Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular & Biochemical Parasitology 143 (2005) 67–79
MOLECULAR & BIOCHEMICAL PARASITOLOGY

The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification

Jason A. Young[a,1], Quinton L. Fivelman[b,1], Peter L. Blair[c,1,3], Patricia de la Vega[c], Karine G. Le Roch[d], Yingyao Zhou[d,2], Daniel J. Carucci[c,4], David A. Baker[b], Elizabeth A. Winzeler[a,d,*]

[a] *Department of Cell Biology ICND202, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA*
[b] *Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK*
[c] *Malaria Program, Naval Medical Research Center, Silver Spring, MD 20910-7500, USA*
[d] *Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121, USA*

Received 8 February 2005; received in revised form 29 April 2005; accepted 20 May 2005
Available online 17 June 2005

Abstract

The sexual stages of malarial parasites are essential for the mosquito transmission of the disease and therefore are the focus of transmission-blocking drug and vaccine development. In order to better understand genes important to the sexual development process, the transcriptomes of high-purity stage I–V *Plasmodium falciparum* gametocytes were comprehensively profiled using a full-genome high-density oligonucleotide microarray. The interpretation of this transcriptional data was aided by applying a novel knowledge-based data-mining algorithm termed ontology-based pattern identification (OPI) using current information regarding known sexual stage genes as a guide. This analysis resulted in the identification of a sexual development cluster containing 246 genes, of which ~75% were hypothetical, exhibiting highly-correlated, gametocyte-specific expression patterns. Inspection of the upstream promoter regions of these 246 genes revealed putative *cis*-regulatory elements for sexual development transcriptional control mechanisms. Furthermore, OPI analysis was extended using current annotations provided by the Gene Ontology Consortium to identify 380 statistically significant clusters containing genes with expression patterns characteristic of various biological processes, cellular components, and molecular functions. Collectively, these results, available as part of a web-accessible OPI database (http://carrier.gnf.org/publications/Gametocyte), shed light on the components of molecular mechanisms underlying parasite sexual development and other areas of malarial parasite biology.

Keywords: *Plasmodium falciparum*; Gametocyte; Sexual development; Microarray; Ontology-based pattern identification; Gene ontology

1. Introduction

Malaria continues to be a devastating infectious disease, responsible for approximately 500 million clinical episodes and millions of deaths each year worldwide [1]. Development of drugs and vaccines to combat the disease traditionally has focused heavily on the intraerythrocytic stages of the malarial parasite life cycle, as these stages are responsible for the clinical symptoms associated with the illness. However, in recent decades, it has become apparent that any successful strategy for controlling malaria will most likely require a multifaceted approach that also includes drugs and vaccines

Abbreviations: OPI, ontology-based pattern identification; GO, gene ontology; ANOVA, analysis of variance; MOID, match-only integral distribution; GBA, guilt-by-association; GlcNAc, *N*-acetyl glucosamine; PBS, phosphate-buffered saline; BSA, bovine serum albumin

* Corresponding author. Tel.: +1 858 784 9468; fax: +1 858 784 9860.
E-mail address: winzeler@scripps.edu (E.A. Winzeler).

[1] J.A.Y., Q.L.F., and P.L.B. contributed equally to this work.

[2] To whom specific correspondence regarding OPI should be addressed.

[3] Biology Department, Earlham College, 801 National Road West, Richmond, IN 47374, USA.

[4] Foundation for the National Institutes of Health, Natcher Building, 45 Center Drive (3An-44), Bethesda, MD 20892-6300, USA.

