

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

## Supplementary information

### Table of Contents

1. Whole genome sequencing.....	1
2. Sequence alignment and genome region classification.....	2
3. Variant discovery and genotype calling.....	4
3.1. Alignment-based calling method (BWA/GATK).....	4
3.2. Assembly-based calling method (Cortex).....	6
3.3. Combined callset.....	6
3.4. Genotype concordance between biological replicates.....	7
3.5. Estimation of FDR and sensitivity.....	7
4. Recombination analyses.....	8
4.1. Determination of maximal block length for conversion tracts.....	8
5. Tables.....	9
6. Figures.....	13
7. References.....	22

### Index of Figures

Figure S1.....	14
Figure S2.....	15
Figure S3.....	16
Figure S4.....	17
Figure S5.....	18
Figure S6.....	19
Figure S7.....	20
Figure S8.....	21

### Index of Tables

Table S1.....	10
Table S2.....	11
Table S3.....	12

## 1. Whole genome sequencing

Sample preparation and sequencing was performed as described in (Manske, Miotto et al., 2012) except that PCR-free library preparation was used throughout (Kozarewa et al., 2009). All sequences were deposited in the European Nucleotide Archive and a mapping from sample

identifiers to ENA accessions is given in **Table S1** and in the web application at @@URL.

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 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).

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`<sup>1</sup>. 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 visualised using the LookSeq web application (Manske & Kwiatkowski, 2009), which can be

---

1 <https://github.com/alimanfoo/pysamstats>

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 S2**.

### 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 S3**). In spite of these criteria there remained some discordance between the Birren et al. assembly and 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 S4**.

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. Calling CO and NCO recombination events

Two types of recombination event are expected: crossover (CO) and non-crossover (NCO). A CO is a reciprocal exchange accompanied by a conversion tract, whereas a NCO is a conversion tract without reciprocal exchange (Youds & Boulton, 2011). A conversion tract can either be simple (all alleles converted to the same parent) or complex (containing switches between parental alleles converted). In studies of yeast or other organisms where all four daughters of a single meiosis can be captured and genotyped, NCO events can be inferred directly from a non-Mendelian ratio of segregation of alleles. In this study each progeny clone is the result of an independent meiosis and thus unequal segregation cannot be observed. However CO and NCO events can be inferred from the patterns of allelic inheritance in the progeny of each cross. Two or more CO events are unlikely to appear in the same progeny clone in close proximity, and thus two or more nearby switches in allelic inheritance are more likely to indicate a conversion tract.

To determine an appropriate threshold for differentiating CO from NCO events, for each cross we first identified contiguous blocks of markers within each progeny where alleles were all inherited from the same parent. Boundaries between such blocks thus indicate switches in parental inheritance. Each such block has a minimal size, given by the distance between the outer markers within the block, and a maximal size, given by the distance between the markers flanking the block. The distribution of minimal block sizes was plotted for each cross (**Figure S9**). The resulting distributions were bimodal for all three crosses with a minor peak of blocks around ~@@kb extending upward to ~@@kb. This minor peak would not be expected from CO events, and suggests an expected size range for NCO conversion tracts, although at this stage we have not accounted for complex conversion tracts (which will appear as multiple adjacent short blocks).

To determine a size limit below which to assume that blocks indicate conversion tracts, we computed the number of blocks of a given size that would be expected from CO events alone, using previously published estimates for the CO recombination rate, which should be reasonably accurate given that a high marker resolution is not required to ascertain CO events. Assuming a uniform CO recombination rate of @@ Morgan/Mb and @@ independent meioses we would expect to observe @@ CO events within 5kb, @@ within 10kb and @@ within 20kb. This model is overly simplistic but serves to provide an estimate for the smallest block size expected from CO (double cross-over) events which is conservative because CO interference is likely to reduce further the true probability of observing smaller blocks. We thus assumed that all blocks with minimal length shorter than 10kb were either whole or part of conversion tracts.

