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

# Figures and tables

**Table 1.** Summary of sequence and variation data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 3D7 x HB3 | HB3 x Dd2 | 7G8 x GB4 |
| No. progeny[[1]](#footnote-1) | 15 | 35 | 28 |
| Coverage[[2]](#footnote-2) | 108X (41-173) | 110X (22-637) | 119X (61-250) |
| No. SNPs[[3]](#footnote-3) | 15388 | 14885 | 14392 |
| No. INDELs[[4]](#footnote-4) | 26699 | 21576 | 20079 |
| Nucleotide diversity (kbp-1)[[5]](#footnote-5) | 0.5 (0.1-1.6) | 0.5 (0.1-1.4) | 0.5 (0.1-1.4) |
| Indel diversity (kbp-1)[[6]](#footnote-6) | 1.3 (0.5-2.1) | 1.0 (0.4-1.8) | 0.9 (0.4-1.7) |
| Marker spacing (bp)[[7]](#footnote-7) | 286 (2-1699) | 304 (2-2027) | 324 (2-2173) |

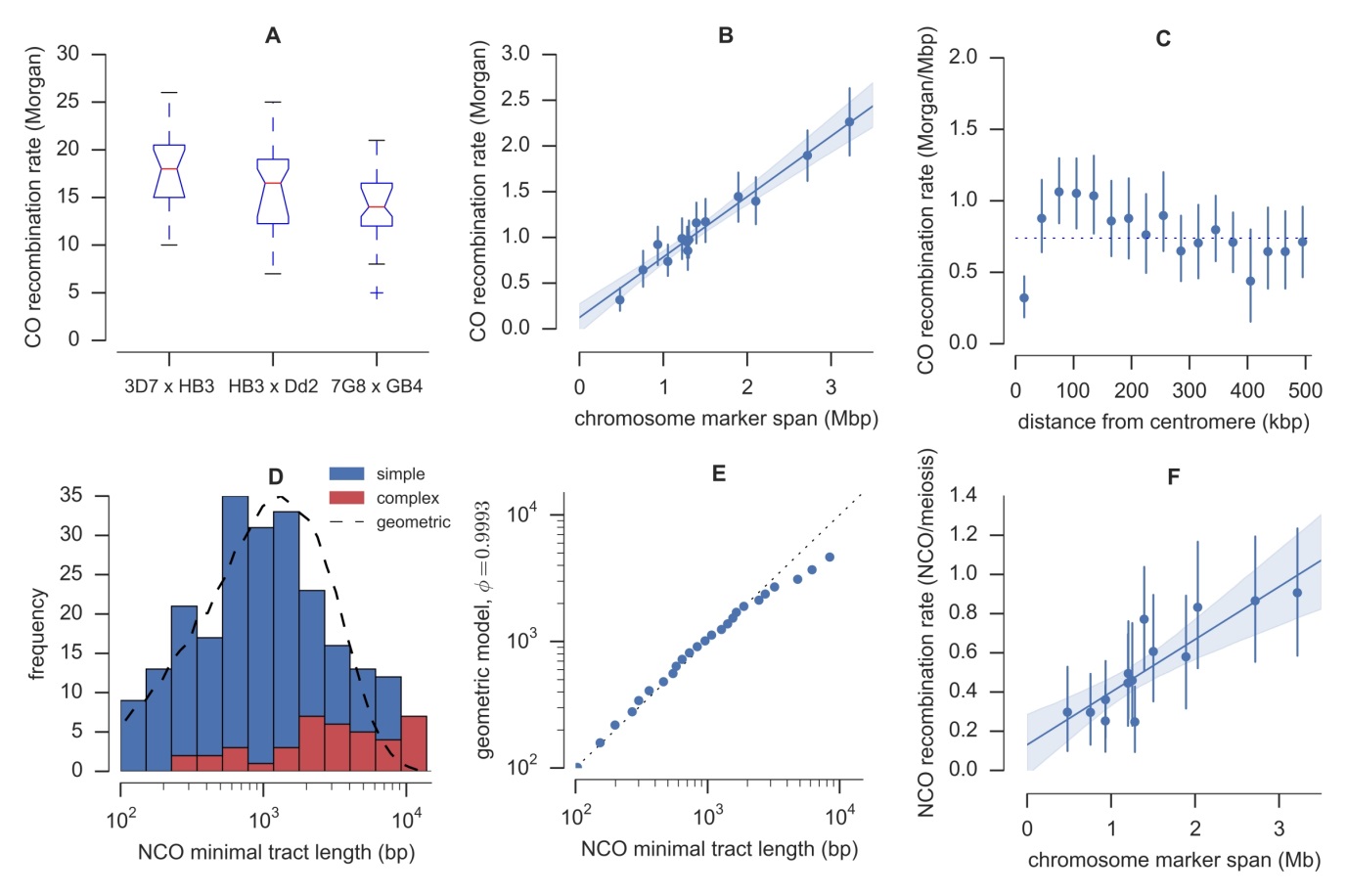
**Figure 1.** Properties of INDELs. **A**, INDEL size distribution (size > 0 are insertions, size < 0 are deletions). Solid black bars represent the frequency of INDELs that are expansions or contractions of short tandem repeats (STR), solid white bars represent the frequency of non-STR INDELs. Most coding INDELs are size multiples of 3, preserving the reading frame. Most non-coding INDELs are size-multiples of 2, reflecting the abundance of poly(AT) repeats in non-coding regions. **B**, amino acids inserted and deleted (relative to the 3D7 reference genome). **C**, INDEL diversity in intergenic regions relative to the position of predicted core promoters (@@REF). Each point represents the mean INDEL diversity in a 50bp window at a given distance from the centre of a core promoter. Vertical bars represent the 95% confidence interval from 1000 bootstraps. The dashed line is at the mean intergenic diversity for the given INDEL class (STR/non-STR).



