BIOINFORMATICS

Vol. 24 ISMB 2008, pages i383–i389
doi:10.1093/bioinformatics/btn140

Assessment and improvement of the *Plasmodium yoelii yoelii* genome annotation through comparative analysis

Ashley Vaughan[1], Sum-Ying Chiu[2], Gowthaman Ramasamy[1], Ling Li[1], Malcolm J. Gardner[1], Alice S. Tarun[1], Stefan H.I. Kappe[1,3,*] and Xinxia Peng[1,*]

[1]Seattle Biomedical Research Institute, Seattle, WA 98109, [2]Department of Biology and [3]Department of Pathobiology, University of Washington, Seattle, WA 98195, USA

ABSTRACT

Motivation: The sequencing of the *Plasmodium yoelii* genome, a model rodent malaria parasite, has greatly facilitated research for the development of new drug and vaccine candidates against malaria. Unfortunately, only preliminary gene models were annotated on the partially sequenced genome, mostly by *in silico* gene prediction, and there has been no major improvement of the annotation since 2002.

Results: Here we report on a systematic assessment of the accuracy of the genome annotation based on a detailed analysis of a comprehensive set of cDNA sequences and proteomics data. We found that the coverage of the current annotation tends to be biased toward genes expressed in the blood stages of the parasite life cycle. Based on our proteomic analysis, we estimate that about 15% of the liver stage proteome data we have generated is absent from the current annotation. Through comparative analysis we identified and manually curated a further 510 *P. yoelii* genes which have clear orthologs in the *P. falciparum* genome, but were not present or incorrectly annotated in the current annotation. This study suggests that improvements of the current *P. yoelii* genome annotation should focus on genes expressed in stages other than blood stages. Comparative analysis will be critically helpful for this re-annotation. The addition of newly annotated genes will facilitate the use of *P. yoelii* as a model system for studying human malaria.

Contact: xinxia.peng@sbri.org; stefan.kappe@sbri.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Malaria, a disease that causes more than 500 million clinical cases each year and kills more than 1 million worldwide in developing countries, is caused by unicellular *Plasmodium* parasites (Snow *et al*., 2005). *Plasmodium* is an obligatory intracellular parasite and its progression through the life cycle involves invasion of and intracellular replication in different tissues of both the mosquito vector and the mammalian host. Within the mammalian host, mousquito-inoculated salivary gland sporozoites of the parasite invade the liver and undergo an intracellular replication phase (schizogony) to produce merozoite stages that are competent for red blood cell infection. This initial blood stage infection establishes the blood stage cycle responsible for the pathology of malaria.

Facilitated by the *P. falciparum* genome sequencing project (Gardner *et al*., 2002), the genome of *Plasmodium yoelii* was sequenced to 5-fold genome coverage using the whole-genome shotgun sequencing approach and was annotated in 2002 (Carlton *et al*., 2002). Comparative genome analysis has revealed high gene sequence similarities and genome-wide synteny among the human parasite *P. falciparum* and the rodent malaria models (Kooij *et al*., 2005; Roy and Hartl, 2006). *P. yoelii* is an important animal model for studying malaria biology, particularly of the liver stages and host–parasite interactions. The *P. yoelii* pre-erythrocytic infection of the mouse represents an accepted model closely reflecting human malaria (Luke and Hoffman, 2003). It is also a better model than the rodent model *P. berghei* for studying the elusive liver stages because it induces less inflammation in murine livers and produces more liver stage parasites (Tarun *et al*., 2006).

The partial genome sequencing approach has enabled the development of genome-wide approaches such as microarrays and proteomics to study parasite biology (Tarun *et al*., 2008). However, the genes in the *P. yoelii* genome database were mostly identified by *in silico* gene prediction and there is an inherent propensity for error using this methodology, especially when the genome is AT rich. For example, the *P. yoelii* S4 gene (Kaiser *et al*., 2004) is not included in the current annotation. Instead, a three-exon gene (PY00180 in PlasmoDB, http://www.plasmodb.org) is annotated on the opposite strand in the same genomic location. Later, S4 was reported to mediate parasite infection in its mosquito and vertebrate hosts (Kariu *et al*., 2006), clearly demonstrating the importance of this un-annotated gene in parasite lifecycle progression. A recent study estimated that ~24% of the annotated *P. falciparum* genes contain errors of varying severity (Lu *et al*., 2007). Since the *P. yoelii* genome was partially sequenced and therefore not fully assembled, the error rate in the gene annotation is likely to be even higher when compared to the annotation of the fully sequenced and assembled *P. falciparum* genome (Salzberg, 2007). As an example of this error rate, the intensely studied gene encoding the circumsporozoite protein (Kumar *et al*., 2006) has been incorrectly annotated as two partial genes in the *P. yoelii* genome annotation (PY07368 and PY03168 in PlasmoDB).

