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

Alistair Miles[[1]](#footnote-1),[[2]](#footnote-2), Zamin Iqbal[[3]](#footnote-3), Paul Vauterin, Richard Pearson1,2, Susana Campino2, Michel Theron2, Kelda Gould2, Daniel Mead2, Eleanor Drury2, John O’Brien[[4]](#footnote-4), Valentin Ruano Rubio[[5]](#footnote-5), Bronwyn MacInnis5, Lisa Ranford-Cartwright[[6]](#footnote-6), Michael Ferdig[[7]](#footnote-7), Karen Hayton8, Xinzhuan Su[[8]](#footnote-8), Thomas Wellems8, Julian Rayner2, Gil McVean3,[[9]](#footnote-9) and Dominic Kwiatkowski2,1

## 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 addressing 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 (Ashley et al. 2014). 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 (Manske, Miotto et al., 2012; Miotto et al., 2013), the generation of antigenic diversity (Claessens, Hamilton et al., 2014) and the genetic basis for artemisinin resistance (Ariey et al. 2014). 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 23Mb genome has an extremely biased nucleotide composition, with 80.6% (A+T) content overall and ~90% (A+T) in non-coding regions (Gardner et al. 2002). 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 (Gardner et al. 2002; Zilversmit et al. 2010; Muralidharan & Goldberg 2013; DePristo et al. 2006). Several multi-copy gene families encoding erythrocyte surface antigens are hypervariable, diversifying within the course of a single infection via ectopic recombination (Claessens, Hamilton et al., 2014; Bopp et al., 2013; Freitas-Junior et al., 2000). 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 (Roy et al. 2008). 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, although there is evidence that INDELs may be unusually abundant (Jeffares et al. 2007; Haerty & Golding 2011; Tan et al. 2010). Little is known about variation in non-coding regions, which could have a significant impact on phenotype by regulating gene expression (Mok et al. 2014). Knowledge of complex variation, where haplotypes are highly diverged from the reference genome, is constrained to a few well-studied loci.

There are many potential sources of error in the process of high throughput sequencing and variant calling (Robasky et al. 2014). Sequencing of replicates provides one strategy for mitigating errors (ibid.). Also, by sequencing related individuals, genotypes that are in violation of Mendelian inheritance can be identified and used as an indicator of error (Saunders et al. 2007). Unfortunately, although pedigrees can be readily obtained for many species, this is currently not the case for *P. falciparum*. The parasite is sexually reproducing but remains haploid for the majority of its life cycle, existing only briefly as a diploid when taken up into the mosquito mid-gut where it undergoes meiosis (Walliker et al. 1987). 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 (Wellems et al. 1991) and host specificity (Hayton et al. 2008). However, although only three crosses have been performed, each cross yields a large number of progeny clones, each of which is the result of an independent meiosis. The large number of progeny provides a higher power to observe Mendelian errors than smaller pedigrees or trios, and thus to identify variants which are spurious or where genotyping is unreliable. These three crosses are therefore a precious resource for studying genome variation, because biological replicates can be readily obtained, and because they represent the only *P. falciparum* experimental system within which Mendelian errors can be observed.

Here we describe a data resource comprising whole genome deep sequencing of parents and progeny of three experimental *P. 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 (DePristo et al. 2011; Li & Durbin 2009; McKenna et al. 2010) and reference-free sequence assembly (Iqbal et al. 2012) 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. Because these data could serve as a 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 address open questions regarding meiotic recombination in *P. falciparum*. Previous studies using lower-resolution genotyping methods have estimated crossover (CO) recombination rates (Su et al. 1999; Jiang et al. 2011; Walker-Jonah et al. 1992) and provided evidence for non-crossover (NCO) recombination (Su et al. 1999). One study used high-throughput sequencing to resolve recombination events in two progeny of the HB3xDd2 cross, finding evidence that rates of CO and NCO recombination may be similar (Samarakoon, Regier, et al. 2011). Previous studies have also demonstrated that at least some recombination events occur within coding regions (Kerr et al. 1994) and suggested that recombination events are not uniformly distributed over the genome (Jiang et al. 2011). 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 (Wellems et al. 1990; Samarakoon, Gonzales, et al. 2011). 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 by point mutation alone.

