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 accessibility....1
3. Variant discovery and genotype calling....5
    3.1. Alignment-based calling method (BWA/GATK)....5
    3.2. Assembly-based calling method (Cortex)....5
4. Recombination analyses....5
    4.1. Determination of maximal block length for conversion tracts....5
5. References....5

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 accessibility

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 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.

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

- **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.

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.

3.3. Combined callset

@@TODO

4. Recombination analyses

4.1. Determination of maximal block length for conversion tracts

@@TODO

5. Figures and tables

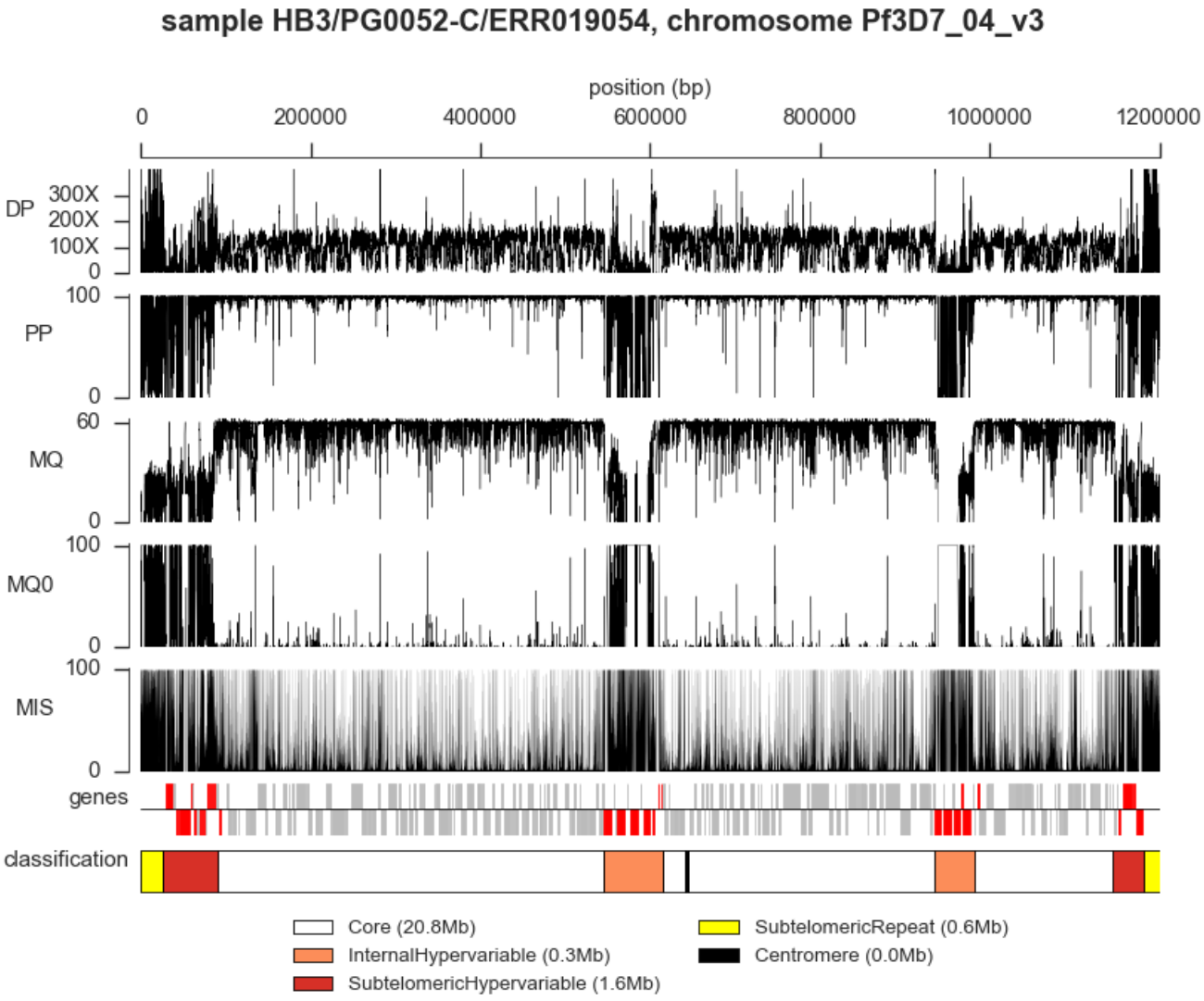

Figure S1: Example of alignment metrics for an individual sample and relationship to genome region classification. The sample shown is HB3/PG0052-C/ERR019054 (parent of 3D7xHB3) and data are shown for the entirety of chromosome 4. DP = total depth of coverage, PP = percent of reads aligned in a proper 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.

