Chapter 4

Understanding mechanisms of *var* gene diversity using next generation sequencing

4.1 Introduction

The constant exposure of surface antigens to the host immune system requires parasites to actively generate and display new variants of *Pf*EMP1 in order to effectively evade host defense mechanisms. In addition, the repertoire of PFEMP1 sequences within the parasite population as a whole needs to be sufficiently diverse such that infection with one genotype does not induce an effective response to other genotypes that may later infect the same patient. Understanding the mechanisms that generate the high level of diversity in *var* genes is therefore a critical step in elucidating the forces that drive their evolution (Frank et al., 2008).

One of the mechanisms suggested as a potential source of *var* diversity is ectopic gene conversion, a non-reciprocal transfer of genetic material between two non-allelic regions of high sequence similarity (Frank et al., 2008; Freitas-Junior et al., 2000). Gene conversion is usually initiated by a double strand break, where homologous recombination is used as a repair mechanism (Chen et al., 2007; Holliday, 1964) (See Chapter 1 for details). Our understanding of

gene conversion in eukaryotes is primarily derived from studies in yeast which show the presence of such events during both the meiotic and mitotic stages of cell division (Gao et al., 2005; Hicks, 2010; Symington et al., 1991). Despite the potentially deleterious effects (Chen et al., 2001, 2007), gene conversion could be advantageous for evolution of gene families and potentially their functional diversification (Maizels, 2005; Nielsen, 2003; Santoyo and Romero, 2005). It is also used by a number of pathogens to generate antigenic variation (Al-Khedery and Allred, 2005; Brayton et al., 2002; Santoyo and Romero, 2005) and also eliminate diversity (Jackson et al., 2012). In *P. falciparum*, double strand break and repair may take place at various points of cell division during both the sexual and asexual stages, providing a platform for meiotic and mitotic recombination events. The haploid nature of the parasite for the most part of its life cycle makes it relatively easy to study rates and mechanisms of recombination.

One method to study how new *var* genes are created would be to monitor the changes that take place on *var* genes as the disease progresses in human patients. However, this is not practical due to the short time scale. A number of genetic cross experiments between a series of parasite clones have been conducted in Chimpanzies. These studies have provided mechanistic and functional insights into the biology of *P. falciparum* parasites (Freitas-Junior et al., 2000; Jiang et al., 2011; Samarakoon et al., 2011; Walliker et al., 1987). Nonetheless, *var* genes pose a great challenge as over 60% are located in subtelomeric regions, where the application of existing laboratory-based and informatics methods is limited.

Previous attempts to understand mechanisms of *var* diversity within a genetic cross were focused on the DBLα region (Taylor et al., 2000a). Presence or absence of hybridisation bands on a Southern blot probed with parental DBLα amplicons to parental and five progeny clones was used to identify recombinant *var* genes. The results revealed 24 non-parental bands in the five progeny, $\sim$10 times more than the expected 2-3 events. However, most array hybridisation based studies exclude markers in subtelomeric regions and hyper variable multigene families to avoid potentially spurious results.

Recently, the 454 sequencing platform was used to study two progeny from a genetic cross between HB3 and DD2 (Samarakoon et al., 2011). The study

used a sliding window approach where a window of 15 SNPs was considered to compute allele frequencies and assign genotypes to either parent. A total of 24 and 27 crossover events were reported for the two progeny. Although this showed an increase from a previously reported number of 20 and 23 events respectively, it was based on ∼30% of the genome as it was not possible to reliably call SNPs form the rest of the genome. Conversely, a total of 25 and 22 non-crossover events were reported for the two progeny. The results demonstrate the value of a sequencing based approach to obtain higher resolution recombination maps. However, *var* genes and other hyper-variable regions were excluded.

In this chapter, five progeny from the original cross between 3D7 and HB3 parental clones (Walliker et al., 1987) were sequenced using Illumina's GAII platform to evaluate the rate of genome wide recombination events and distinguish large scale crossovers from non-cross over gene conversion events. Here, I will explore applications of short read sequencing and the Illumina technology in understanding mechanisms of *var* gene diversity.

4.2 Methods

4.2.1 Sample preparation and sequencing

Five clonal samples of the first genetic cross in *P. falciparum* between 3D7 and HB3 isolates (Walliker et al., 1987) were obtained from Prof. Chris Newbold's laboratory in Oxford. According to Walliker and colleagues, a mixture of infective gametocytes from the two parental clones were fed to ∼1500 mosquitoes which were then used to infect a splenectomized chimpanzee through 500 mosquito bites and intravenous injection. Recombinant parasites were then isolated from the infected chimpanzee and either cloned or passed through Pyrimethamine treatment followed by cloning followed by genotyping to identify recombinant progeny. Five progeny of these cloned progeny and two parental samples were prepared for sequencing on the Illumina GAII platform according to the PCR-free paired-end protocol (Kozarewa et al., 2009) as described in Chapter 2.

4.2.2 Reference genomes

4.2.2.1 Whole genome reference genomes

Reference genomes for the two parental clones 3D7 (version 3) and HB3 were obtained from geneDB (`http://www.genedb.org`) and the Broad Institute (`http://www.broadinstitute.org`) respectively. The HB3 genome is still in thousands of contigs that are joined to create larger sequence fragments (also called Supercontigs). Supercontigs are used to represent the state of a genome assembly that is highly fragmented and potentially incomplete. In order to obtain a better representation for HB3, supercontigs were ordered against the 3D7 genome using ABACAS (Assefa et al., 2009) as described in Chapter 2.

Draft genomes are particularly prone to inaccuracies that could be introduced at any point during the sequencing, assembly or scaffolding stages of the genome finishing process. Iterative Correction Of Reference Nucleotides (ICORN) (Otto et al., 2010a) was developed by Thomas Otto in our group to make use of the high sequence coverage of second generation sequencing platforms in order to detect and correct errors in a reference genome. The process involves iteratively aligning short reads to the reference genome and inspecting aligned reads with mismatches to detect high quality discrepancies. If the majority of the reads supports such discrepancies, a correction step is applied to replace the reference nucleotide by the base called from the reads. Subsequent mapping steps are used for further quality control based on the coverage of mapped reads. ICORN was used to generate a new reference genome for HB3 by iteratively correcting errors using Illumina reads of the HB3 isolate. Illumina reads (76 bp; paired-end reads; coverage $\sim$100X) were obtained from Prof. Dominic Kwiatkowski's group at the Sanger Institute. A combined reference genome of the two parents was generated by concatenating sequences from the 3D7 genome and the new HB3 reference.

4.2.2.2 *Var* gene sequences of 3D7 and HB3 genomes

Var genes of 3D7 were obtained from the latest annotation of version 3 of the genome (`http://www.genedb.org`). A total of 93 genes were obtained

including pseudo-genes and truncated Exon 2 genes. In addition to the *var* genes found from the annotation, HB3 genes were scanned for known conserved *var* motifs that were obtained from three lab adapted *P. falciparum* samples 3D7, HB3 and IT as described in Chapter 3. A multi-FASTA file of a combined *var* reference was generated by concatenating *var* genes from the two parents (93 from 3D7 and 97 from HB3).

4.2.3 Alignment and processing of sequence data

Raw reads were aligned to individual parental genomes and combined references (combined reference genome and combined *var* reference) using SMALT (*-i 500 -r 10 -x -k 13 -s 6*). SMALT is a fast and memory-efficient alignment tool that uses a traditional *k-mer*-based hashing method to index the reference genome. Reads that have a match to indices will then be aligned using a Smith-Waterman algorithm to ensure a highly sensitive output. In our experience, SMALT produces better alignment for *P. falciparum* genomes where the biased A+T content is a major challenge for short read aligners. Although the underlying principle is similar to its predecessor SSAHA (Ning et al., 2001), SMALT is more accurate, faster and user friendly. Special emphasis was given to the parameter *x* in order to map individual reads of a read pair to their best positions irrespective of the insert size between them. This property is unique to SMALT and important in identifying recombination breakpoints. The default option and other aligners will force read pairs to be aligned within the read-pair constraint leading to incorect alignment especially around breakpoints. Raw alignment results were stored in the Sequence Alignment/Map (SAM/BAM) format and processed using Samtools (version 0.1.18) (Li et al., 2009a).

Read pairs that did not align to a single position were excluded due to the difficulty of reliably determining where they should be placed. The Picard Suite from the Broad Institute (`http://www.broadinstitute.org`) was used to mark duplicate reads in the BAM file that were subsequently excluded from further analysis. Samtool's *mpileup* command and BCF tools (`http://www.vcftools.sourceforge.net`) were used for SNP calling (*samtools mpileup*). SNPs were then filtered based on quality (Q$\geq$ 30), number

of reads calling the SNP on each strand (read-depth$\geq$ 2) and position on the genome (eg. SNPs in low complexity regions are less reliable).

4.2.4 Genome-wide scan for recombination breakpoints

4.2.4.1 Detecting homologous recombination using comparative SNP maps

In order to identity regions of the genome that are involved in a reciprocal exchange (crossing-over), reads from the five progeny and the two parental samples were independently aligned to 3D7 and HB3 genomes using SMALT as described in the previous section. High quality SNPs (Q $\geq$ 30) generated from uniquely aligned reads were used to construct comparative SNP maps for each chromosome.

First, parental chromosomes were aligned to each other using BLAST (*blastn -F F -m 8*). The alignment output was then visualized in the Artemis Comparision Tool (ACT) (Carver et al., 2005). SNP files of the progeny and parental genomes were then uploaded to ACT using the Bamview utility (Carver et al., 2010). Recombination breakpoints and regions of homologous exchange were identified by visually inspecting these comparative maps for each chromosome. Regions of high and low SNP density were used to assign segments of chromosomes to either the 3D7 or HB3 genotypes. Accumulation of SNPs on one parent was expected to result in lack of SNPs on the other. Although this approach provides a simple and effective way of detecting regions of interest, it has major limitations in highly polymorphic regions, due to the difficulty of reliably calling SNPs. Differentiating real signals from background noise is therefore a major challenge in relying on visual inspection.

4.2.4.2 Using sliding windows to detect crossover and non-crossover recombination

In order to systematically identify breakpoints at a higher resolution and account for low complexity regions, a sliding window approach was developed. In addition to SNPs used in the previous section, supplementary evidence drawn from paired-end coverage of uniquely mapping reads was considered

prior to assigning genotypes.