## Results

### Whole genome sequencing and genome accessibility

Whole genomes of parent and progeny clones from the crosses 3D7×HB3 (Walliker et al. 1987), HB3×Dd2 (Wellems et al. 1990; Wellems et al. 1991) and 7G8×GB4 (Hayton et al. 2008) 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 (Kozarewa et al. 2009). 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×HB3 and HB3×Dd2 respectively. For several progeny clones 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 acore 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 (Bopp et al. 2013; Claessens, Hamilton et al., 2014), and almost all genes in the associated *rif* and *stevor* families. Hypervariable regions also corresponded closely with regions of heterochromatin (Flueck et al. 2009). 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 data and web application, and may be of use to others studying these immunologically 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 (DePristo et al. 2011; Li & Durbin 2009), the other based on partial assembly of sequence reads and comparison of assembled contigs (Iqbal et al. 2012). 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[[10]](#footnote-10) 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 variant calls can be downloaded from a public FTP site or browsed via the web application at <http://www.malariagen.net/apps/pf-crosses>.

The near-perfect reproducibility of these variant calls, in addition to the highly plausible patterns of recombination (see below), suggests that all discovered variants mark some genuine haplotypic difference between parental clones. However, some variants may be false discoveries, in the sense that the variant allele is not correctly ascertained. To estimate false discovery and sensitivity rates we compared variant calls obtained for clone HB3 from the high throughput sequence data with the HB3 draft genome assembly (Birren et al. 2006) and publicly available HB3 gene sequences (supplementary information). We restricted the analysis to a subset of 32 genes where both a single contig from the draft assembly and a single HB3 gene sequence could be unambiguously aligned to the 3D7 reference spanning the entire gene. Even within this limited subset we found discordances between the draft assembly and the gene sequences, particularly involving INDELs (supplementary figure @@N), suggesting that both pre-existing sources may be affected by errors. @@N/@@M SNP and @@N/@@M INDEL alleles discovered in HB3(1) were not found in either the draft assembly or the gene sequences, giving an FDR of @@X% for SNPs and @@Y% for INDELs. FDRs were similar for the HB3(2) replicate (supplementary figure @@N). @@N/@@M SNPs and @@N/@@M INDELs found in both the draft assembly and the gene sequences were not discovered in HB3(1), giving a sensitivity of @@X% for SNPs and @@Y% for INDELs. Sensitivity was lower for HB3(2) because it was genotyped as part of the HB3xDd2 cross where many INDELs were multiallelic (parents have different non-reference alleles) but only the alignment method was capable of genotyping multiallelic variants. This technical limitation of the assembly method accounts for the lower number of segregating INDELs discovered in the HB3xDd2 and 7G8xGB4 crosses compared with 3D7xHB3 (table 1). Additional details are given in supplementary table @@N and supplementary figure @@S, and interpretation of these results is discussed below.

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

Analysis of the combined variant call sets revealed that, within the core genome, segregating INDELs were more abundant than SNPs in all three crosses (table 1). 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 2:1. This abundance of INDELs is exceptionally high when compared with other species for which data are available, for example, in humans SNPs are approximately 30 times more common than INDELs (Abecasis et al. 2012). 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) tracts (figure 1B). Tandem repeat sequences are prone to slipped strand mis-pairing during DNA replication (Li et al. 2002; Lovett 2004) and are known to be associated with high rates of INDEL mutation (Montgomery et al. 2013). Poly(AT) repeats are very common in the non-coding regions of the *P. falciparum* genome (Gardner et al. 2002) and poly(asparagine) repeats are unusually abundant within the exome (Muralidharan & Goldberg 2013), 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, Brick et al. (2008) predicted the location of core promoters upstream of genes based on a training set of known transcription start sites. 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). 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 19 loci with highly diverged alleles, where local diversity over a region up to 2kb was up to 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*, a highly expressed protein located at the merozoite surface, where several regions of the gene are known to exhibit deep allelic dimorphism (Ferreira et al. 2003). 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 two main divergent regions of msp1 (blocks 4-11 and blocks 13-16 (Ferreira et al. 2003)) as verified by comparison with capillary sequence data (supplementary information). 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 *surf* 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, suggesting involvement in the early stages of infection (Lasonder et al. 2008). 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 results 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 provide 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 ~300bp in all three crosses, an order of magnitude greater resolution than available previously (Jiang et al. 2011). In eukaryotes, programmed double strand breaks (DSBs) during meiosis are resolved by either crossover (CO) or non-crossover (NCO) between homologous chromosomes (Hastings 1992; Youds & Boulton 2011; Mancera et al. 2008; Baudat & de Massy 2007). A CO is a reciprocal exchange accompanied by a conversion tract, whereas an NCO is a conversion tract without reciprocal exchange (also known in the literature as a gene conversion, although NCO events can occur in either coding or non-coding regions). An algorithm was used to infer CO and NCO events from the size and arrangement of parental inheritance 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 inheritance 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 crossovers play an essential role in chromosome segregation and thus the recombination rate is calibrated to produce at least one CO per bivalent (Baudat & de Massy 2007; Martinez-Perez & Colaiácovo 2009; Mancera et al. 2008).

