO

3. Variant discovery and genotype calling

3.1. Alignment-based calling method (BWA/GATK)

The alignment-based calling method used the Genome Analysis Tool Kit version 2.6-4-g3e5ff60 (McKenna et al., 2010) and followed best practice recommendations as published at the time (DePristo et al., 2011; Van der Auwera et al., 2013).

Starting from the reads aligned to the 3D7 version 3 reference genome as described above, the following steps were performed to prepare the BAM files. Using Picard tools version 1.77 the commands CleanSam, FixMateInformation, AddOrReplaceReadGroups and MarkDuplicates were run on each BAM file in that order.

Base quality score recalibration (BQSR) was then applied to the BAM files. BQSR empirically recalibrates the base quality scores reported for each base in each sequence read, by observing the correlation between mismatches in the aligned sequence reads and various covariates, including the original base quality reported by the sequencing machine, in addition to other factors like the local sequence context. BQSR thus relies on the assumption that a substantial number of bases mismatching the reference in aligned sequence reads are due to sequencing error and not true variation, alignment error or some other type of artefact. From a visual inspection of the alignments for the parental clones (see, e.g., Figure S1) it was apparent that the mismatch rate within hypervariable regions was extremely high, and given the other alignment symptoms in hypervariable regions including patchy coverage and ambiguous mapping, we assumed the vast majority of these mismatches were due to divergence between clones and not sequencing error. To avoid hypervariable regions overwhelming BQSR we limited the building of the covariates table to the core genome. BQSR also requires a set of known variant positions to exclude when building the covariates table. To bootstrap BQSR we created an initial set of variant calls for each cross from the raw BAM files using UnifiedGenotyper, then filtered these calls to exclude any that had less than 2 confident (GQ = 99) ALT calls, contained Mendelian errors, had more than 2 missing calls or were part of a homopolymer run of length 5 or more.

We then applied INDEL realignment to the recalibrated BAMs. Each BAM file was realigned separately, but to improve the sensitivity of INDEL realignment we provided as input the set of bootstrap INDEL calls obtained from the previous BQSR step, which has the effect of sharing information about possible INDEL alleles between samples. All other settings were default.

We then generated a raw variant callset using UnifiedGenotyper run under a haploid model (-ploidy 1).

The next step was to empirically recalibrate variant quality scores (VQSR). VQSR requires at least a positive training set of known true variants, and optionally one or more negative training sets of sites where variant calls are likely to be spurious. We defined a positive training set for each cross by selecting variants from the raw callset that segregated within the cross according to Mendelian inheritance (i.e., parents had different genotypes, progeny had no Mendelian errors) and also produced highly parsimonious patterns of inheritance (i.e., did not induce an unrealistically high rate of recombination). Specifically, the positive training sets included only SNP and INDEL variants within the Core genome, with no missing calls, no non-Mendelian calls, and no calls inducing an apparent double-crossover at a single variant. We also created two negative training sets for each cross, the first containing variants with Mendelian errors, the second containing variants inducing single-variant double-crossovers in one or more samples.

We then applied VQSR to each cross separately. VQSR was run for SNPs with the following options:

- -an QD -an DP -an MQ -an UQ -an HaplotypeScore -an ReadPosRankSum -an FS --target_titv 1.0 --percentBadVariants 0.1 --stdThreshold 10.0 --maxGaussians 6

VQSR for INDELs was run with the following options:

- -an QD -an DP -an MQ -an UQ -an HaplotypeScore --target_titv 1.0 --percentBadVariants 0 --stdThreshold 10.0 --maxGaussians 6

"UQ" is the non-uniqueness score define above and the other annotations are standard INFO annotations produced by GATK.

To verify that the VQSR runs had been effective we plotted the rate of Mendelian error against the number of variants for different thresholds of the VQSLOD score (similar to an ROC curve) (Figure S4). For all three crosses and for both SNPs and INDELs, we observed an inflection point in these curves, corresponding to a Mendelian error rate of approximately 0.05% or ~1 Mendelian error in 2000 genotype calls. Thresholds (minimum values) were chosen for the VQSLOD separately for SNPs and INDELs in each of the three crosses at the inflection point in the curve. For SNPs the thresholds were 3D7xHB3: 2.5, HB3xDd2: 3, 7G8xGB4: 4; for INDELs the thresholds were 3D7xHB3: 1, HB3xDd2: 1.5, 7G8xGB4: 1.8.

We generated a final, analysis-ready VCF for each cross by adding the following filter annotations:

- LOW_CONFIDENCE – Variant confidence is low (VQSLOD falls below the chosen threshold).
- NON_MENDELIAN – Variant calls are not consistent with Mendelian segregation because one or more progeny have an allele not found in either parent.
- MISSING_PARENT – One or both parents have a missing genotype call.
- NON_SEGREGATING – Variant is fixed within the sample set (not necessarily a spurious variant but a useful filter annotation as most analyses shown here use only segregating variants).
- DUP_SITE – Variant position coincides with another.
- NON_CORE – Variant is not within the core genome.
- LOW_CONFIDENCE_PARENT – Genotype confidence for one or both parents is low (GQ < 99).
- CNV – There is evidence for copy number variation at this locus.

The CNV filter was applied based on evidence from depth of coverage data, described in the section on CNV analysis below.

For all downstream analyses we also treated genotype calls with a genotype quality (GQ) of less than 99 as missing, although this annotation is not included in the VCF files.

Figure S6 illustrates variant calls from the alignment-based method before and after filtering for a single cross and chromosome. Variant calls for all crosses and chromosomes can be browser via the web application at @@URL, both with and without filters.

3.2. Assembly-based calling method (Cortex)

@@TODO Zam to complete: method to generate the Cortex VCF files.

We plotted the rate of Mendelian error against the number of variants for different thresholds of the SITE_CONF score (Figure S5). Based on these plots we used a target Mendelian error rate of ~0.05% to decide variant and call filtering strategies. For SNPs we chose a SITE_CONF threshold of 50 and for INDELs we chose a SITE_CONF threshold of 200. These thresholds were the same for all crosses.

We generated a final, analysis-ready VCF for each cross by adding the following filter annotations:

- LOW_CONFIDENCE – Variant confidence is low (SITE_CONF falls below the chosen threshold).
- NON_MENDELIAN – Variant calls are not consistent with Mendelian segregation because one or more progeny have an allele not found in either parent.
- MISSING_PARENT – One or both parents have a missing genotype call.
- NON_SEGREGATING – Variant is fixed within the sample set (not necessarily a spurious variant but a useful filter annotation as most analyses shown here use only segregating variants).
- DUP_SITE – Variant position coincides with another.
- NON_CORE – Variant is not within the core genome.
- LOW_CONFIDENCE_PARENT – Genotype confidence for one or both parents is low (GT_CONF < 50).
- CNV – There is evidence for copy number variation at this locus.

Note that these are in addition to a number of filter annotations previously added as a standard part of the Cortex pipeline.

For all downstream analyses we also treated genotype calls with a GT_CONF of less than 50 as missing, although this annotation is not included in the VCF files.

