Chapter 5

Assembly of *var* genes from clinical samples

5.1 Introduction

The study of *var* genes especially in clinical samples taken directly from patients will significantly improve our understanding of their diversity and evolutionary history. Previous studies that involve sequence analysis have mainly looked at diversity and expression of *var* genes on a limited number of clinical isolates. A number of studies have also shown association of sequence features with specific disease phenotypes (Ariey et al., 2001; Bull et al., 2005; Cham et al., 2010; Falk et al., 2009; Jensen, 2004; Kaestli et al., 2004, 2006; Kalmbach et al., 2010; Kirchgatter and Portilo, 2002; Kyriacou et al., 2006; Lavstsen et al., 2005; Montgomery et al., 2007; Nielsen et al., 2002; Normark et al., 2007; Rottmann et al., 2006).

However, all except two of the previous studies that used a sequence analysis approach of *var* genes have focused on the DBLα region (Kraemer et al., 2007; Rask et al., 2010). Although it was possible to accurately classify *var* genes into existing groups and make associations with disease severity using sequences taken from the DBLα domain, a large proportion of the *var* repertoire is still excluded. In the first comparative study to use full length genes, Kraemer and colleagues (Kraemer et al., 2007) analysed a near complete *var* repertoire of

three culture-adapted samples 3D7, IT and HB3. While 3D7 has a complete set of genes (i.e. the expected 60 protein coding genes), the other two had an incomplete repertoire. The results confirmed the presence of extreme diversity in *var* genes with only three genes (var1CSA, var2CSA and Type 3 *var*) showing a higher degree of conservation in all the three genomes. The second study by Rask and colleagues (Rask et al., 2010) used an additional four genomes (DD2, PFCLIN, RAJ116 and IGH) to provide a new definition of domain boundaries and identify recombination hotspots. A combination of phylogenetic and iterative homology block detection methods was used to define 628 homology blocks that could represent *var* genes with a better resolution than existing domain boundaries. However, due to the limits on the number and diversity of sequences used (311), these blocks may not accurately represent *var* genes in natural populations.

Understanding the order of sequence blocks and mosaic domains is of great importance. Recent studies have associated specific domain cassettes (as defined by Rask and colleagues (Rask et al., 2010)) with disease severity and a rosetting phenotype (Avril et al., 2012; Claessens et al., 2012; Lavstsen et al., 2012). Such association studies may facilitate the discovery of important antigens that could be used as potential vaccine targets for severe malaria. Obtaining full-length sequence information on *var* genes may thus be a step forward in such attempts. Moreover, lack of full-length information continues to be a major roadblock in understanding *var* gene diversity.

The focus of this thesis was to develop a new approach for the assembly of *var* genes from short reads of second generation sequencing platforms. As described in previous chapters, assembly of *var* genes using existing tools was not practical due to high polymorphism and the mosaic nature of *var* genes (Chapter 2). An iterative assembly approach that takes advantage of the inherent mosaicism in *var* genes was thus developed (Chapter 3) and evaluated on culture-adapted and a small number of clinical samples (50 samples). Here, the new approach is applied on a larger number of clinical samples.

Assembly of clinical samples adds another layer of complexity due to a number of difficulties associated with the quality of the input DNA and raw sequence data. Contamination with host DNA could result in a lower amount

of starting material, and therefore low yield of sequence data. In addition, systematic errors and bias towards certain sequence features due to the sequencing chemistry may affect data quality and result in reads with errors. Multiple genotypes circulating in a single individual contain highly similar as well as polymorphic haplotypes that affect the structure of the de Bruijn graph and quality of the resulting assembly. Although some of the challenges are being addressed by improvements in library production protocols used for sample preparation and sequencing (Oyola et al., 2013), the effect of poor quality data and uneven coverage still poses a unique challenge in assembly of *var* genes.

In this chapter, the iterative assembly approach described in Chapter 3 was applied to a larger collection of clinical isolates consisting of 743 samples taken from Africa, South East Asia and South America.

5.2 Methods

5.2.1 Sequence data

Clinical samples of *P. falciparum* were obtained from the Plasmodium Genome Variation (PGV) project at the Sanger Institutes malaria programme (`www.sanger.ac.uk/research/areas/malariaprogramme/`). Methods of sample preparation and the sequencing technology have seen a significant improvement over the last few years. Samples sequenced during the early days of the project were especially of low yield and poor quality with shorter read lengths of 37 and 54 bp. It was therefore decided to exclude samples that had a read length of below 76 bp. Samples that were not prepared using the PCR-free protocol (Chapter 2) were also excluded.

5.2.2 Initial Motifs and iterative assembly of clinical samples

The assembly work flow for a large number of clinical samples is illustrated in Figure 5.1. As described in Chapter 3, a total of 50 clinical samples were assembled for 20 iterations to evaluate the iterative assembly approach developed in this thesis (Chapter 3).

To obtain the maximum number of seed motifs for the clinical sample assembly, the assembly of the 50 samples was repeated for another iteration. Initial motifs to assemble 743 samples were thus obtained from the 21st iteration by translating contigs with the DBLα tag (*var*-contigs). Although the open reading frame (ORF) that contains the DBLα tag could be used to identify the correct reading frame, presence of frame-shifts meant some of the long DBLα may be excluded. *Var*-contigs were thus translated in all the six frames to generate the initial set of shared motifs. Once started with these motifs, the assembly of the clinical samples was repeated for three iterations, with each iteration involving sub-iterations of scaffolding and extension. Seed contigs were generated by optimising *k-mer* sizes for the assembly in two categories. For reads with a length of 76 bp, *k-mer* sizes of 51, 65 and 71 were used. For reads of length 100 bp and above, an additional *k-mer* size of 81 was used. Assembly results with different *k-mer* values were compared based on the number and

N50 sizes of *var*-contigs. For each sample the *k-mer* value that resulted in the highest number of contigs with DBLα and the highest N50 value was chosen to generate seed contigs. A final list of motifs (length=10 aa) was generated from *var*-contigs of the third iteration by considering a single frame with the longest Open Reading Frame (ORF) in either the forward or reverse strand. If the longest frame does not contain the DBLα tag, ORFs longer than 300 aa from the three frames on the strand of choice were chosen.

5.2.3 QC and filtering

Assembly quality was measured using the number of *var*-contigs, sum of *var*-contigs, N50 and largest contig sizes. The number of *var*-contigs was used as a measure of repertoire completeness for each assembled sample. Initially, samples that had below 30 *var*-contigs were excluded as they were found to be a result of low yield or poor quality sequence data. Samples with more than 70 *var*-contigs were defined as having multiple infections. Initially, these cutoff values were determined based on the expected number of *var* genes ($\sim$60) from previous studies on laboratory clones and clinical isolates. Additional quality checks include comparing assembly statistics of *var*-contigs with expected values from *var* genes of the reference genome 3D7. Using the method proposed by Bull et al. (2007), *var*-contigs were grouped into one of the six groups. The count of contigs in each group was compared with that of assembly results from the 50 samples (Chapter 3) and the reference genome 3D7. The most reliable method of checking assembly quality of *var*-contigs would be to compare with *var* genes of the reference genome. However, as described in the previous chapters, such approaches are not practical for the highly polymorphic *var* gene family. One approach adopted in my thesis to overcome this limitation was to look at ORFs instead of nucleotide sequences of contigs. In addition to providing a better measure of contiguity, using ORFs could minimize the effect of low complexity regions in introns and upstream and downstream regions of *var*-contigs. ORFs with a minimum length of 300 aa were obtained from each *var*-contig and stored as separate entries. For example, two ORFs of a *var*-contig (VAR1) from Sample1 were represented as follows:

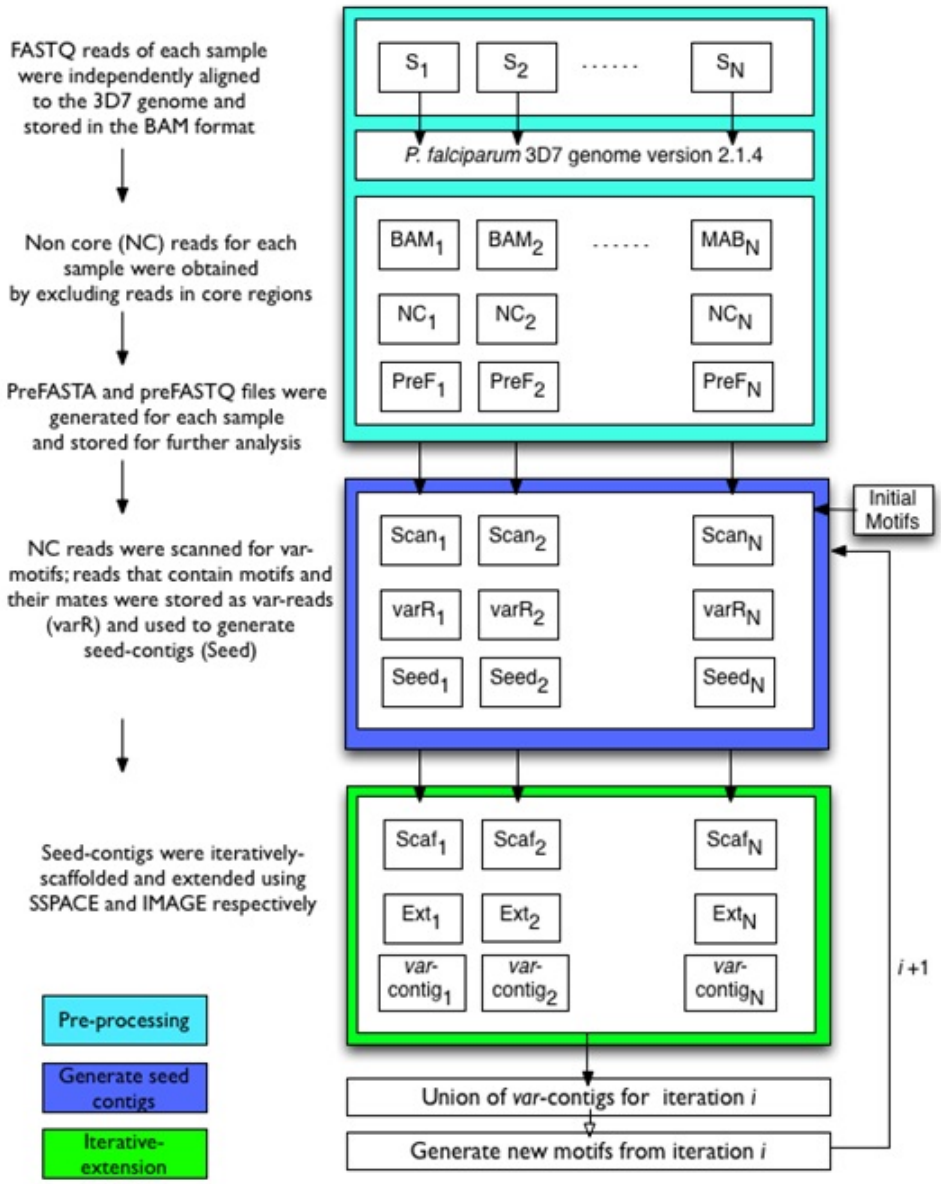

Figure 5.1: Iterative assembly work flow for *var* genes in clinical samples. Processes of the three stages of the *var* assembly are shown in boxes with cyan, blue and green backgrounds. Decisions on further iterations are made based on the quality of *var*-contigs from the current iteration. Assembly results of N=743 clinical samples are presented in this chapter (i=3).