Sliding windows of various sizes were analysed to assign regions of the genome to either 3D7 or HB3 based on the number of SNPs and proportion of each window covered by read-pairs. The following sections outline the approaches used to improve the signal to noise ratio while detecting recombination breakpoints.

Overlapping vs non-overlapping windows

Initially, non-overlapping sliding windows were considered to count SNPs and compute average paired-end coverage values. However, non-overlapping windows were found to underestimate breakpoints as described below (Figure 4.1). Assuming a minimum number of three SNPs (shown as red arrows), windows **a** and **b** (Figure 4.1A) will not be considered as they have below the expected number (two and one respectively). Conversely, window **e** of the overlapping panel is able to capture all three SNPs (Figure 4.1B).

Improving signal-to-noise ratio for smaller window sizes

In order to improve the signal-to-noise ratio for small windows, it was necessary to use a number of consecutive windows (n) instead of considering single windows. The minimum number of such windows was determined by the fragment size of the sequencing library (F) and the chosen window size (w). In order to capture small gene conversion events using paired reads, the length of the region covered by n windows (L) should not be greater than the fragment size, F. Assuming the length of the slide (s) to be half of the window size, the minimum number of windows could be obtained as follows:

$$s = \frac{w}{2} \tag{4.1}$$

$$\frac{w}{2}(n+1) = L \leq F \tag{4.2}$$

$$n \leq (\frac{2F}{w} - 1) \tag{4.3}$$

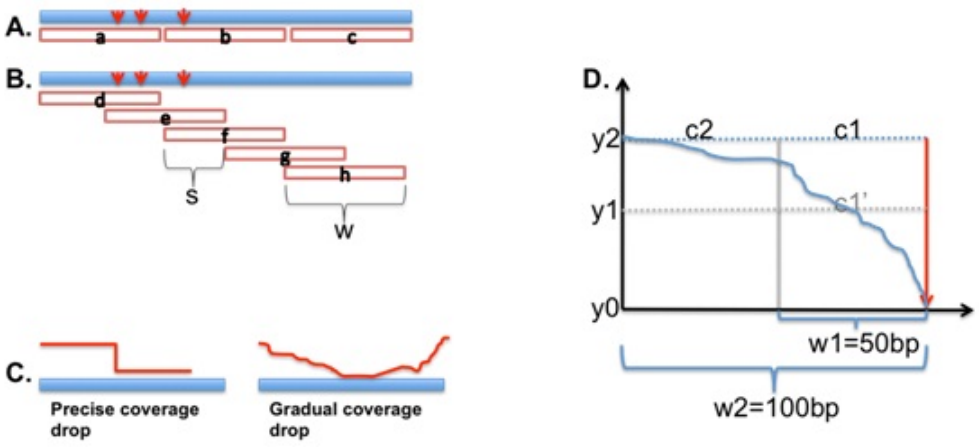

Figure 4.1: Breakpoint detection using SNPs and paired end coverage. **A)**. If a minimum of three SNPs were required to determine breakpoints, non-overlapping windows **a**, **b** and **c** are likely to miss genuine regions of recombination and gene conversion. **B)**. A sliding window approach overcomes limitations of non-overlapping windows. The window e is able to capture a region with three SNPs (red arrows) that would otherwise be missed by windows a and b in the non-overlapping approach. **w** represents the size of the window and **s** represent the sliding length. **C)**. Genome-wide scans for coverage drops are used to detect signal of recombination. However, it was necessary to distinguish between genuine coverage drops (precise coverage drops) and those with a gradual decrease in coverage. **D)**. The ratio of slopes over two regions length 50 bp and 100 bp was used to differentiate between precise and gradual coverage drops. (See below for details).

Gradient filter on coverage drops

While a decrease in coverage below the threshold could be an indication of a break-point, it could also be a result of poor-mapping due to the low complexity of the region in consideration. Distinguishing between precise and gradual coverage drops (Figure 4.1C) in a systematic way was thus adopted to minimise false positive calls. A filter based on differences in gradient (slope of the coverage plot) was developed as described below.

Two windows of lengths w_1 (50 bp) and w_2 (100 bp) were arbitrarily defined to the left of a given drop-point (Shown by the red arrows, Figure 4.1D). The x-axis represents position on the genome, while the y-axis represents paired-read coverage. The average mate pair coverage for windows w_1 and w_2 (i.e. c_1 and c_2); are represented by y_1 and y_2 on the y-axis of Figure 4.1D.

The slopes over windows w_1 and w_2 are defined as:

$$slope_1 = \frac{y_0 - y_1}{w_1} \tag{4.4}$$

$$slope_2 = \frac{y_0 - y_2}{w_2} \tag{4.5}$$

The ratio of the two slopes, r_S is therefore:

$$\begin{aligned} r_s &= \frac{slope_1}{slope_2} \\ &= \frac{w_2}{w_1} * \frac{y_1}{y_2} \\ &= 2(\frac{y_1}{y_2}) \end{aligned} \tag{4.6}$$

To determine whether a drop in coverage is gradual or precise, three possible values are considered for y_1 and y_2

1. $y_1 = y_2$
    - This represents the ideal case and shows no gradual drop in coverage (Figure 4.1C).

- Ratio of slopes (r_S) = 2

2. $y_1 > y_2$
   - The second scenario represents cases where there is an increase in coverage before the breakpoint.
   - Such cases are also acceptable.
   - Ratio of slopes (r_S) > 2

3. $y_1 < y_2$
   - There is a gradual drop in coverage.
   - Ratio of slopes (r_S) < 2
   - These events will be ignored.

Defining ambiguous regions and further quality filters

Ambiguous regions were defined as windows where both SNP count and coverage are below the minimum cutoff of three and five respectively. Such regions were grouped as non-unique (NU) or ambiguous and excluded from further analyses. In addition, polymorphic and low quality SNPs (Q<30), SNPs that have a strand bias (i.e. SNPs found on only one strand) and those common to all five progeny were excluded.

Final choice of window size

After evaluating the number of breakpoints at various window sizes (the results are shown in Figure 4.4), a window size of 20 kb was chosen to capture both homologous and non-homologous recombination breakpoints. Average Paired End Coverage (PEC), uniqueness (determined using a frequency count of *k-mer*s, k=30), and number of SNPs were computed over 20 kb windows (sliding by 10 kb). Windows that have over three SNPs were assigned to the HB3 parent and the boundaries were identified as break-points. Average PEC of above 5

was used to decide whether windows with less than three SNPs belong to the 3D7 parent or grouped as ambiguous (non-unique).

4.2.4.3 Detection and filtering of breakpoints in *var* genes

Initially, the iterative *de novo* assembly approach developed in Chapter 3 was applied with the aim of obtaining full length *var* genes for the five progeny. However, it was not possible to generate valid contigs due to the poor quality of raw data and the short read length (54 bp). The following sections describe alternative methods used to detect breakpoints in *var* genes.

Visual inspection

BAMview (Carver et al., 2010) was used to visualise coverage of uniquely and perfectly mapping read-pairs over parental *var* genes. Observed coverage plots were compared to expected patterns of recombination (Figure 4.2).

Mapping based detection of breakpoints

Raw reads were aligned to the combined *var* reference using SMALT and BWA to obtain a mapping output that contains uniquely and non-uniquely aligning reads respectively. The BAM files from both aligners were further processed to look for read pairs that map to different *var* genes on the same or different parents. However, the BWA output was used for the final analysis as the strict alignment criteria used in the SMALT mapping was found to exclude reads that align to homologous regions. Although the BWA mapping ensured inclusion of reads that are genuinely located on the *var* genes, the short read length (54 bp) makes it difficult to avoid spurious alignment of reads. Genes were therefore identified as potentially recombinant if they were bridged by a minimum of 50 read pairs. Initially, this cutoff may seem too high, however, it was chosen to account for the effect of spurious alignments. *De novo* assembly of reads that aligned to the genes bridged by paired-reads was then used to reconstruct *var* genes of the progeny and validate events (velvet v.1.2.03, (Zerbino and Birney, 2008)). Finally, the new contigs were ordered using ABACAS (Assefa et al., 2009) against parental genes and manually checked for formation of a valid *var*

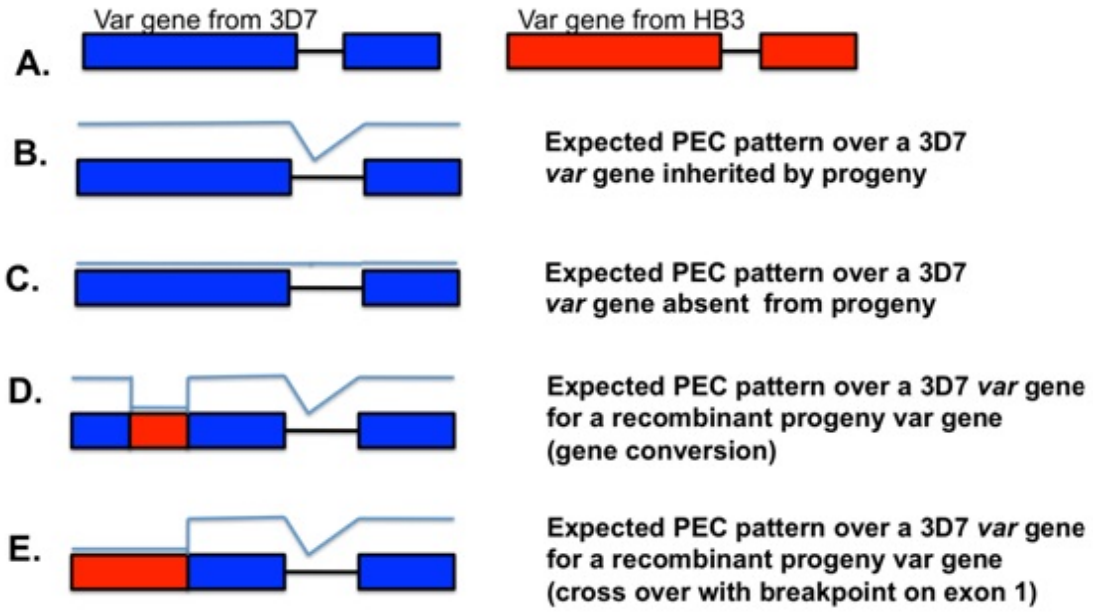

Figure 4.2: Patterns of Paired end coverage (PEC) over a 3D7 *var* gene with/without recombination. **A).** Parental *var* genes from the 3D7 and HB3 are represented in blue and red respectively. **B).** PEC over a 3D7 *var* gene suggests that the progeny had inherited the gene from the 3D7 parent. **C).** Conversely, lack of coverage may suggest that the 3D7 *var* gene was not inherited by the progeny. **D** and **E** show potential signals of recombination in the form of gene conversion and crossing over respectively. The drop in coverage on the first exon is expected to extend to the left for kb or Mb of sequence to establish a crossing-over event.

molecule.

