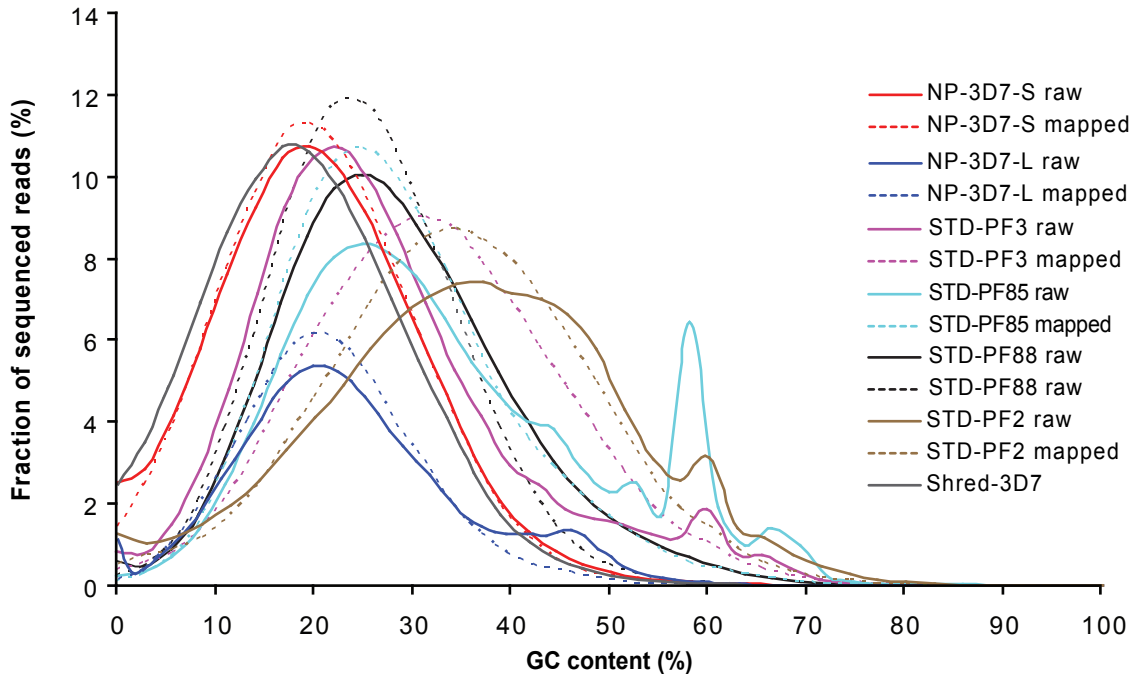
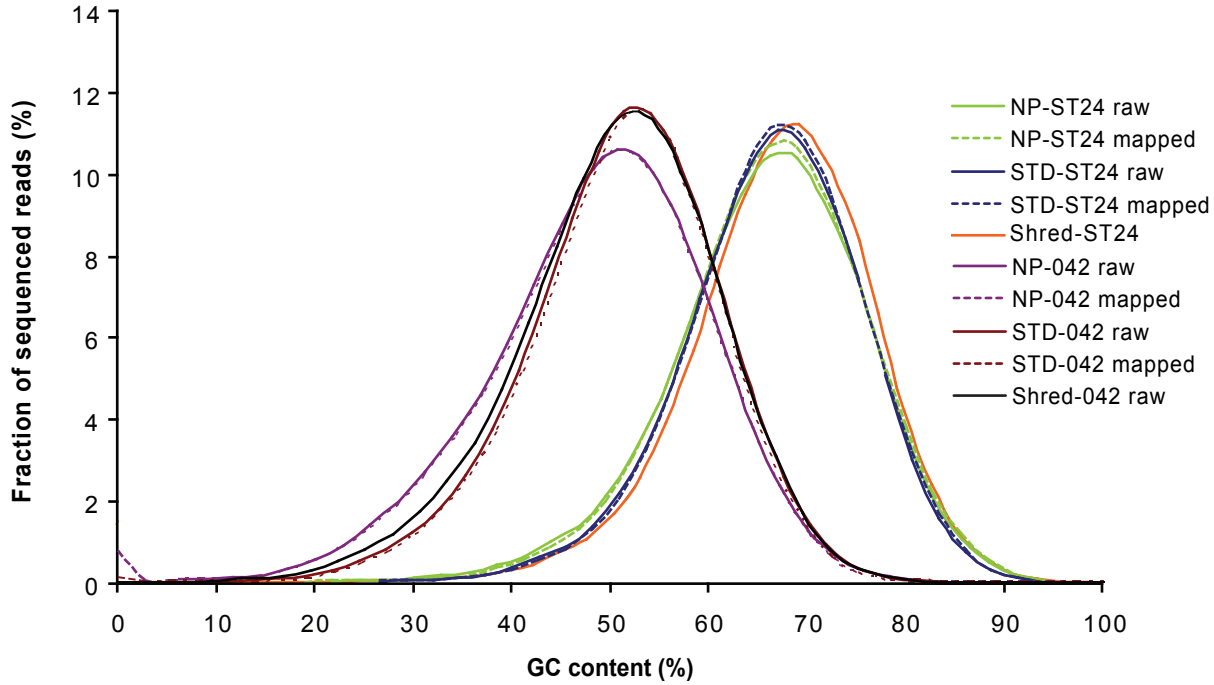
Supplementary Figure 1