Sample1_VAR1.ORF1
Sample1_VAR1.ORF2

The ratio of ORFs to *var*-contigs was used as a further quality control measure. Ideally, a full length *var*-contig should result in two ORFs representing each exon. It is however expected to find one ORF as most *var*-contigs may only capture the first exon. Introns and second exons are likely to cause ambiguities due to the high A+T content, and therefore generate smaller contigs that do not contain the DBLα domain.

5.2.4 Similarity between *var*-contigs

Similarity and relatedness of *var*-contig were analysed on three levels. Initially, we intended to use short-motifs generated from *var*-contigs during the assembly process. However, as the quality of contigs improved, it was possible to use longer matches using Pmatch (for perfect matches) and BLAST (allowing mismatches) as described below.

5.2.4.1 Pmatch analysis

Perfectly matching sequences were detected between any two *var*-contigs using Pmatch (minimum length=14 aa). Pmatch is written in C (Richard Durbin, Sanger Institute; unpublished) and rapidly identifies pairwise identical matches given two multi-fasta files of amino acid sequences. Amino acid translations of *var*-contigs were used as both query and subject for the pmatch analysis (i.e. all against all matching).

Var-contigs were translated by choosing the longest ORF in the strand where the DBLα tag was found. As mentioned in the previous section, if the longest ORF did not contain the DBLα tag, ORFs above 300 amino acids on the three frames of the chosen strand were concatenated. This method of detecting ORFs by jumping across the three frames of the main strand (i.e. the strand with DBLα) minimised the risk of missing ORFs due to frame shifts caused by misassemblies. Although this approach may also introduce the possibility of chimeric ORFs, the minimum length requirement of 300 amino acids was used

to account for such effect. The output of Pmatch required up to ∼250 Gb of storage disc space for a single analysis. It was thus necessary to convert the results to a simple motif-sharing format as defined in Chapter 3.

5.2.4.2 BLAST

Initially, nucleotide and amino acid BLAST (Altschul et al., 1990) databases of all *var*-contigs were generated using *formatdb*. In order to speed up the matching process, *var*-contigs of each sample were separately stored in a file and used as query during the BLAST search (*blastall -p blastp/blastn -e 0.001 -F F -m 8*). The output was compressed (*gzip -9*) prior to storage for further analysis.

5.2.4.3 Defining similarity between *var*-contigs and repertoires

Mosaic blocks of *var* genes result in a fragmented alignment profile between *var*-contigs. Similarity between two *var*-contigs was thus defined as a function of the total number of positions matched (i.e. the proportion of identical positions between the two *var*-contigs to the total length aligned). Similarity between *var* repertories was then computed as the average of all pairwise similarity values.

Similarity between two *var*-contigs V_1 and V_2 with n different blocks of matching sequences $m_1 \ldots m_n$ is defined as:

$$S_{v_1,v_2} = \frac{2 * \sum m_i}{L_1 + L_2}; \tag{5.1}$$

where m_i is length of match i, $i \in$ [1,n];
and L_1 L_2 are the full lengths of the two *var* contigs

5.2.5 Network analysis and clustering

Analysis of social networks was first used in *var* gene studies by Bull and colleagues (Bull et al., 2008). It was shown to be a better approach to study population structures and recombination hierarchies in the DBlα region than the phylogenetic tree based approaches which were shown to be impractical due to higher rates of recombination (Barry et al., 2007). The results of BLAST matching between *var*-contigs were processed to generate a graph of connected

var-contigs. The first set of samples that have a single infection (i.e. samples that contain below 70 *var*-contigs) were analysed. A graph was constructed by considering *var*-contigs as nodes. An edge between two nodes was added to the graph if two contigs have a match that fulfills the minimum identity and length requirements (eg. 99% and 1000 aa respectively for amino acid networks shown in the results section). A pairwise similarity index for two *var*-contigs was computed as described in the previous section. Each edge was thus updated according to weights obtained from the pairwise similarity values. These values range from 0 (no match) to 1 (two contigs are identical). A customised script (*blast*2*Gexf.pl*) was written to convert BLAST output files to the Graph Exchange XML Format (GEXF) (`http://gexf.net/format/`). First developed at the Gephi project in 2007, the GEXF format is widely used in representing a complex graph structure in terms of nodes and edges of a graph. In addition, a number of attributes such as weight and colour of nodes could be included in the graph file. Node colours were defined according to country of origin. In addition to visualising clusters, the Markov Clustering Algorithm (`http://micans.org/mcl/`) was used to generate clusters of *var*-contigs that share identical sequence blocks (Inflation parameter was tested at I=0.2, 1.2, 2, 4 and 6; final choice=1.2).

5.3 Results

5.3.1 Samples and sequence data

A total of 725 samples passed the selection criteria (i.e. samples prepared using PCR-free protocol and with a minimum read length 76 bp) at the beginning of this analysis (Figure 5.2, Table 5.1). These samples represented 13 countries from West Africa, East Africa, South East Asia and South America. The majority of samples came from The Gambia, Ghana and Cambodia. An overview of samples used in this chapter and their geographical origins are shown in Figure 5.2 and Table 5.1.

Figure 5.2: A global map of clinical samples used in this chapter. 725 samples were obtained from 13 countries representing West Africa (WA), East Africa (EA), South East Asia (SEA) and South America (SA). In addition to the 725 samples with known countries of origin, 18 samples from various countries that became available during the course of the study were also included.

5.3.2 Initial Motifs

Assembly of the 50 clinical samples of Chapter 3 plus an additional iteration (10 countries; 21 iterations) resulted in a total of 10.7×10^6 motifs from *var*-contigs.

These motifs were used to initiate the assembly process on the 743 samples (Table 5.1).

5.3.3 Iterations and additional motifs

The initial set of motifs were generated using amino acid translations in all the six frames. Although it was important to have a large number of motifs during assembly in order to increase the efficiency of the iterative process, such a high number is not required for the final analysis of shared motifs, as it will lead to unnecessary redundancy and an inflated count of overlapping genes.

At the end of the third iteration, a final list of 3.5 x10^6 motifs (10 aa long sequences as described in Chapter 3) were obtained from *var*-contigs using the longest ORF with the DBLα tag. Compared to generating motifs from amino acid translations of all the six frames, this approach reduced the number of motifs by ∼70%.

ID	Country	Region	#samples
PA	Gambia	WA	168
PF	Ghana	WA	122
PM	Mali	WA	32
PK	Burkina Faso	WA	3
PT	Malawi	EA	55
PC	Kenya	EA	25
PE	Tanzania	EA	15
PR	Bangladesh	SEA	3
PD	Thailand	SEA	82
PH	Cambodia	SEA	191
PN	Papua New Guinea	SEA	7
PV	Vietnam	SEA	11
PP	Peru	SA	11
Others			18
Total			**743**

Table 5.1: Samples used for initial assembly of *var* genes in clinical samples. A total of 743 samples were obtained from 13 countries as shown in Figure 5.1 (725 samples). Additional 18 samples from various countries became available during the course of the project and were also included.

5.3.4 Results of the initial assembly

Assembly results for the 743 samples were summarised by four commonly used assembly metrics: sum of contigs, N50 size of contigs, number of contigs and largest contig size (Figure 5.3). The variation observed in the quality of each assembly is shown by the distribution of values for these four measures. The sum and N50 of *var*-scaffolds show a wider distribution range reflecting the poor quality assembly on the extreme left side of the distribution as well as a mixture of genotypes (multiple infections) on the far right end of the distribution. Conversely, the number of *var*-scaffolds and the largest scaffold size were narrowly distributed with median values of ∼60 and ∼10 kb respectively, reflecting a high quality assembly in terms of repertoire completeness.

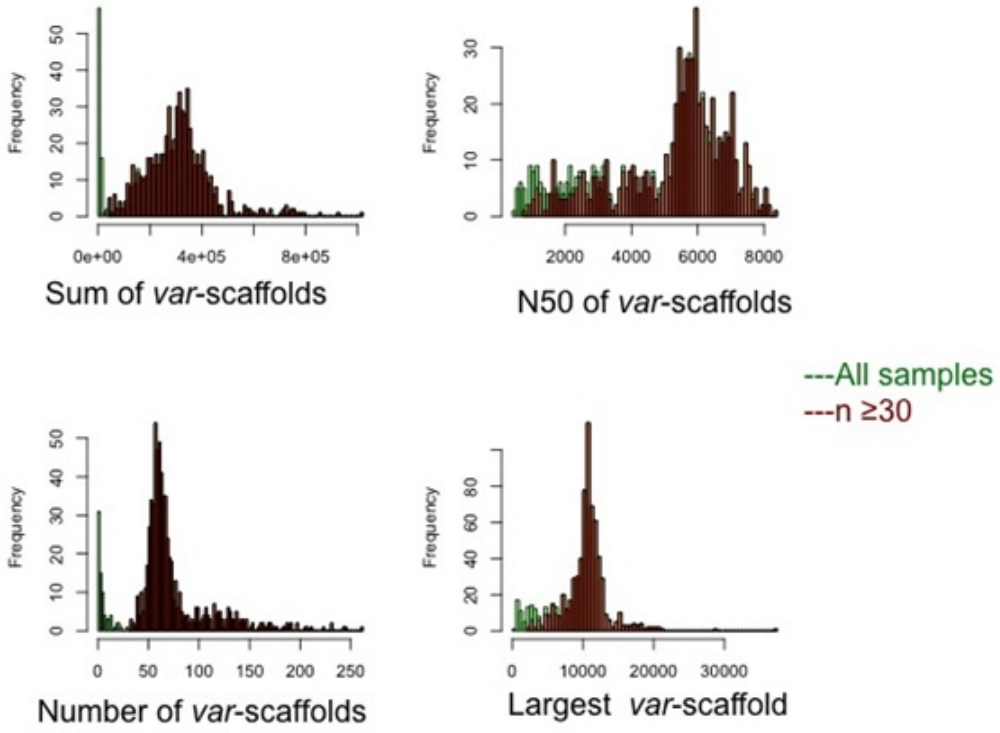

Figure 5.3: Assembly stats of the initial 743 samples. Green shades represent all samples while dark red shades represent samples with above 30 *var*-contigs. Sum of scaffolds, N50 and largest scaffold sizes were measured in base pairs (bp). An additional summary of the four measures is shown in Table 5.2.

In addition, summary of assembly results were shown by the range, mean and median values of five assembly statistics (Table 5.2). Sum of *var*-contigs was used as a measure of overall coverage of the *var* repertoire. The N50 and number of *var*-contigs measure the contiguity and repertoire completeness of *var*-contigs respectively. Mean values for the sum of *var*-contigs, N50 size and number of *var*-contigs for the 743 samples (∼351 kb, ∼5 kb and 68 respectively) revealed an overall highly representative assembly. The initial histogram plots of all 743 samples were shown in green bars in Figure 5.3. A total of 647 samples had above 30 *var*-contigs (i.e. 647 of the 725 samples of interest). The remaining 78 samples had a poor quality assembly with as few as one contig containing the DBLα tag. The count of non-core reads was investigated to find a reason for such fewer number of *var*-contigs in the 78 samples. Overall, there was a positive correlation between non-core read count and the four assembly measures (R^2 = ∼0.5; p<0.0001). The number of non-core reads was also noticeably low for the 78 samples (Figure 5.4 E, F). Samples with the least non-core read count came either from very recent multiplexed libraries (eg. PH0553-C, PH0581-C, PF0539-C had less than 300,000 read-pairs) or libraries that were sequenced at the beginning of the project (eg. PP0011-C, PF0007-C, PK0032-C and PC0034-C had over 1 Million read-pairs but with poor read quality). Multiplexed libraries were observed to generate inconsistent yield within the different samples that are sequenced in one lane (Magnus Manske, personal communication and preliminary assessment of recent multiplexed libraries). Conversely, sample PC0034-C had the fourth largest number of non-core reads (∼75% of total reads) suggesting issues with data quality instead of yield. Further investigation revealed unusually long insert sizes and a large number of duplicates (∼8% of total reads) and reads where the mate aligns to a different chromosome (∼10% of total reads).