Figure S7 illustrates variant calls from the assembly-based method before and after filtering for a single cross and chromosome. Variant calls for all crosses and chromosomes can be browser via the web application at @@URL, both with and without filters.

3.3. Combined callset

A single callset of segregating variants was constructed for each cross by combining variant calls from the alignment and assembly-based methods as follows. For each calling method, a VCF was derived from the full analysis-ready VCF by selecting only variants that passed all filters and segregated within the cross. These two VCFs were then combined into a single VCF using the GATK CombineVariants task, taking genotype calls from the alignment-based calling method where both methods reported the same variant (because the alignment-based method had lower levels of missingness). This produced a single combined VCF of segregating variation for each cross. These VCFs were then post-processed to add a DUP_SITE filter annotation to any variant that coincided with another variant but reported different alleles.

3.4. Genotype concordance between biological replicates

In the 3D7xHB3 cross one replicate for clone C01 and 3 replicates for clone C02 were sequenced and genotyped independently. This provided 6 replicate pairs for analysis of genotype concordance. In the 7G8xGB4 cross a single replicate was obtained for each of 10 progeny clones, providing 10 replicate pairs. We computed genotype concordance for each replicate pair and for each of the three available callsets (alignment-based method, assembly-based method, combined) after filtering variants and genotype calls as described above. We computed concordance for each replicate pair as the number of sites where both samples had a matching genotype call divided by the number of sites where both samples had a non-missing genotype call. The results are given in Table S1.

3.5. Estimation of FDR and sensitivity

To estimate false discovery rate (FDR) and sensitivity, we compared the variant calls generated in this study with pre-existing sequence data resources for the clone HB3. We downloaded contigs from the HB3 genome assembly produced from shotgun sequencing by Birren et al. (2006). We also downloaded HB3 sequences for individual genes deposited in GenBank. We aligned both the HB3 contigs and the gene sequences to the 3D7 reference genome using `bwa mem` with the `-x intractg` option (parameters tuned for mapping contigs within a species). We limited further analyses to a set of 32 genes that were completely covered by a single uniquely mapped contig from the Birren et al. assembly and by a gene sequence (Table S2). In spite of these criteria there remained some discordance between the Birren et al. assembly and the the gene sequences, particularly regarding INDELs. Given that both of these sources may themselves contain errors, we used the following methods to estimate FDR and sensitivity. To estimate FDR we compared variants discovered in this study with the union of variants found in the Birren et al. assembly and the gene sequences. Thus a true positive is a variant discovered in this study and also found in either of the other sources, and a false positive is a variant discovered in this study but not present in either of the other sources. To estimate sensitivity we compared variants discovered in this study with the intersection of variants found in the Birren et al. assembly and the gene sequences. Thus a false negative is a variant not discovered in this study but present in both of the other sources.

FDR and sensitivity were computed for the replicates HB3(1) and HB3(2) separately, and for each of the two variant calling methods. For INDELs these metrics were computed under two different matching schemes: “position match” where we require the position and type (insertion/deletion) of the variant to match but allow the allele to be different, and “allele match” where we require the position and allele to match perfectly. The results are reported in Table S3.

Note that for these comparisons we included all variant alleles called for an HB3 sample, regardless of whether they segregated within a cross (i.e., we ignored the NON_SEGREGATING filter annotations). This is particularly relevant for the HB3(2) sample which was genotyped as part of the HB3xDd2 cross and where many alternate alleles were shared with clone Dd2 and were fixed in all progeny.

4. Recombination analyses

4.1. Determination of maximal block length for conversion tracts

@@TODO

5. Tables

Index of Tables

Cross	Clone	Replicate pair	Genotype discordance		
			BWA/GATK callset	Cortex callset	Combined callset
3D7 x HB3	C01	C01/PG0062-C/ERR019070 vs C01/PG0065-C/ERR019064	3/36567	1/27152	3/42021
3D7 x HB3	C02	C02/PG0053-C/ERR019067 vs C02/PG0055-C/ERR019066	1/36551	0/27008	1/41977
3D7 x HB3	C02	C02/PG0053-C/ERR019067 vs C02/PG0056-C/ERR019068	1/36530	0/26943	1/41948
3D7 x HB3	C02	C02/PG0053-C/ERR019067 vs C02/PG0067-C/ERR019073	3/36569	0/27068	3/42010
3D7 x HB3	C02	C02/PG0055-C/ERR019066 vs C02/PG0056-C/ERR019068	2/36527	0/27022	2/41949
3D7 x HB3	C02	C02/PG0055-C/ERR019066 vs C02/PG0067-C/ERR019073	5/36573	1/27172	6/42029
3D7 x HB3	C02	C02/PG0056-C/ERR019068 vs C02/PG0067-C/ERR019073	1/36545	0/27090	1/41985
7G8 x GB4	AUD	AUD/PG0112-C/ERR029406 vs AUD/PG0112-CW/ERR045639	32/27524	7/22423	15/33814
7G8 x GB4	JC9	JC9/PG0111-C/ERR029409 vs JC9/PG0111-CW/ERR045634	28/27556	8/22700	8/33998
7G8 x GB4	JE11	JE11/PG0100-C/ERR029404 vs JE11/PG0100-CW/ERR045630	30/27182	2/20800	9/32703
7G8 x GB4	JF6	JF6/PG0079-C/ERR027102 vs JF6/PG0079-CW/ERR045637	25/27529	8/22544	10/33878
7G8 x GB4	KB8	KB8/PG0104-C/ERR029148 vs KB8/PG0104-CW/ERR045642	25/27256	6/21939	13/33296
7G8 x GB4	LA10	LA10/PG0086-C/ERR029090 vs LA10/PG0086-CW/ERR045629	26/27393	2/21724	11/33365
7G8 x GB4	NIC	NIC/PG0095-C/ERR027107 vs NIC/PG0095-CW/ERR045631	32/26991	3/19531	10/31909
7G8 x GB4	QF5	QF5/PG0078-C/ERR029092 vs QF5/PG0078-CW/ERR045638	34/27422	6/22349	18/33682
7G8 x GB4	XD8	XD8/PG0105-C/ERR029144 vs XD8/PG0105-CW/ERR045628	29/27562	13/22572	17/33917
7G8 x GB4	XF12	XF12/PG0102-C/ERR029143 vs XF12/PG0102-CW/ERR045635	32/27507	5/22459	18/33801

Table S1: Genotype discordance between biological replicates. Each row reports discordance data for a single replicate pair. Values given for each callset are [number of variants with a discordant genotype call]/[total number of variants with non-missing genotype calls in both members of the pair].

