Chapter 2

Evaluating existing short read assembly methods

2.1 Introduction

Advances in next generation sequencing technologies have significantly improved read length, yield and quality of whole genome shotgun sequencing data. There has also been a considerable development of algorithms and software to deal with the problem of piecing together sequence fragments into a longer contiguous sequence. However, assembly attempts of whole genome *P. falciparum* sequence data are faced with unique challenges due to the high A+T content (∼80% in coding and ∼90% in non-coding regions) (Gardner et al., 2002). *De novo* assembly of *P. falciparum* genomes in general and subtelomeric gene families in particular is thus extremely challenging due to the base composition bias and presence of repeat sequences. *Var* genes are highly polymorphic and composed of mosaic blocks that have regions of high similarity. The quality of the raw sequence data is often affected by systematic errors from the sequencing instruments. Systematic errors and inherent sequence features are thus expected to have a significant impact on the assembly of *var* genes.

The aim of this part of the thesis (year 1) was to see if it was possible to assemble *var* genes using existing short read assembly tools. In this chapter, I will evaluate the feasibility of existing approaches to assemble *var* genes

followed by investigation of potential reasons for poor quality assembly in *var* genes.

2.2 Methods

2.2.1 Library preparation and sequencing

Library preparation and sequencing of samples used in this thesis were done by the Research and Development and Sequencing Production teams at the Sanger Institute. The protocols used for library preparation and sequencing are briefly summarised below.

Library preparation
Standard genomic DNA library preparation
DNA samples were initially quantified on the Invitrogen Qubit and then fragmented using the Covaris Adaptive Focused Acoustics technology (fragment sizes of 200-300 bp and 300-400 bp). End-repairing of fragments and creation of blunt-ends were done using T4 and Klenow DNA polymerases, and T4 polynucleotide kinase respectively. This was followed by A-tailing, addition of a single 3′ A nucleotide to the repaired ends using Klenow exo- and dATP. Standard adapters were then ligated according to the manufacturers guidelines. Size selection of ligated fragments was done using Agencourt AMPure XP beads. The libraries were then enriched by 8 cycles of PCR and quantified using Agilent Bioanalyser chip and Kapa Illumina SYBR Fast qPCR kit.

PCR-free Genomic DNA library (NoPCR) preparation
The PCR-free library preparation protocol (Kozarewa et al., 2009) was developed to minimize the effect of amplification artifacts especially in genomes with G+C bias such as *P. falciparum*. However, it also requires more input DNA. This method uses similar protocols of DNA qualification, shearing, end-repairing and A-tailing. However, instead of the standard adapters, NoPCR adapters were ligated (containing primer sites for sequencing and flowcell surface annealing) according to the Amplification-free Illumina sequencing

protocol (Kozarewa et al., 2009). Subsequent steps of size selection and DNA extraction are similar to the standard library preparation protocol above. Both standard and PCR-free libraries were used in this chapter (Table 2.1). Clinical samples used in Chapters 3 and 5 were all prepared using the PCR-free protocol.

Cluster generation and Sequencing

Initially, in order to allow hybridization of template strands to adaptors that are attached on the flowcell, libraries were denatured with sodium hydroxide and diluted in a hybridisation buffer. Cluster amplification was performed on the Illumina cluster station (changed to Illumina cBOT after April 2010) using the V4 cluster generation kit following the manufacturer's protocol. Cluster density was measured using SYBRGreen to determine whether a flowcell had enough DNA for sequencing. This was followed by consecutive linearization, blocking and hybridisation of R1 and R2 sequencing primers for (i.e. for the forward and reverse reads). Sequencing-by-synthesis was then performed for 75 to 150 cycles depending on the sequencing instrument. These steps were performed using proprietary reagents according to manufacturer's recommended protocol (`https://icom.illumina.com/`)

Samples used in this chapter

Laboratory adapted and cultured clinical isolates of *P. falciparum* were used with the aim of evaluating and optimising existing short read assembly methods in polymorphic gene families, specifically the *var* gene family. Due to the availability of a complete and nearly base-perfect reference genome, re-sequencing of the 3D7 isolate provides an ideal benchmarking standard for new sequencing technologies and protocols. DNA for the 3D7 isolate was obtained from Prof. Chris Newbold's lab in Oxford.

2.2.2 Choice of short read assemblers

Velvet (Zerbino and Birney, 2008) was one of the first short read assemblers to popularise de Bruijn graphs. It is easy to install and run with a growing community of users. The major limitation is the large amount of Random Access

	Libraries	Insert size(bp)	Library type	Read length(bp)	Machine	Cov. (X)
3D7	S_i200	200	Standard	76	GAII	58
	NP_i200.1	200	NoPCR	76	GAII	163
	NP_i200.2	200	NoPCR	76	GAII	184
	HS_i500.2	500	NoPCR	75	HiSeq2000	607
	NP_i500	500	NoPCR	76	GAII	338
	HS_i500.1	500	NoPCR	75	HiSeq2000	724
	MS_i3k.1	3000	Large insert	75	MiSeq	34
	MS_i3k.2	3000	Large insert	150	MiSeq	20
	MS_i3k.3	3000	Large insert	150	MiSeq	70
	MS_i3k.4	3000	Large insert	150	MiSeq	64
Field samples	F1	200	Standard	76	GAII	
	F2	200	Standard	76	GAII	
	F3	200	Standard	76	GAII	
	F4	200	Standard	76	GAII	
	F5	200	Standard	76	GAII	
P. reichenowi	*Pr*	200	Standard	76	GAII	

Table 2.1: Samples used to evaluate existing assembly approaches and determine error rates of the Illumina sequencing technology in *P. falciparum*. Libraries were named according to the library preparation protocol (S for standard, NP for NoPCR/PCR-free; HS for HiSeq and MS for MiSeq) and the insert sizes of the libraries (eg. i200 for 200 bp library)

Memory (RAM) required for large genome assemblies. However, assembly of *var* genes and the *P. falciparum* genome are within the limits of Velvet's memory requirement. Velvet was therefore a good starting point for the purpose of comparisons. Abyss and SOAPdenovo were especially designed to address Velvet's limitation with large genomes. Li and colleagues (Li et al., 2010; Ruiqiang Li, 2010) reported better assembly results for SOAPdenovo compared to Abyss using both small (*Ecoli*) and large (African human) genomes. Velvet (version 1.1.04) and SOAPdenovo (version 1.05) were therefore chosen to represent existing short read assemblers with the intention of identifying a program that could generate high quality assembly for polymorphic gene families in *P. falciparum*.

2.2.3 Assembly benchmarking using error free in silico reads

Error free reads were generated using a Python script written by Martin Hunt in our laboratory (*simulate_pe_reads.py*). Reads were simulated assuming a Gaussian distribution of fragment sizes given a mean fragment size, standard deviation of the distribution and the average read coverage. This script is used to generate in silico reads in this thesis unless stated otherwise.

To compare the performance of Velvet and SOAPdenovo, paired reads of length 76 bp (coverage=80x; mean fragment size=200 bp, standard deviation=30) were simulated from a total of 93 *var* genes of the 3D7 genome (including pseudogenes and truncated exon 2 sequences). The two assemblers were compared on the number and quality of contigs they produced. To obtain the best result for each assembler prior to comparison, the two assemblers were first optimised by running each tool at *k-mer* sizes of 51, 55, 61, 65 and 71. Ideally, the best assembler will generate the least number of contigs, with high N50 values, a total number of bases as close to the expected length of the target sequence (i.e. $\sim$450 kb for *var* genes in the 3D7 genome), and the fewest errors. Assembly quality was assessed based on the coverage of contigs and the repertoire completeness of *var* genes as described below.

2.2.3.1 Accuracy and coverage of contigs

ABACAS (Algorithm Based Automatic Contiguation of Assembled Contigs) was used to assess the quality of assembled contigs (Assefa et al., 2009). If a good quality reference sequence is available, comparing assembled contigs to the reference provides the best measure of completeness and accuracy. A similar approach is used during the process of finishing genomes whereby contigs are aligned to a reference sequence of a close or distant relative with the aim of establishing a relative order of contigs using regions of conserved synteny. However, there was no readily available software to automate the process. I originally developed ABACAS to address this issue by rapidly aligning contigs to a reference genome in order to determine the order and orientation of contigs relative to a reference genome. A multi-FASTA file of contigs was first aligned as nucleotide (*option -p nucmer*) or six-frame translated amino acid sequence

(*option -p promer*). Based on the alignment output, contigs are then ordered and orientated to generate the new reference, termed the pseudo-molecule. Output files include an ordered FASTA file, a feature file and comparison files for visualisation in the Artemis comparison tool (ACT) (Carver et al., 2005). The comparison file displays an ordered list of contigs that were aligned to the reference genome. The proportion of each contig aligned to the reference (i.e. percent coverage) is reported alongside the percent identity of the match.

Additional use-cases of ABACAS such as checking the quality of assembled contigs became apparent during the course of this project. Misassemblies were therefore identified by looking for unordered contigs of length above 1 kb and ordered contigs with a coverage and percent identity values lower than 95%. Despite the reliability of this approach, its application in the absence of a reference sequence is very limited. It was thus necessary to develop an alternative method to determine the completeness of *var* gene repertoires in a wide range of assembly projects including *de novo* assembly of clinical samples.