A closer look at the assembly results was obtained by breaking the analysis down to regions (Figure 5.5) and countries (Figure 5.6). The total number of bases in each assembly provided a measure of how well the *var* repertoire is covered. Sum of *var*-contigs for each region revealed that samples from West Africa and East Africa had the largest range (733 bp to 835 kb and 1.1 kb to 640

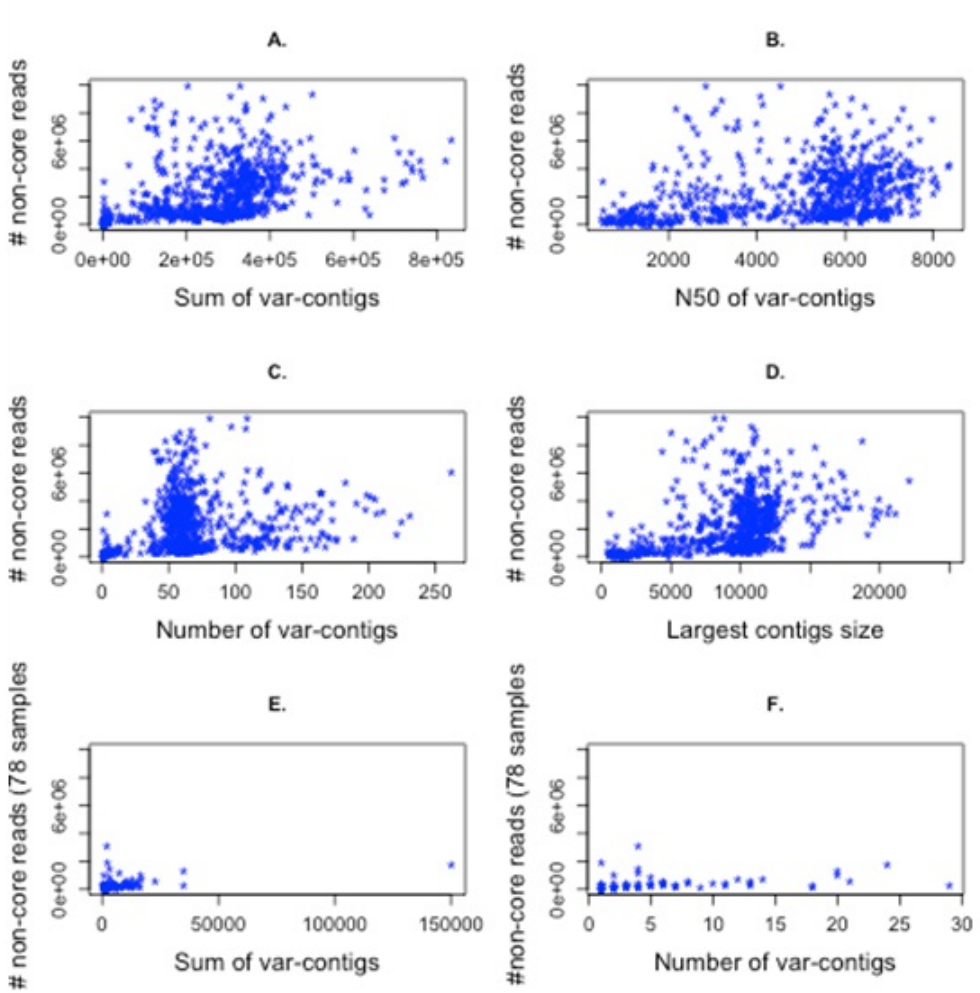

Figure 5.4: Scatter plots of non-core read counts with the four assembly statistics. **A-D)**. Sum of *var*-contigs, N50 contig size, number of *var*-contigs and Largest contig size for all samples. **E-F)**. Sum and number of *var*-contigs are separately shown for samples with less than 30 *var*-contigs.

	Min	Median	Mean	Max
Sum(bp)	477	288,500	350,800	835,100
N50(bp)	477	5,599	4,893	8,362
Num. contigs	1	61	68	262
Largest(bp)	477	10,420	9,705	37,040
N-count(bp)	0	2	403	15,010

Table 5.2: Summary of assembly results for the initial 743 samples. A graphical representation of the Sum, N50, Number of *var*-contigs and Largest contig size is shown in the four histograms of Figure 5.3.

kb respectively) compared to samples from South East Asia and South America. At first sight this may appear to be due to the large number of samples from West Africa (n=325) and East Africa (n=95). However, the narrow distribution in South East Asia can not be accounted for as they have a comparable number of samples (n=294). It is therefore likely that the distribution of sum of contigs as well as *var*-scaffolds is indicative of multiplicity of infection (MOI). West and East African populations were found to display higher values of multiplicity of infection, with up to a five-fold increase in the number of *var*-scaffolds (Figure 5.6). In addition, a visual inspection of aligned reads over the MSP1 gene for samples with the highest number of *var*-contigs confirmed more than one haplotype (Appendix B, Figure B-4). Samples that contain less than 30 *var*-contigs were excluded from further analysis in this chapter. However, they will be included in the future when improving the assembly by, for example, using additional iterative steps.

5.3.5 Initial quality control steps

Quality of assembled contigs was initially assessed using three approaches: count of ambiguous (unknown) bases, size of ORFs, and distribution of *var*-contigs in to the six groups (Bull et al., 2007).

Number of 'N's

Firstly, the total count of 'N's in each sample (i.e. number of gaps in the sum of *var*-scaffolds) was considered and found to be extremely low (median=2,

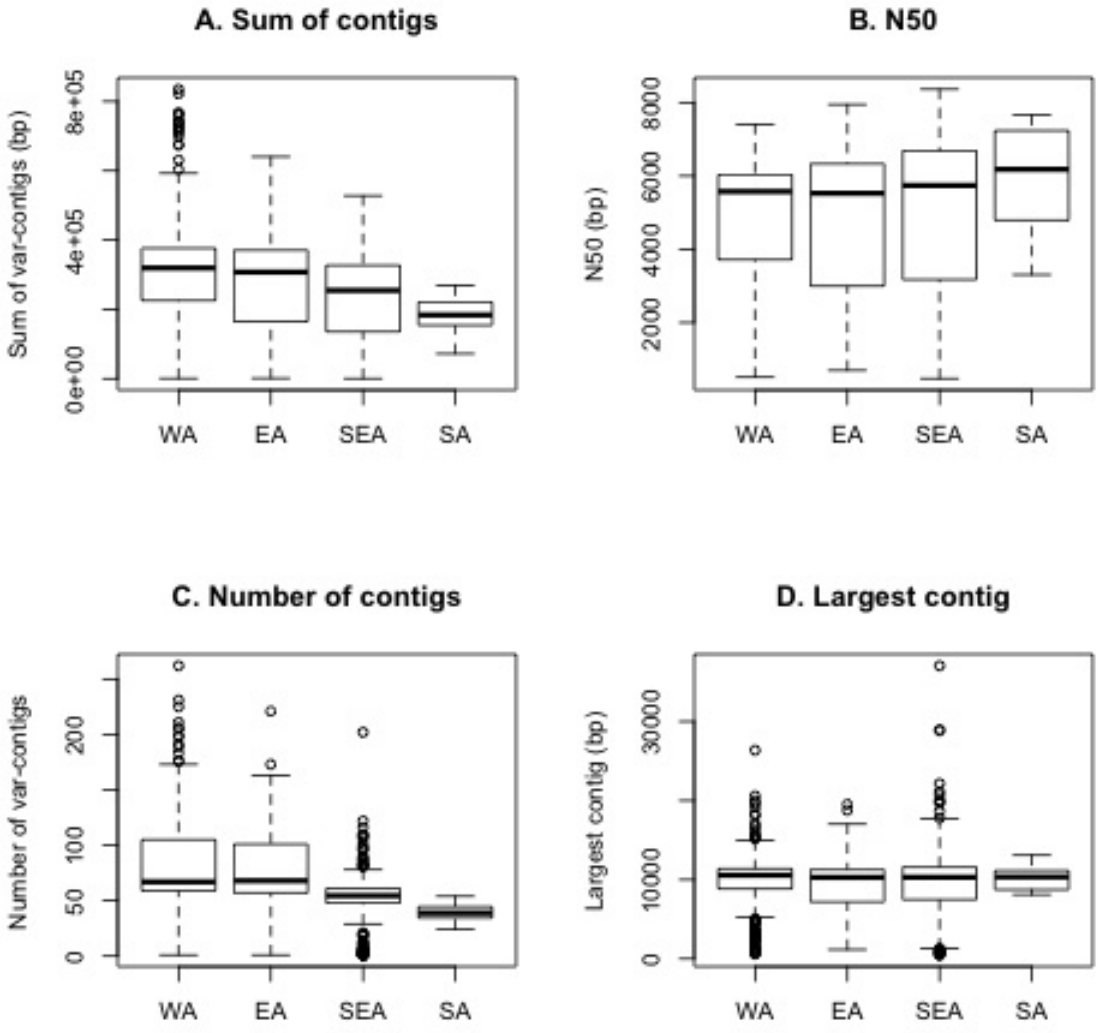

Figure 5.5: Box plots showing assembly statistics by geographical region. The four statistics were separately shown for West Africa, East Africa, South East Asia and South America. Box limits represent median, first and third quantiles; whiskers represent the upper and lower bounds while outliers are shown by the dots. A). Sum of contigs represents the total number of bases in *var*-contigs for the four regions B). N50 contig size distribution of *var*-contigs was $\sim$5 kb on average and consistent across the four regions. C). The number of *var*-contigs showed a similar pattern of variation with West African samples displaying a higher degree of variability compared to South East Asia and South American samples.

