Available online at www.sciencedirect.com
SCIENCE DIRECT
Molecular & Biochemical Parasitology 148 (2006) 144–160
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Heterologous expression of proteins from *Plasmodium falciparum*: Results from 1000 genes

Christopher Mehlin[a,*], Erica Boni[a], Frederick S. Buckner[a,b], Linnea Engel[a], Tiffany Feist[a], Michael H. Gelb[a,c,d], Lutfiyah Haji[a], David Kim[a,d], Colleen Liu[a], Natascha Mueller[a], Peter J. Myler[a,g], J.T. Reddy[a,d], Joshua N. Sampson[h], E. Subramanian[d,1], Wesley C. Van Voorhis[a,b,f], Elizabeth Worthey[a,g], Frank Zucker[a,d], Wim G.J. Hol[a,d,e]

[a] *Structural Genomics of Pathogenic Protozoa (SGPP), Box 357350, University of Washington, Seattle, WA 98195, USA*
[b] *Department of Medicine, University of Washington, Seattle, WA 98195, USA*
[c] *Department of Chemistry, University of Washington, Seattle, WA 98195, USA*
[d] *Department of Biochemistry, University of Washington, Seattle, WA 98195, USA*
[e] *Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA*
[f] *Department of Pathobiology, University of Washington,Seattle, WA 98195, USA*
[g] *Seattle Biomedical Research Institute, Seattle, WA, USA*
[h] *Department of Biostatistics, University of Washington, Seattle, WA 98195, USA*

Received 28 December 2005; received in revised form 16 March 2006; accepted 21 March 2006
Available online 18 April 2006

Abstract

As part of a structural genomics initiative, 1000 open reading frames from *Plasmodium falciparum*, the causative agent of the most deadly form of malaria, were tested in an *E. coli* protein expression system. Three hundred and thirty-seven of these targets were observed to express, although typically the protein was insoluble. Sixty-three of the targets provided soluble protein in yields ranging from 0.9 to 406.6 mg from one liter of rich media. Higher molecular weight, greater protein disorder (segmental analysis, SEG), more basic isoelectric point (pI), and a lack of homology to *E. coli* proteins were all highly and independently correlated with difficulties in expression. Surprisingly, codon usage and the percentage of adenosines and thymidines (%AT) did not appear to play a significant role. Of those proteins which expressed, high pI and a hypothetical annotation were both strongly and independently correlated with insolubility. The overwhelmingly important role of pI in both expression and solubility appears to be a surprising and fundamental issue in the heterologous expression of *P. falciparum* proteins in *E. coli*. Twelve targets which did not express in *E. coli* from the native gene sequence were codon-optimized through whole gene synthesis, resulting in the (insoluble) expression of three of these proteins. Seventeen targets which were expressed insolubly in *E. coli* were moved into a baculovirus/*Sf*-21 system, resulting in the soluble expression of one protein at a high level and six others at a low level. A variety of factors conspire to make the heterologous expression of *P. falciparum* proteins challenging, and these observations lay the groundwork for a rational approach to prioritizing and, ultimately, eliminating these impediments.

Keywords: *Plasmodium falciparum*; Protein expression; *E. coli*; Structural genomics; Ligase independent; Cloning; Insect cells

Abbreviations: %AT, percent adenosine and thymidine; CDK, cyclin dependent kinase; DMSO, dimethyl sulfoxide; EDTA, ethylamine diamine tetraacetic acid; GRAVY, Grand Average of Hydrophobicity; GST, Glutathione S-transferase; HEPES, 4-(2-hydroxyethyl)piperazine-1-ethane-sulfonic acid; IPTG, isopropyl β-D-1-thiogalactopyranoside; kDa, kilodalton; LIC, ligation independent cloning; MBP, maltose binding protein; PCR, polymerase chain reaction; pI, isoelectric point; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; SEG, segmental analysis of disorder; SGPP, Structural Genomics of Pathogenic Protozoa; WGS, whole gene synthesis

* Corresponding author. Tel.: +1 206 265 7237.
E-mail address: cmehlin@u.washington.edu (C. Mehlin).
[1] Visiting scholar from the Department of Crystallography and Biophysics, University of Madras, India.

0166-6851/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2006.03.011

1. Introduction

Despite enormous efforts to control malaria, it remains one of the world's most devastating diseases. With an annual infection rate of over half a billion people worldwide, a mortality rate of over a million, mostly children, and the ever-escalating issue of drug resistance, malaria presents an immediate and increasing challenge to the scientific and medical communities [1]. The impact of the disease is largely confined to impoverished regions of the world, leaving it considerably less attractive to pharmaceutical companies and necessitating a high level of collaboration among the research groups working in this field. Organizations such as the Gates Foundation and the Medicines for Malaria Venture (MMV) have helped the field progress in recent years [2], in part by facilitating these collaborative efforts. However, the urgent need for new drugs, especially against novel targets, has continued to outstrip efforts in development. One fundamental issue is that obtaining large amounts of putative drug targets from *Plasmodium falciparum* for screening and structural study has frequently proven to be difficult with heterologous expression efforts, particularly in *E. coli*, typically resulting in a lack of expression or in the expression of the protein as insoluble inclusion bodies [3–5].

The recent completion of the genome sequencing project for *P. falciparum* [6], the parasite responsible for the most deadly form of malaria, has enabled a great deal of research which previously was not possible. Among this research has been a structural genomics project: the Structural Genomics of Pathogenic Protozoa (SGPP, www.sgpp.org). The SGPP is an NIGMS-funded [7] effort to determine X-ray crystallographic structures of proteins from pathogenic protozoa, and *P. falciparum* figures prominently among the target organisms of this consortium. The methodology of the SGPP borrowed heavily from the successful efforts of other structural genomics groups, including the use of Ligation Independent Cloning (LIC) in the Midwest Center for Structural Genomics (MCSG) [8], autoinduction media in the New York Structural Genomix Research Consortium (NYSGRC) [9], 96-well *E. coli* growth and SDS-PAGE screening employed within the TB Structural Genomics Consortium [10], and a large amount of advice from the structural genomics community provided in the context of regular NIGMS Protein Production and Crystallization workshops. Most structural genomics groups have chosen to focus on prokaryotic organisms, in large part because these proteins are relatively facile to express in *E. coli*, but the SGPP has worked strictly on eukaryotic parasites. *P. falciparum* is well known for having proteins which are particularly resistant to heterologous expression. A variety of factors are thought to account for this: the genes average one intron each, requiring the use of cDNA for cloning; the genome is 80% AT-rich and has a codon bias well removed from *E. coli*; the genes are characterized by long, continuous stretches of adenosines and thymidines; glycosylation patterns are utilized which are unique to the parasites [11]; and the *P. falciparum* proteins are generally larger than homologues in other species, as much as 50% larger than homologues in *S. cerevisiae* [12], and often possess long, disordered loops [13]. A Cyclin Dependent Kinase homologue from *P. falciparum*, for example, has 83 asparagines in a row [12]. Long stretches of asparagines and glutamines are predisposed to forming amyloid-like aggregates, and about 24% of *P. falciparum* genes carry these repeats [14]. *E. coli* has been shown to sporadically substitute amino acids in low-complexity *P. falciparum* sequences, resulting in non-homogenous protein which appears homogenous by SDS-PAGE [15]. There is some evidence that *P. falciparum* proteins may bind their own mRNA [16]. In addition, it is not uncommon for *P. falciparum* genes to contain cryptic start sites for *E. coli*, resulting in multiple, truncated products when overexpressed in bacteria [17].

Despite these difficulties, there are many examples of *P. falciparum* proteins expressed in *E. coli*, including over 30 different proteins in the Protein Data Bank. The simplicity and speed of the T7 expression system has made *E. coli* the host of choice for essentially all structural genomics groups, including the SGPP. Because cloning and screening for soluble expression can be accomplished in a high-throughput manner, the fastest way to produce a large number of proteins is to rapidly clone a large number of targets and screen for those which will be amenable to soluble expression in *E. coli*. Presented here are our efforts to apply this approach to 1000 open reading frames from *P. falciparum.* In addition, 12 proteins which were not expressed by *E. coli* were codon-optimized through the whole gene synthesis (WGS), and 17 proteins which were insolubly expressed in *E. coli* were moved to a baculovirus expression system. While the ideal result of this endeavor would have been to discover a universally applicable method for the expression of these proteins, this was not achieved. Certain qualities of the genes and proteins were, however, found to profoundly affect their soluble expression, and these observations may contribute to future efforts to study the structure and function of *P. falciparum* proteins and the development of rational strategies for their heterologous expression.

2. Results

2.1. Target selection

Open reading frames were selected from the annotated database at PlasmoDB [18] and are listed in the supplementary materials for this publication. The bulk of the targets was selected using standard parameters for structural genomics projects: relatively small (<450aa), non-membrane proteins with <30% identity to anything currently in the PDB. The *var* and *rif* multigene surface proteins exported to the surface of the infected red blood cell were specifically excluded. The rest were chosen because they were predicted to be soluble enzymes and therefore have a higher likelihood of chemotherapeutic accessibility. Other than a bias towards smaller proteins (median 35.3 kDa in the target set versus 52.7 kDa in the genome [19]), the target set was fairly representative of *P. falciparum* open reading frames [6]. 60.8% of the target set and 60.9% of the genome were annotated as "hypothetical" and 47.9% of the target set contained introns, slightly less than the 54% with introns in the genome. The coding regions of the *P. falciparum* genome con-

tain 76.5% adenosine and thymidine (AT), while those of the target set had a mean of 74.8% AT content.

2.2. Amplification, cloning, and sequencing

Targets were amplified by Polymerase Chain Reaction (PCR) in batches of 94 at a time. All PCR products were visualized and purified by agarose gel electrophoresis, and to facilitate this step the PCR reactions were pre-arranged by expected size. All targets were cloned into a modified pET vector and screened for soluble expression in *E. coli*.

In order to ensure that the cloning was proceeding as expected, a subset of 108 of the clones was sequenced from both ends. Of these, 85 (79%) were found to contain the correct target. Of the 23 which did not contain the insert, 17 had a weak or invisible band by PCR, suggesting that poor PCR amplification was the source of the problem. The high cloning success rate when visible bands were present underscores the utility of ligation independent cloning for large-scale cloning projects like this. Even when a PCR band was not seen at all, the pre-sorting of targets by size allowed for extraction of the gel slice where the expected product should have been; there were three cases where no PCR product was observed but the proper product was cloned nonetheless. It was found that visualization was easiest when the products were arrayed by size in groups of six, from low to high, generating "ladders" which served both to provide a simple verification of the size of the band and the identity of the well it was taken from.