Chromosome	Start	Stop	ID	Name	Previous ID	Genbank Accession
Pf3D7_01_v3	265208	269173	PF3D7_0106300	ATP6	PFA0310c	gi|56342158|dbj|AB121052.1|
Pf3D7_02_v3	290168	292703	PF3D7_0207300	SERA8	PFB0325c	gi|803375251|dbj|AB733715.1|
Pf3D7_02_v3	294273	297616	PF3D7_0207400	SERA7	PFB0330c	gi|803375249|dbj|AB733714.1|
Pf3D7_02_v3	298897	302564	PF3D7_0207500	SERA6	PFB0335c	gi|803375247|dbj|AB733713.1|
Pf3D7_02_v3	303593	307027	PF3D7_0207600	SERA5	PFB0340c	gi|803375245|dbj|AB733712.1|
Pf3D7_02_v3	308847	312155	PF3D7_0207700	SERA4	PFB0345c	gi|803375243|dbj|AB733711.1|
Pf3D7_02_v3	313449	316741	PF3D7_0207800	SERA3	PFB0350c	gi|803375241|dbj|AB733710.1|
Pf3D7_02_v3	322338	325723	PF3D7_0208000	SERA1	PFB0360c	gi|803375237|dbj|AB733708.1|
Pf3D7_03_v3	221323	222516	PF3D7_0304600	CSP	PFC0210c	gi|56342142|dbj|AB121018.1|
Pf3D7_04_v3	137640	146653	PF3D7_0402300	RH1	PFD0110w	gi|33414602|gb|AF411930.2|
Pf3D7_04_v3	748088	749914	PF3D7_0417200	DHFR-TS	PFD0830w	gi|340507|gb|J03772.1|PFADHFRTSE
Pf3D7_04_v3	1085979	1091277	PF3D7_0424200	RH4	PFD1150c	gi|21321386|gb|AF420310.1|
Pf3D7_05_v3	328666	329715	PF3D7_0508000	P38	PFE0395c	gi|133900606|gb|EF137222.1|
Pf3D7_06_v3	851378	852955	PF3D7_0620400	MSP10	PFF0995c	gi|237664869|gb|FJ406615.1|
Pf3D7_07_v3	381592	384614	PF3D7_0708400	HSP90	PF07_0029	gi|505339|gb|L34028.1|PFAHSP86B
Pf3D7_07_v3	408215	411961	PF3D7_0709100		PF07_0035	gi|2642510|gb|AF030690.1|
Pf3D7_07_v3	413560	421749	PF3D7_0709300		PF07_0037	gi|2642515|gb|AF030693.1|
Pf3D7_08_v3	278381	279034	PF3D7_0804800	CYP24	PF08_0121	gi|1000520|gb|U10322.1|PFU10322
Pf3D7_08_v3	1358314	1363618	PF3D7_0831600	CLAG8	MAL7P1.229	gi|167962700|dbj|AB250802.1|
Pf3D7_09_v3	121621	125006	PF3D7_0902800	SERA9	PFI0135c	gi|803375253|dbj|AB733716.1|
Pf3D7_09_v3	270740	274789	PF3D7_0905400	RhopH3	PFI0265c	gi|167962547|dbj|AB250806.1|
Pf3D7_09_v3	1175203	1180762	PF3D7_0929400	RhopH2	PFI1445w	gi|167963178|dbj|AB250805.1|
Pf3D7_09_v3	1413840	1419754	PF3D7_0935800	CLAG9	PFI1730w	gi|167962308|dbj|AB250804.1|
Pf3D7_11_v3	592130	593584	PF3D7_1115700		PF11_0165	gi|9719453|gb|AF282979.1|
Pf3D7_11_v3	1293856	1295724	PF3D7_1133400	AMA1	PF11_0344	gi|182407599|gb|EU586393.1|
Pf3D7_12_v3	1915749	1917798	PF3D7_1246100	ALAS	PFL2210w	gi|1220442|gb|L46348.1|PFADAAS
Pf3D7_13_v3	975403	977175	PF3D7_1323500	PMV	PF13_0133	gi|58372444|gb|AY878742.1|
Pf3D7_13_v3	1416316	1417458	PF3D7_1335000	MSRP1	PF13_0196	gi|237665051|gb|FJ406706.1|
Pf3D7_13_v3	1419086	1420141	PF3D7_1335100	MSP7	PF13_0197	gi|116109338|gb|DQ987539.1|
Pf3D7_13_v3	1497877	1501494	PF3D7_1337200		MAL13P1.186	gi|6690111|gb|AF111814.2|
Pf3D7_14_v3	1368815	1369796	PF3D7_1434200	CAM	PF14_0323	gi|160125|gb|M59349.1|PFACALMOD
Pf3D7_14_v3	1954601	1957675	PF3D7_1447900	MDR2	PF14_0455	gi|294166|gb|L13381.1|PFAMDR2X

Table S2: Genes used for the estimation of FDR and sensitivity.

Sample	Callset	Variant Type	TP	FP	FN	FDR	Sensitivity
HB3(1)	BWA/GATK	SNPs	178	5	33	2.7%	84.4%
		INDELs	45	3	18	6.2%	71.4%
		INDELs (allele match)	42	6	18	12.5%	70.0%
	Cortex	SNPs	188	2	22	1.1%	89.5%
		INDELs	38	4	15	9.5%	71.7%
		INDELs (allele match)	38	4	12	9.5%	76.0%
HB3(2)	BWA/GATK	SNPs	171	1	39	0.6%	81.4%
		INDELs	36	2	21	5.3%	63.2%
		INDELs (allele match)	34	4	19	10.5%	64.2%
	Cortex	SNPs	57	0	137	0.0%	29.4%
		INDELs	11	1	35	8.3%	23.9%
		INDELs (allele match)	11	1	29	8.3%	27.5%

Table S3: FDR and sensitivity estimates for the two replicate samples of clone HB3. See supplementary text for estimation methods.