0166-6851/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.molbiopara.2005.05.007

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Plasmodium falciparum sexual development transcriptome: A microarray analysis using ontology-based pattern identification	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Cell Biology ICND202, The Scripps Research Institute, 10550 North Torrey Pines Road La Jolla, CA 92037**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **Naval Medical Research Center 503 Robert Grant Avenue Silver Spring, MD 20910-7500**	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**13**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

against other stages of the complex parasite life cycle. Consequently, the sexual stages of the parasite essential for the mosquito transmission of the disease are considered attractive targets for the development of new transmission-blocking drugs and vaccines that aim to prevent the spread of malaria in human populations [2,3]. Such strategies are thought to be especially promising, as it has been hypothesized that the parasite may be more vulnerable to vaccine and drug intervention during the sexual part of its life cycle due to its passage through a numerical bottleneck and the limited exposure to the human immune system it receives during these stages [3].

Of the four *Plasmodium* parasite species responsible for malaria in humans, the progression of sexual development is best understood morphologically for the most lethal species, *Plasmodium falciparum* [4,5]. The switch from an asexual to sexual mode of replication begins in the haploid intraerythrocytic stages, where a sub-population of asexual parasites begin to develop into male and female gametocytes. This process of gametocyte development continues in the human host over a period of approximately 10 days, encompassing five morphologically defined gametocyte stages (stage I–V) and ending with the formation of mature male and female gametocytes. When mature gametocytes are taken up by a mosquito as part of the bloodmeal the process of sexual development continues in the mosquito midgut, where male gametocytes undergo exflagellation to form highly motile male gametes and female gametocytes enlarge and emerge from red blood cells to form female gametes. The subsequent fusion of a single male and female gamete results in fertilization and the formation of a diploid zygote that then differentiates into a motile ookinete. It is this ookinete stage that transverses the mosquito midgut wall to form an oocyst, where asexual sporogonic development is once again initiated.

Despite this detailed morphological description, comprehensive understandings of most of the fundamental biological mechanisms driving the parasite sexual development process remain elusive. For example, we still do not fully understand how gene and protein expression is regulated during sexual development and what specific metabolic differences exist between the asexual and sexual stages. As a result, current transmission-blocking vaccine development has focused on only a handful of known sexual-specific proteins, and chemotherapeutics that selectively kill sexual stage parasites have yet to be discovered [2,3].

The availability of the full genome sequence for *P. falciparum* since 2002 has allowed for the investigation of gene functions in the malarial parasite using high-throughput genomic and proteomic approaches that circumvent some obstacles associated with the application of more traditional genetic and biochemical methods to malaria research [6–8]. Previously, we conducted a genome-wide transcriptional analysis of various stages of the *P. falciparum* life cycle using a high-density oligonucleotide microarray coupled with a *k*-means clustering approach to identify 15 groups of genes sharing mRNA expression patterns characteristic of various biological processes such as antigenic variation and cell invasion [9]. In an analogous study, Bozdech and co-workers utilized a spotted oligonucleotide array and Fourier transformation analysis to demonstrate that ~80% of the *P. falciparum* genes expressed during the asexual intraerythrocytic cell cycle exhibit a single-peak periodic pattern in their transcript levels, with genes involved in common biological processes also sharing similar phase [10]. Together, these studies provided many novel insights into potential functions for the approximately 3000 hypothetical genes in the *P. falciparum* genome. However, due to the inclusion of only a single late-stage gametocyte time point in the former work, insights provided by these studies regarding sexual stage gene function were limited.

To better understand genes involved in *P. falciparum* sexual development using an expression microarray approach, we have obtained new transcriptional data on detailed time courses of gametocyte development, including high-purity early-stage gametocytes. To aid in the interpretation of this new microarray data within the context of our previous asexual stage data [9], we applied a recently developed knowledge-based clustering algorithm called ontology-based pattern identification (OPI) [11]. OPI utilizes classifications of gene function in the form of systematic gene annotations to generate gene clusters with greater specificity and statistical confidence than is possible using more routinely-used clustering methods such as the *k*-means clustering approach employed in the past [7,9,10]. For example, hierarchical or *k*-means methods require the user to arbitrarily select a correlation coefficient, a fold-change, or a number of *k*-clusters to determine cluster sizes. In contrast, OPI empirically selects the best parameters to give the highest concentration of genes belonging to a particular classification in the smallest cluster size. Using manually-assigned gene annotations for sexual development to guide OPI, a cluster of 246 genes possessing highly gametocyte-specific mRNA expression patterns was identified. Furthermore, inspections of promoter regions upstream of these genes revealed conserved sequences that potentially play roles as *cis*-regulatory elements in sexual development transcriptional control mechanisms. Beyond sexual development, OPI was also extended to identify genes most likely to be involved in 380 biological processes, cellular components, and molecular functions using information provided by the Gene Ontology (GO) Consortium [12]. Collectively, these 381 clusters, accessible as part of an OPI database at http://carrier.gnf.org/publications/Gametocyte, provide new insights into *P. falciparum* gene function with direct implications for rational drug and vaccine development against malaria.

