Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Molecular & Biochemical Parasitology 151 (2007) 100–110
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Genome-scale protein expression and structural biology of *Plasmodium falciparum* and related Apicomplexan organisms

Masoud Vedadi [a], Jocelyne Lew [a], Jennifer Artz [a], Mehrnaz Amani [a], Yong Zhao [a], Aiping Dong [a], Gregory A. Wasney [a], Mian Gao [a], Tanya Hills [a], Stephen Brokx [a], Wei Qiu [a], Sujata Sharma [a], Angelina Diassiti [b], Zahoor Alam [a], Michelle Melone [a], Anne Mulichak [c], Amy Wernimont [a], James Bray [d], Peter Loppnau [a], Olga Plotnikova [a], Kate Newberry [a], Emayavaram Sundararajan [a], Simon Houston [a], John Walker [a], Wolfram Tempel [a], Alexey Bochkarev [a], Ivona Kozieradzki [a], Aled Edwards [a], Cheryl Arrowsmith [a], David Roos [e], Kevin Kain [a], Raymond Hui [a,*]

[a] *Structural Genomics Consortium, U. of Toronto, 100 College St. Rm 522B, Toronto, Ont., Canada M5G 1L5*
[b] *McLaughlin-Rotman Centre for Global Health, Toronto General Hospital, Eaton Wing Ground Floor, Room 224, 200 Elizabeth Street, Toronto, Ont., Canada M5G 2C4*
[c] *APS, Argonne National Lab Building 435A, Sector 17, 9700 South Cass Avenue Argonne, IL 60439, United States*
[d] *Structural Genomics Consortium (SGC), U. of Oxford, Oxford, United Kingdom*
[e] *Department of Biology, U. of Pennsylvania, 301 Goddard Laboratories, Philadelphia, PA 19104, United States*

Received 26 April 2006; received in revised form 19 October 2006; accepted 20 October 2006
Available online 13 November 2006

Abstract

Parasites from the protozoan phylum *Apicomplexa* are responsible for diseases, such as malaria, toxoplasmosis and cryptosporidiosis, all of which have significantly higher rates of mortality and morbidity in economically underdeveloped regions of the world. Advances in vaccine development and drug discovery are urgently needed to control these diseases and can be facilitated by production of purified recombinant proteins from Apicomplexan genomes and determination of their 3D structures. To date, both heterologous expression and crystallization of Apicomplexan proteins have seen only limited success. In an effort to explore the effectiveness of producing and crystallizing proteins on a genome-scale using a standardized methodology, over 400 distinct *Plasmodium falciparum* target genes were chosen representing different cellular classes, along with select orthologues from four other *Plasmodium* species as well as *Cryptosporidium parvum* and *Toxoplasma gondii*. From a total of 1008 genes from the seven genomes, 304 (30.2%) produced purified soluble proteins and 97 (9.6%) crystallized, culminating in 36 crystal structures. These results demonstrate that, contrary to previous findings, a standardized platform using *Escherichia coli* can be effective for genome-scale production and crystallography of Apicomplexan proteins. Predictably, orthologous proteins from different Apicomplexan genomes behaved differently in expression, purification and crystallization, although the overall success rates of *Plasmodium* orthologues do not differ significantly. Their differences were effectively exploited to elevate the overall productivity to levels comparable to the most successful ongoing structural genomics projects: 229 of the 468 target genes produced purified soluble protein from one or more organisms, with 80 and 32 of the purified targets, respectively, leading to crystals and ultimately structures from one or more orthologues.

Keywords: Structural genomics; Heterologous protein expression; Apicomplexa; Malaria; Crystallography; Crystallization

1. Introduction

The *Apicomplexa* is a protozoan phylum of obligate intracellular parasites characterized by an apical complex, where organelles, such as micronemes and rhoptries reside. While *Eimeria*, *Neospora*, *Babesia* and *Theileria* are causes of mainly veterinary health concerns, organisms from other apicomplexan

Abbreviations: Pf, *Plasmodium falciparum*; Py, *Plasmodium yoelii*; Pb, *Plasmodium berghei*; Pv, *Plasmodium vivax*; Pk, *Plasmodium knowlesi*; Cp, *Cryptosporidium parvum*; Tg, *Toxoplasma gondii*

* Corresponding author. Tel.: +1 416 946 7182; fax: +1 416 946 0588.
E-mail address: raymond.hui@utoronto.ca (R. Hui).

0166-6851/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2006.10.011

genera, such as *Plasmodium*, *Cryptosporidium* and *Toxoplasma* are responsible for high rates of morbidity and mortality in humans, particularly in economically underdeveloped regions of the world. *Plasmodium* parasites alone are annually responsible for over 300 million cases of human malaria, resulting in up to 3 million deaths [1]. Cryptosporidiosis and toxoplasmosis – mediated, respectively, by *Cryptosporidium parvum* and *Toxoplasma gondii* – are opportunistic infections and major causes of morbidity and mortality amongst immuno-compromised patients, particularly in those infected with HIV. Co-infection with HIV and malaria is also particularly common in Africa, resulting in severe malaria anemia [2]. Simply put, Apicomplexan parasites are ravaging the parts of the world without the infrastructure to control them.

Apicomplexan diseases lack effective treatment. While there is currently no cure for toxoplasmosis and cryptosporidiosis, various anti-malarial drugs exist; however, economic, geopolitical and scientific factors have conspired to allow malaria to re-emerge in the last decade as the leading global cause of child mortality. Scientifically, the challenge is manifold: the paucity of validated drug targets, complexity of Apicomplexan life cycles as well as highly adaptable gene expression mechanisms leading to a unique ability to develop drug resistance and evade immune response.

The urgently needed advances in vaccine development and drug discovery for Apicomplexan diseases can be significantly facilitated by genome-scale production of purified recombinant proteins from *Apicomplexa* and determination of their 3D structures. This structural genomics approach has already been proven effective for a number of genomes. For example, 70% of 424 targets were expressed as soluble proteins in a project on the thermophilic archaeon *Methanobacterium thermoautotrophicum* [3,4]. Of these, 47% and 40%, respectively, of small (<20 kDa) and large proteins yielded 2 mg or more of soluble protein. In addition, 19% and 9%, respectively, of small and large proteins crystallized. A total of 36 structures (8.4%) were derived from these crystals or NMR samples. Similarly, application of structural genomics to 1376 *Thermotoga maritima* clones produced 542 purified targets, 432 crystallized proteins and 24 unique structures [5,6]. Eukaryotic genomes are also being tackled. From a set of 250 *Saccharomyces cerevisiae* proteins, 88% were found to be expressed, with 60 out of 250 or 24% yielding sufficient soluble protein for crystal trial [7]. This culminated in 22 crystallized proteins (8.8%) and 14 structures (5.6%). The *Arabidopsis thaliana* genome has also yielded 496 pure recombinant proteins out of 632 targets, with the MBP fusion tag used to aid solubilization [8]. These projects focused primarily on non-membrane proteins. In a high throughput membrane protein project involving 280 proteins from *Escherichia coli* and *T. maritima*, 30% of the cloned proteins expressed in *E. coli*, leading to 22 pure proteins (7.9%), 2 crystals (0.7%) and 1 structure [9]. Clearly, genome-scale protein expression and structural biology have been successfully implemented in both prokaryotic and eukaryotic organisms.

