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

Alistair Miles, Zamin Iqbal, Paul Vauterin, Richard Pearson, Susana Campino, Michel Theron, Kelda Gould, Daniel Mead, Eleanor Drury, John O’Brien, Valentin Ruano Rubio, Bronwyn MacInnis, Karen Hayton, Lisa Ranford-Cartwright, Michael Ferdig, Xinzhuan Su, Thomas Wellems, Julian Rayner, Gil McVean and Dominic Kwiatkowski

## Abstract

The human malaria parasite *Plasmodium falciparum* continues to be a major focus for public health research, as elimination efforts confront the spread of artemisinin resistance in South-East Asia. These efforts stand to gain from a deeper knowledge of genetic diversity and recombination in parasite populations, yet *P. falciparum* remains a difficult genome to study. To provide a foundation for ongoing research in these and other areas of *P. falciparum* genome biology, we describe a data resource comprising whole genome deep sequencing of haploid parents and progeny from three genetic crosses. We use the sequence data to construct a map of genome variation within each cross, spanning both coding and non-coding regions, and integrating SNP, INDEL and complex polymorphisms. We confirm that INDELs are exceptionally abundant, being more common than SNPs and thus the dominant mode of polymorphism within the core genome. This map of variation can serve as a robust reference point for studies of variation in natural populations, as well as a tool for fine-grained genotype-phenotype studies within the crosses themselves. We further illustrate the value of these resources by exploring open questions regarding meiotic recombination, including the relative contribution of crossover and non-crossover events, and the factors influencing variation in recombination rate over the genome. Finally, we study a peculiar phenomenon within these crosses whereby several progeny have undergone meiotic recombination within regions of copy number amplification associated with drug resistance, and discuss the implications for the potential of this species to undergo bursts of evolution within genome regions under strong selection.

## Introduction

Genome variation in the eukaryotic pathogen *Plasmodium falciparum* underpins both fundamental biology, such as the ability of the parasite to evade the human immune response, and clinical outcomes, through the evolution of antimalarial drug resistance. This is of particular concern with the recent spread of resistance to front-line therapies in South-East Asia (@@REF). High throughput sequencing is a proven technology for the study of genome variation in *P. falciparum*, and has yielded insights into natural patterns of variation and population structure (@@REFs), the mutational processes that generate diversity (@@REF) and the genetic basis for artemisinin resistance (@@REF). However, knowledge of genome variation is still incomplete, in part because several factors make the *P. falciparum* genome difficult to study in its entirety. The highly compact @@2XMb genome has an extremely biased nucleotide composition, with 80.X% (A+T) content overall and over @@X% (A+T) in non-coding regions (@@REF). Despite being free of transposable elements, the genome is highly repetitive, with short tandem repeats and other low complexity sequences unusually abundant in both coding and non-coding regions (@@REF). Several multi-copy gene families encoding erythrocyte surface antigens are hypervariable, diversifying within the course of a single infection via ectopic recombination (@@REF). Other genes associated with the blood-stage exhibit deep and possibly ancient allelic dimorphisms under balancing selection where one or more allele families are be highly diverged from the reference sequence (@@REF). These factors challenge both the sequencing technology itself and methods for variant discovery and genotyping from high throughput sequence data. Data on insertion/deletion (INDEL) variation are limited (@@REF) although several lines of evidence suggest that small INDELs may be unusually abundant (@@REFs). Little is known about variation in non-coding regions, which could have a significant impact on phenotype by regulating gene expression. Knowledge of complex variation, where haplotypes are highly diverged from the reference genome, is constrained to a few well-studied loci (e.g., @@REFs).

In studies of genome variation in humans and other species (@@TODO species examples) trios, crosses and other types of pedigree have proved an invaluable resource. Pedigree information can be used to identify errors in the process of high throughput sequencing of genomic DNA and variant discovery. By sequencing related individuals, variants that are in violation of Mendelian segregation can be identified. Although such variants may occasionally arise naturally, for example through *de novo* mutation, in general they provide a good indicator of genotyping error (@@REF). Mendelian errors have therefore been widely used to calibrate and validate methods for variant discovery and genotype calling. Unfortunately, although pedigrees can be readily obtained for most species, this is currently not the case for *P. falciparum*. Although sexually reproducing, the parasite is haploid for the majority of its life cycle, and exists only briefly as a diploid when taken up into the mosquito mid-gut, where it undergoes meiosis (@@REF Walliker). Crossing *P. falciparum* parasites currently involves a laborious and difficult process of mixing clones that are fed to mosquitoes, which are then used to generate blood stage parasites by biting non-human primates. As a result, only three experimental crosses have been published to date, where they have led to key discoveries regarding the causes of drug resistance (@@REF) and host specificity (@@REF). However, although only three crosses have been performed, each cross yields a large number of progeny clones, each the result of an independent meiosis, and therefore provides a much higher power to observe Mendelian errors than smaller pedigrees or trios. These three crosses are therefore a precious resource for studying genome variation in *Plasmodium falciparum*, because they represent the only controlled Mendelian system within which methods for variant discovery can be calibrated to produce robust results.