The inaccurate annotation of the *P. yoelii* genome greatly compromises genome-wide studies. Thus, there is clearly a need for an updated and accurately annotated *P. yoelii* genome sequence. Fortunately, additional *Plasmodium* species have been sequenced (Hall *et al*., 2005), or are being sequenced (http://www.tigr.org/tdb/e2k1/pva1/). Furthermore, the *Plasmodium* cDNA sequences available in Genbank have also grown dramatically since 2002. These provide an opportunity to examine the existing *P. yoelii* genome annotation using comparative analysis, to improve the current annotation and to discover missed genes. Here we describe

[*] To whom correspondence should be addressed.

a systematic assessment of the current *P. yoelii* annotation based on the publicly available cDNA sequence data and locally generated proteomic data. Based on the comparative analysis of published *Plasmodium* genome sequences, we have identified and manually curated 510 *P. yoelii* genes that have clear orthologs in *P. falciparum*, but are absent or incorrectly annotated in the current database. The improved annotation will greatly facilitate the use of the rodent malaria system in the modeling of human *Plasmodium* biology.

2 METHODS

2.1 cDNA sequence trimming and assembly

All *Plasmodium* expressed sequence tag (EST) sequences were downloaded from Genbank (August 2, 2007). To identify segments of EST sequences of vector origin, EST sequences were blasted against NCBI's UniVec Database using similar parameters as VecScreen (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html). Segments of vector origin were trimmed and only sequences with at least 100 bp were used. Trimmed EST sequences were then blasted against *Plasmodium* ribosomal gene sequences. Possible ribosomal RNA sequences were identified (with a match to a ribosomal gene of 40 bp or longer with a minimum of 90% identity) and removed. The ESTs of possible host origin were similarly identified and removed. The remaining EST sequences were clustered and assembled using TGI clustering tools (Pertea *et al*., 2003). Consensus sequences (contigs) were built based on two or more EST sequences that overlap for at least 40 bases with at least 94% sequence identity. These strict criteria helped to minimize the creation of chimeric contigs.

2.2 Proteomics analysis

Proteomics data were generated as described (Tarun *et al*., 2008). In brief, liver stage-infected hepatocytes were isolated from GFP-expressing *P. yoelii*-infected BALB/c mice by fluorescence-activated cell sorting. Total protein was extracted from sorted liver stage-infected hepatocytes and separated by polyacrylamide gel electrophoresis. Proteins were cut from the gel and digested in-gel with sequencing grade-modified trypsin. Mass spectra were obtained by nano-flow liquid chromatography tandem mass spectrometry (nano LC-MS/MS).

In an attempt to account for missing, partial or erroneous *P. yoelii* gene models, we assembled two additional databases besides a standard sequence database that combined the currently annotated *P. yoelii* protein sequences with the currently annotated *Mus musculus* protein sequences. The first database was assembled with annotated *P. berghei* protein sequences (PlasmoDB version 5.2) supplemented with *M.musculus* protein sequences. The *M.musculus* genome database was downloaded from the NCBI FTP site (http://www.ncbi.nlm.nih.gov) on May 4, 2006. We constructed a second database to search for additional peptides that might not be found using current annotations. To do this, we translated all the *P. yoelii* genomic sequences downloaded from Genbank in all six frames and extracted all stop-to-stop open reading frames (ORFs). We used this approach since many *P. yoelii* genes have multiple exons and thus, not all peptides will be identified by searching a straightforward six-frame translation for contiguous sequences that contain start and stop codons. For instance, those peptides spanning exon boundaries will be missed with a standard six-frame ORF search, but they are critical for confirming and annotating splicing. To overcome the limitations of this straight-forward search, we also translated both the contigs and singlets derived from the EST analysis as described above in all six frames and again extracted all stop-to-stop ORFs and added these peptides to the second database. The six-frame translation was achieved using the EMBOSS package (Rice *et al*., 2000). All ORFs over 30 amino acids are merged. Redundant entries were removed using the WU-BLAST package [Gish W. (1996–2004) http://blast.wustl.edu]. This second sequence database was supplemented with *M.musculus* protein sequences as before.

The nano LC-MS/MS data (Tarun *et al*., 2008) were searched against each of the three sequence databases separately using BioworksBrowser 3.1 SR1 (San Jose, CA, ThermoElectron), which uses the SEQUEST algorithm (Eng *et al*., 1994). The search results were validated using PeptideProphet version 3.0 (Keller *et al*., 2002) for peptide identification. The PeptideProphet program scores the assignment of a particular MS spectrum to a known peptide. We considered only those PeptideProphet scores with a minimum P-value of 0.85, which in this case corresponds to a probability >98% of the peptide being correct. Depending on the database used for the analysis, single MS spectra could be assigned to different peptides. We kept only those spectra to which all peptides assigned were similar (in this case, they had no more than two amino acid differences between them). A total of 12 276 spectra with a minimum *P*-value of 0.85 were successfully assigned to *Plasmodium* protein sequences and 585 of the 12276 were assigned with two or more different peptides. Where common peptides differed slightly, a single representative peptide for each spectrum was chosen. Identified peptides were then mapped to the annotated *P. yoelii* protein sequences, genomic DNA sequences and EST sequences using tools provided by EMBOSS (Rice *et al*., 2000).