In contrast, there is no report of successful large-scale protein production or structural biology in any Apicomplexan genomes. Instead, the relevant literature centers around problems of obtaining purified *Plasmodium* proteins, including codon mismatch [10,11] and toxicity of plasmodium proteins [12]. There are also reports of isolated successes using custom techniques, such as specialized expression vectors, codon optimization and refolding [11–13], which may not be effective for a majority of proteins. On a larger scale, two independent pilot projects on expression have, respectively, yielded 13 purified soluble proteins from 368 *Pf* genes [14] and 9 purified proteins from 95 *Pf* genes [10]—a level of success lower even than that achieved with membrane proteins. Most recently and most significantly, Mehlin et al. reported [15] successful expression of soluble proteins from only 63 of 1000 open reading frames from *Plasmodium falciparum*. This study spanned various classes of proteins and yielded instructive findings from comprehensive statistical analysis: (a) smaller proteins, proteins with p*I* lower than 6 and those with relatively higher homology to *E. coli* homologues are more likely to express in soluble form in *E. coli*; (b) codon usage and AT-contents do not affect expression. The general conclusion from previous works is that *E. coli* is not an effective expression host for *Plasmodium* proteins.

Problems in expressing Apicomplexan proteins could be presaged from genomic analysis. The *Pf* genome is the most AT-rich of all genomes sequenced to date [1], at 80%, with the *Py* [16] and *Cp* [17] genomes not far behind at 78% and 70%, respectively. The unusually high AT bias translates into some codons rarely required by *E. coli* proteins, e.g. AGA and AGG [18], highlighting an intrinsic problem with heterologous expression. In addition, *Pf* genes have a mean length of 2.3 kb—1 kb longer than homologues from other organisms [1], with the extra length often featuring unique inserts. Many *Plasmodium* proteins have low complexity regions consisting of long hydrophobic stretches, amino acid repeats or segments highly rich in amino acids encoded by AT-biased codons, notably lysine and asparagine. Furthermore, 10% of the *Pf* proteins are targeted to the apicoplast, with another 10% predicted to be secreted [19,20] to the host. The N-termini of these proteins typically contain sequence motifs regulating their localization. These peptide regions influence localization rather than function but are membrane-like in their effects on expression and folding. Codon bias, size, sequence inserts, signal and transit peptides are all features that can be predicted to affect not only recombinant protein expression, but also crystallization.

Predictably, limited success in recombinant protein expression has translated into a relatively low number of protein structures to date. As of December 31st, 2005, the Protein Data Bank (http://www.pdb.org) included 78 unique structures (sharing less than 90% in sequence identity) from all *Plasmodium* species, 8 from *T. gondii* and 6 from *Cryptosporidium hominis* and *C. parvum* combined. From 2001 to 2004, the number of novel *Plasmodium* structures deposited in the PDB was 6, 8, 10 and 14. The number jumped to 24 in 2005, spurred in part by the release of genomic data.

While the *Py* and *Cp* genomes are close to *Pf* in AT-contents, they differ in other respects. Both *Py* and *Cp* genes are on average shorter [16,17], with fewer proteins exported. Furthermore, other Apicomplexan genomes, including *Plasmodium vivax*, *Plasmodium knowlesi* and *T. gondii* are more balanced in their

AT-GC quotient, while maintaining a high level of sequence identity and similarity between orthologous proteins. This offers an opportunity to overcome the difficulties associated with the expression, purification and/or crystallization of *Pf* proteins by exploiting the different behaviors of orthologous proteins from other Apicomplexan organisms. Orthologues have been successfully utilized previously in structural genomics of *E. coli* and *T. maritima* [21] to produce a higher number of proteins and crystals than obtainable with either genome alone. With *Py* and *Pb* both furnishing over 3000 orthologous genes [16,22] to the *Pf* genome and at least another 1800 orthologues from *Cp* [17], the Apicomplexan genomes are well suited for this approach.

Here, we report a successful application of structural genomics to five *Plasmodium* species, *P. falciparum* (human parasite), *P. vivax* (human parasite), *Plasmodium yoelii* (rodent parasite), *P. knowlesi* (simian parasite) and *Plasmodium berghei* (rodent parasite), as well as *T. gondii* (human parasite and causative agent of toxoplasmosis) and *C. parvum* (human and animal parasite and causative agent of cryptosporidiosis). These results demonstrate that, in spite of challenges reported in the literature and summarized above, a standardized platform using *E. coli* can be effective for genome-scale production and crystallography of Apicomplexan proteins. As can be expected, orthologous proteins from different Apicomplexan genomes behaved differently in expression, purification and crystallization trials. These differences were effectively exploited to elevate the overall productivity to levels comparable to the most successful ongoing structural genomics projects, e.g. proteins not expressed from *P. falciparum* were expressed from one or more other Apicomplexan genomes. With 304 proteins successfully purified from the 7 genomes, at least 97 proteins crystallized and 36 distinct structures determined to date, these results establish the ground work for enabling post-genomic research in malaria and related parasitic diseases.

2. Methods

2.1. Cloning

Coding sequences were amplified by PCR in 96-well format from genomic DNA (all *Plasmodium* and *Cryptosporidium* proteins) or cDNA (*T. gondii* proteins and some of the *P. vivax* proteins) using Platinum Pfx from Invitrogen. Three alternative vectors, all based on the pET expression vector with His6 tag and T7-lacO promoter, and encoding proteolytic sites for either TEV or thrombin, were used. Descriptions for DNA and vectors used are provided as supplementary data.