Here we describe a data resource comprising whole genome deep sequencing of parents and progeny of three experimental *Plasmodium falciparum* genetic crosses, involving the parental clones 3D7, HB3, Dd2, 7G8 and GB4, representing three distinct continental origins. We use a combination of methods for variant discovery, leveraging both alignment of sequence reads to the 3D7 reference genome and reference-free sequence assembly, to build a map of genome variation within each cross incorporating SNP, INDEL and complex polymorphisms. All of the variants included in the final call sets are highly consistent with Mendelian inheritance and show almost perfect genotype concordance between biological replicates, and hence could serve as a positive training set to underpin variant discovery in other studies, particularly those surveying variation in natural populations. These data also reveal some interesting features of genome variation between the parental clones, including an exceptionally high abundance of INDELs relative to SNPs, and evidence for the non-neutral impact of INDELs within both coding and non-coding regions. Because these data represent a valuable reference resource for both research and education, we describe a novel web application providing a means for exploring and interacting with these data in an intuitive way.

We also take this opportunity to explore some open questions regarding meiotic recombination in *P. falciparum*. Previous studies have estimated the crossover recombination rate (@@REFs), provided evidence that non-crossover events occur (@@REFs), demonstrated that at least some recombination events occur within coding regions (@@REFs) and suggested that recombination events are not uniformly distributed over the genome (@@REF). Here we combine SNP and INDEL markers to obtain a resolution of ~300bp within each cross, which is sufficient to differentiate between crossovers and non-crossovers, and estimate rates for both types of recombination event. This resolution is also sufficient to resolve the location of most recombination events relative to gene and exon boundaries and study spatial variation in recombination rates. Finally we investigate recombination in the context of two large regions of copy number amplification, both of which segregate in the crosses and are associated with drug resistance (@@REFs). We find evidence for crossover recombination within these amplifications leading to regions of pseudo-heterozygosity within progeny clones. This has a number of implications for evolutionary biology, because it demonstrates a mechanism whereby fitness costs associated with a drug resistance mutation could be compensated, and different copies of an amplified gene could diverge more rapidly than if limited by mutation alone.

## Results

### Whole genome sequencing and genome accessibility

Whole genomes a parent and progeny clones from the crosses 3D7×HB3 (@@REF), HB3×Dd2 (@@REF) and 7G8×GB4 (@@REF) were sequenced using Illumina high throughput technology, with the majority of samples obtaining an average depth above 100×. All DNA libraries were derived from haploid parasite clones in culture, and sufficient DNA was available to use PCR-free library preparation throughout, which has been shown to reduce some of the biases associated with the AT-rich *P. falciparum* genome and hence improve the evenness of coverage across both coding and non-coding regions (@@REF). The clone HB3 is a parent in two crosses, however because DNA samples were obtained from different sources and had different culturing histories the two HB3 clones were sequenced and genotyped separately, and are here labelled HB3(1) and HB3(2) for crosses 3D7 x HB3 and HB3 x Dd2 respectively. The progeny clones from all crosses had previously been isolated via limiting dilution and shown to each be derived from independent meioses (@@REF). For several progeny biological replicates were obtained, where multiple libraries were created from DNA extracted from the same clonal culture at different points in time. These were also sequenced and genotyped separately, to enable analysis of concordance between replicates. All sequence data has been deposited at the European Nucleotide Archive and a mapping from clone identifiers to ENA accessions can be obtained via the web application at <http://www.malariagen.net/apps/pf-crosses> where sequence alignments can also be browsed interactively.

Sequence reads from all samples were aligned to the 3D7 reference genome and various metrics were calculated per genome position including depth of coverage and average mapping quality. Visual examination of these data revealed an obvious, qualitative difference between a **core genome** (20.8Mb) comprising regions of near-complete coverage and unambiguous alignments in all samples; **hypervariable regions** (1.9Mb) where accessibility is severely affected by both extensive paralogy and extreme divergence of parental clones other than 3D7 from the reference genome; and **subtelomeric repeat regions** (0.6Mb) where accessibility is limited by repetitive sequence (supplementary figure @@N; a BED file defining the boundaries of these regions is provided as supplementary file @@N). Hypervariable regions contained all genes in the antigenic *var* family, which are known to undergo frequent non-allelic recombination during mitosis (@@REF), and almost all genes in the associated *rif* and *stevor* families. Hypervariable regions also corresponded closely with regions of heterochromatin (@@REF). All samples exhibited some degree of bias such that coverage was lower where (A+T) content was above 80% (supplementary figure @@N), however the high depth of sequencing meant that coverage was sufficient to support variant calling in both coding and non-coding regions across the entire core genome. Because of the poor accessibility of hypervariable and subtelomeric repeat regions we excluded them from further study and limit ourselves to the core genome for the remainder of this paper. However, raw variant calls and other data for hypervariable regions can be obtained via the accompanying web application, and may be of use to others studying these important regions.