2.2.3.2 Estimating *var* gene repertoire completeness

As described in Chapter 1, the DBLα domain is present in nearly all *var* genes of the 3D7 genome and other sequenced isolates (Kraemer et al., 2007; Rask et al., 2010). The number of contigs that have a complete DBLα domain were counted using a customised Perl-script (*getDBL.pl*). The script was initially written for DBLα sequence tags by Dr. Pete Bull of Kemri-Wellcome Research Unit, Kilifi Kenya and improved to look for DBLα start and end motifs in all the six frames instead of one frame each from the forward and reverse strands. Outputs of the script include amino acid sequence file of DBLα tags and a FASTA-format file of contigs that contain DBLα. Counting such contigs will therefore rsult in the closest approximation of repertoire completeness that is robust in clinical samples. Although the var2CSA genes (gene that encodes a *Pf*EMP1 variant responsible for pregnancy associated malaria) do not contain the DBLα domain, they are highly conserved and could easily be identified using their 3D7 homologues.

2.2.4 Evaluating the velvet assembler on real and simulated reads

In order to evaluate Velvet on real data, a PCR-free library of 3D7 (Table 2.1, libraries NP_i200.1 and NP_i200.2) was sequenced on Illimina's GAII platform. One aim of this analysis was to investigate the quality of assembled contigs while decreasing the dataset from the whole-genome to a chromosome and finally to reads that belong to *var* genes.

2.2.4.1 Whole genome assembly

First, to find optimal results, a whole genome *de novo* assembly was done by varying the *k-mer* size and coverage cutoff values. A *k-mer* of 61 and coverage cutoff dynamically determined by Velvet (*i.e. coverage_cutoff=auto*) resulted in the best assembly results as determined by the highest N50 and least number of contigs. Other parameters were kept to the default settings. In order to investigate the effect of poor quality reads on assembly, a filtered set of reads was generated by aligning raw reads to the 3D7 reference genome (version 2.1.4) using Bowtie (Langmead et al., 2009) (*using the -v alignment mood, -v 2*) allowing a maximum of two mismatches. Bowtie is a memory efficient and fast short read alignment tool that uses the Burrows-Wheeler Transformation (BWT) to index the reference genome. The version of Bowtie used in this thesis (version 1) did not support gapped alignment, which contributed to its increased speed compared with other BWT based aligners such as BWA (Li and Durbin, 2009).

Filtered reads were then assembled using a similar set of parameters as the raw reads (*velveth -k=61; velvetg -cov_cutoff auto*).

2.2.4.2 Chromosome 1 assembly

To further investigate the effect of errors and reducing the dataset to a single chromosome, sequences that aligned to chromosome 1 were obtained from the filtered set of reads generated in the previous section (library NP_i200.1). These reads were assembled using velvet (*velveth -k 61; velvetg -cov_cutoff auto*). In addition, error free synthetic reads were evenly generated with a similar number

of reads and mean fragment size as the real data (mean=166 bp, standard deviation=35). A second set of synthetic reads was generated by mimicking the coverage of real data over chromosome 1 followed by a random introduction of mismatch errors. Both sets of simulated reads were assembled using similar parameters as the real data (*velveth -k=61; velvetg -cov_cutoff auto*).

2.2.4.3 Assembly of *var* genes

In order to evaluate the performance of Velvet on *var* genes from the real data, reads that aligned to *var* genes of the 3D7 genome were obtained from a PCR-free library (Table 2.1, Library NP_i200.1) and assembled (*Velvet, k-mers 25 - 65, cov-cutoff=auto*). Potential reasons for a poor quality assembly of *var* genes were investigated in two steps.

First, raw reads were aligned to a concatemer of 3D7 *var* genes with the aim of assessing the effect of sequencing errors and uneven coverage. The alignment output was stored in the Binary Alignment/Map (BAM) (Li et al., 2009a) format and visualised in ACT using the BamView utility (Carver et al., 2010). A graphical representation of read coverage and SNPs (errors) was used to look at their correlations with assembly quality.

Second, to look at the effect of repeats in assembly, shared sequences within *var* genes of the 3D7 genome were identified using a pairwise blast search (*all-against-all blast; blastn, -F F, -e=1x10*$^{-3}$). Regions of genes that have a perfect match with a length above the fragment size of the library (200 bp) were identified as repeat regions. Reads that aligned to such regions were excluded to generate a second set of *var* reads. Assembly of the two sets were compared by looking at the underlying de Bruijn graphs which were visualised using a python script contributed to the Velvet package by Paul Harrison (*graph2.py*) .

2.2.5 Evaluating mapping based assembly approaches

To assess the feasibility of a reference guided assembly approach, short reads from the 3D7 clone, three field samples and *Plasmodium reichenowi (Pr)* (Table 2.1) were aligned to the 3D7 reference genome (BWA; *version 0.5.5, default parameters*). Reads that aligned to *var* genes in proper pairs (i.e. read pairs aligned in the

correct orientation and within the expected insert size) were counted for each gene. The number of mapped reads per thousand bases (kb) was used as a measure of mappability over *var* genes and computed by normalizing the number of properly aligned paired reads over a gene by the length of the gene.

2.2.6 Sequencing errors

A total of 10 genomic DNA libraries were used for error profiling of Illumina's GA2, HiSeq and MiSeq instruments (Table 2.1). In order to assess the improvements and changes in error rates, libraries from early Illumina (GAII) runs and the latest MiSeq runs were used. Raw reads from the reference genome were aligned to the most recent version of the genome (version 3) using Bowtie1 allowing a maximum of three mismatches. The output file was processed using a purposely written Perl-script to identify mismatch/error positions and find error rates at low ($Q5$), medium ($Q15$) and high ($Q25$) quality thresholds. Quality values represent a confidence score assigned by Illumina's base calling algorithm (Bustard) which assigns quality scores (Q) based on an expected error probability P such that :

$$Q_{\text{Solexa prior to v.1.3}} = -10log_{10}(P/(1-P)) \tag{2.1}$$

or

$$Q_{\text{Illumina v.1.3+}} = -10log_{10}(P). \tag{2.2}$$

It is not unusual to observe incorrect base calls or unknown bases (Ns) with high quality score. It was therefore important to look at sequencing errors at low and high quality cutoff values. First, the overall error rate (E) for the forward and reverse reads were computed over the full length of each read R at a quality cutoff Q as follows:

$$ER = (\textit{mismatched bases above } Q\ )/(\textit{mapped bases above } Q) \tag{2.3}$$

Similarly, error rates were computed for each position P on the population of forward and reverse reads at a quality cutoff Q as follows:

$$ERP = (\textit{mismatched bases above } Q \textit{ at } P) / (\textit{mapped bases above } Q \textit{ at } P) \quad (2.4)$$

Finally, substitution profiles were obtained for the 12 mismatch types (A-C, A-G, A-T, C-A, C-G, C-T, G-A, G-C, G-T, T-A, T-C and T-G) by computing the average number of mismatches that exhibited a given patterns across all cycles.

2.3 Results

The standard approach to any assembly problem would be to use *de novo* or reference guided assembly on all fragments (reads). This chapter presents tests performed to evaluate the feasibility of reconstructing *var* genes from short reads using existing assembly tools. A comparison of two representative assemblers, Velvet and SOAPdenovo, is presented using synthetic reads simulated from *var* genes followed by further evaluations on real and simulated data. An investigation into the potential reasons of low quality assembly is also presented with a focus on errors specific to *P. falciparum* sequences.

2.3.1 Comparing *de novo* assembly tools

A total of 507,812 read-pairs were simulated from *var* genes of the 3D7 reference isolate to compare SOAPdenovo and Velvet. To account for the variability in assembly quality with changes in *k-mer* size, both assemblers were first run with *k-mer* sizes of 51, 61, 65 and 71. Optimal assembly values were obtained at a *k-mer* size of 65 for both assemblers (Figure 2.1). Assembly results from *k-mer* sizes of smaller and larger than 65 were highly fragmented. In addition to generating the highest number of contigs, a k-mer of 71 resulted in the shortest size for the Largest-contig (Table 2.2). The highest contig N50 size (5,713 bp) was obtained from the Velvet assembly with 252 contigs compared to SOAPdenovo's N50 of 2,346 bp and 1,398 contigs.