The centromeres were cold-spots of CO recombination, as expected from studies in other eukaryotes and confirming previous data from the 7G8xGB4 cross (Jiang et al. 2011), 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 within a gene (intragenic COs), 162 (13%) were within an intergenic region, and 636 (53%) were ambiguous (flanking markers spanned a gene boundary). The number of intragenic CO events was significantly higher than expected if CO events were distributed uniformly over the genome (P=0.001 by Monte Carlo simulation). Of the 396 intragenic COs observed, 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 an exon 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 were intragenic and occurred within coding regions, in contrast with humans where the majority of recombination occurs within hotspots that preferentially occur near genes but outside of the transcribed domain (Myers et al. 2005).

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

Of the 331 conversion tracts observed, an outlying group of 7 very long (>18kb) complex tracts was found, described further below. 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 *Drosophila* the distribution of conversion tract lengths has been found to fit a geometric model, with parameter *phi* determining the per-base-pair probability of extending a tract (Hilliker et al. 1994). 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; @@TODO consider main figure?). 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 C04 (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 do not fit well with conventional DSB repair pathways leading to normal CO and NCO events, suggesting other repair pathways may also be utilised during meiosis (Mancera et al. 2008) and may have radical results in terms of generating novel haplotypes.

### Recombination within 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 (Samarakoon, Gonzales, et al. 2011; Gonzales et al. 2008; Wellems et al. 1990). Amplifications spanning *mdr1* are found in clinical isolates, particularly from South-East Asia, and both point mutations and amplifications of *mdr1* have been shown to alter sensitivity to a range of drugs (Price et al. 2004; Anderson et al. 2009). On chromosome 12 amplifications spanning *gch1*, the gene encoding the first enzyme of the folate biosynthesis pathway, have been found in both lab and clinical isolates and have been shown to confer resistance to anti-folate drugs (Kidgell et al. 2006; Anderson et al. 2009; Heinberg et al. 2013; Samarakoon, Gonzales, et al. 2011; Nair et al. 2008; Sepúlveda et al. 2013). 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 (Samarakoon, Gonzales, et al. 2011). High throughput sequencing data can be used to infer copy number variation by observing variations in the depth of read coverage and the relative orientation and distance between aligned read pairs (Mills et al. 2011; Zhao et al. 2013). Here we present evidence for core genome 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 *mdr1* and *gch1* were clearly visible from depth of coverage of sequence reads aligned to the 3D7 reference genome (figure 4; supplementary figure @@N). 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) (Medvedev et al. 2009). 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 (Wellems et al. 1990; Samarakoon, Gonzales, et al. 2011; Sepúlveda et al. 2013). On chromosome 12 all parent clones including 3D7 had evidence for some form of amplification relative to the 3D7 reference genome spanning *gch1* (figure 4). 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 except for two progeny of 3D7 x HB3 (C05, C06) 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 (Sepúlveda et al. 2013; Chung et al. 2007; Iriko et al. 2008). 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 (Hinterberg et al. 1994). 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 4). 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 5). 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 or the other parental allele should be found almost exclusively.