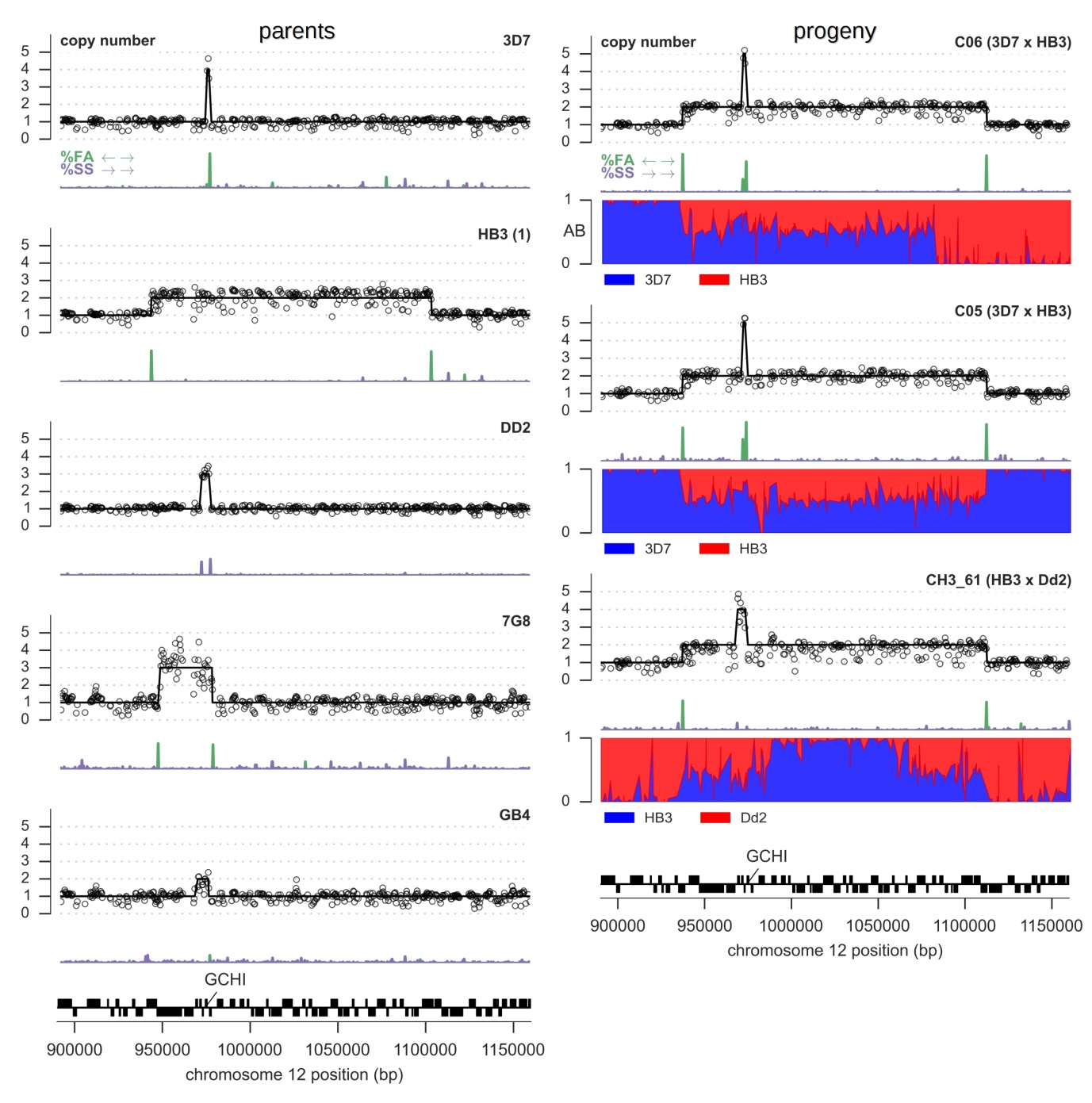
**Figure 2.** Nucleotide diversity and sites of complex variation in the core genome. Nucleotide diversity is shown for each cross in 500bp half-overlapping windows across the core genome (which excludes hypervariable regions containing *var*, *rif* or *stevor* genes) using SNPs discovered by either variant calling method and passing all quality filters. The peak of nucleotide diversity in chromosome 10 is expanded to show four distinct peaks due to the merozoite surface antigens *MSP3*, *MSP6*, *DBLMSP* and *DBLMSP2.* All labelled loci (with the exception of AMA1) are sites of complex variation where assembly of sequence reads is required to determine the non-reference alleles.