### SNP, INDEL and complex variation within the core genome

SNPs, small INDELs and regions of complex polymorphism were discovered and genotyped within each cross by two independent methods, one based on alignment of sequence reads to the 3D7 reference genome (@@REFs), the other based on partial assembly of sequence reads and comparison of assembled contigs (@@REF). Variants where genotype calls in one or more progeny clones were inconsistent with Mendelian segregation (Mendelian errors) were used to calibrate the outputs of both calling methods and decide variant filtering strategies (see methods and supplementary information). After variants were filtered, both methods achieved near-perfect concordance between biological replicates for both SNPs and INDELs, and concordance with previous array-based SNP genotyping platforms was high (supplementary information). The patterns of inheritance of parental alleles and recombination within each cross were also nearly identical when comparing SNPs with INDELs, or comparing results of the two variant calling methods. Many variants were discovered by both methods, however there were some notable differences. The assembly method was able to access variation in regions where alignment failed because the sample was too diverged from the reference (see below). On the other hand, the alignment method was able to access more variation in non-coding regions, which typically suffer from both lower coverage and lower sequence complexity. A deeper comparison of these and other methods for variant calling in *P. falciparum* using the crosses and other data is being carried out as part of a separate project (@@REF Pf3k) and will be reported elsewhere. To provide the greatest possible resolution for the present study, filtered variants called by each method were combined into a single call set for each cross (see methods). All SNP and INDEL calls can be downloaded from a public FTP site (@@REF) or browsed via the web application at <http://www.malariagen.net/apps/pf-crosses>.

#### INDELs are the most abundant form of polymorphism

Analysis of the combined variant call sets revealed that, within the core genome, INDELs were more abundant than SNPs in all three crosses (table @@N). Overall 83% of INDELs were found in non-coding regions, where INDELs were 3 times more abundant than SNPs. INDELs were also relatively abundant in coding regions, with the ratio of SNPs to INDELs being approximately @@N:M. This level of INDEL diversity is exceptionally high when compared with other species for which data are available, for example, @@TODO comparison SNP:INDEL rates in other species@@. The vast majority of INDELs found in the crosses were expansions or contractions of short tandem repeats (STRs), i.e., microsatellites (figure 1A). In non-coding regions 83% of INDELs were STR length variations, of which 71% were variations within poly(AT) repeats. In coding regions 77% of INDELs were STR variations, of which the majority were within poly(asparagine) or poly(isoleucine) tracts. Tandem repeat sequences are prone to slipped strand mis-pairing during DNA replication (@@REF) and are known to be associated with high rates of INDEL mutation in other species (@@REF). Poly(AT) repeats are very common in the non-coding regions of the *P. falciparum* genome, and poly(asparagine) and poly(isoleucine) repeats are unusually abundant within the exome (@@REF), hence the high INDEL diversity overall may be accounted for by the abundance of STRs in the genome, coupled with the high mutability of tandem repeats due to replication slippage.

Regarding the phenotypic consequences of INDEL variation, frame shift mutations within coding regions are expected to have severe consequences and hence be negatively selected. Consistent with this expectation, 94% of coding INDELs were found to be size multiples of 3 and hence preserved the reading frame (figure 1A). Within non-coding regions, the potential consequences of INDEL mutations are harder to predict. Relatively little is known about the transcription machinery in *P. falciparum*, however, @@REF predicted the location of core promoters upstream of @@N genes based on a training set of known transcription start sites (@@REF). Intergenic INDEL diversity in the crosses was found to display a specific architecture relative to the central positions of these predicted promoters, with an excess of non-STR INDELs within the first 50bp upstream of the promoter centre, and a deficit of STR INDELs 100-200bp upstream and 0-100bp downstream of the promoter centre (figure 1C; supplementary figure @@TODO). Further experimental work will be required to determine whether these INDELs have any impact on promoter activity.

#### Low nucleotide diversity is punctuated by complex variation in merozoite-stage genes

Average nucleotide diversity across the core genome was similarly low (5x10-4) in all three crosses (table 1). However, this low diversity was punctuated by a handful of loci with highly diverged alleles, where local diversity over a region up to 2kb was between 2 and 3 orders of magnitude greater (figure 2). These divergent loci were found almost exclusively within coding regions of genes associated with the merozoite life cycle stage, where the parasite is briefly exposed to the host immune system before invading another erythrocyte, and include several well-studied merozoite surface antigens. The most extreme example was MSP1 (@@ID), a highly expressed protein located at the merozoite surface (@@REF), where several regions of the gene are known to exhibit deep allelic dimorphism (@@REF). The complex variation at these loci could not be accessed by the alignment method, because sequences were two diverged from the reference genome, and hence coverage was locally patchy or non-existent (supplementary figure @@N). However, the assembly method was able to construct complete and correct sequences for all parents and progeny in the divergent regions of MSP1 (@@TODO clarify which parts) as verified by comparison with capillary sequence data (@@Genbank, @@REFs). Other genes where peaks of diversity were found and alleles could be assembled include merozoite surface proteins in the MSP3 family (MSP3, MSP6, DBLMSP, DBLMSP2), members of the *surfin* family (SURF1.2, SURF4.1, SURF4.2, SURF8.2, SURF13.1, SURF14.1) and PF3D7\_0113800 (a DBL-containing protein with unknown function on chromosome 1). A notable exception to the pattern of merozoite expression is PF3D7\_0104100 which is transcribed by the sporozoite specifically within the mosquito salivary gland (@@REF) suggesting involvement in the early stages of infection. Several of these genes are blood-stage vaccine candidates and/or are being actively studied for their role in erythrocyte invasion, and comprehensive knowledge of variation at these loci is essential for the design of effective vaccines and invasion assays. These data show that complex variation in a number of key *P. falciparum* genes can be ascertained in a high throughput manner via assembly of short sequence reads, thus it should be possible to use the same method to carry out large scale surveys of variation at these loci in clinical isolates from different geographical regions.

