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

@@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. Here 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 as separate samples, with separate DNA library preparation and sequencing runs. In the remainder of this document the words “sample” and “sequencing run” can thus be considered equivalent, because 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.

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 non-default 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 larger 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], and included 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, 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 @@X and chromosome @@X is shown in Figure @@SX. The alignments themselves were also visualised using the LookSeq web application (Manske & Kwiatkowski, 2009), an example of which is shown in Figure @@SX.

From these visualisations a clear, qualitative distinction could be seen between regions of the genome with consistent coverage across all samples, and regions with significant alignment issues in one or more samples. 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 negligible proportion of reads aligned ambiguously.
- **Subtelomeric Hypervariable** – Gene-containing regions towards the subtelomere of a chromosome, with patchy and/or highly variable coverage in one or more samples and/or a high percentage of ambiguous alignments.
- **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 high percentage of ambiguous alignments.
- **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

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

alignments.

- **Centromere** – Centromere position as given in the genome annotation.

Within each chromosome we defined boundaries for these regions by eye from the visualisations described above. Figure @@SX shows a map of the genome regions defined, and Figure @@SX gives a summary of the alignment statistics for each parental clone within each 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.

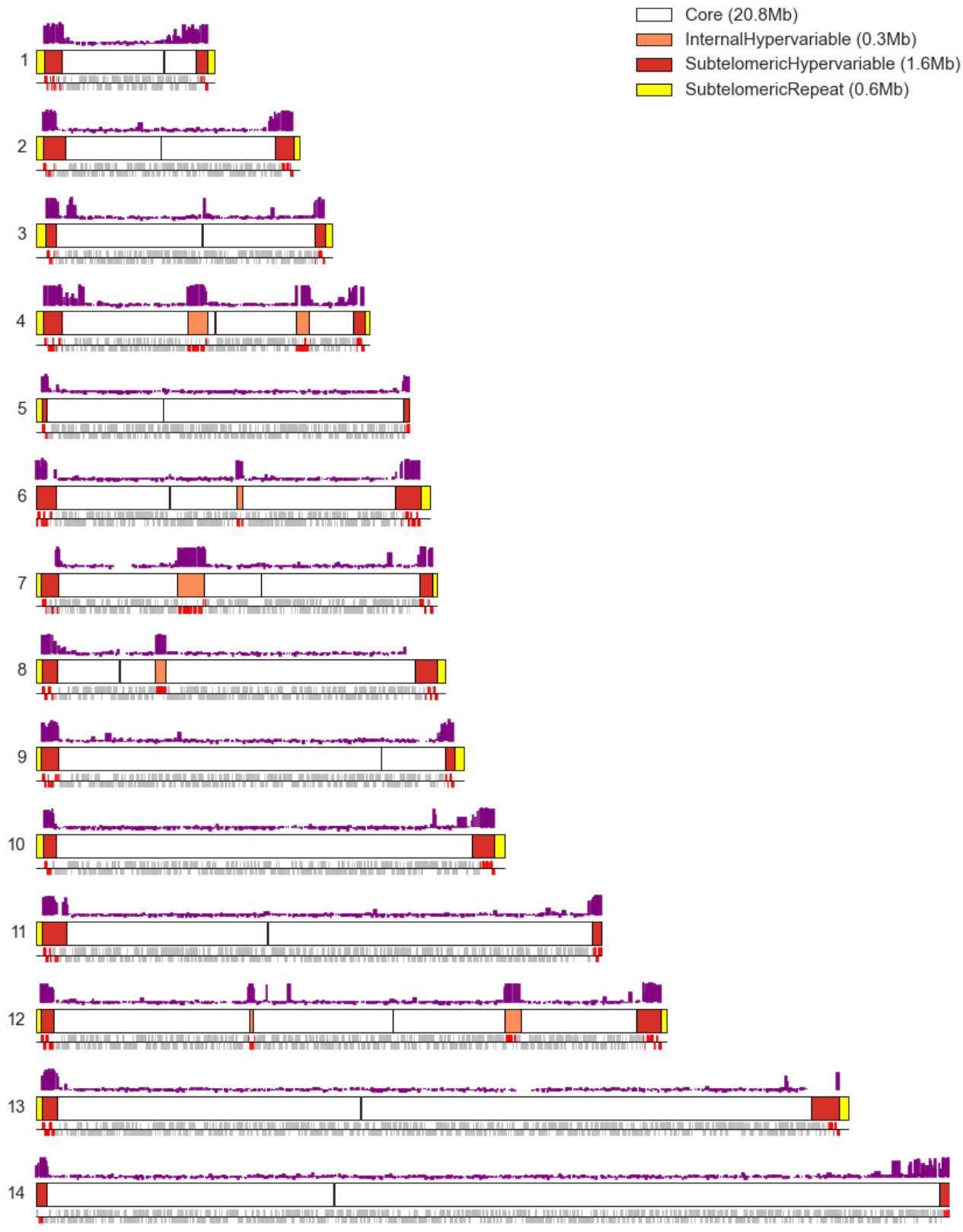
Our definition of the core genome is subjective and more sophisticated methods could be devised to partition the genome into regions with different alignment characteristics. However the contrast between these different regions of the genome is very striking and we believe the definitions given here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

