DATA NOTE

Progression of the canonical reference malaria parasite genome from 2002–2019 [version 1; peer review: awaiting peer review]

Ulrike Böhme, Thomas D. Otto, Mandy Sanders, Chris I. Newbold, Matthew Berriman

1Parasite Genomics, Wellcome Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
2Institute of Infection, Immunity and Inflammation, MVLS, University of Glasgow, Glasgow, G12 8QQ, UK
3Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK

Abstract

Here we describe the ways in which the sequence and annotation of the Plasmodium falciparum reference genome has changed since its publication in 2002. As the malaria species responsible for the most deaths worldwide, the richness of annotation and accuracy of the sequence are important resources for the P. falciparum research community as well as the basis for interpreting the genomes of subsequently sequenced species. At the time of publication in 2002 over 60% of predicted genes had unknown functions. As of March 2019, this number has been significantly decreased to 33%. The reduction is due to the inclusion of genes that were subsequently characterised experimentally and genes with significant similarity to others with known functions. In addition, the structural annotation of genes has been significantly refined; 27% of gene structures have been changed since 2002, comprising changes in exon-intron boundaries, addition or deletion of exons and the addition or deletion of genes. The sequence has also undergone significant improvements. In addition to the correction of a large number of singe-base and insertion or deletion errors, a major miss-assembly between the subtelomeres of chromosome 7 and 8 has been corrected. As the number of sequenced isolates continues to grow rapidly, a single reference genome will not be an adequate basis for interpreting intra-species sequence diversity. We therefore describe in this publication a population reference genome of P. falciparum, called Pfref1. This reference will enable the community to map to regions that are not present in the current assembly. P. falciparum 3D7 will be continued to be maintained with ongoing curation ensuring continual improvements in annotation quality.

Keywords

Plasmodium, falciparum, genome, reference, annotation, curation
Introduction

The genome of *Plasmodium falciparum* 3D7, the species responsible for the most severe form of malaria, was the first reference genome published to support *Plasmodium* research. Its publication more than almost two decades ago (Gardner et al., 2002) was a milestone, the impact of which is reflected in several thousand citations that mention the *P. falciparum* 3D7 genome. The sequencing of *P. falciparum* was initially accompanied by the draft genome of a rodent malaria species, *P. yoelii* (Carlton et al., 2002). These genomes were followed by those of several other *Plasmodium* spp, sequenced using Sanger sequencing technology, including human-infective species *P. vivax* (Carlton et al., 2008), the monkey and human malaria parasite *P. knowlesi* (Pain et al., 2008) and further rodent *Plasmodium* spp (Hall et al., 2005). With the advent of much cheaper short-read technology, many more genomes have been sequenced, including the chimpanzee parasite *P. reichenowi* (Otto et al., 2014), the monkey malaria parasites *P. cynomolgi* (Tachibana et al., 2012), *P. coatneyi* (Chien et al., 2016), *P. inui* and *P. fragile*, the murine parasite *P. vincki*, the human parasites *P. malariae, P. ovale* (Rutledge et al., 2017) and the avian malaria parasites *P. gallinaceum* and *P. relictum* (Böhme et al., 2018). Although, many of these genomes are highly fragmented draft assemblies, algorithms that use high coverage of aligned short reads have enabled a variety of cost-effective genome-assembly improvements for several species (Swain et al., 2012).

*P. falciparum* 3D7 is a major focus of malaria research and the accuracy of its reference genome and annotation are vital for accelerating hypothesis-driven research. Moreover, the availability of a reference genome has additional importance: it underpins genome comparisons, across the suite of genome sequences that are now available for multiple *Plasmodium* species, and the global efforts to analyse genome variation amongst thousands of clinical and lab isolates. The need for a commitment to maintain and improve this genome has long been recognized by the Wellcome Sanger Institute. Through careful manual curation, highly accurate predictions of coding and non-coding genes have been added. Functional descriptions of genes have also been kept up to date, to reflect the growing volume of *P. falciparum* related scientific literature. In many ways the depth of annotation is similar to that more commonly associated with model organisms. For instance, Gene Ontology terms have been manually selected that capture from the scientific literature the richness of gene roles in a format that can be easily queried or used for inference in genome-wide analyses. Recent examples include the genome-wide analysis of transcriptional dynamics (Painter et al., 2018) and the uncovering of common functions in essential genes (Zhang et al., 2018).