@@TODO



# 5. Tables

## Index of Tables

Table S1.....	11
Table S2.....	12
Table S3.....	13
Table S4.....	14

Cross	Clone	Sample	Run	Instrument	Coverage
3D7 x HB3	3D7	PG0051-C	ERR019061	Illumina Genome Analyzer II	122X
3D7 x HB3	C01	PG0065-C	ERR019064	Illumina Genome Analyzer II	163X
3D7 x HB3	C01	PG0062-C	ERR019070	Illumina Genome Analyzer II	108X
3D7 x HB3	C02	PG0055-C	ERR019066	Illumina Genome Analyzer II	102X
3D7 x HB3	C02	PG0053-C	ERR019067	Illumina Genome Analyzer II	73X
3D7 x HB3	C02	PG0056-C	ERR019068	Illumina Genome Analyzer II	84X
3D7 x HB3	C02	PG0067-C	ERR019073	Illumina Genome Analyzer II	126X
3D7 x HB3	C03	PG0066-C	ERR019072	Illumina Genome Analyzer II	79X
3D7 x HB3	C04	PG0061-C	ERR019059	Illumina Genome Analyzer II	165X
3D7 x HB3	C05	PG0068-C	ERR019065	Illumina Genome Analyzer II	41X
3D7 x HB3	C06	PG0069-C	ERR019055	Illumina Genome Analyzer II	135X
3D7 x HB3	C07	PG0070-C	ERR019056	Illumina Genome Analyzer II	144X
3D7 x HB3	C08	PG0071-C	ERR019074	Illumina Genome Analyzer II	120X
3D7 x HB3	C09	PG0072-C	ERR019057	Illumina Genome Analyzer II	173X
3D7 x HB3	C10	PG0063-C	ERR019060	Illumina Genome Analyzer II	108X
3D7 x HB3	C11	PG0064-C	ERR019071	Illumina Genome Analyzer II	48X
3D7 x HB3	C12	PG0058-C	ERR019063	Illumina Genome Analyzer II	51X
3D7 x HB3	C13	PG0054-C	ERR019062	Illumina Genome Analyzer II	95X
3D7 x HB3	C14	PG0060-C	ERR019058	Illumina Genome Analyzer II	102X
3D7 x HB3	C15	PG0057-C	ERR019069	Illumina Genome Analyzer II	56X
3D7 x HB3	HB3	PG0052-C	ERR019054	Illumina Genome Analyzer II	100X
7G8 x GB4	7G8	PG0083-C	ERR027099	Illumina Genome Analyzer II	87X
7G8 x GB4	AL2	PG0103-CW	ERR045627	Illumina HiSeq 2000	127X
7G8 x GB4	AUD	PG0112-C	ERR029406	Illumina Genome Analyzer II	129X
7G8 x GB4	AUD	PG0112-CW	ERR045639	Illumina HiSeq 2000	88X
7G8 x GB4	D2	PG0094-CW	ERR045632	Illumina HiSeq 2000	153X
7G8 x GB4	DAN	PG0098-C	ERR027110	Illumina Genome Analyzer II	140X
7G8 x GB4	DEV	PG0081-CW	ERR045633	Illumina HiSeq 2000	89X
7G8 x GB4	GB4	PG0084-C	ERR027100	Illumina Genome Analyzer II	104X
7G8 x GB4	JB12	PG0099-C	ERR029146	Illumina Genome Analyzer II	120X
7G8 x GB4	JB8	PG0087-C	ERR029091	Illumina Genome Analyzer II	103X
7G8 x GB4	JC3	PG0077-CW	ERR045636	Illumina HiSeq 2000	94X
7G8 x GB4	JC9	PG0111-C	ERR029409	Illumina Genome Analyzer II	122X
7G8 x GB4	JC9	PG0111-CW	ERR045634	Illumina HiSeq 2000	121X
7G8 x GB4	JE11	PG0100-C	ERR029404	Illumina Genome Analyzer II	134X
7G8 x GB4	JE11	PG0100-CW	ERR045630	Illumina HiSeq 2000	55X
7G8 x GB4	JF6	PG0079-C	ERR027102	Illumina Genome Analyzer II	181X
7G8 x GB4	JF6	PG0079-CW	ERR045637	Illumina HiSeq 2000	94X
7G8 x GB4	JON	PG0107-C	ERR029408	Illumina Genome Analyzer II	180X
7G8 x GB4	KA6	PG0091-C	ERR027117	Illumina Genome Analyzer II	80X
7G8 x GB4	KB8	PG0104-C	ERR029148	Illumina Genome Analyzer II	116X
7G8 x GB4	KB8	PG0104-CW	ERR045642	Illumina HiSeq 2000	81X
7G8 x GB4	KH7	PG0088-C	ERR027111	Illumina Genome Analyzer II	96X
7G8 x GB4	LA10	PG0086-C	ERR029090	Illumina Genome Analyzer II	119X
7G8 x GB4	LA10	PG0086-CW	ERR045629	Illumina HiSeq 2000	66X
7G8 x GB4	NF10	PG0096-C	ERR027108	Illumina Genome Analyzer II	75X
7G8 x GB4	NIC	PG0095-C	ERR027107	Illumina Genome Analyzer II	70X
7G8 x GB4	NIC	PG0095-CW	ERR045631	Illumina HiSeq 2000	80X
7G8 x GB4	QF5	PG0078-C	ERR029092	Illumina Genome Analyzer II	147X
7G8 x GB4	QF5	PG0078-CW	ERR045638	Illumina HiSeq 2000	82X
7G8 x GB4	TF1	PG0080-C	ERR027103	Illumina Genome Analyzer II	73X
7G8 x GB4	WC4	PG0082-C	ERR029093	Illumina Genome Analyzer II	78X
7G8 x GB4	WE2	PG0085-C	ERR027101	Illumina Genome Analyzer II	124X
7G8 x GB4	WF12	PG0097-C	ERR027109	Illumina Genome Analyzer II	109X
7G8 x GB4	XB3	PG0093-C	ERR029105	Illumina Genome Analyzer II	214X
7G8 x GB4	XD8	PG0105-C	ERR029144	Illumina Genome Analyzer II	121X
7G8 x GB4	XD8	PG0105-CW	ERR045628	Illumina HiSeq 2000	122X
7G8 x GB4	XE7	PG0106-C	ERR029407	Illumina Genome Analyzer II	250X
7G8 x GB4	XF12	PG0102-C	ERR029143	Illumina Genome Analyzer II	141X