4.3 Results

4.3.1 Sequence data and mapping to reference genomes

Five progeny obtained from the first genetic cross in *P. falciparum* between 3D7 and HB3 parental clones were sequenced using the Illumina GAII platform to a depth of 60- to 80-fold (paired reads of length of 54 bp; expected insert size of 200 bp). Raw reads were aligned to the two parental genomes and a combined reference genome. The 3D7 genome is the reference isolate (Gardner et al., 2002) with near-perfect base accuracy while HB3 is still highly fragmented and incomplete. A total of 93 and 47 genes were annotated as *var* (including pseudogenes and truncated exons) in 3D7 and HB3 respectively. Additional 50 genes were identified for HB3 using conserved *var* motifs resulting in 97 *var* genes for HB3.

During the initial whole genome analysis, reads that aligned to multiple positions were excluded, resulting in 34 to 60% of paired-reads mapping to the 3D7 parent (Table 4.1). The number of reads aligned to the HB3 reference was lower (26 to 52%) compared to 3D7 reflecting the poorer state of the HB3 reference rather than potential allelic bias. On the other hand, the reduction in mapping against the combined reference is primarily attributed to the high sequence similarity in the central regions of the parental genomes.

	X1	**X2**	**X3**	**X4**	**X5**
Raw reads ($x10^6$)	33.4	34.7	34.1	26.3	29.2
% Mapped	[52,46,26]*	39,29,22	47,38,21	68,58,29	61,52,30
% Mapped in pairs	44,39,23	34,26,20	42,34,19	60,52,27	52,46,27
Read length	54	54	54	54	54
Original ID	X33	XP8	X10	X4	XP2

*[3D7,HB3,Combined reference]

Table 4.1: Raw reads and mapping statistics to 3D7, HB3 and combined reference genomes. X1, X2, X3, X4 and X5 represent the five progeny.

4.3.2 Detecting genome wide crossover events

4.3.2.1 Visual identification

Initially, high quality SNPs (excluding SNPs with a quality score of below 30 and SNPs that were identified as heterozygous) generated by aligning progeny reads to the 3D7 parent were visualised using bamview. Figure 4.3 shows SNP views of the five progeny for chromosomes 1-5.

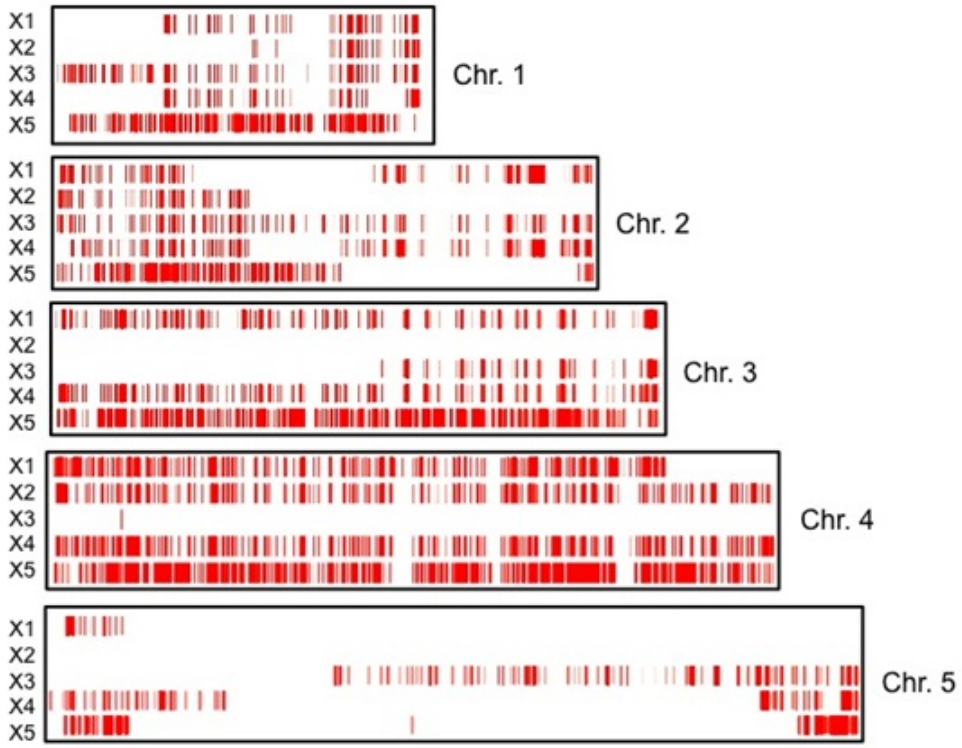

Figure 4.3: Visualising SNPs of the five progeny using the 3D7 parent as a reference. Here, SNPs are shown for the chromosomes 1 to 5. The quality of each SNP is reflected by the intensity of the red bars. The white horizontal blocks represent regions where no SNPs are called.

Although the red bars represented regions of chromosomes that were inherited from the HB3 parent, the white regions could either be inherited from the 3D7 parent or were ambiguous (i.e. non unique regions where reads could not be reliably aligned). In order to address this limitation, additional informa-

tion obtained from the reciprocal alignment (i.e. aligning reads from the five progeny to the HB3 reference) was used. In addition, Illumina reads of the 3D7 and HB3 clones were aligned to the 3D7 and HB3 reference genomes and used as control samples.

Comparative chromosome maps were thus constructed using the five progeny and both parental genomes (see Figure 4.4 for a map of Chromosome 3). A visual analysis of such maps at a chromosome level revealed regions of high (red bars) and low (white regions) SNP density. In most cases, regions of high SNP density on the 3D7 parent were found to have a lower density against the HB3 parent due to the reciprocal nature of homologous recombination during a crossover. SNP dense areas of a progeny against the 3D7 parent could be easily assigned to HB3 and vice versa. A total of 15 to 21 large-scale crossing over events were detected in each progeny with an average of $\sim$1 event per chromosome (Table 4.2).

Chr	**X1**	**X2**	**X3**	**X4**	**X5**
1	1	1	0	1	0
2	2	1	0	2	2
3	0	0	1	0	0
4	1	0	0	0	0
5	1	0	1	2	2
6	3	0	1	1	0
7	1	2	1	1	0
8	2	0	1	2	0
9	0	0	1	1	3
10	3	3	1	2	0
11	2	2	2	3	3
12	2	2	0	0	1
13	2	0	3	1	2
14	1	4	3	2	5
Total	**21**	**15**	**15**	**18**	**18**

Table 4.2: Number of cross-over events per chromosome. Large scale cross over events were detected using a visual inspection of comparative chromosome maps as shown in Figure 4.4.

Although initially the comparative chromosome maps as shown in Figure 4.4 may appear to clearly identify breakpoints, the white regions are still highly

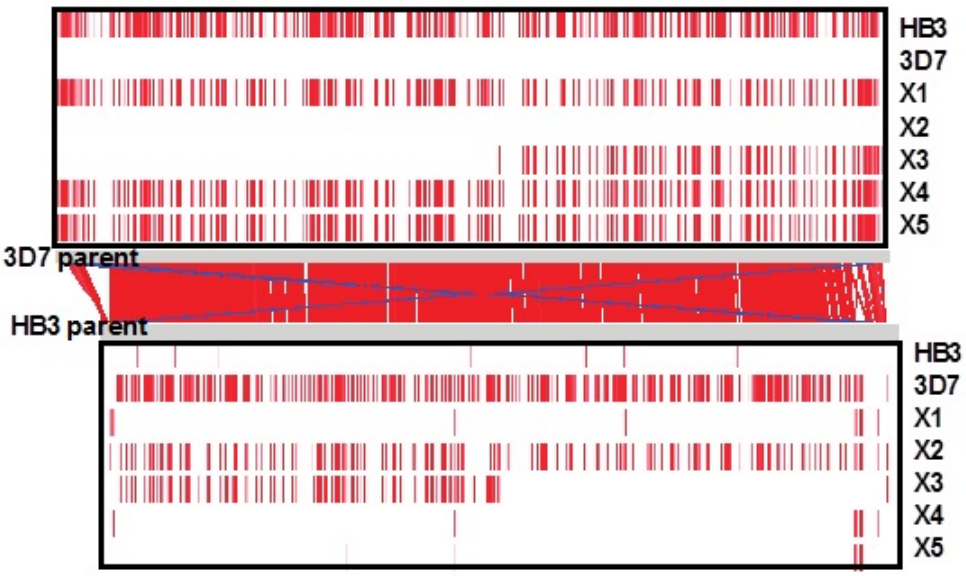

Figure 4.4: Comparative SNP-map of Chromosome 3: The top and bottom panels show SNPs called from progeny reads against Chromosome 3 of 3D7 and HB3 genomes respectively. The middle panel shows an Artemis Comparison Tool (ACT) view of synteny blocks (red lines) between the two chromosomes (generated by aligning the two chromosomes using *blastn -F F -m 8*). Copies of chromosome 3 in progeny X1, X4 and X5 are inherited from HB3; in progeny X2 it is inherited from 3D7; and in progeny X3, chromosome 3 is a result of a cross-over event. Although this figure looks cleaner compared to Figure 4.3 and the informatics approach described in the next section (Figure 4.6), assigning genotypes to the 'white' regions remains to be a challenge as they could either be from the 3D7 parent or non-unique regions.

ambiguous as they represent either the 3D7 genotype or non-unique regions. An informatic approach was therefore developed to merge the evidence from SNPs, paired-read coverage and reciprocal chromosome maps in order to detect large-scale crossover events as well as smaller gene conversion events.