The number of contigs aligned to the reference set of 93 *var* genes (by ABACAS) was higher in Velvet (72% vs SOAPdenovo's 45%) (Table 2.3). No

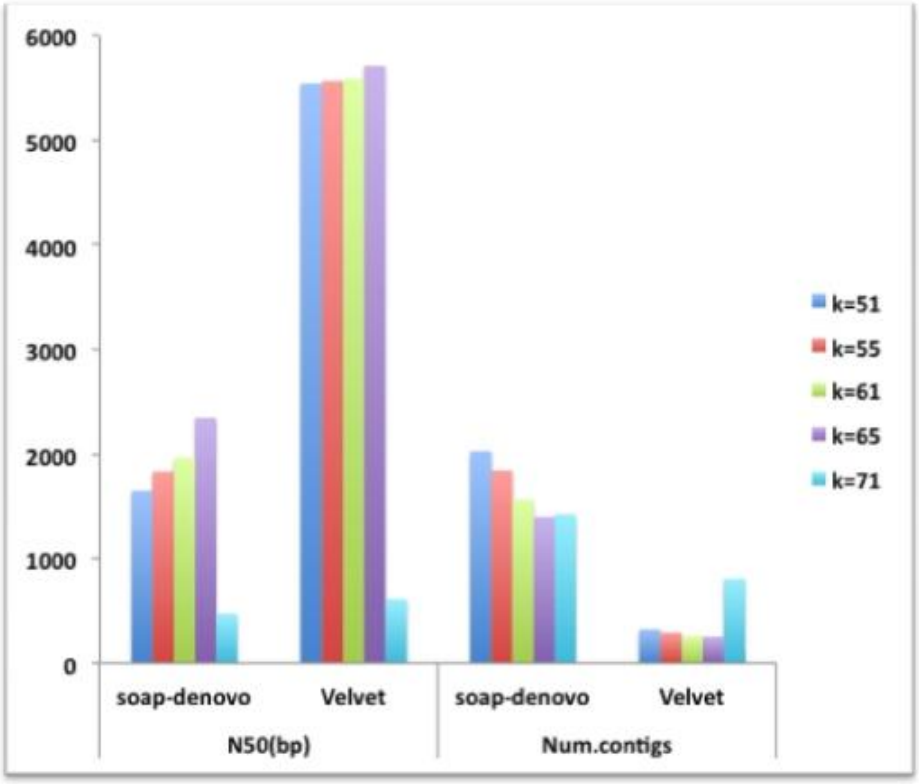

Figure 2.1: Comparing SOAPdenovo and Velvet on synthetic reads simulated from *var* genes of the 3D7 genome. Optimal assembly results were obtained at a *k-mer* size of 65 with Velvet generating a better assembly with the highest N50 and least number of contigs.

k	SOAPdenovo				Velvet			
	Sum (kb)	N50 (bp)	Num.contigs	Largest (bp)	Sum (kb)	N50 (bp)	Num.contigs	Largest (bp)
51	514	1648	2027	9517	21	5543	324	10418
55	514	1831	1841	9525	425	5567	291	10422
61	512	1963	1560	10489	430	5586	258	10489
65	**509**	**2346**	**1398**	**10493**	**433**	**5713**	**252**	**10493**
71	464	471	1422	3842	420	612	804	4521

Table 2.2: Assembly statistics of Velvet and SOAPdenovo at different *k-mer* sizes. Optimal assembly results were found at *k-mer* size of 65.

	Sum (kb)	N50 (bp)	#contigs with DBLα
SOAPdenovo	192	5277	47
Velvet	293	6738	50

Table 2.3: Assembly statistics of contigs with the DBLα domain. Velvet generated the highest number of contigs with the DBLα domain suggesting a better assembly with better representation of the *var* repertoire.

misassembled contigs were detected in both assemblies. A total of 50 contigs contained a complete DBLα domain in the Velvet assembly. The sum of these contigs accounted for $\sim$68% of the expected sum of *var* sequences with DBLα in the 3D7 genome as determined by our method ($\sim$428 kb, 54 genes). Conversely, SOAPdenovo assembled 47 contigs with the DBLα domain that only contained 45% of the expected total sequence (192 kb compared to Velvet's 293 kb).

Velvet was therefore chosen as a better tool to further optimise assembly of *var* genes in *P. falciparum*. The main objective of these experiments was to establish a robust method that could be used in the context of clinical samples. However, the initial comparisons were done using simulated data and real reads from the reference genome as separating out the complete *var*-specific reads would not be possible for a real clinical sample due to high polymorphism. Having established Velvet as a better choice, the following sections present further tests on real and simulated data from the 3D7 genome. Assembly results were examined in a decreasing order of complexity from the whole genome to a single chromosome and finally the *var* gene family.

2.3.2 Velvet assembly of whole genome data

A whole genome *de novo* assembly of 26.6 million reads (paired 76 bp, two lanes) from a PCR-free library (Table 2.1, NP_i200.1 and NP_i200.2) of the reference clone 3D7 was extremely fragmented (Table 2.4). Assembly results from the two lanes were comparable in the number and size distribution of the contigs generated. The number of contigs was ∼20,000 with an average N50 size of 1,370 bp. The total number of assembled bases (i.e. sum of contigs) was however closer to the expected genome size of ∼23 Mb (∼18 Mb to 19 Mb).

A filtered set of reads was obtained by excluding reads that aligned to the reference genome with more than two mismatches. Reads that passed the mapping-based filtering accounted for 73% of the total and covered 97% of the genome with at least one read (94% covered with at least 5 reads). The assembly of filtered reads was slightly better after removing 27% of the reads that were either contaminants or had a lower quality (Table 2.4). However, the results show that whole genome *de novo* assembly of short reads is still impractical for *P. falciparum*.

	NP_i200.1		**NP_i200.2**	
	All reads	**Filtered reads**	**All reads**	**Filtered reads**
Total bases (Mb)	18.87	18.25	19.59	19.2
Num. contigs	19676	18276	19899	18508
N50 (bp)	1361	1442	1387	1491
Average (bp)	958.9	998.8	984.7	1037.6
Largest (bp)	20771	16989	19471	23005
Unused reads (%)	31.22	18.1	41.3	16.87

Table 2.4: Whole genome Velvet assembly of real data from two PCR-free libraries of 3D7. The second data set of 'filtered reads' was obtained by removing reads that aligned to the genome with more than two mismatches. Assembly results were highly fragmented for both libraries.

2.3.3 Velvet assembly on Chromosome 1 of 3D7

Reads that aligned to chromosome 1 of the reference genome were obtained from the 'filtered set' described in the previous section. A mapping coverage

of 94% was observed on comosome 1. The average fragment size (166 bp) was shorter than the expected standard library fragment size of 200 bp. Assembly of these reads was highly fragmented (N50=1 kb, sum of contigs=402 kb) compared to assembly of error free simulated data with a similar number of reads and insert size distribution (N50=15 kb, sum of contigs=614 kb). However, introducing errors and uneven coverage to simulated reads had a significant effect on assembly quality dropping the N50 contig size to 1.3 kb (Table 2.5). These results suggest that sequencing errors and uneven coverage may explain the low quality assemblies.

	Real reads	**Simulated with error**	**Error free simulation**
N50	1076	1255	15278
Average	849	1019	3593
Larges	4979	5457	40793
Total bases	401616	637948	614369

Table 2.5: Assembly results of simulated and real reads on chromosome 1 of 3D7. Real reads and uneven-simulated reads with errors (column 3) had comparable results. Reads simulated without errors and with even-coverage (column 4) resulted in better assembly.

2.3.4 Velvet assembly on *var* genes

Initially, 1.9 million read-pairs where one or both of the reads aligned to *var* genes were assembled generating the best assembly at a k-mer size of 65 with N50 contig size of ∼1.7 kb (sum of contigs=457 kb, Number of contigs=603, Largest contig=9.85 kb). The number of contigs that contained the DBLα domain was 40 (∼74% of the expected) generating a total of 165 kb bases (∼37% of the expected ∼450 kb) lower than reported for simulated reads (Num. contigs with DBLα=50, sum of contigs=293 kb). A closer look at errors and read coverage shows that regions with highest errors correspond with contig breakpoints (Figure 2.2). In addition, low and uneven coverage also affected assembly quality. Next, the effect of repeated sequences on *var* assembly was investigated. Although the most commonly shared sequences were smaller than 100 bp, stretches of above 1 kb were also identified from the pairwise blast search of 3D7 *var* genes (Figure 2.3). A visual inspection of the underlying de Bruijn graph

revealed extremely dense nodes which represented highly similar sequences within members of the family (Figure 2.4A). The effect of repetitive sequences on the *var* assembly graph was examined by removing reads that aligned to genes that contained shared sequences above the fragment size of 200 bp. The graph was significantly simplified although still far from ideal (Figure 2.4B). However, such simplifications are likely to affect quality by generating gaps in the final assembly.

