Appendices

Appendix A:
Storage space estimates for an iterative assembly of *var* genes in *P. falciparum*

ID (from Figure A-1)	Process	Disc Space per process (Gb)	Temporary storage (Gb)	Sum (Gb)
1	Get non-core reads	1.5	0	1.5
2	Get var-reads	0.4	0.4	
3	Assembly	15	15	
4	Scaffolding	15	15	
5	IMAGE	12	12	
6	Scaffolding	15	15	
7	IMAGE	12	12	
	Output files (Permanent storage)			0.5 2

Table A-1: Storage (Disc) space estimates for one lane of sequence data per iteration. Disc space required for each step of the seven processes are shown for one iteration. Although temporary files were deleted after each process (as shown in column 4), processing ∼800 samples (Chapter 5) required up to ∼12 Tb of temporary storage space at a given time.

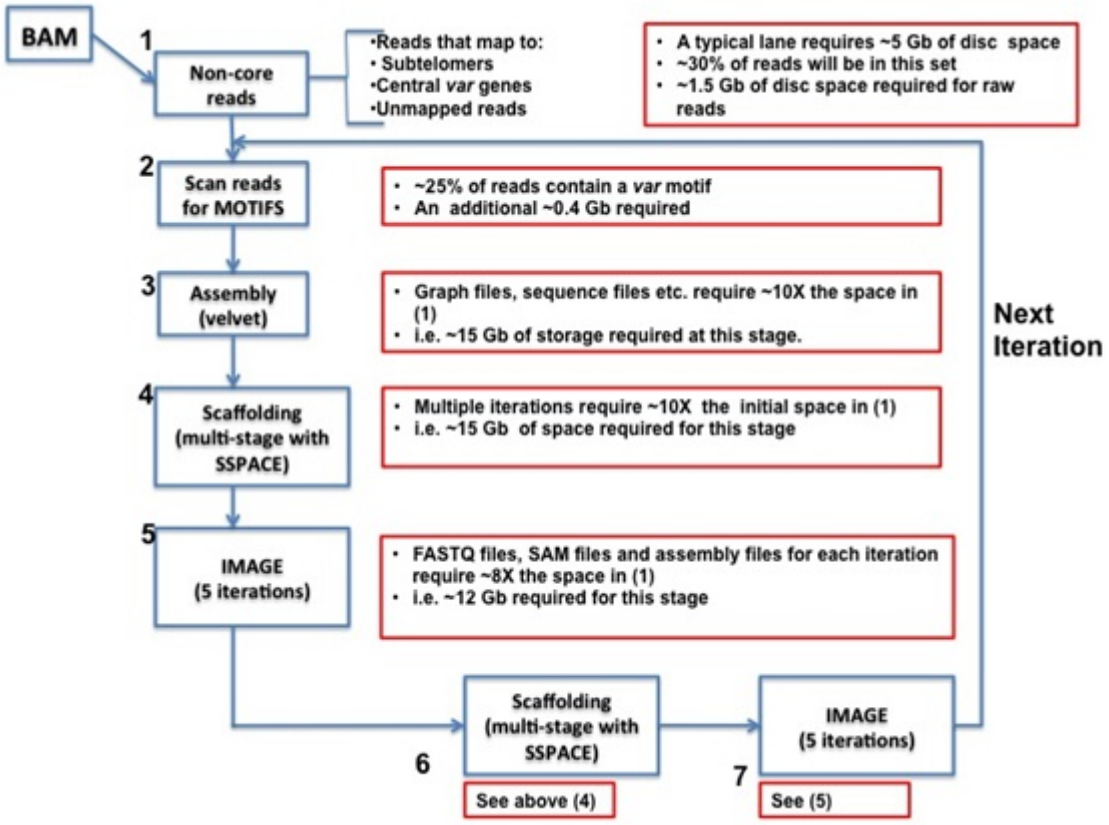

Figure A-1: Storage (Disc space) estimates for one iteration of *var* assembly per sample. Temporaty and permanent storage required for a single iteration of assembly is shown (red boxes) for each process. Additional information on the temporary and permanent storage estimates is shown in Table A-1.

Appendix B:
Quality of raw data and its effect on assembly

Quality and insert size plots were obtained from the Sanger Institute's raw data quality control archives.