4.3.2.2 Informatic identification

Genome-wide breakpoints were identified using a combination of high quality SNPs and paired-end coverage over overlapping sliding windows on the 3D7 genome. One aim of this experiment was to see if the shape of the distribution could be used to differentiate false positives from real events. The number of breakpoints detected was a function of window size and dropped from $\sim$500 to $\sim$14 per progeny with an increase in window size from 10 kb to 300 kb (Figure 4.5). The right tailed distribution was indicative of small scale gene conversion events (left tail) and the large scale crossover events (right tail). Although smaller window sizes ($<$ 20 kb) may provide a better resolution on the number and position of breakpoints, it was not possible to reliably assign genotypes mainly due to the poor data quality.

A window size of 20 kb was thus chosen to visualise patterns of recombination and gene conversion events across the 14 chromosomes (Figure 4.6)

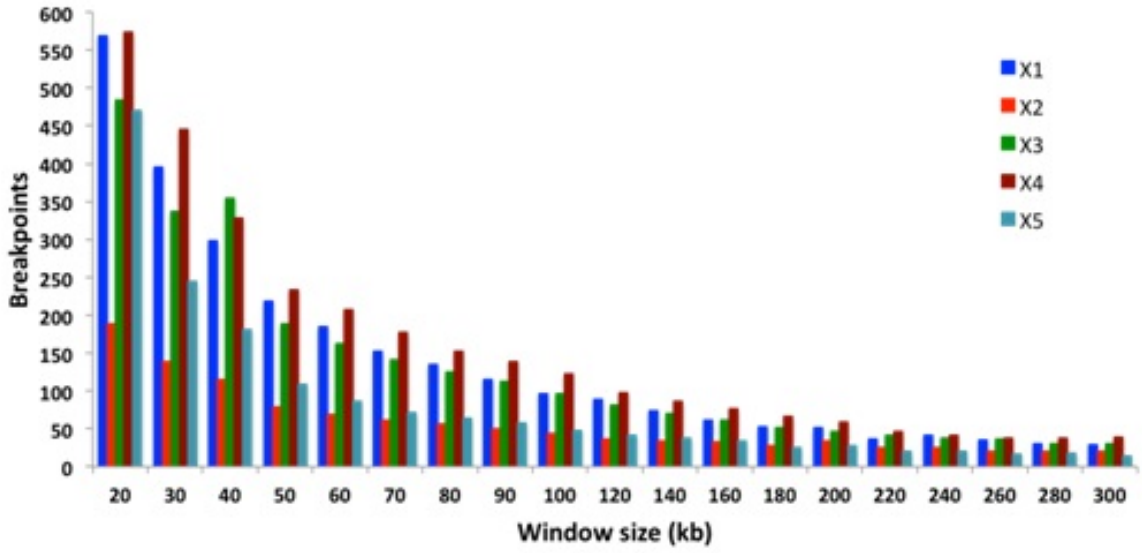

Figure 4.5: Genome wide breakpoints per window size. Breakpoints detected at larger window sizes were comparable with results of the visual identification.

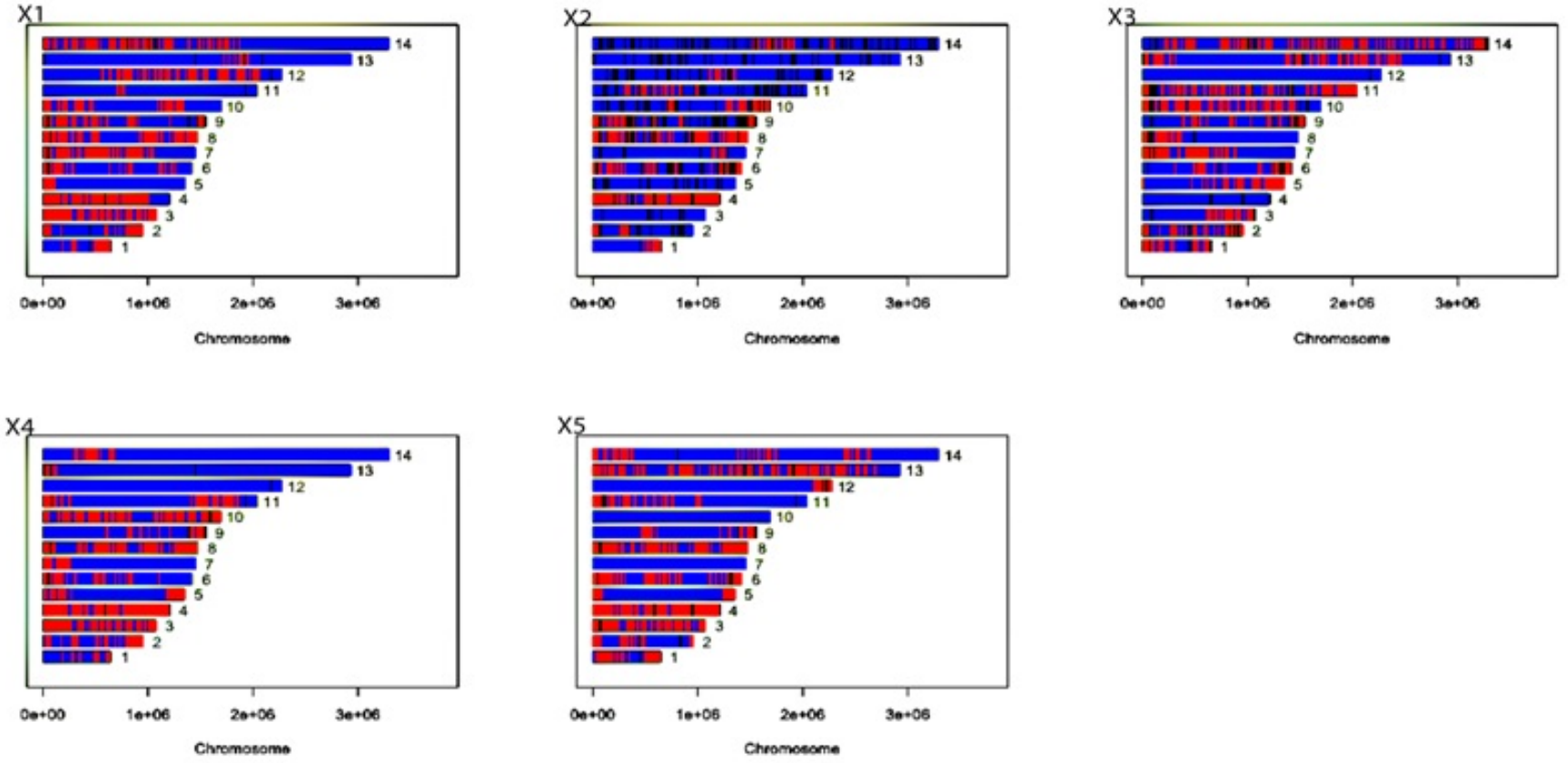

Figure 4.6: Breakpoints per chromosome at a window size of 20 kb (sliding by 10 kb) for the five progeny X1 to X5: Blue blocks show regions of chromosomes that were inherited from the 3D7 parent; red blocks show regions that were inherited from the HB3 parent; black regions represent non-unique (ambiguous) regions. Smaller window sizes provided a better resolution to identify large scale recombination events as well as potential non-crossover events.

Breakpoints detected at higher window sizes of ∼300 kb (Figure 4.6) were comparable to crossover events identified using a visual inspection of chromosome maps (Table 4.2). However, the numbers were not identical as the informatics identification included additional filtering and was done using sliding windows. Non-unique regions were also separately identified making the informatics approach more reliable instead of immediately assigning regions of low SNP density to the 3D7 parent. A smaller window size of 20 kb (sliding by 10 kb) was used to gain a better resolution of both crossover and putative gene conversion events (Figure 4.6).

Fewer breakpoints were observed in $X2(183)$ compared to other progeny $(X1 = 563, X3 = 478, X4 = 568, X5 = 465)$ due to lower read mapping (Table 4.1), which resulted in a larger proportion of non-unique regions. Further analysis of breakpoints at a window size of 20 kb revealed that most of the breakpoints (∼51%) were unique to one progeny compared to breakpoints shared by two (∼28%), three (∼18%) and four (∼3%) progeny. After excluding 7 breakpoints that were common to all, a total of 1,304 breakpoints were identified of which 270 were found in three or four of the five progeny.

In order to detect biases in inheritance, the proportion of the progeny genome assigned to each parent was analysed (Figure 4.7). The average inheritance over the progeny (Figure 4.7F and G) shows that chromosomes 5, 7 and 12 were predominantly inherited from 3D7 whereas the majority of chromosomes 2, 6, 8 and 9 came from HB3.

4.3.3 Signature of recombination in *var* genes

This section focuses on identifying crossovers and gene conversion events in *var* genes. As described previously, approaches based on read coverage and SNPs have a limited application in highly polymorphic regions such as *var* genes. Here, both a visual inspection and informatics approaches of detecting breakpoints are explored.