6. Figures

Index of figures

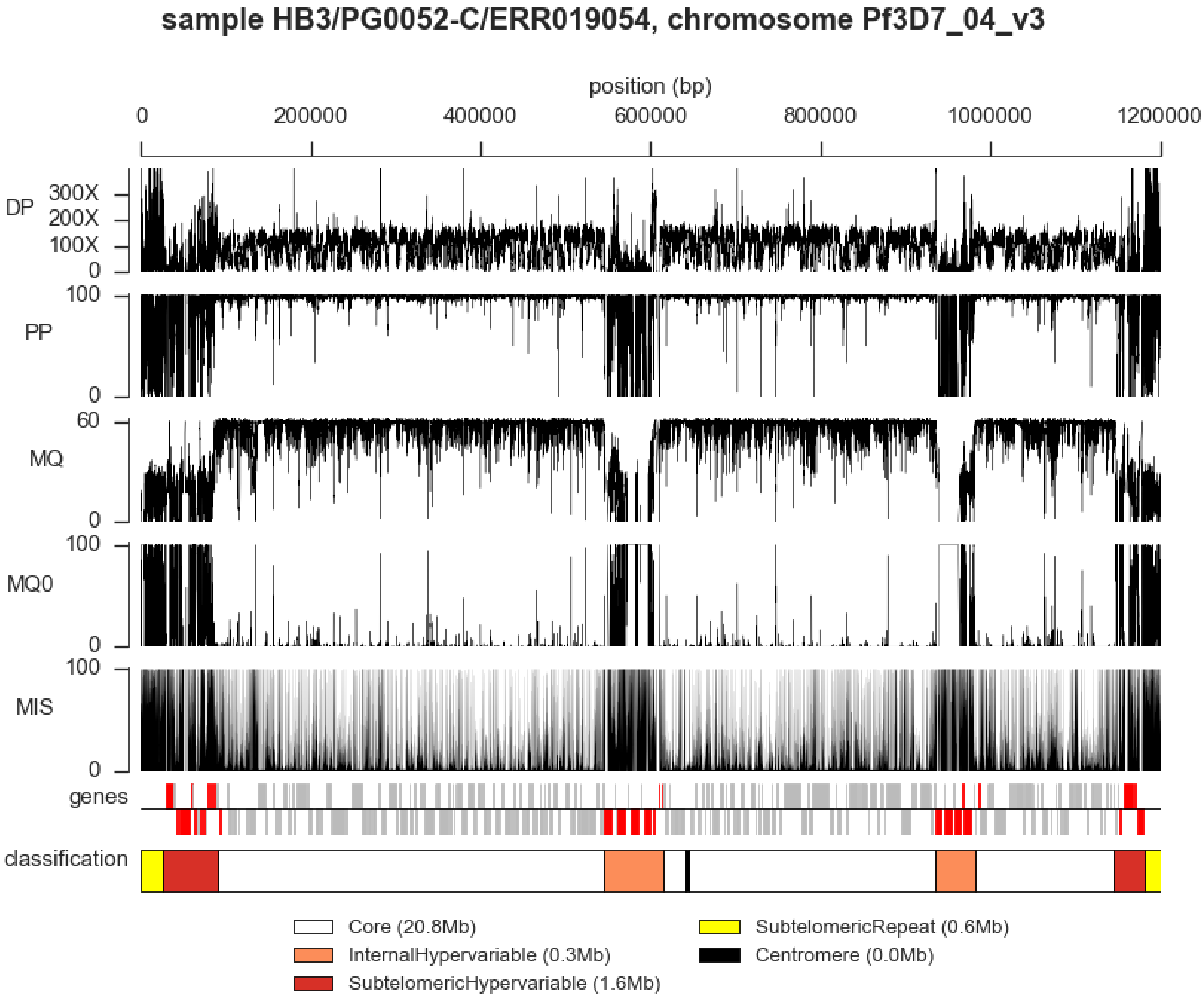


Figure S1: Example of alignment metrics for an individual sample and relationship to genome region classification. The sample shown is HB3/PG0052-C/ERR019054 (parent of 3D7xHB3) and data are shown for the entirety of chromosome 4. DP = total depth of coverage, PP = percent of reads aligned in a proper pair; MQ = root mean square mapping quality of aligned reads; MQ0 = percent of reads aligned ambiguously (mapping quality zero); MIS = percent of reads aligned with a base mismatching the reference. Genes tracks shows forward strand above the line, reverse strand below the line; genes in red are var/rif/stevor. Genome region classification is shown in the bottom track, colours as in the legend.

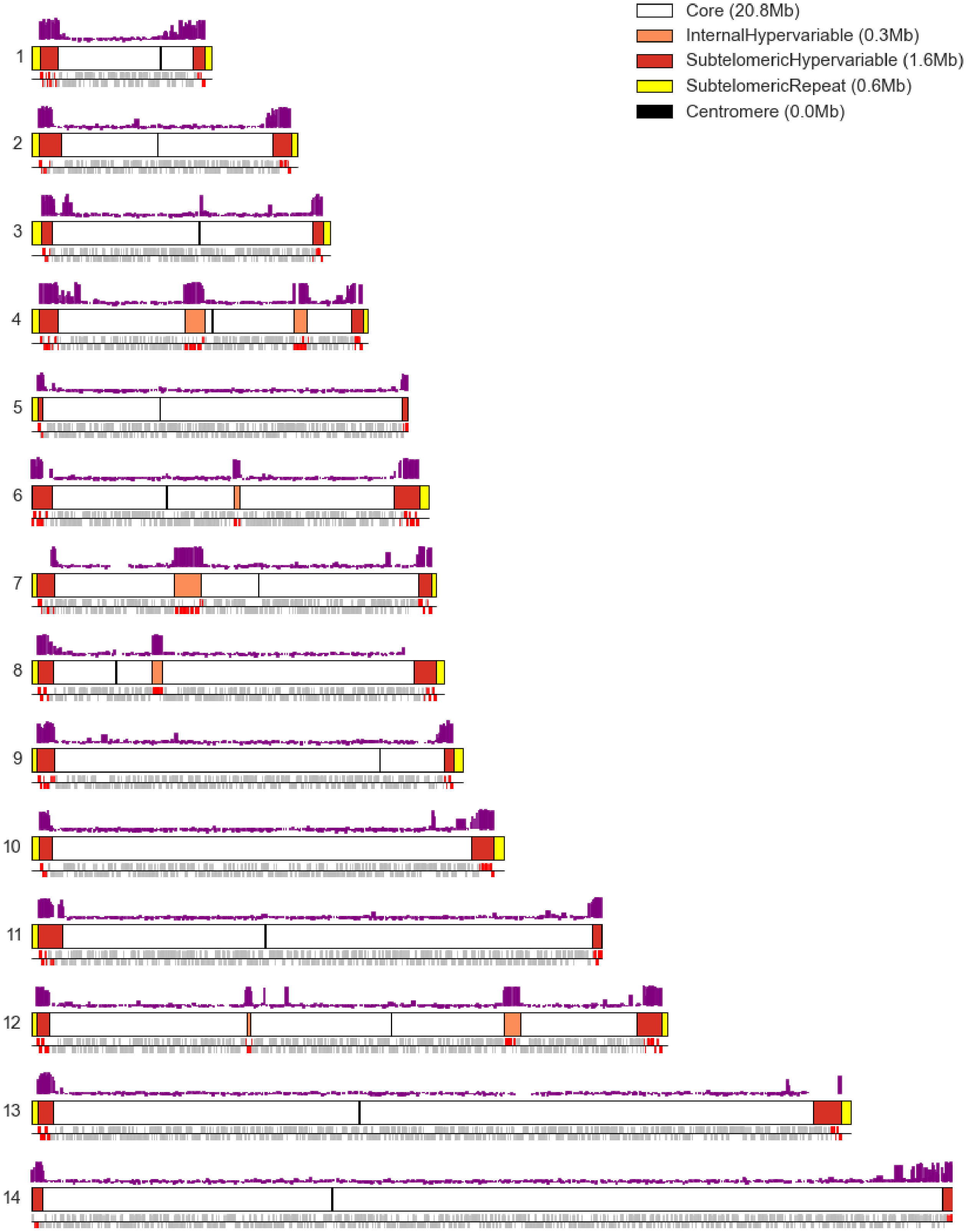


Figure S2: Genome region classification. Each sub-plot corresponds to one of the fourteen nuclear chromosomes. The central bar in each sub-plot shows the genome region classification coloured according to the legend. Above the central bar in purple are levels of heterochromatin protein 1 (HP1) per gene from (Flueck et al., 2009). Below in grey are genes, with positive and negative strands plotted above and below the line respectively; genes in the rif, stevor and var families are shown in red.