### Meiotic crossover and non-crossover recombination

As mentioned earlier, these crosses are currently the only available experimental system for *P. falciparum* where meiotic recombination can be directly observed. No previous study has had comparable data for all three crosses, or had the resolution available from deep whole genome sequencing, making this study uniquely able to derive definitive answers regarding the extent and features of meiotic recombination inthis species. For each cross, SNP and INDEL variants combined from both calling methods were used as a set of segregating markers for analyses of meiotic recombination (table 1). The average distance between markers was @@-@@bp, approximately an order of magnitude greater resolution than that available previously (@@REF; @@CHECK). In eukaryotes, programmed double strand breaks (DSBs) during meiosis are resolved by either crossover (CO) or non-crossover (NCO) between homologous chromosomes (@@REFs). A CO is a reciprocal exchange accompanied by a conversion tract, whereas an NCO is a conversion tract without reciprocal exchange. An algorithm was used to infer CO and NCO events from the size and arrangement of parental haplotype blocks found in the progeny, and to identify both simple and complex conversion tracts (see methods). Because occasional genotyping errors could also manifest as short haplotype blocks, all putative conversion tracts supported by only a single marker or with a minimal length less than 100bp were excluded. This yielded a total of 1194 COs, 230 NCOs and 331 conversion tracts for further analyses.

#### Gene coding regions are warm-spots and centromeres are cold-spots of CO recombination

Combining CO events from all three crosses, the total map length of the core genome was 15.7 Morgan (95% confidence interval: 14.8-16.6). The total marker span was 21.16Mb giving an average CO recombination rate of 13.5 kb/cM (95% confidence interval: 12.7-14.3). The CO recombination rate varied between crosses, with 3D7 x HB3 highest (17.7 Morgan) and 7G8 x GB4 lowest (14.3 Morgan) although this difference was marginally significant (P=0.06, Kruskal-Wallis H-test) (figure 3A). There was a strong linear correlation between chromosome size and CO recombination rate, with a rate of 0.55 predicted for the smallest chromosome (figure 3B) consistent with 0.5 expected if the recombination rate is calibrated to produce at least one CO per chromosome pair (@@REF).

The centromeres were cold-spots of CO recombination, as expected from studies in other eukaryotes, although the effect was highly localised (figure 3C). Within ~30kb of the centromere the CO rate was significantly lower, however between ~80-120kb from the centromere the rate was slightly higher than average. Due to the high marker density, in many cases it was possible to resolve the location of CO events relative to individual gene and exon boundaries. Of the 1194 CO events, 396 (33%) were observed with both flanking markers within a gene, 162 (13%) were within an intergenic region, and 636 (53%) were ambiguous (flanking markers spanned a gene boundary). The number of CO events observed within a gene was significantly higher than expected if CO events were distributed uniformly over the genome (P=0.001 by Monte Carlo simulation). Of the 396 COs observed within a gene, 298 (75%) occurred within an exon, 3 (1%) were within an intron, and 95 (24%) spanned an exon boundary. The number of COs observed within exons was also significantly higher than expected if COs occurred uniformly within genes (P<0.001 by Monte Carlo simulation). Thus a substantial fraction of all CO events occur in coding regions, in contrast with higher eukaryotes where recombination occurs almost exclusively within intergenic regions (@@REFs).

#### Estimation of conversion tract length and NCO recombination rate

Although previous studies using these crosses have provided evidence for non-crossover conversion (@@REFs) no estimates have been made for the rate of NCO recombination or the distribution of conversion tract lengths. Of the 331 conversion tracts observed, an outlying group of 7 very long (>18kb) complex tracts was found, described further in the next sub-section. Of the remaining 324 tracts, 94 were associated with a CO and 230 were assumed to be NCO conversion tracts. The majority of observed NCO conversion tracts had a minimal size less than 1kb, and 73% were smaller than 2kb (figure 3D). The relatively small size of conversion tracts and the available marker density means that some NCO events will not have been observed, because we required tracts to span at least two markers and more than 100bp. To estimate the NCO recombination rate and true tract length distribution, the incomplete discovery of NCO events and bias towards discovery of longer tracts has to be taken into account. In other organisms the distribution of conversion tract lengths has been found to fit a geometric model, with a single parameter *phi* determining the per-base-pair probability of extending a tract (@@REF). By Monte Carlo simulation we found that a geometric model also provided a good fit for the observed distribution of tract lengths in the present study with *phi* = 0.9993, although there was a small excess of tracts observed with minimal length greater than 3kb (figure 3D, figure 3E). Assuming this model for the tract length distribution, simulations predicted an NCO discovery rate of 40% for HB3 x Dd2, 39% for 7G8 x GB4, and 45% for 3D7 x HB3 where the marker density was slightly higher.

Adjusting for incomplete discovery, the average rate of NCO recombination over all three crosses was estimated at 7.5 NCO/meiosis (0.36 NCO/meiosis/Mb), thus COs are roughly twice as common as NCOs events. The 95% confidence interval for the NCO recombination rate based on sampling error is 6.8-8.1 NCO/meiosis, however this does not account for additional uncertainty in the estimation of NCO discovery rates for each cross. There was a linear correlation between chromosome size and NCO rate (figure 3F) however the correlation was weaker than for the CO recombination rate, presumably due to the fewer number of observed NCO events and thus greater sampling error. As with CO events there was a significant enrichment of NCO events within genes (P=0.002 by Monte Carlo simulation) with 37 (16%) of NCO conversion tracts falling entirely within a gene, 110 (48%) spanning a gene boundary, 35 (15%) entirely spanning a gene, and 14 (6%) intergenic.