7G8 x GB4	XF12	PG0102-CW	ERR045635	Illumina HiSeq 2000	96X
7G8 x GB4	XG10	PG0109-C	ERR029405	Illumina Genome Analyzer II	61X
HB3 x Dd2	1BB5	PG0023-C	ERR015449	Illumina Genome Analyzer II	22X
HB3 x Dd2	3BA6	PG0022-Cx	ERR126027	Illumina HiSeq 2000	32X
HB3 x Dd2	3BD5	PG0024-C	ERR019053	Illumina Genome Analyzer II	92X
HB3 x Dd2	7C101	PG0074-C	ERR019048	Illumina Genome Analyzer II	98X
HB3 x Dd2	7C111	PG0038-C	ERR015457	Illumina Genome Analyzer II	148X
HB3 x Dd2	7C12	PG0035-Cx	ERR037704	Illumina HiSeq 2000	637X
HB3 x Dd2	7C126	PG0047-C	ERR015452	Illumina Genome Analyzer II	187X
HB3 x Dd2	7C140	PG0039-C	ERR015454	Illumina Genome Analyzer II	78X
HB3 x Dd2	7C159	PG0040-Cx	ERR107475	Illumina HiSeq 2000	59X
HB3 x Dd2	7C16	PG0036-C	ERR015455	Illumina Genome Analyzer II	26X
HB3 x Dd2	7C170	PG0041-C	ERR015446	Illumina Genome Analyzer II	130X
HB3 x Dd2	7C183	PG0042-C	ERR015448	Illumina Genome Analyzer II	118X
HB3 x Dd2	7C188	PG0030-C	ERR019046	Illumina Genome Analyzer II	171X
HB3 x Dd2	7C20	PG0037-C	ERR015451	Illumina Genome Analyzer II	82X
HB3 x Dd2	7C3	PG0034-C	ERR019047	Illumina Genome Analyzer II	142X
HB3 x Dd2	7C408	PG0031-C	ERR015458	Illumina Genome Analyzer II	51X
HB3 x Dd2	7C421	PG0043-C	ERR015459	Illumina Genome Analyzer II	164X
HB3 x Dd2	7C424	PG0044-C	ERR019043	Illumina Genome Analyzer II	172X
HB3 x Dd2	7C46	PG0046-Cx	ERR107476	Illumina HiSeq 2000	62X
HB3 x Dd2	7C7	PG0048-C	ERR019049	Illumina Genome Analyzer II	110X
HB3 x Dd2	B1SD	PG0015-C	ERR019044	Illumina Genome Analyzer II	91X
HB3 x Dd2	B4R3	PG0018-C	ERR019042	Illumina Genome Analyzer II	115X
HB3 x Dd2	CH3_116	PG0032-Cx	ERR037703	Illumina HiSeq 2000	186X
HB3 x Dd2	CH3_61	PG0033-Cx	ERR175544	Illumina HiSeq 2000	68X
HB3 x Dd2	D43	PG0029-Cx	ERR107474	Illumina HiSeq 2000	34X
HB3 x Dd2	DD2	PG0008-CW	ERR012840	Illumina Genome Analyzer II	122X
HB3 x Dd2	GC03	PG0021-C	ERR015447	Illumina Genome Analyzer II	152X
HB3 x Dd2	GC06	PG0028-C	ERR015456	Illumina Genome Analyzer II	54X
HB3 x Dd2	HB3	PG0004-CW	ERR012788	Illumina Genome Analyzer II	80X
HB3 x Dd2	QC01	PG0017-C	ERR019050	Illumina Genome Analyzer II	117X
HB3 x Dd2	QC13	PG0016-C	ERR012895	Illumina Genome Analyzer II	68X
HB3 x Dd2	QC23	PG0045-C	ERR012892	Illumina Genome Analyzer II	115X
HB3 x Dd2	QC34	PG0026-C	ERR015453	Illumina Genome Analyzer II	55X
HB3 x Dd2	SC01	PG0025-C	ERR019045	Illumina Genome Analyzer II	149X
HB3 x Dd2	SC05	PG0019-C	ERR019051	Illumina Genome Analyzer II	97X
HB3 x Dd2	TC05	PG0027-C	ERR015450	Illumina Genome Analyzer II	115X
HB3 x Dd2	TC08	PG0020-C	ERR019052	Illumina Genome Analyzer II	144X

*Table S1: Samples and sequencing runs used in this study.*

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 S2: 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 PFADHFR-TS
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 S3: Genes used for the estimation of FDR and sensitivity.

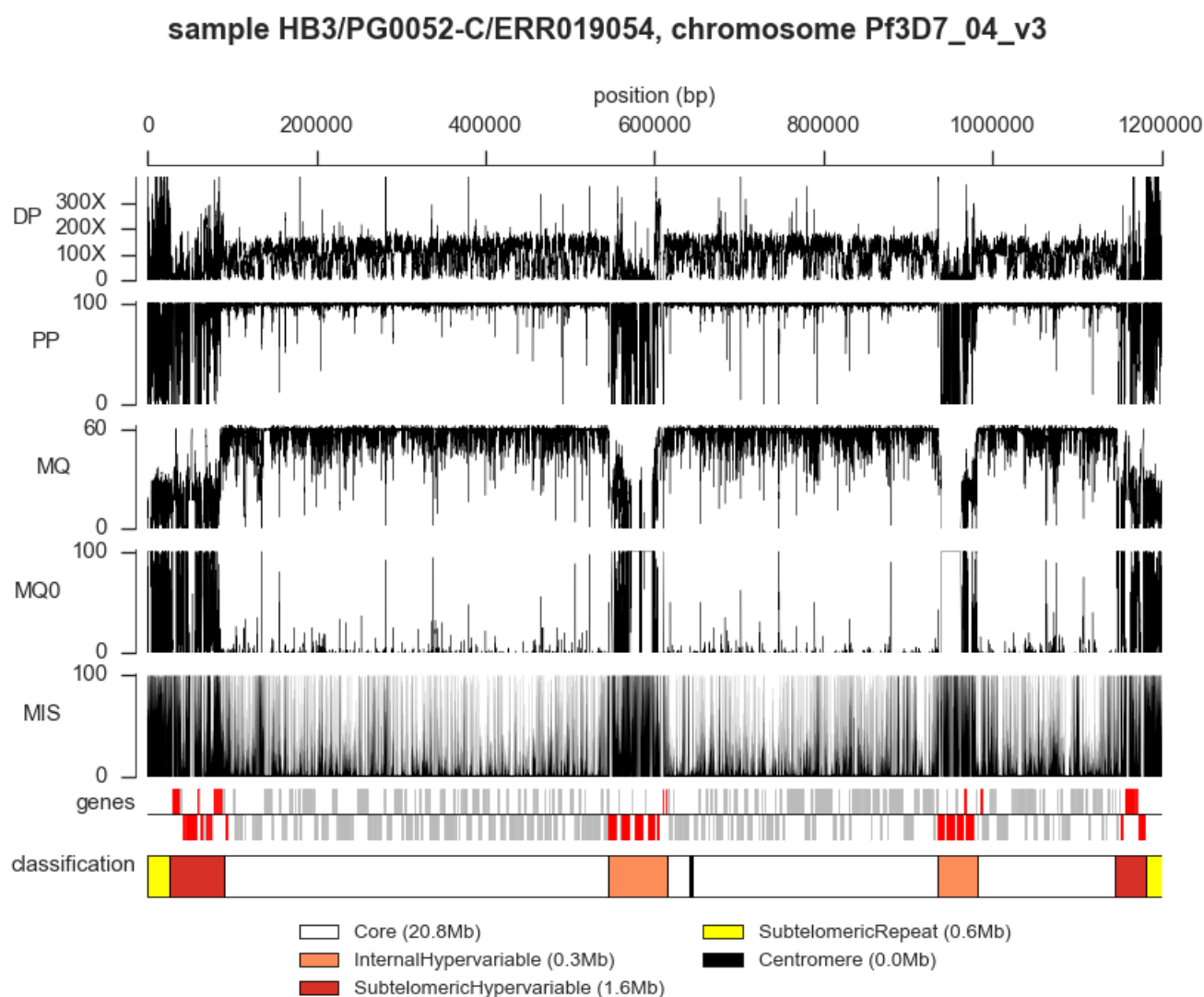
Sample	Callset	Variant Type	TP	FP	FN	FDR	Sensitivity
HB3(1)	BWA/GATK	SNPs	178	5	33	2.7%	84.4%
		INDELs	45	3	18	6.2%	71.4%
		INDELs (allele match)	42	6	18	12.5%	70.0%
	Cortex	SNPs	188	2	22	1.1%	89.5%
		INDELs	38	4	15	9.5%	71.7%
		INDELs (allele match)	38	4	12	9.5%	76.0%
	BWA/GATK	SNPs	171	1	39	0.6%	81.4%
		INDELs	36	2	21	5.3%	63.2%
		INDELs (allele match)	34	4	19	10.5%	64.2%
HB3(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 S4: 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.....	11
Figure S2.....	12
Figure S3.....	13
Figure S4.....	14
Figure S5.....	15
Figure S6.....	16
Figure S7.....	17
Figure S8.....	18



