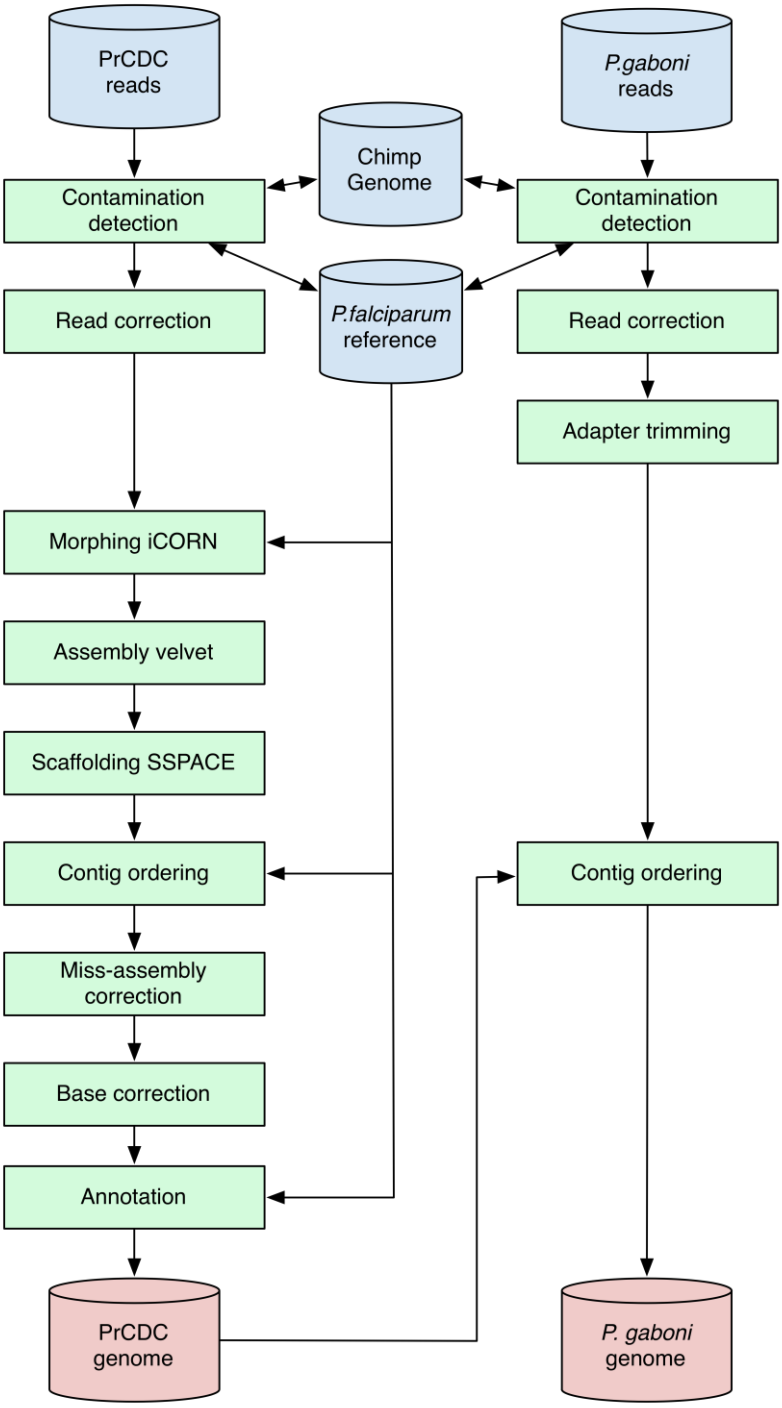
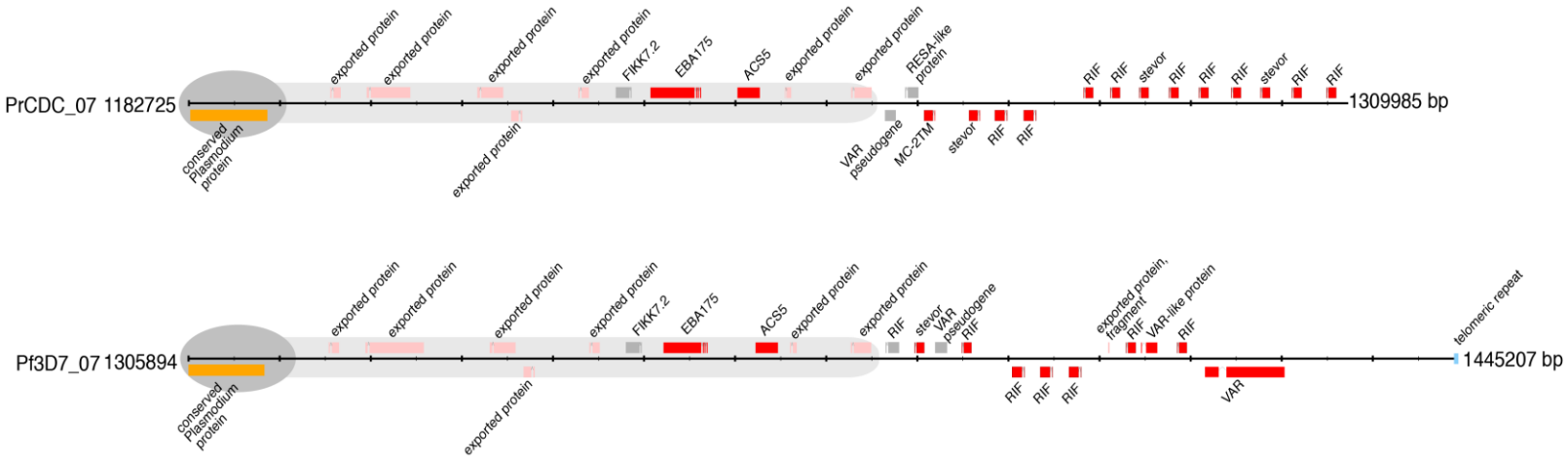
# SUPPLEMENTARY INFORMATION