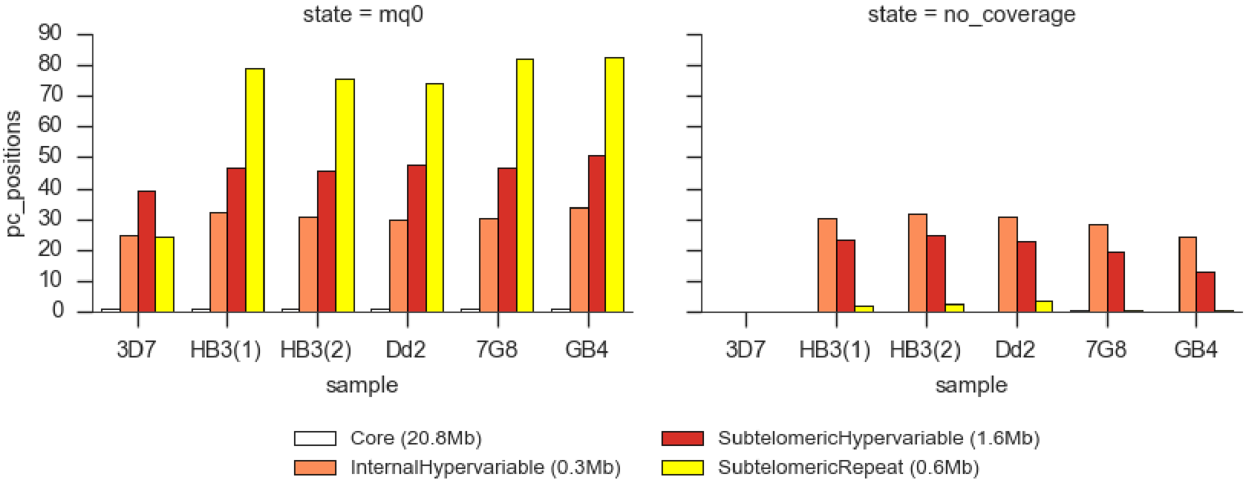


Figure S3: Summary of alignment characteristics for different genome region classes. The left-hand sub-plot shows the percentage of positions with more than 10% of reads aligned ambiguously (mapping quality zero). The right-hand sub-plot shows the percentage of positions without any coverage whatsoever.

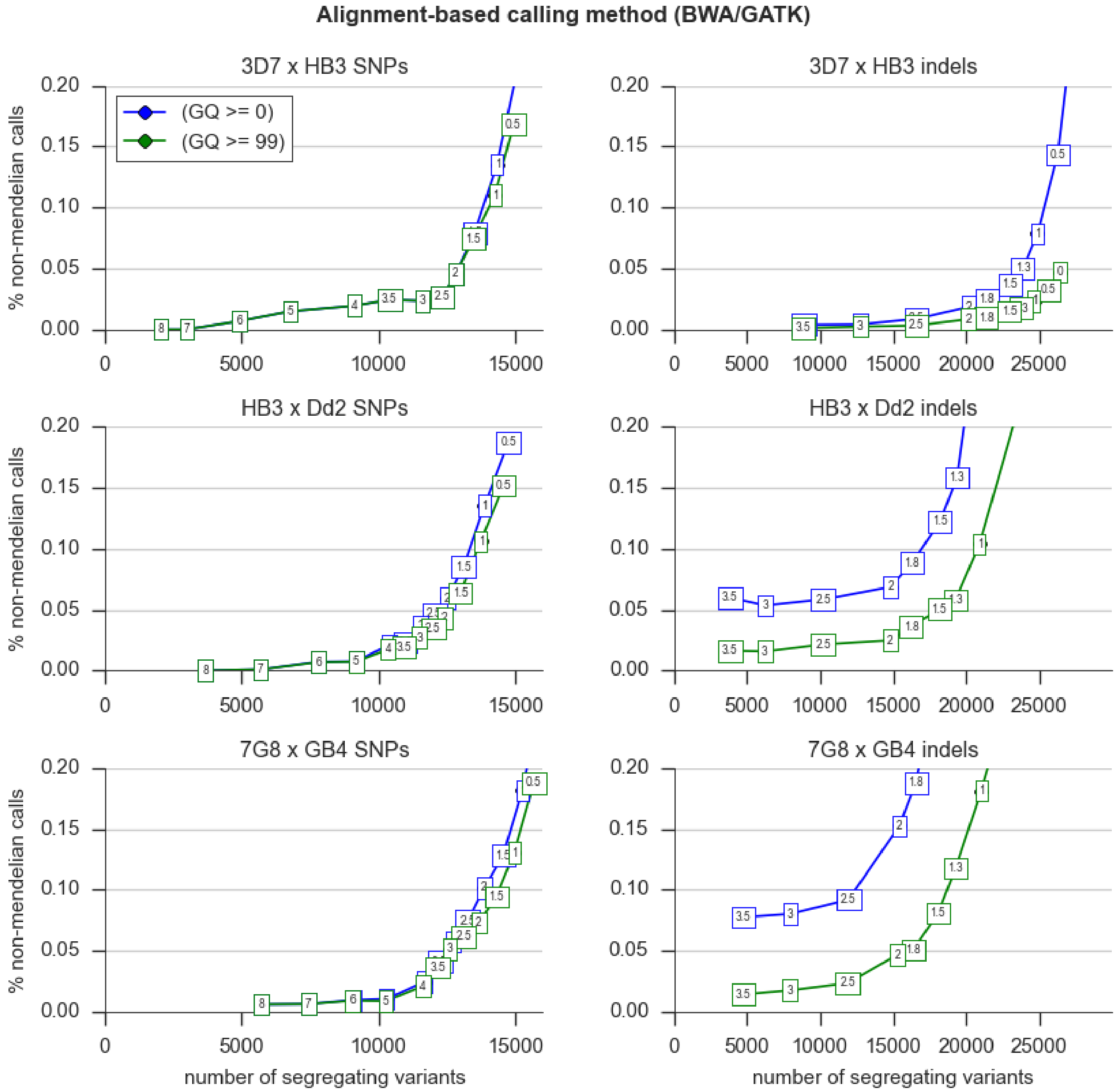


Figure S4: Using Mendelian error as a guide to filtering variants and genotype calls from the alignment-based calling method. Each point plotted corresponds to variants filtered according to a minimum value of the VQSLOD annotation and genotype calls filtered according to a minimum value of GQ. The VQSLOD threshold value is shown labelling the point, the colour indicates the GQ threshold according to the legend.