As mentioned above, 7 apparently long (>18kb) complex conversion tracts were found. Two of these tracts occurred in clone JF6 (7G8 x GB4) within a 60kb region on chromosome 11, and thus appear to be part of a single complex long-range recombination event involving a total of 20 switches in inheritance (supplementary figure @@N). Two biological replicates of clone JF6 were sequenced and genotyped in this study, and the pattern of recombination was identical in both replicates. Similar observations were made for clone C@@ (3D7 x HB3) where a 70kb region on chromosome 14 accounted for 12 switches in inheritance, and clone 3BD5 (HB3 x Dd2) where an 80kb region on chromosome 10 contained 13 switches (supplementary figure @@N). At all of these loci there was no evidence of copy number variation or other artefacts that could manifest as an apparent excess of recombination. These observations suggest that, in addition to the two established DSB repair pathways leading to normal CO and NCO events (@@REF), other repair pathways may also be utilised during meiosis, and may have radical results in terms of generating novel haplotypes.

### Recombination with regions of copy number variation

Two independent loci associated with drug resistance are known to exhibit copy number variation (CNV) within these crosses and in *P. falciparum* clinical isolates. On chromosome 5 amplifications spanning the multi-drug resistance homologue *mdr1* have been found in Dd2 and several progeny of the HB3xDd2 cross (@@REFs), and in a large number of clinical isolates, particularly from South-East Asia (@@REFs). Both point mutations and amplifications of *mdr1* have been shown to alter sensitivity to a range of drugs (@@REFs). On chromosome 12 amplifications spanning *gch1*, the first enzyme in the folate biosynthesis pathway, have been found in 3D7, HB3 and Dd2 (@@check other parents) and clinical isolates (@@REF) and have been shown to confer resistance to anti-folate drugs (@@REF). The amplifications spanning *gch1* differ between 3D7, HB3 and Dd2 in size and copy number (@@REF). A previous study of CNVs within the HB3xDd2 cross using CGH technology observed Mendelian segregation of *gch1* amplifications in all progeny except CH3\_61 where both parental alleles appeared to be inherited together, which can be explained if one or more meiotic recombination events had occurred within the amplified region. High throughput sequencing data can be used to infer copy number variation by observing variations in the depth of read coverage (@@REFs). The relative orientation and distance between aligned read pairs can also be used to infer the arrangement of amplified segments, for example, tandem duplications or inversions (@@REF). Here we present the evidence for CNVs within the crosses provided by the sequence data resource, and study the inheritance of amplifications spanning *mdr1* and *gch1* and recombination within these amplified regions.

#### Evidence for copy number variations spanning drug resistance genes mdr1 and gch1

Amplifications spanning *gch1* and *mdr1* were clearly visible from depth of coverage of sequence reads aligned to the 3D7 reference genome. Orientation of read pairs was also informative in most cases, providing evidence for either tandem duplication (paired reads aligned facing away from each other) or tandem inversion (paired reads aligned in the same direction). On chromosome 5 a three-fold amplification of an 82kb region spanning *mdr1* and 13 other genes was evident in Dd2 and as either 2 or 3 copies in 14 progeny of HB3 x Dd2 (supplementary figure @@N) consistent with previous reports (@@REF). On chromosome 12 all parent clones including 3D7 had evidence for some form of amplification relative to the 3D7 reference genome spanning *gch1* (figure @@). The amplifications were different in each of the 5 parents in terms of the extent, copy number, and arrangement of amplified segments. The 3D7 clone had four copies of a 2kb segment spanning *gch1* only (the 3D7 reference genome to which all reads were aligned has only a single copy of *gch1*) and read pair alignments were consistent with a tandem array. HB3(1) had two copies of a 161kb tandem duplication spanning *gch1* and 38 other genes. Dd2 had 3 copies of a 5kb segment spanning *gch1* and 2 upstream genes and alignments indicated a tandem inversion. 7G8 had a 3-fold tandem amplification of a 31kb segment spanning *gch1* and 6 other genes. GB4 had 2 copies of a segment spanning *gch1* and 3 upstream genes, although the arrangement of segments could not be determined from read alignments. The HB3(2) sample appeared to be a mixture with approximately 20% of parasites retaining the duplication found in HB3(1) and 80% having no amplification. The *gch1* CNVs segregated in the progeny of all three crosses (supplementary figures @@N) except for two progeny of 3D7 x HB3 (C@@, C@@) and one progeny of HB3 x Dd2 (CH3\_61) where both parental alleles appeared to be inherited together, described further below.

Known copy number variations at two other loci within the core genome were also evident from the sequence data. On chromosome 3 a deletion of one of the two CLAG3 paralogs and the two intervening pseudogenes (one of which is a *var*) was evident in GB4 and 13 progeny of 7G8 x GB4 (supplementary figure @@N) consistent with the previously reported absence of CLAG3.2 in GB4 and other clones (@@REF). On chromosome 11 a 70kb segment spanning 12 genes was duplicated in both HB3 replicates and 5 progeny of HB3 x Dd2 (supplementary figure @@N). This amplification has been shown to involve a translocation of one copy of the region to chromosome 13 (@@REF). No further evidence for copy number variation spanning one or more genes was found within the core genome, although evidence for CNVs was rife within hypervariable regions.