2. Materials and methods

2.1. Parasite cultivation

Early passages of *P. falciparum* clone 3D7 and isolate NF54 were cultured with human A^+ or O^+ erythrocytes,

respectively, as previously described [13]. Prior to induction of gametocyte development, the cultures underwent two rounds of synchronization with 5% D-sorbitol [14]. Synchronized sexual development was induced by a sudden increase to the hematocrit of a fast growing ring stage culture in the presence of partially-spent medium. Fresh erythrocytes were added to the media at the schizont stage on the following day (Day-1). *N*-Acetyl glucosamine (GlcNAc, 50 mM) was added to the media exchanged from Day 0 to eliminate residual asexual stages [5]. Both sets of cultures were maintained for 14 days to obtain stage I through stage V gametocytes, with NF54 GlcNAc treatment being withdrawn on Day 10. NF54 stage V gametocytes were subsequently fed to *Anopheles stephensi* mosquitoes on Day 15 to confirm infectivity. Parasite stages for all cultures were monitored using Giemsa-stained blood smears.

In order to obtain enriched early 3D7 gametocyte stages I and II, contaminating asexual stages were removed using CS magnetic affinity columns on Day 2 (Miltenyi Biotech, Germany). A mixed culture of erythrocytes, asexual ring stages, and early gametocytes were passed through a CS column. Stage I and II gametocytes containing greater amounts of hemozoin were retained on the column while the younger rings and red blood cells passed through. The column was separated from the magnet and the retained cells were eluted as an enriched, positively selected early-stage gametocyte cell fraction with a removal of ∼80% asexual parasites. The enriched gametocytes were returned to culture at 2% parasitemia in the presence of GlcNAc. Giemsa-stained blood smear cell counts of enriched gametocytes were confirmed using immunofluorescence cell counts with a gametocyte specific antibody. Parasite preparations were washed three times in phosphate-buffered saline (PBS; pH 7.4) containing 0.1% bovine serum albumin (BSA) and spotted on to each of 12 wells on Teflon-coated slides. Dried slides were fixed in acetone on ice for 30 min and stored at −20 °C. After removal from −20 °C, slides were washed in PBS, and blocked in PBS-0.1% BSA for 30 min at 22 °C in the dark. Anti-Pfs16 primary antibody (mouse monoclonal antibody 93A3A2) diluted 1:250 was applied in PBS-0.1% BSA for 2 h, the slides were then washed and fluorescence-labeled secondary antibody (goat anti-mouse immunoglobulin G conjugated to AlexaFluor 594, Molecular Probes, USA) diluted 1:400 was added for 1 h. The parasites were then given a final wash in PBS, the slides mounted in Vectashield with nucleic acid stain DAPI (Vector Laboratories, USA), and stored in the dark at 4 °C until visualized using fluorescence microscopy. Use of the anti-Pfs16 antibody allowed a comparison between the number of sexual parasites (Pfs16 fluorescent) and overall infected erythrocytes (DAPI fluorescent).