2.3 Orthology mapping

To identify *P. yoelii* genes that were not included in the original annotation or incorrectly annotated, we created a database of homologous proteins, which mapped proteins from four *Plasmodium* species (*P. falciparum*, *P. yoelii*, *P. berghei* and *P. chabaudi*) into one-to-one orthologous relationships. The annotated protein sequences of the four *Plasmodium* species were downloaded from PlasmoDB version 5.2. All pair-wise comparisons among protein sequences from any two of the four genomes were performed using NCBI BLAST. For each protein, the best hit in each of the other genomes was detected. Orthologous proteins from two genomes were defined as reciprocal best hits (BLAST E-value $< 1e{-}15$ as cutoff).

2.4 Protein features

Signal peptides were predicted using the SignalP 3.0 Server (Bendtsen *et al*., 2004) and only ORFs with a start codon were considered. Transmembrane domains were predicted using TMHMM Server v. 2.0 (Krogh *et al*., 2001). For Pfam domain annotations, protein sequences were searched against the Interpro database Release 15.0 (Mulder *et al*., 2007) using InterProScan web service.

2.5 PCR verification of newly annotated genes

Genomic and total RNA were prepared from mixed blood stage parasites obtained from mice infected with *P. yoelii* at 3–5% parasitemia. Genomic DNA was extracted using the DNeasy kit (Valencia, CA, Qiagen) while total RNA was extracted using Trizol (Carlsbad, CA, Invitrogen) according to manufacturer's instructions. Total RNA was treated with Turbo-free RNAse-free DNAse (Austin, TX, Ambion) to remove contaminating genomic DNA in the preparation. First-strand cDNA was synthesized from 500 ng of total RNA using the Superscript III Platinum RT kit (Invitrogen). For each 25 µl PCR reaction, genomic DNA (25 ng) or cDNA (from 5 ng of total RNA) was used as template with 25 pmol of the forward and reverse primers (Table 5) and 12.5 µl of the Bioline Red PCR mix (Randolph, MA, Bioline) using the following cycling conditions: initial denaturation at 95°C for 3 min; 30 cycles of 94°C for 30 s, 55°C for 45 s and 72°C for 1 min; final extension of 72°C for 7 min.

3 RESULTS

3.1 Assessment of the *P. yoelii* genome annotation using cDNA sequences

EST sequences have been shown to be a powerful aid in *Plasmodium* gene identification and genome annotation (Carlton *et al*., 2002;

Table 1. Summary of the enrichment and pre-processing of *Plasmodium* EST data from Genbank

Organism	10/2002	8/2007	Cleaned	Contigs (No. of ESTs)	Singlets	Total	Year of Coverage	Publication	No. of full-length genes
All four *Plasmodium*	39 848	138 827	121 308	15 605 (93 558)	27 750	43 355			
P. yoelii	15 562	18 932	18 319	2401 (12 293)	6026	8427	2002	5×	5878
P. berghei	5544	58 955	43 665	5379 (32 528)	11 137	16 516	2005	4×	5864
P. falciparum	18 742	38 702	37 959	5118 (30 875)	7084	12 202	2002	14.5×	5268
P.vivax	0	22 238	21 365	2707 (17 862)	3503	6210	2005	4×	5698

Table 2. Overview of the available EST evidence for 7861 currently annotated and 510 re-annotated *P. yoelii* CDS

Shorter than	Complete			Partial			Re-annotated		
	Total	w EST	Percentage	Total	w EST	Percentage	Total	w EST	Percentage
Total	5904	4766	81	1957	1095	56	510	452	89
200 aa	1826	962	53	946	237	25	248	220	89
180 aa	1676	832	50	910	214	24	213	189	89
160 aa	1509	690	46	874	182	21	182	159	87
140 aa	1355	558	41	844	161	19	149	128	86
120 aa	1180	412	35	797	126	16	107	92	86
100 aa	1015	286	28	760	99	13	74	64	86
80 aa	861	180	21	706	66	9	42	35	83

Gardner *et al*., 2002; Hall *et al*., 2005; Li *et al*., 2003). Since the publication of the annotated *P. yoelii* genome in 2002 (Carlton *et al*., 2002), the number of *Plasmodium* cDNA sequences has increased dramatically (Table 1). In October 2002, there were 39 848 *Plasmodium* ESTs in Genbank. By August 2007, this number had increased by ∼350% to 138 827. Of note, is the generation of liver stage EST libraries (Ishino *et al*., GenBank accession DC195411-DC201252; Sacci *et al*., 2005; Wang *et al*., 2004). The liver stage has historically been extremely difficult to study (Blair and Carucci, 2005), but new tools have enabled isolation of rodent liver stages for analysis. The enrichment of *Plasmodium* cDNA sequences from all life stages has allowed us to examine the current *P. yoelii* annotation in far greater detail than previously possible.

We downloaded all available *Plasmodium* ESTs from Genbank, cleaned them, and then assembled them into large contigs after clustering (Table 1). We then mapped the contigs and singlets onto the current *P. yoelii* genome annotation by sequence similarity analysis using NCBI BLAST (an E-value $<1e-15$ cutoff was used).