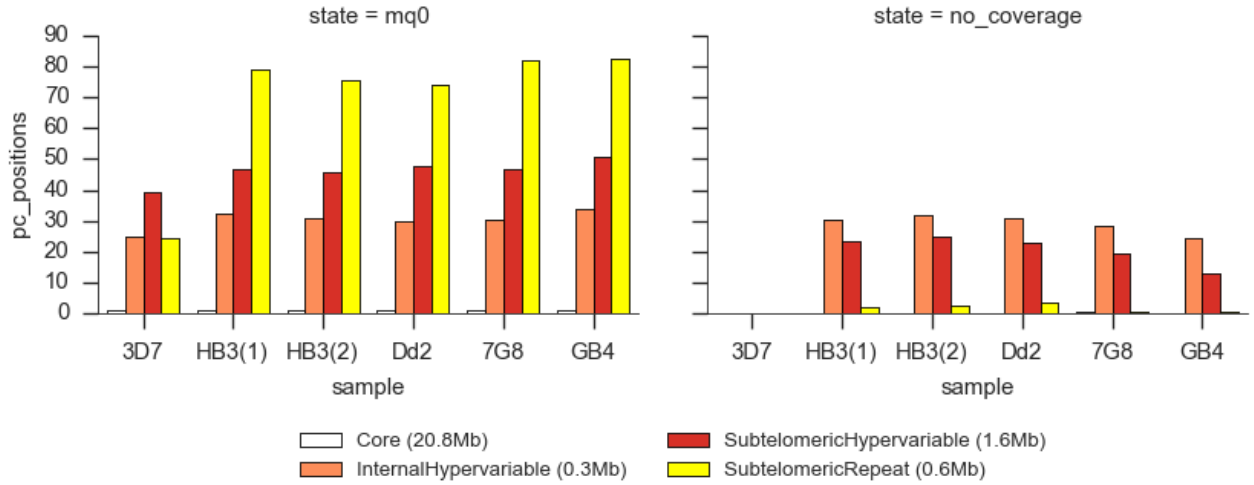

Figure S3: Summary of alignment characteristics for different genome region classes. The left-hand sub-plot shows the percentage of positions with more than 10% of reads aligned ambiguously (mapping quality zero). The right-hand sub-plot shows the percentage of positions without any coverage whatsoever.