2.2. Small-scale text expression

Small-scale protein expression was used to select the best candidates for scale-up in *E. coli*. Apicomplexan clones were transformed into *E. coli* BL21-CodonPlus (DE3)-RIL cells (Stratagene) in a 96-well deep wall plate. After incubation at 4 °C for 30 min and heat shock at 42 °C (25 s), 900 μL of SOC medium was added and cells were incubated at 37 °C with shaking for 1 h. An aliquot of 20 μL of this cell culture was used directly to inoculate 500 μL LB (with 50 μg/mL kanamycin and 34 μg/mL chloramphenicol added) in a fresh 96-well plate, followed by incubation overnight at 37 °C. Subsequently, 20 μL of this culture was transferred to 2 mL TB medium supplemented with antibiotics and trace elements (8.3 mM $MgSO_4$, 6.3 μM $CoCl_2$, 47.3 μM $MnSO_4$, 8.1 μM H_3BO_3, 8.3 μM Na_2MoO_4, 7.0 μM $ZnSO_4$, 108 μM $FeSO_4$, 68 μM $CaCl_2$, 4.1 μM $AlCl_3$, 4.2 mM $NiCl_2$ and 5.9 μM $CuCl_2$) in four 24-well blocks (Qiagen). This culture was incubated at 37 °C with shaking until OD_{600} reached ~1.5, at which point the temperature of the culture was lowered to 15 °C and isopropyl-β-D-thiogalactopyranoside (IPTG) was added to 1 mM. Following overnight incubation with shaking at 15 °C, the plates were centrifuged at 1450 × *g*. The resulting cell pellets were incubated at −80 °C for 10 min. After adding 250 μL of suspension buffer consisting of *Binding Buffer* (50 mM HEPES pH 7.5, 0.5 M NaCl, 5 mM imidazole and 5%, v/v, glycerol) mixed with protease inhibitors (1 mM phenylmethanesulfonyl fluoride (PMSF),[1] 1 mM benzamidine), the cells were suspended by shaking at 800 rpm for 10 min at room temperature. An aliquot of 200 μL of this suspension was transferred to 1 mL suspension buffer, with the addition of 0.5% (w/v) CHAPS and 50 U benzonase (Sigma), in a 96-well plate, followed by shaking at 800 rpm for 30 min at room temperature. This lysed cell suspension was centrifuged at 2100 × *g* for 5 min, and 40 μL of lysate was reserved for SDS-PAGE analysis while 800 μL of the supernatant was transferred to a fresh 96-well block. To this lysate, 150 μL (30 μL bed volume) of Ni-NTA Superflow resin (Qiagen), equilibrated in suspension buffer, was added. After shaking at 800 rpm for 30 min at room temperature, the plate was centrifuged at 230 × *g* for 5 min to pellet the nickel resin. Most of the supernatant was removed and the resin was suspended in the remaining ~200 μL of supernatant. The suspension was transferred to a 96-well AcroPrep filter plate (Pall Corp.) and centrifuged at 930 × *g* for 5 min. The resin, adhering to the filter plate, was then washed three times with 220 μL *Binding Buffer* (but with 30 mM imidazole), each time followed by centrifugation at 930 × *g* for 5 min. The histidine-tagged protein was eluted from the resin by addition of 80 μL of the same buffer containing 500 mM imidazole, followed by centrifugation and collection of the eluate. Whole cell lysate and eluted purified protein samples were analyzed by SDS-PAGE using 26-well Criterion gels (Bio-Rad), and the gels were stained with SimplyBlue reagent (Invitrogen) to determine which proteins were expressed and which of the expressed proteins were visibly soluble.

2.3. "Large Scale" expression

Based on test expression results, soluble clones were selected for expression in 2 or 4 L. A single colony from each transformation was inoculated in 100 mL LB medium with antibiotics (50 μg/mL kanamycin and 25 μg/mL chloramphenicol) in a 250 mL baffled flask overnight at 37 °C. Fifty millilitres of this

[1] PMSF was not added in cases where the targeted protein was a serine protease.

culture was added to 1.8 L of TB medium with kanamycin (50 μg/mL) and antifoam (200 μL) in a 2 L bottle arranged in LEX bubbling system from Harbinger Biotech. The LEX system consists of an enclosure capable of housing 24 × 2 or 48 × 1 L round media bottles. Each bottle was connected to an air manifold via a quick disconnect and a manual flow regulator. In turn, the manifold was fed by a centralized building air compressor. Typically, the air flow into each bottle was adjusted to 4–6 L/min, sufficient for oxygenation and mixing of cultures.

Cultures were left to grow in LEX at 37 °C until OD_{600} reached at least 5 (typically 5 or above after 3–4 h). The temperature was quenched to 15 °C by a combination of ice and dropping the water bath temperature. Once 15 °C was attained, 600 μL 1 M IPTG was added for induction. The next morning, the pellets were harvested by centrifugation at 4000 rpm for 15 min at 4 °C. Pellets were re-suspended to approximately 40 mL/L of cell culture in *Binding Buffer* with protease inhibitors, flash frozen in liquid nitrogen and stored at −80 °C.

In some cases, incorporation of seleno-methionine was necessary. The "Se-Met" medium was prepared by combination of 100 mL 10× M9, 1 mL 1M MgSO4, 10 mL 40% glucose, 100 μL 0.5% thiamine, 300 μL 12.5 mg/mL $FeSO_4$, 300 μL 0.166% biotin, 1 mL 100 mg/mL ampicillin, 0.5 mL 100 mg/mL kanamycin. The final volume was adjusted to 1 L using sterile water. The amino acid mixture was prepared freshly before each use by mixing 100 mg L-lysine, 100 mg L-phenylalanine, 100 mg L-threonine, 50 mg L-isoleucine, 50 mg L-valine, 50 mg L-leucine and 50 mg L-selenomethionine in a total volume of 10 mL. After the amino acids were completely dissolved, the solution was filter-sterilized.

Overnight pre-culture of the transformed *E. coli* BL21 (DE3) in 1–2 mL LB medium with kanamycin (50 mg/mL) and chloramphenicol (25 mg/mL) were incubated in a glass tube at 30 °C. Next morning, the cells were collected by centrifugation (2500 rpm for 3–4 min at room temperature), re-suspended in 1–2 mL of M9 medium at 37 °C, transferred to 1 L "Se-Met" culture and shaken at 150–220 rpm at 37 °C. The OD_{600} of the culture reached 0.3 after 7–8 h. The temperature was decreased to 15 °C and 10 mL amino acid mixture was added to the culture. Fifteen minutes later, IPTG was added to a final concentration of 1 mM. The incubation was continued for another 15–16 h at 15 °C.

2.4. Purification

Proteins were purified typically in batches of 12 through lysis, affinity chromatography, gel filtration and concentration steps. The initial cell pellets were initially re-suspended to approximately 40 mL/L of cell culture in *Binding Buffer* with the addition of protease inhibitors. Re-suspended pellets were stored at −80 °C and thawed overnight at 4 °C on the day before purification. After pre-treating each pellet from 1 L of culture with 0.5% CHAPS and 500 units of benzonase for 40 min at room temperature, cells were mechanically lysed using a high shear fluid processor (Microfluidizer Processor, M-110EH) at 18,000 psi. The lysate was then clarified by centrifugation at ∼75,000 × *g* (24,000 rpm) for 20 min at 10 °C. The cleared lysate was loaded onto a column pre-packed with 10 g DE52 (Whatman) anion exchange resin (previously activated with 2.5 M NaCl and equilibrated with *Binding Buffer*) and subsequently onto a 1.0–2.5 mL Ni-NTA (Qiagen) column pre-equilibrated with *Binding Buffer* at approximately 1–1.5 mL/min. The presence of 0.5 M NaCl in the *Binding Buffer* limited any potential protein binding to the DE52 resin to a negligible level. The volume of the Ni-NTA resin was predetermined by the predicted protein yield from test expression analysis. After the lysate was loaded, the DE52 was further washed with 20 mL of *Binding Buffer*. Each Ni-NTA column was then washed with 200 mL of *Wash Buffer* (50 mM HEPES pH 7.5, 500 mM NaCl, 30 mM imidazole and 5% glycerol) at 2–2.5 mL/min. After washing, the protein was eluted with 15 mL of *Elution Buffer* (50 mM HEPES pH 7.5, 500 mM NaCl, 250 mM imidazole and 5% glycerol). EDTA was immediately added to the elution fraction to 1 mM; and DTT was added to 1–5 mM after approximately 15 more minutes. The purity of the eluted protein was immediately evaluated by SDS-PAGE gel; and the collected protein was desalted and/or further purified by gel filtration, cleavage of the His_6-tag or dialysis.