We classified the annotated genes into two categories (Table 2). A gene was considered as ‘complete’ if its annotated coding sequence (CDS) contained a start and stop codon. Otherwise the gene was considered ‘partial’. Having EST evidence suggests that the annotated gene is very likely to encode a transcript. However, the exact boundaries of the gene cannot be verified explicitly unless an EST or EST contig contains a full length CDS. As shown in Table 2, longer gene sequences were better supported by an EST. It is possible that longer annotated genes were more likely to have representative ESTs. It is also possible that smaller genes were more likely to be incorrectly annotated, since short CDSs are more challenging for accurate computational prediction. When genes were of *similar* sizes, the gene was more likely to have a representative EST if its annotation was complete, rather than partial (Table 2). This implies that complete gene models are more likely to be correctly annotated.

Table 3. Summary of mapping *P. yoelii* ESTs to annotated *P. yoelii* CDSs

cDNA library	Total	Cleaned	Mapped	Unmapped (%)
Asexual blood stages	12 471	12 043	10 474	1569 (13)
Salivary gland sporozoite	3091	3072	2005	1067 (34.7)
Axenic early liver stages (<24 h; trophozoite)	1452	1387	681	706 (**50.9**)
Liver stage (40 h; schizont)	1916	1815	845	970 (**53.4**)

Highlighted in bold are the percentage of ESTs from liver stage libraries which were not mapped to any annotated *P. yoelii* CDSs.

We found that there were significant numbers of *P. yoelii* ESTs that were not mapped to currently annotated *P. yoelii* CDSs (Table 3). Some un-mapped ESTs are likely to correspond to the untranslated regions of genes (both 5′ and 3′), since only CDSs were used in the analysis. Interestingly there were relatively more ESTs from liver stage libraries that were not mapped to any annotated *P. yoelii* CDSs compared to those ESTs from blood stages (Table 3). About 50% of the liver stage EST sequences were not mapped, while only about 13% of blood stage EST sequences were not mapped. This suggests that the current annotation has a better coverage of transcripts expressed in blood stages than those expressed in liver stages.

3.2 Assessment of the *P. yoelii* genome annotation by proteomic analysis

To further investigate the impact of current annotations on the studies of liver stage parasites, we used proteome data we generated from

Table 4. Analysis of peptides identified in liver stage by proteomics using different databases

	Summary		Peptides mapped to *P. yoelii* CDSs, genomic sequences, ESTs and *plasmodium* ESTs (%)				Peptides mapped to *P. yoelii* or *P. berghei* ESTs from liver stage libraries[a]			
	Total	Not in Py (%)[b]	PyCDS	PyGdna	PyEST	PlasEST	PlasESTLS	PyLCM	PyAxenic	PbLS31
PbCDS	2644	87 (3.3)	2308 (87.3)	2331 (88.2)	1666 (63)	2249 (85.1)	952 ***112 (11.8)***	240 ***23 (9.6)***	156 ***18 (11.5)***	862 ***104 (12.1)***
6f	3512	244 (6.9)	2813 (80.1)	3006 (85.6)	2002 (57)	2763 (78.7)	993 ***148 (14.9)***	284 ***46 (16.2)***	199 ***30 (15.1)***	805 ***115 (14.3)***
PyCDS	3459	NA	3459 (NA)	3331 (96.3)	2191 (63.3)	2687 (77.7)	1005 ***NA***	300 ***NA***	196 ***NA***	810 ***NA***
Combined[c]	4234	309 (7.3)	3547 (83.8)	3717 (87.8)	2453 (57.9)	3264 (77.1)	1190 ***172 (14.5)***	355 ***48 (13.5)***	238 ***35 (14.7)***	963 ***138 (14.3)***

[a] Highlighted in bold and italics are the number (and the percentage) of peptides identified which can be mapped to ESTs from the indicated liver stage library, but not to any annotated *P. yoelii* protein sequence.

[b] The number (and the percentage) of peptides identified in each database search which cannot be mapped to any *P. yoelii* sequence in any of PyCDS, PyGdna and PyEST.

[c] The total number of peptides (non-redundant) identified from all three database searches.

PyCDS: annotated *P. yoelii* protein sequences. 6f: protein sequences from translation of *P. yoelii* genomic sequences and all *Plasmodium* EST sequences in six-frames. PbCDS: annotated *P. berghei* protein sequences. PyGdna: *P. yoelii* genomic sequences. PyEST: *P. yoelii* EST sequences. PlasEST: all *Plasmodium* EST sequences. PyLCM: *P. yoelii* EST sequences from Sacci *et al*. (2005). PyAxenic: *P. yoelii* EST sequences from Wang *et al*. (2004). PbLS31: *P. berghei* EST sequences from (Ishino *et al*. GenBank accession DC195411-DC201252). PlasESTLS: liver EST sequences from three liver stage libraries: PyLCM, PyAxenic and PbLS31. NA: not applicable. See text for details.