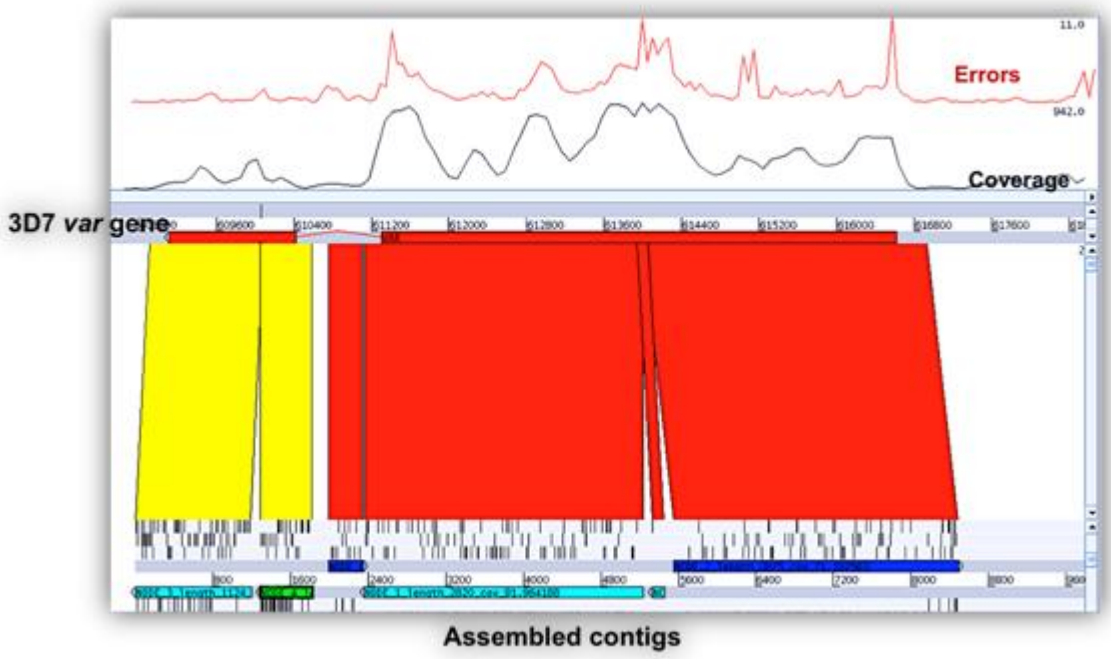

Figure 2.2: The effect of sequencing errors and uneven coverage on assembly of *var* genes. A comparative view of assembled contigs with *var* genes of the reference genome is shown. The top panel shows mismatch (error) count plot (red) and read mapping coverage plots (black). The bottom panel shows assembled contigs (bottom) that were ordered and orientated against a 3D7 *var* gene (top). The red and yellow blocks represent synteny matches. The color of contigs indicates whether they align in the forward (green) or reverse (blue) strands and if there is an overlap between neighboring contigs (cyan). Black bars represent stop codons.

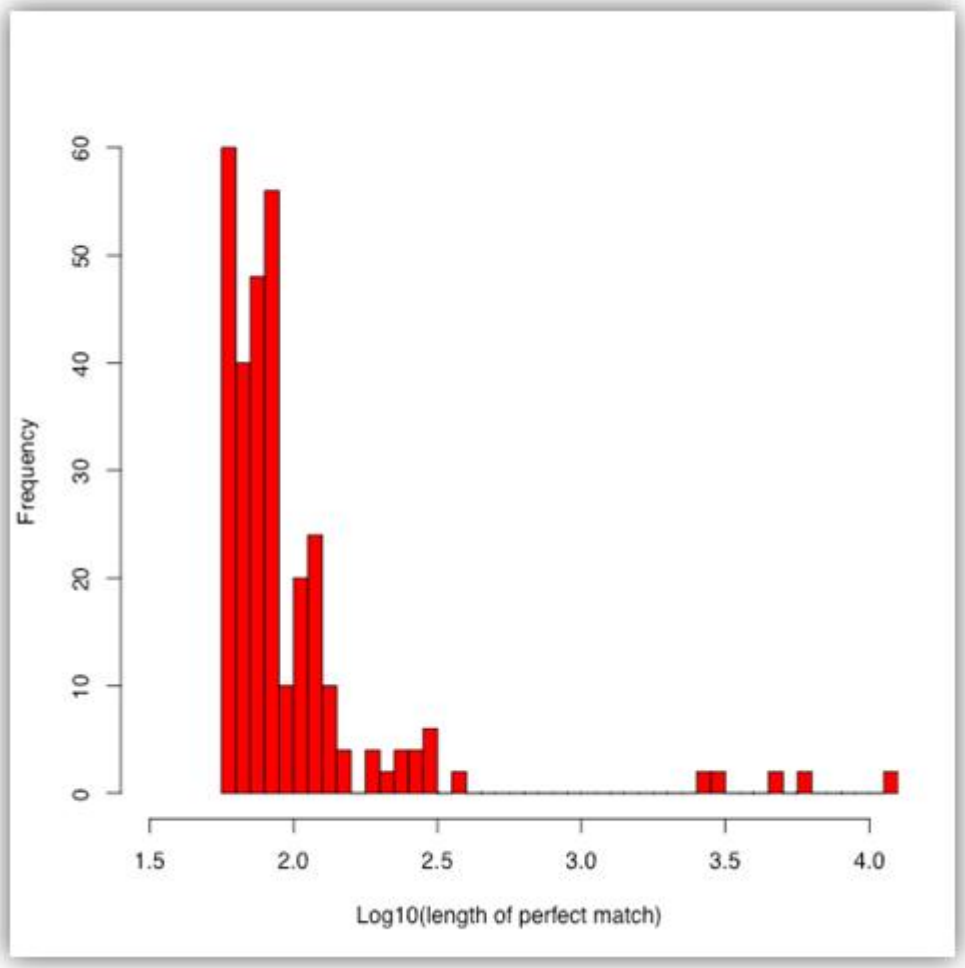

Figure 2.3: A histogram of shared sequences in 3D7 *var* genes identified by a pairwise blast alignment. Perfectly matching sequence blocks of length up to 4 kb were found between *var* genes of the 3D7 genome.

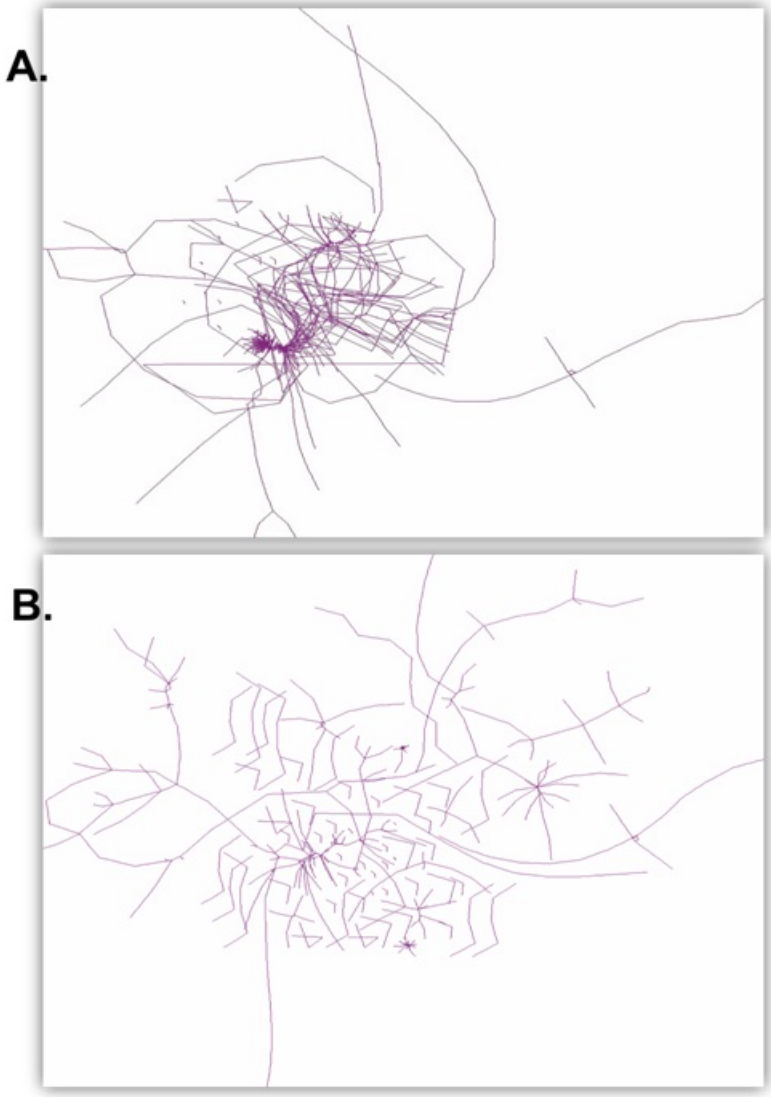

Figure 2.4: Visualizing the assembly graph of all *var* genes from a velvet assembly (real reads, k=61). In these plots, node sequences were represented as lines and curves that are joined with other nodes at their tips. The dense regions (nodes) on the graph represent repetitive sequences that have multiple connections with other nodes A) The assembly graph of real reads that align to *var* genes of the the 3D7 genome. B) A simplified assembly graph after removing reads that align to shared sequences between *var* genes with a match length of above 200 bp (as shown in Figure 2.3).

2.3.5 Reference guided assembly

In addition to *de novo* assembly, a possibility of using mapping based assembly approaches was investigated. The challenge to genome assembly posed by the inherent features of the genome (particularly the high A+T bias) is illustrated using a *k-mer*-based uniqueness plot computed by counting the frequency of sequences of length 30 bp. The uniqueness plot indicates how well short reads could be uniquely mapped to genomic regions. The correlation between sequence complexity, read coverage and G+C content is shown for chromosome 1 (Figure 2.5) and the left subtelomeric region of chromosome 8 (Figure 2.6). The increase in G+C indicates higher information content and therefore mappabiliy. Although both alignment and assembly benefit from paired-end information, current assemblers construct graphs using *k-mer*s from all reads independently. Read pair information is used at later stages of the assembly to simplify the graph and resolve repeats. A *k-mer*-based uniqueness plot therefore shows regions of the genome where the assembly would terminate contigs due to ambiguity. The subtelomeres of *P. falciparum* contain repeat blocks which have a lower uniqueness compared to core regions of the genome. The overall spikiness of uniqueness and G+C content affects mapping coverage across the genome more specifically in subtelomeric regions (Figures 2.5 and 2.6).