At the *gch1* locus, both clones C05 and C06 inherited the large 161kb duplication from parent HB3 as well as the smaller 2kb 4-fold amplification from parent 3D7 spanning *gch1* only. C06 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 4). The most parsimonious explanation is that a single crossover occurred within the region duplicated in HB3 (figure 5). Clone C05 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 4). This is harder to explain, as it would require two crossover events at or close to the borders of the duplicated region (supplementary figure @@N), which seems improbable unless the CNV breakpoints are also particularly prone to meiotic crossover (Völker et al. 2010). 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 C05. For both clones C05 and C06 *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 4). Two separate regions of heterozygosity were visible at either ends of the HB3 duplicated region, which can be explained if two crossover events occurred (supplementary figure @@N). Again *gch1* was within the region of heterozygosity, and thus 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 each cross; a tool for browsing the genome, allowing the location of variants to be 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 (Manske & Kwiatkowski 2009). 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[[11]](#footnote-11) that will be described in detail elsewhere.

## Discussion

@@TODO discuss apparent preference for recombination in coding regions, due to higher GC?

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 resource for experimental and genome biology. 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. Meanwhile, these data are a valuable foundation for studies that use current sequencing technology to investigate *P. falciparum* in the lab and the field, and provide 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 verified 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, considering the abundance of tandem repeats and other low complexity sequences in both coding and non-coding regions of the *P. falciparum* genome (Gardner et al. 2002; DePristo et al. 2006; Zilversmit et al. 2010; Muralidharan & Goldberg 2013) and the evidence that replication slippage causes high variability in short tandem repeats (Lovett 2004; Li et al. 2002; Tan et al. 2010). Nevertheless, the comprehensive data on INDEL variation presented here fill an important gap in current knowledge. The 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 (Mok et al. 2014). Increased expression of genes through copy number amplification is known to be an important component of resistance to anti-malarial compounds (Price et al. 2004; Heinberg et al. 2013) although comparable increases of expression can also be obtained through variation in promoter and regulatory elements (Li et al. 2002; Anderson et al. 2009). In particular, short tandem repeat length variations within regulatory elements are known to affect gene activity across a number of species (Li et al. 2002). We found that INDELs are exceptionally abundant in non-coding regions within *P. falciparum*, and furthermore that INDEL diversity has a particular architecture in relation to predicted core promoters, which may indicate that these mutations are non-neutral. Experimental work is now required to investigate the phenotypic consequences of these mutations.

It has previously been difficult to perform high throughput studies of complex polymorphisms observed in genes expressed at the merozoite surface, including *msp1* (Roy et al. 2008; Ferreira et al. 2003) and members of the *msp3* family (Singh et al. 2009). High levels of variability have also been reported in other genes associated with the merozoite stage, such as the *surf* family (Kidgell et al. 2006; Winter et al. 2005), although it is not 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 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 *surf* 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 demonstration that parasites undergo meiotic recombination whilst in the mosquito (Walliker et al. 1987). 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 (Su et al. 1999; Jiang et al. 2011) being approximately 50 times greater than the average recombination rate in humans. 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 12.7-14.3 kb/cM and estimate that CO events are approximately twice as frequent as NCO events. Conversion tract lengths are comparable to yeast (Mancera et al. 2008) but longer than humans (Jeffreys & May 2004) and *Drosophila* (Hilliker et al. 1994) and the observation of long-range complex recombination events spanning >60kb 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 (Myers et al. 2005). Previous work on the 7G8xGB4 cross suggested that the rate of recombination may not be uniform over the *P. falciparum* genome (Jiang et al. 2011). Here we confirmed that the CO recombination rate was lower within a highly localised region surrounding the centromeres, and found 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 (Claessens, Hamilton, et al. 2014).

