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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 3D7xHB3 | HB3xDd2 | 7G8xGB4 |
| No. progeny[[1]](#footnote-1) | 15 | 35 | 28 |
| Coverage[[2]](#footnote-2) | 102X (41-173) | 110X (22-637) | 107X (55-250) |
| 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) |

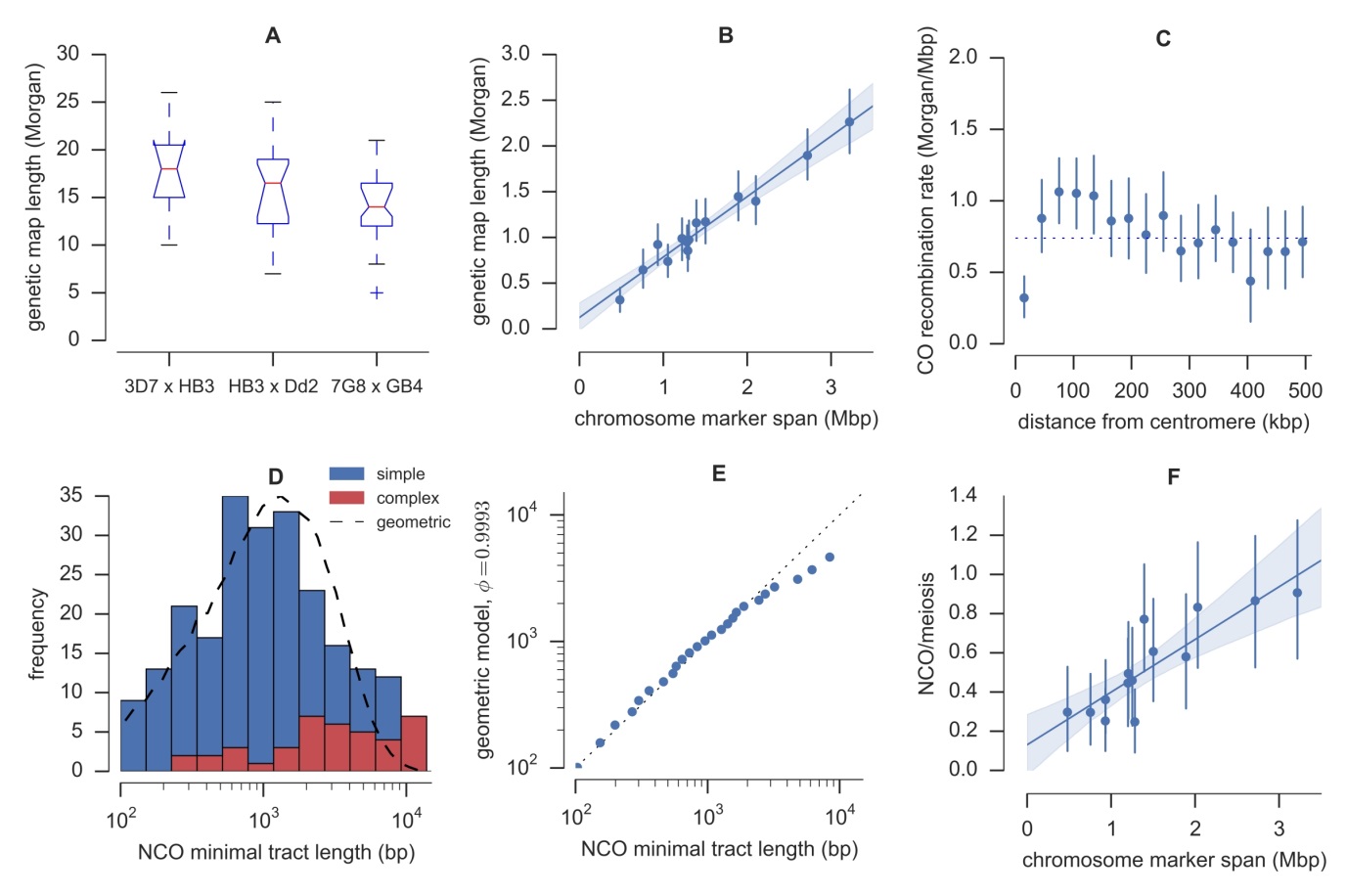
**Figure 1.** Properties of INDELs. **A**, INDEL size distribution (size > 0 are insertions, size < 0 are deletions). Solid black bars represent the frequency of INDELs that are expansions or contractions of short tandem repeats (STR), solid white bars represent the frequency of non-STR INDELs. Most 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 core promoters predicted by Brick et al. (2008). 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 genes *msp3*, *msp6*, *dblmsp* and *dblmsp2* encoding merozoite surface antigens*.* 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.