Genome improvement and curation has resulted in thousands of individual changes over more than 15 years. In particular, the resolution of subtelomeric regions has been transformed along with the ability to annotate important multigene families that are often found in those regions.

This is the first paper to describe the changes since the *P. falciparum* 3D7 genome was first published. Originally 5,268 protein-coding genes were annotated and of those, over 60% (3,465 genes) of predicted genes had unknown functions (Table 1). Despite the fact that there still seems to be a common perception that over 50% of genes remain functionally unannotated (Briquet et al., 2018; Tang et al., 2019) the number of predicted genes has risen to 5,438 and the proportion without ascribed functions has now shrunk to 33% (1,776 genes) (Table 1). Since 2002, 27% of genes have undergone structural

| Table 1 | Table showing the main differences between *P. falciparum* 3D7 version 1 (2002), version 2 (2005), version 2.1.4 (2007) and version 3.2 (March 2019). *Numbers includes partial genes, pseudogenes are not included. ** Nucleotide ambiguities present in genome v1 and v2 have been changed to gaps. |
|---|---|---|---|---|
| **Nuclear genome** | 2002 (version 1) | 2005 (version 2) | 2007 (version 2.1.4) | 2019 (version 3.2) |
| Genome size (bp) | 22,853,764 | 23,289,065 | 23,264,337 | 23,292,622 |
| Gaps | 93 | 10 | 160** | 0 |
| Genes* | 5268 | 5414 | 5387 | 5280 |
| Pseudogenes | NA | 70 | 73 | 158 |
| ncRNAs | 0 | 0 | 606 | 103 |
| Hypothetical proteins | 3465 | 3545 | 3099 | 1776 |
| tRNAs | 43 | 27 | 44 | 45 |
| **Apicoplast genome** | | | | |
| Assembly size (kb) | Absent | Absent | Absent | 34250 |
| genes | Absent | Absent | Absent | 30 |
| **Mitochondrial genome** | | | | |
| Assembly size | Absent | Absent | Absent | 5967 |
| genes | Absent | Absent | Absent | 3 |
changes or have been added based on RNAseq data and other data from publications. New ncRNAs have also been added and complete apicoplast and mitochondrial genomes have been assembled. One of the many purposes of a reference genome is to interpret natural variation data. In the latest version, we have therefore included alternative contigs representing major haplotype differences. This reference dataset has been named PfRef1 to reflect the fact that it does not simply comprise *P. falciparum* 3D7 data but has been supplemented with other reference data to better represent the pan-genome for this species. The aim of PfRef1 is to enable robust mapping to analyse genome variation in regions of *Plasmodium* genomes where the current Pf3D7 genome (v3.2) is an unsuitable reference.

**Methods**

**Curation and annotation**

Changes to the genome annotation reflect ongoing work at the Wellcome Sanger Institute. The software *Artemis* (version 10 to version 18) was adapted to use a CHADO database schema (Carver et al., 2008) and has been used continuously for manual curation and annotation. This database system is directly connected to GeneDB. Every 4 to 6 months data is transferred to PlasmoDB. To update functional annotation, Pubmed was searched (search terms *Plasmodium* and apicomplexa) on a regular basis for publications related to *Plasmodium*. Relevant information, i.e. gene product descriptions, EC numbers, gene names and functional descriptions to be captured by Gene Ontology terms, was extracted and changes manually added in Artemis. RNA-Seq data and TBLASTX comparisons were the primary supporting evidence for manual improvements to gene models. Information from user comments that were submitted to gene record pages in PlasmoDB were assessed and where relevant used to update annotation. Evidence codes that support product descriptions are available as GFF format genome annotation files from the following FTP site: ftp://ftp.sanger.ac.uk/pub/genedb/releases/latest/Pfalciparum/. To find annotation differences between genome versions we’ve performed pair-wise TBLASTX comparisons using the Artemis Comparison Tool (ACT) (version 13) (Carver et al., 2005).