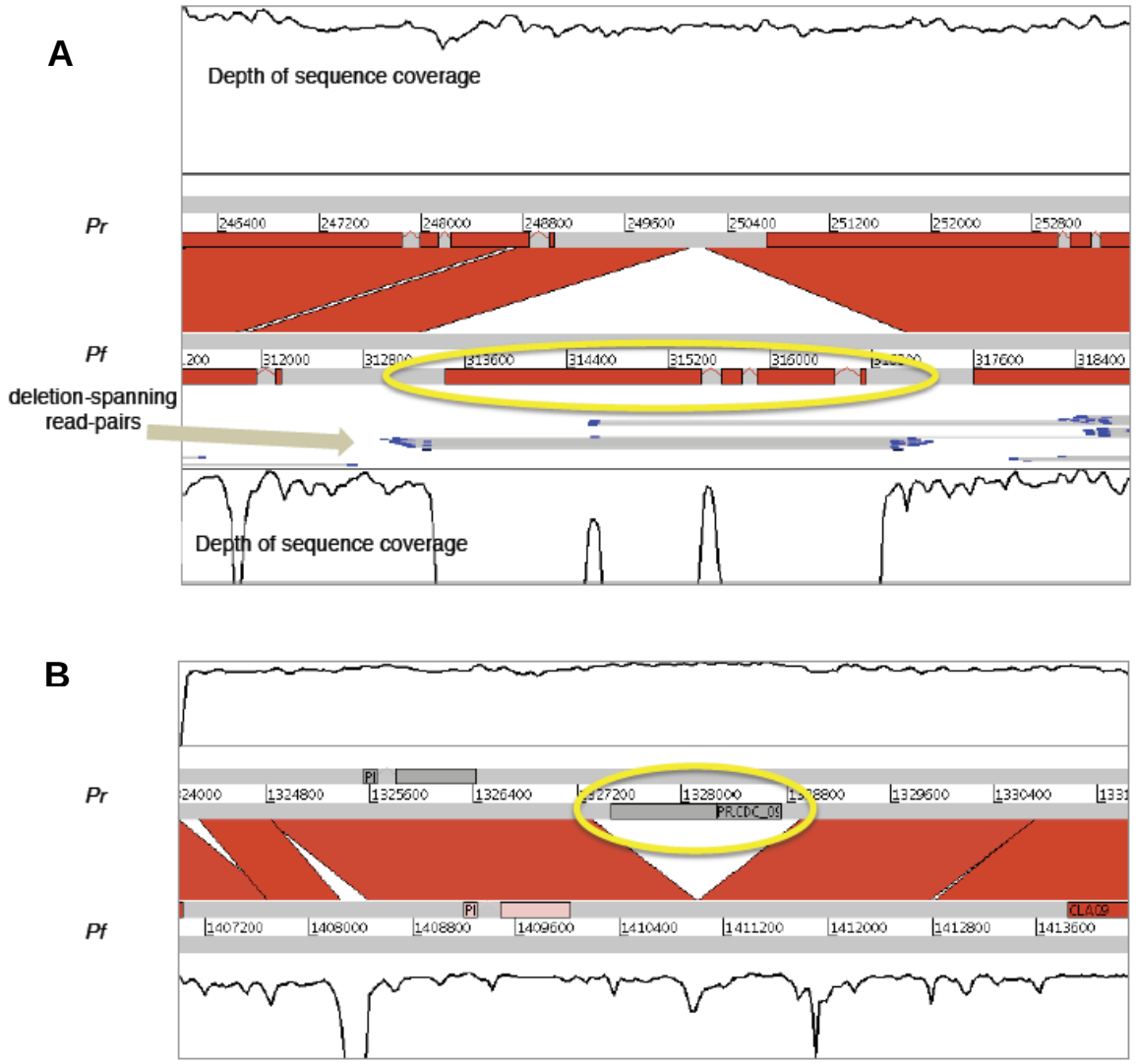
## Supplementary Figures



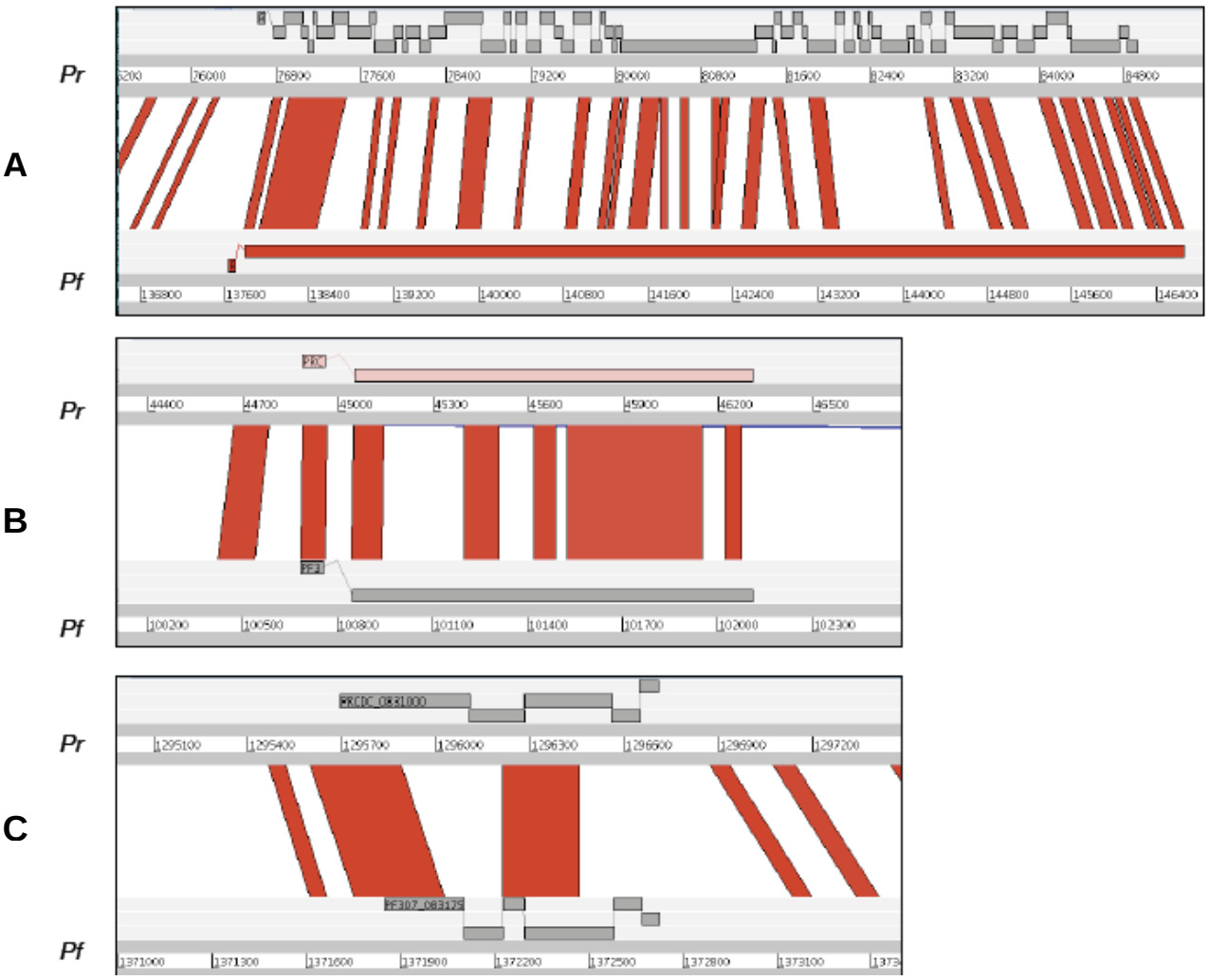

**Supplementary Figure 1** Assembly of the P. reichenowi and P. gaboni genome sequences. process. The process by which the genomes were assembled, are detailed in the above flow chart.



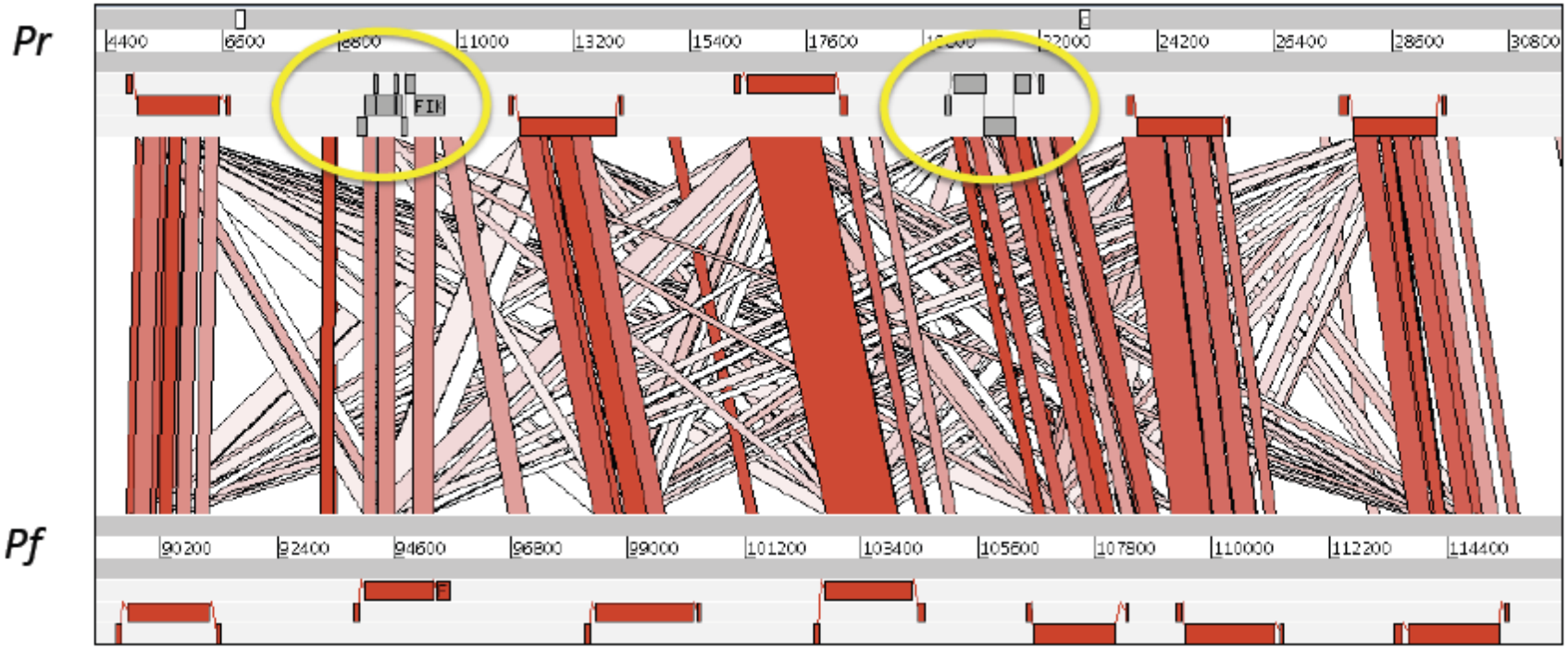

**Supplementary Figure 2** Subtelomere organization of *P. reichenowi* CDC and *P. falciparum* 3D7. The order and orientation of the genes on the right hand subtelomeres of chromosome 7 of *P. reichenowi* (PrCDC_07) and *P. falciparum* (Pf3D7_07) are shown. Exons are shown as coloured boxes with introns as linking lines. The shaded area in grey marks the synteny between *P. falciparum* 3D7 and *P. reichenowi* CDC. The dark shaded area marks the synteny between other *Plasmodium* species, e.g. *P. knowlesi* H and *P. vivax* SalI.