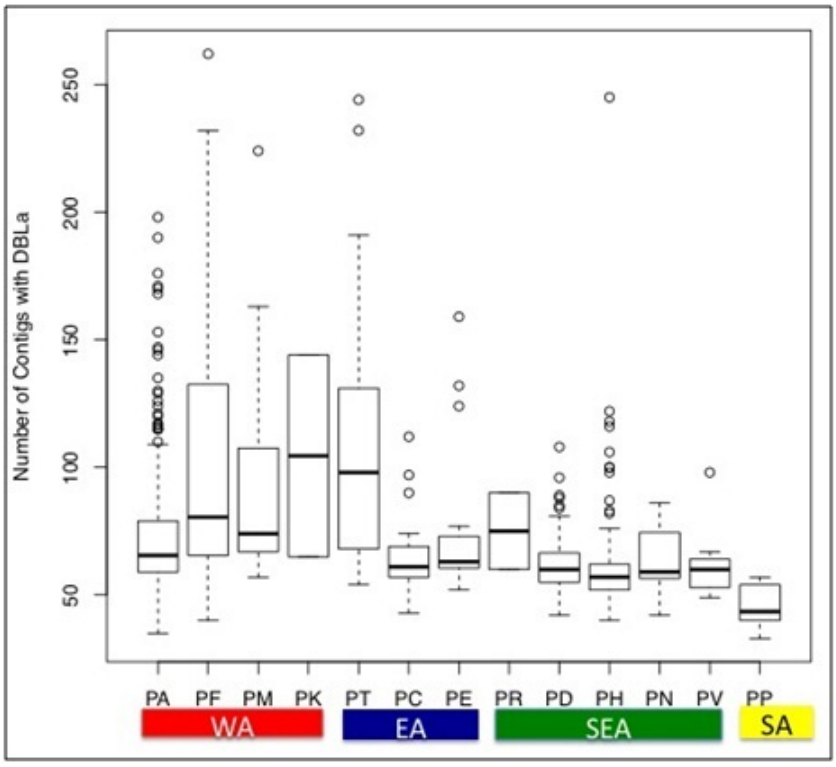

Figure 5.6: Number of *var*-contigs by country of origin. A better resolution on the distribution of the number of *var*-contigs is shown using box plots on a country level. Box limits represent median, first and third quantiles; whiskers represent the upper and lower bounds while outliers are shown by the dots. West African (WA) and East African (EA) samples had higher variability in the number of *var*-contigs than samples from South East Asia (SEA) and South America (SA).

mean=∼200). Together with the high N50 values (mean=∼5 kb), the fewer gaps observed in the assembly are indicative of a higher level of contiguity. The top ten samples with the highest number of Ns were found in The Gambia. However, these samples also had above 120 *var*-scaffolds and a sum of *var*-scaffolds between ∼500 kb and ∼840 kb suggesting the presence of multiple genotypes in the samples. The size of the largest *var*-scaffold was between ∼9 kb and ∼13 kb. In order to make sure the largest contigs are segregated by individual genotypes instead of creating false joins, a further quality cheek was conducted by investigating ORFs.

Size of ORFs

Secondly, in order to evaluate the effects of misassembly on assembly contiguity, the number and size of ORFs was examined. The number of ORFs for each sample was expected to be higher than the number of *var*-contigs (or *var*-scaffolds; but the term *var*-contigs is used here after in this section for simplicity) as *var*-contigs may have multiple ORF entries. If the ORF of a given *var*-contig is not interrupted by stop codons due to false joins that result in frame shifts, the upper limit for the number of ORFs is expected to be twice the number of contigs. A total of 51,140 ORFs were obtained with a minimum length of 300 amino acids. The ratio of ORFs to *var*-contigs was expected to be between one and two for samples with a good quality assembly. A ratio lower than one indicates that most contigs are shorter than ∼900bp. Conversely, a ratio of above two is a sign of frame-shifts as a result of potential mis-assembly. The overall ratio of ORFs (n=51,140) to *var*-contigs (n=50,131) was nearly one as the assembly process mainly captures exon 1 of the *var* repertoire. The sum of ORFs was equivalent to ∼91% of the sum of *var*-contigs (∼205 Mb). The remaining 9% of sequence is due to UTRs, introns and exon 2 sequences. N50 size of ORFs (1,761 aa) was also comparable with the N50 size of *var*-contigs (5,705 bp). The size distribution of ORFs from the assembled clinical sample was also comparable with ORFs from 3D7 and IT genomes (Figure 5.7).

In addition to the overall ORF distribution for all *var*-contigs, a closer look at the ratio of ORFs to *var*-contigs for each sample revealed that two samples PA0106 and PA0107 had the highest number of ORFs (170 and 156 respectively),

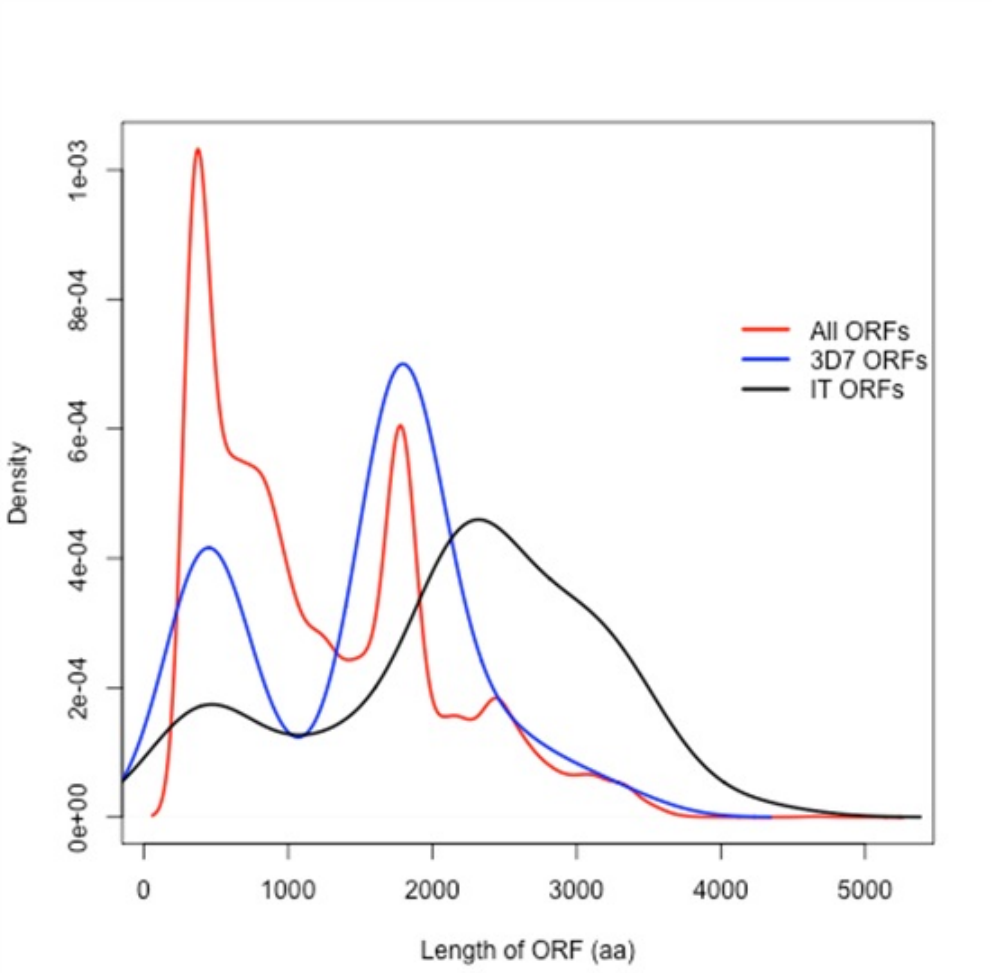

Figure 5.7: Density plots of ORF sizes for clinical samples, 3D7 and IT. The size distribution of ∼50,000 ORFs (red) is shown together with two culture-adapted samples 3D7 and IT with complete repertoires containing 83 and 74 ORFs respectively. The mean ORF sizes were 1217, 1484 and 2168 for all, 3D7 and IT respectively.

although the number of *var*-contigs was 59 and 56 respectively. The remaining samples showed no evidence of excessive frame shifts as the ratio of ORFs to *var*-contigs was within the expected range of one and two.

Grouping var-contigs

Finally, the number of *var*-contigs that are represented by the six groups (as defined by Bull et al. (2007)) followed a similar distribution as that of the 50 samples (Chapter 3) and the three culture-adapted samples 3D7, IT and HB3 where the majority of the genes fall into group 4 (Figure 5.8).

Taken together, these results confirm that the contigs and scaffolds generated from clinical samples were of a high quality.

5.3.6 A first look at the motif sharing *var*-contigs

Motifs generated from *var*-contigs of the third assembly iteration revealed that ∼40% of the total were unique to single samples. The remaining motifs were shared by a minimum of two samples with a heavily right-tailed distribution (Figure 5.9A).

5.3.7 Using full length sequences

5.3.7.1 Pmatch

The pmatch output file was converted to a shared-motif format, revealing perfect amino acid matches of length 14 to 3,404 aa. A large proportion (∼98%) of shared motifs were between 14 and 100 aa (Figure 5.10), and shared by the majority of *var*-contigs (Figure 5.11).

However, unexpected long perfect matches of length above 1,000 aa were also observed (∼600 motifs). Interestingly, these long motifs were shared between *var*-contigs of the same population as well as different populations. The longest motif shared by samples from different countries was 3,404 aa long and found in three samples (PA0036, PA0020 and PH0136), two from The Gambia and one from Cambodia.

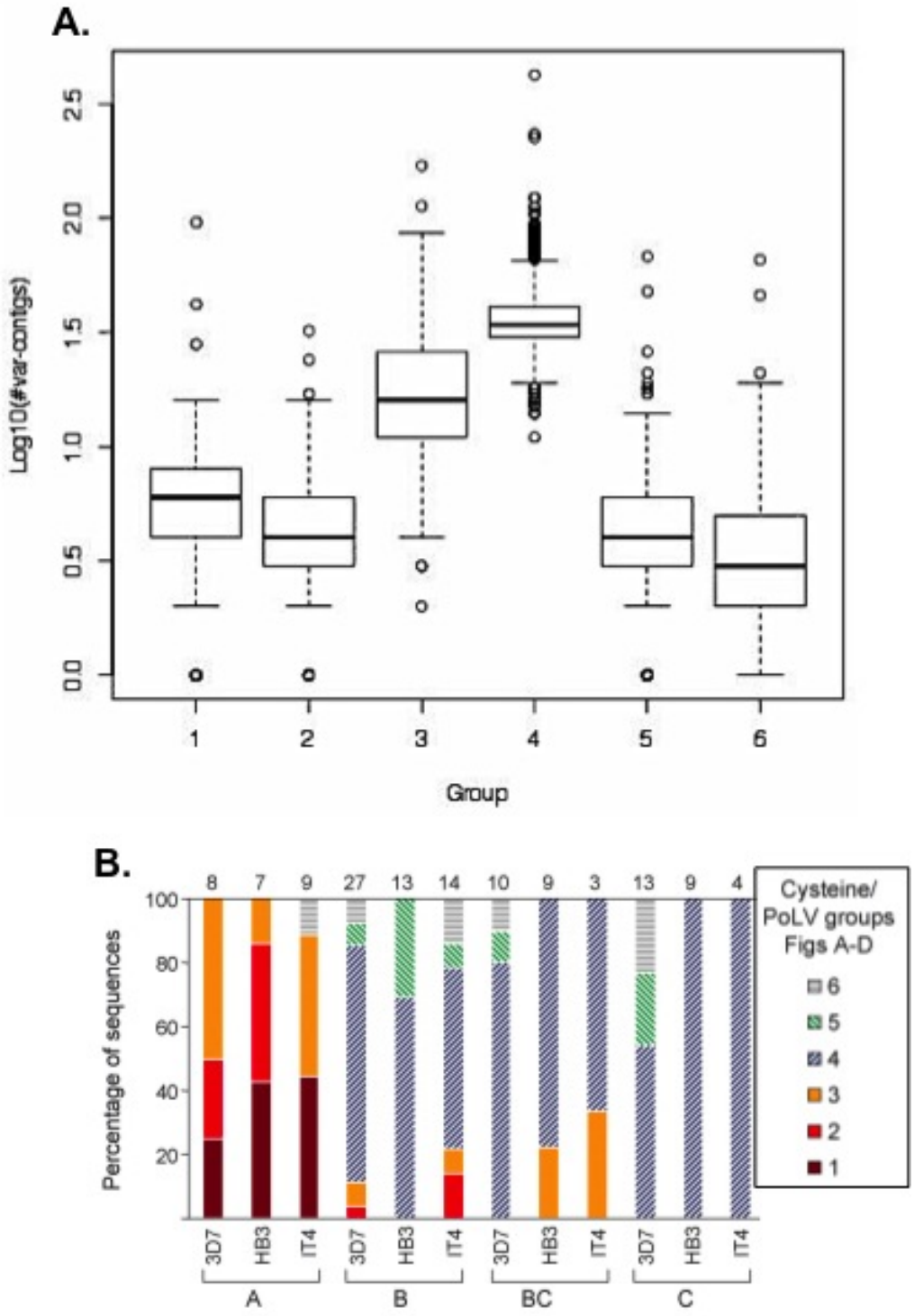

Figure 5.8: Grouping ∼50,000 *var*-contigs using the method proposed by Bull et al. (2007). A) Box plots show the distribution of *var*-contigs in the six groups. B) Correlation of the six groups with existing classification (A, B, C, BC) based on three culture-adapted samples (Adapted from Bull et al. (2007)).