#### Recombination within amplified regions leads to pseudo-heterozygosity

As mentioned earlier, CNV alleles did not segregate perfectly at the *gch1* locus and some of the progeny clones appeared to inherit both parental amplifications (figure @@N). This can be explained if recombination events occurred within the amplified region. Depending how homologous chromosomes align, a crossover within a region that is duplicated in one parent can result in an individual that maintains the same duplication but inherits one copy from either parent for some portion of the amplified region (figure @@). Within such a segment a haploid individual is effectively diploid and also becomes effectively heterozygous for any SNP, indel or smaller CNV variants within the segment that segregate between the two parents. In the analyses of genome variation and recombination described in the preceding sections, all SNP and indel variants within regions of copy number variation were filtered, because both of the variant calling methods used here assumed that samples are haploid, and therefore genotypes are not correctly modelled where effective ploidy is higher. To study crossover events within CNVs and identify regions of effective heterozygosity, we used the alignments of sequence reads from each progeny clone to the 3D7 reference genome, and examined the ratio between the coverage of reads supporting the two parental alleles at each segregating variant site. Within regions of effective heterozygosity reads supporting each parental allele should appear in a roughly 1:1 ratio, whereas elsewhere one parental allele should be found almost exclusively.

At the GCH1 locus, both clones C@@ and C@@ inherited the large 161kb duplication from parent HB3 as well as the smaller 2kb 4-fold amplification from parent 3D7 spanning GCH1 only. C@@ had a region of heterozygosity spanning the leftmost 130kb of the region duplicated in HB3, but was apparently homozygous for the remainder of this region (figure @@N). The most parsimonious explanation is that a single crossover occurred within the region duplicated in HB3. Clone C@@ had a region of heterozygosity spanning the entire region duplicated in HB3, with borders that appeared to coincide closely with the breakpoints of the duplication (figure @@N). This is harder to explain, as it would require two crossover events at or close to the borders of the duplicated region, which seems improbable unless the CNV breakpoints are also particularly prone to meiotic crossover. An alternative explanation is that the CNV is not a tandem duplication and one copy of the region has been translocated to a different chromosome, however read orientation evidence clearly indicated that the region is tandemly arrayed in both HB3 and C@@. For both clones C@@ and C@@ *gch1* itself lay within the region of effective heterozygosity, thus one copy of *gch1* was inherited from HB3 and 4 copies from 3D7. At the same locus clone CH3\_61 inherited the 161kb duplication from HB3 as well as the 5kb 3-fold tandem inversion from Dd2 (figure @@N). Two separate regions of heterozygosity were visible at either ends of the HB3 duplicated region, which can be explained if two crossover events occurred. Again *gch1* was within the region of heterozygosity, and CH3\_61 acquired 1 copy from HB3 and 3 copies from Dd2. In addition to these three cases, we found two further progeny of HB3 x Dd2 where crossovers occurred within the 82kb region spanning *mdr1* on chromosome 5that is amplified in Dd2 (supplementary figure @@N). Clone QC23 had a region of heterozygosity spanning the leftmost 16kb of the segment, and CH3\_61 was heterozygous for the rightmost 40kb spanning *mdr1* itself, both of which are consistent with a single crossover having occurred within the amplified region. These data show that crossover recombination within amplified regions can lead to effective heterozygosity at genes involved in drug resistance, the evolutionary implications of which are discussed below.

### A web application to facilitate data exploration and re-use

To facilitate re-use of the data reported here, we developed a web application that provides a number of tools for intuitive, interactive data exploration, available at <http://www.malariagen.net/apps/pf-crosses>. The application includes a tool for browsing and querying a table of variants for each call set; a tool for visualising and browsing the genotype calls at individual samples and patterns of inheritance and recombination within a cross; a tool for browsing the genome, allowing the location of variants to viewed in the context of genome features and alignment metrics; and a browser for visualising the sequence alignments themselves, implemented by embedding the LookSeq software (@@REF). The underlying technologies for this web application are being developed as a generic framework so that they can be used with other datasets, as part of an open source project (@@REF) that will be described in detail elsewhere.

## Discussion

We have presented a resource comprising deep sequence data for three *P. falciparum* genetic crosses. These three crosses are a unique asset in the study of *P. falciparum* and remain an important model for experimental biology as well as a source of high quality reference data on genome variation and meiotic recombination. The development of sequencing technology continues apace, and within the next few years it may be possible to re-sequence these crosses with longer reads and ultimately to fully assemble genomes for all parents and progeny. However, for the time being, we hope these data will provide a valuable foundation for studies using current sequencing technology to study *P. falciparum* genome biology in the lab and the field, as well as a useful benchmark against which new sequencing technologies and methods for variant discovery and assembly can be evaluated.