**DNA sequencing and correction of the P. falciparum 3D7 genome**

After publication of the genome in 2002, manual finishing continued. Sequence changes and gap closure between genome versions 1 and 2 have been described (Berry et al., 2004). Using a combination of capillary sequencing, Illumina sequencing, Pacific Biosciences sequencing and automated sequence correction, we have continued to improve the reference genome. Sequence errors on all chromosomes were corrected using the iCORN algorithm (version 1) (Otto et al., 2010a). Two ambiguous regions towards the telomeres of chromosomes 7 and chromosome 8 were resolved using a PCR tiling path. The tiling path spanned the region between genes PF3D7_0805400 (MAL8P1.200) to PF3D7_0831200 (MAL8P1.204) on chromosome 8 and from PF3D7_0833500 (MAL7P1.212) to PF3D7_0701900 (PF07_0004) on chromosome 7. From the results of these the assembly of the left hand side of chromosome 8 was confirmed. There was a gap in chromosome 7 over a highly repetitive region of ~20kb. Read pair information from a 3kb insert Illumina library was used to identify unassembled contigs that could close the gap and PCR was used to confirm the assembly. Unassembled contigs were also searched for any unique coding sequence. One contig was found and this was linked to a region on the right hand side of chromosome 8 by Illumina read pairs. This was also confirmed by PCR. For genome version 3.1, the complete apicoplast was included. Resolving this sequence has been described (Hunt et al., 2015). For genome version 3.2, the *P. falciparum* 3D7 mitochondrial was also included. This mitochondrion was part of a recently described PacBio assembly of *P. falciparum* 3D7 (Otto et al., 2018).

The mitochondrion was corrected using the iCORN algorithm (Otto et al., 2010a) followed by circulator (Hunt et al., 2015).

**PfRef1 reference genome**

The *P. falciparum* 3D7 version 3.2 assembly was compared with PacBio assemblies that we have recently described for several other isolates (Otto et al., 2018) to create a population reference that we have termed PfRef1. BLASTN comparisons of 3 lab isolates (*P. falciparum* IT, *P. falciparum* DD2, *P. falciparum* HB3) were inspected together using ACT software (Carver et al., 2005). Excluding single nucleotide polymorphisms and small insertions and deletions, major sequence differences were identified, manually extracted and are provided as small EMBL files. Each file contains a sequence variant with a 500 bp anchor sequence on either side to enable unambiguous placement along the 3D7 reference. The files are available on the following FTP site: ftp://ftp.sanger.ac.uk/pub/project/eukaryotes/Plasmodium/falciparum/PfRef1

**Results**

**Improvement of the *P. falciparum* 3D7 assembly**

An update to the published version of the *P. falciparum* 3D7 genome was released in 2005 (version 2), this included gap closure and completion of many chromosomes from telomere to telomere (Table 1) (Berry et al., 2004). In September 2011, version 3 of the genome was released that included the correction of two major miss-assemblies on chromosome 7 and 8, the replacement of all “N”s in the genome with corrected sequence and the correction of hundreds of sequencing errors. These improvements were brought about largely due to the availability of Illumina technology and the development of a genome correction algorithm (Otto et al., 2010a). In addition, gene locus identifiers were changed to bring them into line with guidelines published by the European Nucleotide Archive (ENA): systematic identifiers were changed to start with the prefix PF3D7 followed by an underscore and the chromosome number. All previous identifiers were retained as searchable synonyms.

Version 3.1 included a complete apicoplast genome with a length of 34 kb. In the original genome project that was published in 2002 the *P. falciparum* 3D7 apicoplast was not sequenced. For the re-annotation process the apicoplast from *P. falciparum* isolate C10 was temporarily used (Genbank X95275.2, X95276.2) (Wilson et al., 1996) and included in genome version 2.1.4. The published apicoplast from isolate C10 consisted of two
large contigs that could not be assembled due to a large deletion caused by an almost identical inverted repeat. As previously reported (Hunt et al., 2015), version 3.1 of the genome includes a complete apicoplast genome that includes the resolved 5kb repeat.

For the reannotation process the mitochondrial genome with Genbank ID M76611 was included. This Genbank entry only reports the source as *P. falciparum*. Our recently described *P. falciparum* 3D7 PacBio assembly (Otto et al., 2018) contained a mitochondrial genome. Comparing this PacBio genome with M76611 revealed a difference of two bases: a SNP at base 772 (T to C) and at base 4952 (C to T). The SNP at base 772 is non-synonymous and results in an Isolucine to Valine substitution in COX3 (PF3D7_MIT01400) at amino acid 250. The genome version with the correct mitochondrial genome has been named version 3.2.