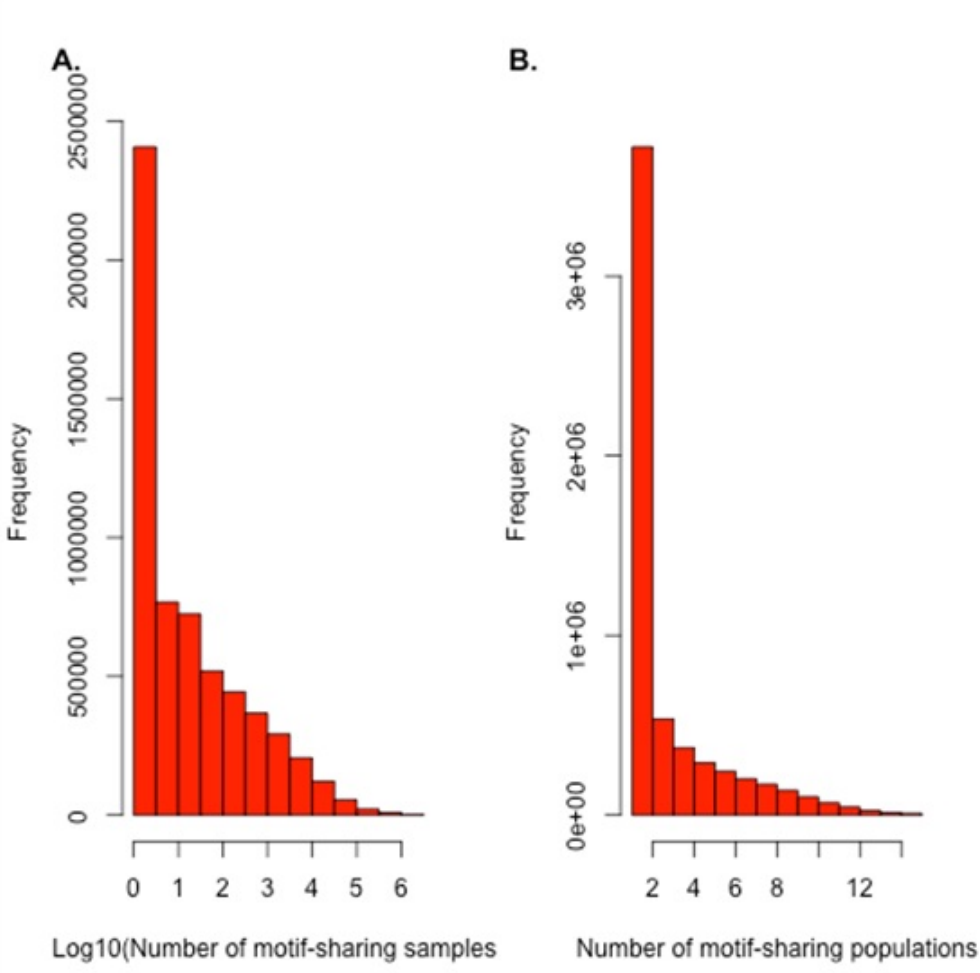

Figure 5.9: A histogram of samples **(A)** and populations **(B)** that share 10 aa long motifs. The majority of motifs were shared by one or two samples and populations with a heavy right-tailed distribution.

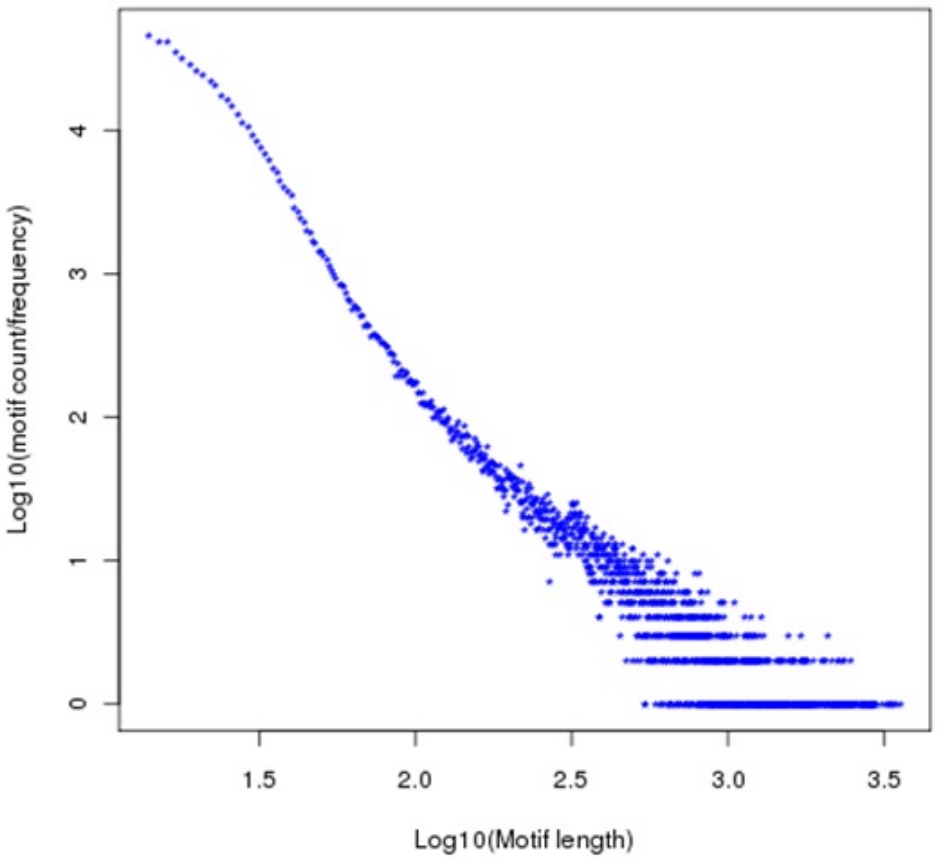

Figure 5.10: Frequency of shared motifs. This plot shows the number of shared motifs as a function of motif-length (x-axis). A large proportion of shared motifs were below 100 aa.

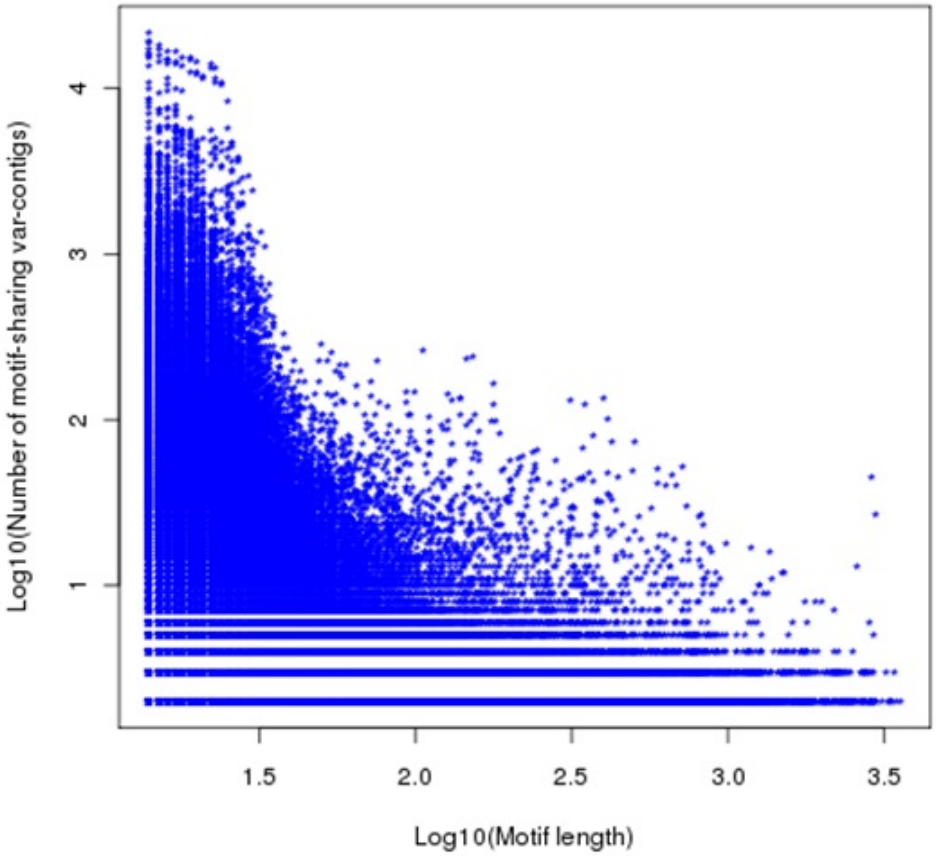

Figure 5.11: Motif sharing *var*-contigs from a pmatch analysis of all-vs-all *var*-contigs. The scatter plot shows a negative correlation (R^2=-0.2; p-value $< 2.2x10^{-16}$) between motif length and the number of *var*-contigs that share a motif.

Initially, we hypothesised that the three identical *var*-contigs were a result of contamination during sample handling. In order to verify this, raw reads from the three samples (PA0036, PA0020 and PH0136) and an additional control sample (from Thailand) were aligned to all *var*-contigs of the sample PA0036 (Figure 5.12). If there was contamination, non-core reads of the other samples would align to *var*-contigs other than PA0036_VAR1. However, the alignment results showed a distinct full coverage signal over PA0036_VAR1 from all the three samples. Other *var*-contigs of PA0036 were only covered by non-core reads of PA0036 (i.e. mapping to itself as expected). It was reassuring to confirm that non-core reads from a control sample (from Thailand) did not align to any of the *var*-contigs. The long perfect matches between single genes therefore appear to represent genuine biological events. It is expected to see a higher degree of conservation between *var* genes of the central clusters. It is thus intuitive to assume *var*-contigs with long perfect matches come from a central region of chromosomes. However, aligning the 10 largest motifs to the *P. falciparum* genome (via BLAST (Altschul et al., 1990)) revealed top matches to subtelomeric *var* genes such as PF3D7_0632500 on chromosome 6 (Figure 5.13). It is important to note that this may not be the best way of identifying central *var* clusters as the target (i.e. 3D7) is only one genome. A flanking sequence of *var* genes could provide a better marker to identify central *var* genes based on similarities of Ups sequences (see Chapter 1 for details).

To investigate whether the long motifs are associated with specific *var* groups (1 to 6), we looked at the number of distinct *var*-groups that are represented by a motif. The results show that the majority of *var*-contigs that share longer motifs were represented by fewer groups (1 to 2) than shorter motifs which can contain up to all six groups (Figure 5.14).