Despite the problem of uniqueness in the sub-telomeric regions, at first sight it appears that the relatively high G+C content and uniqueness of the *var* genes should aid their assembly by mapping. However the extreme polymorphism of these genes presents a much greater additional problem when sequence reads from different genotypes are used. Figure 2.7 demonstrates this point by comparing the homologous mapping coverage of the reference genotype 3D7 to the coverage of the reference from three field isolates of *P. falciparum* and to its closest known relative, the chimpanzee parasite *P. reichenowi*.

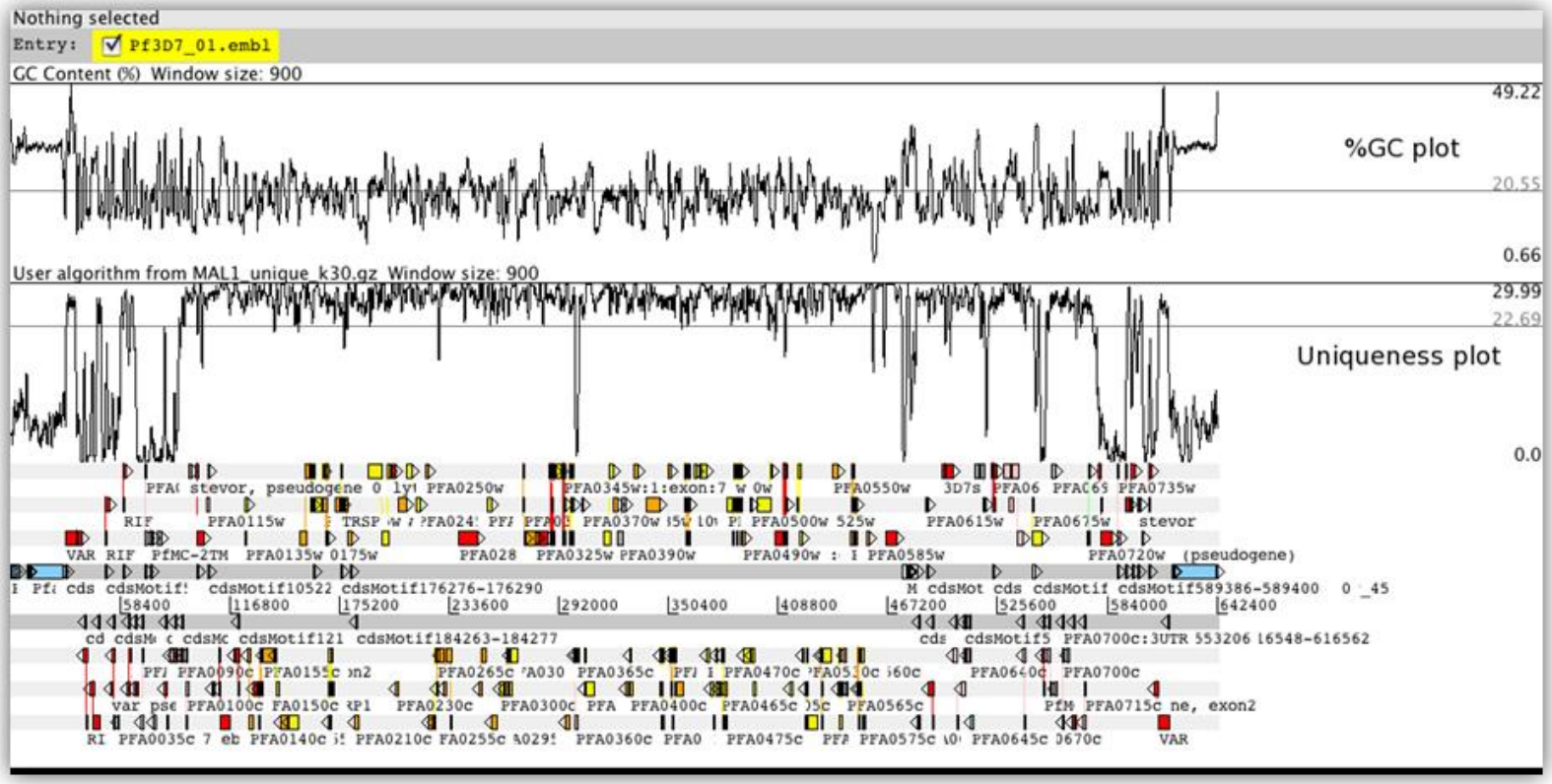

Figure 2.5: A plot of G+C content and uniqueness (based on a *k-mer* size of 30 bp) on Chromosome 1 of the 3D7 genome. An Artemis view of G+C content (top panel) and a uniqueness plot (middle panel). The bottom panel shows annotation information with the different blocks representing protein coding genes, pseudogenes and repeats in all the six reading frames.

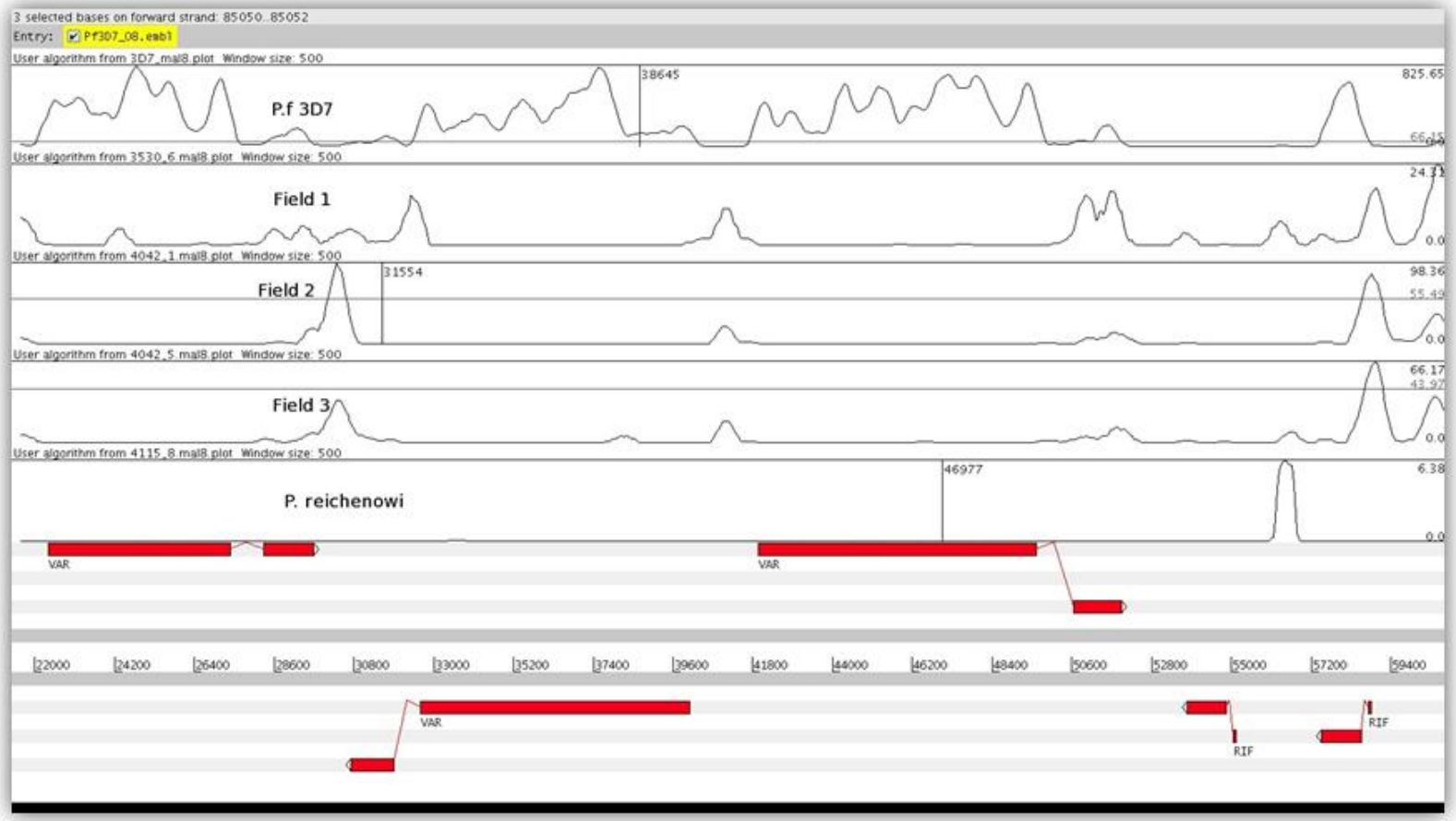

Figure 2.6: Read mapping coverage plots over *var* genes on the left subtelomere of Chromosome 8. Illumina reads from 3D7, three field samples and *P. reichenowi* were aligned to the 3D7 genome.This figure shows the difficulty of reliably aligning reads obtained from clinical samples to *var* genes (shown in red over the forward and reverse strands) of the reference genome 3D7.