There are multiple lines of evidence that gene amplification plays a key role in the evolution of drug resistance in *P. falciparum* (Price et al. 2004; Anderson et al. 2009; Nair et al. 2008; Kidgell et al. 2006; Heinberg et al. 2013; Triglia et al. 1991; Cowman et al. 1994). We have confirmed that amplifications spanning the multi-drug resistance gene *mdr1* segregate in the HB3xDd2 cross, and shown that 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 (Samarakoon, Gonzales, et al. 2011) 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 (Kondrashov 2012; Anderson et al. 2009) and can also confer both resistance to one class of drugs and sensitivity to another (Anderson et al. 2009). 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 (Nair et al. 2007). The abundance of tandem repeats in the *P. falciparum* core genome creates a rich network of non-allelic sequence 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 (ibid.) 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 persisted long enough for parasites to be taken up by a mosquito, meiotic recombination provides an opportunity for copies of the amplified region from different lineages to be acquired within the same genome, creating a region of pseudo-heterozygosity. 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 of *P. falciparum* thus appears stable yet poised 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 indicate that all known modes of mutation and recombination should be integrated when modelling 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 (Manske et al. 2012) except that the PCR-free method of library preparation as described in (Kozarewa et al. 2009) was used.

### Variant calling

Variants were called by two independent methods. The alignment method used BWA (Li & Durbin 2009) to align reads to the 3D7 version 3 reference genome, then applied GATK (McKenna et al. 2010) base quality score recalibration, indel realignment, duplicate removal, and performed SNP and INDEL discovery and genotyping across samples within each cross simultaneously, then used variant quality score recalibration to filter variants, following GATK best practice recommendations (DePristo et al. 2011; Van der Auwera et al. 2002) with some adaptations for *P. falciparum* described in supplementary information. The assembly method used Cortex (Iqbal et al. 2012) following the independent workflow. Filtered variants from both calling methods were then combined into a single 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 inheritance 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 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

Abecasis, G.R. et al., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), pp.56–65. Available at: http://dx.doi.org/10.1038/nature11632 [Accessed July 9, 2014].

Anderson, T.J.C., Patel, J. & Ferdig, M.T., 2009. Gene copy number and malaria biology. *Trends in parasitology*, 25(7), pp.336–43. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2839409&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2014].

Antoine Claessens, William L. Hamilton, Mihir Kekre, Thomas D. Otto, Adnan Faizullabhoy, Julian C. Rayner, D.K., 2014. Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of var genes during mitosis. *PLoS genetics*.

Ariey, F. et al., 2014. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature*, 505(7481), pp.50–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24352242 [Accessed July 11, 2014].

Ashley, E.A. et al., 2014. Spread of Artemisinin Resistance in Plasmodium falciparum Malaria. *New England Journal of Medicine*, 371(5), pp.411–423. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25075834 [Accessed July 31, 2014].

Van der Auwera, G.A. et al., 2002. *Current Protocols in Bioinformatics* A. Bateman et al., eds., Hoboken, NJ, USA: John Wiley & Sons, Inc. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4243306&tool=pmcentrez&rendertype=abstract [Accessed July 16, 2014].

Baudat, F. & de Massy, B., 2007. Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 15(5), pp.565–77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17674146 [Accessed December 1, 2014].

Birren, B. et al., 2006. Plasmodium falciparum HB3, whole genome shotgun sequencing project.

Bopp, S.E.R. et al., 2013. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. N. Maizels, ed. *PLoS genetics*, 9(2), p.e1003293. Available at: http://dx.plos.org/10.1371/journal.pgen.1003293 [Accessed December 15, 2014].

Brick, K., Watanabe, J. & Pizzi, E., 2008. Core promoters are predicted by their distinct physicochemical properties in the genome of Plasmodium falciparum. *Genome biology*, 9(12), p.R178. Available at: http://genomebiology.com/2008/9/12/R178 [Accessed December 15, 2014].

Chung, W.Y. et al., 2007. The CLAG/RhopH1 locus on chromosome 3 of Plasmodium falciparum: two genes or two alleles of the same gene? *Molecular and biochemical parasitology*, 151(2), pp.229–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17166605 [Accessed December 16, 2014].

Cowman, A.F., Galatis, D. & Thompson, J.K., 1994. Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. *Proceedings of the National Academy of Sciences of the United States of America*, 91(3), pp.1143–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=521470&tool=pmcentrez&rendertype=abstract [Accessed November 12, 2014].