For gel filtration, an AKTA system (from GE Healthcare) was customized to run up to seven protein samples in series. Proteins judged to be more than 90% pure by SDS-PAGE were loaded onto a gel filtration column (either Hi-Load Superdex S200 26/60 or Hi-Load Superdex S75 26/60 from GE Healthcare depending on the protein size). The gel filtration column was pre-equilibrated with *Crystal Buffer* (10 mM HEPES pH 7.5, 500 mM NaCl). If the protein was less than 90% pure by SDS-PAGE gel, then the His_6-tag was cut with either thrombin or TEV protease (depending on the vector used) at 4 °C. The protein was dialyzed overnight and passed through another Ni-NTA column pre-equilibrated with 10 mM HEPES pH 7.5, 500 mM NaCl and 15 mM imidazole. Proteins with low extinction coefficients or insufficient yields to be purified by gel filtration were dialyzed overnight against *Crystal Buffer*. For proteins with yields greater than 30 mg and 90% pure (as judged by SDS-PAGE gel), half of the protein would be purified by gel filtration and half of the protein would be cut with thrombin or TEV.

All purified proteins (cut and uncut) were concentrated using Amicon spin concentrators (from Millipore) to a concentration determined by the Pre-Crystallization Test (Hampton Research) and put into crystal trial. The remaining protein was flash frozen in 100–300 μL aliquots in liquid N_2 and stored at −80 °C. Each protein sample was evaluated by: (i) mass spectrometry to confirm the molecular weight and identify post-translational modifications; (ii) thermo-stability analysis; (iii) SDS-PAGE gel to assess the final purity of the sample.

2.5. Crystallization

Initial crystal trials were performed using two custom SGC screens formulated from the commercial conditions reported [23] or found internally to be most effective. The conditions for

our primary screens can be found online (http://www.thesgc.com/SGC-WebPages/toronto-technology-crystallization.php). Protein samples were dispensed in sitting drops in 1:1 volume ratio on Intelliplates from Art Robbins Instruments or CrystalQuick™ plates from Greiner, ranging from 1 to 2 μL in total volume using the Mosquito liquid handling robot (TTP LabTech), typically with 100 μL of mother liquor added using the Beckman Biomek FX robot (Beckman Coulter). Crystal plates were stored in 18–20 °C in a temperature-controlled room.

2.6. Data collection

The crystals were tested and data collected using the FR-E SuperBright X-ray generator from Rigaku, equipped with two imaging plate detectors—one R-AXIS IV and one R-AXIS HTC. For higher resolution and for selenomethionine crystals, data were collected at the APS (http://www.aps.anl.gov/) and CHESS (http://www.chess.cornell.edu/) synchrotrons.

2.7. Structure determination

For cases with pre-existing protein structures with sufficient homology, we used the FFAS03 server (http://ffas.burnham.org) in order to generate molecular replacement models. For model building and refinement, we used the automated model building programs *RESOLVE* and/or *ARP/wARP* while *O*, *refmac*, *Coot* and *CNS* were used for manual data refinement.

3. Results and discussion

3.1. Project design and target selection

With significant input from experts specializing in research in malaria research and Apicomplexa, a set of 468 *Plasmodium* genes, henceforth referred to as the *reference targets*, were chosen for this project. This list includes putative drug and vaccine targets involved in hemoglobin metabolism and biosynthesis, exported proteins, proteins without significant homology to proteins from other genomes and consequently no current functional annotation, as well as apicoplast-targeted proteins. With *P. falciparum* as the model genome, select orthologous genes from *P. vivax, P. yoelii, P. berghei, P. knowlesi, T. gondii* and *C. parvum* were then identified based on an inclusion threshold of ≥25% in sequence identity and BLAST [24] *E*-value $\leq 1 \times 10^{-20}$. A total of 1008 Apicomplexan genes were included in the results discussed herein.

The coding sequences for the selected genes were truncated at one or both termini, where necessary, to remove transmembrane domains, signal and transit peptides (in cases of apicoplast proteins) and export motifs (in cases of proteins predicted to be exported by the presence of the *Pexel* motif [19,20]). In some cases, multiple terminally truncated constructs were cloned to increase the probability of successful expression and crystallization, as well as to mitigate the uncertainty of boundaries of the functional domains.

3.2. Overall results—E. coli can be an effective expression host for Apicomplexan proteins

As of July 31st, 2852 expression constructs representing the 1008 genes from *Pf, Pv, Py, Pb, Pk, Tg* and *Cp* have been successfully cloned with an N-terminal hexa-histidine fusion tag. Genomic DNA (all *Plasmodium* proteins and *Cp*) and cDNA (for all *Tg* proteins), generously provided by various members of the scientific community, were used as the templates (see supporting material for DNA templates and their biological sources). The clones were transformed into either BL21-CodonPlus® (DE3)-RIL[2] from Stratagene or BL21 (DE3) R3.[3]

Using a small-scale expression screening system, 2086 or 73.1% of the expression clones were selected for large-scale expression in 2 L or more of TB culture and purified by affinity chromatography and gel filtration. Purified soluble protein with yield of 2 mg or more per liter of culture[4] were obtained for 652 protein constructs or 22.9% of the initial set of 2852 expression clones. Crystals emerged from 97 distinct purified apicomplexan proteins, culminating in 36 crystal structures, including 9 from *P. falciparum*, 11 from *P. yoelii*, 7 from *C. parvum*, 4 from *P. vivax,* 2 each from *P. knowlesi* and *T. gondii* and 1 from *P. berghei*. The structures are listed in Table 1 while the overall statistical summary is provided in Table 2. In addition, these structures are shown in Figs. 1–4, with detailed description of each provided at the SGC's malaria portal (http://www.thesgc.com/malaria/), including functional descriptions, structural analysis and methods used for protein production and crystallography.

The 652 purified soluble protein samples represented 304 of the starting 1008 Apicomplexan genes. Furthermore, the 97 crystallized protein constructs were derived from 80 distinct Apicomplexan genes. In other words, 30.2% and 9.6% of the Apicomplexan genes yielded, respectively, purified soluble proteins and protein crystals. In addition, the total of 36 structures represents 3.6% of the starting total number of Apicomplexan genes. Overall, this level of effectiveness compares favourably against other structural genomics projects and clearly demonstrates that the use of a codon-enhanced but commercially available strain of *E. coli* can form an effective platform for genome-scale production and crystallization of Apicomplexan proteins.

3.3. The power of orthologues

In total, 229 or 49% of the 468 reference genes produced purified soluble proteins from one or more of the seven Apicomplexan genomes studied in this project, while 80% or 17% produced one or more crystals from *Pf, Py, Pv, Pb, Pk, Cp* or *Tg*. By regarding all seven genomes as a super genome and

[2] This is a strain of *E. coli* enriched with extra copies of the tRNA genes *argU* (AGA and AGG), *ileY* (AUA) and *leuW* (CUA).

[3] This is a custom strain engineered by transforming the Rosetta2 plasmid into BL21 (DE3) cells selected for phage resistance (available from the SGC) with extra copies of AGA, AGG, AUA, CUA, GGA, CCC and CGG.

[4] The size and purity of the purified proteins were verified by means of SDS-PAGE gel or mass spectrometry.