**Manual annotation of *P. falciparum* 3D7**

At the start of the *P. falciparum* 3D7 genome project manual gene finding was heavily reliant on sequence composition, especially GC content differences, due to an absence of alignable evidence. Over time additional *Plasmodium* genomes became available, in 2005 the rodent malaria genomes (Hall et al., 2005) and in 2008 *P. knowlesi* (Pain et al., 2008) and *P. vivax* (Carlton et al., 2008). Having these additional *Plasmodium* genomes enabled gene structures to be revisited. A comprehensive reannotation process began with a workshop in 2007, involving approximately 40 members of the *Plasmodium* research community (see Box 1). One of the goals of the workshop was to ascribe updated functions to predicted proteins, check gene structures and systematically revisit the nomenclature for large gene families. A major new addition to the evidence was genome wide TBLASTX comparisons between species that were used to highlight conserved regions at the protein level and therefore identify positionally conserved orthologues and refine their exon-intron boundaries. In 2010, we published the first RNA-Seq data for this species (Otto et al., 2010b). These data were used to further evaluate gene models and improve the accuracy of gene structures. As a result, 27% of genes have been added or had their structure changed since 2002 (Figure 1). 1255 genes had changes to exon-intron boundaries or exons added or removed (Figure 2A); this number include genes that were merged (Figure 2B) or split (Figure 2C). Since 2002, 244 genes have been added (Figure 2D, Extended data: Table 1 (Böhme, 2019)) and 36 predicted genes have been deleted due to a lack of evidence supporting their earlier prediction in regions of repetitive or unusual sequence, or because later RNAseq evidence (including strand-specific information) suggested that they were ncRNAs rather than protein-coding (Figure 2E, Extended data: Table 2 (Böhme, 2019)). In addition, a number of genes were created after 2002 based on algorithmic predictions but subsequently deleted due to a lack of supporting evidence (Extended data: Table 3 (Böhme, 2019)). Figure 3 shows the number of changes to predicted gene structures, as well as the addition and deletion of new genes at four different time points.

New non-coding RNAs have been annotated based on transcriptome data, examples published in the scientific literature (Chakrabarti et al., 2007; Guizetti et al., 2016; Raabe et al., 2010) and new predictive models in Rfam (Kalvari et al., 2018). In 2007, 603 automatically predicted ncRNAs were included in the annotation (Table 1) (Mourier et al., 2008). Since 2002 the amount of aligned functional genomics data has increased enormously for this species providing deeper evidence support for most genuine transcripts. However, due to a lack of supporting evidence, the majority of previously predicted ncRNAs were subsequently removed from the reference annotation. Currently there are 103 ncRNAs annotated (Extended data: Table 2 (Böhme, 2019)), including a recently described spliced antisense ncRNA that acts as a negative regulator (PF3D7_0935390), (Filarsky et al., 2018).

Using a combination of manual and automated methods, functional annotation has been radically improved since the initial publication in 2002. Revised annotation is based on a combination of literature searching, comments received from the research community, InterPro (Jones et al., 2014) and sequence-similarity searches. Using these approaches the number of proteins with unknown function has almost halved, from 60% in 2002 to 33% in March 2019 (Figure 4). The number of experimentally verified genes has changed from 597 genes in 2002 to 1296 in 2019 and the number of genes with putative functions has risen from 1215 to 2206. The richness of the annotation is reflected in the number of Gene Ontology (GO) terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2017) that have been manually assigned to genes based on published experiments, reflected in the following evidence codes: Inferred from direct assay (IDA), physical interaction (IP), mutant phenotype (IMP) and genetic interaction (IGI).

---

**Box 1. *Plasmodium falciparum* community reannotation workshop**

A *Plasmodium falciparum* community reannotation workshop, co-organised by EuPathDB (David Roos) and GeneDB (Matthew Berrien) took place at the Wellcome Genome Campus Conference Center in Hinxton in October 2007. During this week-long workshop several hundred gene products were changed, including the annotation of AP2 proteins.