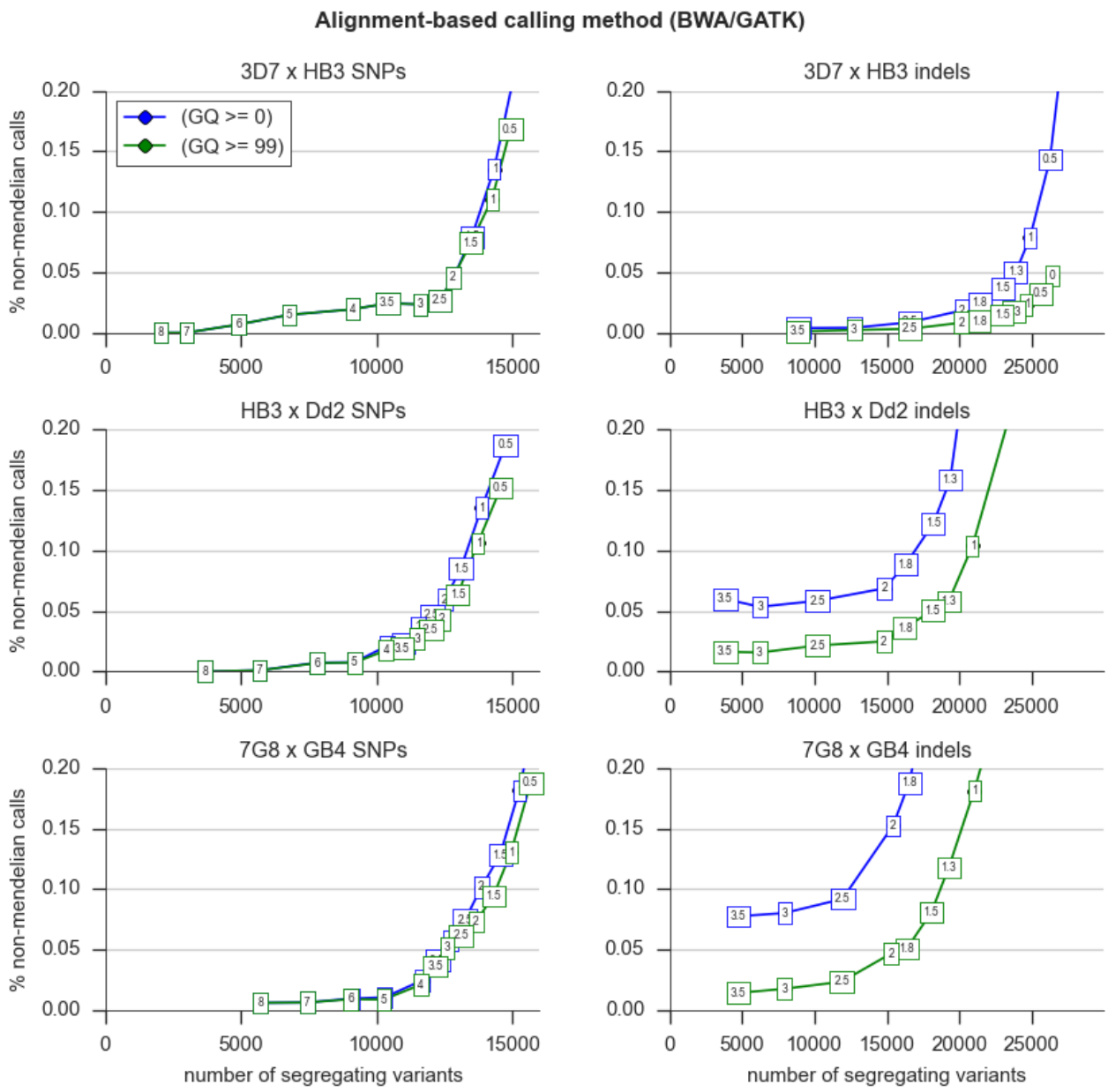

Figure S4: Using Mendelian error as a guide to filtering variants and genotype calls from the alignment-based calling method. Each point plotted corresponds to variants filtered according to a minimum value of the VQSLOD annotation and genotype calls filtered according to a minimum value of GQ. The VQSLOD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

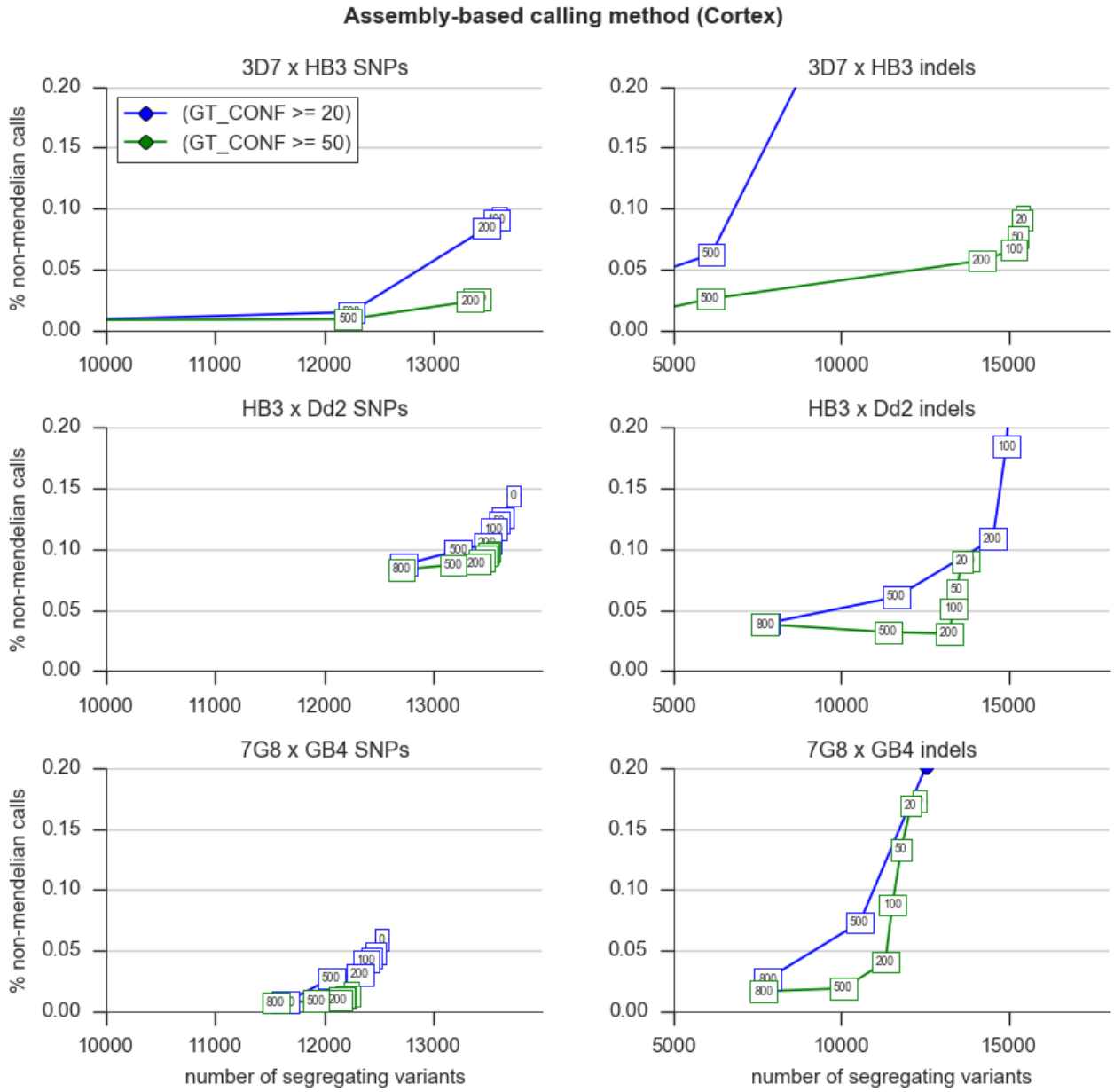

Figure S5: Using Mendelian error as a guide to filtering variants and genotype calls from the assembly-based calling method. Each point plotted corresponds to variants filtered according to a minimum value of the SITE_CONF annotation and genotype calls filtered according to a minimum value of GT_CONF.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

6. References

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