Next, analysis of the six groups represented by *var*-contigs that share a motif revealed that the majority of *var*-contigs that shared long motifs were of groups 1, 2 and 3 ($\sim$60% of *var*-contigs for a motif length of above 500 aa and $\sim$75% for motifs above 1,000 aa). These three groups are shown to contain a DBLα with two cysteine residues and correspond with Type A *var* genes (Bull et al., 2005, 2007, 2008).

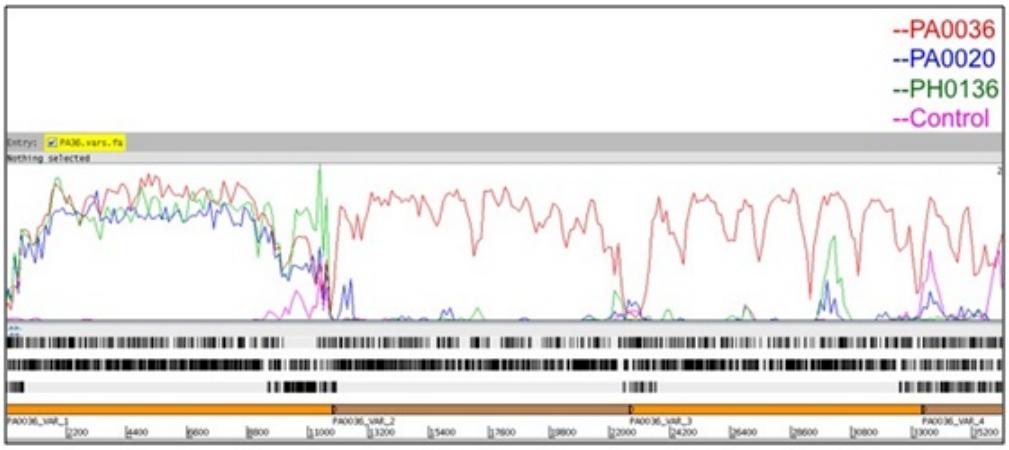

Figure 5.12: Artemis view of non-core reads from four samples (PA0036, PA0020,PH0136 and Control sample) aligned to *var*-contigs of PA0036. The top panel shows paired-read coverage plots for PA0036 (red), PA0020 (blue), PH0136 (green) and the Control sample. The middle panel shows the three reading frames of the forward strand for *var*-contigs of PA0036. Black bars represent stop codons, ORFs are represented by long open white blocks. The bottom panel shows *var*-contigs of PA0036 starting at PA0036_VAR1. Read coverage from samples PA0020 (blue), PH0136(green) was visible over PA0036_VAR1 while the remaining *var*-contigs of PA0036 remain uncovered.

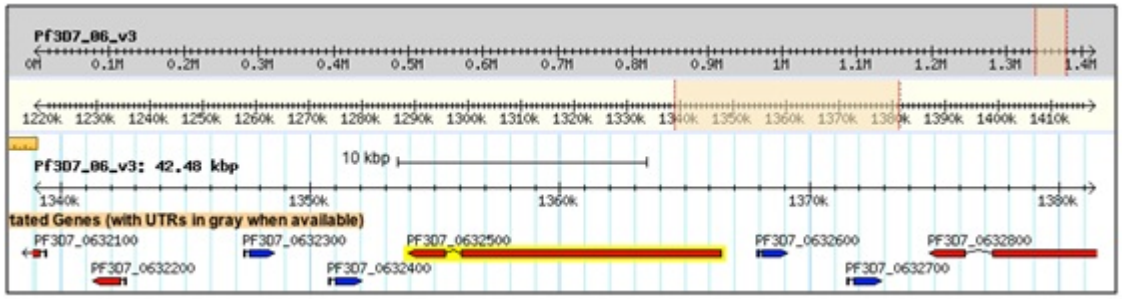

Figure 5.13: A screen shot of PF3D7_0632500 from PlasmoDB: The subtelomeric gene PF3D7_0632500 was the closest match (∼47% identity over the full length) to the long motif shared by three samples from different geographical regions.

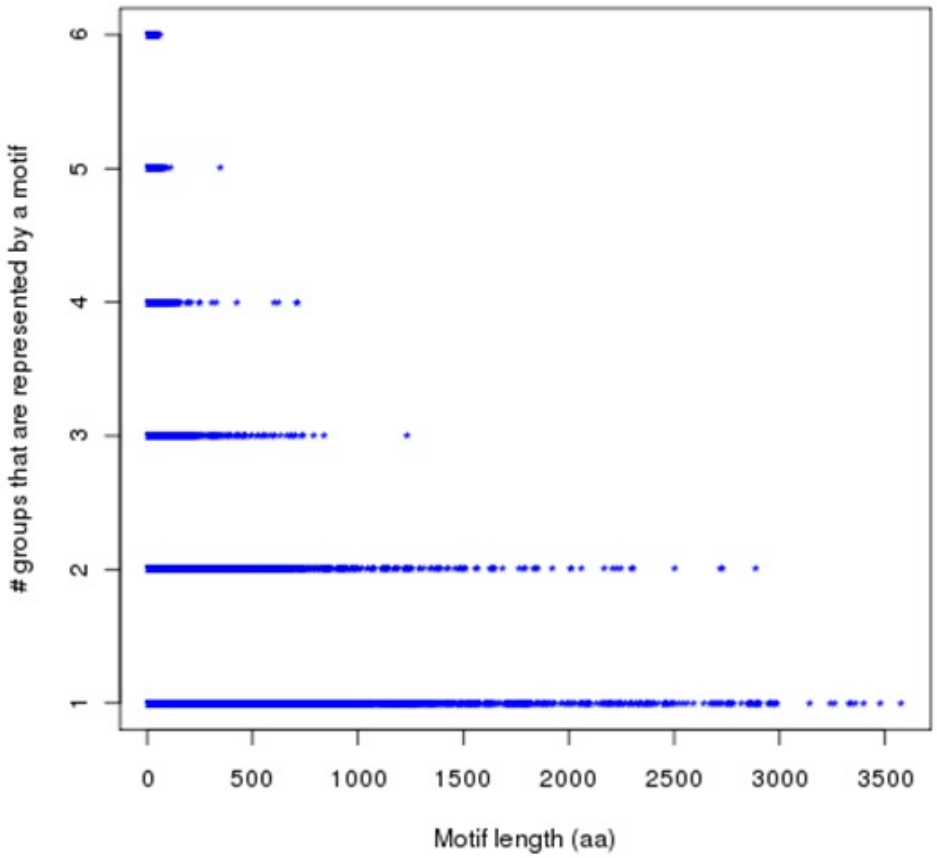

Figure 5.14: Number of groups represented by *var*-contigs that share a pmatch motif for 426 samples that had a minimum of 30 *var*-contigs. Long motifs were shared by *var*-contigs that belong to one or two distinct groups. Conversely, the short motifs are shared by *var*-contigs from all groups.

5.3.7.2 Amino acid and nucleotide BLAST matches

The pmatch analysis was only able to show perfect matches between *var*-contigs or conserved blocks within *var*-contigs. In order to investigate similarity between *var*-contigs while allowing mismatches, a BLAST search of all-against-all *var*-contigs was conducted. The results revealed long identical matches over the full length of *var*-contigs, confirming observations of the pmatch analysis. A closer look at nucleotide alignments of *var*-contigs also showed long identical matches that span over exons, introns and upstream regions. The observed sequence similarity between *var*-contigs of the same and different geographical origins(match lengh >5 kb; identity >99%) was higher than expected from previous studies. The results of the BLAST search provided a better way of processing the output and quickly identify matches of a given *var*-contig. For example, *var*-contig PA0036_VAR1 had a match with total of 59 *var*-contigs from 57 samples in 6 countries at a minimum identity of 99% and match length of 5 kb. The 59 *var*-contigs represented The Gambia (n=21), Ghana (n=6), Mali (n=4), Burkina Faso (n=1), Thailand (n=2) and Cambodia (n=25). 54 of the 59 *var*-contigs were of group 3 *var* genes, while the remaining five were of group 1. As mentioned previously, these two groups belong to Type A *var* genes.

Match length vs percent identity of a match

A scatter plot of nucleotide matches of PA0036_VAR1 to other *var*-contigs revealed a positive correlation ($R^2 = 0.3; p - value < 2.2x10^{-16}$) between match length and the percent identity of a match (Figure 5.15). As the match length decreased, the identity of a match also decreased. The observed relationship between match length and percentage identity of a match was further investigated by looking at all *var*-contigs at various identity cutoff values. Longer matches were predominantly found at higher percent identity thresholds (Figure 5.16). These results are interesting as they have implications on the time-scale of events that contributed to maintaining diversity in *var* genes. For example, the long perfect matches may be a result of recent population expansion events where there was not enough time for recombination to break these long haplotype blocks. Conversely, shorter matches indicate a longer time scale since the

event allowing more SNPs to accumulate.

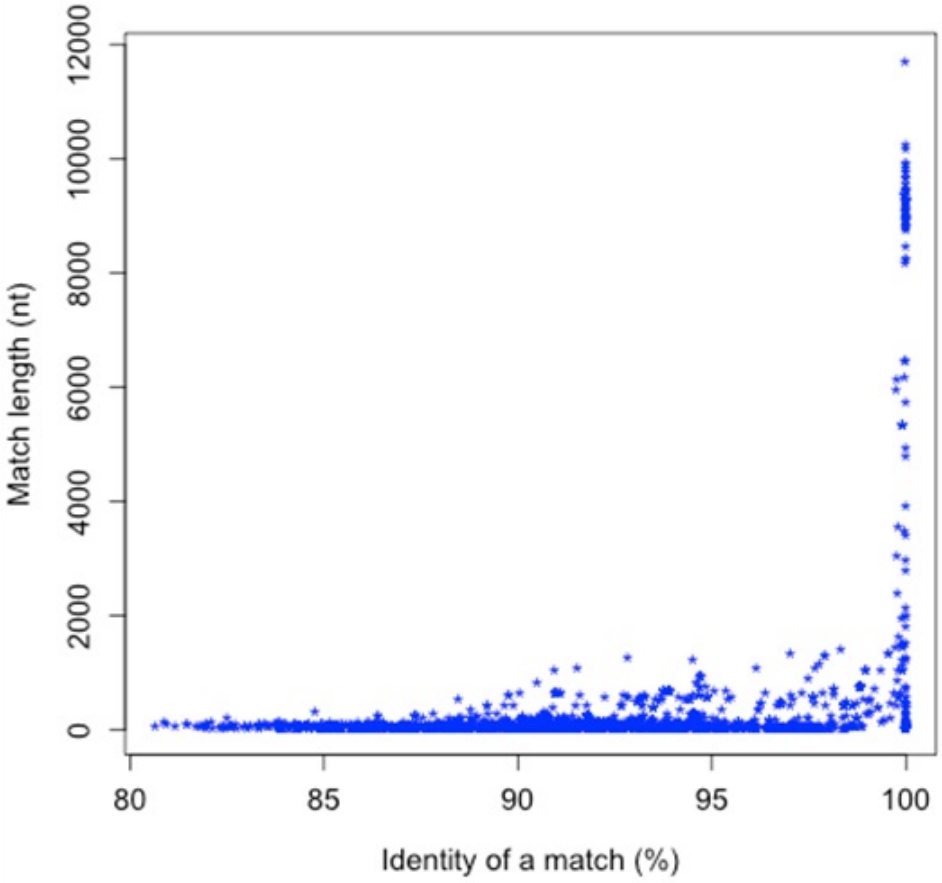

Figure 5.15: A scatter plot of match length and percent identity for the *var*-contig PA0036_VAR_1. Longer matches had the highest percent-identity (R^2=0.3, p-value $< 2.2x10^{-16}$)

In both cases (Figure 5.15 and Figure 5.16), the highest match length values were observed at high identity thresholds.