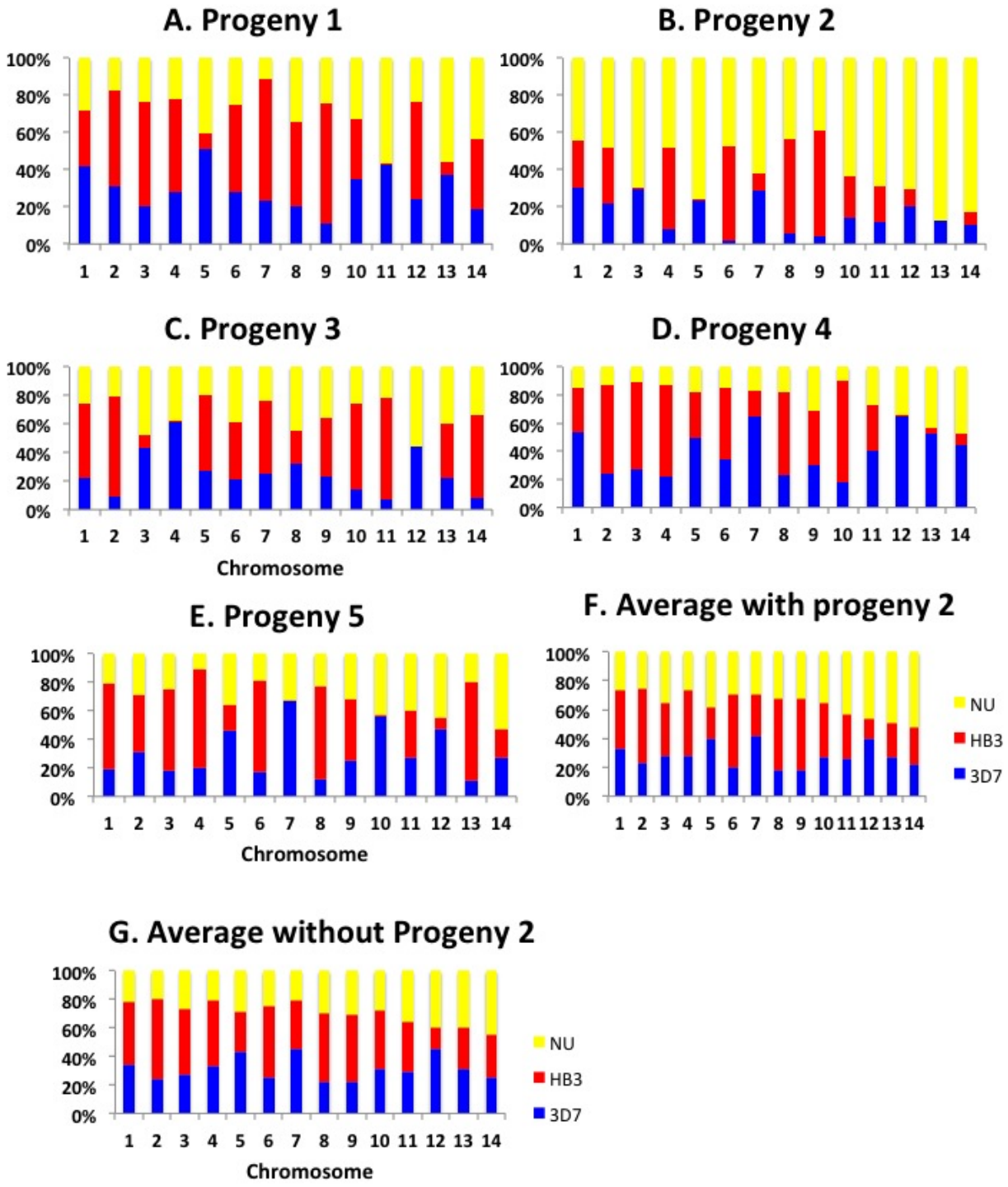

Figure 4.7: Per-chromosome inheritance patterns for each progeny and average inheritance profiles with / without progeny 2 to see the effect of poor quality data on the results. Blue and red blocks represent the proportion of chromosomes inherited from the 3D7 and HB3 parents respectively. Yellow bars represent regions of chromosomes that could not be reliably assigned to either parent.

4.3.3.1 Visual inspection of paired-read coverage

Firstly, BAM files of the five progeny were visualised and ambiguous patterns such as those characterised by lack of uniqueness and loss of coverage in all five progeny were excluded. Paired end coverage (PEC) plots for each *var* gene of the 3D7 genome were compared to expected patterns of recombination and gene conversion as shown in Figure 4.2. Regions of a gene that exhibited a clear loss of PEC were identified as breakpoints. Ideally, loss of PEC around a breakpoint is accompanied by an increased proportion of orphaned reads i.e. reads whose mates map to a different gene. Mates of orphaned reads were then investigated to find other genes that were involved in forming the recombinant in the progeny (Figure 4.8).

Although detecting regions of higher orphaned read coverage could be used to identify potential signature of recombination (Figure 4.8A), flanking regions of breakpoints often have higher levels of homology with donor and acceptor regions. Such homology, however essential to initiate recombination, may result in reads that align to multiple locations. It was therefore not possible to use this test in a genome-wide context. A closer look at read coverage over *var* genes in the 3D7 parent revealed that 24-50% of the genes were inherited from the 3D7 parent (Figure 4.9, Table 4.3).

A total of four genes in two progeny fulfilled both criterion i.e. loss of PEC and presence of unique orphaned reads that span two different genes (Table 4.3). The four genes with non-parental patterns of coverage were located on opposite ends of chromosomes one (PFA0005w, PFA0765c) and two (PFB1055c, PFB0010w) (Figure 4.10) in close proximity to telomeric-associated repetitive elements (TAREs). All four genes were of the group Type A, as they are transcribed away from the telomeres.

A detailed investigation of reads and their mates that mapped to these genes revealed that sequence blocks from two parental *var* genes were ordered to generate new genes in the progeny (Figures 4.11 and 4.12).

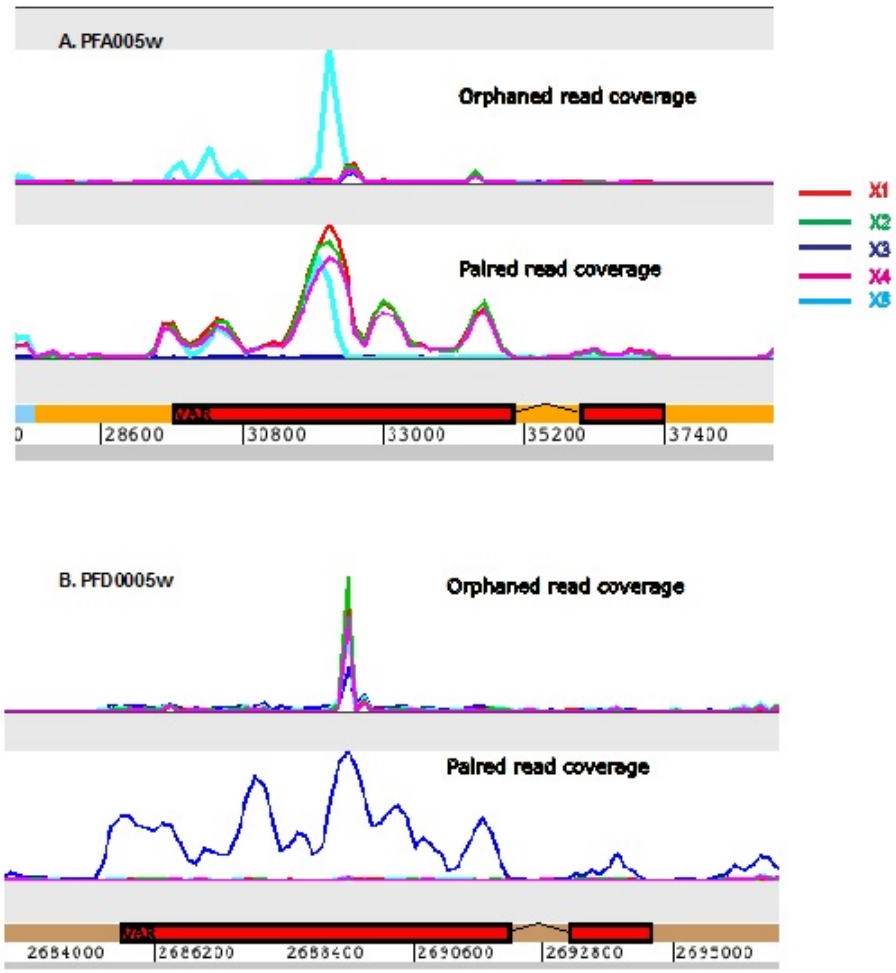

Figure 4.8: Paired and Orphaned read coverage over 3D7 *var* genes. **A)** Coverage pattern on progeny 5 shows a drop in PEC accompanied by an increase in orphaned read coverage indicating signatures of recombination for a gene located on Chromosome 1 of 3D7 (PFA005w). **B).** Coverage over a chromosome 4 *var* gene, PFD0005. The paired end coverage plot for progeny 3 (X3) indicated that PFD0005w was inherited from the 3D7 parent. Lack of coverage from other progeny may suggest that the gene was inherited from HB3 for progeny X1, X2, X4 and X5. No evidence of recombination is shown from the coverage plots.