Despite using a proofreading polymerase, a very large number of point mutations was observed. Even though not all of the sequences were complete, 59 of the 85 correct targets were found to contain mutations relative to the expected sequence. Importantly, 28 contained mutations which resulted in a frame shift, typically in long stretches of adenosines or thymidines. In an effort to determine whether these frameshifts were being introduced at the PCR stage or in the process of being replicated in *E. coli*, several of the PCR products were sequenced prior to cloning. The frameshifts were not observed in sequences which were obtained straight from the PCR, suggesting that they were introduced by the *E. coli* (data not shown). It is possible that the frameshift mutations were occurring in a small percentage of each PCR product; this small percentage would be invisible to direct sequencing but could be selected for during cloning. It was somewhat of a surprise that errors in the primers were not a problem; none of these targets was observed to have a frameshift mutation in the primer region.

2.3. Screening, epression, and purification of targets

The rate of attrition of these targets was very high in both expression and solubility. Three hundred and thirty-seven (33.7%) were expressed in *E. coli*, as shown by a band on an SDS-PAGE gel of the urea-solubilized pellet obtained following sonication in small-scale screening. Of those which expressed only 18.7% were soluble at sufficient levels for purification, resulting in a 6.3% overall soluble expression rate. Thus, for every soluble protein obtained, more than four were observed to express insolubly and more than 10 did not express at all.

A total of 63 soluble proteins, listed in Table 1, were obtained in yields ranging from 0.9 to 406 mg from a liter of culture. Ninty-nine different targets appeared to have some solubility on a small-scale and were attempted on a large-scale. When a target screened positive on a small-scale but failed on a large-scale the cause was approximately evenly divided between either insufficient solubility or insufficient yield. Every target which exhibited some solubility in the small-scale screen was tried on a large-scale at least twice before it was abandoned. As these proteins were being generated for structural study and relatively large amounts of protein were required, those which did not produce high levels of soluble protein were not pursued further. It should be noted that the amounts of protein recovered from the top-producing targets were several times higher than the theoretical capacity of the nickel resin and may be an indication of protein aggregation. Clearly, the soluble expression of protein was rare, and the expression of high levels of soluble protein was even less frequent.

2.4. Data analysis

The factors which appeared likely to influence expression and solubility (e.g. codon usage, protein disorder, protein size) are highly interrelated, complicating the analysis of what the root issues are in heterologous protein expression for *P. falciparum.* Table 2 shows univariate, unadjusted relationships between various genetic and physical characteristics of these targets with their outcomes in expression trials. Table 2 also lists univariate relationships between target characteristics and solubility separately for the complete set proteins and only the expressed proteins. The latter decouples expression from solubility. Although examining all 1000 proteins technically has greater power, such an analysis does not directly and specifically address solubility. Therefore, the majority of solubility analyses focus on solubility among expressed proteins.

The factors under consideration were broken down into three categories: protein characteristics, gene composition, and relationships. Protein characteristics included the molecular weight of the target, the isoelectric point (pI), predicted disorder, and hydrophobicity. Gene composition encompassed the overall adenosine/thymidine content (AT%), the number of introns, and the longest stretch of adenosines or thymidines. The similarity of these proteins to *E. coli* homologues, the presence of hits in the Pfam database, and the annotation of the targets were considered as part of relationships.

In order to reduce as much as possible the confounding relationships of these factors, those which had a significant effect on protein expression or the solubility of expressed targets were further scrutinized using multivariate logistic regression to determine a final model (Table 3), discussed below.

2.5. Protein characteristics associated with protein expression and solubility

The univariate analysis of the correlation between protein characteristics and protein expression and solubility suggested that the physical characteristics of the proteins were very impor-

Table 1
Plasmodium falciparum proteins solubly produced in *E. coli*

SGPP name	PlasmoDB name	Annotation	Yield (mg/l)
Pfal008495AAA	PFE1035c	BIS(5′-nucleosyl)-tetraphosphatase, putative	406.6
Pfal009132AAA	PFL0780w	Glycerol-3-phosphate dehydrogenase, putative	187.2
Pfal001782AAA	PFL0210c	Eukaryotic initiation factor 5a, putative	109.9
Pfal004546AAA	PFF1360w	6-Pyruvoyl tetrahydropterin synthase, putative	97.8
Pfal003231AAA	PF14_0210	Hypothetical protein	92.4
Pfal007493AAA	PFA0315w	Hypothetical protein	86.6
Pfal006645AAA	PF13_0349	Nucleoside diphosphate kinase b; putative	76.2
Pfal007201AAA	PF14_0545	Thioredoxin	75.6
Pfal004616AAA	PFF1050w	Nascent polypeptide associated complex alpha chain, putative	72.4
Pfal008828AAA	PFI1090w	s-Adenosylmethionine synthetase, putative	68.6
Pfal003122AAA	MAL13P1.103	Hypothetical protein	64.5
Pfal001447AAA	PF11_0282	Deoxyuridine 5′-triphosphate nucleotidohydrolase, putative	62.4
Pfal008498AAA	PFE1050w	Adenosylhomocysteinase	62
Pfal004654AAA	PFF0880c	Hypothetical protein	61.5
Pfal007922AAA	PFC0525c	Glycogen synthase kinase, putative	61.2
Pfal000066AAA	PF10_0022	Hypothetical protein	59.2
PfAl008421AAA	PFE0660c	Uridine phosphorylase, putative	58.2
Pfal005333AAA	PF08_0085	Ubiquitin-conjugating enzyme, putative	55.6
Pfal006597AAA	PF13_0301	Ubiquitin-conjugating enzyme, putative	51.1
Pfal008844AAA	PFI1170c	Thioredoxin reductase (NADPH) EC 1.6.4.5	47.9
Pfal008831AAA	PFI1105w	Phosphoglycerate kinase	47.4
Pfal009014AAA	PFL0190w	Ubiquitin-conjugating enzyme e2, putative	43.3
Pfal002645AAA	PFL1850c	Hypothetical protein	36.8
Pfal004331AAA	MAL13P1.257	Hypothetical protein, conserved	34.8
Pfal005717AAA	PF10_0330	Ubiquitin-conjugating enzyme, putative	30
Pfal006638AAA	PF13_0342	Hypothetical protein	28.6
Pfal006386AAA	PF13_0085	Serine/threonine protein kinase, putative	28.1
Pfal006583AAA	PF13_0287	Adenylosuccinate synthetase	27.8
Pfal006626AAA	PF13_0330	ATP-dependent DNA helicase, putative	27
Pfal004761AAA	PFF0260w	ST kinase, putative	26.4
Pfal003552AAA	PF14_0257	Leucine-rich acidic nuclear protein	26
Pfal006677AAA	PF14_0020	Choline kinase, putative	22.6
Pfal001147AAA	PF11_0503	Hypothetical protein	22.4
Pfal003589AAA	PF14_0064	Vacuolar sorting protein VPS29	22
Pfal004253AAA	MAL13P1.178	Hypothetical protein	21.9
Pfal004414AAA	MAL13P1.339	Hypothetical protein	21.3
Pfal008434AAA	PFE0730c	Ribose 5-phosphate epimerase, putative	20.7
Pfal00023AAA	PFE0625w	Putative GTPase (rab1b gene)	16.7
Pfal002849AAA	PFL2225w	Myosin A tail domain interacting protein MTIP, putative	16.5
Pfal004964AAA	MAL8P1.108	Protein phosphatase, putative	16.2
Pfal003403AAA	PF14_0490	Hypothetical protein	16.1
Pfal008494AAA	PFE1030c	Phosphomethylpyrimidine kinase, putative	14.6
Pfal009167AAA	PFL0960w	D-ribulose-5-phosphate 3-epimerase, putative	14.3
Pfal000954AAA	PF11_0434	Hypothetical protein	14.2
Pfal005921AAA	PF11_0145	Glyoxalase I, putative	13.4
Pfal005676AAA	PF10_0289	Adenosine deaminase, putative	13.2
Pfal005312AAA	PF08_0064	Hypothetical protein, conserved	11.1
Pfal008670AAA	PFI0300w	Developmental protein, putative	10.6
Pfal008572AAA	PFE1430c	Cyclophilin, putative	9
Pfal007254AAA	PF14_0598	Glyceraldehyde-3-phosphate dehydrogenase	8.6
Pfal002361AAA	PFL1340c	Hypothetical protein	8.4
Pfal006371AAA	PF13_0070	Branched-chain alpha keto-acid dehydrogenase, putative	7.9
Pfal007858AAA	PFC0255c	Ubiquitin-conjugating enzyme E2, putative	7.2
Pfal001042AAA	PF11_0393	Hypothetical protein	7.1
Pfal005257AAA	PF08_0009	Translation initiation factor EIF-2b alpha subunit, putative	6.8
Pfal001263AAA	PF11_0293	Multiprotein bridging factor type 1, putative	6.6
Pfal002997AAA	PFL2545c	Hypothetical protein	6.5
Pfal005255AAA	PF08_0007	Hypothetical protein	5.2
Pfal000304AAA	PF10_0225	Orotidine-5′-monophosphate decarboxylase	4.1
Pfal004233AAA	MAL13P1.159	Hypothetical protein, conserved	3.9
Pfal002331AAA	PFL1275c	Hypothetical protein	2.6
Pfal003458AAA	PF14_0465	Hypothetical protein	2.4
Pfal000154AAA	PF10_0151	Conserved hypothetical protein	0.9

The above 63 proteins were those which were successfully expressed in *E. coli*, grown on a 1 l scale, and purified by both nickel chromatography and size-exclusion chromatography.