*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.

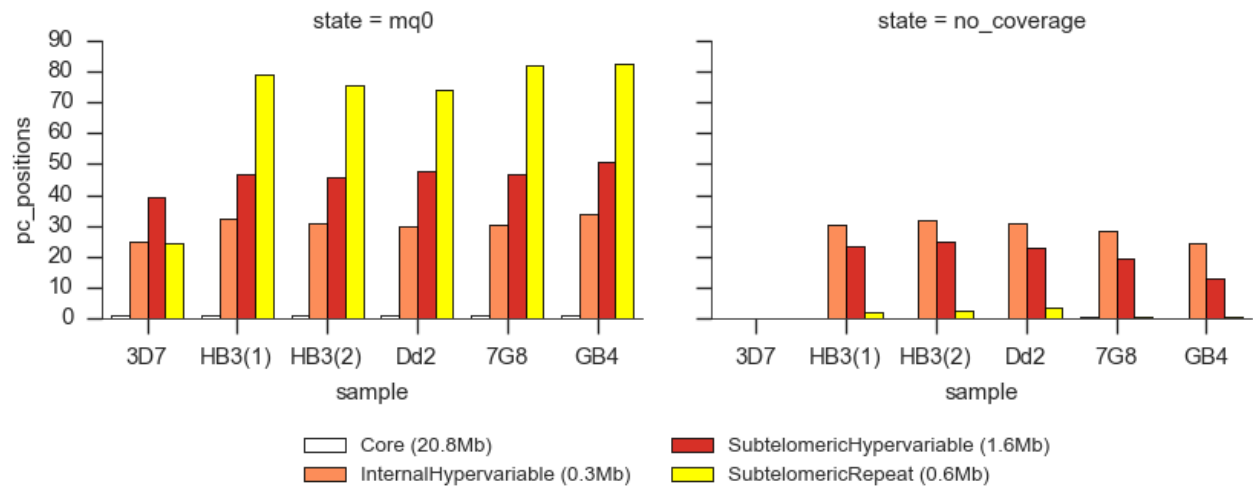


Figure S3: Summary of alignment characteristics for different genome region classes. The left-hand sub-plot shows the percentage of positions with more than 10% of reads aligned ambiguously (mapping quality zero). The right-hand sub-plot shows the percentage of positions without any coverage whatsoever.