a)

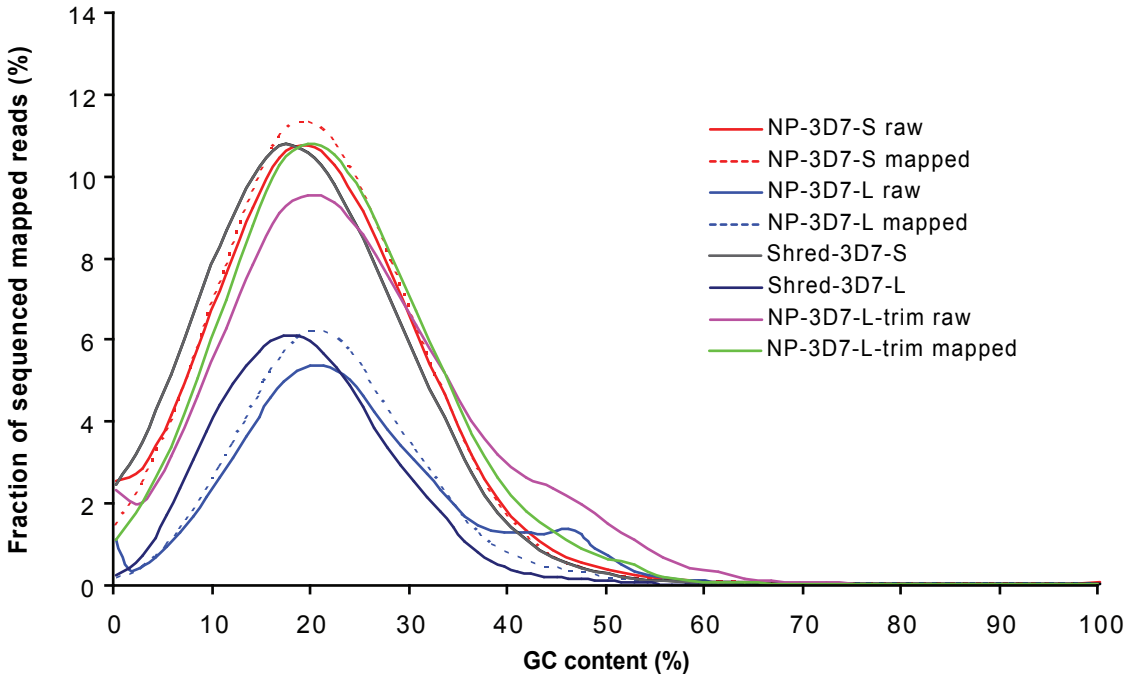


b)



Supplementary Figure 1. Distribution of sequenced reads for different values of GC content. (**a**) GC profiles for raw and mapped sequence data for all *P. falciparum* libraries, along with simulated data ('Shred-3D7') for comparison. GC levels are calculated in a window size of read length, so the peak of fraction reads is dependent upon the read length. If, for example, read length = 1 bp, there would be two points in the GC curve, corresponding to the mean % AT at 0 % GC, and the mean % GC at 100 % GC. Conversely, for the NP-3D7-L data, read length is approximately twice that of the other datasets, so and so the peak height is halved. (**b**) GC profiles for *E. coli* and *B. pertussis* datasets alongside simulated data ('Shred-ST24' and 'Shred-042').

Supplementary Figure 2



Supplementary Figure 2. Trimmed 76-base reads.