Table 2
Expression and solubility by characteristic

Category	*N*	All proteins				Expressed proteins	
		N (%) expressed	*p*	*N* (%) soluble	*p*	% Soluble	*p*
All proteins	1000	337 (33.7)		63 (6.3)		18.7	
Protein characteristics							
Fusion MW			<0.01		<0.01		0.01
9153.41–26229	250	106 (42.4)		28 (11.2)		26.4	
26230–35323	250	86 (34.4)		7 (2.8)		8.1	
35324–47207	250	86 (34.4)		18 (7.2)		20.9	
47208–224845	250	59 (23.6)		10 (4.0)		16.9	
pI			0.02		<0.01		<0.01
3.45–6.8	261	104 (39.8)		32 (12.3)		30.8	
6.9–8.4	232	83 (35.8)		19 (8.2)		22.9	
8.5–9.5	264	85 (32.2)		11 (4.2)		12.9	
9.6–12.1	243	65 (26.7)		1 (0.4)		1.5	
Disopred %			<0.01		0.03		0.27
0–9	240	102 (42.5)		24 (10.0)		23.5	
10–19	251	81 (32.3)		17 (6.8)		21	
20–30	248	79 (31.9)		12 (4.8)		15.2	
31–101	261	75 (28.7)		10 (3.8)		13.3	
SEG percent			<0.01		<0.01		0.02
0	226	98 (43.4)		27 (11.9)		27.6	
1–20	560	179 (32.0)		29 (5.2)		16.2	
21–81.08	214	60 (28.0)		7 (3.3)		11.7	
GRAVY			0.1		0.05		0.16
0.27–0.35	76	28 (36.8)		4 (5.3)		14.3	
0.36-0.40	243	74 (30.5)		8 (3.3)		10.8	
0.41-0.45	462	147 (31.8)		30 (6.5)		20.4	
0.46–0.56	219	88 (40.2)		21 (9.6)		23.9	
Gene composition							
AT %			<0.01		<0.01		<0.01
60–72	306	129 (42.2)		37 (12.1)		28.7	
73–75	277	87 (31.4)		12 (4.3)		13.8	
76–77	200	58 (29.0)		8 (4.0)		13.8	
78–84	217	63 (29.0)		6 (2.8)		9.5	
Number of introns			0.16		0.2		0.41
0	502	179 (35.7)		33 (6.6)		18.4	
1–2	267	78 (29.2)		12 (4.5)		15.4	
>2	212	76 (35.8)		18 (8.5)		23.7	
Longest A or T stretch			0.03		0.03		0.1
2–6	509	184 (36.1)		42 (8.3)		22.8	
7–9	367	124 (33.8)		17 (4.6)		13.7	
10–30	124	29 (23.4)		4 (3.2)		13.8	
Relationships							
E. coli			<0.01		0.04		0.12
0–5	143	39 (27.3)		10 (7.0)		25.6	
6–8	114	46 (40.4)		5 (4.4)		10.9	
9–14	119	40 (33.6)		11 (9.2)		27.5	
15–65	118	57 (48.3)		17 (14.4)		29.8	
None	506	155 (30.6)		20 (4.0)		12.9	
Pfam size			0.24		<0.01		<0.01
0	629	199 (31.6)		26 (4.1)		13.1	
1–100	112	42 (37.5)		5 (4.5)		11.9	
101–500	117	48 (41.0)		15 (12.8)		31.2	
501–1000	55	21 (38.2)		8 (14.5)		38.1	
>1000	87	27 (31.0)		9 (10.3)		33.3	
Annotation			0.21		<0.01		<0.01
Actual protein	420	151 (36.0)		42 (10.0)		27.8	
Hypothetical protein	579	185 (32.0)		21 (3.6)		11.4	

The 1000 proteins are divided into categories according to various characteristics. Columns list the number in each group, the number (and %) expressed, the number (and %) soluble, and the % of the expressed proteins which were soluble. *p*-Values are calculated from a chi-squared test performed separately for each variable.

Table 3
Multivariate analysis of expression and solubility

Variable	Odds ratio	95% CI	*p*-Value
A. Expression			
pI	0.91	(0.85–0.98)	0.01
SEG percent (per 10%)	0.86	(0.75–0.97)	0.02
Number of introns			0.07
0	Reference		
1–2	0.68	(0.48–0.94)	
>2	0.9	(0.64–1.28)	
Fusion MW (per 10,000)	0.85	(0.78–0.92)	<0.01
E. coli similarity			0.04
<1%	Reference		
1–5	1.14	(0.72–1.78)	
6–8	1.44	(0.94–2.22)	
9–14	1.02	(0.66–1.59)	
15–64	1.89	(1.23–2.9)	
B. Solubility of expressed proteins			
pI	0.66	(0.56–0.78)	<0.01
Annotation	0.32	(0.17–0.59)	<0.01

Multivariate logistic regression was employed to distil the interrelated factors for expression (A) and solubility (B) to the most predictive ones. These factors were then combined into the final models above. This simplified model was then re-tested using the factors which had been discounted; the number of introns was the only factor which came close to significance in this re-analysis and it is included for comparison.

tant (Table 2). Increasing protein size, pI, and disorder (by both the SEG [20] and DISOPRED2 [21] methods) were all found to be negatively correlated with expression. With the exception of the DISOPRED2 algorithm, these same characteristics similarly impacted the solubility of expressed proteins. In contrast, the hydrophobicity of proteins (Grand Average of Hydrophobicity, GRAVY [22]) was barely correlated with overall soluble expression ($p = 0.05$) and did not correlate specifically with either expression or the solubility of those proteins which expressed.

The one physical protein quality which was associated with both expression and solubility following multivariate logistic regression was pI (Fig. 1, Table 3). The trend towards insolubility with increasing pI was striking: of the 288 targets with a pI above 10 only one was soluble. At very low pI values, there appeared to be a tendency towards a lack of expression which was compensated for by greater level of solubility among those which did express. The mean pI of these targets, including the tags, was 8.2. Since DNAse is used to help clarify samples in the small-scale screening, where the same trend was observed, it does not seem likely that high pI proteins were lost due to DNA binding. pI is the only physical characteristic of these targets which was correlated with solubility amongst the expressed proteins following multivariate analysis (Table 3B). This pI effect does not appear to be due to any one amino acid; the percentage content of aspartic acid, glutamic acid, and histidine went up and the percentage of tyrosine, lysine, and arginine went down amongst the expressed and the solubly expressed proteins (data not shown).

Attrition was highly correlated with protein size ($p < 0.01$), with larger proteins failing in expression at higher rates than smaller proteins (Fig. 2). It is apparent from the figure that the

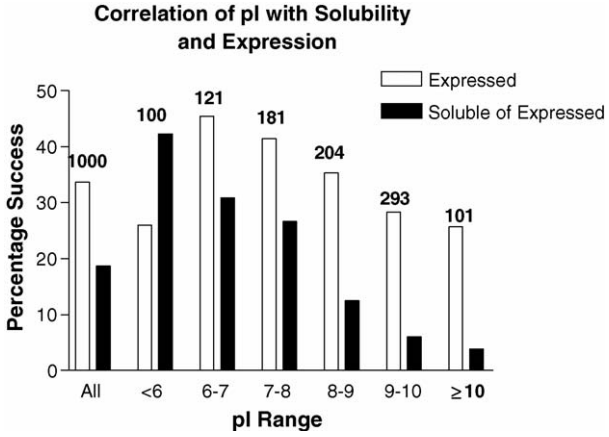

Fig. 1. Protein isoelectric point (pI) is inversely correlated with both expression and solubility. The 1000 targets were separated into categories based upon their calculated pI with the tags included. Shown is the percentage expressed and the percentage of those expressed which did so solubly (clear and black bars, respectively). The numbers above each category are the number of targets within the indicated pI range, and the values for all 1000 targets are shown to the left for comparison. Each category includes the lower whole pI value up to but not including the upper whole pI value.

size effect is especially pronounced at the extremes: 42% (41/98) of those proteins under 20 kDa expressed and 14% (14/98) were soluble while of those over 60 kDa only 20% (20/100) expressed and 3% were soluble (3/100). It is not surprising that protein size had a negative impact upon expression; among the set of 334 proteins which expressed, however, there was no correlation between gene/protein size and solubility.

Gel chromatography of the soluble proteins following nickel chromatography revealed their aggregation state. Thirt-two of the proteins (51%) were observed to migrate on the gel filtration column with an apparent mass greater than 1.5 times their theoretical mass, an indication of protein self-association. Three targets were observed to have migration consistent with the formation of large aggregates (>100 kDa), although two of these proteins were over 40 kDa, making it difficult to know, given

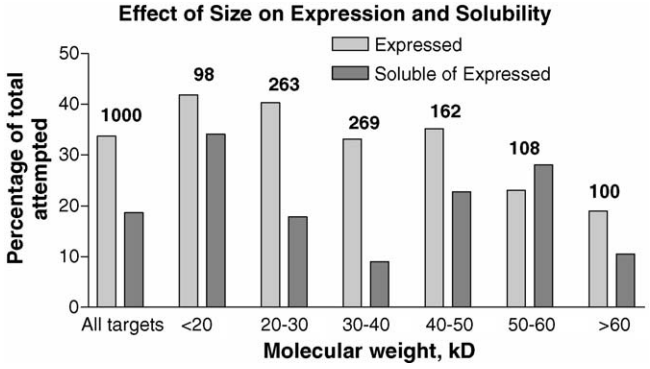

Fig. 2. Protein size is inversely correlated with expression but does not have a consistent effect on the solubility of expressed proteins. Targets were divided into categories by weight, and the percentage which expressed and soluble is shown for each category. The numbers above the bars are the number of samples in that category. "Expressed" contains both those which were expressed solubly and those expressed insolubly, and "soluble of expressed" is the percentage soluble of the total which expressed in that size category.

	ALL		Expressed			Soluble		
	Mean	SEM	Mean	SEM	p	Mean	SEM	p
SEG, longest	19.7	0.60	15.7	0.82	<.0001	11.4	1.6	0.0070
SEG, total	43.3	1.7	31.3	2.3	<.0001	21.4	4.2	0.0161
SEG, percent	12.1	0.37	10.3	0.62	0.0037	6.81	1	0.0016
Disorder, regions of >9	2.42	0.72	1.98	0.44	<.0001	1.6	0.22	0.0894
Disorder, max. length	37.2	1.4	29.6	1.9	<.0001	23.8	3	0.0581
Disorder, percent	22.3	0.55	20	0.9	0.0107	17.3	2	0.2096

(A)

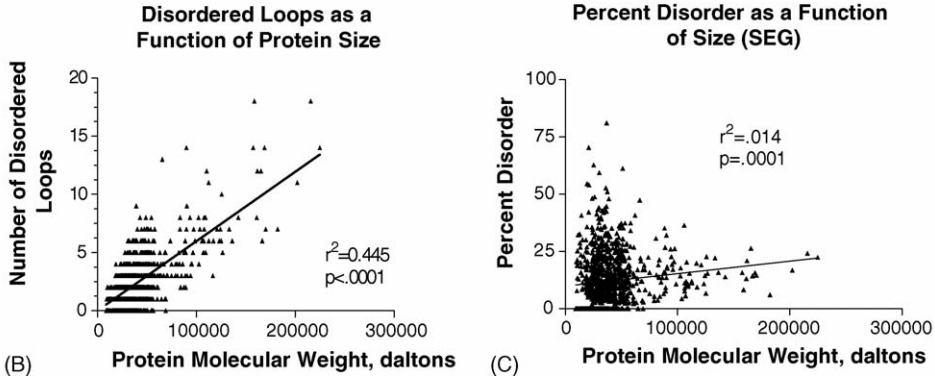

Fig. 3. Calculated disorder of the proteins is associated with downstream difficulties in expression and solubility. (A) SEG and disorder were calculated via the referenced methods. The *p*-value is the result of a two-tailed *t*-test for the mean of the expressed samples vs. the mean of the entire set and the mean of the soluble proteins versus the mean of the expressed proteins. "SEG, longest" is the longest continuous stretch of disordered amino acids. "SEG, total" is the total number of disordered amino acids, and "SEG, percent" is the number of disordered amino acids as a percentage of the total number of amino acids in the target. (B) Disordered loops were regions of 10 or more disordered amino acids as predicted by DISOPRED2. (C) The predicted disorder using SEG is plotted as a function of protein size; *p* values are for a non-zero slope as predicted by an *F*-test.

the columns used, whether this was gross aggregation or the formation of (for example) trimers. It is clear that while a significant proportion of these targets were self-associating, most did not form large aggregates.