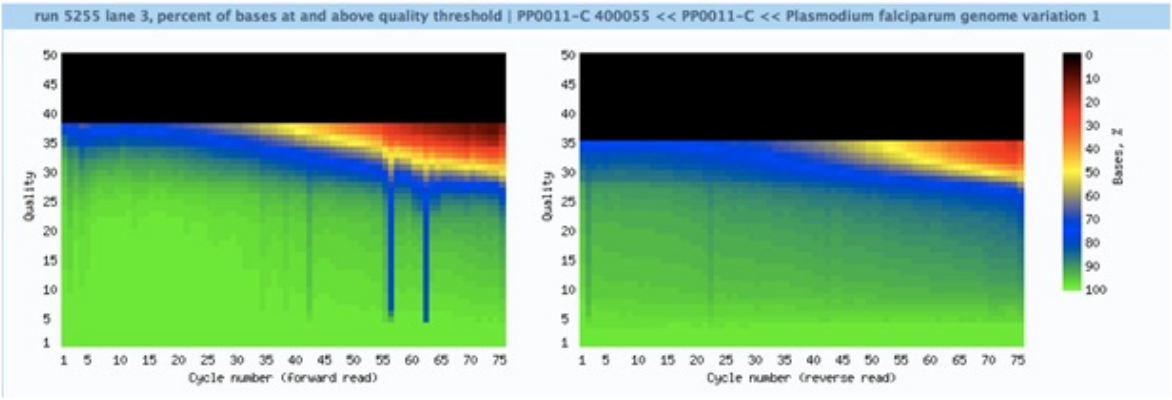

Figure B-1: Base quality plot for the forward (left panel) and reverse (right panel) reads of sample PP0011. Assembly results for PP0011 (Chapter 3, 50 clinical samples) were affected by the drop in quality on cycles 55 and 73. The number of *var*-contigs was 26 for PP0011 (Chapter 3, section 3.3.3). The quality of each base (y-axis) is shown for the 76 cycles (x-axis). The colors represent the percentage of bases that have a quality value shown on the y-axis.

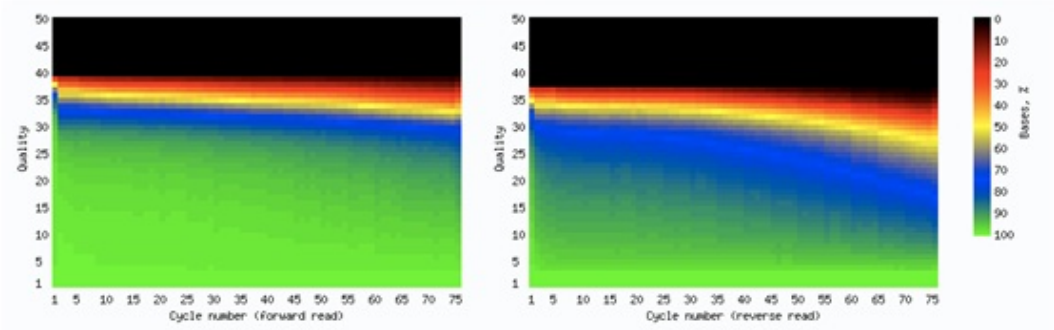

Figure B-2: Base quality plot for the HB3 genome used in Chapter 3. The second read (right panel) shows a decay in quality from cycle ∼45 onwards. See Figure B-1 for description. The HB3 genome had a significantly higher number of non-core reads due to the larger proportion of reads that did not align to the 3D7 genome (∼33%). The decrease in quality on the second read has affected the number of reads that aligned to the reference genome (Chapter 3, section 3.3.2.1).

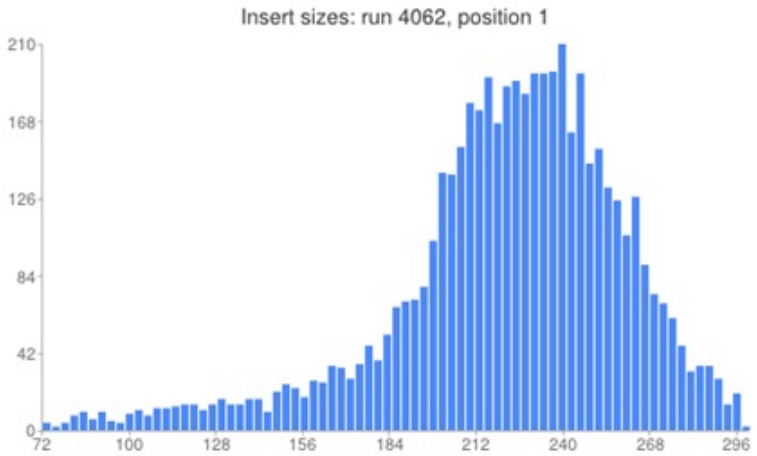

Figure B-3: Hb3 insert sizes (normal sizes)