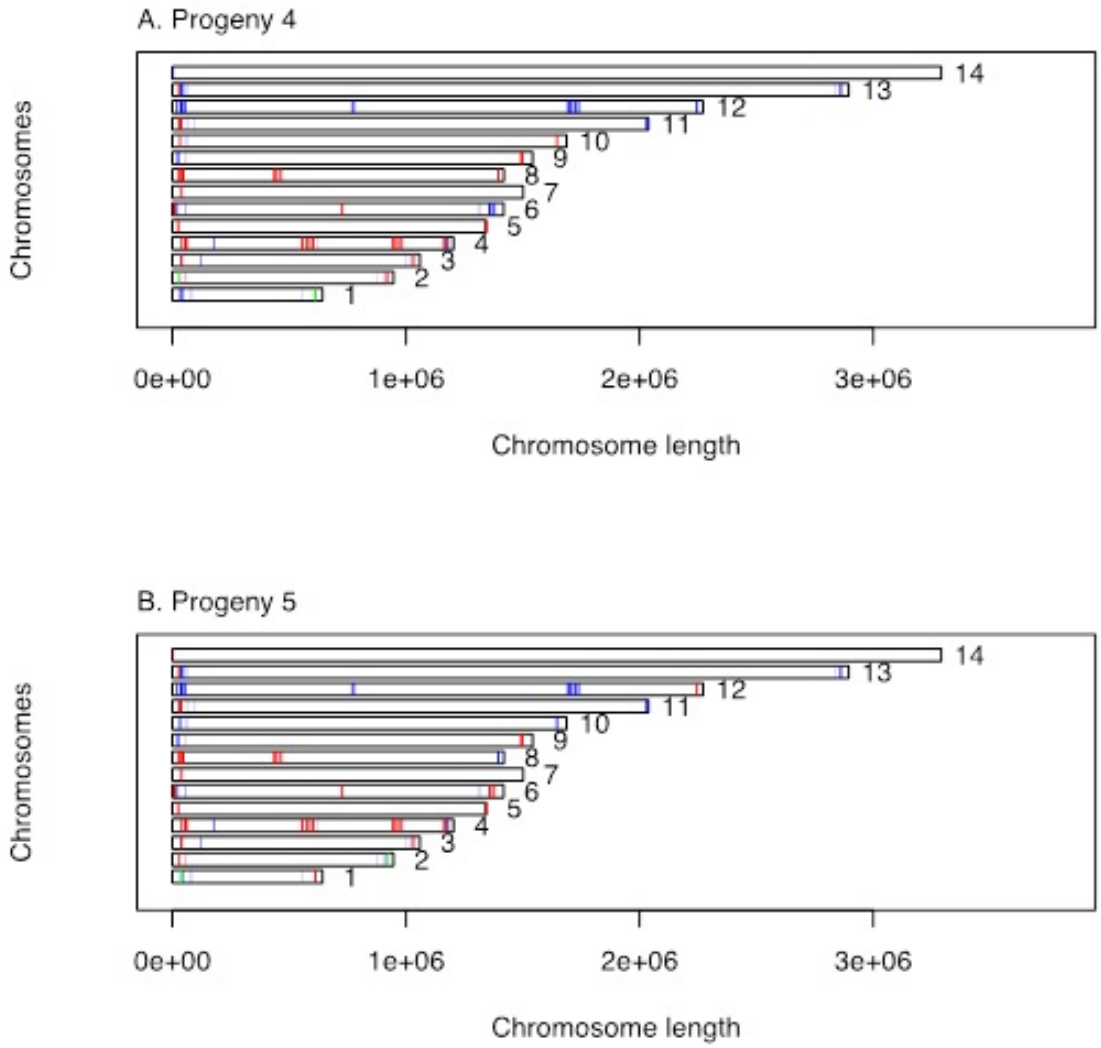

Figure 4.9: Chromosome view of *var* genes for progeny 4 and 5. The majority of *var* genes were inherited from either the 3D7 (shown in blue) or HB3 (shown in red) parents. The *var* genes on chromosomes 1 and 2 were involved in a non-homologous recombination (gene conversion) as indicated by the green bars. The results are consistent with the crossover data shown in Figure 4.6.

Var gene ID	chr.	X1	X2	X3	X4	X5
PFA0005w	1	3D7	3D7	HB3	3D7	**GC**
PFA0765c	1	HB3	HB3	HB3	**GC**	HB3
PFB0010w	2	HB3	HB3	HB3	**GC**	HB3
PFB0020c	2	HB3	HB3	HB3	HB3	HB3
PFB0045c	2	HB3	HB3	HB3	HB3	HB3
PFB0974c	2	HB3	3D7	HB3	HB3	3D7
PFB0975c	2	HB3	HB3	HB3	HB3	HB3
PFB1025c	2	HB3	HB3	HB3	HB3	HB3
PFB1045w	2	HB3	3D7	HB3	HB3	3D7
PFB1055c	2	HB3	3D7	HB3	HB3	**GC**
PFC0005w	3	HB3	3D7	HB3	HB3	HB3
PFC1120c	3	HB3	HB3	HB3	HB3	HB3
PFD0005w	4	HB3	HB3	3D7	HB3	HB3
PFD0020c	4	HB3	HB3	3D7	HB3	HB3
PFD0615c	4	HB3	HB3	3D7	HB3	HB3
PFD0625c	4	HB3	HB3	3D7	HB3	HB3
PFD0630c	4	HB3	HB3	3D7	HB3	HB3
PFD0635c	4	HB3	HB3	3D7	HB3	HB3
PFD0995c	4	HB3	HB3	3D7	HB3	HB3
PFD1000c	4	HB3	HB3	HB3	HB3	HB3
PFD1005c	4	HB3	HB3	3D7	HB3	HB3
PFD1015c	4	HB3	HB3	3D7	HB3	HB3
PFD1235w	4	HB3	HB3	HB3	HB3	HB3
PFD1254c	4	3D7	HB3	3D7	HB3	HB3
PFE0005w	5	HB3	3D7	3D7	HB3	HB3
PFE1640w	5	3D7	3D7	HB3	HB3	HB3
PFF0010w	6	HB3	HB3	3D7	HB3	HB3
PFF0030c	6	HB3	HB3	3D7	HB3	HB3
PFF0845c	6	HB3	HB3	HB3	HB3	HB3
PFF1580c	6	3D7	HB3	HB3	3D7	HB3
PFF1595c	6	3D7	HB3	HB3	3D7	HB3
MAL7P1.212	7	HB3	HB3	HB3	HB3	HB3
PF07_0048	7	HB3	3D7	HB3	3D7	3D7
PF07_0049	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.50	7	HB3	3D7	HB3	3D7	3D7
PF07_0050	7	HB3	3D7	HB3	3D7	3D7
PF07_0051	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.55	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.56	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.187	7	3D7	3D7	3D7	3D7	3D7
PF08_0142	8	HB3	HB3	3D7	HB3	HB3
PF08_0141	8	HB3	HB3	3D7	HB3	HB3
PF08_0140	8	HB3	HB3	3D7	HB3	HB3
PF08_0107	8	HB3	HB3	3D7	HB3	HB3
PF08_0106	8	HB3	HB3	3D7	HB3	HB3
PF08_0103	8	HB3	HB3	3D7	HB3	HB3
MAL8P1.207	8	HB3	HB3	HB3	HB3	HB3
MAL8P1.220	8	HB3	3D7	HB3	HB3	3D7
PFI0005w	9	HB3	HB3	3D7	3D7	3D7
PFI1820w	9	HB3	HB3	HB3	HB3	HB3
PFI1830c	9	HB3	HB3	HB3	HB3	HB3
PF10_0001	10	HB3	3D7	HB3	HB3	3D7
PF10_0406	10	3D7	HB3	3D7	HB3	3D7
PF11_0007	11	3D7	3D7	HB3	HB3	HB3
PF11_0008	11	3D7	3D7	HB3	HB3	HB3
PF11_0521	11	3D7	3D7	HB3	3D7	3D7
PFL0005w	12	3D7	3D7	3D7	3D7	3D7
PFL0020w	12	3D7	3D7	3D7	3D7	3D7
PFL0030c	12	3D7	3D7	3D7	3D7	3D7
PFL0935c	12	HB3	3D7	3D7	3D7	3D7
PFL0940c	12	HB3	3D7	3D7	3D7	3D7
PFL0947c	12	HB3	3D7	3D7	3D7	3D7
PFL1950w	12	HB3	3D7	3D7	3D7	3D7
PFL1955w	12	HB3	3D7	3D7	3D7	3D7
PFL1960w	12	HB3	3D7	3D7	3D7	3D7
PFL1970w	12	HB3	3D7	3D7	3D7	3D7
PFL2665c	12	3D7	3D7	3D7	3D7	HB3
MAL13P1.1	13	3D7	3D7	HB3	HB3	HB3
MAL13_0003	13	3D7	3D7	HB3	HB3	HB3
MAL13P1.356	13	3D7	3D7	3D7	3D7	3D7
PF14_0001	14	HB3	3D7	3D7	3D7	HB3

Table 4.3: Assigning *var* genes of the progeny to either the 3D7 or HB3 parent based on a visual inspection of paired read coverage. Four genes that were involved in a non-reciprocal recombination event were shown as GC.

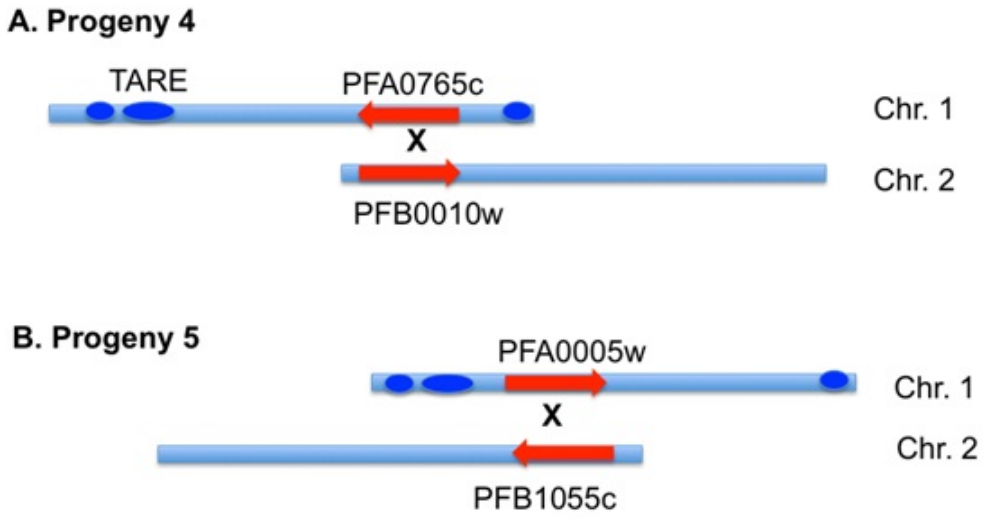

Figure 4.10: Subtelomeric *var* genes involved in ectopic recombination in progeny 4 and 5. *Var* genes on chromosomes 1 and 2 of the 3D7 parent showed evidence of a non-homologous recombination. The four genes were found to be of the group Type A.

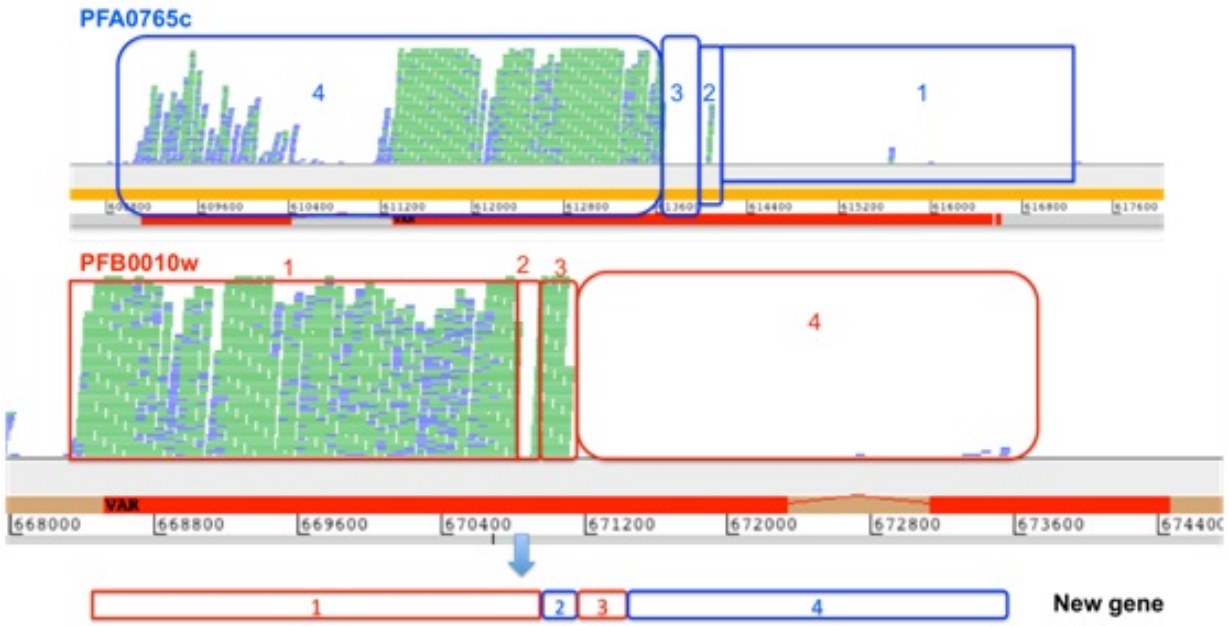

Figure 4.11: Evidence of gene conversion in progeny 4 (X4) from paired end coverage analysis. Coverage of paired reads is represented by a pileup of reads (blue and green stacks). A new *var* gene was formed by acquiring small blocks (1 to 4) from two genes on chromosomes 1 (PFA0765c) and 2 (PFB0019w). The upstream region and first half of Exon 1 was inherited from PFB0019w as shown by the increase in coverage over block 1 (red) on chromosome 2. Similarly, alternating blocks inherited from the two genes constituted the remaining regions of the new gene.