Calculated disorder, determined via several methods, was negatively associated with expression (Fig. 3A), and the SEG analysis had more predictive value than did DISOPRED2. The correlation of SEG with expression held through multivariate analysis of expression, although all disorder prediction methods failed to be useful for solubility prediction amongst those proteins which expressed (Table 3). The detrimental effects of disorder on expression are likely to be a particularly vexing problem for the expression of *P. falciparum* proteins, as these frequently have large, disordered loops. As shown in Fig. 3B, the frequency of these disordered loops is highly correlated with protein size. The SEG score is correlated with protein size (Fig. 3C) more than is the DISOPRED2 score (data not shown); the size is not, however, a good predictor of disorder, as shown by the low r^2-value. The disorder of a protein, as determined by SEG, added predictive value to a multivariate model of expression which included protein size (Table 3A, $p = 0.02$), indicating that the correlation of disorder with a lack of expression extends beyond the confounding relationship of size.

Several factors appeared to have little correlation with expression and solubility. Hydrophobicity, calculated as GRAVY score, was not found to have a significant effect even as an independent variable in expression or solubility (Table 2). Higher serine content of proteins has been associated with insolubility [23], but in this *P. falciparum* dataset, serine content was 6.16% for insoluble proteins and 6.67% for soluble proteins, a statistically insignificant difference.

2.6. Gene composition, codon usage and protein expression

Given the nearly 80% AT content of the *P. falciparum* genome and a codon bias which is divergent from *E. coli* [24,25], it seemed likely that the genetic makeup of the parasite would contribute significantly to the failure rate in protein expression. The genetic characteristics investigated included the %AT content, the length of long stretches of A or T, and the number of introns. We excluded the %AT of introns from our analysis. Also, a separate model was constructed to look at the effect of specific codons on protein expression.

The statistically significant correlation of %AT with expression and solubility is shown in Table 2, but given the high degree of correlation between %AT and both pI and size (Fig. 4A and B) and its direct relationship with amino acid sequence[26], it appears that the effect of %AT is more likely due to its confounding relationship with these other factors. The multivariate analysis bears this out for both expression and solubility (Table 3), where %AT did not add meaningfully to the predictive value of either the expression or solubility model. Altogether, within this set of high %AT targets, differences in the %AT did not in themselves appear to affect expression or solubility.

The observation that frameshift mutations often occurred in continuous stretches of A or T prompted a closer examination of these regions: of the 334 targets which had 10 or more A or T in a row, 80 (24%) were observed to express, and 10 (3.0%) were

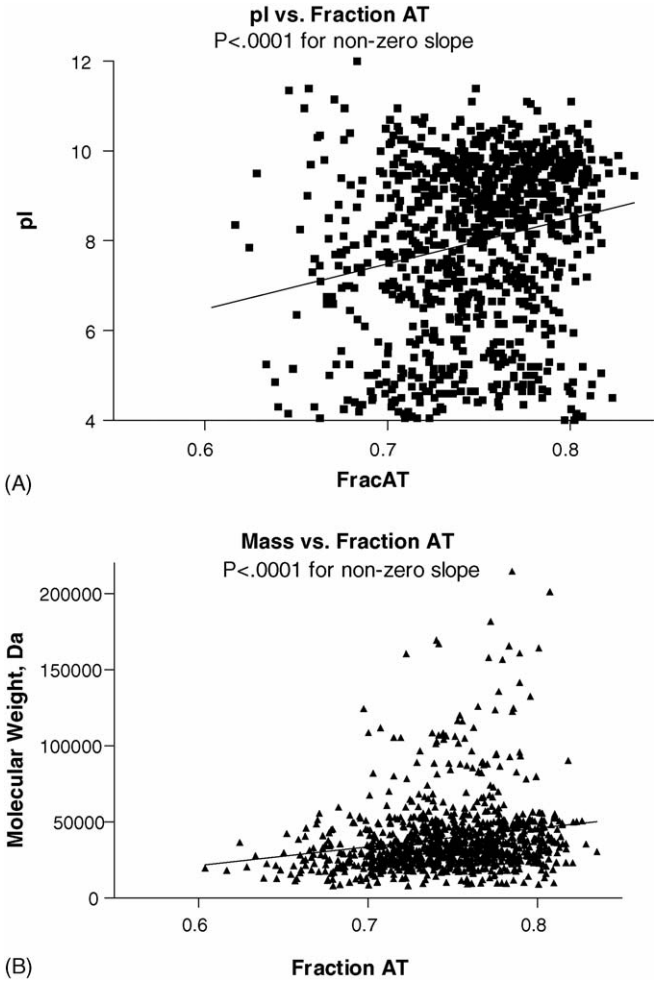

Fig. 4. Percentage of AT content in the genes was highly correlated with (A) isoelectric point (pI) and (B) protein size. *p* Values are for a non-zero slope.

observed to do so solubly. Although these figures are lower than the average rates for all targets (33.7% and 6.3%, respectively), it is worth noting that these 10 targets tended to express to high levels, ranging from 16.2 to 92.4 mg purified yield from a liter of culture.

One technical point which is of some interest for Ligation Independent Cloning (LIC) is that the technique appeared to be insensitive to gaps introduced into the 3′ end. With the vector employed here, the 3′ end of the target had no pre-defined stop point for the T4 polymerase used to make single-stranded overhangs from the PCR products. The enzyme's proofreading function removed bases from the 3′ end of the PCR product until it reached a cytosine naturally occurring in the target. Given the high AT content of the genes, this cytosine was often far from the 3′ (carboxy-terminal) end of the gene, as far back as 120 bases (median six bases), leaving a single-stranded gap. There was no correlation observed between the size of this gap and the expression of the proteins (data not shown), indicating that the *E. coli* have a robust ability to fill in such gaps or that the proofreading processivity of the T4 polymerase is slow or incomplete under these conditions.

In order to determine whether specific codons are responsible for difficulty in heterologous expression, the 10 codons in *P. falciparum* which were the rarest with respect to *E. coli* were assessed for their correlation with a lack of expression. As shown in Table 4, none of these codons was significantly associated with difficulties in protein expression.

2.7. Predictive models based upon annotation and other organisms

Having considered the characteristics of the proteins and genes themselves, the utility of models based upon relationships of these targets to those of other organisms was investigated. Pfam [27], a multiple-alignment-driven tool used for genome annotation, the *P. falciparum* annotation itself (divided broadly into "hypothetical" and "non-hypothetical"), and the percent identity of these targets to *E. coli* homologues were all considered as part of this analysis. Although these factors do not reflect the physical qualities of the targets per se and are likely to have a considerable overlap with each other, they have the potential to lend predictive value to the analysis of expression and solubility, and this value is likely to increase as more genomes are sequenced and these tools become more complete.

Nearly 2/3 of the proteins which were targeted for cloning and expression tests were annotated as hypothetical, reflecting the 60% of the *P. falciparum* genes with a hypothetical annotation following the genome sequencing project. Although a

Table 4
Codon usage and gene expression

Codon	Amino acid	Genome usage per 1000 codons			Statistical analysis		*p*-Value
		P. falciparum	*E. coli*	Ratio	Odds ratio	95% CI	
ATA	Isoleucine	49.9	4.3	12	0.96	(0.89–1.03)	0.26
AGA	Arginine	15.9	2.1	7.6	0.96	(0.87–1.07)	0.49
AAT	Asparagine	123	17.7	6.9	0.97	(0.94–1.01)	0.15
AGG	Arginine	4.3	1.2	3.6	1.07	(0.82–1.41)	0.61
TTA	Leucine	47.3	13.9	3.4	1.07	(1–1.15)	0.06
TAT	Tyrosine	50.4	16.3	3.1	0.96	(0.89–1.03)	0.25
ACA	Threonine	21.7	7.1	3.1	1.08	(0.94–1.23)	0.28
TGT	Cysteine	15.2	5.2	2.9	0.98	(0.86–1.11)	0.75
AAA	Lysine	95.6	33.6	2.8	0.99	(0.95–1.03)	0.61

Shown are the usage rates for the 10 rarest codons in *P. falciparum* relative to their use in *E. coli*. The significance was determined via ANOVA analysis for codon usage grouped by 10 per 1000.

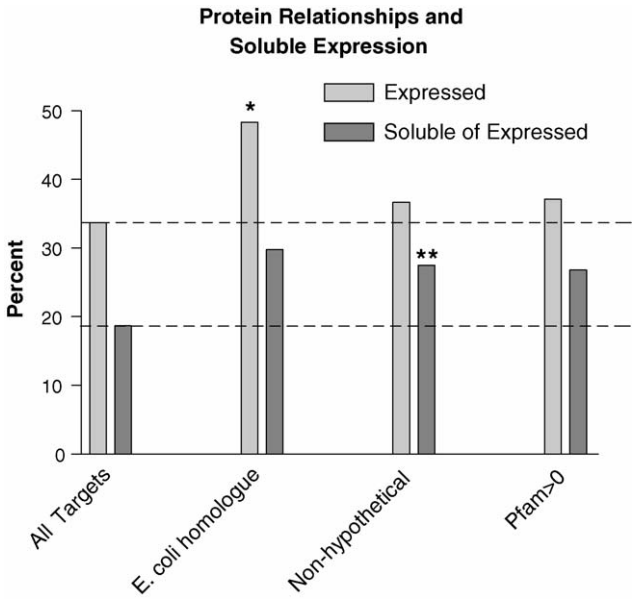

Fig. 5. Known proteins and proteins with homologues were more likely to be expressed and soluble. The categories investigated were *P. falciparum* targets with a homologue in *E. coli*, an annotation other than "hypothetical," or membership in Pfam. Multivariate logistic regression was applied to determine which relationships were the most predictive. Targets with a homologue in *E. coli* were much more likely to be expressed ($^{*}p = 0.004$), and targets with a non-hypothetical annotation were more likely to be soluble ($^{**}p = 0.048$). None of the other relationships added statistically significant power to these predictions.

hypothetical annotation was not significantly associated with expression (Table 2), the correlation with *soluble* expression was more pronounced, with significantly higher success rates seen with non-hypothetical proteins (Fig. 5). Following multivariate logistic regression, a non-hypothetical annotation and pI were the only two factors which were significantly correlated with solubility amongst the expressed proteins (Table 3B).

A similar kind of relationship can be drawn between expression and those proteins which have homologues in *E. coli* (Fig. 5). Half of the targets (50.6%) had no identifiable homologue in *E. coli*, even at the 1% identity level. It should be noted that the *P. falciparum* homologue of an *E. coli* protein typically shares a conserved core area but has additional loops which drive the overall homology down. A comparison of overall identity revealed that targets with *E. coli* homologues are much more likely to be expressed, although there was not a significant correlation with protein solubility (Table 2, Fig. 5). The 25 targets which were in the 95th percentile (>33% identical) with regard to *E. coli* homologues expressed at a rate of 82%. Interestingly, there was no correlation between the level of expression and the percent identity to an *E. coli* protein (data not shown). One potential confounding factor is that the proteins with a high percent identity to an *E. coli* homologue also tended to have a low %AT (data not shown); homology to an *E. coli* protein had a much higher correlation with expression, however, eclipsing %AT in the multivariate analysis (Table 3A).