We trimmed the 76 bp reads back to 36 bp and plotted the fraction of mapped reads against % GC. It can be seen that mapped curve (NP-3D7-L-trim mapped) agrees well with the Shred-3D7-L curve. For the trimmed raw data (NP-3D7-L-trim raw), however, we see a tail shifting away from the theoretical predictions, showing that the data from the 76 bp run has a higher level of bias than the 36 bp run. Because the same library was used in both runs, it indicates a problem with the longer sequencing run.

Supplementary Table 1

Library	Organism	Strain	Genome size (Mb)	Insert size (bp)	Run_id	Sequencing platform	Read length (bases)	Number of tiles
NP-3D7-S	*P. falciparum*	3D7	23	200	1022_s_5	GAI	36	300
			23		1022_s_8		36	300
NP-3D7-L	*P. falciparum*	3D7	23	200	1105_s_5	GAII	76	100
			23	200	1105_s_7	GAII	76	100
STD-PF88	*P. falciparum*	clinical isolate	23	200	883_s_1	GAI	36	330
			23		883_s_2		36	330
			23		883_s_3		36	330
			23		883_s_5		36	330
			23		883_s_6		36	330
			23		883_s_7		36	330
			23		883_s_8		36	330
STD-PF3	*P. falciparum*	3D7	23	200	368_s_1	GAI	35	330
			23		368_s_2		35	330
			23		368_s_3		35	330
			23		368_s_5		35	330
			23		368_s_6		35	330
			23		368_s_7		35	330
			23		368_s_8		35	330
STD-PF2	*P. falciparum*	3D7	23	200	245_s_1	GAI	35	330
			23		245_s_2		35	330
			23		245_s_3		35	330
			23		245_s_5		35	330
			23		245_s_6		35	330
			23		245_s_7		35	330
			23		245_s_8		35	330
STD-PF85	*P. falciparum*	3D7	23	200	851_s_1	GAI	35	330
			23		851_s_2		35	330
NP-042	*E. coli*	042	5.3	200	1022_s_1	GAI	36	300
STD-042	*E. coli*	042	5.3	200	1603_s_5	GAII	37	100
NP-ST24	*B. pertussis*	ST24	4.0	200	1022_s_3	GAI	36	300
STD-ST24	*B. pertussis*	ST24	4.0	200	1603_s_7	GAII	37	100
			4.0		1603_s_8		37	100

Library	Mean clusters per tile	Mean PF clusters per tile	Total PF clusters	% PF clusters	Number of reads/lane	Number of unique PE reads[†]	Total number of reads
NP-3D7-S	4.3E+04	2.2E+04	6.7E+06	51.2	1.3E+07	1.3E+07	
	3.7E+04	2.4E+04	7.3E+06	66.1	1.5E+07	1.4E+07	2.8E+07
NP-3D7-L	7.1E+04	5.3E+04	5.3E+06	74.6	1.1E+07	1.0E+07	
	5.7E+04	4.5E+04	4.5E+06	78.3	8.9E+06	8.0E+06	2.0E+07
STD-PF88	4.7E+04	2.4E+04	8.0E+06	51.8	1.6E+07	1.6E+07	
	4.6E+04	2.5E+04	8.3E+06	55.1	1.7E+07	1.6E+07	
	4.5E+04	2.5E+04	8.2E+06	55.0	1.6E+07	1.6E+07	
	4.6E+04	2.4E+04	8.1E+06	53.5	1.6E+07	1.6E+07	