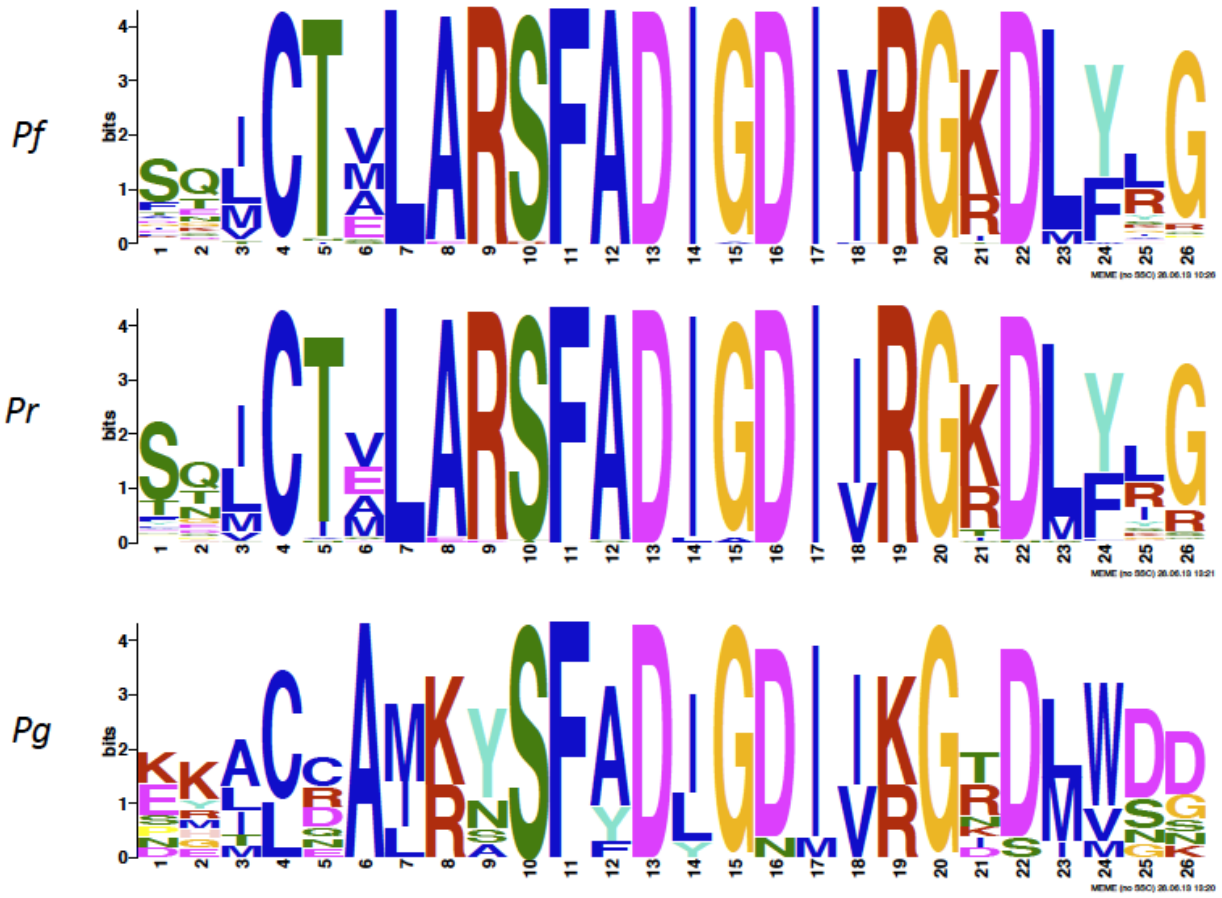

**Supplementary Figure 3** Deleted genes in the *P. reichenowi* and *P. falciparum* genomes. Regions of the *P. reichenowi* genome (Pr; top) and the *P. falciparum* 3D7 genome (Pf; bottom) are compared using Artemis Comparison Tool [1]. The vertical red lines between the genomes are BLASTN hits with bit scores ≥ 100. The coloured boxes in each genome represent genes or pseudogenes (grey). Graphs above and below each comparison show the depth of coverage of mapped, paired-end sequencing reads (only including "proper" pairs within the expected mapping distance and in the correct orientation) on a log scale. An orthologue of *P. falciparum* gene PF3D7_020780 appears to be deleted in *P. reichenowi* (A). No *P. reichenowi* reads map against the gene in *P. falciparum*, and pair-end reads from *P. reichenowi* span the deletion. The deleted/inserted region is 4kb long. In (B) PRCDC_0933800 is shown with the orthologous region deleted from *P. falciparum*.