Related to the issue of annotation is the association of targets with Pfam, which was used as a tool to aid the genome annotation. Six hundred and twenty-nine of the 1000 targets (62.9%) had a Pfam size of 0, and 525 (83.4%) of these also had a hypothetical annotation. While belonging to a Pfam family (Fig. 5) was only modestly associated with expression (37% for those with Pfam members versus 32% for those without), targets with Pfam matches were much more likely to be soluble (overall rates 10% for those with matches versus 4.1% for those without). When multivariate logistic regression was applied, however, the Pfam association was trumped by *E. coli* homologues in expression and non-hypothetical annotation in solubility (Table 3). As Pfam is a tool employed for annotation, it is not surprising that there is a large overlap amongst the targets categorized as being part of a Pfam family and categorized as being annotated: 81.7% with a Pfam score greater than zero had a non-hypothetical annotation. Thus Pfam family membership and non-hypothetical annotation are associated in much the same way with solubility, although non-hypothetical annotation was more highly correlated, and a multivariate model which included annotation was not improved by Pfam.

Table 5
Synthetic, codon-optimized genes and *E. coli* expression

SGPPname	PlasmoDB	Annotation	Fusion MW	Result
Pfal006912AAA	PF14_0256	Exosome complex exonuclease rrp41	28550	Nothing
Pfal004600AAA	MAL6P1.197	Ferredoxin-NADP reductase	44670	Expressed
Pfal007889AAA	PFC0380w	Dual-specificity protein phosphatase	68464	Nothing
Pfal006547AAA	PF13_0251	DNA topoisomerase III	85195	Nothing
Pfal008792AAA	PFI0910w	DNA helicase	86398	Truncated
Pfal007173AAA	PF14_0517	Peptidase	89392	Truncated
Pfal005232AAA	PF07_0123	mRNA (N6-adenosine)-methyltransferase	90524	Nothing
Pfal008996AAA	PFL0100c	ATP dependent RNA helicase	96755	Truncated
Pfal008637AAA	PFI0135c	Papain family cysteine protease	106432	Truncated
Pfal006604AAA	PF13_0308	DNA helicase	107178	Nothing
Pfal005964AAA	PF11_0188	Heat shock protein 90	109353	Expressed
Pfal004033AAA	PFB0420w	YgbB protein	109569	Expressed

In some cases, synthetic genes can help to get *P. falciparum* genes to express. The 12 genes above were cloned from *P. falciparum* and sequenced. None of the genes in their native form provided expression in *E. coli* as visualized on an SDS-PAGE gel. The full-length genes were made synthetically as codon-optimized constructs, sequenced completely, and tested for expression in *E. coli*. None of the constructs provided soluble protein, but several yielded insoluble protein, either full-length or truncated as shown. All genes, both native and synthetic, were tested at least three times from independent *E. coli* colonies.

Table 6
Baculovirus expression of native *P. falciparum* genes

SGPP ID	PlasmoDB	Annotation	Fusion MW	Yield, mg
Pfal000259AAA	Pf334_t00005	Ard1 family	17804	
Pfal000429AAA	Pf392_t00001	Hypothetical protein	23020	1.2
Pfal005358AAA	PF08_0110	Rab18 GTPase	23163	1.5
Pfal000561AAA	Pf1136.t00001	Hypothetical protein	23330	0.9
Pfal000115AAA	Pf313_t00011	Hypothetical protein	24130	1.2
Pfal005915AAA	PF11_0139	Protein tyrosine phosphatase, putative	25102	7.2
Pfal004314AAA	MAL13P1.241	GTPase, putative	26061	
Pfal009000AAA	PFL0120c	Cyclophilin, putative	26429	11.7
Pfal000255AAA	Pf334_t00001	Hypothetical protein	32524	0.6
Pfal009033AAA	PFL0285w	Glyoxalase II family protein, putative	38414	
Pfal005766AAA	PF10_0379	Phospholipase, putative	41770	0.6
Pfal008186AAA	PFD0725c	Arsenical pump-driving ATPase, putative	43317	
Pfal006641AAA	PF13_0345	Aminomethyltransferase, mitochondrial	46903	
Pfal007204AAA	PF14_0548	ATPase, putative	48259	3
Pfal005857AAA	PF11_0071	RuvB DNA helicase, putative	53406	
Pfal000149AAA	Pf318_t00009	Hypothetical protein	53804	
Pfal006636AAA	PF13_0340	Exosome complex exonuclease, putative	55027	

A baculovirus expression system rarely helped for proteins insolubly expressed in *E. coli*. Seventeen sequence-verified genes which were observed to express insolubly in *E. coli* were moved into a baculovirus expression system and grown on a 500 ml scale. Shown are the relatively crude yields following a single nickel chromatography step.

2.8. Alternate approaches to protein cloning and expression: whole gene synthesis and baculovirus

Twelve genes which were not observed to express in *E. coli* were synthesized as codon-optimized versions and cloned into the same vector. The amino acid sequence, including any low-complexity loops, etc. was not altered. Both the native and synthetic genes were fully sequenced to ensure that there were no frameshifts, and the resulting constructs were appraised for expression by SDS-PAGE. Table 5 shows the results of this experiment. Out of these 12 artificial genes, three were found to express as full-length products and four expressed as truncated products, as judged by SDS-PAGE. None of these expressed solubly, although WGS is not likely to address issues of solubility unless the native amino acid sequence is altered.

Seventeen genes which were observed to express insolubly in *E. coli* were moved into a baculovirus expression system. The native *P. falciparum* gene sequence was retained, the genes were all sequenced, and the Sf-21 cells were grown on a 500 ml scale. Seven of the 17 (41.2%, Table 6) provided 0.9 mg or more of soluble protein from these targets following nickel chromatography. It should be noted that these proteins were considerably less pure than those produced by *E. coli*, and therefore yields are likely to be somewhat of an overestimate. It is apparent that a baculovirus system can provide soluble *P. falciparum* proteins in cases where *E. coli* produce only insoluble material, although expression via baculovirus involves considerably higher investment in time and will probably necessitate a high volume of cell culture in order to obtain sufficient yield.

3. Discussion

In most cases, the soluble, heterologous expression of milligram amounts of *P. falciparum* proteins will be challenging. It is worth noting, however, that crystallography is particularly demanding in regard to protein supply, typically requiring at least 5 mg of rigorously pure material. The expression of eukaryotic proteins in *E. coli* is typically associated with a lack of solubility [28], and *P. falciparum* has proven to be no exception to this. It is likely that variations in cloning vector, expression temperature, *E. coli* strain, etc. might well allow access to some of these proteins which were missed in our survey. Although the percentage of proteins solubly expressed would go up with such methodology, it would likely be disproportionate to the amount of effort required to obtain them.

The high correlation of pI and insolubility (Fig. 1, Table 3) was particularly striking and unexpected. Some loss of solubility was anticipated near 7.25, the pH of the solubilization buffer employed; this, however, was not the case with these targets. There was also no significant correlation between protein size and pI (data not shown). It has been observed that integral membrane proteins typically have high pI values [29], so it may be that these proteins are associated with membranes. A similar trend was also reported by the Midwest Center for structural genomics for *T. maritima* proteins, although there also was a dip in the soluble expression of proteins near neutrality [30]. The pI emerged as one of the most important factors in the soluble expression of these *P. falciparum* targets, and the mechanism and implications of this represent the largest remaining puzzles left by this work.

It certainly does not come as much of a surprise that size has a profound effect on the likelihood that a protein will be expressed, and one might expect, too, that larger proteins would have more opportunities to misfold and aggregate than would smaller proteins. It appears, however, that the protein size has a much greater impact on expression than on solubility (Fig. 1). The apparent lower solubility of proteins around 30–40 kDa is puzzling, though; there is not any clear factor associated with these intermediate-size targets (such as hypothetical annotation) which might account for their relative insolubility (data

not shown). It should be recognized, too, that "expression" in this study is a combination of PCR, cloning, and expression. Larger targets should be harder to get as full-length PCR clones and should be less stable in *E. coli*, two factors which would be understood in this study as a lack of expression.

An unexpected result was the relatively weak association between the %AT and protein expression. Although the initial analysis (Table 2) did show a strong correlation between %AT and both expression and solubility, the %AT dropped out completely during multivariate analysis (Table 3). Given that the %AT does correlate with both pI and molecular weight (Fig. 4), it stands to reason that there is a bit of guilt by association, and it is difficult to explain why the %AT would have an effect on protein solubility per se unless this correlation reflects a confounding relationship, such as an association between the %AT and pI or sequence. The %AT is known to have an effect on the codons and amino acids employed in *P. falciparum* [26,31]. It was surprising that there was so little correlation between long stretches of adenosine or thymidine and failure in expression. The large number of mutations and frameshifts observed implies that the genes have features which make it difficult for the *E. coli* to faithfully maintain them or code for proteins which are toxic. This is in contrast to another effort to clone *P. falciparum* genes: in this case only five of 23 genes contained mutations, and none contained frame shifts [5]. Whether this was because only intron-free genes were employed, the 3D7 strain had been passaged fewer times, or the expression vectors used had less leaky expression, and hence, less selective pressure on the *E. coli*, is unclear. *P. falciparum* targets as a group did experience a lower rate of expression than targets picked from other protozoan pathogens with a more balanced AT content, such as *L. major* and *T. brucei* (unpublished observation). It should be noted that while, following multivariate analysis, the %AT fell out as a factor affecting expression in *E. coli*, these *P. falciparum* targets as a whole were all quite high in %AT, and differences in %AT may not mean as much for a group which already exceeds a critical %AT threshold.