**Workshop participants:**


**Facilitators during the workshop included:**

- Andrew Berry, Céline Carret, Al Ivens, Arnab Pain, Adrian Tivey, Brian Brunk, Zhongqiang Chen, Mark Heiges and Lucia Peixoto.

**Maintenance and setup of the Artemis Chado instance:**

- Tim Carver, Adrian Tivey, Chinnay Patel and Robin Houston.
Figure 1. Distribution of genes with structural changes and new genes on chromosomes 1 to 14 of *P. falciparum* 3D7. The positions of new genes (shown in red), genes that have undergone structural changes (shown in blue) and genes that stayed the same since 2002 (shown in grey) are shown on the 14 chromosomes. The values along the right of each chromosome indicate the total sequence length in base pairs.

Throughout the annotation improvement process, engagement of the malaria research community has played an important role. The process started with the workshop in 2007 (Box 1) but has continued through the activities of a dedicated full time curator aided by direct feedback and through comments that can be added by the community to gene record pages at PlasmoDB. These comments are constantly being evaluated and incorporated where relevant. The ongoing annotation is physically housed at the Sanger Institute, with updates regularly passed on to PlasmoDB (every 4 to 6 months).

**Population reference genome Pfref1**

One of the many purposes of a reference is to interpret natural variation data, the aim being to enable robust mapping of re-sequencing reads from subsequent isolates. In the latest version we have incorporated sequence differences derived from 3 lab
isolates assembled de novo as part of a collection of 15 PacBio reference genomes (Otto et al., 2018). The differences have been incorporated into the reference as three classes (Figure 5). The first are “patches” to correct errors in Pf3D7 (version 3.2), for example a missing centromere on chromosome 10 and a missing gene on chromosome 13 (Figure 5, Figure 6A), the second are core genes that are present in other sequenced isolates, i.e. P. falciparum IT or P. falciparum DD2 but are missing in Pf3D7 (Figure 5, Figure 6B) and the third are dimorphic genes where alternative alleles cannot be mapped to the one currently present in Pf3D7 (Figure 5, Figure 6C, Figure 6D). In total, there are now 17 type-1, four type-2 and 17 type-3 patches. The type 2 patches include genes encoding gamete associated protein (GAP), CLAG and hypothetical proteins. Type 3 include dimorphic genes encoding DBL-containing protein (PF3D7_0113800), Surfin 1.2, Surfin 8.3, Surfin 13.1, Surfin 14.1, MSP1, MSP2, MSP3, MSP6, S-antigen, EBA175, CLAG3.1, CLAG3.2, DBLMSP, DBLMSP2, CSP and VAR1CSA. For the new population reference, EMBL files containing the different types of patches including a 500 bp alignment on each side of the patch are provided. To reflect the fact that changes to 3D7 as well as data from other isolates have been included, we have termed the new reference sequence Pfref1.

**Discussion**

In this paper we have provided an overview of the sequence and annotation changes the *P. falciparum* 3D7 genome has undergone since the initial publication in 2002. The inclusion...
**Figure 3.** Diagram showing gene structure changes. Number of genes that have been added, deleted or changed are shown over four different time frames: October 2002 (genome version 1) and 2005 (genome version 2), between 2005 (version 2) and September 2007 (version 2.1.4), between September 2007 (genome version 2.1.4) and February 2010 (version 2.1.5) and between February 2010 and March 2019 (version 3.2). The number of changed genes includes gene models that have been merged, split or had a deletion/addition of exons or change of exon boundaries.

**Figure 4.** *P. falciparum* 3D7 annotation changes between October 2002 and March 2019. The number of protein coding genes between October 2002 and March 2019 are compared. The number of genes with unknown function is shown (blue), genes with experimental evidence (red), genes with putative function (yellow) and the complete number of genes (light blue). Genes with unknown function have the following product description: conserved Plasmodium protein, unknown function; conserved protein, unknown function; conserved Plasmodium membrane protein, unknown function; Plasmodium exported protein, unknown function; probable protein, unknown function; hypothetical protein.
Table 2. Gene Ontology (GO) annotation of *P. falciparum* 3D7 (version 3.1, 14.02.2019).