Table 1
List of structures

Gene ID	Description	PDB
cgd3_300	*Cp* nuclear transport factor 2	1ZO2
cgd5_440	*Cp* cyclase associated protein	2B0R
cgd6_3850	*Cp* high mobility protein NHP2	2AIF
cgd7_4580	*Cp* small nuclear ribonuclear protein LSm5	2FWK
cgd4_2550	*Cp* farnesyl pyrophosphate synthase	2HER
cgd7_470	*Cp* malate dehydrogenase	2HJR
cgd7_2060	*Cp* vacuolar protein sorting 29	2A22
XP_679107	*Pb* orotidine monophosphate decarboxylase (OMPDC)	2FDS
PF14_0417	*Pf* heat shock protein	1Y6Z
PFL0660w	*Pf* dynein light chain 1	1YO3
PFI1420w	*Pf* guanylate kinase	1Z6G
PFC0310c	*Pf* ClpP protease	2F6I
PFL2275c	*Pf* FKBP, TPR domain	2FBN
PFE0505w	*Pf* cyclophilin	2FU0
PF14_0156	*Pf* dimethyladenosine transferase	2H1R
MAL13P1.227	*Pf* ubiquitin conjugating enzyme E2	2H2Y
PF11_0301	*Pf* spermidine synthase	2HTE
PK5_1010c	*Pk* translationally controlled tumour protein	1TXJ
PK8_1460w	*Pk* Fe-superoxide dismutase	2AWP
PY04285	*Py* 1-cys peroxiredoxin	1XCC
PY00104	*Py* ornithine aminotransferase	1Z7D
PY00382	*Py* cyclophilin	1Z81
PY02076	*Py* adenosine deaminase	2AMX
PY02252	*Py* deoxiribose phosphate aldolase	2A4A
PY01515	*Py* OMPDC	2AQW
PY07357	*Py* thioredoxin-like protein 4A	2AV4
PY00693	*Py* cyclophilin	2B71
PY06285	*Py* holo-ACP synthase	2BDD
PY06285	*Py* multidrug resistance 2, nucleotide binding domain	2GHI
PY00414	*Py* 2-cys peroxiredoxin	2H01
Pv111555	*Pv*-OMPDC	2FFC
Pv083175	*Pv* ubiquitin conjugating enzyme E2	2FO3
Pv118545	*Pv* 2cys-peroxiredoxin	2H66
Pv003765	*Pv* adenylosuccinate lyase	2HVG
TgTwinScan_2721	*Tg* ubiquitin conjugating enzyme	2F4Z
TgTwinScan_2218	*Tg* ubiquitin conjugating enzyme UBE2E2	2AYV

Accessions for *Pf, Pv, Pb* and *Py* genes are based on www.plasmodb.org identification scheme. Accessions for *Cp* and *Tg* genes are based on www.cryptodb.org and www.toxodb.org, respectively. Accessions for the *Pk* genes are based on the most recent data released by the Sanger (http://www.sanger.ac.uk/Projects/P_knowlesi/). The accession for the *Pb* gene is based on NCBI.

orthologous genes as alternate expression constructs, the use of orthologues produced more proteins, crystals and structures than in any specific genome. The effect of orthologues is further magnified by isolating 209 targets[5] for which there were at least two orthologues cloned. As shown in Table 2, this group (gene identities provided as supplementary data) includes 182, 158, 104, 90, 151 and 69 proteins from *P. falciparum, P. yoelii, P. vivax, P. berghei, P. knowlesi* and *C. parvum*, respectively, a total

Table 2
Overall project statistical summary

	Total no.	Percent of clones
Apicomplexan genes from all organisms	1008	
Purified soluble Apicomplexan proteins	304	30.2% of all cloned orthologues
Crystallized Apicomplexan proteins	97	9.6%
Apicomplexan genes with structures	36	3.6%
Reference target genes	468	
Reference genes with purified soluble protein from *Pf* or at least one orthologue	229	48.9% of reference genes
Reference genes with crystals from *Pf* or at least one orthologue	80	17.1%
Reference genes with structures	32	6.8%

of 755 Apicomplexan proteins.[6] *T. gondii* proteins have been excluded because far fewer *Tg* genes were cloned in comparison to other genomes.

Clones, proteins, crystals and structures were derived from all six genomes, albeit not in equal numbers. The combined output of all genomes (counting each output as only one in cases where a reference target was, for example, crystallized in more than one genome) yielded at least twice the percentage of purified proteins (46.2% versus 20.4%), crystals (21.0% versus 10.5%) and structures (11.4% versus 3.4%) than obtained from *Pf* alone. Moreover, 52 of 92 soluble reference targets yielded purified protein from one genome only. A similar trend is more emphatic in the crystallization data: only 12 targets of 44, which crystallized yielded crystals from more than 1 genome. Clearly, without employment of orthologues, not only would there have been fewer purified proteins, crystals and structures, there would be no protein, crystal or structure for more than half of the targets.

3.4. Protein size and pI are weak predictors of expression and crystallization

It is generally expected that larger proteins are more challenging for a heterologous expression system. The protein constructs in this project ranged in size as follows: <20 kDa (25.2%), between 20 and 30 kDa (22.3%), between 30 and 40 kDa (19.3%), between 40 and 50 kDa (15.6%) and heavier than 50 kDa (17.6%). As shown in Fig. 5, the percentage of soluble and crystallized proteins decreased with increasing size, although the drop off is not steep. There was similarly weak correlation between soluble expression, crystallization and p*I* (Fig. 6). Clearly, there are other factors that more directly affect expression of Apicomplexan proteins.

3.5. AT contents does not affect heterologous expression

AT-richness of the *P. falciparum* genome has been proposed as a major factor for the resistance of its proteins to heterologous expression. Our results suggest that this factor may not be

[5] For the other reference targets, cloning from additional genomes has not yet begun at the time of analysis.

[6] The remaining Apicomplexan genes from the starting list of 1008 have been cloned only in one genome at the time of this analysis.

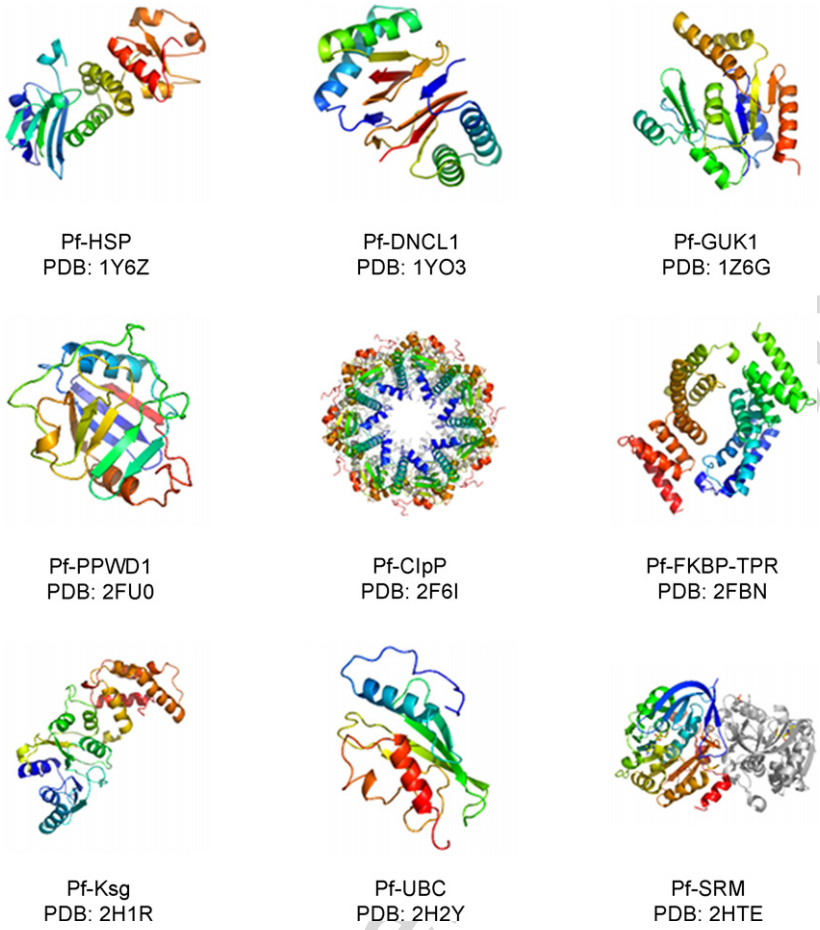

Fig. 1. *P. falciparum* structures from SGC.