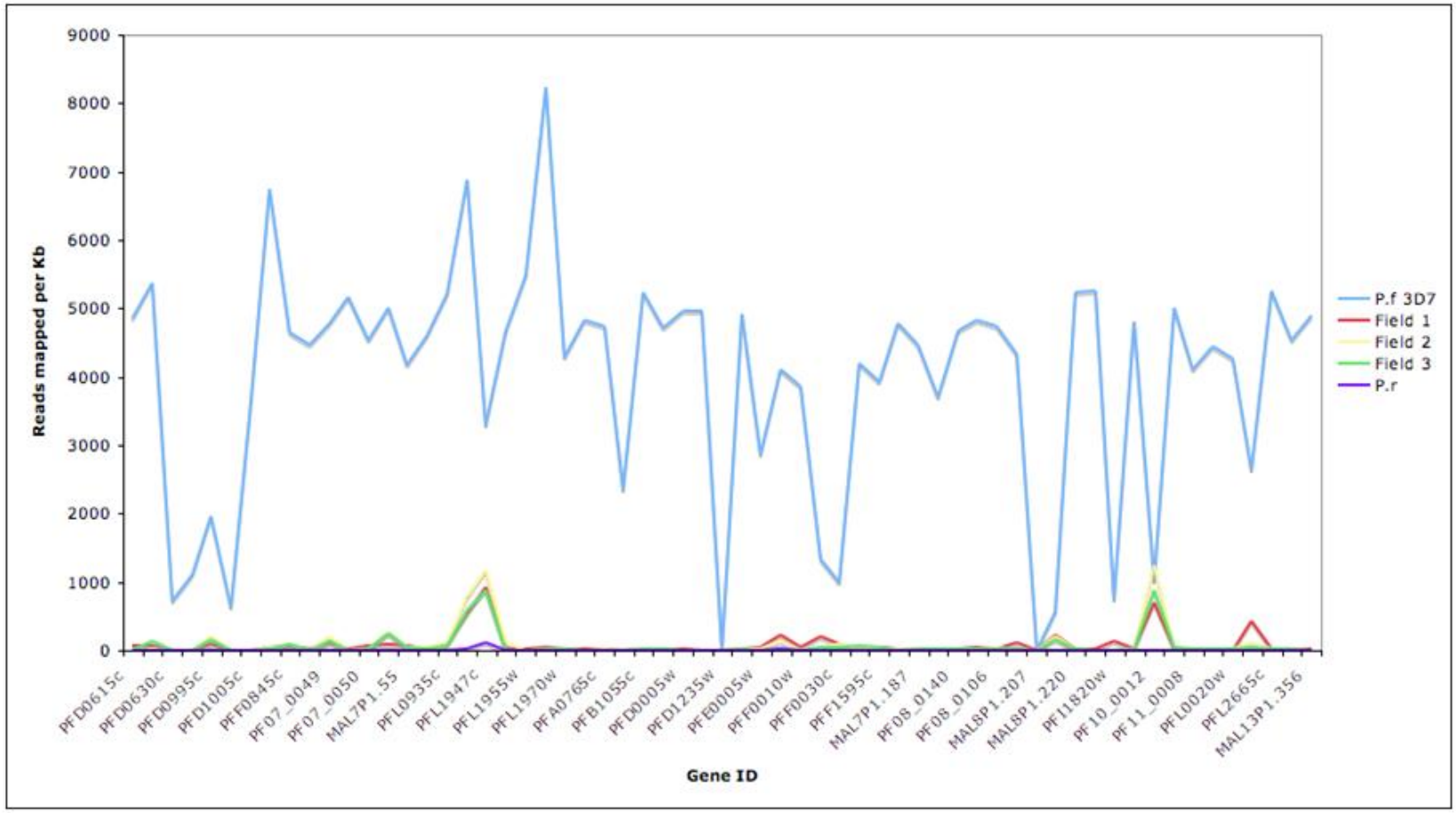

Figure 2.7: Number of reads mapped per kb over *var* genes of the 3D7 genome. Reads from 3D7, three field samaples and *P. reichenowi (P.r)* were uniquely aligned to version 2.1.4 of the reference genome 3D7. A count of reads mapped per kb is shown for the five genomes. The effect of repetitive sequences (shared matches between *var* genes) is shown by the lack of coverage over some *var* genes. Read mapping was very poor for field samples due to high polymorphism in *var* genes.

2.3.6 Understanding sequencing errors in *P. falciparum*

Following observations that suggest a role for sequencing errors in assembly quality, one aim of this section was to look at the extent and distribution of substitution errors in *P. falciparum* sequencing.

2.3.6.1 Overall sequencing errors

Initially, sequencing errors were independently computed for the forward and reverse reads at low ($Q5$), medium ($Q15$) and high ($Q25$) quality bins. Error rates were variable between runs and lanes (Figure 2.8). Low quality bins had higher error rates in all libraries. The variation in errors on the forward and reverse reads was not consistent between instruments. For examples, the Genome Analyser showed higher error rates on the second read at low ($Q5$) quality bins with the exception of NP_i500. On the other hand, error rates were comparable between the forward and reverse reads in HiSeq and MiSeq at all quality bins.

Libraries from the HiSeq 2000 instrument had the lowest error rates ($\sim$0.7%) compared to the Genome Analyser and MiSeq. The four MiSeq libraries (MS_i3K.1-4) had the highest error rates ($\sim$1 to 1.3%) in both the forward and reverse reads across all quality bins. However, these were part of a research and development experiment on long insert protocols and had a lower yield (Table 2.1). They were therefore excluded from further comparisons, as they may not reflect the quality of standard production libraries. The remaining six libraries were used for the analyses described in the following sections.

2.3.6.2 Per-cycle error rates

In order to identify positions that are particularly prone to errors in *P. falciparum*, error rates were computed for each position on both the forward and reverse reads (Figures 2.9 and 2.10).

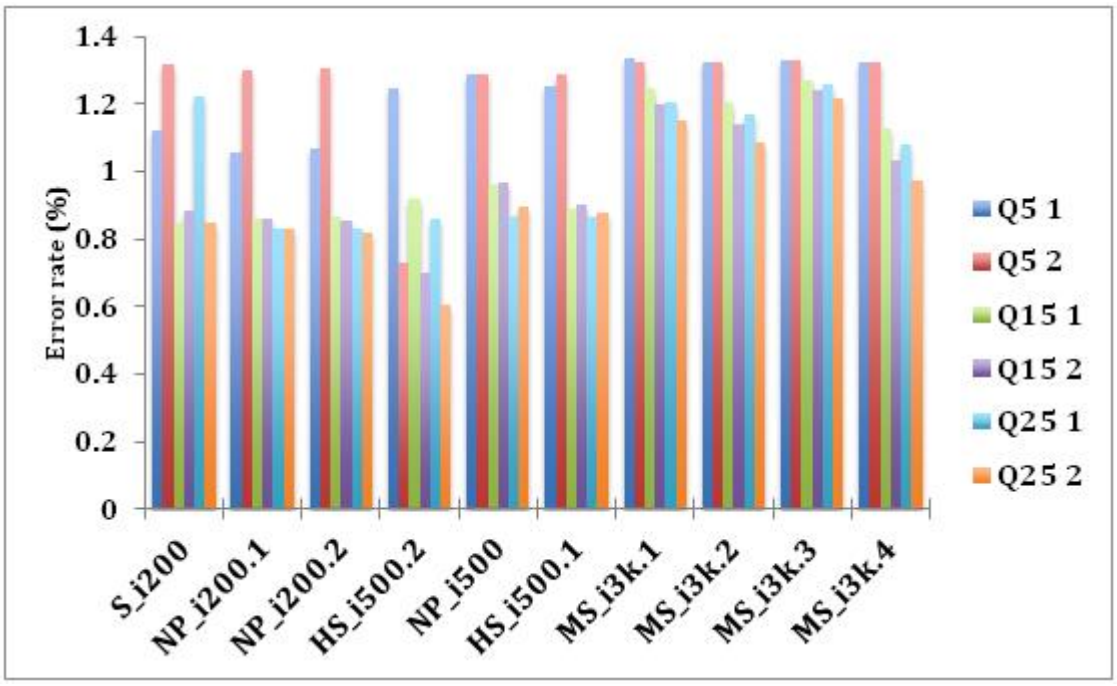

Figure 2.8: Overall error rates at low ($Q5$), medium ($Q15$) and high ($Q25$) quality bins for forward (eg. $Q5$ 1) and reverse (eg/ $Q5$ 2) reads. Error rates were computed for 10 libraries sequenced on GAII, HiSeq and MiSeq platforms representing standard and PCR-free library preparation protocols.