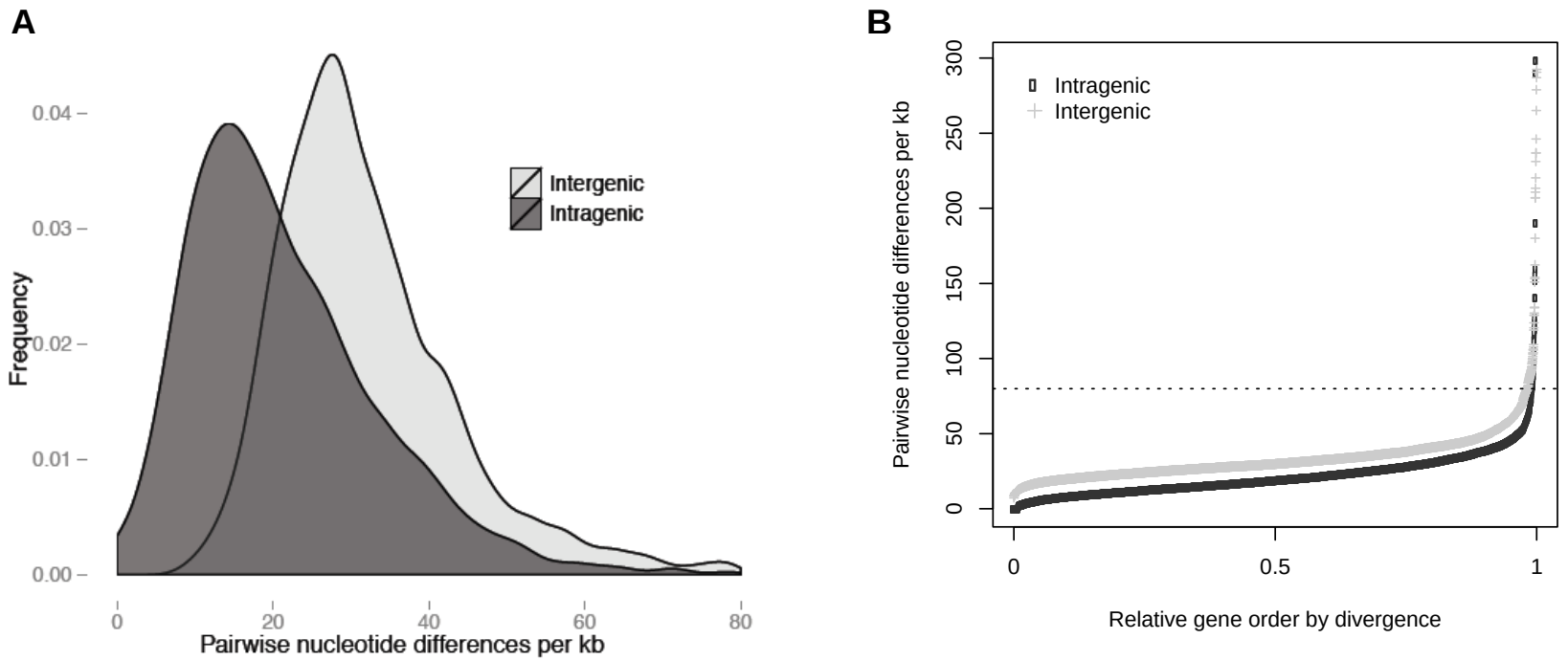

**Supplementary Figure 4** Examples of pseudogenes in *P. reichenowi* and *P. falciparum*. Loci containing pseudogenes are compared between *P. reichenowi* (top) and *P. falciparum* (bottom) using Artemis Comparison Tool [1]. Grey bars represent the forward and reverse strands of DNA. The red lines between sequences represent tBLASTx hits. (A) In *P. reichenowi*, RH1 is a pseudogene (PRCDC_0005400) but the orthologous locus in *P. falciparum* (PF3D7_0402300) contains an intact RH1 (reticulocyte binding protein homologue 1). (B) An intact gene (PRCDC_0005400) in *P. reichenowi* encoding a putative exported protein but the orthologous locus in Pf3D7 contains a pseudogene (PF3D7_0402300). At the PHISTa locus (C), a pseudogene is present in both *P. reichenowi* (PRCDC_0831000) and *P. falciparum* (PF3D7_0831750).



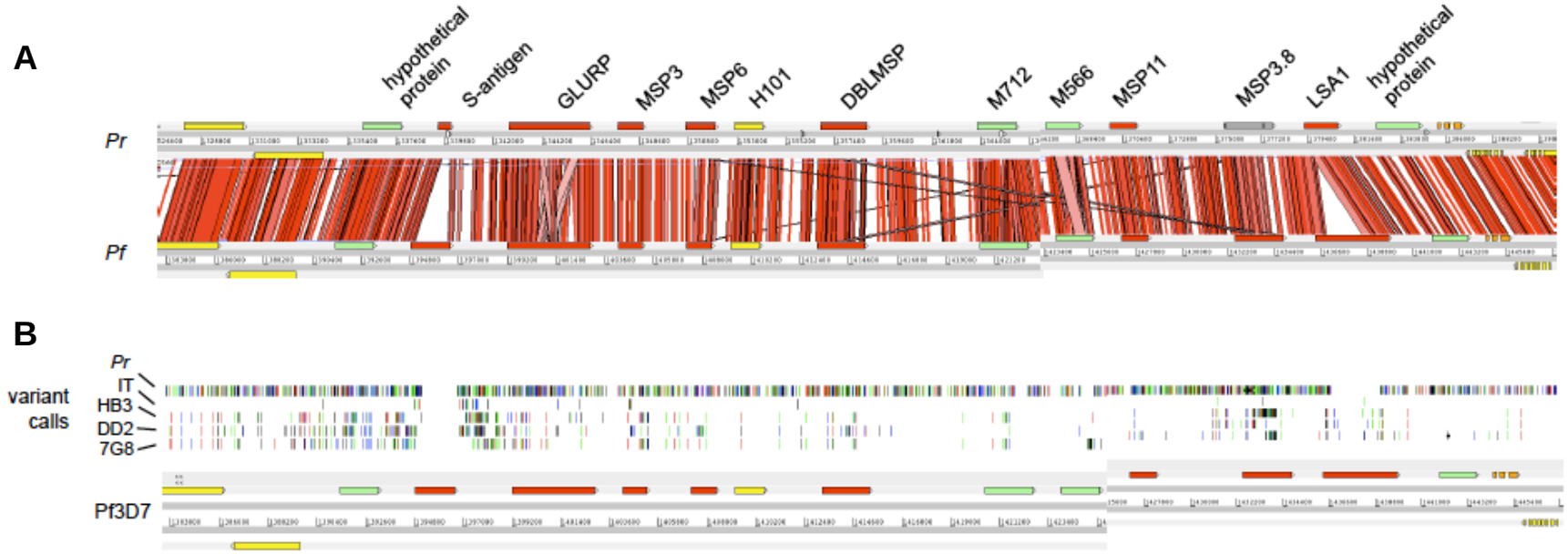

**Supplementary Figure 5** Presence of pseudogenes within a tandem array of *fikk* genes on chromosome 9. The PrCDC locus (top) is compared with Pf3D7 (bottom) using Artemis Comparison Tool [1]. Pseudogenes (PF3D7_0902100, PF3D7_0902400) are shown in grey. tBLASTx hits with a bit score ≥ 200 are shown as parallelograms connecting the two sequences and are coloured by identity: 40% (white) to 100% (red).