2.2. Microarray gene expression profiling

Total RNA was isolated on each day for 13 days starting on Day 1 for the NF54 parasites and on Days 1–3, 6, 8, and 12 for the 3D7 parasites as previously described [15]. For the purified 3D7 early-stage gametocyte samples, RNA was isolated on Day 1, immediately after CS column separation on Day 2, and for the following 2 days (Days 3 and 4). Labeling of the RNA and hybridization to a custom-designed *P. falciparum* full-genome high-density oligonucleotide array (Affymetrix, USA) containing 25 mer probes to 5159 *P. falciparum* genes was conducted as previously described [9]. The raw data files from this match-only microarray were analyzed using the Match-Only Integral Distribution algorithm (MOID) [16]. The background noise distributions for the MOID algorithm were calculated with the probe intensities of 100 negative control genes. Background was then subtracted from each raw probe intensity probabilistically according to the above distribution. During normalization, only probe sets having more than 10 probes and having a signal 1.5-fold higher than the 70 percentile of background distribution were taken into account. Normalization factors were determined by scaling the average intensity of genes in the range between 30 and 90 percentile to 200. All background levels and normalization factors are available at http://carrier.gnf.org/publications/Gametocyte.

2.3. Semi-quantitative RT-PCR

Three micrograms of total RNA, isolated from mixed asexual NF54 parasites and Day 7 NF54 gametocytes, was used for cDNA synthesis using superscript reverse transcriptase and oligo(dT)$_{12–18}$ (Invitrogen, USA). PCR using Extaq polymerase (Takara, Japan) was conducted on 1 μl of each cDNA template for 20 cycles using an annealing temperature of 56 °C with primers specific to three hypothetical genes (PF11_0214, PF14_0067, and MAL8P1.16) identified from the microarray data as expressed only during sexual development (PF11_0214: forward primer 5′-GTA ACG AAC AGG AGA TAA ATG GGA TTT GTA-3′, reverse primer 5′-GGG TAT TAC CAT AAC TTT TCA ATT TGT CTC G-3′; PF14_0067: forward primer 5′-GCA TAT TGT AAA GAA TGG TGT AAA GCC ACG-3′, reverse primer 5′-CCA GCA CTA AAT CCT TCA TGT AAT ACT TTA G-3′; MAL8P1.16: forward primer 5′-GGT GTT GTA TCA TCT GAA AAT ATT AAG CTC C-3′, reverse primer 5′-CTT AGC CTT TCT TGT TAA ATT CTT CCA AGC-3′) and the known asexual-stage-specific gene *eba-175* (PF07_0128) as an asexual stage positive control (PF07_0128: forward primer 5′-TCA TAG TCA TCA TGG AAA CAG ACA AGA TCG-3′, reverse primer 5′-GTA AAA TAG CTC ATA CAG TAA TCT GAT ACT GC-3′). Whenever possible, primers were designed such that they spanned an intron to allow for the detection of any possible contaminating genomic DNA template. For the genomic DNA positive control reactions, approximately 100 ng genomic DNA isolated from mixed asexual stage NF54 parasites was used as template.

2.4. Ontology-based pattern identification

For a detailed description of OPI, please refer to Zhou et al. [11]. In this study, 38 microarray expression data sets