The number of manually curated experimentally verified GO terms, manually curated terms that are based on sequence similarity and GO terms based on automated searches are listed (Ashburner et al., 2000; The Gene Ontology Consortium, 2017). IDA, inferred from direct assay; IPI, inferred from physical interaction; IMP, inferred from mutant phenotype; IGI, inferred from genetic interaction; HAD, inferred from high throughput direct assay; ISS, inferred from sequence or structural similarity; ISO, inferred from sequence orthology; ISM, inferred from sequence model; RCA, Reviewed computation analysis; IEA, inferred from electronic annotation. The IEA evidence code is either based on InterPro (Jones et al., 2014) or based on Chitale et al., 2009 (Chitale et al., 2009). The GO annotation for 2002 was taken from Gardner et al., 2002.

<table>
<thead>
<tr>
<th>Gene Ontology</th>
<th>Number of GO annotations (2018)</th>
<th>Number of GO annotations (2002)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cellular Component</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental evidence code: IDA, IPI, IMP, IGI</td>
<td>6437</td>
<td>2,412</td>
</tr>
<tr>
<td>Evidence code based on sequence similarity: ISS, ISO, ISM</td>
<td>1,867</td>
<td></td>
</tr>
<tr>
<td>Evidence code: RCA</td>
<td>1,187</td>
<td></td>
</tr>
<tr>
<td>Evidence code: HDA</td>
<td>962</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (InterPro)</td>
<td>175</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (PMID:19435743)</td>
<td>1527</td>
<td></td>
</tr>
<tr>
<td><strong>Molecular Function</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental evidence code: IDA, IPI, IMP, IGI</td>
<td>4,921</td>
<td>1,244</td>
</tr>
<tr>
<td>Evidence code based on sequence similarity: ISS, ISO, ISM</td>
<td>979</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (InterPro)</td>
<td>1,111</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (PMID:19435743)</td>
<td>1,732</td>
<td></td>
</tr>
<tr>
<td><strong>Biological Process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental evidence code: IDA, IMP, IGI</td>
<td>3,803</td>
<td>1,301</td>
</tr>
<tr>
<td>Evidence code based on sequence similarity: ISS, ISO, ISM</td>
<td>857</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (InterPro)</td>
<td>1,270</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (PMID:19435743)</td>
<td>1,117</td>
<td></td>
</tr>
<tr>
<td>Evidence code: IEA (PMID:19435743)</td>
<td>559</td>
<td></td>
</tr>
</tbody>
</table>

of long-read sequencing has been critically important for spanning gaps that persisted for years in the reference assembly due to their extreme AT-richness and length. A previous attempt to produce an improved *P. falciparum* 3D7 reference assembly used Pacific Biosciences sequencing data assembled de novo (Vembar et al., 2016). Although the assembly contiguity metrics were impressive, the authors did not attempt error-correction. As a consequence, a high proportion of gene sequences contained frameshifts and there were many unresolved repetitive sequences. In the present study, we have used automated error-correction assisted by a high coverage of aligned short reads, plus extensive manual review of individual read alignments. This has enabled us to drive the accuracy of underlying reference sequence, bringing a range of benefits. First, end users interested in individual genes have access to the most up to date information. Second, users interested in high throughput functional genomics or genome variation need the most up to date and complete sequence for mapping purposes to be available and used by all labs. Third, detailed curation in *P. falciparum* has a knock on effect across other important species of the genus because functional insights from one species can be projected to others based on homology. It is inevitable that there is a law of diminishing returns. However, with 33% of genes still of unknown function it is essential that ongoing maintenance, annotation and curation are continued. In particular, our future plans include the annotation of untranslated regions (UTRs) and the annotation of additional common alternative splice-forms for genes. We also plan to provide better visibility for evidence codes that support protein descriptions but are currently only available as GFF format genome annotation files. New possibilities are now being explored for the community to get involved with annotation. GeneDB will soon provide an opportunity for the community to contribute directly to structural annotation. Equivalent to that of an annotator’s
Type 1 (sequence differences in Pf3D7 PacBio)

Type 2 (new gene)
Type 2.1 (anchored on both sides)

Type 2.2 (not anchored)

Type 3 (dimorphic gene)
Type 3.1 (anchored on both sides)

Type 3.2 (anchored on one side)

**Figure 5.** Diagram showing different types of patches created for the *P. falciparum* 3D7 population reference (PfRef1). Type 1 are sequence differences between the current *P. falciparum* 3D7 assembly version 3.2 and a new Pf3D7 PacBio assembly. 500 bp are provided on each side as anchor (shown in blue). Type 2 are new genes, that are either anchored on both sides (type 2.1), or not anchored (type 2.2). Type 3 are dimorphic genes that are either anchored on both sides (type 3.1) or anchored on one side (type 3.2).

view, the user will be able to view the curated *Plasmodium* genomes in Apollo (Lee et al., 2013), a collaborative genomic annotation editor which allows multiple users to access the data.