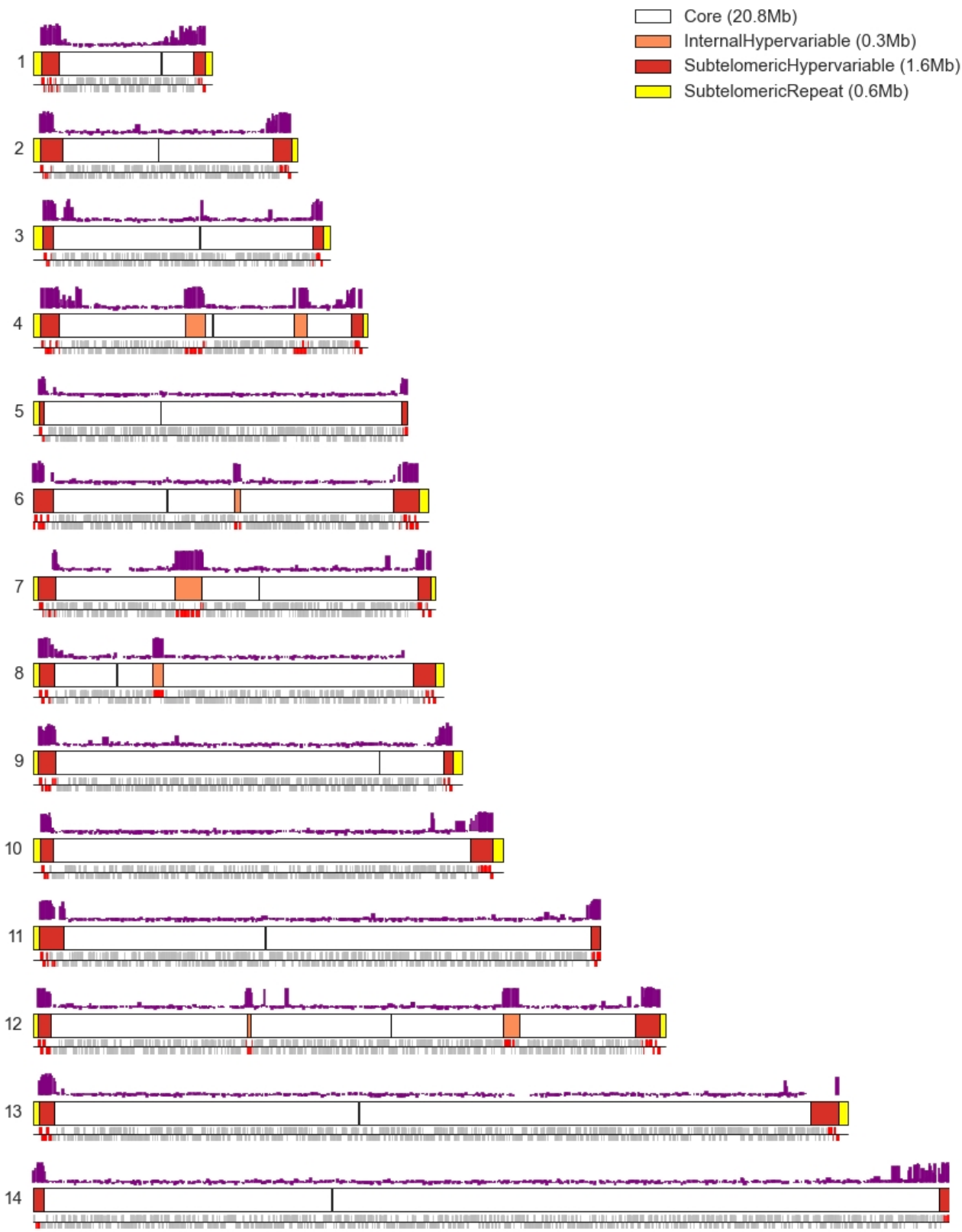
The genome region classification can be browsed alongside coverage, mapping quality and other metrics via the web application at the following URL:

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

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

@@TODO


Core (20.8Mb)
InternalHypervariable (0.3Mb)
SubtelomericHypervariable (1.6Mb)
SubtelomericRepeat (0.6Mb)
1
2
3
4
5
6
7
8
9
10
11
12
13
14



3. Variant discovery and genotype calling

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

@@TODO (Li & Durbin, 2009)

3.2. Assembly-based calling method (Cortex)

@@TODO

4. Recombination analyses

4.1. Determination of maximal block length for conversion tracts

@@TODO

5. References

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

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