### Alignment-based calling method (BWA/GATK)

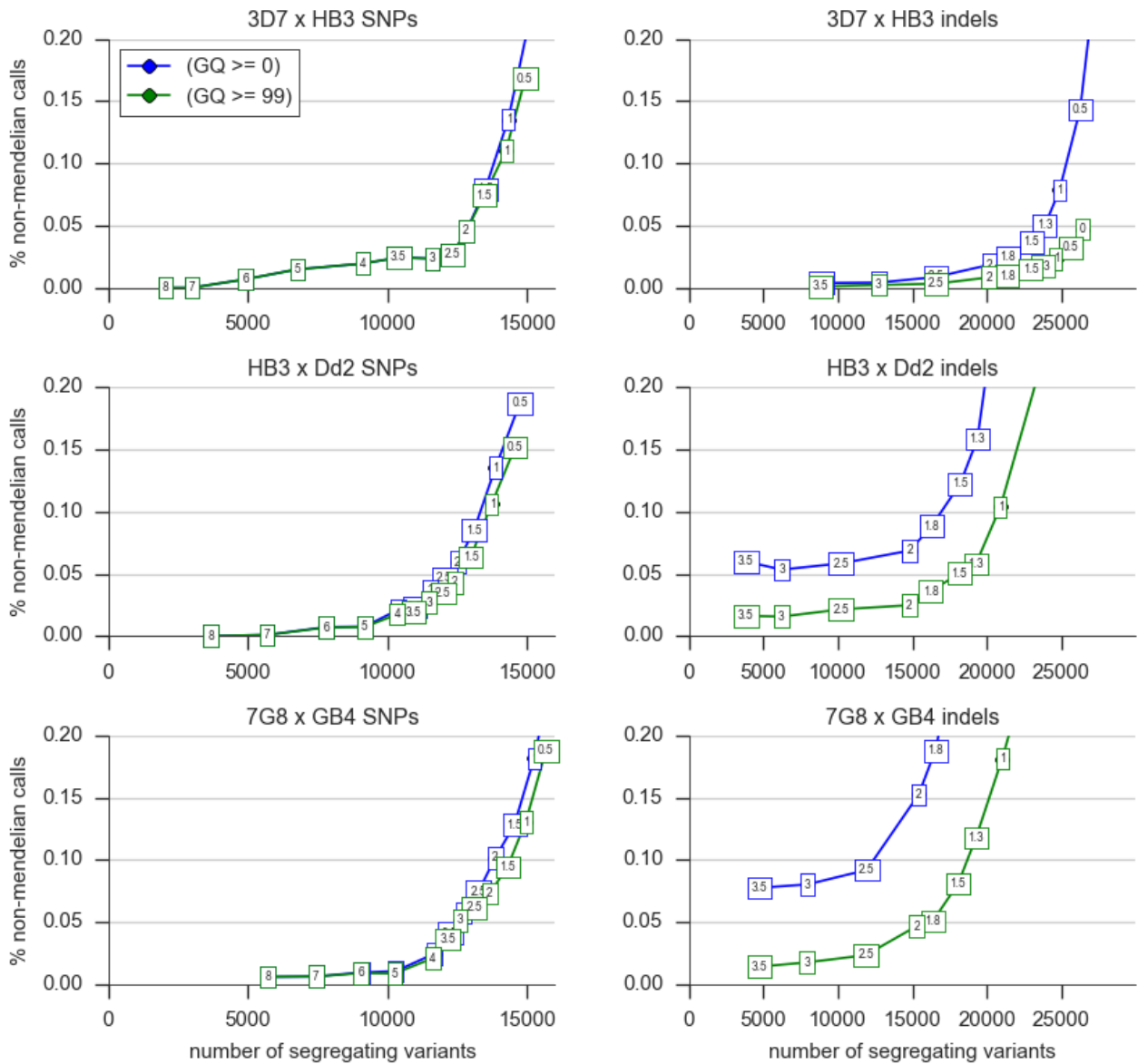


Figure S4: Using Mendelian error as a guide to filtering variants and genotype calls from the alignment-based calling method. Each point plotted corresponds to variants filtered according to a minimum value of the VQSLOD annotation and genotype calls filtered according to a minimum value of GQ. The VQSLOD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

### Assembly-based calling method (Cortex)

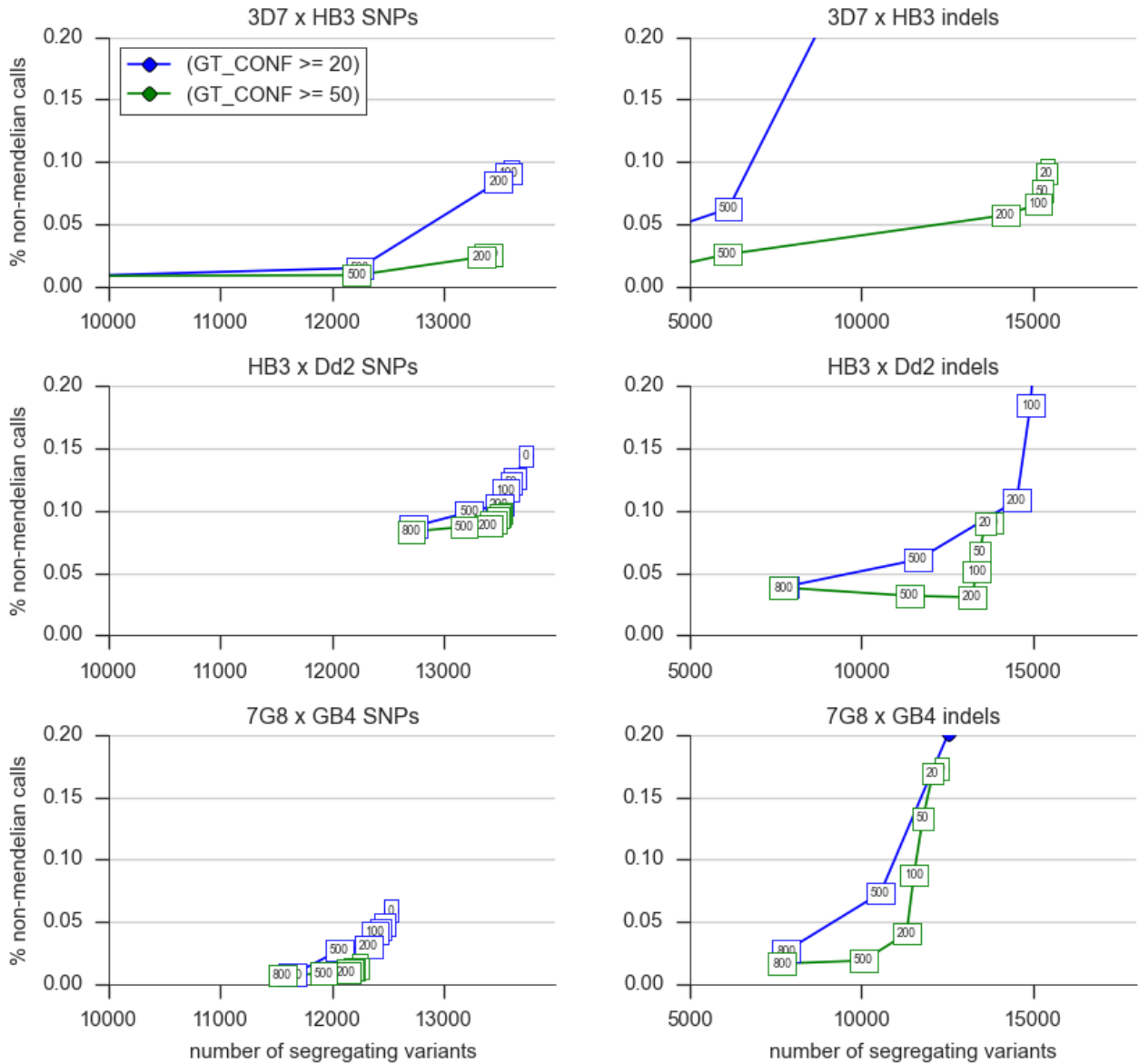


Figure S5: Using Mendelian error as a guide to filtering variants and genotype calls from the assembly-based calling method. Each point plotted corresponds to variants filtered according to a minimum value of the  $SITE\_CONF$  annotation and genotype calls filtered according to a minimum value of  $GT\_CONF$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