as directly important as suspected. To wit, while the genomes of *P. falciparum, P. yoelii* and *P. berghei* are around 80% AT-rich, *P. vivax* and *P. knowlesi* are significantly lower in their AT contents. As shown in Table 3, all *Plasmodium* genomes expressed proteins at a rate between 20% and 25%. The lack of significant difference in the results is not surprising given that, regardless of AT coding, *Plasmodium* genomes feature low complexity regions and *Plasmodium* specific inserts that are directly problematic for heterologous expression of soluble protein.

Despite being 70% AT-rich, the *C. parvum* genome is more amenable to heterologous expression than all Plasmodium species. Overall, almost 50% of *Cp* clones yielded soluble protein. Within the set of 209, only 20 or 30% of the 69 *Cp* clones were found to be soluble, which is nevertheless noticeably better than the *Plasmodium* proteins. Furthermore, 16 of these 20 were expressed as full length proteins.

3.6. Full length versus truncated constructs

Of the 172 distinct Apicomplexan genes expressed, 117 were full length proteins, i.e. cloned and expressed from complete open reading frames. Nevertheless, the use of truncated

Table 3
Summary of orthologue statistics

	Pf	*Py*	*Pv*	*Pb*	*Pk*	*Cp*	Total	Unique reference targets
No. of proteins	182	159	104	90	151	69	755	209
No. of proteins expressed and purified	38 (20.9%)	38 (23.9%)	25 (24.0%)	18 (20.0%)	33 (21.9%)	20 (29.0%)	172 (22.8%)	97 (46.4%)
No. of proteins with crystals	20 (11.0%)	14 (8.8%)	11 (10.6%)	4 (4.4%)	6 (4.6%)	7 (10.1%)	63 (8.3%)	45 (21.5%)
No. of proteins with structures	8 (4.4%)	11 (6.9%)	4 (3.8%)	1 (1.1%)	2 (1.3%)	2 (2.9%)	28 (3.7%)	25 (12.0%)

This table includes only cases where a given reference target was successfully cloned in more than one of the six genomes. Only 19 targets were cloned from all 6 genomes; 29 were cloned from 5 genomes; 59 were cloned from 4 genomes; 53 were cloned from 3 genomes. *Cp* proteins expressed more readily than *Plasmodium* proteins. Only one protein was expressed from five genomes. Seven targets yielded soluble protein from four genomes, with at least one orthologue crystallizing in each case. *Pb* and *Pk* proteins crystallize less readily, but that could be influenced by the amount of effort applied, particularly after crystals or structures emerged from other orthologues. Proteins crystallized from two or more organisms in 12 cases. Overall, the use of orthologues produced well over twice the number of purified proteins, over three times the number of crystals and structures than obtained from *Pf* alone.

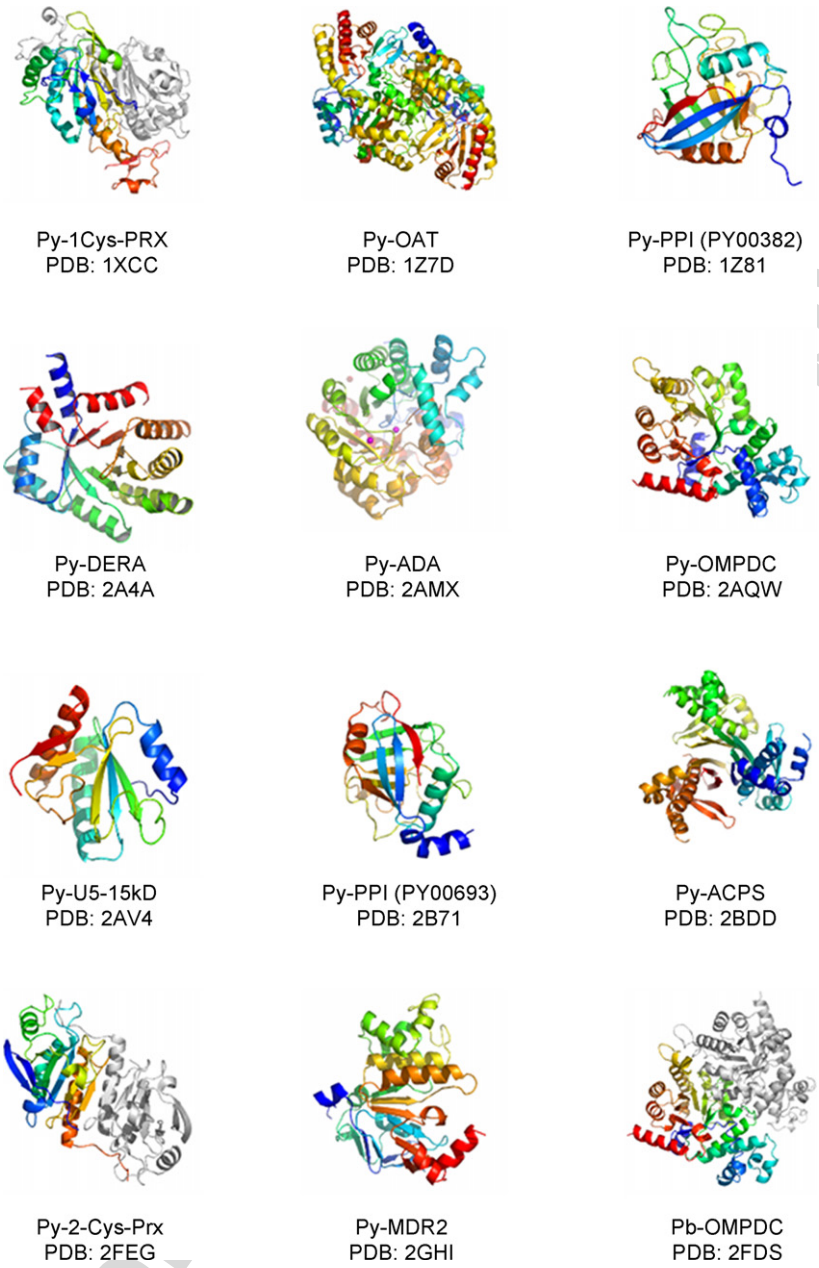

Fig. 2. *P. yoelii* and *P. berghei* structures from SGC.

constructs was instrumental in this project. Many proteins and crystals resulted from sequence truncation at either the N-terminal or the C-terminal or both (but there was no mid-sequence truncation). Specifically, removal of transmembrane regions as well as signal peptides, transit peptides and export motifs was essential to expression, not to mention crystallization. It is particularly notable that we did not obtain soluble expression of any full-length apicoplast-targeted proteins.