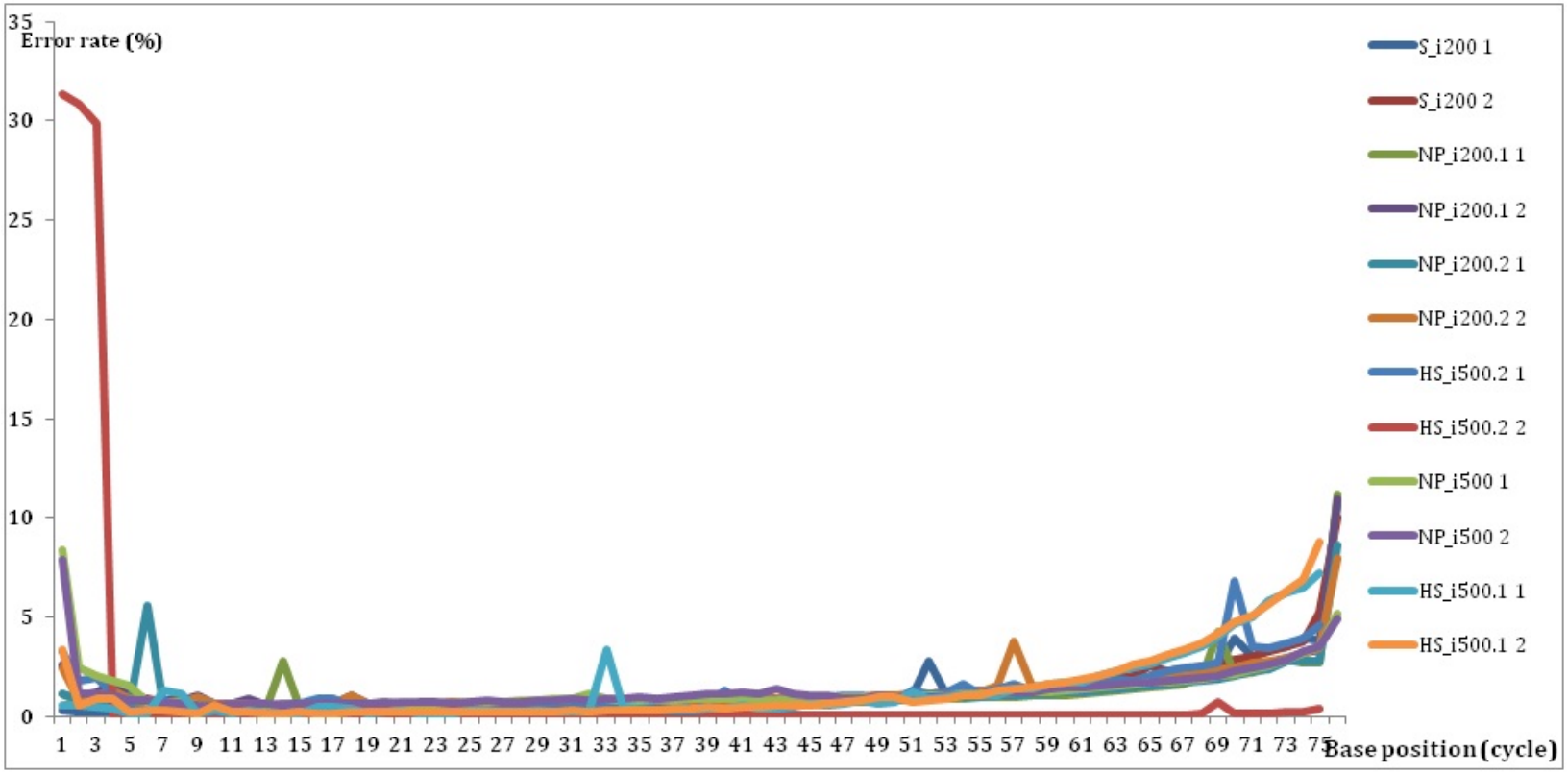

Figure 2.9: Error rates per-cycle for Illumina's GA2 and HiSeq platforms at a quality cutoff of 5 (Q5). Here, error rates (y-axis) were computed for each cycle (x-axis) or base position of the forward and reverse reads. Forward reads are represented by the suffix "1" (eg S_i200 1 represents the forward reads of the library S_i200).

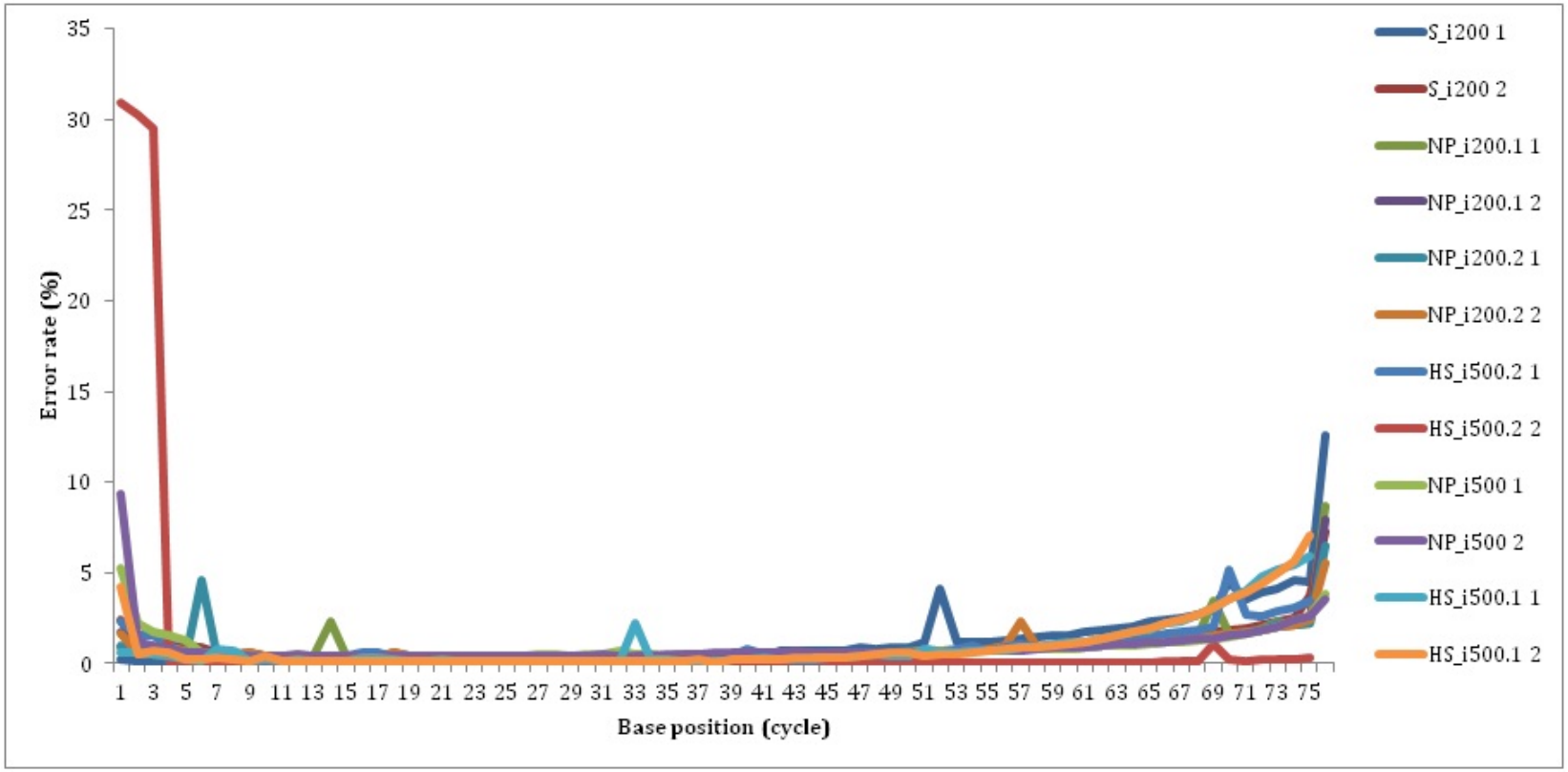

Figure 2.10: Error rates per-cycle of Illumina's GA2 and HiSeqplatforms at a quality cutoff of 25 (Q25). See Figure 2.9 for details.

The highest proportion of errors was found on the final five to seven positions. Surprisingly, a similar trend of increased errors were also observed at the beginning of reads in all libraries affecting both low ($Q5$) and high ($Q25$) quality bins. In five of the six libraries, errors in the first five bases accounted for $\sim$1 to $\sim$17% of the total while the last seven bases contributed a higher 24 to 43% of the total. The second read of the HiSeq library HS_i500.2 had an exceptionally high error rate in the first three bases causing $\sim$92% of all errors. Although error rate in the rest of the positions was very low, high error rates concentrated around a few bases will still affect the overall quality of the read. The occasional spikes in error rates such as those found on base 33 of the first read in library NP_i500.1 were potential indicators of random errors due to a number of reasons including tile-specific problems and issues associated with imaging. Despite the decrease in values as quality increases, the trends of error rates per cycle were consistent at low and high quality bins.

In addition to quantifying the extent of errors for each cycle of the sequencing process, patterns of substitution were investigated in order to understand systematic sources of bias. Proportion of errors due to the 12 potential substitutions (A$\rightarrow C, A \rightarrow G, A \rightarrow T \ldots T \rightarrow A, T \rightarrow C, T \rightarrow G$) were computed for each read of the six libraries (Figure 2.11A and 2.11B). Substitutions A-T, T-G and A-G were over-represented at low quality bins while T-G, A-T, A-G and T-C dominated high quality substitutions.