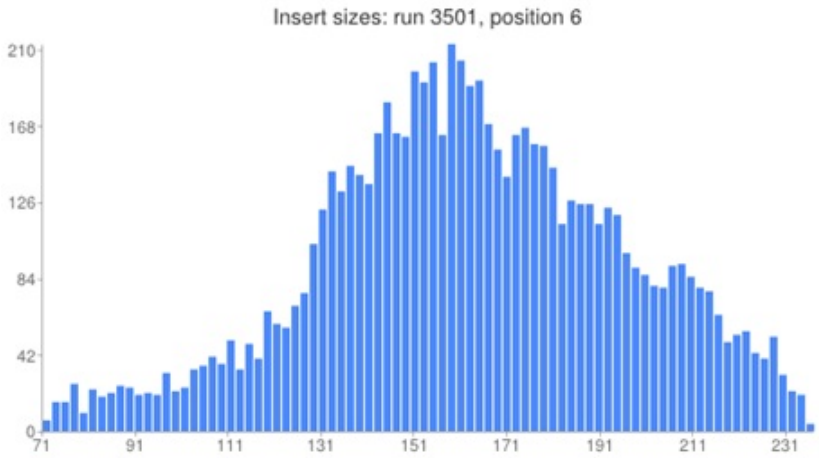

Figure B-4: A distribution of insert sizes for the 3D7 genome used in Chapters 2 and 3. The poor quality assembly for *var* genes of the 3D7 genome (Chapter 3, testing the iterative assembly approach on culture-adapted samples) was partly due to the shorter than the expected insert sizes. As shown here, the actual insert sizes were ∼160 bp, while the expected sizes were 200-300 bp.

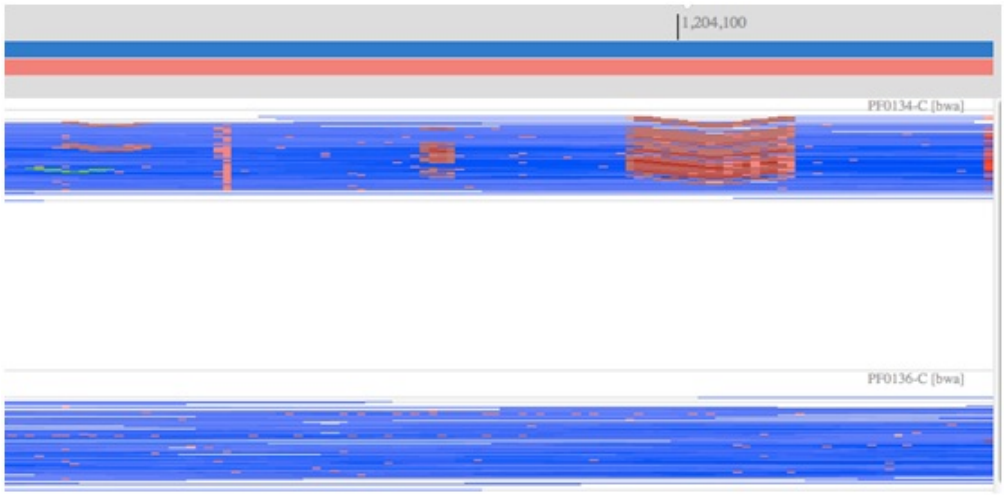

Figure B-5: Number of *var*-contigs is indicative of multiplicity of infection: Samples PF0134-C and PF0136-C had 262 and 65 *var*-contigs respectively. This figure shows a LookSeq view of SNPs (shown in red) for the two samples (Top panel for PF0134-C and bottom panel for PF0136-C). Read coverage and SNPs over the MSP1 gene show segregating haplotypes for sample PC0134-C suggesting multiple infections.

Appendix C:
Availability of Software developed to assemble *var* genes

Software developed in this thesis, additional scripts and documentation are available from:

`https://sourceforge.net/projects/varassembly/`