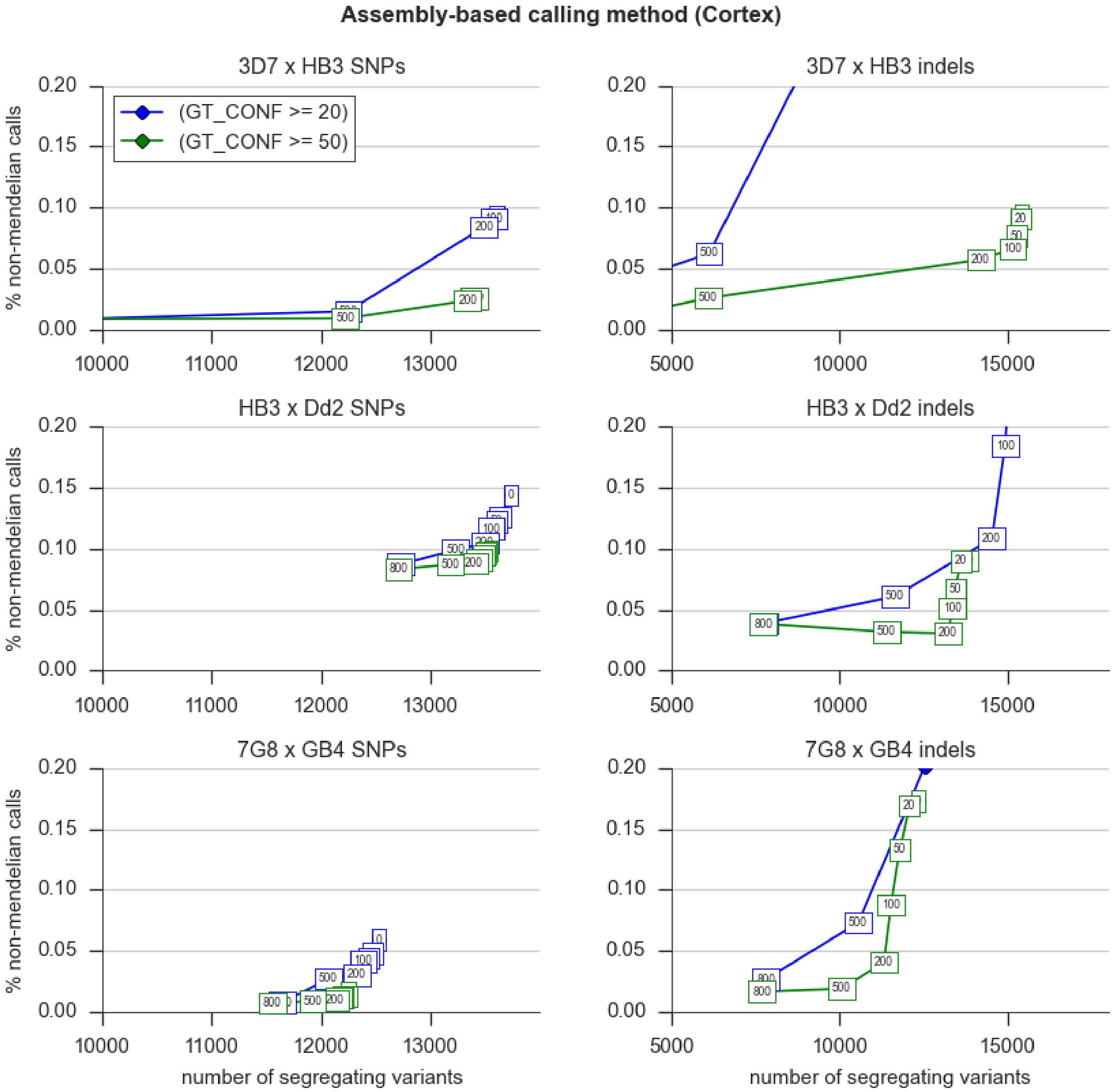
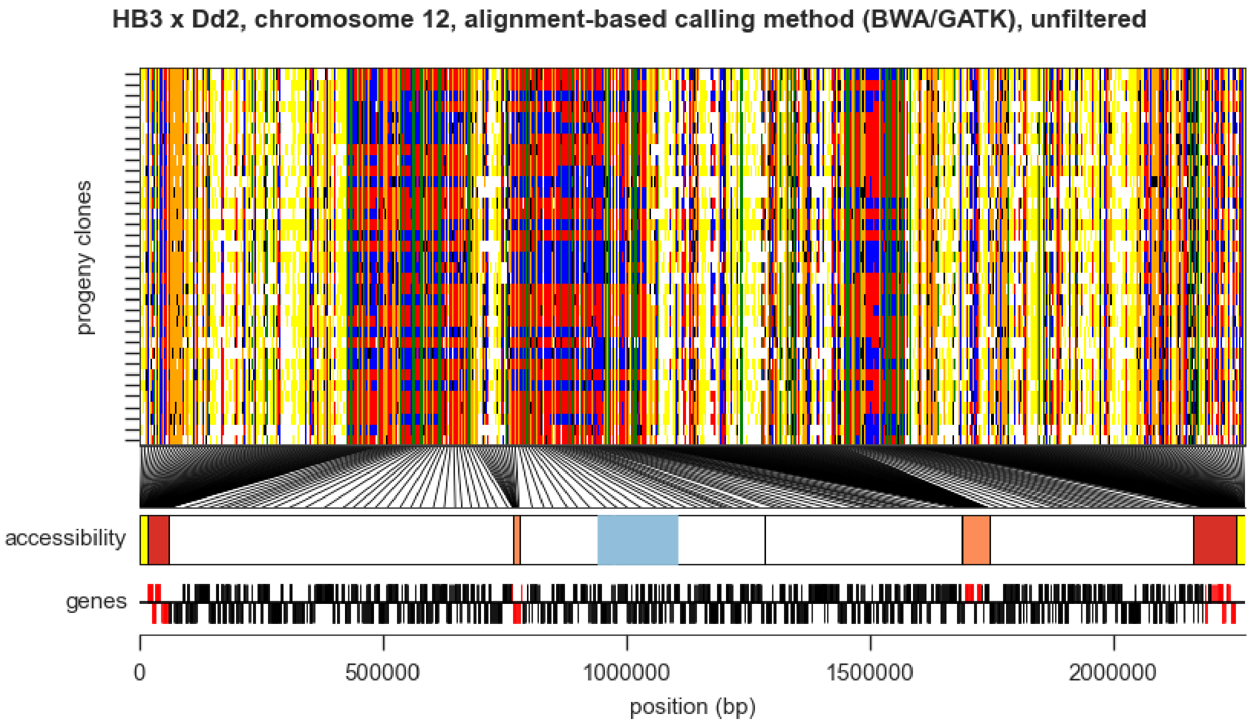


Figure S5: Using Mendelian error as a guide to filtering variants and genotype calls from the assembly-based calling method. Each point plotted corresponds to variants filtered according to a minimum value of the SITE_CONF annotation and genotype calls filtered according to a minimum value of GT_CONF. The SITE_CONF threshold value is shown labelling the point, the colour indicates the GT_CONF threshold according to the legend.