Using a combination of alignment and assembly methods we have constructed a high resolution map of genome variation within each cross, integrating SNP, INDEL and complex polymorphisms, and spanning both coding and non-coding regions of the core genome. These data have been calibrated using Mendelian errors and validated using sequencing of biological replicates, and thus represent a well-validated resource. They demonstrate that INDELs are the most abundant form of polymorphism within the core genome, in stark contrast to other eukaryotes for which genome-wide data are available. This is perhaps not surprising, given that the abundance of tandem repeats and low complexity sequence in both coding and non-coding regions of the *P. falciparum* genome is well-documented (@@REFs) and short tandem repeats are known to be highly variable due to replication slippage (@@REFs). Nevertheless, these are the most comprehensive and robust data available on INDEL variation in *P. falciparum* and fill an important gap in current knowledge. These data also provide the first genome-wide view of non-coding variation, which is almost completely unexplored in *P. falciparum*, yet could have important phenotypic consequences. For example, increased expression of genes through copy number amplification is known to be an important component of resistance to a number of anti-malarial compounds, and data from other species have shown that comparable increases in expression levels can also be obtained through variation in promoter and regulatory elements (@@REF). We found that INDEL diversity has a particular architecture in relation to predicted core promoters, which may indicate that these mutations are not neutral and have phenotypic consequences, although this will require experimental confirmation.

Another class of variation which has previously been difficult to study in a high throughput manner are complex polymorphisms observed in genes expressed at the merozoite surface, including MSP1 and members of the MSP3 family. High levels of variability have also been reported in other genes associated with the merozoite stage, such as the *surfin* family (@@REFs), although it is not yet clear whether these loci exhibit deep allelic dimorphism as seen in MSP1. The presence of diverged allelic forms and the fact that many of these genes appear to be under balancing selection (@@REFs) indicates their clinical importance as potential targets of the immune system and/or components of the erythrocyte invasion machinery. These loci are typically inaccessible to alignment-based methods, because some alleles are too diverged from the reference genome. Thus studies of these genes have so-far relied on more labour-intensive methods, and genome-widediscovery of other loci with similar variation has not been possible. Here we have shown that an assembly-based variant-calling method can access variation at a number of divergent loci, including MSP1 and members of the MSP3 and *surfin* families. Some known divergent loci, including MSP2 and S-antigen could not be accessed by either calling method due to the presence of longer repeats, and longer sequence reads will be required to access complex variation at all antigenic loci. Also the crosses studied here are clonal samples, and further work will be required to adapt these methods for use on clinical isolates which may comprise a mixture of parasite genotypes. However, these data provide a foundation for high-throughput surveys of complex variation at a number of clinically important loci across natural populations.

*P. falciparum* is a sexually reproducing eukaryotic pathogen, and these crosses originally provided the first demonstrations that parasites undergo meiotic recombination whilst in the mosquito (@@REF). These crosses remain the only available experimental system for studying meiotic recombination in *P. falciparum*, due to the practical difficulties associated with generating a cross. Previous work using individual crosses demonstrated a high rate of crossover recombination relative to eukaryotes with larger genomes (@@REF), being approximately 50 times greater than the average recombination rate in humans, although the recombination rate relative to the size of the smallest chromosome is almost identical (~0.5 Morgan) as expected if recombination rates are primarily calibrated to ensure one crossover per bivalent. Here we have pooled data from all three crosses and used a ~300bp resolution map of SNP and INDEL variation to provide the first estimates for both crossover (CO) and non-crossover (NCO) recombination rates within the core genome and the distribution of conversion tract lengths. We confirm a CO recombination rate in the range @@-@@ and estimate that CO events are approximately twice as frequent as NCO events. Conversion tract lengths are comparable to yeast but longer than humans and @@Drosophila@@ (@@REFs) and the observation of long-range complex recombination events observed in some progeny remains to be explained. In higher eukaryotes the recombination rate is known to be highly variable over the genome, with most recombination concentrated within narrow hotspots within non-coding regions (@@REFs). Previous work on the 7G8xGB4 cross suggested that the rate of recombination may not be uniform over the *P. falciparum* genome (@@REFs) although the number of recombination events available in a single cross was not sufficient to prove the existence of hotspots. Here we found that the CO recombination rate was lower within a highly localised region surrounding the centromeres, and that CO events occurred with appreciable frequency in both coding and non-coding regions, with a significant excess of events within coding regions relative to a model of uniform recombination. While population data will be required to robustly evaluate the support for different hotspot models, the fact that we observe recombination events frequently in both coding and non-coding regions of the core genome, and that these different regions have very different nucleotide composition and sequence characteristics, suggests that a model of highly punctuate recombination targeted at specific sequence motifs is unlikely, and that the location of recombination events may be less constrained in *P. falciparum*. Note that these findings apply only to the core genome, and entirely different processes may operate within hypervariable regions (@@REF).

There are multiple lines of evidence that gene amplification plays a key role in the evolution of drug resistance in *P. falciparum* (@@REFs). Amplifications spanning the multi-drug resistance gene *mdr1* segregate in the HB3xDd2 cross, and pleiomorphic amplifications spanning the anti-folate resistance gene *gch1* segregate in all three crosses, thus the data resource described here provides a controlled setting in which to study copy number variation at drug resistance loci and observe it in the context of meiotic recombination. We have extended the previous observation of a recombination event within the *gch1* amplification in the HB3xDd2 cross (@@REF) to illustrate four other cases of recombination within amplified regions, and shown that all of these events are associated with regions of effective heterozygosity within a progeny clone where both parental sequences are inherited and maintained within a single haploid genome. Such events could have several evolutionary consequences. Firstly, drug resistance mutations generally confer a fitness cost relative to the wild type allele in the absence of drug pressure (@@REF) and can also confer both resistance to one class of drugs and sensitivity to another (@@REF). The process of amplification followed by homologous recombination could provide a mechanism by which both mutant and wild type alleles are acquired, compensating for fitness costs associated with either allele alone. Secondly, the same process could enable copies of an amplified region to diverge in jumps and thus more rapidly than if by the gradual accumulation of mutations alone.