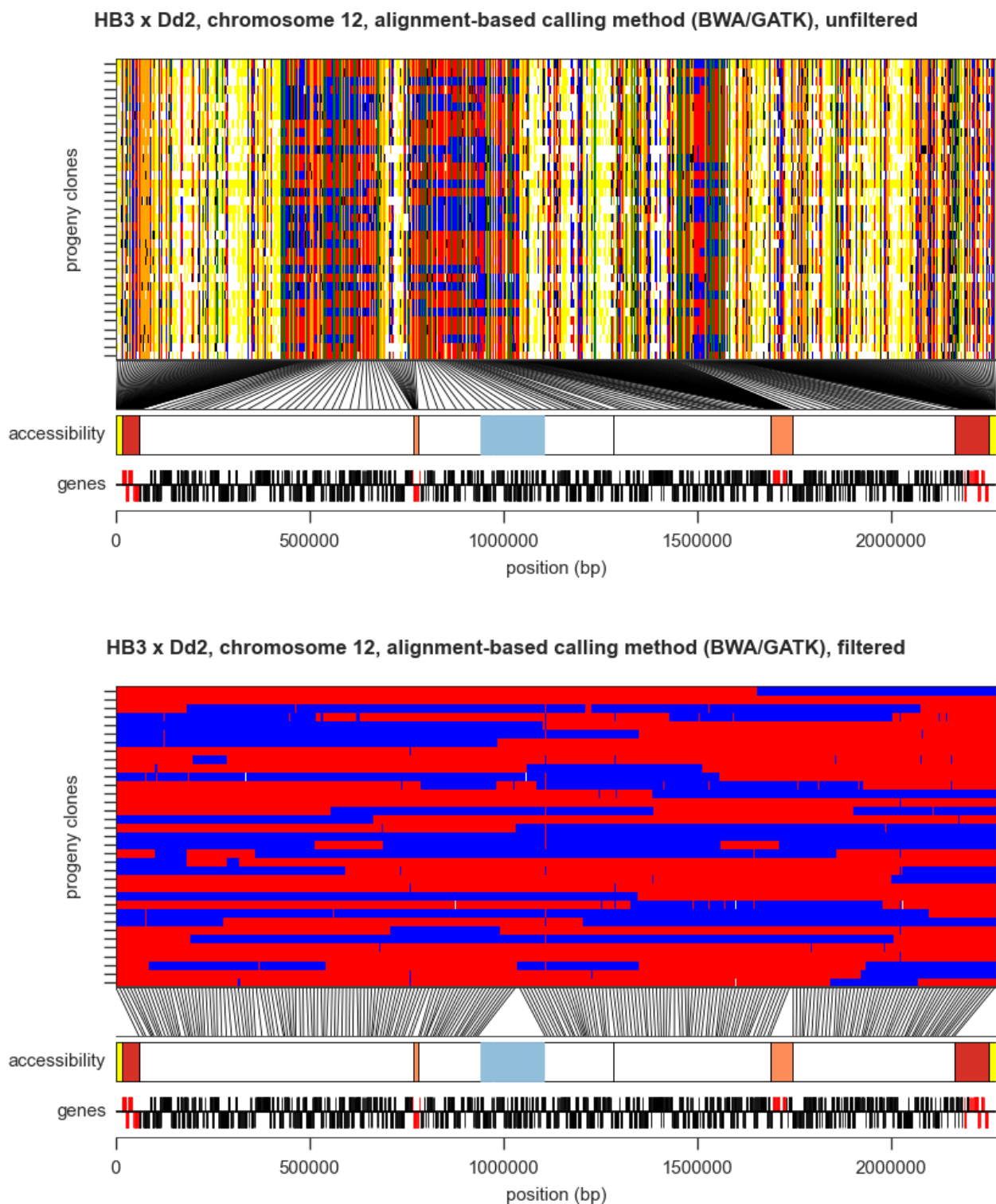
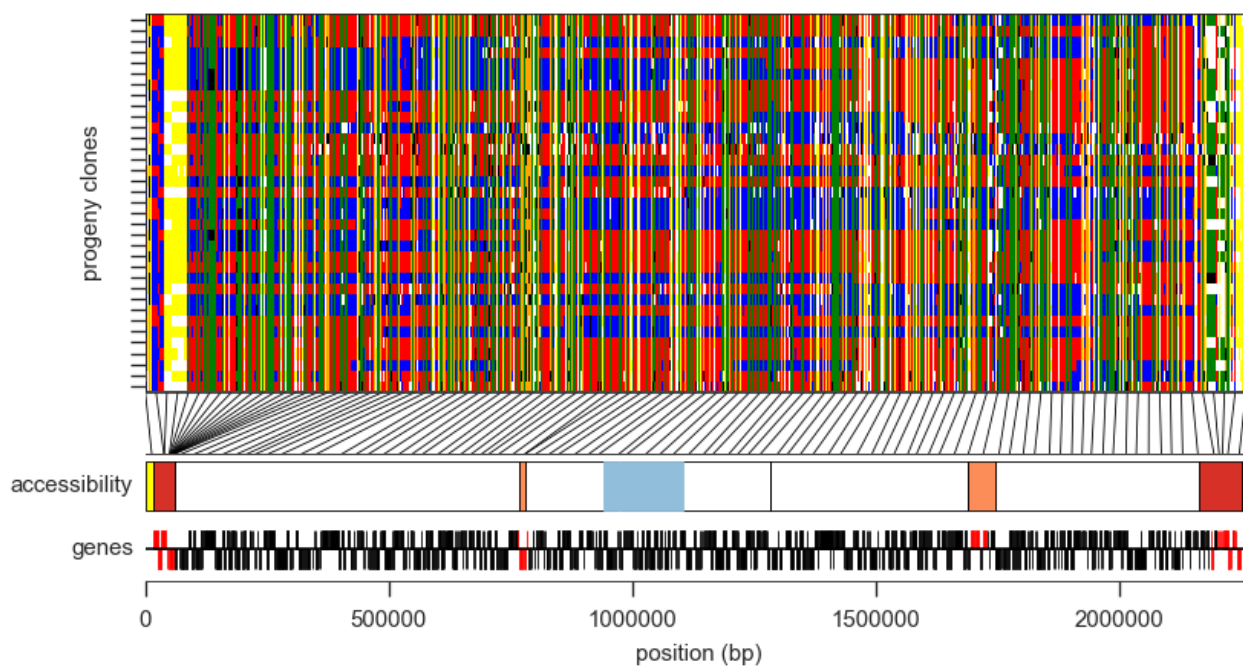