3.7. Overview of structures

On average, the structures resulting from this project to date are 48% identical in sequence to pre-existing PDB structures. Most of the structures align tightly with structures of homologous proteins from other organisms. The most notable *Plasmodium* specific feature appears in the structures of *Py, Pb* and *Pv* orotidine 5′-monophosphate decarboxylase (OMPDC). In addition to the TIM-barrel fold characterizing OMPDC structures from other organisms, there is an extra-helical fold outside the network of alternating alpha helices and beta sheets [25].

In addition to *Plasmodium* OMPDC, we also obtained structures from multiple *Plasmodium* orthologues for 2-cys-peroxiredoxin – *Py* (2H01) and *Pv* (2H66) and an E2 ubiquitin-conjugating enzyme – *Pf* (2H2Y) and *Pv* (2FO3). In all three cases, the overall structures and the active sites of the orthologues structures align tightly, confirming that orthologues are an efficient and practical tool in structural biology of *Plasmodium* parasites.

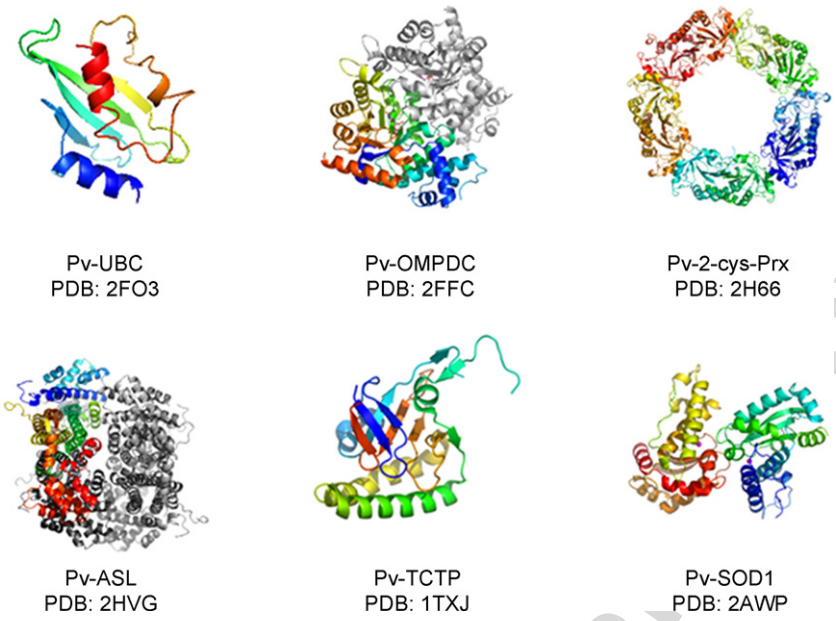

Fig. 3. *P. vivax* and *P. knowlesi* structures from SGC.

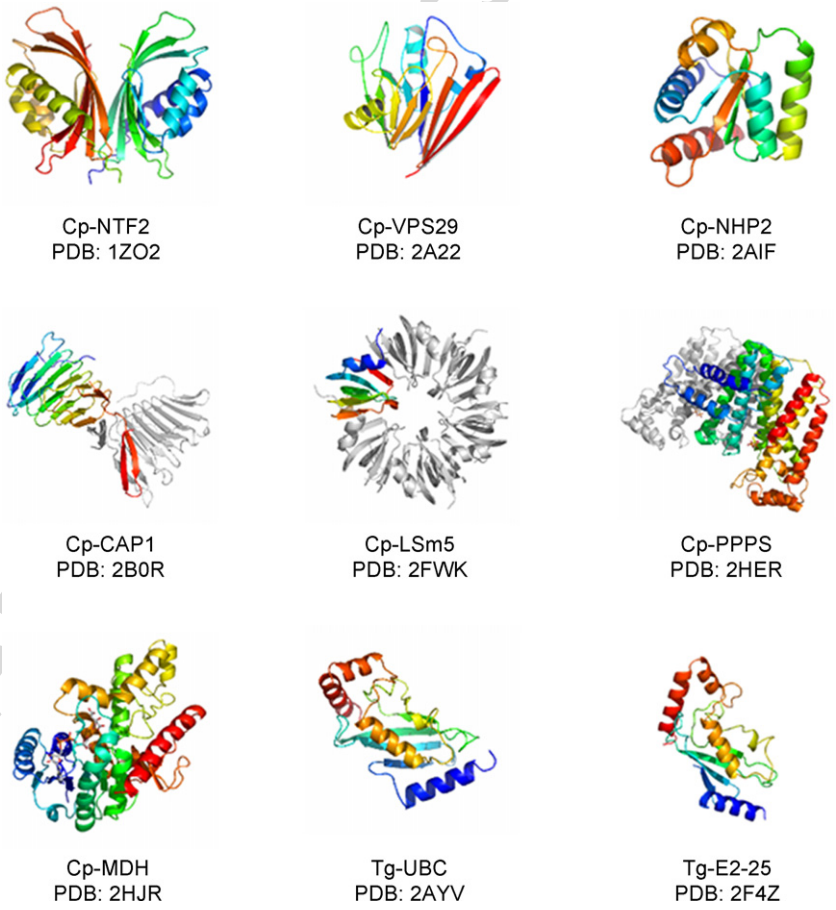

Fig. 4. *C. parvum* and *T. gondii* structures from SGC.

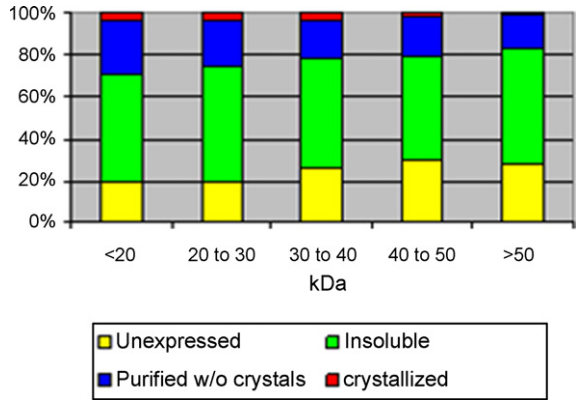

Fig. 5. Distribution of unexpressed, expressed, purified and crystallized proteins as a function of size. The total percentage of soluble proteins is the sum of the blue and red bars. There is a weak trend of decreasing percentage of soluble expression with increasing protein size.

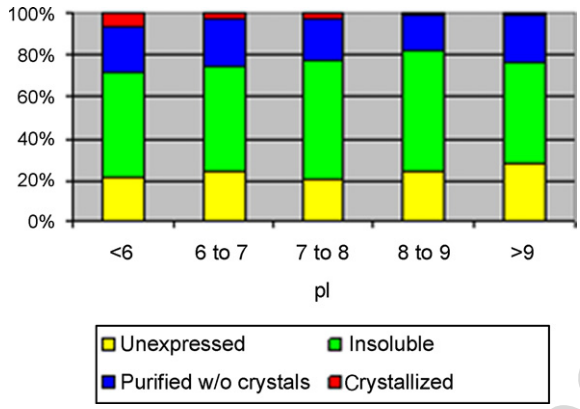

Fig. 6. Distribution of unexpressed, insoluble, purified and crystallized proteins as a function of p*I*. The total percentage of soluble proteins is the sum of blue and red bars, which shows a mild trend of decreasing with increasing p*I*.