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5), pp.491–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3083463&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

DePristo, M.A., Zilversmit, M.M. & Hartl, D.L., 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene*, 378, pp.19–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16806741 [Accessed November 3, 2014].

Ferreira, M.U. et al., 2003. Sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-1 (MSP-1) of Plasmodium falciparum. *Gene*, 304, pp.65–75. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12568716 [Accessed November 19, 2014].

Flueck, C. et al., 2009. Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS pathogens*, 5(9), p.e1000569. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2731224&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Freitas-Junior, L.H. et al., 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature*, 407(6807), pp.1018–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11069183 [Accessed December 15, 2014].

Gardner, M.J. et al., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906), pp.498–511. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3836256&tool=pmcentrez&rendertype=abstract [Accessed October 22, 2014].

Gonzales, J.M. et al., 2008. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS biology*, 6(9), p.e238. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2553844&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2014].

Haerty, W. & Golding, G.B., 2011. Increased polymorphism near low-complexity sequences across the genomes of Plasmodium falciparum isolates. *Genome biology and evolution*, 3, pp.539–50. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3140889&tool=pmcentrez&rendertype=abstract [Accessed November 19, 2014].

Hastings, P.J., 1992. Mechanism and control of recombination in fungi. *Mutation research*, 284(1), pp.97–110. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1279396 [Accessed November 19, 2014].

Hayton, K. et al., 2008. Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of Plasmodium falciparum invasion. *Cell host & microbe*, 4(1), pp.40–51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2677973&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Heinberg, A. et al., 2013. Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in Plasmodium falciparum. *Molecular microbiology*, 88(4), pp.702–12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3654098&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Hilliker, A.J. et al., 1994. Meiotic gene conversion tract length distribution within the rosy locus of Drosophila melanogaster. *Genetics*, 137(4), pp.1019–26. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1206049&tool=pmcentrez&rendertype=abstract [Accessed December 16, 2014].

Hinterberg, K. et al., 1994. Interchromosomal exchange of a large subtelomeric segment in a Plasmodium falciparum cross. *The EMBO journal*, 13(17), pp.4174–80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395340&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Iqbal, Z. et al., 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44(2), pp.226–32. Available at: http://dx.doi.org/10.1038/ng.1028 [Accessed July 9, 2014].

Iriko, H. et al., 2008. Diversity and evolution of the rhoph1/clag multigene family of Plasmodium falciparum. *Molecular and biochemical parasitology*, 158(1), pp.11–21. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2268843&tool=pmcentrez&rendertype=abstract [Accessed December 16, 2014].

Jeffares, D.C. et al., 2007. Genome variation and evolution of the malaria parasite Plasmodium falciparum. *Nature genetics*, 39(1), pp.120–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2663918&tool=pmcentrez&rendertype=abstract [Accessed November 13, 2014].

Jeffreys, A.J. & May, C.A., 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature genetics*, 36(2), pp.151–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14704667 [Accessed December 17, 2014].

Jiang, H. et al., 2011. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. *Genome biology*, 12(4), p.R33. Available at: http://genomebiology.com/2011/12/4/R33 [Accessed November 11, 2014].

Kerr, P.J., Ranford-Cartwright, L.C. & Walliker, D., 1994. Proof of intragenic recombination in Plasmodium falciparum. *Molecular and biochemical parasitology*, 66(2), pp.241–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7808474 [Accessed November 19, 2014].

Kidgell, C. et al., 2006. A systematic map of genetic variation in Plasmodium falciparum. *PLoS pathogens*, 2(6), p.e57. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1480597&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Kondrashov, F.A., 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings. Biological sciences / The Royal Society*, 279(1749), pp.5048–57. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3497230&tool=pmcentrez&rendertype=abstract [Accessed November 12, 2014].

Kozarewa, I. et al., 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods*, 6(4), pp.291–5. Available at: http://dx.doi.org/10.1038/nmeth.1311 [Accessed October 31, 2014].

Lasonder, E. et al., 2008. Proteomic profiling of Plasmodium sporozoite maturation identifies new proteins essential for parasite development and infectivity. D. E. Goldberg, ed. *PLoS pathogens*, 4(10), p.e1000195. Available at: http://dx.plos.org/10.1371/journal.ppat.1000195 [Accessed December 16, 2014].