from various stages of the *P. falciparum* life cycle were analyzed concurrently. These included the newly obtained data on thirteen NF54 gametocyte development stages, six 3D7 gametocyte development stages, and four 3D7 gametocyte development samples enriched for early-stage gametocytes, in addition to previously published data on seven sorbitol and seven temperature synchronized 3D7 erythrocytic stages (early ring, late ring, early trophozoite, late trophozoite, early schizont, late schizont, and merozoite), and one 3D7 sporozoite stage [9]. From here, OPI analysis proceeded conceptually as follows: (i) 3059 differentially expressed genes were identified using the criteria of P_{ANOVA} less than 0.2 and fold change in expression greater than 1.5 across the life cycle stages sampled; (ii) gene annotations for functionally classified genes were obtained from the GO Consortium [12] as of October, 2004, with the sole exception being the sexual development annotation (GO:GNF0004), which was assigned manually to 15 genes based on a review of the current literature; (iii) for each classification category, all genes sharing the same annotation were identified; (iv) the gene with the expression pattern that best correlated with all other genes of the same classification based on Pearson's correlation coefficients was identified and used as the reference gene; (v) all other differentially expressed genes were then ranked based on the correlation coefficients between their life-cycle-dependent expression profiles and that of the reference gene; (vi) the gene that had the highest correlation with the reference gene and also shared the same classification was identified and a cluster was created that included both of these genes as well as any other genes that did not share the same annotation, but fell in between these two annotated genes in the sorted gene list; (vii) a *p*-value was assigned to this cluster based on the hypergeometric distribution representing the probability that these two genes in a cluster of size N genes would share the same classification by chance; (viii) this cluster was then expanded to include all the genes in the sorted gene list up to the next one possessing the same classification as the reference gene; (ix) a *p*-value was assigned to this new cluster just as before using the hypergeometric distribution; (x) this process of cluster expansion and *p*-value assignment (steps viii and ix) was repeated until all genes sharing the annotation were grouped together yielding the largest possible cluster for this classification; (xi) the cluster with the lowest *p*-value was identified, containing the highest concentration of genes belonging to a particular classification in the smallest cluster size. The entire process was also repeated on log-transformed expression data, for several filtering P_{ANOVA} cut-offs, and using the average expression pattern of all genes in each classification category instead of the single reference gene in the ranking step. The process that yielded the cluster with the minimal hypergeometrical *p*-value for each classification category was chosen as the best method to describe the category. In order to identify those clusters that were statistically significant, the entire process was then repeated 100 times on 100 permutated datasets where the associations between gene annotations and the gene expression profiles were randomized. Statistically significant clusters were defined as those having *p*-values less than 90% of the corresponding values in the 100 permutation runs.

2.5. GBSSR

For a detailed description of GBSSR, please refer to Kreps et al. [17]. In this study, 1000 bootstraps were used to establish the frequency distribution of all eligible 6–12 base pair strings in the 1000 base pair upstream regions of all genes in the genome. Comparison of this genome-wide string frequency distribution to the frequency of strings in the 1000 bases upstream regions of the 246 genes in the OPI sexual development cluster identified those overrepresented sequences in the sexual development gene promoter regions. These sequences were considered putative *cis*-regulatory elements for parasite sexual development transcriptional control mechanisms.

3. Results

3.1. Gametocyte cultivation

Two 14-day time courses of in vitro gametocyte development were obtained, one using isolate NF54 and one using clone 3D7. Progression and purity of gametocyte development from stage I to stage V was monitored by Giemsa-stained blood smears (Fig. 1 and Fig. S1). Although overall gametocyte development was the same for both time-courses, a slight difference in the rate of development was observed between NF54 and 3D7 with NF54 stage III and stage IV parasites peaking on Days 4 and 6, respectively, approximately 2 days before 3D7, which peaked for stage III and stage IV parasites on Days 6 and 8, respectively. Despite taking measures to standardize gametocyte production for both time courses, this discrepancy could be due to biological differences inherent to NF54 and 3D7 or differences in