A related issue is that of codon bias, where the triplet codons are considered instead of the overall A and T percentages. It is commonly held that the codon bias of *P. falciparum* genes is an impediment to heterologous expression in *E. coli* [3,32]. Various approaches to circumventing this have been utilized, including the use of plasmids bearing extra tRNAs for the rare codons and the generation of synthetic genes which are codon-optimized for *E. coli*. Evidence has accumulated for the utility of both of these solutions, although this evidence has remained limited due to the small number of genes attempted. The RIG plasmid, which encodes tRNAs for arginine, isoleucine, and glycine, has been demonstrated to boost the expression levels of several *P. falciparum* proteins in *E. coli* [32]. Several codon-enhanced *E. coli* cell lines are commercially available and are commonly used for the expression of *P. falciparum* genes, although more recent work suggests that the stability and secondary structure of the encoded RNA might be more of an issue than the codons themselves [33]. The creation of synthetic genes allows for the optimization of every codon and elimination of the long adenosine and thymidine repeats in *P. falciparum* genes; this approach boosted the production of dihydrofolate reductase–thymidylate synthase (DHFR-TS) 10-fold [34]. Full-length merozoite surface protein 1 (MSP-1), which holds promise as a vaccine candidate, could only be made in *E. coli* after an artificial gene was generated [35]. Enoyl-acyl carrier reductase was not expressed well in *E. coli* until a stretch of 10 adenosines was modified with silent mutations [36]. Given that among the 1000 targets tested here there was so little apparent effect of codon bias, it stands to reason that there would be little benefit to using codon-optimized *E. coli* strains. The notorious AGG (arginine) codon which has been implicated in expression problems in the past [37] appeared to have no effect on these targets ($p = 0.61$, Table 4) despite being used in the *P. falciparum* targets 3.6× more often than in *E. coli*. It is apparent that whole gene synthesis (WGS) can help to address the troublesome gene structure of *P. falciparum*. It must be recognized, however, that WGS was successful for only three of 12 targets tested in this study, and even if a non-expressing gene is made to express there is still no guarantee of solubility; none of the three that were expressed were soluble. Many of these unexpressed targets were quite large (Table 5), so the bar was set fairly high for obtaining soluble protein from these constructs. The fact that WGS worked for any of these is testimony to the contribution of *P. falciparum* genetic structure to difficulties in expression. There are several cases of codon-optimized *P. falciparum* genes which expressed at higher levels in *E. coli* [34,38,39], but there have been no large-scale studies for *P. falciparum* proteins to the best of our knowledge. For high-value targets which do not express, WGS would certainly be among the routes worth trying.

In this set of 1000 targets, there was a high degree of correlation between protein disorder and problems in soluble expression (Table 3, Fig. 3). This has been observed in other large-scale expression efforts as well [40], leading to suggestions that highly disordered targets should simply be excluded from structural genomics pipelines altogether [41]. SEG, which uses a mathematical analysis of amino acid compositional complexity, was better at predicting difficulties in expression and solubility than was DISOPRED2, which utilizes a predictive algorithm based upon disordered regions of known crystal structures. Given that the target set is biased towards those proteins with no known crystal structure, it may be that the DISOPRED2 was less well suited to these targets. SEG disorder percentage was also more closely correlated with protein size (Fig. 3C). Thus the higher predictive value of SEG disorder percentage than DISOPRED2 disorder percentage for protein expression (Fig. 3A) may be due to the closer relationship of SEG to molecular weight.

Some methodologies which have been applied to the production of *P. falciparum* proteins in *E. coli* include refolding of insoluble aggregates and the use of solubility-enhancing fusion partners. Plasmepsins I and II, aspartic proteases from the food vacuole of *P. falciparum*, are refolded as part of their heterologous expression and isolation [42,43]. Refolding has also played an important role in the generation of dihydrofolate reductase mutants [44] and falcipain-2 [4]. The use of fusion partners to enhance the solubility of target proteins does have some notable success stories, although cleavage of the fusion partner can be troublesome or lead to precipitation of one's target

protein [45]. *P. falciparum* falcipain-2 has been produced in *E. coli* both via a refolding protocol and as a fusion with maltose binding protein (MBP), and although cleavage of the MBP severely reduced the yield, the falcipain-2 retained good enzymatic activity even without cleavage [46]. Interestingly, fusions to glutathione-s-transferase (GST) and thioredoxin (TRX) were ineffective in promoting the solubility of falcipain-2. In another study, GST was fused to *P. falciparum* erythrocyte membrane protein domains in order to enhance their soluble expression [3]. Clearly, though, a universal answer has not been found.

The use of alternative hosts for *P. falciparum* protein expression has been shown to be an effective route for many targets, especially vaccine candidates. A wide variety of systems have been employed, from baculovirus (insect cells) [47–49] and yeast [50,51] to secretion in the milk of transgenic mice [52]. All of these systems are considerably more technically demanding than *E. coli*, and no rigorous, comparative studies have been done to show which of these systems is most likely to be effective. The C-terminal fragment of merozoite surface protein could be expressed in either yeast or baculovirus systems, but expression in *E. coli* could only be obtained as an MBP-fusion [53]. The baculovirus system has probably had the most widespread use as an alternative to *E. coli*, although the AT-rich genome structure of *P. falciparum* may still pose problems, and a more effective approach might well be to employ artificial genes in baculovirus. In addition to addressing the AT-richness of the genes, the whole gene synthesis would also allow one to mutate glycosylation sites, as *P. falciparum* proteins generated in baculovirus systems to have been shown to be overglycosylated [54]. Our test of seventeen proteins in a baculovirus system was disappointing, as only one of these targets produced sufficiently high levels of protein for structural study from a 0.5 l culture, but the native gene sequence was used for these and rigorous optimization was not performed.

3.1. Comparison to other large-scale screening and expression efforts

In many ways, the workflow paralleled that of other structural genomics protein expression groups working on prokaryotic systems. However, the considerably higher difficulty associated with expression of these eukaryotic systems required some steps not employed by these other groups. We gel purified all PCR products, for example, and all SGPP plasmids were cloned twice, first in recombinase minus (*Rec*-) *E. coli* used for expansion of the plasmid and second in an expression (BL-21) strain. This double cloning was done in order to ensure that the plasmid was generated in the absence of selective pressure from leaky expression. Groups working on prokaryotic systems are typically able to move from PCR to expression without cloning steps. Other features of our pipeline mirrored other structural genomics groups quite closely: small-scale expression trials were done in 96-well blocks, expression was induced automatically through the use of Studier autoinduction media [9], and purification was done by nickel chromatography followed by size-exclusion chromatography. Importantly, the general philosophy of not dwelling on any particular target was also shared: the approach was a once-through, single-condition trial. A single clone was picked for each target, and those that screened positively for soluble expression were moved forward. Those that failed were quickly abandoned.

Sixty-three targets out of a total of 1000 (6.3%) is a remarkably low rate of soluble expression, but it is comparable to expression rates of other eukaryotic genes in *E. coli* [45,55–58]. For the sake of comparison, it is important to remember that these 63 were proteins which provided 0.9 mg or more of purified protein from a liter of culture, not those which simply screened soluble by ELISA or Western blot. The only other published study of a large number of *P. falciparum* proteins expressed this way obtained five soluble proteins from 95 single-exon constructs fused to MBP (maltose binding protein) or GST (glutathione-S-transferase) (5.2%) [5]. Interestingly, these authors employed the Gateway cloning system, and despite picking four colonies per target and mostly going after single-exon proteins, they were able to do only slightly better in cloning these genes than our LIC cloning, 85% as opposed to 79%. A recent paper detailing the expression screening of 10,167 random *C. elegans* genes found a solubility rate of 15.1% by ELISA [55]; the final percentage providing useful levels of protein upon scaleup will almost certainly be lower. In a study of 50 human proteins selected for relatively small size and low hydrophobicity, four (8%) could be obtained in sufficient yield to be purified[56]. Research conducted on 30 proteins selected from a variety of mammalian species found three (10%) which appeared to have 0.9 mg/l expression levels or greater by Western blot when expressed with a simple 10× His tag [57]. This number was raised to 17 (56%) when an MBP-fusion was employed, underscoring the utility of fusion partners when they will not interfere with downstream applications. A similar study using 32 human proteins with a 6× His tag yielded no pure proteins under non-denaturing conditions but considerably better yields with GST and MBP tags [58]. Likewise, an effort to clone and express 135 small proteins from *Oryza sativa* (rice) obtained a 53% solubility rate with the use of fusion proteins; about half of these proteins were not stable, however, after the fusion partner was cleaved [45].

Genome scale efforts in protein production have been underway in several laboratories, and some factors repeatedly are associated with the ability to heterologously express a protein. A primary factor appears to be the size of the target protein, as this has been reported by several groups working in both prokaryotic and eukaryotic systems [8,23,57]. For this reason, it is commonplace for structural genomics groups to bias their target selection towards smaller targets.

The strong relationship we observed between protein isoelectric point and insolubility (Fig. 1, Table 3) is one which has not received a lot of attention, although a high isoelectric point has been shown to be a negative predictor of crystallization success [40]. pI has also came up as a node in a decision tree analysis of structural genomics [23], although the weight of pI for this set of largely prokaryotic proteins was not nearly as great as what has been observed for these *P. falciparum* targets. Although the number of targets included in the tree-based analysis was impressive (over 27,000 targets from over 120 organisms), it was not possible for the authors to tease apart targets which had actually failed

from those which were still undergoing trials. This analysis suggested that the number of serines in a protein was correlated with insolubility; amongst the *P. falciparum* proteins, no such correlation was observed (data not shown). The correlation between pI and *P. falciparum* protein failure in expression and solubility was one of the strongest relationships in this study, and although a mechanistic explanation is readily apparent, this relationship might be indicative of a fundamental biological mechanism in *P. falciparum* or *E. coli* and is certainly worthy of further research.

The observation that well-known proteins are more likely to be solubly expressed (e.g. Pfam score greater than zero, non-hypothetical annotation) has been correlated with a positive outcome in other structural genomics programs [23,55]. Given that Pfam is a multiple-alignment based tool which is employed in the annotation of sequenced genomes, Pfam and annotation are very closely related and share most of their targets (over 80% of the annotated targets have a Pfam score greater than zero). While it stands to reason that proteins which can be solubly expressed in *E. coli* are likely to have been better characterized, it should also be recognized that targets which were annotated as membrane proteins were routinely excluded from our pipeline, and it is likely that some membrane proteins were annotated as hypothetical but were not predicted to be membrane-associated. It is also to be expected that among the hypothetical proteins will be members of protein complexes which require proper protein partners for solubility. Some may also be pseudogenes which are not actually expressed in the parasite. A related quantity is the percent identity to an *E. coli* homologue, a factor which one might expect to be highly correlated with annotation and Pfam hits. Surprisingly, about half of the *P. falciparum* targets with *E. coli* homologues were not represented either with an annotation or with Pfam hits. Of the 494 proteins with *E. coli* homologues, 43% were annotated hypothetical and 50% had Pfam score of zero. Of the 275 proteins with *E. coli* homologues and having a hypothetical annotation or a Pfam score of zero, the expression rate was 33.8%, almost exactly the same as the 33.7% expression rate observed for the 1000 targets as a whole. The presence of a homologue in *E. coli* was a better predictor of expression than either Pfam score or annotation (Table 3A), although once a protein was expressed, a non-hypothetical annotation was a better predictor of solubility (Table 3B).

The *C. elegans* structural genomics project demonstrated a high degree of correlation between increasing hydrophobicity (GRAVY) and insolubility [55]. A similar relationship was observed for these *P. falciparum* proteins only in univaraite analysis of soluble expression (Table 2); the relationship was not nearly as strong as that observed for *C. elegans* and dropped out completely during multivariate analysis. Dyson et al. [57] did not see this trend with a panel of 125 eukaryotic proteins; this may be because the sample size was smaller or because, as the authors suggest, the *C. elegans* experiment contained membrane proteins, a class which for which efforts were made to exclude in both the Dyson et al. study and in ours.