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–60. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Li, Y.-C. et al., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular ecology*, 11(12), pp.2453–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12453231 [Accessed November 18, 2014].

Lovett, S.T., 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Molecular microbiology*, 52(5), pp.1243–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15165229 [Accessed November 19, 2014].

Mancera, E. et al., 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203), pp.479–85. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2780006&tool=pmcentrez&rendertype=abstract [Accessed November 18, 2014].

Manske, H.M. & Kwiatkowski, D.P., 2009. LookSeq: a browser-based viewer for deep sequencing data. *Genome research*, 19(11), pp.2125–32. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2775587&tool=pmcentrez&rendertype=abstract [Accessed December 5, 2014].

Manske, M. et al., 2012. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407), pp.375–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3738909&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Martinez-Perez, E. & Colaiácovo, M.P., 2009. Distribution of meiotic recombination events: talking to your neighbors. *Current opinion in genetics & development*, 19(2), pp.105–12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2729281&tool=pmcentrez&rendertype=abstract [Accessed December 16, 2014].

McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), pp.1297–303. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Medvedev, P., Stanciu, M. & Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6(11 Suppl), pp.S13–20. Available at: http://dx.doi.org/10.1038/nmeth.1374 [Accessed January 17, 2015].

Mills, R.E. et al., 2011. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332), pp.59–65. Available at: http://dx.doi.org/10.1038/nature09708 [Accessed July 9, 2014].

Miotto, O. et al., 2013. Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nature genetics*, 45(6), pp.648–55. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3807790&tool=pmcentrez&rendertype=abstract [Accessed September 30, 2014].

Mok, S. et al., 2014. Structural polymorphism in the promoter of pfmrp2 confers Plasmodium falciparum tolerance to quinoline drugs. *Molecular microbiology*, 91(5), pp.918–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24372851 [Accessed November 13, 2014].

Montgomery, S.B. et al., 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome research*, 23(5), pp.749–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3638132&tool=pmcentrez&rendertype=abstract [Accessed August 21, 2014].

Muralidharan, V. & Goldberg, D.E., 2013. Asparagine repeats in Plasmodium falciparum proteins: good for nothing? L. J. Knoll, ed. *PLoS pathogens*, 9(8), p.e1003488. Available at: http://dx.plos.org/10.1371/journal.ppat.1003488 [Accessed December 8, 2014].

Myers, S. et al., 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)*, 310(5746), pp.321–4. Available at: http://www.sciencemag.org/content/310/5746/321 [Accessed July 23, 2014].

Nair, S. et al., 2008. Adaptive copy number evolution in malaria parasites. *PLoS genetics*, 4(10), p.e1000243. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2570623&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2014].

Nair, S. et al., 2007. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Molecular biology and evolution*, 24(2), pp.562–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17124182 [Accessed November 12, 2014].

Price, R.N. et al., 2004. Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number. *Lancet*, 364(9432), pp.438–47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15288742 [Accessed November 11, 2014].

Robasky, K., Lewis, N.E. & Church, G.M., 2014. The role of replicates for error mitigation in next-generation sequencing. *Nature reviews. Genetics*, 15(1), pp.56–62. Available at: http://dx.doi.org/10.1038/nrg3655 [Accessed July 15, 2014].

Roy, S.W., Ferreira, M.U. & Hartl, D.L., 2008. Evolution of allelic dimorphism in malarial surface antigens. *Heredity*, 100(2), pp.103–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17021615 [Accessed November 19, 2014].

Samarakoon, U., Regier, A., et al., 2011. High-throughput 454 resequencing for allele discovery and recombination mapping in Plasmodium falciparum. *BMC genomics*, 12, p.116. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3055840&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Samarakoon, U., Gonzales, J.M., et al., 2011. The landscape of inherited and de novo copy number variants in a Plasmodium falciparum genetic cross. *BMC genomics*, 12, p.457. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3191341&tool=pmcentrez&rendertype=abstract [Accessed October 18, 2014].

Saunders, I.W., Brohede, J. & Hannan, G.N., 2007. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*, 90(3), pp.291–6. Available at: http://www.sciencedirect.com/science/article/pii/S088875430700136X [Accessed December 15, 2014].