	4.5E+04	2.4E+04	7.9E+06	53.1	1.6E+07	1.6E+07	
	4.5E+04	2.3E+04	7.7E+06	51.1	1.5E+07	1.5E+07	
	4.5E+04	2.2E+04	7.4E+06	49.0	1.5E+07	1.5E+07	1.1E+08
STD-PF3	3.0E+04	1.5E+04	5.1E+06	51.9	1.0E+07	8.7E+06	
	2.8E+04	1.4E+04	4.7E+06	51.1	9.4E+06	7.8E+06	
	2.9E+04	1.8E+04	6.0E+06	63.6	1.2E+07	1.1E+07	
	2.4E+04	1.6E+04	5.3E+06	68.3	1.1E+07	9.2E+06	
	2.4E+04	1.7E+04	5.7E+06	72.2	1.1E+07	1.0E+07	
	2.4E+04	1.6E+04	5.4E+06	69.4	1.1E+07	9.7E+06	
	2.3E+04	1.6E+04	5.3E+06	69.0	1.1E+07	9.5E+06	7.5E+07
STD-PF2	1.6E+04	1.1E+04	3.5E+06	66.2	7.0E+06	6.2E+06	
	1.5E+04	1.0E+04	3.3E+06	69.7	6.7E+06	5.6E+06	
	2.3E+04	1.6E+04	5.4E+06	69.4	1.1E+07	1.1E+07	
	2.1E+04	1.5E+04	5.1E+06	73.7	1.0E+07	9.8E+06	
	2.1E+04	1.6E+04	5.2E+06	75.6	1.0E+07	1.0E+07	
	1.8E+04	1.4E+04	4.5E+06	76.4	8.9E+06	8.3E+06	
	1.8E+04	1.3E+04	4.4E+06	75.6	8.8E+06	8.4E+06	6.3E+07
STD-PF85	1.5E+04	1.0E+04	3.4E+06	68.6	6.9E+06	5.1E+06	
	1.4E+04	1.0E+04	3.3E+06	72.7	6.7E+06	4.8E+06	1.4E+07
NP-042	4.2E+04	2.4E+04	7.1E+06	55.5	1.4E+07	1.4E+07	1.4E+07
STD-042	6.4E+04	5.4E+04	5.4E+06	83.5	1.1E+07	1.0E+07	1.1E+07
NP-ST24	3.2E+04	2.1E+04	6.3E+06	65.6	1.3E+07	1.2E+07	1.3E+07
STD-ST24	3.7E+04	3.0E+04	3.0E+06	82.5	6.0E+06	5.7E+06	
	3.6E+04	2.9E+04	2.9E+06	79.3	5.7E+06	5.4E+06	1.2E+07

Library	% unique reads per library	Yield per read / lane (Mb)	Yield per lane (Mb)	Total sequence generated (Mb)	Number of reads mapped / lane	Total number of reads mapped	% mapping / lane
NP-3D7-S		240	480		1.2E+07		88.0
	96.4	264	529	1008	1.3E+07	2.5E+07	89.2
NP-3D7-L		404	809		9.5E+06		88.8
	93.8	339	678	1486	6.5E+06	1.6E+07	73.4
STD-PF88		287	574		1.2E+07		72.8
		298	597		1.2E+07		71.9
		296	592		1.2E+07		71.1
		290	581		1.2E+07		71.3
		284	568		1.1E+07		71.3
		276	552		1.1E+07		71.4
	99.3	265	530	3994	1.1E+07	7.9E+07	71.6
STD-PF3		179	357		6.7E+06		65.9
		165	331		6.1E+06		65.2
		211	422		8.8E+06		73.3
		186	372		7.2E+06		67.8
		198	396		7.7E+06		68.5
		190	379		7.9E+06		72.9
	88.3	187	374	2631	7.8E+06	5.2E+07	72.9
STD-PF2		123	245		3.3E+06		46.9
		117	234		3.2E+06		47.2
		188	377		6.9E+06		63.9
		178	356		5.3E+06		52.5

		182	365		4.7E+06		44.7
		156	312		5.9E+06		66.8
	93.8	155	310	2199	5.7E+06	3.5E+07	64.9
STD-PF85		120	240		4.2E+06		61.2
	73.7	117	233	474	4.0E+06	8.2E+06	60.1
NP-042	98.5	254	508	508	1.3E+07	1.3E+07	94.2
STD-042	97.9	198	397	397	1.0E+07	1.0E+07	96.9
NP-ST24	96.4	226	452	452	1.1E+07	1.1E+07	88.8
STD-ST24		112	223		5.3E+06		87.4
	94.2	106	212	435	4.8E+06	1.0E+07	84.5

Library	Mean % mapping	Mean raw coverage per lane	Total mean raw coverage	Bases not covered by unique mapping	% bases not covered by unique mapping
NP-3D7-S		21			
	89	23	44	447927	1.9
NP-3D7-L		35			
	82	29	65	221275	1.0
STD-PF88		25			
		26			
		26			
		25			
		25			
		24			
	72	23	174	1513500	6.6
STD-PF3		16			
		14			
		18			
		16			
		17			
		16			
	70	16	114	1104037	4.8
STD-PF2		11			
		10			
		16			
		15			
		16			
		14			
	56	13	96	2214439	9.6
STD-85		10			
	61	10	21	4582724	19.9
NP-042	94	96	96	43719	0.8
STD-042	97	75	75	48303	0.9
NP-ST24	89	113	113	227603	5.7
STD-ST24		56			
	86	53	109	270697	6.8

Supplementary Table 1. Detailed library and run information for standard and no-PCR libraries.