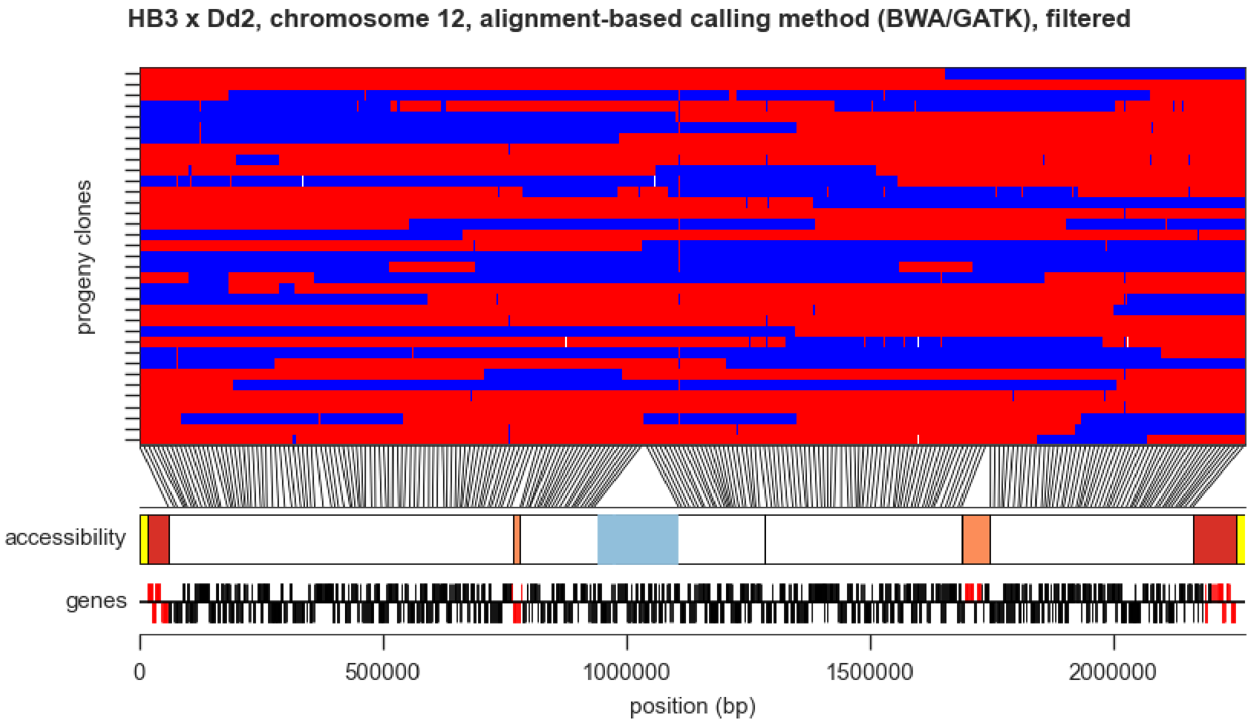


Figure S6: Illustration of the alignment-based callset before and after variant filtration. The upper plot shows the raw variant calls, the lower plot shows the filtered variant calls. The main subplot in each shows each sample as a row and each variant as a column, painting genotype calls as follow: red: parent 1 allele; blue: parent 2 allele; white: missing genotype call; grey: filtered genotype call; yellow: parent genotype missing; black: non-Mendelian genotype; orange: reference allele and both parents reference also; green: alternate allele and both parents alternate also. Lines from the inheritance subplot to the accessibility track indicate the physical position of variants, with one line drawn for every 100 variants in the upper (unfiltered) plot and one line for every 10 variants in the lower (filtered) plot.

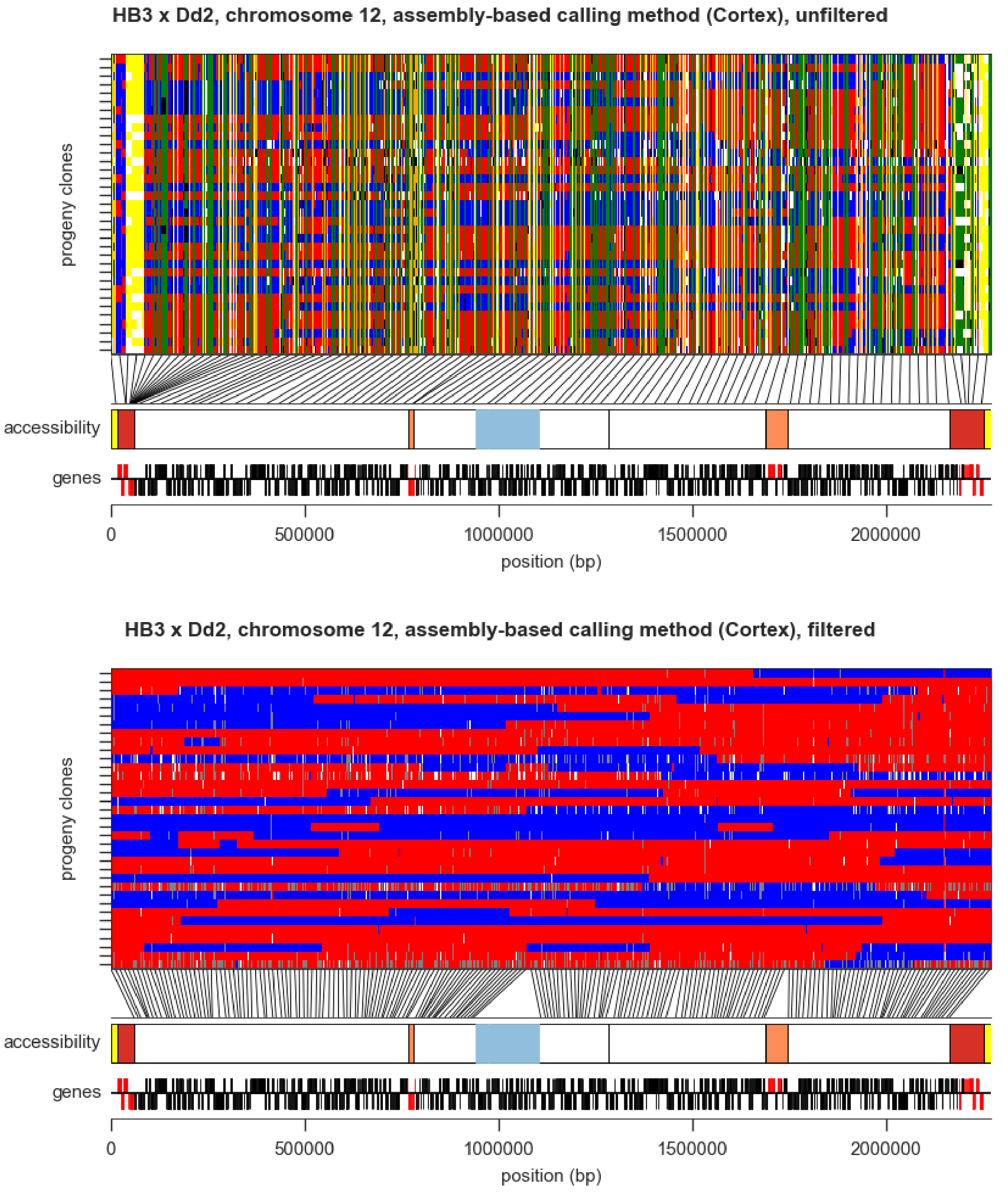
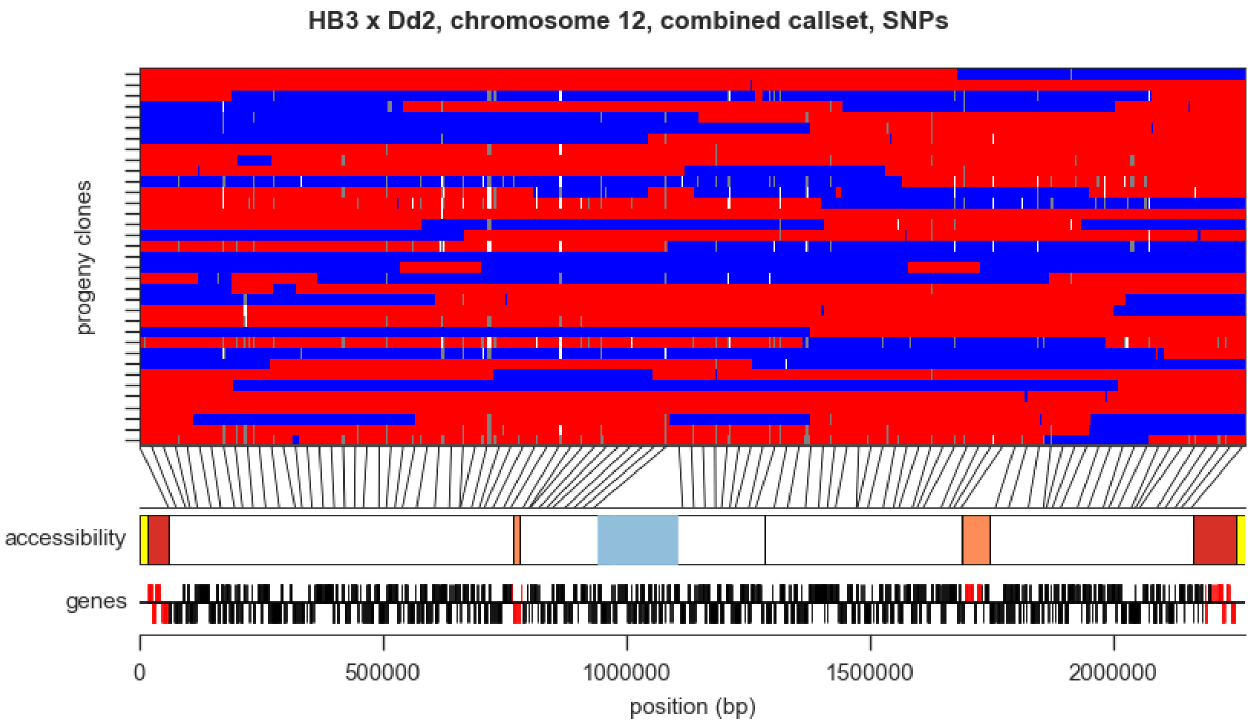