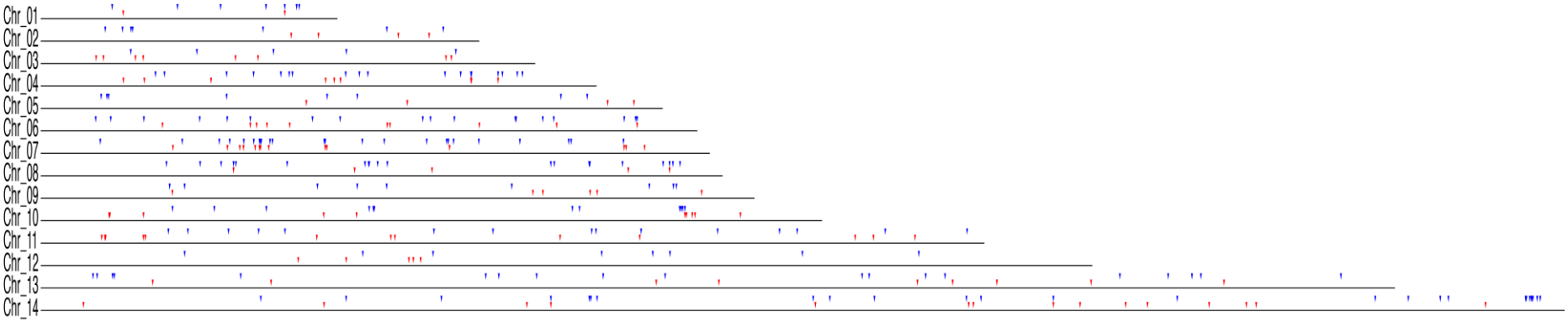

**Supplementary Figure 6** Conservation at the most highly conserved motif in the *var* DBLα domain. The LARSFADIIG amino acid motif is conserved across *P. falciparum* and *P. reichenowi,* but has diverged in *P. gaboni*. To represent the motifs, the coding sequences of *var* genes from Pf, Pr and Pg (longer than 500aa) were aligned and regions aligning with the known LARSFADIIG motif were extracted and run through MEME [2].



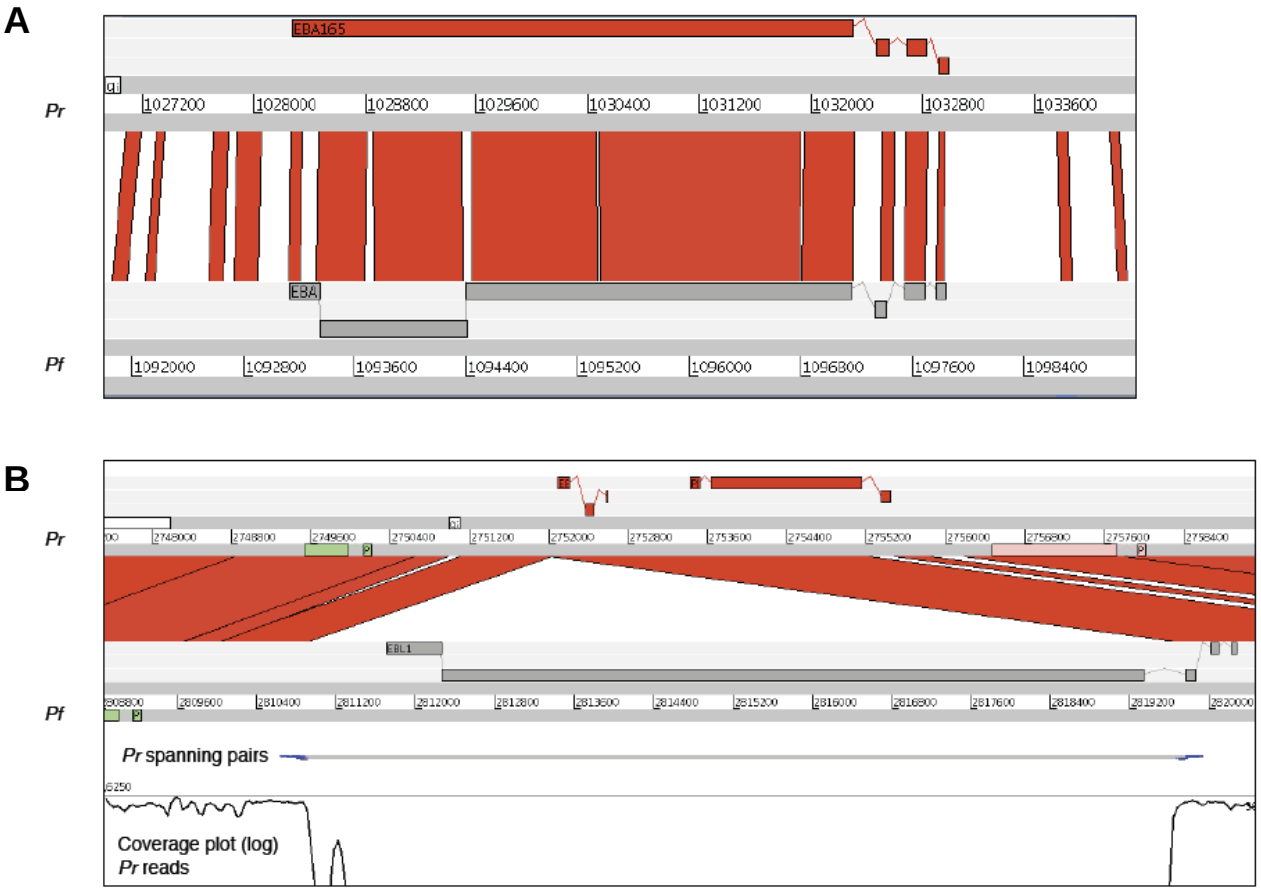

**Supplementary Figure 7** Distribution of pairwise difference in inter- and intra-genic regions. Values are taken from Tables S3 and S4 and plotted using R and the ggplot2 package. In (A) the frequency of pairwise differences (below 80 $kb^{-1}$) is is less skewed for intragenic divergence. In (B), genes are ranked by divergence. Values are shown for approx. 5000 genes and 4500 intergenic regions with the ranges normalised. The threshold of 80 $kb^{-1}$ used in (A) is marked as a dotted line on plot (B). (N.B. Divergence K is the pairwise nucleotide version *per base*)