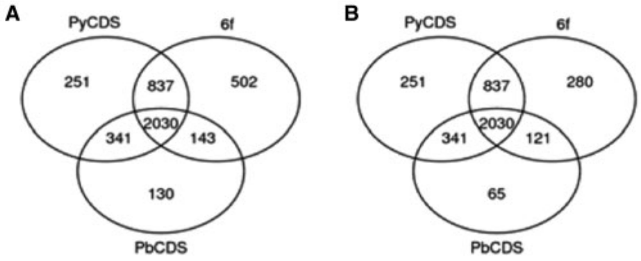

Fig. 1. Venn diagram of peptides identified using different sequence databases. (**A**) Total number of peptides identified in each database search. (**B**) The subset of peptides identified which can be mapped to a *P. yoelii* sequence in any of PyCDS, PyGdna and PyEST. PyCDS, 6f, PbCDS, PyGdna and PyEST: see Table 4 legend.

purified *P. yoelii* liver stages (Tarun *et al*., 2008). Since there is a high level of sequence similarity (91.3% identity at nucleotide level) between the *P. yoelii* and *P. berghei* genomes (Kooij *et al*., 2005), we hypothesized that a number of missing *P. yoelii* protein annotations would also be found by searching the annotated genome of the closely related *P. berghei*.

We prepared a protein sequence database that included the annotated *P. berghei* protein sequences from the 4× coverage genome sequences (Hall *et al*., 2005) and the updated mouse protein sequences from NCBI. We then searched the spectra generated from our liver stage *P. yoelii* samples against this combined database. In total, 2644 unique *P. berghei* peptides were identified that had matches with our *P. yoelii* liver stage proteome data (Table 4 and Fig. 1). Only 87 (3.3%) of the peptides could not be perfectly mapped to any available *P. yoelii* sequence, suggesting that database searching *P. berghei* sequences or sequences from related *Plasmodium* species will supplement available *P. yoelii* annotations. Of the 2644 peptides, 2308 (87%), 2331 (88%) and 1666 (63%) peptides were perfectly mapped to *P. yoelii* annotated CDSs (PyCDS), genomic DNA (PyGdna) and EST (PyEST) sequences, respectively (Table 4). A total of 2249 peptides perfectly mapped to *Plasmodium* ESTs (Table 4, PlasEST) and of those, 952 matched to ESTs from three liver stage libraries (Table 4): PyLCM (Sacci *et al*., 2005), PyAxenic (Wang *et al*., 2004) and PbLS31 (Ishino *et al*. GenBank accession DC195411-DC201252). Of the 952 liver stage matched ESTs, 112 did not have a perfect match to any annotated *P. yoelii* proteins—about 12%. Similar estimations could be obtained when individual liver stage libraries were considered separately (9.6% for PyLCM, 11.5% for PyAxenic and 12.1% for PbLS31). Since the *P. berghei* genome sequence was only 4× coverage, we hypothesized that additional peptides were not identified by this analysis. The complete list of identified peptides is shown in Supplementary Table 1.

To search for additional peptides that were not identified because of the limitations of the current annotations, we compiled a second protein sequence database. The second database contained: (1) all stop-to-stop ORFs over 30 amino acids after translating available *P. yoelii* genomic sequences from Genbank in six-frames, (2) stop-to-stop ORFs over 30 amino acids after translating pre-processed *Plasmodium* cDNA contigs and singlets in all six-frames and (3) the updated mouse protein sequences. The peptides were identified in the same way as above with this combined sequence database.

In total, 3512 unique peptides were identified (Table 4 and Fig. 1). Of these 3512 peptides, 2763 mapped perfectly to *Plasmodium* ESTs and 993 perfectly mapped to ESTs from liver stage libraries (PyLCM, PyAxenic and PbLS31). However, as was seen with the *P. berghei* database search, a certain percentage of peptides that matched to liver stage ESTs did not have a perfect match with the current *P. yoelii* protein annotation (15%, 148 of 993). As before, we obtained similar estimations when individual liver stage libraries were considered separately (16% for PyLCM, 15.1% for PyAxenic and 14.3% for PbLS31). Since the second sequence database was more comprehensive compared to the annotated *P. berghei* sequence

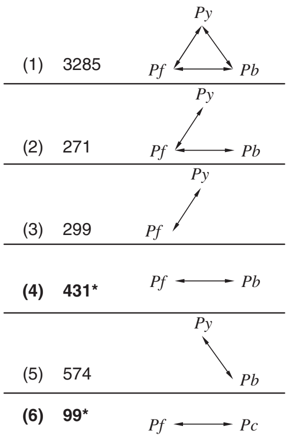

Fig. 2. The strategy for the identification of orthologous *P. yoelii* genes potentially absent from the current annotation by comparative analysis. Orthologous genes from four *Plasmodium* species were identified based on the reciprocal best BLAST hit approach, and classified into six categories. Py: *P. yoelii*. Pb: *P. berghei*. Pf: *P. falciparum*. Pc: *P.chabaudi*. Solid lines with two arrows indicate proteins from two corresponding species are reciprocal best blast hits. The number on the left indicates the number of orthologous groups in each category. The asterisk indicates the orthologous groups from which potential P. *yoelii* genes orthologous to *P. falciparum* genes were identified (also highlighted in bold).

database, and used sequences from the same species that the proteome data came from (*P. yoelii*), we consider this estimation to be closer to the actual situation. Thus, 15% of our liver stage proteome data that matches available liver stage EST data is not contained in the current *P. yoelii* annotation, suggesting that a significant number of *P. yoelii* genes/exons are not represented in the current annotation. There were 406 peptides were identified by searching annotated *P. berghei* protein sequences but not translated *P. yoelii* genome (Fig. 1B), showing that database searching of sequences from related *Plasmodium* species can supplement the partially sequenced *P. yoelii* genome. The complete list of identified peptides is shown in Supplementary Table 1.