Figure S7: Illustration of the assembly-based callset before and after variant filtration. The upper plot shows the raw variant calls, the lower plot shows the filtered variant calls. The main subplot in each shows each sample as a row and each variant as a column, painting genotype calls as described in Figure S6.



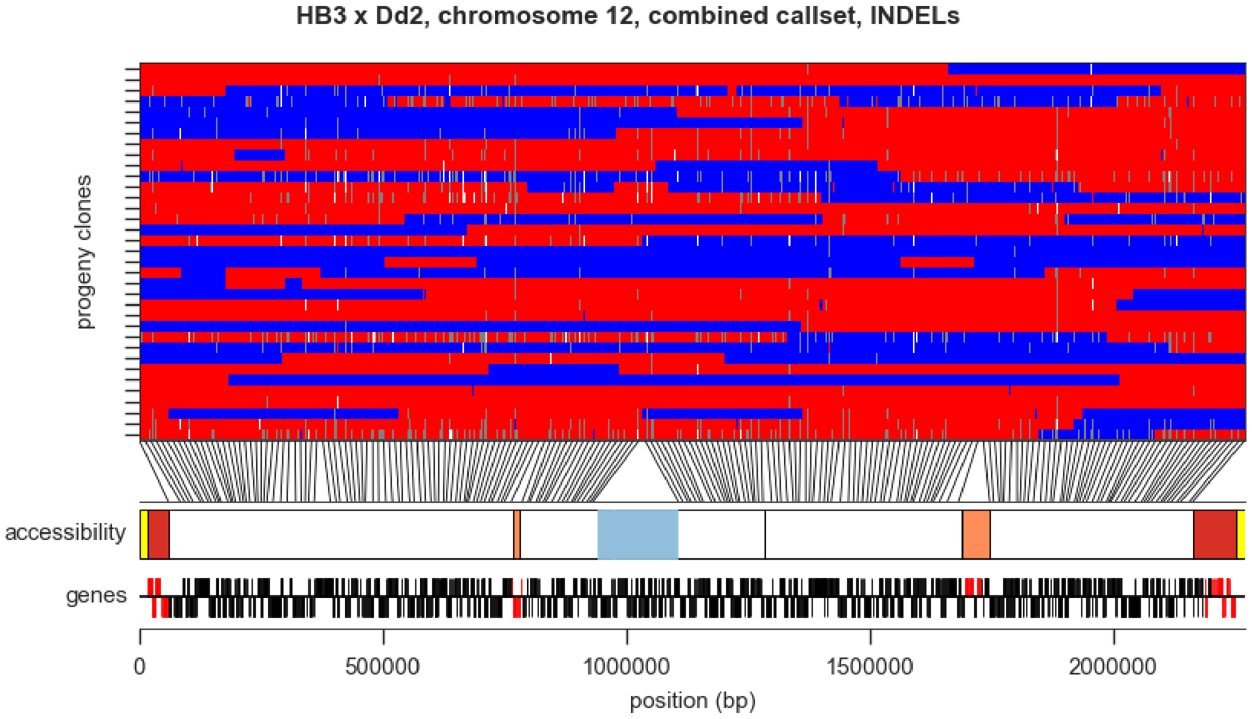


Figure S8: Comparison of SNP and INDEL calls. The main subplot in each plot shows each sample as a row and each variant as a column, painting genotype calls as described in Figure S6. Lines from the inheritance subplot to the accessibility track indicate the physical position of variants, with one line drawn for every 10 variants.

7. References

Birren, B., Lander, E., Galagan, J., Nusbaum, C., Devon, K., Henn, M., … Hartl, D. (2006). Plasmodium falciparum HB3, whole genome shotgun sequencing project. Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, MA 02141, USA. Retrieved from http://www.ncbi.nlm.nih.gov/nuccore/AANS00000000

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. doi:10.1038/ng.806

Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., … Wellems, T. E. (2000). Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Molecular Cell*, *6*(4), 861–71. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2944663&tool=pmcentrez&rendertype=abstract

Flueck, C., Bartfai, R., Volz, J., Niederwieser, I., Salcedo-Amaya, A. M., Alako, B. T. F., … Voss, T. S. (2009). Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathogens*, *5*(9), e1000569. doi:10.1371/journal.ppat.1000569

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–60. doi:10.1093/bioinformatics/btp324

Manske, H. M., & Kwiatkowski, D. P. (2009). LookSeq: a browser-based viewer for deep sequencing data. *Genome Research*, *19*(11), 2125–32. doi:10.1101/gr.093443.109

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–303. doi:10.1101/gr.107524.110

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). *Current Protocols in Bioinformatics*. (A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, & J. R. Yates, Eds.)*Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* (Vol. 11). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471250953