Figure S6: Illustration of the alignment-based callset before and after variant filtration. The upper plot shows the raw variant calls, the lower plot shows the filtered variant calls. The main subplot in each shows each sample as a row and each variant as a column, painting genotype calls as follow: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

HB3 x Dd2, chromosome 12, assembly-based calling method (Cortex), unfiltered



HB3 x Dd2, chromosome 12, assembly-based calling method (Cortex), filtered

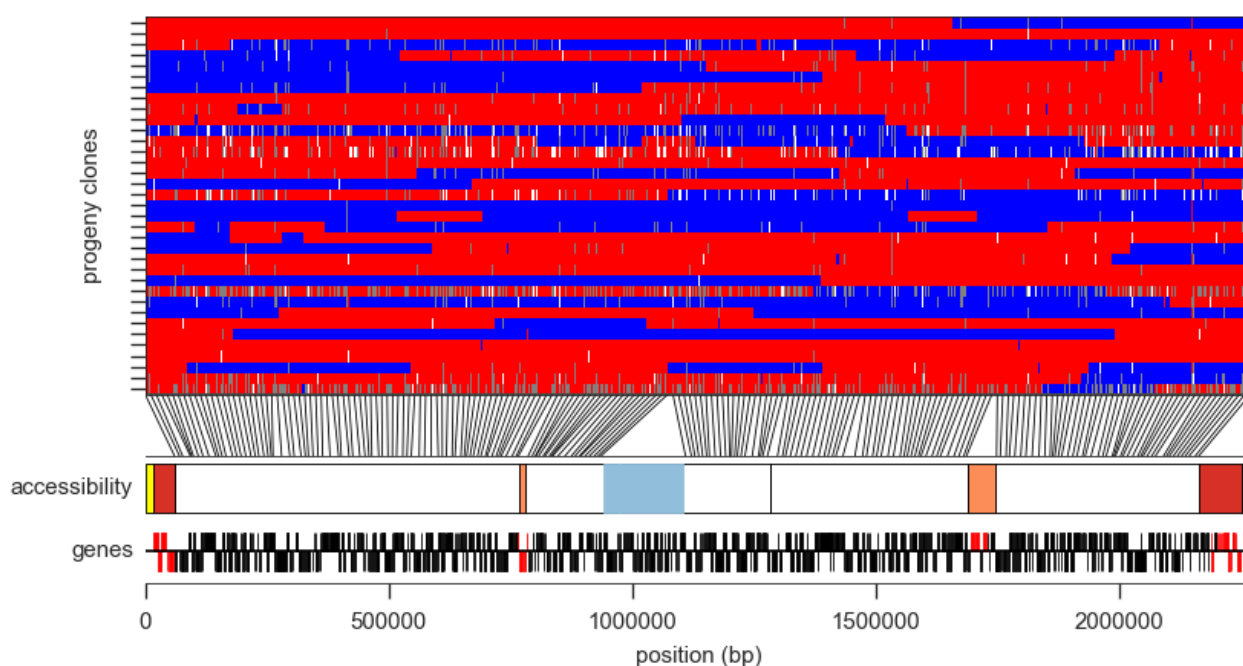
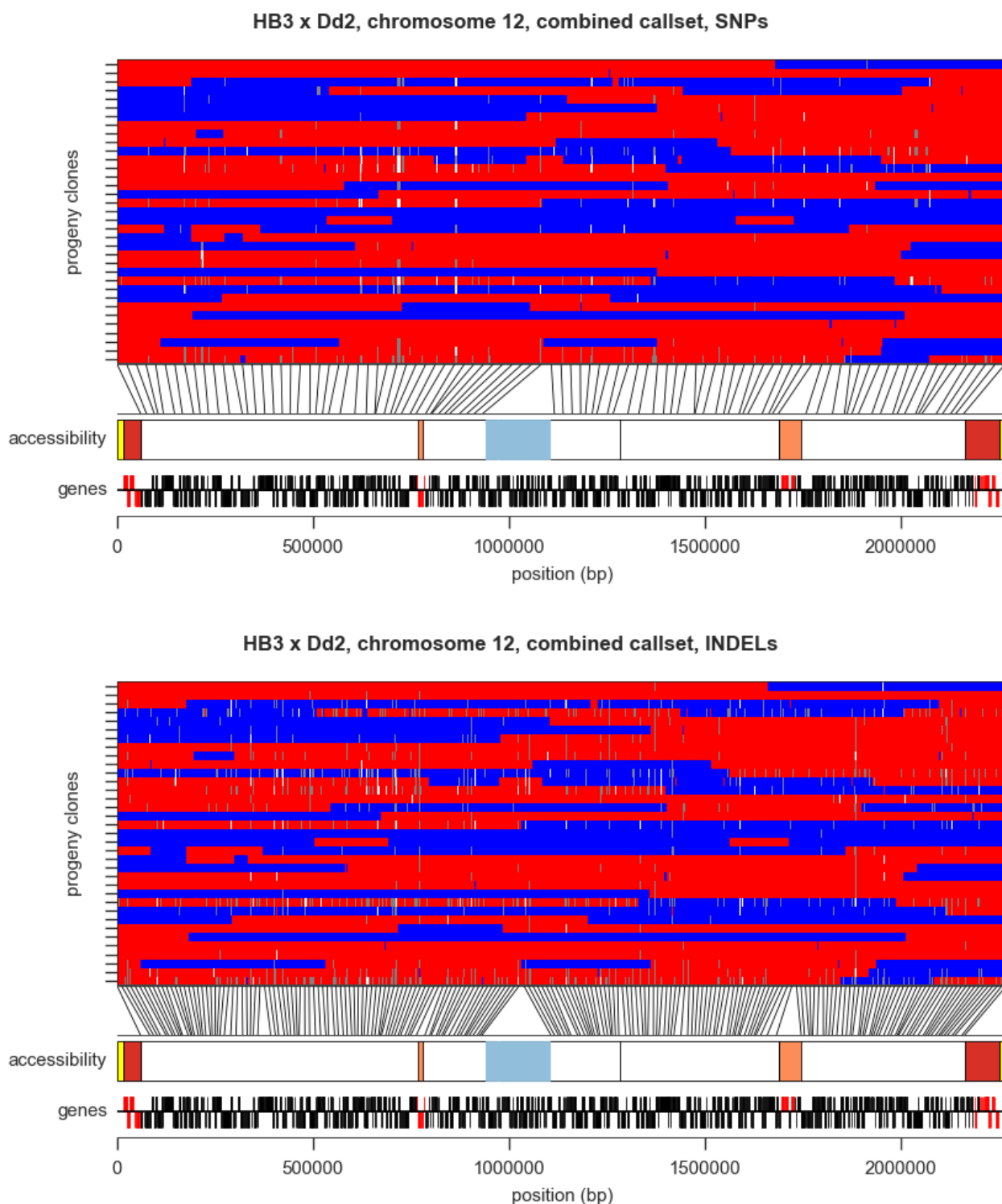