3.3 Identification of missing and incorrectly annotated *P. yoelii* genes by comparative analysis

To identify missing and incorrectly annotated genes in the annotated *P. yoelii* genome we took advantage of the recent genome publication of two closely related rodent *Plasmodium* species—*P. berghei* and *P.chabaudi* (Hall *et al.*, 2005). Additional species such as *P.vivax* and *P.knowlesi* are also being sequenced. These additional sequences provide a great opportunity to improve current annotation by comparative analysis. Because rodent malaria parasites are used as model systems for studying human malaria parasites, we chose to focus on those genes that are well conserved between the human malaria parasite *P. falciparum* and the three rodent malaria parasites (*P. yoelii*, *P. berghei* and *P. chabaudi*) using the reciprocal best BLAST hit approach (Tatusov *et al.*, 1997). First we identified a core set of genes conserved between *P. falciparum* and three rodent malaria parasites. Centered around *P. falciparum* proteins, we classified the mapping of orthologous proteins into six categories in the following order (Fig. 2): (1) those with three-way best BLAST hits among proteins from *P. yoelii*, *P. berghei* and *P. falciparum* (3285 proteins), (2) two-way best BLAST hits between proteins from *P. falciparum* and *P. yoelii* and *P. falciparum* and *P. berghei* only (271 proteins), (3) two-way best BLAST hits between only *P. falciparum* and *P. yoelii* (299 proteins) (4) two-way best BLAST hits between proteins from only *P. falciparum* and *P. berghei* (431 proteins), (5) two-way best BLAST hits between proteins from *P. yoelii* and *P. berghei* only (574 proteins) and (6) the two-way best BLAST hits between proteins from *P. falciparum* and *P.chabaudi*.

Table 5. Summary characteristics of re-annotated orthologous *P. yoelii* genes

Category	Re-annotated genes
Total number	510
Span two or more Contigs	109
AAs (Mean, Min, Max)	310, 19, 4783
Transmembrane domain	116
Signal peptide	80
EST hit	452
Pfam domain	157

The analysis enabled us to identify 530 *P. falciparum* genes which had clear *P. berghei* and/or *P.chabaudi* orthologs, but no *P. yoelii* orthologs based on the reciprocal best BLAST hit analysis (Fig. 2). This analysis provided strong evidence that those orthologous genes were absent from or incorrectly annotated in the current *P. yoelii* genome annotation, based on the assumption that the three rodent malaria parasite species are so closely related that genes are more conserved among the three than between any of the three and the *P. falciparum* genome. The list of candidate *P. falciparum* genes is provided in Supplementary Table 2.

For each of the 530 candidate *P. falciparum* genes identified above, we attempted to manually annotate the orthologous gene in *P. yoelii* based on available genome sequences. First we searched candidate *P. yoelii* contigs by blasting the *P. falciparum* gene and its orthologs in other rodent species against the *P. yoelii* genomic sequences. Related EST sequences were also identified. The alignments were examined manually. For genes with multiple exons, we considered only those introns with the conserved eukaryotic intron splice site GT-AG (Lu *et al.*, 2007). Publicly available software, databases and customized scripts were used to facilitate the manual annotation (Brejova *et al.*, 2005; Rice *et al.*, 2000; Slater and Birney, 2005). An internal MySQL database and the associated web interface were constructed to manage the data flow.

In total, we were able to manually annotate 510 orthologous *P. yoelii* genes from the 530 candidate *P. falciparum* genes (Table 5). About 90% (452/510) of these newly annotated genes had EST support, which is much higher than that of the existing annotated genes (Table 2). Interestingly, this new set of genes did not show the trend that larger genes tended to be better supported by EST evidence, as was seen in the existing annotated genes (Table 2). This implies that our manual curation is more accurate than the *in silico* prediction. About 21% (109/510) of the newly annotated genes spanned two or more contigs. The submission of newly annotated genes to public databases is in process. The current detailed annotations are provided in Supplementary Table 3.

Table 6. PCR verification of newly annotated genes

Py Gene ID	Forward primer	Reverse primer	Genom	No. of intron	cDNA
PY_PFL0415w	CACAATCGTTATGCGAAAATG	TCCCATGCTCTATTATCTTTGG	356	0	356
PY_PF14_0205	AAAAACCTTCATTTTATTTTATCTCCA	TTTTTGAGAGTGACTTTGAATGC	651	1	317[a]
PY_PF14_0623	GCGAACTTAACGGGATCTCA	GCACACCGATCCTTTCTCTT	1388	4	756
PY_PF14_0612	CGGCGGGCTTATATTAAAAA	AGCAGCTCGTAATGCATCCT	809	3	310[a]
PY_PFD0775c	CCCCCAAAGATTTGTCTGAA	TGCTCCTAAAACATTTCCCATA	1775	4	1415[a]
PY_PFB0885w	TGGGTAAGTTTAAAGCGATTTTT	ACTTTTGAGTTAGGCCCTTTTT	516	1	175[a]
1.PY_PFB0505c[b]	TGGTCATTCTTATCCTTCACATGAA	ACATTCCAGCCCCAAAACC	∼1500	∼5	∼900[a]
2.PY_PFB0505c[b]	GATAATTTAGATGCCCCAAACCAA	ACATTCCAGCCCCAAAACC	645	2	350
3.PY_PFB0505c[b]	AAACACATCAGCAGCTTCAATACC	ACATTCCAGCCCCAAAACC	252	1	92[a]

[a] Multiple PCR products observed from cDNA suggesting alternative splicing.