No-PCR libraries have the prefix ‘NP’, whereas standard libraries have the prefix ‘STD’. Suffixes ‘L’ and ‘S’ indicate Long and Short different sequencing runs performed on the same library. *B. pertussis* ST24 is not a finished assembly and some contigs are vector contamination. As a consequence, the % of regions with zero coverage is high.

[†.] Number of unique paired end reads for all sequences, before mapping.

Supplementary Methods

DNA preparation and adapter ligation

We fragmented 4.5 µg genomic DNA (quantified by NanoDrop) to approximately 200 bp using Covaris Adaptive Focused Acoustics technology, using the settings: 5 % Duty Cycle; Intensity 10; 200 Cycles per burst over the course of 12 minutes. Following this, we requantified DNA using an Agilent Bioanalyzer 2100 DNA 1000 chip, to confirm the mass of starting material. After end repair and A-tailing following the standard Illumina protocols, we set up ligation reactions in a total volume of 50 µl containing 10 µl template DNA, 8 µM adapters, 1x Illumina DNA ligation buffer, 5 µl Illumina DNA ligase, and incubated reactions for 15 minutes at 20 °C. Ligated samples were then run in a 2 % agarose gel, a size selection was performed, and DNA was extracted [1].

Adapter sequences

We designed adapters were designed so that the 5’ sequence allows hybridisation to the Illumina flowcell, and the 3’ sequence allows hybridisation of the read 1 or read 2 sequencing primer [1].

A_adapter_t (Sigma)

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC

GATC*T, * indicates phosphorothioate

A_adapter_b (Sigma)

GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCT

TCTGCTTG

Both adapters were HPLC purified.

Quantification and sequencing

Libraries were quantified by qPCR [2], alongside three dilutions of a concentration standard library – i.e. a similar library that we had sequenced previously and for which we knew the precise cluster number based on its Bioanalyzer concentration.

Libraries were sequenced on Illumina GAI and GAII Analyzers following the manufacturer's standard cluster generation and sequencing protocols, for 35-76 cycles of sequencing per read [1].

Standard library preparation

Standard sequencing libraries were prepared following the manufacturer's recommended protocol [1], except that genomic DNA was fragmented using Covaris Adaptive Focused Acoustics rather than nebulisation, as described above. After 12 cycles of PCR, we quantified these libraries using an Agilent Bioanalyser 2100 DNA 1000 chip, and by qPCR, as described above, prepared PE flowcells and sequenced for 35 or 36 cycles on an Illumina Genome Analyzer and 36, 37 or 76 cycles on a Genome Analyzer II fitted with PE modules. 76 cycle runs were performed using an alternative deblock reagent, supplied by Illumina.

Illumina read alignment, SNP calling and de novo assembly

For read mapping, we used our modified SSAHA (**S**equence **S**earch and **A**lignment by **H**ashing **A**lgorithm) program http://www.sanger.ac.uk/Software/analysis/SSAHA2/ [3], which has been optimized for short-reads. The alignment files were processed further for SNP detection using a variation detection pipeline ssaha_pileup (ftp://ftp.sanger.ac.uk/pub/zn1/ssaha_pileup/). It is expected that the high sensitivity offered by the alignment tool should improve read mapping, particularly, for those extremely AT biased genomes such as *Plasmodium falciparum*.

The availability of paired end Illumina data provides challenging, but exciting prospects for de novo assemblies, where requirement for read coverage across the genome is more than that for variation detections. Within a generated contig, every base has to be covered several times from raw reads in order to make a consensus. We performed assembly using the Velvet (version 0.7.26) short read assembler [4] to produce assemblies from all datasets. Input parameters such as "-ins_length" and "exp_cov" were adjusted according to the mapping values while setting "-min_pair_count 20", "-min_contig_length 100" and "-cov_cutoff 10". Other parameters were set by default.

References

1. Bentley D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
2. Quail M.A. *et al.* A large genome centre's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005-1010 (2008).

3. Ning Z., Cox A.J. & Mullikin J.C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729 (2001).

4. Zerbino D.R. & Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).