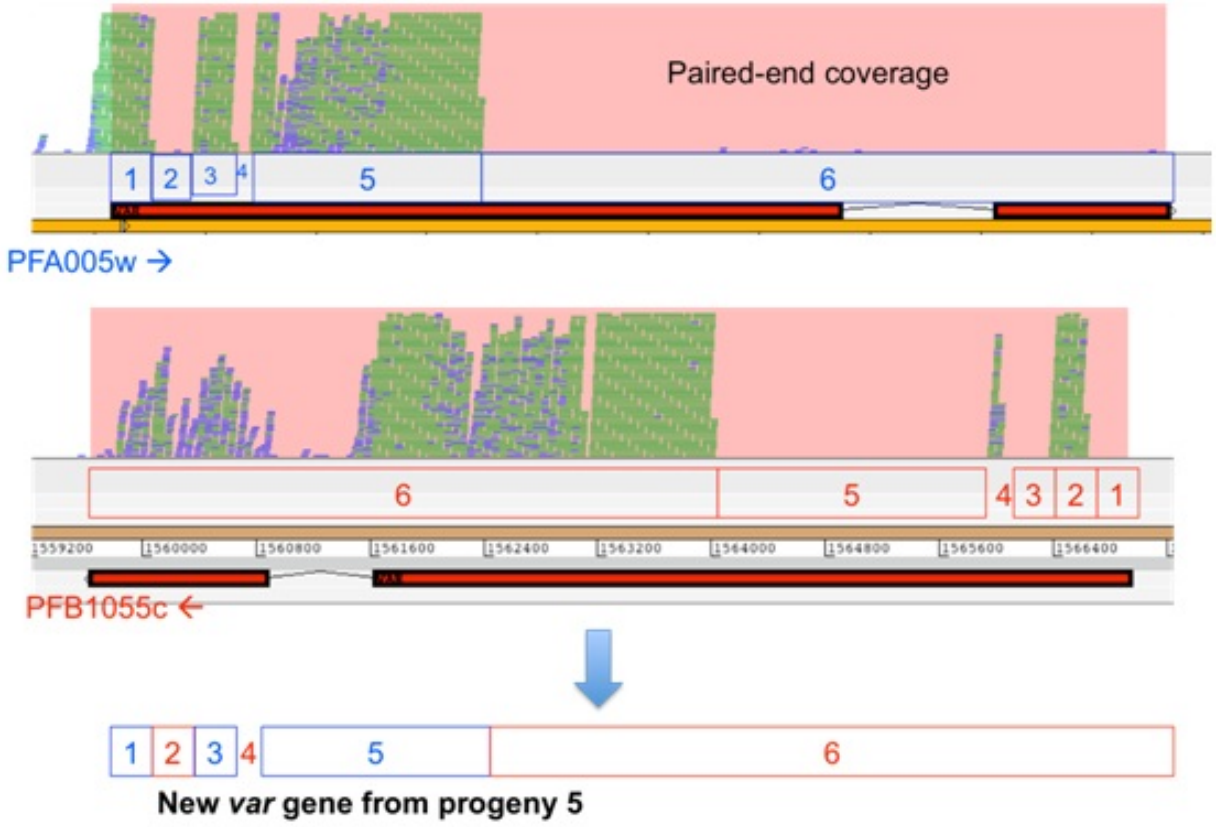

Figure 4.12: Evidence of gene conversion in progeny 5 (X5) from paired end coverage analysis. Description as above (Figure 4.11). Here, six alternating blocks from two *var* genes on chromosomes 1 and 2 of the 3D7 parent were used to generate a new *var* gene.

4.3.3.2 Reads mapping and *de novo* assembly to detect recombination in *var* genes

In order to systematically identify recombination breakpoints in *var* genes, progeny reads were aligned to a combined *var* reference (i.e. a multi-FASTA file containing *var* genes from both parents) allowing for a random placement of non-unique reads to one of the matching positions. Reads from the 3D7 parent were also aligned to the combined *var* reference and used as controls to detect false positive results. In addition to the previously identified two events: *PFA0005w X PFB1055c* and *PFA0765c X PFB0010w* (Table 4.3), a number of putative recombination events were detected (Table 4.4) by looking for read pairs that bridge two *var* genes of the same or different parent. Central *var* gene clusters on chromosomes 4 and 12 were also found in the putative list.

Events found in all or most of the progeny as well as those found in the 3D7 control sample were excluded, as they are likely to be caused by regions of high sequence similarity between *var* genes. The remaining events were further evaluated using *de novo* assembly of raw reads that align to both genes. The two events previously identified in progeny X4 and X5 were confirmed by *de novo* assembly of short reads (Figures 4.13 and 4.14). Lack of evidence from the control samples and formation of contigs with long open reading frames prove that the events are genuine.

The two events previously identified in progeny X4 and X5 (Figures 4.11 and 4.12) were confirmed by *de novo* assembly of short reads as shown in Figures 4.13 and 4.14. Lack of evidence from the control samples and formation of contigs with long open reading frames prove that the events are genuine. Both cases represent non-allelic gene conversions between telomeric *var* genes of chromosomes one and two of the 3D7 parent. Recombinant *var* genes were made of four to six alternating sequence blocks obtained from the parental genes. Block sizes as short as $\sim$200 bp were detected in both events. It was however not possible to determine whether these events were mieotic or mitotic as the parental clones underwent both sexual and asexual development stages. Gene conversion events could thus be a result of mieotic or mitotic non-homologous recombination exchanges.

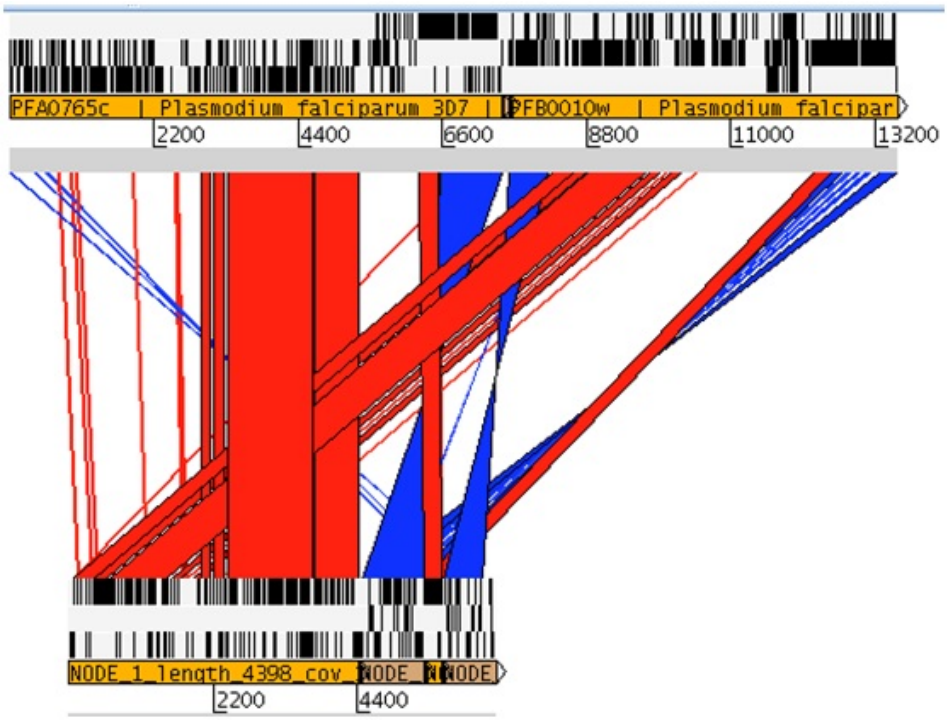

Figure 4.13: *De novo* assembly of reads that map to PFA0765c and PFB0010w in progeny 4 (X4). The top panel shows sequences of the two genes PFA0765c and PFB0010w in the forward strand. The three frames of the forward strand are shown where black lines indicate stop codons and long white regions indicate open reading frames. The bottom panel shows four contigs that were generated by the *de novo* assembly of reads that mapped to PFA0765c and PFB0010w. The red and blue bars indicate matches between sequences from the top and bottom panels (blue bars for reverse complement matches).