[b] Three different forward primers were used to verify the intron positions and confirm that the gene spanned two contigs (see text).

Genom, expected size in base pair from genomic DNA; cDNA, expected size in base pair from cDNA.

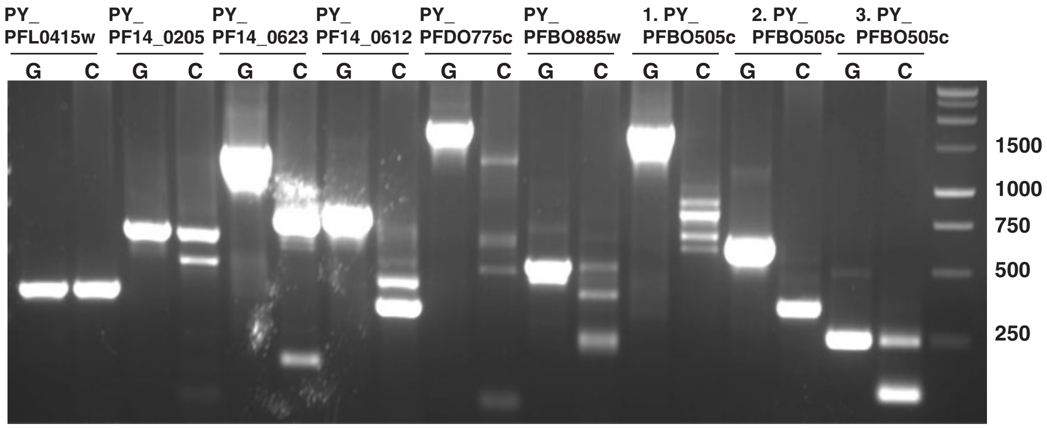

Fig. 3. PCR products confirming the expression of newly annotated genes. Oligonucleotide primers flanking selected predicted genes were amplified from genomic DNA (gDNA, **G**) and reverse-transcribed mRNA of mixed asexual stages (cDNA, **C**).

We have verified the expression of seven of these re-annotated genes by PCR (Table 6 and Fig. 3). PCR amplification of cDNA prepared from mixed blood stages of *P. yoelii* confirms the presence of introns since the amplified products are smaller than the corresponding amplified DNA from genomic DNA. In six of the eight amplifications where introns are predicted, multiple bands in the resulting amplified cDNA were observed suggestive of alternative splicing. The gene for PY_PFB0505c, encoding beta-ketoacyl-acyl carrier protein synthase III, is predicted to be spanned by two separate contigs (Genbank ID 23485908 and 83285534) that encode the amino and carboxyl sequences of the protein. Although sequence information for the central portion of the gene is missing, using a forward primer from contig 23485908 and a reverse primer from contig 83285534, a ∼1.5 kb fragment is amplified from genomic DNA while multiple bands are observed from cDNA confirming that these two contigs are contiguous (1.PY_PFB0505c in Table 6 and Fig. 3).

4 DISCUSSION AND CONCLUSION

The accumulation of publicly available *Plasmodium* genomic and cDNA sequences enabled us to carefully examine the existing *P. yoelii* genomic annotations using comparative analysis. Our analysis confirmed that the incompleteness of the *P. yoelii* genome data has a considerable effect on the accuracy of the gene annotation (Salzberg, 2007). There were 5812 gaps left in the *P. yoelii* genome sequences when the sequencing project was finished (Carlton *et al*., 2002). Subsequent annotation was limited to ∼87% (20 of 23 Mb) of the genome. Therefore, many genes ‘run-off’ at the end of contigs. About 21% of the genes that we re-annotated were split across two or more contigs. Many partial genes, i.e. those genes without either an annotated start or stop codon, were predicted from these short fragments of DNA. Partial genes were less likely to be supported by EST sequences as shown in Table 2, and we feel that these predictions are less accurate than those for complete genes. Additional sequencing to close these gaps will greatly facilitate the improvement of the current annotation. In addition, it was recently estimated that ∼24% of the *P. falciparum* genes in the current database have been incorrectly predicted (Lu *et al*., 2007). Since the *P. yoelii* genes were predicted using similar bioinformatics software, we expect that improved gene finders may be able to generate more accurate gene models. For example, our limited experience with a gene finder known as ExonHunter (Brejova *et al*., 2005) indicates that similar methodologies using additional sources of information, especially *Plasmodium* EST sequences and other *Plasmodium* genome annotations, will greatly improve the prediction accuracy on the available sequences.