Figure S7: Illustration of the assembly-based callset before and after variant filtration. The upper plot shows the raw variant calls, the lower plot shows the filtered variant calls. The main subplot in each shows each sample as a row and each variant as a column, painting genotype calls as described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

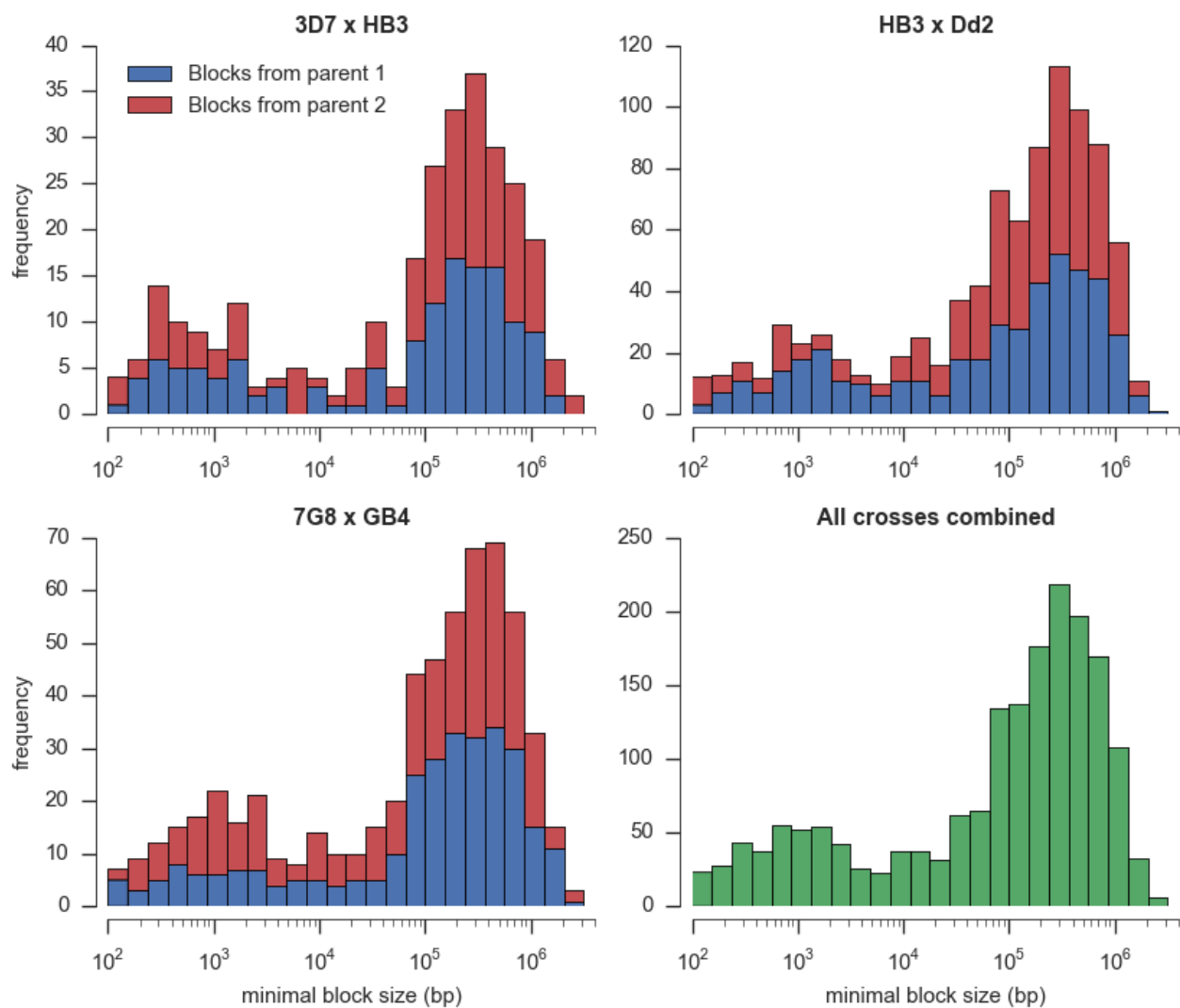


Figure S9: Size distribution for haplotype blocks transmitted from parents to progeny in the three crosses.



## 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/AANS000000000>
- 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., ... Wellem, 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
- Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, 6(4), 291–5. doi:10.1038/nmeth.1311
- 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
- Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., ... Kwiatkowski, D. P. (2012). Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, 487(7407), 375–9. doi:10.1038/nature11174
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, 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 / editorial board, Andreas D. Baxevanis ... [et al.]* (Vol. 11). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471250953

Youds, J. L., & Boulton, S. J. (2011). The choice in meiosis - defining the factors that influence crossover or non-crossover formation. *Journal of Cell Science*, 124(Pt 4), 501–13.  
doi:10.1242/jcs.074427