Sepúlveda, N. et al., 2013. A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC genomics*, 14, p.128. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679970&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Singh, S. et al., 2009. A conserved multi-gene family induces cross-reactive antibodies effective in defense against Plasmodium falciparum. V. Moorthy, ed. *PloS one*, 4(4), p.e5410. Available at: http://dx.plos.org/10.1371/journal.pone.0005410 [Accessed December 17, 2014].

Su, X. et al., 1999. A genetic map and recombination parameters of the human malaria parasite Plasmodium falciparum. *Science (New York, N.Y.)*, 286(5443), pp.1351–3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10558988 [Accessed November 19, 2014].

Tan, J.C. et al., 2010. Variable numbers of tandem repeats in Plasmodium falciparum genes. *Journal of molecular evolution*, 71(4), pp.268–78. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3205454&tool=pmcentrez&rendertype=abstract [Accessed November 21, 2014].

Triglia, T. et al., 1991. Amplification of the multidrug resistance gene pfmdr1 in Plasmodium falciparum has arisen as multiple independent events. *Molecular and cellular biology*, 11(10), pp.5244–50. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=361573&tool=pmcentrez&rendertype=abstract [Accessed November 12, 2014].

Völker, M. et al., 2010. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome research*, 20(4), pp.503–11. Available at: http://genome.cshlp.org/content/20/4/503.full [Accessed December 15, 2014].

Walker-Jonah, A. et al., 1992. An RFLP map of the Plasmodium falciparum genome, recombination rates and favored linkage groups in a genetic cross. *Molecular and biochemical parasitology*, 51(2), pp.313–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1349423 [Accessed November 19, 2014].

Walliker, D. et al., 1987. Genetic analysis of the human malaria parasite Plasmodium falciparum. *Science (New York, N.Y.)*, 236(4809), pp.1661–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3299700 [Accessed November 11, 2014].

Wellems, T.E. et al., 1990. Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. *Nature*, 345(6272), pp.253–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1970614 [Accessed November 11, 2014].

Wellems, T.E., Walker-Jonah, A. & Panton, L.J., 1991. Genetic mapping of the chloroquine-resistance locus on Plasmodium falciparum chromosome 7. *Proceedings of the National Academy of Sciences of the United States of America*, 88(8), pp.3382–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51451&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

Winter, G. et al., 2005. SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes. *The Journal of experimental medicine*, 201(11), pp.1853–63. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2213267&tool=pmcentrez&rendertype=abstract [Accessed December 17, 2014].

Youds, J.L. & Boulton, S.J., 2011. The choice in meiosis - defining the factors that influence crossover or non-crossover formation. *Journal of cell science*, 124(Pt 4), pp.501–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21282472 [Accessed December 12, 2014].

Zhao, M. et al., 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14 Suppl 1(Suppl 11), p.S1. Available at: http://www.biomedcentral.com/1471-2105/14/S11/S1 [Accessed November 6, 2014].

Zilversmit, M.M. et al., 2010. Low-complexity regions in Plasmodium falciparum: missing links in the evolution of an extreme genome. *Molecular biology and evolution*, 27(9), pp.2198–209. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2922621&tool=pmcentrez&rendertype=abstract [Accessed October 20, 2014].

1. MRC Centre for Genomics and Global Health, University of Oxford, Oxford, UK [↑](#footnote-ref-1)
2. Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, UK [↑](#footnote-ref-2)
3. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK [↑](#footnote-ref-3)
4. Bowdoin College, Brunswick, Maine, USA [↑](#footnote-ref-4)
5. Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA [↑](#footnote-ref-5)
6. Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, UK [↑](#footnote-ref-6)
7. Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana, USA [↑](#footnote-ref-7)
8. Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, Maryland, USA [↑](#footnote-ref-8)
9. Department of Statistics, University of Oxford, Oxford, UK [↑](#footnote-ref-9)
10. <http://www.malariagen.net/projects/parasite/pf3k> [↑](#footnote-ref-10)
11. <http://github.com/cggh/panoptes> [↑](#footnote-ref-11)