Liver stages of the parasite life cycle are considered ideal targets for malaria vaccine and drug development. Our analysis indicates that the genes expressed in liver stages are under-represented in the current *P. yoelii* genome annotation compared to blood stages. This bias could further compromise the study of liver stages (Blair and Carucci, 2005). The likely reason for the bias is that many blood stage ESTs were generated before the *P. yoelii* genome was annotated and were subsequently incorporated into the genome annotation. However, liver stage EST sequences only became available after the annotation of the *P. yoelii* genome was completed. It is unlikely that genes expressed in liver stages are dramatically different to those expressed in other stages, thus a separate training dataset for gene finders will not be necessary to detect these genes in the genome. The availability of the liver stage ESTs, however, assist gene identification and the selection of correct gene models from alternative predictions. Since there are no *P. falciparum* liver stage EST sequences, it is likely that the *P. falciparum* genome annotation is also lacking in liver stage proteins. Therefore, annotating genes expressed in liver stages will benefit greatly from liver stage proteome and EST data from rodent malaria models.

Based on the proteomic analysis, we estimated the percentage of protein sequences that are not covered by current annotation to be 12∼15%. Since the database from six-frame translations was more comprehensive compared to the one with annotated *P. berghei* protein sequences and used the sequences from the same species, we consider that the 15% is more realistic. Like the ESTs, the liver stage peptides presented here will provide a great resource for the future improvement of *P. yoelii* annotation and the identification of proteins expressed in the liver stages.

We have provided here a systematic assessment of the accuracy of the *P. yoelii* genome annotation based on a detailed analysis of a comprehensive *Plasmodium* EST sequence database. The impact of the sub-optimal genome annotation on genome-wide experiments was evaluated using locally generated proteomics data. We identified a subset of *P. yoelii* genes that have clear orthologs in the *P. falciparum* genome but are absent from, or incorrectly annotated in current databases. These genes were manually annotated. This set of newly annotated genes will greatly facilitate the use of *P. yoelii* as a model system for studying human malaria parasite *P. falciparum*.

ACKNOWLEDGEMENTS

We thank Brona Brejova and Tomas Vinar for providing ExonHunter and the training on *Plasmodium* sequences.

Funding: This bioinformatics work is funded in part by a Grant from the Foundation for the National Institutes of Health through the Grand Challenges in Global Health initiative and the Ellison Medical Foundation (S.H.I.K.). Proteome analysis is supported by a Seattle Biomedical Research Institute Innovation grant (S.H.I.K.).

Conflict of Interest: none declared.

REFERENCES

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Blair,P.L. and Carucci,D.J. (2005) Functional proteome and expression analysis of sporozoites and hepatic stages of malaria development. *Curr. Top. Microbiol. Immunol.*, **295**, 417–438.

Brejova,B. *et al.* (2005) ExonHunter: a comprehensive approach to gene finding. *Bioinformatics*, **21**(Suppl. 1), i57–i65.

Carlton,J.M. *et al.* (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii. Nature*, **419**, 512–519.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spect.*, **5**, 976–989.

Gardner,M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature*, **419**, 498–511.

Hall,N. *et al.* (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, **307**, 82–86.

Kaiser,K. *et al.* (2004) Differential transcriptome profiling identifies *Plasmodium* genes encoding pre-erythrocytic stage-specific proteins. *Mol. Microbiol.*, **51**, 1221–1232.

Kariu,T. *et al.* (2006) CelTOS, a novel malarial protein that mediates transmission to mosquito and vertebrate hosts. *Mol. Microbiol.*, **59**, 1369–1379.

Keller,A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

Kooij,T.W. *et al.* (2005) A *Plasmodium* whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathogens*, **1**, e44.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Kumar,K.A. *et al.* (2006) The circumsporozoite protein is an immunodominant protective antigen in irradiated sporozoites. *Nature*, **444**, 937–940.

Li,L. *et al.* (2003) Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res.*, **13**, 443–454.

Lu,F. *et al.* (2007) cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics*, **8**, 255.

Luke,T.C. and Hoffman,S.L. (2003) Rationale and plans for developing a non-replicating, metabolically active, radiation-attenuated *Plasmodium falciparum* sporozoite vaccine. *J. Experi. Biol.*, **206**, 3803–3808.

Mulder,N.J. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.

Pertea,G. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Rice,P. *et al.* (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Roy,S.W. and Hartl,D.L. (2006) Very little intron loss/gain in *Plasmodium*: intron loss/gain mutation rates and intron number. *Genome Res.*, **16**, 750–756.

Sacci,J.B. Jr. *et al.* (2005) Transcriptional analysis of in vivo *Plasmodium yoelii* liver stage gene expression. *Mol. Biochem. Parasitol.*, **142**, 177–183.

Salzberg,S.L. (2007) Genome re-annotation: a WIKI solution?, *Genome Biol.*, **8**, 102.

Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Snow,R.W. *et al.* (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, **434**, 214–217.

Tarun,A.S. *et al.* (2006) Quantitative isolation and in vivo imaging of malaria parasite liver stages. *Int. J. Parasitol.*, **36**, 1283–1293.

Tarun,A.S. *et al.* (2008) A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc. Natl Acad. Sci. USA*, **105**, 305–310.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Wang,Q. *et al.* (2004) Transcriptome of axenic liver stages of *Plasmodium yoelii*. *Mol. Biochem. Parasitol.*, **137**, 161–168.