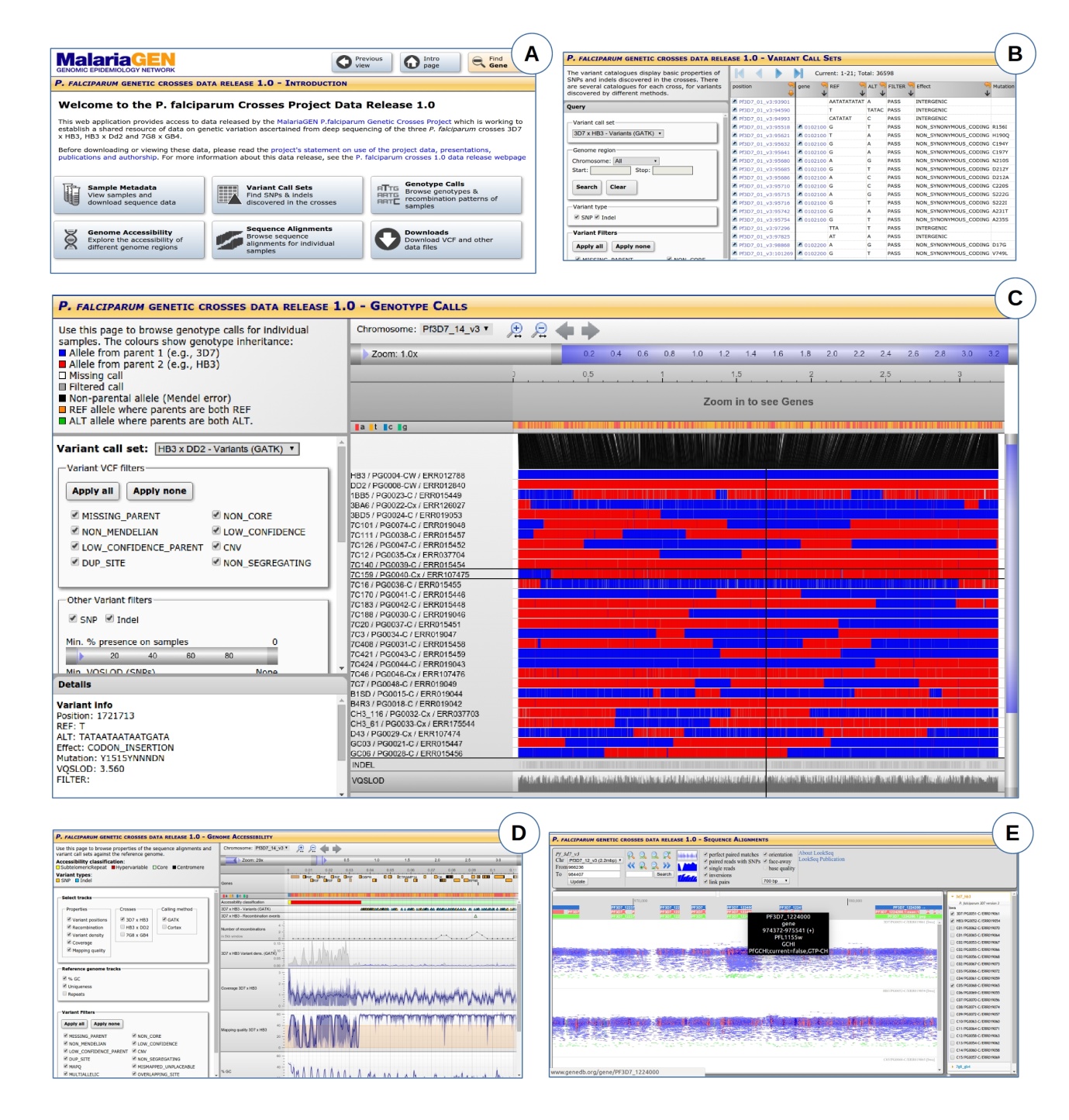
**Figure 3.** Crossover (CO) and non-crossover (NCO) recombination parameters. **A,** CO recombination rate by cross. **B,** CO recombination rate by chromosome. **C,** CO recombination rate relative to centromere position. Error bars show the 95% confidence interval from 1000 bootstraps. **D,** NCO 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. **A**, CNVs in the 3D7 and HB3(1) parental clones; *α* labels the segment amplified in HB3, *β* labels the segment amplified in 3D7. **B**, CNV and recombination in clone C06, progeny of 3D7xHB3. **C**, CNV and recombination in clone C05, progeny of 3D7xHB3. **D**, CNVs in the HB3(2) and Dd2 parental clones; *γ* labels the segment amplified in Dd2. **E**, CNV and recombination in clone CH3\_61, progeny of HB3xDd2. **F**, CNVs in the 7G8 and GB4 parental clones. CN = copy number, black 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; solid black line is the copy number state predicted by fitting a Gaussian hidden Markov model to the coverage data (Supplementary Information). FA = reads aligned facing away from each other (expected at boundaries of a tandem array), SS = reads aligned in the same orientation (expected at boundaries of a tandem inversion), scale is depth of coverage (i.e., 0-40X). AB = fraction of aligned reads containing the first parent’s allele; recombination events within an 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.** Screenshots from the web application at [www.malariagen.net/apps/pf-crosses/](http://www.malariagen.net/apps/pf-crosses/) providing access to sequence and variation data on the three crosses. **A**, introduction page, providing navigation to different tools for data exploration. **B**, browse and query data on variants (SNPs and INDELs) discovered in the crosses by different calling methods. **C**, browse genotype calls in parents and progeny and visualise patterns of allelic inheritance and recombination. **D**, genome browser, providing multi-resolution views of various data tracks including coverage and mapping quality. **E**, sequence alignment browser (LookSeq).



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)