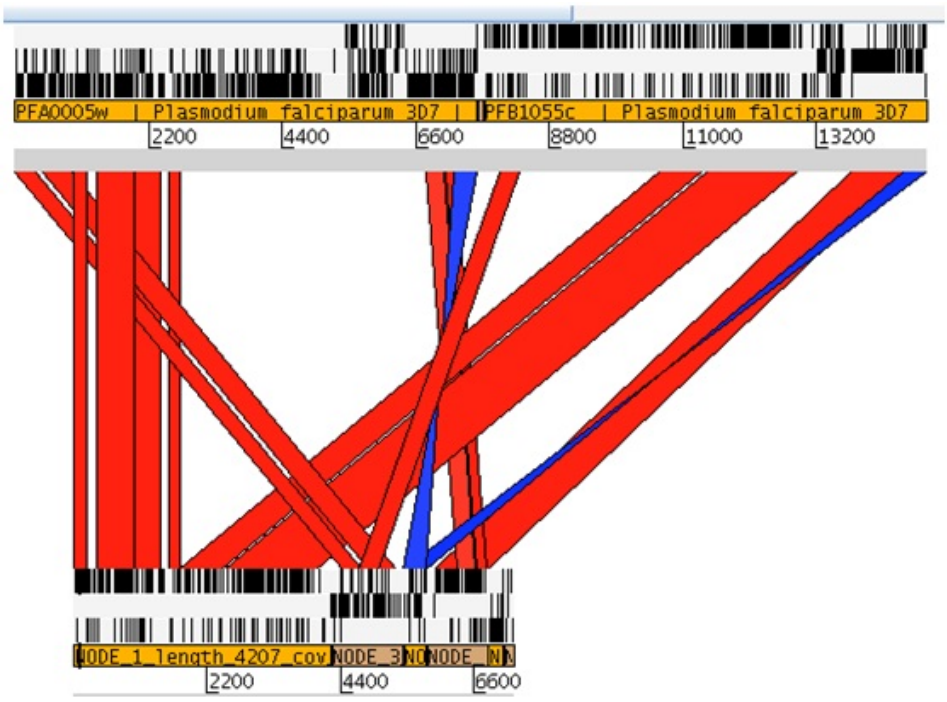

Figure 4.14: *De novo* assembly of reads that map to PFA0005w and PFB1055c in progeny 5 (X5). See description for Figure 4.13

Gene 1	Gene2	X1	X2	X3	X4	X5	3D7
PFA0005w	**PFB1055c**	-	-	-	-	**3295**	-
PFA0765c	**PFB0010w**	-	-	-	**1071**	-	-
PFC0115c	PFD1025w	87	97	-	92	84	-
PFD0140w	PFD0615c	-	-	**216**	-	-	-
PFD0615c	PFD0625c	-	-	**217**	-	-	-
PFD0625c	PFHG_02419	-	-	**302**	-	-	-
PFD0630c	PFHG_02419	-	-	**214**	-	-	-
PFD0635c	PFHG_02419	-	-	**93**	-	-	-
PFD0655w	PFD0995c	-	-	**69**	-	-	-
PFD0995c	PFD1015c	-	-	**129**	-	-	-
PFD1000c	PFD1015c	-	-	**122**	-	-	-
PFD1005c	PFD1015c	-	-	**493**	-	-	-
PFD1235w	PFL1955w	-	**62**	-	-	-	-
PFF1580c	PFF1595c	-	-	-	171	-	-
PF08_0107	PFHG_02419	-	-	77	-	-	-
PFL0005w	PFHG_04770	71	69	-	-	58	-
PFL0005w	PFL0020w	240	-	-	-	64	-
PFL1955w	PFL1960w	-	483	724	379	404	395
PFL1960w	PFHG_02419	-	321	480	228	325	-
PFL1960w	PFL1970w	-	540	690	306	265	136
MAL13P1.1	PF13_0003	83	70	-	-	-	-
PF13_0003	PFHG_03234	**56**	-	-	-	-	-
PFHG_02272	PFHG_04910	-	**68**	-	-	-	-
PFHG_02429	PFHG_04081	94	80	-	-	-	-
PFHG_03234	PFHG_05483	-	**125**	-	-	-	-
PFHG_03671	PFHG_05483	-	406	-	215	-	-
PFHG_03839	PFHG_03999	244	138	-	163	253	-
PFHG_03839	PFHG_04859	242	128	-	206	269	-
PFHG_03999	PFHG_04859	236	204	-	300	260	-
PFHG_04081	PFHG_04368	160	180	-	109	-	-
PFHG_04081	PFHG_04928	105	130	-	101	82	-
PFHG_05483	PFHG_05502	-	**86**	-	-	-	-

Table 4.4: Genes bridged by mate-pairs: Number of reads where mates map to different genes (mismatch ≤ 2). Reads from the 3D7 parent were also aligned to *var* parental genes and used as a control. Gene-pairs that were bridged by mate-pairs in all progeny or in the 3D7 were thus excluded.Gene pairs shown in boldface were recombinant genes that were confirmed by *de novo* assembly as shown in Figure 4.13. Values shown in boldface represent gene pairs that were bridged by a minimum of 50 read-pairs and not common to all progeny.

The remaining potential recombination events shown in Table 4.4 could not be confirmed using *de novo* assembly due to the formation of two or more separate contigs representing individual genes instead of a recombinant gene.

4.4 Discussion

This chapter explored applications of the Illumina sequencing technology to investigate the mechanisms used by parasites to generate diversity in *var* genes. The following conclusions summarise the results of sequence analysis on five progeny from a genetic cross experiment between the 3D7 and HB3 parental clones.

Conclusion 1: *Despite challenges of using very short (54 bp) reads, it was possible to obtain a global view of recombination, even in subtelomeric regions.*

The main challenge in using short reads for progeny sequence analysis was mappability. On average, the number of reads mapped to individual parental genomes is expected to be ∼50% of the total. However, variation from the expected value was observed due to recombination events that favor over-representation of one parent. The lowest percentage of read mapping to the combined reference genome was primarily due to the high sequence identity between core regions of parental chromosomes. A lower proportion of reads aligned to the HB3 reference (26 to 52%) compared to 3D7 potentially due to the incomplete nature of the genome.

Similar results were obtained from the two complementary methods of detecting genome wide crossover events (15 to 21 events). These observations are in agreement with the expected number of meiotic crossover events and previous reports (Jiang et al., 2011). An approach based on detecting SNP dense areas of a progeny may provide a quick overview of large-scale breakpoints. However, shorter fragments of low SNP density may also be caused by lack of coverage as a result of low complexity and the difficulty of reliably calling variants. It is therefore crucial to understand the genomic context of the region and decide whether a region is ambiguous prior to assigning genotypes. While some of such ambiguous regions may be common to all progeny due to inherent sequence features of the genomes, others are specific to a given sample as a result of the issues with the library preparation and the sequencing.

Despite the advantages of using short reads to obtain a nucleotide level resolution of breakpoints, notable limitations were observed in detecting subtelomeric breakpoints with a read length of 54 bp. The synteny between core

regions of 3D7 and HB3 allows reads to be aligned to either parent with mismatches (SNPs) that could be analysed to identify regions inherited from each parent. This approach was found to be adequate for detection of larger crossing over events. However, due to high polymorphism in *var* genes and subtelomeres, aligning short reads within tolerable allowance of mismatches was not possible. A different approach was thus required to detect smaller gene conversion events. For conversion tracts larger than the fragment size, following mates of reads that map near breakpoints was a better way of identifying donor and acceptor regions. Longer reads and larger insert libraries will be needed to reliably detect breakpoints as well as indels in subtelomeric regions where recombination rate is ∼10X higher than core regions of the genome (Taylor et al., 2000c).

Conclusion 2: *This chapter demonstrates the use of short-read sequencing to obtain a detailed resolution of ectopic gene conversion and shuffling of sequence blocks employed by parasites to generate new* var *genes (Figures 4.11 to 4.14)*

In addition to chromosome level breakpoints, short reads were used to identify segments of *var* genes that were inherited by each progeny. Regions of high read coverage when aligned to the 3D7 genome represented segments of the genome that were transferred to the progeny. Low (or zero) coverage on the other hand may imply a genuine absence of those regions in the progeny, implying they were inherited from the HB3 parent. However, lack of coverage could also be a result of low-complexity and repeats that prevent short reads from uniquely aligning in the subtelomeric regions where most *var* genes reside (Gardner et al., 2002). Signatures of recombination within and between parental *var* genes were thus detected using a combination of paired-end coverage inspection and targeted mapping of reads.

Although ectopic gene conversion was reported as a potential mechanism of generating new variants (Frank et al., 2008; Freitas-Junior et al., 2000), previous approaches based on hybridisation of the DBLα domain offered a limited insight on the extent of shuffling over the full length of genes.

Var genes that were involved in recombination via shuffling of sequence blocks were identified in two of the five progeny on chromosomes 1 and 2 of the

3D7 parent. The genes on Chromosome 1 were located adjacent to telomeric-associated repetitive elements (TAREs), which may be involved in initiating homologous recombination. This finding is consistent with previous studies that associated repeats and low complexity regions with hot-spots and elevated rates of recombination (Jiang et al., 2011). Random clustering of telomeres, during both the sexual and asexual stages of the parasite, is believed to provide a means for enhanced shuffling of sequence blocks in *var* genes (Freitas-Junior et al., 2000).

The genes that were identified to be recombining were members of the group Type A in both progeny. In addition, the genes were located on the 3D7 parent and between chromosomes 1 and 2. Recombination in *var* genes is believed to occur within defined hierarchies, such that members of Type A genes recombine more often with other type A than non-Type A genes (Bull et al., 2008; Kraemer et al., 2007). Our findings are also consistent with this observation, and also shed a new light on the extent of recombination via shuffling of blocks. However, understanding mechanisms of such frequent events (up to six recombining blocks were detected between two *var* genes in progeny 5) at the molecular level needs further investigation.

Central *var* clusters are known to be involved in spontaneous recombination (Deitsch et al., 1999), and may be ideal regions for gene conversion due to their organisation and high sequence similarity. We thus expected to see central *var* gene clusters in our putative list of recombinant genes (Table 4.4). However, it was not possible to reliably confirm these recombination events due to limitations imposed by the short read length. In addition, the fragment size (200-300 bp) was shorter than identical sequence fragments (repeated sequences) that are characteristic of these genes as described in Chapter 2.

In terms of absolute numbers, fewer than expected non-parental *var* genes were identified to be involved in recombination (only two of the progeny *var* genes were confirmed to have signatures of recombination) compared to an earlier study by Taylor and colleagues (Taylor et al., 2000a) that identified 24 events. The short read length (54 bp) and small fragment size of the library (200 bp) were major limiting factors. It was also not possible to *de novo* assemble *var* genes using approaches described in Chapter 3. Additional potential reasons

include low complexity, high percent identity between parental genomes and poor sequence quality resulting in unreliable read mapping.