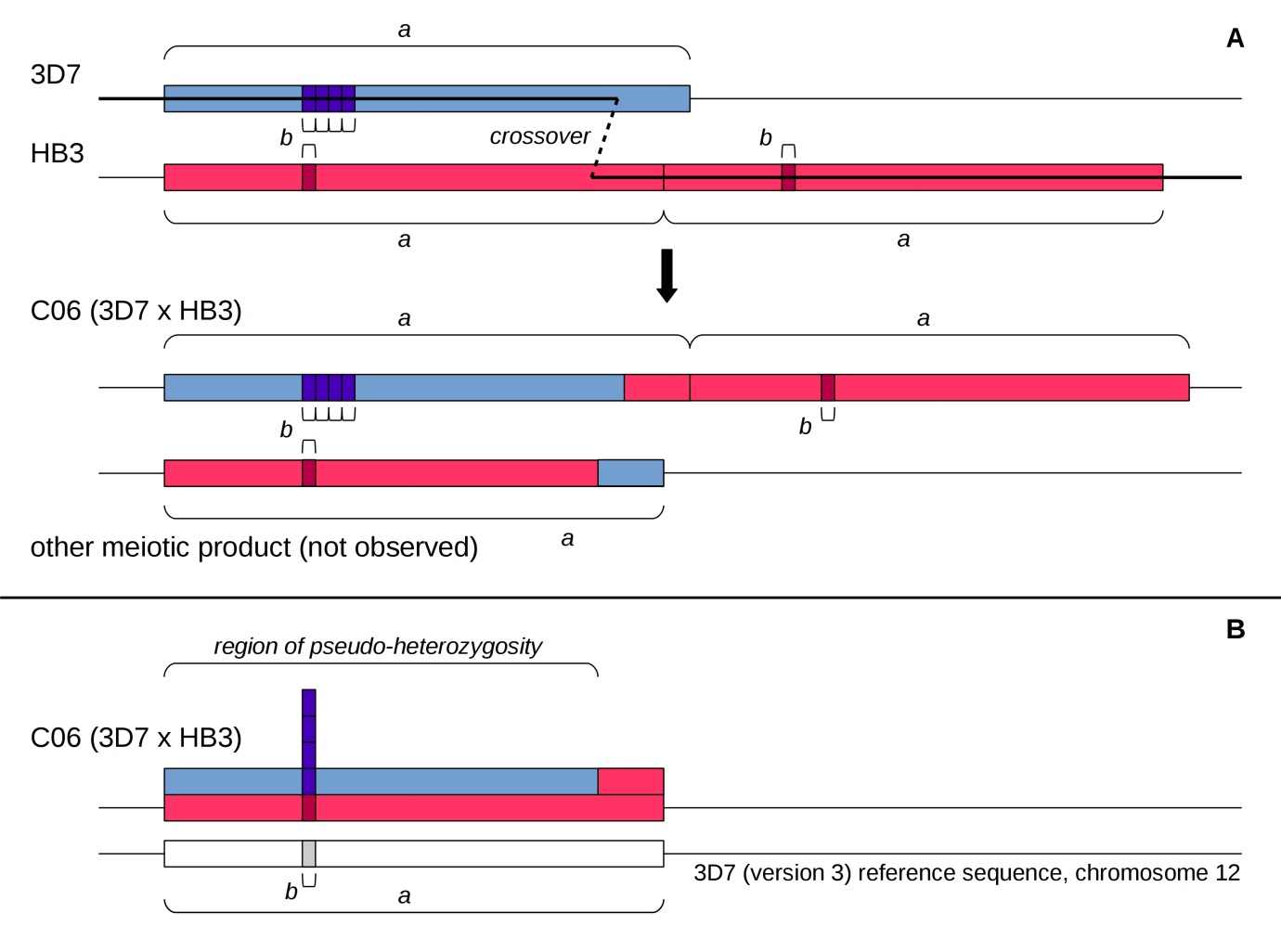
**Figure 3.** Crossover (CO) and non-crossover (NCO) recombination parameters. **A,** CO recombination rate by cross. **B,** CO recombination rate by chromosome. **C,** CO recombination rate relative to centromere position. Error bars show the 95% confidence interval from 1000 bootstraps. **D,** NCO tract length distribution. The dashed line shows the distribution of minimal tract lengths that would be observed with the available markers if NCO tract lengths follow a geometric distribution with parameter *phi* = 0.9993. **E,** Quantile-quantile plot of actual NCO minimal tract lengths versus the expected distribution of minimal tract lengths that would be observed with the given markers if NCO tract length is modelled as a geometric distribution with parameter *phi* = 0.9993. The data fit the model well except for an excess of tracts with minimal length greater than ~3kb. **F,** NCO recombination rate by chromosome, adjusted for incomplete discovery of NCO events.