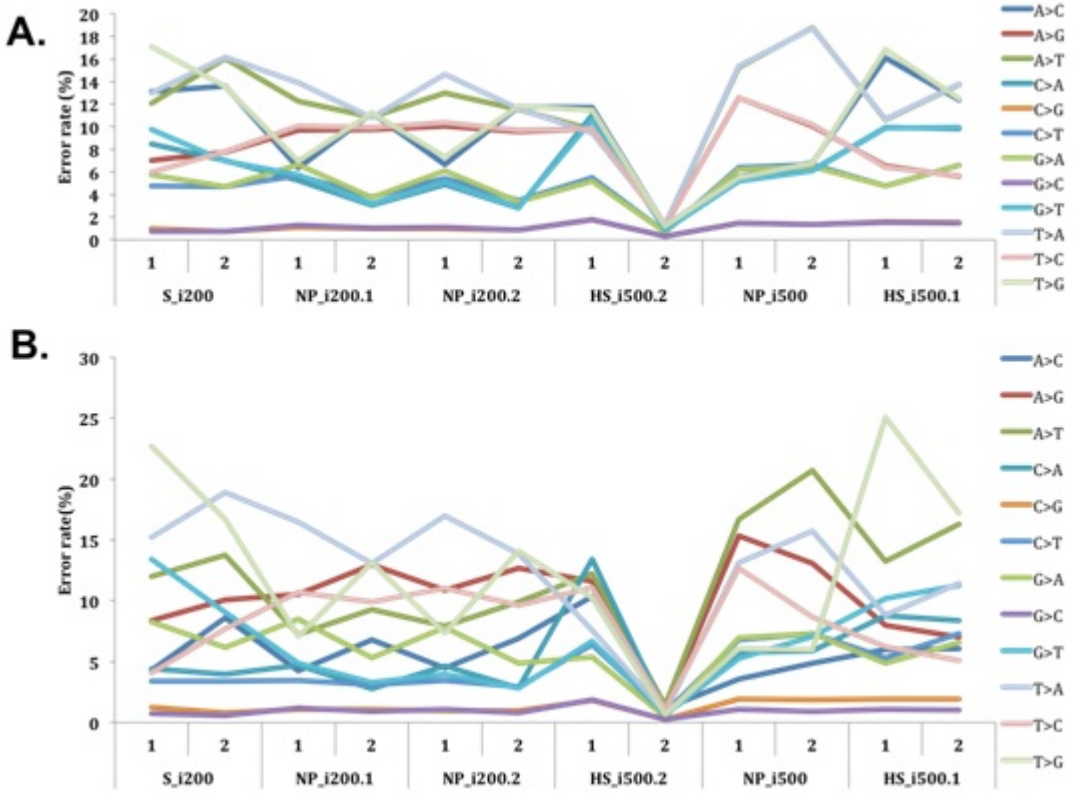

Figure 2.11: Overall substitution patterns of Illumina reads from six libraries. Substitution profiles of the forward and reverse reads of five libraries (as shown on the x-axis) and the rate of such substitutions were plotted for lower base quality of 5 (A) and a higher base quality cutoff of 25 (B).

2.4 Discussion

This chapter was aimed at evaluating the feasibility of using existing short read assembly methods to reconstruct the *var* gene family in *P. falciparum*. A method that works on *var* genes is expected to be effective on other gene families of *P. falciparum* such as the *rif* and *stevor* gene families.

Conclusion 1: *A comparison of short read assembly programs suggested that Velvet is a better choice than SOAPdenovo. However, assembly results from Velvet were not satisfactory.*

A larger proportion of the sequencing and analysis for this chapter was done in 2009 where development of short read assembly tools was in its early days. Assemblers that employ a de Bruijn graph structure were preferred over those with the traditional overlap-layout-consensus approach due to their potential to deal with the increase in sequence yield from next generation sequencing technologies. Although the underlying algorithm used to represent sequences is similar, de Bruijn graph-based assemblers have subtle differences in their pre and post graph processing of reads. For instance, SOAPdenovo applies a read filtering and error correction step based on a predefined *k-mer* frequency cutoff prior to building the graph. On the other hand, Velvet uses a similar approach to remove erroneous reads without correcting base calling errors. In addition, Velvet puts an extra effort in simplifying the assembly graph during both the construction stage and the assembly process by removing singletons. Further reduction in graph complexity and error correction is achieved by removing tips (a node or chain of nodes that have one loose end), bursting bubbles (i.e. merging paths based on sequence similarity and minimum number of reads represented by each path), removing paths that have fewer than the minimum cutoff and using read pair information. The extremely high A+T content of *P. falciparum* poses unique challenges in short read assembly and requires advanced heuristics of resolving ambiguous paths at various stages of the process.

Velvet's superior assembly results compared to SOAPdenovo were therefore attributed to its ability to simplify the assembly graph and remove erroneous paths. However, further evaluation of Velvet on real reads from whole genome, chromosome 1 and *var* genes of the 3D7 genome revealed that the results are not satisfactory as they are highly fragmented. The *k-mer* size was found to be a very important parameter that needs optimisation for a good quality assembly. Smaller *k-mer* values cause potential false overlaps that lead to ambiguities and assembly breaks. Conversely, larger *k-mer* sizes could generate assemblies with a higher contiguity, they require long overlaps and a higher read coverage.

Filtering reads with mismatches did not significantly improve the whole genome *de novo* assembly. This is due to a number of potential problems associated with read length, fragment size, low complexity, uneven coverage and

ambiguous overlaps from repeats that could not be resolved by the assembler. In addition, the decrease in coverage due to the filtering may also affect assembly quality. The overall results are consistent with previous whole genome *de novo* assembly attempts using a PCR-free sample preparation (Kozarewa, et al., 2009). A whole genome *de novo* assembly from a PCR-free library of *P. chabaudi* - a *Plasmodium* species with a higher G+C content - was substantially better (Thomas Otto, personal communication). Therefore low G+C content is one of the primary limiting factors in short read assemblies of *P. falciparum*.

Conclusion 2: *A reference-guided assembly was also not feasible due to a higher level of polymorphism than is acceptable by methods that employ a comparative assembly approach.*

In order for a reference-guided (or mapping based) assembly approach to work, a nucleotide similarity of above ∼90% is required (Pop, 2009). Subtelomeric regions of and *var* genes of *P. falciparum* are very divergent and not positionally conserved. The concept of using a fixed linear reference to guide the assembly of new sequences is thus not relevant. One consequence of the lack of mapping of reads from different genotypes to the telomeres is that all the reads that do not map to the reference will include those that cover the most polymorphic regions. It is therefore clear that assembly by mapping will be of little value when dealing with such highly polymorphic regions of the genome.

Conclusion 3: *Poor quality assembly of short reads was caused by a combination of technical/systematic reasons from the sequencing process and inherent features of the genome. Technical reasons include sequencing errors and uneven coverage due to an enzyme's limited capability to amplify certain regions of the genome. On the other hand, inherent genome features include biased A+T content and repeated sequences.*

Standard quality control pipelines of the Illumina platform often report error rates computed from a fraction of sequenced reads. Such reports provide a quick overview of quality in order to decide whether the data could be used for downstream analyses. However, the information is not enough to understand where the errors occur and whether some of the data could be

recovered. For instance, a closer look at a lane labeled as 'failed' may identify a subset of reads or cycles that contribute to the increased error rate which could then be excluded from further analysis. It is therefore important to further investigate error profiles for each position on the reads at low and high quality cutoff. Although sequencing errors are known to accumulate towards read-ends (Abnizova et al., 2012), it was surprising to see increased error rates at the beginning of reads.

In addition, unexpected high quality substitution errors specific to *P. falciparum* sequences were observed. The Illumina platform uses two lasers to initiate emission of fluorescence from four channels (A, G, C, T) where A,C and G,T pairs share each laser. Although Illumina's base calling algorithm, Bustard, uses a 16-parameter correction matrix (the cross talk matrix) to account for responses from sources other than incorporation of the intended base, cross talk effects are still visible at lower quality errors particularly for transversions (A-C and G-T). The observed high frequency of substitutions A $\leftrightarrow$ *G*, C $\leftrightarrow$ *T* and A $\leftrightarrow$ *T* is therefore less likely to be due to the crosstalk effect. This is potentially due to the extreme base composition and requires a special attention in applications such as variant calling. However, it is difficult to measure the effect of high/low quality substitution errors in assembly as short read assemblers have yet to take advantage of base quality information. Currently, such errors could be accounted for during the assembly process by trimming-off the first and last error prone bases. However, assembly tests performed by trimming read-ends did not improve the quality of contigs for two potential reasons. Firstly, the Velvet assembler has an efficient algorithm of removing erroneous nodes created due to sequencing errors that accumulate towards read-ends. Trimming of reads may thus have very little improvement over the initial assembly. Secondly, as trimming shortens the effective read length, the size of *k-mer*s that could be used for assembly also becomes smaller. As described previously, shorter *k-mer* sizes are likely to generate false overlaps and poor quality assembly.

Conclusion 4: *A slightly different approach to reference guided and whole genome or whole chromosome de novo assembly was required to reconstruct var genes and subtelomeric regions from the current sequence data.*

In summary, even when the best available methods were used, *var* genes could not be reliably assembled. Optimizing for assembling gene-families in particular and removing technical errors improved the results. Even so, technical and inherent bias meant that the assembly remains challenging, and we conclude that none of the current methods can effectively assemble highly polymorphic gene-families.