**Supplementary Figure 8** Inter- and intra-species differences at the MSP3 gene cluster on chromosome 10**.** *P. reichenowi* CDC (top) and *P. falciparum* 3D7 (bottom) are compared using ACT [1]. (A) Gene structures at this locus are largely conserved between *P. falciparum* and *P. reichenowi with the exception that* MSP3.8 (DBLMSP2) is a pseudogene (grey) in *P. reichenowi* CDC. (B) SNPs called in the same region using sequencing reads from *P. reichenowi* (top track), and four lab clones in the following order: IT, HB3, DD2 and 7G8. Reads were mapped using SMALT (quality 60 and 20 read depth). SNPs were called using mpileup and bcftools from the samtools package. The colours represent individual base substitutions: A (green), G (blue), T (black) and C (red).




**Supplementary Figure 9** Genome-wide distribution of regions with low divergence. Intergenic (blue) and intragenic (red) gegions are displayed, if HKAr > 0.15 or K <0.01.




**Supplementary Figure 10** Inter-species differences in the EBL family of erythrocyte binding proteins. Orthologous loci are compared using ACT [1]. (A) EBA165 is a pseudogene PF3D7_0424300 in P. falciparum (bottom) with two premature codons, but is an intact coding sequence PRCDC_0421500in P. *reichenowi*. (B) The genome assembly suggests a large deletion of EBL1 in *P. reichenowi* (upper track). This is confirmed by a complete absence of coverage (coverage plot), as well as the existence of read pairs (blue) that span the deletion. The total deletion is around 8 kb.

## **Supplementary Note 1** Calling of sequence variants in *P. reichenowi* and *P. gaboni*

### Comprehensive list of variant calls in *P. reichenowi*

To define a reference list of SNP's and indels that differ between the *P. reichenowi* CDC and *P. falciparum* 3D7 genomes the MUMmer package [3] was used. *P. reichenowi* and *P. falciparum* chromosomes were aligned using nucmer (parameter: -g 500 -c 500 -l 10). The alignments were filtered with the delta-filter (parameter: -g -i 75) and variants called using find-snp, with and without the inclusion of SNP calls within repetitive sequence (parameter -C). A merged output file can be found at:

ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/reichenowi/Alleles/Pf3D7_Preich.Alleles.April2014.txt.gz

### Identification of single nucleotide variants in *P. gaboni*

Contigs of *P. gaboni* with read depth ≥ 10 were aligned against *P. falciparum* 3D7 version 3 and putative fixed differences called using *show-snp* from the MUMmer package, as described above. 6,329,559 bases of the *P. falciparum* genome were covered by *P. reichenowi* and *P. gaboni* contigs and ~2,700 *P. falciparum* genes were covered across ≥ 25% of their lengths with *P. gaboni* sequence (after low complexity trimming, see *Methods*). To clean possible alignment artefacts, additional alignments were generated using Gblocks [4] (parameter -t=c -p=n -b4=4). Ka/Ks ratios were calculated (Supplementary Data 3) as described in *Methods*. For comparative purposes Ka/Ks ratios were recalculated for *P. reichenowi* vs *P. falciparum* using this automated Blocks-based method.

# Supplementary Methods

## Contamination & Correction of reads

Reads were screened against the *P. falciparum* reference genome using SSAHA [5]. Those reads that did not map to *P. falciparum* were aligned to the Chimpanzee genome using BWA [6], to identify and exclude contaminating host-derived sequences.

Adapter sequences were mapped against the *P. gaboni* reads and those bases that matched the adapter sequences were deleted. Reads smaller than 60 bp were also discarded.

To correct sequencing errors in the reads we used SGA [7] version 0.9.9. First low quality regions of the reads were trimmed with the preprocess command of SGA using the parameters: --permute-ambiguous (bases randomly assigned to all ambiguous, N, bases within reads); -f 15 (discard reads with > 15 low quality bases); -q 5 (quality cut-off to trim reads). After indexing the reads, the correction program was used to correct the reads of *P. gaboni* with the parameters: -k 21; -x 3; -i 4 (if a k-mer of 21 from a read occurs fewer 3 times, attempt correction four times). For *P. reichenowi*, the same parameters were used with the exception that the correction was run twice.

After processing, 46 million (22.8%) of the initial *P. reichenowi* reads were used for the assembly. For *P. gaboni*, 23 million (6.3%) were used. Within the *P. gaboni* reads, we found 1% contamination with *P. falciparum*.

## Assembly of *P. reichenowi*

Supplementary Figure 1 shows an overview of the assembly process. A working reference of *P. reichenowi* was produced by iteratively mapping the Illumina reads against the *P. falciparum* 3D7 genome (Pf3D7) and changing the sequence of the latter to match apparent fixed differences, using iCORN [8]. After 12 iterations, a total of 850,000 single-base substitutions, or insertions and deletions of up to 3 bases, were made and 88% of the genome was covered by mapped reads at a depth of at least 20-fold coverage.

Using reads mapped by SMALT (http://www.sanger.ac.uk/resources/software/smalt/) to the *P. reichenowi* working reference, a reference-guided assembly was produced using Velvet Columbus [9] (version 1.1.04) and the parameters: -kmer 71 -ins_length 300 -cov_cutoff 3 -min_contig_lgth 200 -min_pair_count 12. The 3 kb reads were excluded from the assembly step because ~90% of reads were composed of PCR replicates. At this stage, the total assembly length was 24.2 Mb, the largest scaffold was 761.7 kb and the N50 was 82.4 kb.