The genetic structure of *P. falciparum* populations is currently studied by aligning whole genome sequencing reads to the Pf3D7 reference (Manske et al., 2012; Neafsey et al., 2012). However, the approach has limitations: genomic regions that are absent or extremely diverged from 3D7 can not be analysed by alignment. This applies to genes that are completely missing as well as alleles of genes that are extremely different (either due to high sequence diversity or dimorphism). By supplementing *P. falciparum* 3D7 v3.2 with alternative loci we have created a combined new reference called Pfref1. Utilising such alternative loci, rather than simply excluding them from population analyses, remains a challenge and will require the further development of variant-calling methods. However, as a first step, the popular short-read alignment tool BWA-MEM (Li, 2013), used in the GATK variant call pipeline, has been able to perform alignments in an alternative-aware mode for several years.
Figure 6. Differences between *P. falciparum* 3D7 genome version 3.2, a PacBio assembly of *P. falciparum* 3D7 and two lab strains *P. falciparum* IT and DD2. ACT comparisons between regions of the above genomes. Coloured boxes represent genes. The red blocks between sequences represent sequence similarity (BLASTn). (A) In the current *P. falciparum* 3D7 genome assembly v3.2, a hypothetical protein on chromosome 13 is missing. This gene is present in a Pf3D7 PacBio assembly (shown in green). (B) Comparison between *P. falciparum* IT chromosome 11 and *P. falciparum* v3.2 chromosome 11. *P. falciparum* 3D7 is missing a hypothetical gene on chr11. This gene is present in *P. falciparum* IT (PfIT_110029300) (shown in green). (C) Comparison between Pf3D7 v3.2 chromosome 7 and *P. falciparum* DD2 chromosome 7. The comparison shows the dimorphic gene EBA175 erythrocyte binding antigen-175 (PF3D7_0731500). (D) Clustalx alignment of EBA175 from PfDD2 and Pf3D7. The area shown is the dimorphic part of the two genes.

Data availability

Underlying data

*P. falciparum* 3D7 is maintained in GeneDB (http://www.genedb.org) and updates are passed on at regular intervals to PlasmoDB (http://www.plasmodb.org).

The annotation (GFF) files are being extracted from the annotation database on a monthly basis. They are available on the following FTP site:
ftp://ftp.sanger.ac.uk/pub/genedb/releases/latest/Pfalciparum/

The latest version of the database at time of publication is available here:
ftp://ftp.sanger.ac.uk/pub/genedb/releases/2019-03/Pfalciparum/

The patches for the population reference are on the following FTP site.
ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/PfRef1

This page also includes an explanation of the naming of the short EMBL files and instructions on how to map them to the reference genome.

Extended data

This project contains the following extended data:

- Table_1.xlsx (P. falciparum 3D7 genes present in 2019 that were missing in 2002)
- Table_2.xlsx (List of ncRNAs present in P. falciparum 3D7 version 3.2)
- Table_3.xlsx (P. falciparum 3D7 genes that were deleted anytime between 2002 and 2019)

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Software availability
The Artemis software used for annotation is available on GitHub: http://sanger-pathogens.github.io/Artemis/Artemis/

Grant information
This work was supported by the Wellcome Trust through a core grant to the Wellcome Trust Sanger Institute [206194] and a Wellcome Investigator Award to CIN [104792].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments
We would like to thank Bart Barrell for contributing to the structural annotation. We would also like to thank Daniel Deplde for his contribution to the GO annotation, Blessing Adamu for her work on ncRNAs and Andrew Berry for Plasmodium curation between 2004 and 2007. We would like to thank Sue Kyes and Zoe Christodolou for resolving the chromosome 7/8 misassembly. We would like to thank PlasmoDB for providing the main portal to access Plasmodium genome data and for setting up an option to add user comments to gene record pages.

References