Taken together, this data resource provides a concrete illustration of how different modes of mutation and recombination could interact to create an evolutionary system that is more than the sum of its parts. Although we have not attempted to precisely map CNV breakpoints here, previous studies have found that CNV breakpoints almost invariably occur at sites with some degree of local homology, due to the presence of tandem repeat sequences found commonly throughout the genome, indicating that amplifications arise due to improper pairing of homologous chromosomes followed by unequal crossover (@@REF). The abundance of tandem repeats in the *P. falciparum* core genome creates a rich network of ectopic homology and thus potential for amplifications to arise. Previous work has also shown that CNV breakpoints are found in repeat regions that are slightly longer than the genome-wide average (@@REF) thus variations in the length of tandem repeat regions could alter the homology network and shift the amplification potential to a different set of loci. We have shown here that INDEL variation within tandem repeat regions is abundant throughout the core genome and thus the homology network is likely to be highly dynamic and variable within populations. Once an amplification has occurred and has persisted long enough for parasites to be taken up by a mosquito, meiotic recombination provides an opportunity for different copies of the amplified region to be acquired within the same genome. Both the initial amplification and the acquisition of pseudo-heterozygosity may have fitness benefits which enable short-term adaptation to strong and/or variable selection pressures, for example, due to introduction of new anti-malarial drugs or other changes in drug policy. Over multiple generations, meiotic recombination could enable amplified genes to diverge far more rapidly than by the accumulation of mutations alone, which in turn could accelerate the evolution of modified or novel gene functions. The core genome thus appears stable yet poised and retains the capacity to undergo bursts of evolution within any region that may come under selection pressure. It remains to be seen whether meiotic recombination in concert with INDEL mutation and gene amplification is playing a major role in adaptation to drug pressure in natural populations. However, these data provide a strong indication that different modes of mutation and recombination cannot be considered in isolation when studying the emergence and spread of drug resistance.

## Methods

### DNA samples

DNA samples from the parents and progeny of the 3D7xHB3 cross were obtained from cultures maintained at Glasgow University. DNA samples from the HB3xDd2 cross were obtained from @@TODO. DNA samples for the 7G8xGB4 cross were obtained from @@TODO, including the provenance of replicate clones that got cultured locally@@. @@TODO say something about DNA extraction methods?

### Whole genome sequencing

All sequencing was carried out using Illumina high throughput technology as described in @@REF except that the PCR-free method of library preparation as described in @@REF was used for all samples.

### Variant calling

Variants were called by two independent methods. The alignment method used the Genome Analysis Toolkit version @@ (@@REF) and followed published best practice (@@REF) with some adaptations for *P. falciparum*. The assembly method used Cortex version @@V (@@REF) following the independent workflow. Filtered variants from both calling methods were then combined into a single call set of segregating variation for each cross.

### Inference of CO and NCO recombination events and conversion tracts

The combined variant call sets was used to infer CO and NCO events and identify conversion tracts within each cross. To identify conversion tracts, parental haplotype blocks with a minimal length shorter than 10kb were identified. Any such blocks occurring in isolation were assumed to be simple conversion tracts. Any such blocks occurring adjacent to each other (indicative of a complex conversion tract) were merged into a single complex conversion tract. To identify CO events, all genotype calls within conversion tracts were first masked, and remaining switches in parental inheritance were called as CO events. Conversion tracts occurring adjacent to a CO were then identified, and the remaining conversion tracts were assumed to be associated with NCO events.

### Recombination analyses

To calculate the CO recombination rate for each cross the unadjusted recombination fraction was summed across all markers. No map function was used because the marker density was so high and thus we assume no crossovers were missed. To estimate the true conversion tract length distribution, conversion tracts were simulated at different values for the parameter *phi*, and the distribution of tract lengths that would be observed given the available markers within each cross was computed. The parameter *phi* was fitted by examining quantile-quantile plots comparing simulated and actual distributions of observed tract lengths. These simulations also predicted the fraction of conversion tracts that would be discovered given the markers available in each cross and the requirement that tracts must span at least 100bp. The rate of NCO recombination was then estimated by adjusting the observed rate of NCO recombination by the discovery rate predicted by simulations.

### Analyses of copy number variation

The genome was divided into 300bp non-overlapping bins and the number of reads whose alignment started within each bin was calculated for each sample. These binned read counts were then normalised by dividing by the median read count found within the core regions of chromosome 14. Bins where the GC content was lower than 20% were excluded from coverage analyses due to coverage bias in most samples. The fraction of aligned reads with face-away orientation and same-strand orientation was calculated per position for each sample using pysamstats (@REF). Copy number state was predicted in all samples by fitting a Gaussian hidden Markov model to the normalised coverage data.

## References

@@TODO