BLAST matches of ORFs

To minimize the effect of low complexity regions, such as introns, amino acid translations of *var* contigs (using the longest ORF with DBLα) were used to

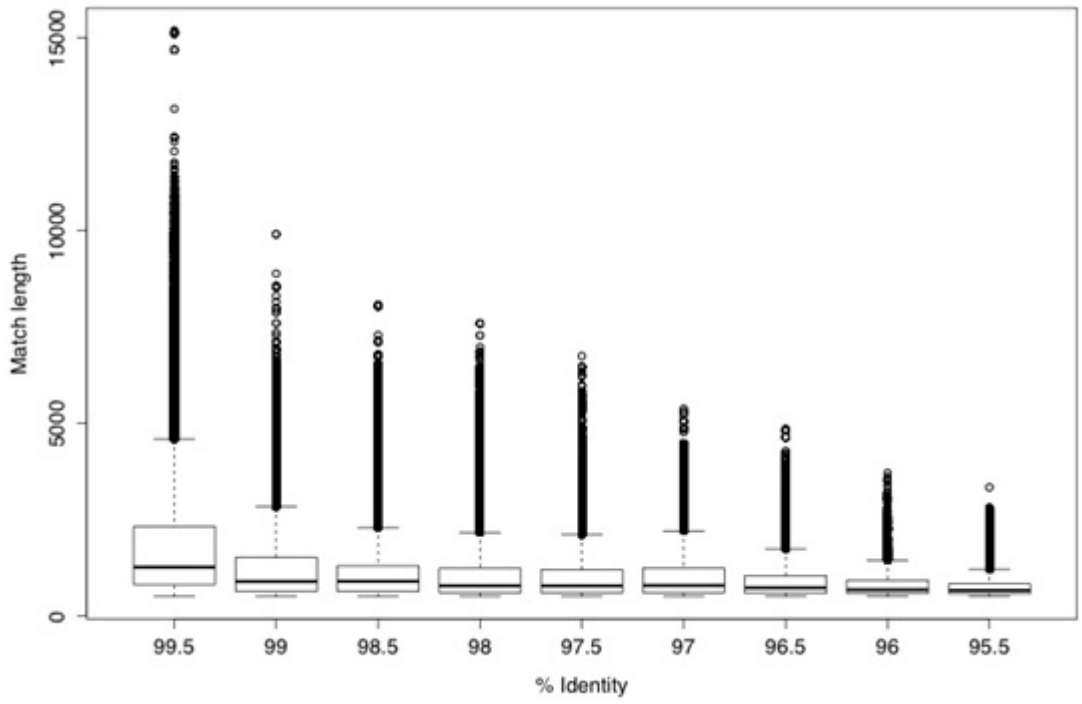

Figure 5.16: Box plots of match length (bp) for different cutoff values of percent identity using all *var*-contigs. Box limits represent median, first and third quantiles; whiskers represent the upper and lower bounds while outliers are shown by the dots.

generate a list of potential matches between *var*-contigs. Long and perfect amino acid matches, similar to those observed from Pmatch, were also identified using the BLAST search. Previous studies have identified three strain-transcending *var* genes: var1CSA, var2CSA and Type 3 *var* genes, that were found in clinical and culture-adapted isolates studied so far including 3D7 and IT (Kraemer et al., 2007). In the 3D7 genome, three *var* genes were identified as Type 3 *var* genes: PFA0015c (PF3D7_0100300), PFF0020c (PF3D7_0600400) and PFI1820w (PF3D7_0937600).

In order to test if the long-perfect matches were homologues of the known strain-transcending *var* genes, ORFs from *var* genes of the 3D7 genome were aligned to ORFs of the ∼50,000 *var*-contigs. As var2CSA (PFL0030c/PF3D7_1200600) does not have the DBLα tag, it was not included in the current list of *var*-contigs. We expected the Type 3 *var* genes and var1CSA (PFE1640w-ps/PF3D7_0533100) to have sequence homology with the highly conserved *var*-contigs. However, no match was observed at higher identity and length thresholds of 99% and 1,000 aa respectively. Lowering the match length cutoff to 50 aa identified matches between 13 *var*-contigs and two of the three Type 3 *var* genes (PFF0020c had a match with two *var*-contigs and PFI1820w to 11 *var*-contigs). The 13 *var*-contigs represented samples from The Gambia, Ghana, Mali, Kenya, Thailand and Cambodia, but they were not part of the long perfect matching *var*-contigs described earlier in this section.

The longest identical ORF matches of length 4,668 aa were observed between *var*-contigs of Thailand and Cambodia. Similarly, *var*-contigs from other countries including Kenya and The Gambia were also found to have perfect matches of up to ∼4,000 aa. The Markov Clustering Algorithm was able to detect distinct clusters of *var*-contigs based on a pairwise similarity measure derived from the BLAST matches of ORFs. A social network analysis was then used to visualise the population level structure of the global collection of *var*-contigs as described in the following section.

5.3.8 Clustering *var*-contigs and detecting population structure

There is no established method for the analysis of diversity and population structure in the *var* gene family. This is mainly due to high levels of recombination that prevent orthology /ancestry from being established. A social network analysis was used a better approach to deal with the complexity resulting from a highly polymorphic nature of *var* genes (Bull et al., 2008). Here results of a preliminary analysis of amino acid similarity networks were presented as an exemplar method of establishing structures in populations of *var* genes.

To simplify the analysis, samples with over 70 *var*-contigs were excluded, as they are likely to have multiple infections (Table 5.3). Populations from Burkina Faso and Bangladesh were also excluded, as they were represented by single samples (Table 5.3). In addition, two of the 102 Gambian samples (PA0106 and PA0107) were excluded, as they had a high ratio of ORFs to *var*-contigs due to a highly fragmented assembly. A total of 424 samples from 11 countries remained for subsequent analysis.

ID	Country	Region	#samples	#PasseedQC1 (DBL $\geq$ 30)	#PassedQC2 ($\leq$ 70)
PA	Gambia	WA	168	162	102
PF	Ghana	WA	122	108	43
PM	Mali	WA	32	32	13
PK	Burkina faso	WA	3	2	1
PT	Malawi	EA	55	43	13
PC	Kenya	EA	25	25	20
PE	Tanzania	EA	15	15	11
PR	Bangladesh	SEA	3	2	1
PD	Thailand	SEA	82	79	63
PH	Cambodia	SEA	191	153	138
PN	Papua New Guinea	SEA	7	7	5
PV	Vietnam	SEA	11	9	8
PP	Peru	SA	11	10	10
			725	**647**	**428**

Table 5.3: A list of samples used for initial assembly (Column 4), pmatch /BLAST analysis (Column 5) and social network analysis (Column 6). Burkina faso and Bangladesh were excluded from the final list as they were represented by a single sample. Two of the 102 Gambian samples were also excluded due to a poor quality assembly.

A total of 26,832 ORFs were obtained from *var*-contigs of the 424 samples. Using a minimum match length of 1,000 aa (identity cutoff of 99%), a total of 1,553 clusters were identified. The majority of clusters (∼68%) only contained 2 or 3 *var*-contigs. They were thus excluded from the network diagram in Figure 5.17. Conversely, the largest cluster (Figure 5.17. Cluster 1) had 141 *var*-contigs representing 127 samples and 10 populations (The Gambia=26, Ghana=14, Mali=4l, Vietnam=2, Thailand=30, Kenya=4, Malawi=2, Peru=4, Thailand=53 and PNG=2). Other clusters contained *var*-contigs that represented variable numbers of samples and populations.

The network diagram (Figure 5.17) represented a total of 6,985 nodes (i.e. ∼26% of the total ORFs in 424 samples) that overlap with other ORFs at a minimum percent identity of 99% and a match length of at least 1,000 aa. In addition to results from the clustering algorithm, the visual inspection revealed interesting identity matches as shown by the five clusters (Figure 5.17, 1-5). The second cluster has a single *var*-contig from Kenya (PC0016_VAR27) clustered with *var*-contigs from Thailand and Cambodia. Similarly, cluster 4 has single *var*-contigs from The Gambia, Malawi and Vietnam mixed with contigs from Thailand and Cambodia. Conversely, most of the *var*-contigs in cluster 3 were from The Gambia and Cambodia, while mixed with single *var*-contigs from Thailand, Ghana and Mali. Finally, in cluster 5, single *var*-contigs from Tanzania, Mali and Thailand were clustered with samples from Peru, Kenya and The Gambia. These observations highlight the widespread distribution of highly conserved *var*-contigs.

A higher degree of overlap between *var*-contigs is observed with the majority of clusters representing samples primarily from South East Asia. These samples also form some of the largest cluster sizes (connected components) compared to African samples. A larger proportion of *var*-contigs from African samples formed smaller clusters and were excluded during the filtering based on degree of a node (i.e. *number of connections*<4). This is an interesting observation as samples from Africa have much older *var* repertoires that have been exposed to natural forces such as mutation and recombination. As a result, a lower degree of similarity is expected in African samples compared to Thailand and Cambodia.