**Figure 4.** Copy number variation and recombination spanning the anti-folate resistance gene *gch1* on chromosome 12. The left side of the plot shows data from the five parental clones, the right side of the plot shows data from three progeny clones where recombination events were observed within amplified regions. For each clone, the upper-most subplot shows the predicted copy number based on depth of coverage of aligned sequence reads. The circular markers show the actual depth of coverage within 300bp non-overlapping windows, excluding windows where GC content was below 20%, and normalised by dividing by the average depth over the core genome regions of chromosome 14. The solid line is the copy number state predicted by fitting a Gaussian hidden Markov model to the coverage data (see supplementary information). Beneath this is a subplot showing evidence for the arrangement of amplified segments, from the alignment orientation of read pairs. %FA: percentage of reads aligned facing away from each other, expected if amplified segments are arranged as a tandem array. %SS: percentage of reads aligned facing in the same direction, expected if amplified segments are arranged as a tandem inversion. For the progeny a third subplot is included showing the allele balance between the two parental alleles at segregating SNP and INDEL sites within the region. AB: fraction of aligned reads containing the first parent’s allele. Recombination events within the amplified region can lead to regions of pseudo-heterozygosity within a progeny clone, where one copy of the amplified region is inherited from each parent, indicated by regions where the allele balance is approximately 0.5.



**Figure 5.** Schematic illustrating how a single crossover recombination event can explain the patterns of copy number amplification and heterozygosity observed in progeny clone C06 from the 3D7 x HB3 cross at the anti-folate resistance locus *gch1* on chromosome 12. **A,** segment *a* is the region amplified in parent HB3 and present as 2 copies; segment *b* is the smaller region with four-fold amplification in parent 3D7. Segment *b* contains gene *gch1*. After fusion of gametes from the parental 3D7 and HB3 haploid clones, a crossover within segment *a* leads to one meiotic product (progeny clone C06) which retains the amplification of segment *a* but inherits a portion of segment *a* from each parent. The other meiotic product is not observed amongst the progeny isolated from the 3D7 x HB3 cross. In this case the crossover event occurred such that progeny C06 acquires two copies of segment *a* and a total of 5 copies of segment *b*, and the meiotic sister reverts to a single copy of segments *a* and *b*. **B,** alignment of the C06 genome to the 3D7 (version 3) reference sequence, which contains only a single copy of segments *a* and *b*. Within segment *a* 2-fold coverage is expected, within segment *b* 5-fold coverage is expected, and within a portion of segment *a* the progeny clone is effectively heterozygous, inheriting one copy from each parent. Similar schematics for the other two progeny clones (C05 and CH3\_61) with recombination events within the amplified region on chromosome 12 are included in supplementary information.



1. Number of independent recombinant progeny that were sequenced as part of this study and yielded usable sequence data. [↑](#footnote-ref-1)
2. Coverage for each sample was calculated as the mean depth of sequenced bases across the whole genome. Values shown are the median (minimum-maximum) values of sample coverage within a cross. [↑](#footnote-ref-2)
3. Total number of segregating SNPs combined from both alignment and assembly calling methods that passed all quality filters. [↑](#footnote-ref-3)
4. Total number of segregating indels combined from both alignment and assembly calling methods that passed all quality filters. [↑](#footnote-ref-4)
5. Nucleotide diversity (number of segregating SNPs per kilo base pair) calculated over 10kb non-overlapping windows within the core genome. Values shown are the median (5-95th percentile). [↑](#footnote-ref-5)
6. Indel diversity (number of segregating indels per kilo base pair) calculated over 10kb non-overlapping windows within the core genome. Values shown are the median (5-95th percentile). [↑](#footnote-ref-6)
7. Average distance between combined SNP and indel markers. Values shown are median (5-95th percentile). [↑](#footnote-ref-7)