The low level of successful expression with eukaryotic targets is in stark contrast to that seen with targets from prokaryotic systems. Eight hundred and fifty *B. subtilis* targets were found to provide soluble protein at a rate of 44% [8]. One thousand eight hundred and seventy-seven targets from *T. maritima*, representing almost the entire genome, were cloned and expressed with 29% soluble expression [40]. Although these numbers are a great deal higher than those obtained from eukaryotic systems, most of the proteins still failed in expression or solubility, underscoring the fundamental limitations of *E. coli*-based expression systems even for prokaryotic targets.

3.2. Combining factors for soluble expression prediction

In structural genomics programs, it is fairly standard to use filters in target selection to enrich the target population for those targets which are most likely to solubly express and crystallize. These filters need not be based upon a causal relationship; as long as there is a positive association between a particular factor and soluble expression then the filter can be useful. Given the interrelatedness of factors such as gene sequence and amino acid sequence, size and disordered loops, etc. it may not be possible to extract causation from even strong correlations. It is unclear, too, whether data derived from "low hanging fruit" will apply to a wider target set. One of the great achievements of this pilot phase of the protein structure initiative has been the development of high-throughput cloning and screening methodology; these techniques reduce the need to pre-select targets for likely soluble expression.

Given that cloning and screening are so rapid, it makes more sense to exclude only those targets which fall into categories which almost always fail. For example, of the 243 targets attempted with a pI greater than 9.5, only one was solubly expressed (0.4%). No soluble proteins were obtained from the 77 targets with SEG scores of 29% or more. And only two soluble proteins were recovered from the 119 targets attempted of over 56 kDa in mass (1.7%). If one simply excluded these three categories, 60 soluble proteins would have been obtained from 617 total targets (9.7% soluble); only three of the targets obtained from the full set of 1000 would have been missed. While this is certainly more efficient, it should be noted that among the 90 targets chosen due to the existence of patented inhibitors, and thus with a certain degree of biomedical relevance, the median size was over 75 kDa (data not shown). The sobering bottom line is that those targets which are the most medically important are likely to be among those which are the most difficult to heterologously express. In addition, it is worth adding that crystallization and structure determination are by no means guaranteed to follow from purified protein: of the 63 proteins obtained in this study, a total of 11 have thus far yielded to structural determination.

Most efforts to clone and express *P. falciparum* proteins will be motivated by interest in a particular protein or pathway, and this kind of study will not have the luxury of excluding potentially troublesome cases. The future for heterologous expression of these proteins in *E. coli* might well reside in protein engineering; one could conceivably excise disordered regions and reduce the mass and the pI of the protein, all of which would be predicted to enhance the chances of soluble expression. While these steps are more difficult for proteins of unknown structure, as more structures are determined from more organisms our ability

to engineer the proteins from troublesome organisms should be enhanced. In much the same manner as the genome sequencing projects, the elucidation of information from one organism can empower investigations into others.

4. Materials and methods

4.1. Cloning

RNA from erythrocytic stage *P. falciparum* strain 3D7 (strain provided by Dr. P.K. Rathod, University of Washington) was extracted using the RNAqueous kit (Ambion) and reverse transcribed into cDNA using Superscript II (Invitrogen) as per the manufacturers' instructions. PCR was employed to amplify the target using the following primers:

Fwd primer: CTCACCACCACCACCACCAT + target specific sequence.
Reverse primer: ATCCTATCTTACTCAC + target specific sequence.

Target specific sequence length was determined by adding bases until the Tm was equal to or greater than 68 °C, with G and C contributing 4° and A and T pairs contributing 2°.

Primer sequences were generated in batches using GelbPrime, a custom Perl script available from M. Gelb, University of Washington: gelb@u.washington.edu Primers were purchased from Invitrogen, shipped in water and stored at −80 °C.

PCR protocol:

7 s at 94 °C.
20 s at 94 °C, 10 s at 50 °C, 10 s at 37 °C, 4.5 min at 60 °C, five times.
20 s at 94 °C, 10 s at 50 °C, 10 s at 42 °C, 4.5 min at 60 °C, 25 times.
10 min at 60 °C, hold at 4 °C, one time.

Targets were arrayed on the plate in size ladders in order to facilitate their identification and size verification. PCR was done using the Expand High Fidelity PCR kit (Roche) on a PTC 200 (MJ Research) thermal cycler, and the ramp speed was set to 70% between all steps except those between 60 and 94 °C, which were set at the maximum speed. The amplified PCR product was purified by agarose gel electrophoresis, visualized with Sybr Green (Invitrogen), extracted using a QiaQuick 96 kit (Qiagen), and spliced into BG1861 [59], a modified pET14b vector which appends MAHHHHHH onto the N-terminus of the protein, by ligation independent cloning (LIC) [60] using T4 polymerase (Novagen). Light exposure of the PCR products was minimized, and UV light was avoided altogether. HT-96 *E. coli* (Novagen) were transformed, and plasmids were extracted with a QiaPrep 96 Turbo kit (Qiagen) following overnight growth of a single colony inoculated into 600 μl of Terrific Broth (TB) with 100 μg/ml ampicillin and carbenicillin. This plasmid was used to transform BL-21 star cells (Invitrogen) in cloning grills [61], and a single colony was expanded in 600 μl of TB media in each well of a 96-well block for 6 h. This "inoculator block" was used to inoculate a second "expression block" filled with autoinduction media, [9] and this expression block was shaken overnight. In the morning, in order to ensure induction for cultures which did not get very dense, IPTG was added to 1 mM, and the expression block was shaken for 6 h at room-temperature. Immediately after inoculation, 60 μl DMSO was added to each well of the inoculator block, and the block was stored frozen at −80 °C. Following the 6 h, room-temperature growth of the expression block, the *E. coli* were pelleted by centrifugation, the media was discarded, and the pellets were sonicated in the sonication buffer detailed below using a custom-built robotic sonicator. These sonicates were cleared by centrifugation, 6× His protein was captured on Ni-NTA Superflow nickel resin (Qiagen), and the presence of soluble protein was assessed by running the supernatant on Criterion 4–12% SDS-PAGE gels (Bio-Rad) and staining with GelCode Blue colloidal Coomassie stain (Pierce). Insoluble expression was determined by solubilizing the pellets in 8 M urea and running these on SDS-PAGE gels. Colonies that showed any evidence of soluble expression by SDS-PAGE were inoculated into a liter of ZYP-5052 autoinduction media via an ice stab from the inoculator block; cultures were grown for 15 h at 37 °C followed by an overnight incubation at 18 °C. *E. coli* were harvested by centrifugation, frozen in liquid nitrogen and stored at −80 °C. The pellet was then resuspended in standard buffer: 25 mM HEPES (pH 7.25), 500 mM NaCl, 5% glycerol, and 0.025% sodium azide to which was added 0.1% cholate, 1 mg/ml lysozyme (Sigma), 1 mM 2ME, protease inhibitors (Roche Complete, EDTA-free), and 750 Units benzonase (Sigma) and sonicated on ice to disrupt *E. coli*. Cellular debris was removed by 20 min of centrifugation at 18,000 × *g*, and the supernatant was tumbled with 5 ml of nickel-NTA resin (Superflow NTA, Qiagen) for 45 min at 4 °C. The resin was allowed to settle, the supernatant discarded, and the resin then rinsed with standard buffer containing 10 mM imidazole and once with standard buffer containing 20 mM imidazole. The resin was then recovered, added to a disposable column and rinsed with 20 and 50 mM imidazole in standard buffer, the protein was eluted with 15 ml of 250 mM imidazole in standard buffer, and the eluent was dialyzed against 4 l of standard buffer overnight at 4 °C. The dialyzed material was concentrated to 10 ml by centrifugal ultrafiltration (Amicon Ultra), DTT was added to 1 mM, and it was then applied to a pre-packed Superdex 75 26/60 gel chromatography column (Amersham Biosciences) at 4 °C in standard buffer. After running at 1 ml/min, peak fractions were collected and pooled, protease inhibitors (Roche Complete, EDTA free) added, and the solution was concentrated to 10–20 mg/ml in standard buffer with 2 mM DTT. Protein was aliquotted into a PCR plate, flash frozen in liquid nitrogen, and stored at −80 °C [62]. Protein concentration was determined using the BioRad protein assay system and a standard curve of bovine serum albumin.

4.2. Statistical analysis

The goal was to determine whether a set of biological characteristics, defined prior to study start, influenced protein expression or solubility. Protein characteristics included molecular

weight, pI, and disorder, measured by both SEG total and disopred percent. pI was calculated from the pK_a of the side chains and termini of the proteins, and no position- or folding-dependent context was taken into account in these calculations. Gene characteristics included percentage of nucleotides which were A or T, number of introns, and largest consecutive stretch of either A or T bases. Also included in gene characteristics was codon usage, described by the percentage of amino acids encoded by one of the following rare codons: ATA, AGA, AAT, AGG, TTA, TAT, ACA, TGT, and AAA. Three other characteristics, based on previously established information, include Pfam size, annotation and percentage *E. coli* similarity.

Initially, the relationship between each variable and outcome was analyzed separately. To allow for flexible, non-linear, associations, all continuous and ordinal variables were transformed into categorical variables. Molecular weight, pI, and A/T percentage were categorized by quartile. SEG total was separated into the four categories: 0, 1–50, 51–100, and >100. Number of introns was classified as 0, 1–2, or >2. The longest A or T stretch was classified as 0–6, 7–9, or >9. Pfam size was classified as 0, 1–100, 101–500, 501–1000, or >1000. For each variable, a χ^2-test was employed to test whether the percentage of protein expressed, the percentage soluble among all proteins, and the percentage soluble among expressed proteins differed across categories.

As many of the characteristics were expected to be highly correlated, multivariate logistic regression was used to establish associations independent of possible confounders. To determine whether the odds of an outcome had an approximately linear relationship with a continuous variable, ANOVA was used to analyze the variance reduction from adding the categorical form of the variable to a model with only the continuous form. If the *F*-statistic was not significant at the 0.05 level, the continuous form was included in all multivariate models. Models compared expressed versus not expressed and expressed/not soluble versus expressed/soluble. First, protein characteristics, gene characteristics, codon usage, and other characteristics were included in four separate models. *p* values were determined by the *F*-statistic. Second, any covariate significant at the 0.05 level was then included a new model, combining information across all four categories. These models determined one set of variables that are correlated with expression and one set correlated with solubility of expressed proteins. Each set, and any other characteristic significant when added to the set, was included in a final model. Proteins missing a covariate in a model were excluded. Multivariate logistic regression was performed with the s-plus function GLM (family: binomial). Statistics were determined using S-PLUS 6.0 (Insightful Corporation, Seattle, WA).

4.3. Synthetic genes

Synthetic genes were constructed by Blue Heron (Bothell, WA) and codon-optimized for *E. coli* expression. Genes were spliced into the same expression vector utilized for the standard expression pipeline and were sequenced in their entirety. At least three clones were grown and tested for solubility via the standard methodology.