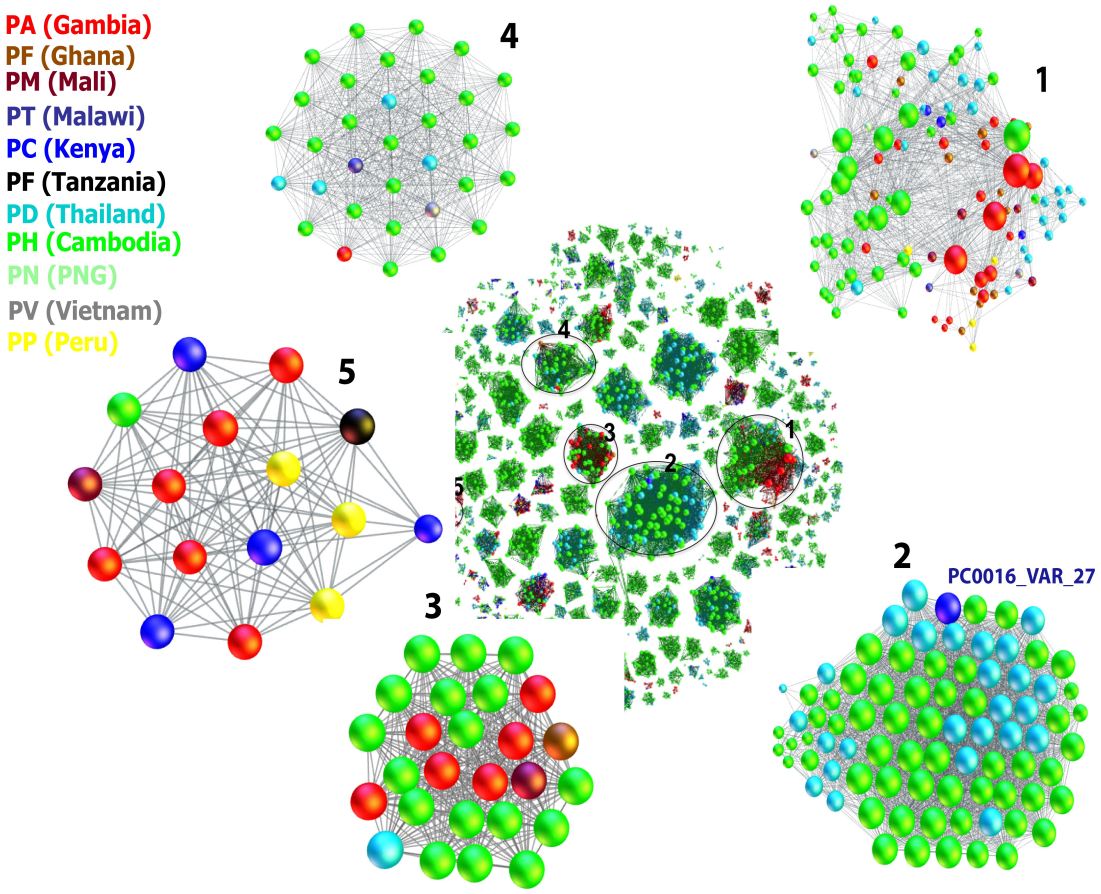

Figure 5.17: Amino acid identity network of *var*-contigs from 424 samples. The main figure shows *var*-contigs that share highly similar sequence blocks (above 1,000aa and a minimum identity of 99%). *Var*-contigs that have fewer than 4 connections with other contigs were excluded to simplify the graph. The number of connections of a node (*var*-contig) is represented by the size of each circle. A closer view of five representative clusters is also shown (1-5). Cluster 1 contained 141 ORFs*var*-contigs representing 127 samples and 10 populations (The Gambia=26, Ghana=14, Mali=4l, Vietnam=2, Thailand=30,Kenya=4, Malawi=2, Peru=4, Thailand=53, and PNG=2). Other clusters contained *var*-contigs that represented variable numbers of samples and populations (See text for details).

Country-specific clustering analysis was done on 138 Cambodian samples revealing sub-populations of *var*-contigs as shown in Figure 5.18. 42% of the total ORFs (n=8,065) were grouped into 629 clusters containing 2 to 55 *var*-contigs. Initially, such clusters may appear to be a result of clonal expansion events. However, there was a very strong positive correlation ($R^2 > 0.99$) between number of *var*-contigs and the number of unique samples in each cluster (Figure 5.19) suggesting a wider distribution of highly conserved *var*-contigs.

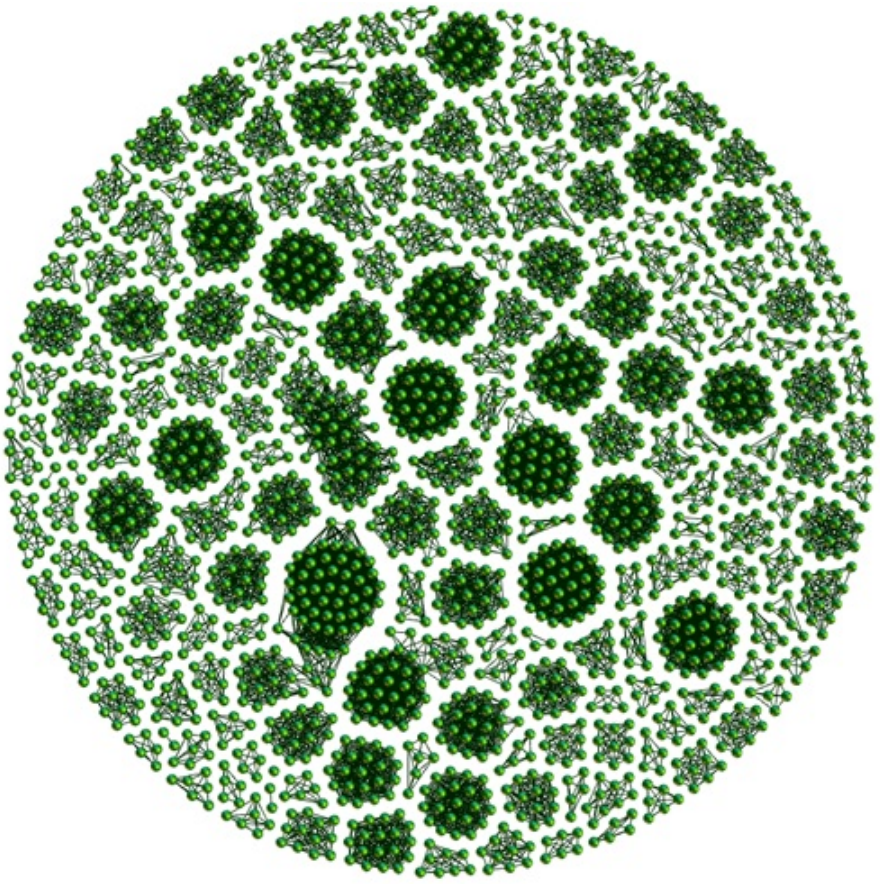

Figure 5.18: A network of *var*-contigs from Cambodian samples ($degree > 3$). A total of 3,380 *var*-contigs (nodes) were represented in this network (number of edges=15,598). Initially, the distinct sub-groups of *var*-contigs may seem due to a clonal expansion event. However, as shown in Figure 5.19, each cluster contains different samples suggesting the presence of long and unexpected conserved sequences in a large number of samples.

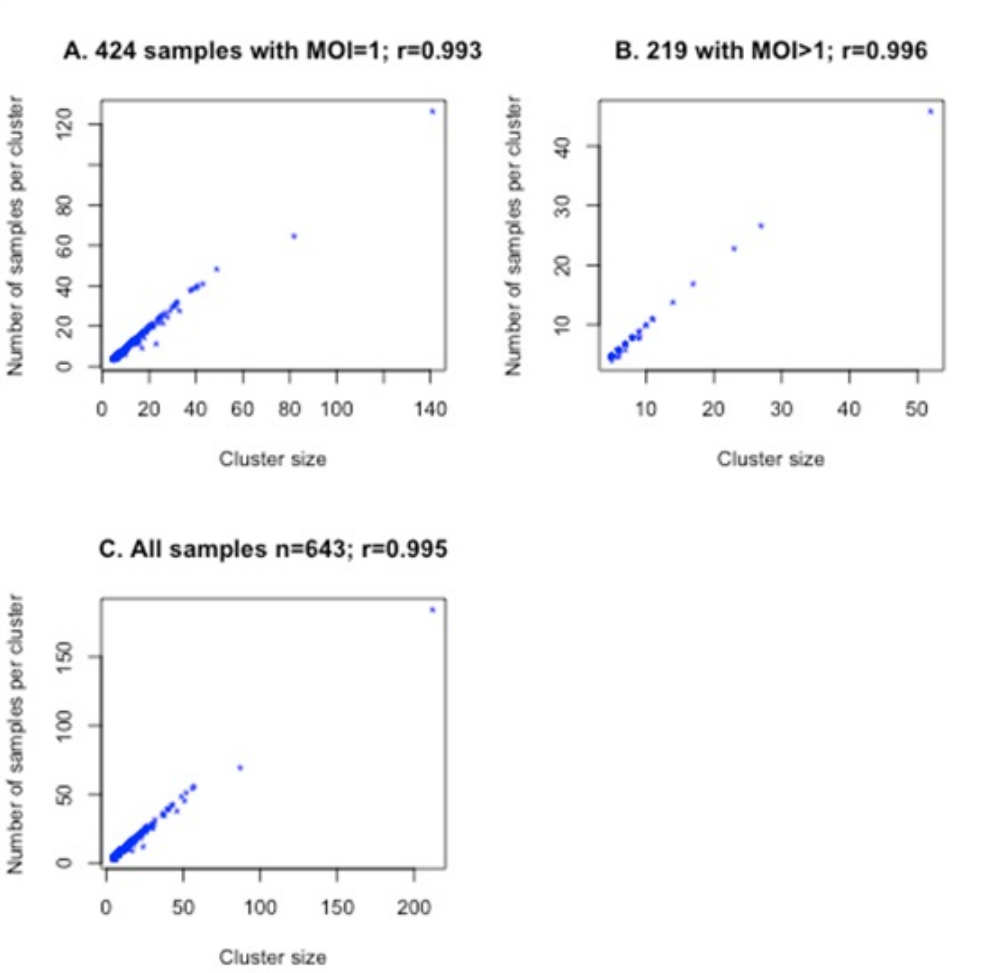

Figure 5.19: Scatter plots of cluster size vs the number of unique samples represented in each cluster. A very strong correlation is observed (as shown by R^2 values) between cluster size and number of distinct samples. **A)**. Scatter plot for samples with a minimum of 30 *var*-contigs and a maximum count of 70 *var*-contigs. These samples are considered to have a single infection with a multiplicity of infection (MOI) of 1. **B)** Scatter plot for samples that have above 70 *var*-contigs (MOI>1). **C)**. Scatter plot for all samples.

5.4 Discussion

This chapter presented assembly results of *var* genes from clinical samples. The following two conclusions summarise the results obtained and their implications.

Conclusion 1: *The results show the first full length assembly of var genes and the largest collection of var genes from clinical samples.*

It was possible to generate the largest collection of full length *var*-contigs using an iterative assembly approach developed to specifically assemble *var* genes. Although the assembly was initiated with a small number of motifs from three lab-adapted samples, it was able to generate high quality assembly of *var* genes on over 700 samples. The increase in the number of samples used to iteratively generate seed motifs is one reason for high quality assembly with as few as three iterations. The results show the first report of a targeted assembly of highly polymorphic gene families in general and of the *var* gene family in particular.

As expected, sequence quality and yield affect quality of the final assembly. This effect is likely to be pronounced in the iterative assembly approach compared to a simple *de novo* assembly of all reads due to the need to have enough reads that contain seed motifs to initiate the process. However, it could also be seen as an early quality check as samples with poor quality and low yield will not have enough reads to proceed with the assembly. Despite the continual improvement in sequencing yield and protocols developed to remove contaminants, variability in quality and yield are characteristics of sequences obtained from natural populations.

The expected number of *var* genes with the presence of a single genotype provided a simple measure of assembly quality during the initial stages of the assembly. Excluding samples with assembled *var*-contigs of below 30 was further justified by looking at the number of non-core reads and quality of the raw data.

As described in Chapter 2, the Illumina platform is prone to substitution errors at the beginning and ends of reads. Initially, we intended to incorporate

trimming of reads in the assembly workflow. However, assembly attempts by trimming reads at error-prone ends did not improve the assembly of *var* genes. Although Velvet was chosen to generate seed contigs, the assembly approach is modular such that a different assembly tool could easily be used if future tests justify the choice. Additional iterations and quality control steps could be included by simply restarting the assembly process from where it stopped (i.e. start from iteration 4).

In summary, the iterative assembly approach together with the quality control steps applied in this chapter were shown to be effective in producing high quality full length draft *var* genes.

Conclusion 2: *Unexpected cross-continental var-contigs were identified between samples of unrelated countries.*

Preliminarily analysis of similarity between *var*-contigs obtained from the *de novo* assembly of clinical samples revealed unexpected and long sequences of high similarity. Nucleotide and amino acid alignments as well as perfect-match searches confirmed the presence of continent-transcending *var*-contigs (CTVs).

We have established that these similarities were not because of DNA contamination and informatics issues (eg. misassemblies). Potential explanations for the existence of continent-transcending *var*-contigs in such diverse parasite populations and their implications are presented in the next chapter.