The assembly was further scaffolded with SSPACE [10], using first the PCR-free library and then the 3 kb library. We found that running SSPACE iteratively improved the performance of the scaffolder. For the PCR-free library, the number of mate-pairs required to form a scaffold was decreased with each iteration as follows: 20, 10, 10, 7, 5 and 5. For the 3 kb library, the number of mate-pairs was similarly decreased: 10, 10, 5, 5, 5, 3 and 3. The PCR-free library reduced the number of scaffolds from 2377 to 1416 and the 3 kb library joined a further 396 scaffolds, increasing the size of the largest scaffold to 2.0 Mb and the N50 to 541.2 kb.

Next the scaffolds were ordered against the existing Pf3D7 using ABACAS [11]. To avoid misplacing scaffolds, we masked the subtelomeric regions and placed a scaffold if its identity and degree of overlap with the Pf3D7 genome were at least 92% and 500 bp, respectively. Scaffolds smaller than 500 bp were discarded, resulting in 145 scaffolds. The sequencing gaps were then closed with IMAGE [12]. In 18 iterations of the 3663 gaps, 2454 (67%) could be closed, using the k-mers 61, 51 and 41.

To find miss-assemblies, we mapped all the reads against the scaffolds (SMALT) and broke them if less than 3 mate-pairs covered a position. This increased the number of scaffolds by 145, to 592. We repeated the ABACAS and correction step, resulting in 439 scaffolds.

Sequences representing host contamination were also removed after assembly. All scaffolds were BLAST-searched against the chimpanzee genome and 65 supercontigs that had a hit with 95% identity and an overlap of >150 bp were excluded, reducing the number of scaffolds to 374.

As a final step, iCORN was used to correct 3142 single-base discrepancies and 1204 insertions or deletions in the assembly.

The final version of the *P. reichenowi* (PrCDC version 3), has 374 scaffolds, consisting in 14 chromosomes, two plastids and a bin containing 358 scaffolds. The assembly length is 24,064,760 bp (including 139,524 N's distributed across 1,574 sequencing gaps).

### Assembly of *P. gaboni*

The assembly of *P. gaboni* resulted in many fragmented small contigs, due to the low amount of template DNA and high level of host and adapter contamination. Although we tested the same methodology as used with the PrCDC genome (i.e.. transforming the genome), better results were obtained by assembling the corrected reads *de novo*.

Velvet was run with following parameters (exp_cov 999; cov_cutoff 3; min_contig_lgth 200). We obtained an assembly of 20.2 Mb in 41,285 scaffolds and an N50 of 574 bp (largest scaffold 9,913 bp). When ordered against the PrCDC assembly 17,876 contigs (11.2 Mb) could be placed against the reference.

During contamination screening, contigs with hits to *Malassezia globosa* were excluded.

The *P. gaboni* assembly is available from:
ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/gaboni/Assembly/Version1/

## Annotation

The annotation of *P. reichenowi* was performed in three steps: transfer of annotation from regions where synteny is conserved with *P. falciparum*, genome-wide *ab initio* gene prediction and manual curation. Using RATT [13] (Parameter: Species), annotation was transferred from the October 2011 update of Pf3D7 version 3.0 onto the PrCDC genome. From 5418 genes in the reference, 5111 were transferred correctly, and the annotation of gene models was kept current until June 2012 as part of ongoing curation efforts.

In regions where synteny was not conserved, genes were predicted using Augustus [14], version 2.5.5. Augustus was trained with the annotation of Pf3D7. Each predicted protein was BLAST-searched against the proteome of *P. falciparum*. Using an E-value threshold of $10^{-6}$, the first hit was used to provide functional annotation. The annotation was manually inspected in Artemis [15].

For *P. gaboni*, due to its highly fragmented state a full genome annotation was not undertaken. Instead we analysed how many *P. falciparum* genes hit the *P. gaboni* assembly. After discarding contigs smaller than 300bp, BLASTX (E-value ≤$10^{-6}$) was run and hits of ≥ 80% identity and ≥ 100 amino acid overlap were included. 6576 *P. gaboni* contigs hit 3297 *P. falciparum* genes but after further filtering, to remove contigs with a median coverage of <10 reads (conservative exclusion of possible contamination), 2061 *P. falciparum* genes had a qualifying BLAST hit. Of those, just 1263 *P. falciparum* genes had matches across ≥ 50% of their lengths to corresponding *P. gaboni* sequences. The *P. gaboni* was therefore only used for specific, focused comparisons.

To annotate the laboratory clones, we used RATT [13] (using the “Assembly” parameter) to transfer the October 2011 annotation of Pf3D7. The gene models of *P. falciparum* IT were systematically corrected manually using Artemis and the Artemis Comparison Tool [15].

For functional annotation (gene product descriptions & gene names), the latest version of the Pf3D7 (July 2013) was used.

*Reference datasets:*

ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.latest_version/version3/October_2011/

ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.latest_version/version3/June_2012/

ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.latest_version/version3/July 2013/ (functional annotation only)

## Supplementary References


1. Carver T*, et al.* Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. *Bioinformatics* 24, 2672-2676 (2008).

2. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34, W369-373 (2006).

3. Kurtz S*, et al.* Versatile and open software for comparing large genomes. *Genome Biol* 5, R12 (2004).

4. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17, 540-552 (2000).

5. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res* 11, 1725-1729 (2001).

6. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589-595 (2010).

7. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22, 549-556 (2012).

8. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* 26, 1704-1707 (2010).

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

10. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578-579 (2011).

11. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 25, 1968-1969 (2009).

12. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* 11, R41 (2010).

13. Otto TD, Dillon GP, Degrave WS, Berriman M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* 39, e57 (2011).

14. Stanke M, Schoffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7, 62 (2006).

15. Carver T*, et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 24, 2672-2676 (2008).