4. Conclusions

Using codon enriched strains of *E. coli* as expression hosts, 304 out of 1008 Apicomplexan genes from *P. falciparum*, *P. vivax*, *P. yoelii*, *P. knowlesi*, *P. berghei*, *C. parvum* and *T. gondii* produced purified soluble proteins, with 97 of them yielding crystallized proteins. From these purified and crystallized proteins, 36 distinct structures were derived. Furthermore, by treating the seven genomes as a super Apicomplexan genome, we were able to obtain purified proteins from 229 of 468 reference target genes, along with 80 crystallized targets and 28 distinct structures.

These results demonstrate clearly that Apicomplexan proteins, including those from *Plasmodium*, *Cryptosporidum* and *Toxoplasma*, can be expressed, purified and crystallized using established high throughput methods. Specifically, it has been demonstrated that *E. coli* is indeed an effective heterologous expression platform for these proteins, even for *Plasmodium* genes with extreme AT-biases.

Although some *Plasmodium* genomes are not as AT-rich as *P. falciparum* and *P. yoelii*, their proteins express at roughly the same rate. This is in agreement with previous observations (Mehlin et al. [15]) that codon biases have less impact on heterologous expression than expected. On the other hand, proteins from *C. parvum* can be expressed in *E. coli* much more readily, likely due to noticeably lower frequency of low complexity regions and signal peptides.

Acknowledgements

The Structural Genomics Consortium is a public private charitable partnership that receives funds from Canada Foundation for Innovation, Canadian Institutes for Health Research, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, the Ontario Innovation Trust, the Ontario Research and Development Challenge Fund, the Wellcome Trust, the Knut and Alice Wallenberg Foundation, the Vinnova Swedish Agency for Innovation, the Swedish Foundation for Strategic Research and Karolinska Intitutet. This research was also supported by funds from the McLaughlin Centre of Molecular Medicine. We are very grateful to a number of experts in malaria and *Apicomplexa* who generously contributed their time and their invaluable insights into selection of organisms and targets—C. Newbold (University of Oxford), A. Fairlamb (University of Dundee), A. Holder (MRC, Mill Hill), D. Carucci (Foundation of US National Institutes of Health), D. Sibley (U. of Washington at St. Louis), P. Keeling (U. of British Columbia) and K. Haldar (Northwestern University). In addition, the following researchers generously donated genomic DNA and/or cDNA which were indispensable in enabling this project—D. Sibley, A. Waters (Leiden University Medical Center), J. Barnwell (US Center for Disease Control), J. Carlton (TIGR), J. Aguiar (US Naval Medical Research Center) and L. Cui (Pennsylvania State University). The Rosetta-R3 strain of *E. coli* used in expressing some of our proteins came from O. Gileadi of SGC at University of Oxford. We also thank S. Dhe-Paganon, F. Marino, R. Yeung, C. Jewell, S. Nalli and K. Yamazaki of the SGC for their roles in the development of the LEX expression system, which was used to grow most of the *E. coli* cultures for this project. Finally, the formulation for 96 of our crystallization conditions was developed by Alexei Savchenko and his Structural Proteomics Group at the University of Toronto.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molbiopara.2006.10.011.

References

[1] Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002;419:498–511.
[2] Otieno RO, Ouma C, Ong'echa JM, et al. Increased severe anemia in hiv-1-exposed and hiv-1-positive infants and children during acute malaria. Aids 2006;20:275–80.
[3] Christendat D, Yee A, Dharamsi A, et al. Structural proteomics: prospects for high throughput sample preparation. Prog Biophys Mol Biol 2000;73:339–45.

[4] Yee A, Pardee K, Christendat D, Savchenko A, Edwards AM, Arrowsmith CH. Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. Acc Chem Res 2003;36:183–9.
[5] Lesley SA, Kuhn P, Godzik A, et al. Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. Proc Natl Acad Sci USA 2002;99:11664–9.
[6] Lesley SA, Wilson IA. Protein production and crystallization at the joint center for structural genomics. J Struct Funct Genomics 2005;6:71–9.
[7] Quevillon-Cheruel S, Liger D, Leulliot N, et al. The paris-sud yeast structural genomics pilot-project: from structure to function. Biochimie 2004;86:617–23.
[8] Jeon WB, Aceti DJ, Bingman CA, et al. High-throughput purification and quality assurance of *Arabidopsis thaliana* proteins for eukaryotic structural genomics. J Struct Funct Genomics 2005;6:143–7.
[9] Dobrovetsky E, Lu ML, Andorn-Broza R, et al. High-throughput production of prokaryotic membrane proteins. J Struct Funct Genomics 2005;6:33–50.
[10] Aguiar JC, LaBaer J, Blair PL, et al. High-throughput generation of *P. falciparum* functional molecules by recombinational cloning. Genome Res 2004;14:2076–82.
[11] Baca AM, Hol WG. Overcoming codon bias: a method for high-level overexpression of *Plasmodium* and other AT-rich parasite genes in *Escherichia coli*. Int J Parasitol 2000;30:113–8.
[12] Cinquin O, Christopherson RI, Menz RI. A hybrid plasmid for expression of toxic malarial proteins in *Escherichia coli*. Mol Biochem Parasitol 2001;117:245–7.
[13] Sijwali PS, Brinen LS, Rosenthal PJ. Systematic optimization of expression and refolding of the *Plasmodium falciparum* cysteine protease falcipain-2. Protein Expr Purif 2001;22:128–34.
[14] Mehlin C. Structure-based drug discovery for *Plasmodium falciparum*. Comb Chem High Throughput Screen 2005;8:5–14.
[15] Mehlin C, Boni E, Buckner FS, et al. Heterologous expression of proteins from *Plasmodium falciparum*: results from 1000 genes. Mol Biochem Parasitol 2006;148:144–60.
[16] Carlton JM, Angiuoli SV, Suh BB, et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature 2002;419:512–9.
[17] Abrahamsen MS, Templeton TJ, Enomoto S, et al. Complete genome sequence of the apicomplexan *Cryptosporidium parvum*. Science 2004;304:441–5.
[18] Sayers JR, Price HP, Fallon PG, Doenhoff MJ. Aga/agg codon usage in parasites: implications for gene expression in *Escherichia coli*. Parasitol Today 1995;11:345–6.
[19] Hiller NL, Bhattacharjee S, van Ooij C, et al. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 2004;306:1934–7.
[20] Marti M, Baum J, Rug M, Tilley L, Cowman AF. Signal-mediated export of proteins from the malaria parasite to the host erythrocyte. J Cell Biol 2005;171:587–92.
[21] Savchenko A, Yee A, Khachatryan A, et al. Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. Proteins 2003;50:392–9.
[22] Hall N, Karras M, Raine JD, et al. A comprehensive survey of the plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science 2005;307:82–6.
[23] Kimber MS, Vallee F, Houston S, et al. Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens. Proteins 2003;51:562–8.
[24] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[25] Wu N, Gillon W, Pai EF. Mapping the active site-ligand interactions of orotidine 5′-monophosphate decarboxylase by crystallography. Biochemistry 2002;41:4002–11.