4.4. Baculovirus expression

Baculovirus expression tests were conducted by Orbigen using their standard methodology (www.orbigen.com) in Sf-21 cells. Targets were selected which appeared to express insolubly in *E. coli*. Soluble expression screens were done by western blot, and cultures which appeared promising were scaled up to 0.5 l. Protein was purified in essentially the same manner as for *E. coli*.

Acknowledgements

This work was supported by NIH grant P50 GM64655 to WGJH. We would like to thank Dr. Elizabeth Grayhack of the University of Rochester for providing the cloning vector BG-1861 and Dr. Pradip K. Rathod of the University of Washington for *P. falciparum* strain 3D7. Thanks also to Kiley Nakamura and Micah Yospe for their help in the lab, to Kevin Bauer, Kasey Rivas, and Carrey Horney for preparing *P. falciparum* cDNA, to Martin Criminale for providing computer support and to Marissa Vignali and Douglas LaCount for the PCR protocol.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molbiopara.2006.03.011.

References

[1] Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. Nature 2005;434(7030):214–7.
[2] www.mmv.org, 2005.
[3] Flick K, Ahuja S, Chene A, Bejarano MT, Chen Q. Optimized expression of *Plasmodium falciparum* erythrocyte membrane protein 1 domains in *Escherichia coli*. Malar J 2004;3(1):50.
[4] 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(1):128–34.
[5] Aguiar JC, LaBaer J, Blair PL, et al. High-throughput generation of *P. falciparum* functional molecules by recombinational cloning. Genome Res 2004;14(10B):2076–82.
[6] Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002;419(6906):498–511.
[7] www.nigms.nih.gov/psi/.
[8] Moy S, Dieckman L, Schiffer M, Maltsev N, Yu GX, Collart FR. Genome-scale expression of proteins from Bacillus subtilis. J Struct Funct Genomics 2004;5(1–2):103–9.
[9] Studier FW. Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 2005;41(1):207–34.
[10] Terwilliger TC, Park MS, Waldo GS, et al. The TB structural genomics consortium: a resource for Mycobacterium tuberculosis biology. Tuberculosis (Edinb) 2003;83(4):223–49.
[11] Gowda DC, Davidson EA. Protein glycosylation in the malaria parasite. Parasitol Today 1999;15(4):147–52.
[12] Aravind L, Iyer LM, Wellems TE, Miller LH. Plasmodium biology: genomic gleanings. Cell 2003;115(7):771–85.
[13] Pizzi E, Frontali C. Low-complexity regions in *Plasmodium falciparum* proteins. Genome Res 2001;11(2):218–29.
[14] Singh GP, Chandra BR, Bhattacharya A, Akhouri RR, Singh SK, Sharma A. Hyper-expansion of asparagines correlates with an abundance of pro-

teins with prion-like domains in *Plasmodium falciparum*. Mol Biochem Parasitol 2004;137(2):307–19.
[15] Schneider EL, King DS, Marletta MA. Amino acid substitution and modification resulting from *Escherichia coli* expression of recombinant *Plasmodium falciparum* histidine-rich protein II. Biochemistry 2005;44(3):987–95.
[16] Zhang K, Rathod PK. Divergent regulation of dihydrofolate reductase between malaria parasite and human host. Science 2002;296(5567): 545–7.
[17] Turgut-Balik D, Shoemark DK, Moreton KM, Sessions RB, Holbrook JJ. Over-production of lactate dehydrogenase from *Plasmodium falciparum* opens a new route to antimalarials. Biotechnol Lett 2001;23:917–21.
[18] Bahl A, Brunk B, Crabtree J, et al. PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. Nucleic Acids Res 2003;31(1):212–5.
[19] Kissinger JC, Brunk BP, Crabtree J, et al. The Plasmodium genome database. Nature 2002;419(6906):490–2.
[20] Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18(3): 269–85.
[21] Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinformatics 2004;20(13):2138–9.
[22] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157(1):105–32.
[23] Goh CS, Lan N, Douglas SM, et al. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J Mol Biol 2004;336(1):115–30.
[24] Saul A, Battistutta D. Codon usage in *Plasmodium falciparum*. Mol Biochem Parasitol 1988;27(1):35–42.
[25] 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(9):345–6.
[26] Singer GA, Hickey DA. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 2000;17(11): 1581–8.
[27] Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. Nucleic Acids Res 2004;32(Database issue):D138–41.
[28] Yokoyama S. Protein expression systems for structural genomics and proteomics. Curr Opin Chem Biol 2003;7(1):39–43.
[29] Schwartz R, Ting CS, King J. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res 2001;11(5):703–9.
[30] Joachimiak A. Establishing a high-throughput platform for structural genomics using synchrotron-based X-ray crystallography. In: Proceedings of the International Conference on Structural Genomics (ICSG). 2004.
[31] Xue HY, Forsdyke DR. Low-complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. Mol Biochem Parasitol 2003;128(1):21–32.
[32] 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(2):113–8.
[33] Wu X, Jornvall H, Berndt KD, Oppermann U. Codon optimization reveals critical factors for high level expression of two rare codon genes in *Escherichia coli*: RNA stability and secondary structure but not tRNA abundance. Biochem Biophys Res Commun 2004;313(1):89–96.
[34] Prapunwattana P, Sirawaraporn W, Yuthavong Y, Santi DV. Chemical synthesis of the *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase gene. Mol Biochem Parasitol 1996;83(1): 93–106.
[35] Pan W, Ravot E, Tolle R, et al. Vaccine candidate MSP-1 from *Plasmodium falciparum*: a redesigned 4917 bp polynucleotide enables synthesis and isolation of full-length protein from *Escherichia coli* and mammalian cells. Nucleic Acids Res 1999;27(4):1094–103.
[36] Perozzo R, Kuo M, Sidhu AS, et al. Structural elucidation of the specificity of the antibacterial agent triclosan for malarial enoyl acyl carrier protein reductase. J Biol Chem 2002;277(15):13106–14.
[37] Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol 1995;6(5):494–500.
[38] Zhou Z, Schnake P, Xiao L, Lal AA. Enhanced expression of a recombinant malaria candidate vaccine in *Escherichia coli* by codon optimization. Protein Expr Purif 2004;34(1):87–94.
[39] Yadava A, Ockenhouse CF. Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems. Infect Immun 2003;71(9):4961–9.
[40] Canaves JM, Page R, Wilson IA, Stevens RC. Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics. J Mol Biol 2004;344(4):977–91.
[41] Brenner SE. Target selection for structural genomics. Nat Struct Biol 2000;7(Suppl.):967–9.
[42] Moon RP, Tyas L, Certa U, et al. Expression and characterisation of plasmepsin I from *Plasmodium falciparum*. Eur J Biochem 1997;244(2):552–60.
[43] Gulnik SV, Afonina EI, Gustchina E, et al. Utility of (His)6 tag for purification and refolding of proplasmepsin-2 and mutants with altered activation properties. Protein Expr Purif 2002;24(3):412–9.
[44] Sirawaraporn W, Yongkiettrakul S, Sirawaraporn R, Yuthavong Y, Santi DV. *Plasmodium falciparum*: asparagine mutant at residue 108 of dihydrofolate reductase is an optimal antifolate-resistant single mutant. Exp Parasitol 1997;87(3):245–52.
[45] Tsunoda Y, Sakai N, Kikuchi K, et al. Improving expression and solubility of rice proteins produced as fusion proteins in *Escherichia coli*. Protein Expr Purif 2005;42(2):268–77.
[46] Goh LL, Loke P, Singh M, Sim TS. Soluble expression of a functionally active *Plasmodium falciparum* falcipain-2 fused to maltose-binding protein in *Escherichia coli*. Protein Expr Purif 2003;32(2):194–201.
[47] Jean L, Hackett F, Martin SR, Blackman MJ. Functional characterization of the propeptide of *Plasmodium falciparum* subtilisin-like protease-1. J Biol Chem 2003;278(31):28572–9.
[48] Li J, Matsuoka H, Mitamura T, Horii T. Characterization of proteases involved in the processing of *Plasmodium falciparum* serine repeat antigen (SERA). Mol Biochem Parasitol 2002;120(2):177–86.
[49] Yuda M, Yano K, Tsuboi T, Torii M, Chinzei Y, von Willebrand. Factor A domain-related protein, a novel microneme protein of the malaria ookinete highly conserved throughout Plasmodium parasites. Mol Biochem Parasitol 2001;116(1):65–72.
[50] Zhang H, Howard EM, Roepe PD. Analysis of the antimalarial drug resistance protein Pfcrt expressed in yeast. J Biol Chem 2002;277(51):49767–75.
[51] Brady CP, Shimp RL, Miles AP, Whitmore M, Stowers AW. High-level production and purification of P30P2MSP1(19), an important vaccine antigen for malaria, expressed in the methylotropic yeast Pichia pastoris. Protein Expr Purif 2001;23(3):468–75.
[52] Stowers AW, Chen LH, Zhang Y, et al. A recombinant vaccine expressed in the milk of transgenic mice protects Aotus monkeys from a lethal challenge with *Plasmodium falciparum*. Proc Natl Acad Sci USA 2002;99(1):339–44.
[53] Planson AG, Guijarro JI, Goldberg ME, Chaffotte AF. Assistance of maltose binding protein to the in vivo folding of the disulfide-rich C-terminal fragment from *Plasmodium falciparum* merozoite surface protein 1 expressed in *Escherichia coli*. Biochemistry 2003;42(45):13202–11.
[54] Kedees MH, Azzouz N, Gerold P, et al. *Plasmodium falciparum*: glycosylation status of *Plasmodium falciparum* circumsporozoite protein expressed in the baculovirus system. Exp Parasitol 2002;101(1):64–8.
[55] Luan CH, Qiu S, Finley JB, et al. High-throughput expression of *C. elegans* proteins. Genome Res 2004;14(10B):2102–10.
[56] Ding HT, Ren H, Chen Q, et al. Parallel cloning, expression, purification and crystallization of human proteins for structural genomics. Acta Crystallogr D Biol Crystallogr 2002;58(Pt. 12):2102–8.
[57] Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. BMC Biotechnol 2004;4(1):32.

[58] Braun P, Hu Y, Shen B, et al. Proteome-scale purification of human proteins from bacteria. Proc Natl Acad Sci USA 2002;99(5):2654–9.
[59] Alexandrov A, Vignali M, LaCount DJ, et al. A facile method for high-throughput co-expression of protein pairs. Mol Cell Proteomics 2004;3(9):934–8.
[60] Aslanidis C, de Jong PJ. Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res 1990;18(20):6069–74.
[61] Mehlin C, Boni EE, Andreyka J, Terry RW. Cloning grills: high throughput cloning for structural genomics. J Struct Funct Genomics 2004;5(1–2):59–61.
[62] Deng J, Davies DR, Wisedchaisri G, Wu M, Hol WG, Mehlin C. An improved protocol for rapid freezing of protein samples for long-term storage. Acta Crystallogr D Biol Crystallogr 2004;60(Pt. 1): 203–4.