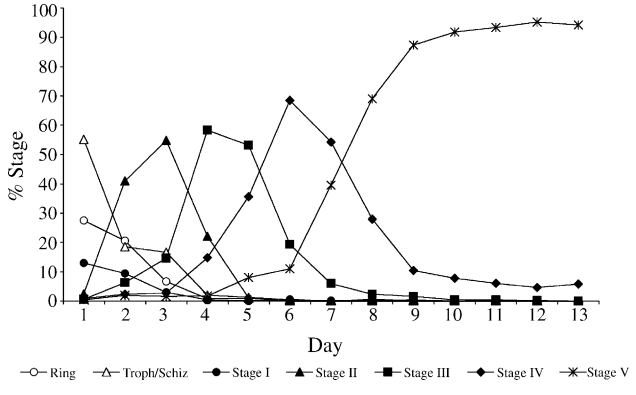
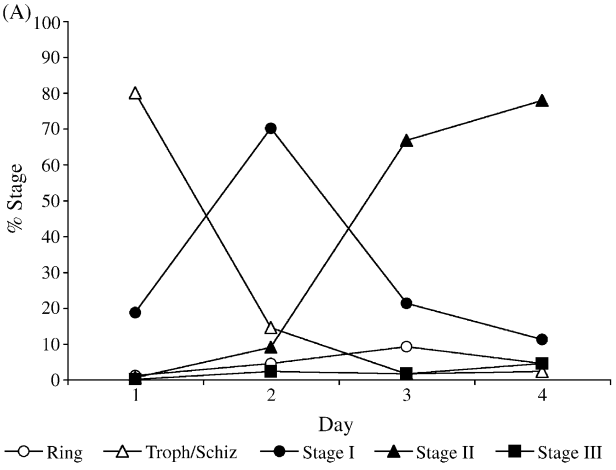


Fig. 1. Percentages of stages present at each day in the NF54 gametocyte development time course were monitored using Giemsa-stained blood smears. RNA was isolated on each day from Day 1 to Day 13 for hybridization to the microarray.

(A)



(B)

Day	Giemsa Gametocyte Count	Pfs16 IF Gametocyte Count
1	19.0%	16.7%
2	81.1%	77.1%
3	89.4%	88.5%
4	93.3%	89.9%

Fig. 2. (A) Percentages of stages present in the CS magnetic affinity column purified 3D7 time course were monitored using Giemsa-stained blood smears. Column separation was conducted to remove contaminating asexual stages immediately after Day 1 sample collection. RNA was isolated on each day for hybridization to the microarray. (B) Giemsa-stained blood smear counts of early-stage gametocytes were confirmed to within 5% for each day using immunofluorescence. Antibody against the gametocyte-specific protein Pfs16 allowed a comparison between the abundance of sexual parasites (Pfs16 fluorescent) vs. overall infected erythrocytes (DAPI fluorescent).

difficult-to-control aspects of the parasite cultivation process such as medium preparation, culture conditions and human donor-specific issues regarding blood and serum that can have an impact on parasite growth rates. Feeding of NF54 stage V gametocytes to *Anopheles stephensi* mosquitoes on Day 15 demonstrated a 40% (2/5) infectivity rate. In addition to the two 14-day time-courses, high purity stage I and stage II 3D7 gametocytes were obtained using CS magnetic affinity columns and quantified by examination of Giemsa-stained blood smears (Fig. 2A). Examination of these blood smears by immunofluorescence using antibodies to Pfs16, which is expressed at the onset of gametocytogenesis, confirmed the Giemsa stain parasites counts to within 5% (Fig. 2B).

3.2. Microarray analysis

RNA isolated from parasites on Days 1–13 for the 2-week NF54 time-course, Days 1–3, 6, 8, and 12 for the 2-week 3D7 time-course, and Days 1–4 for the early-stage 3D7 time-course was hybridized to a full genome high-density oligonucleotide microarray, and expression values were calculated using MOID [16]. Differences in the dynamic range of expression values in the NF54 versus the 3D7 hybridizations were observed with the NF54 having a narrower, but more sensitive range than the 3D7 hybridizations, resulting in more genes being considered present in the NF54 time course than in the 3D7 time course. Although care was taken to ensure that the same protocols were followed for each set of hybridizations, this discrepancy likely arose because different lab personnel conducted hybridizations for each time course at separate times and locations using different microarray scanners. As a result, visual differences occurred between the NF54 and 3D7 colormetric depictions of expression levels for some genes after normalization. However, Pearson's correlation coefficients between all gene expression values in corresponding days in each time course ranged from 0.75 (Day 12) to 0.88 (Day 2), indicating that although the absolute expression values between the data sets may have differed slightly, the overall patterns in gene expression were quite similar (Table S1).

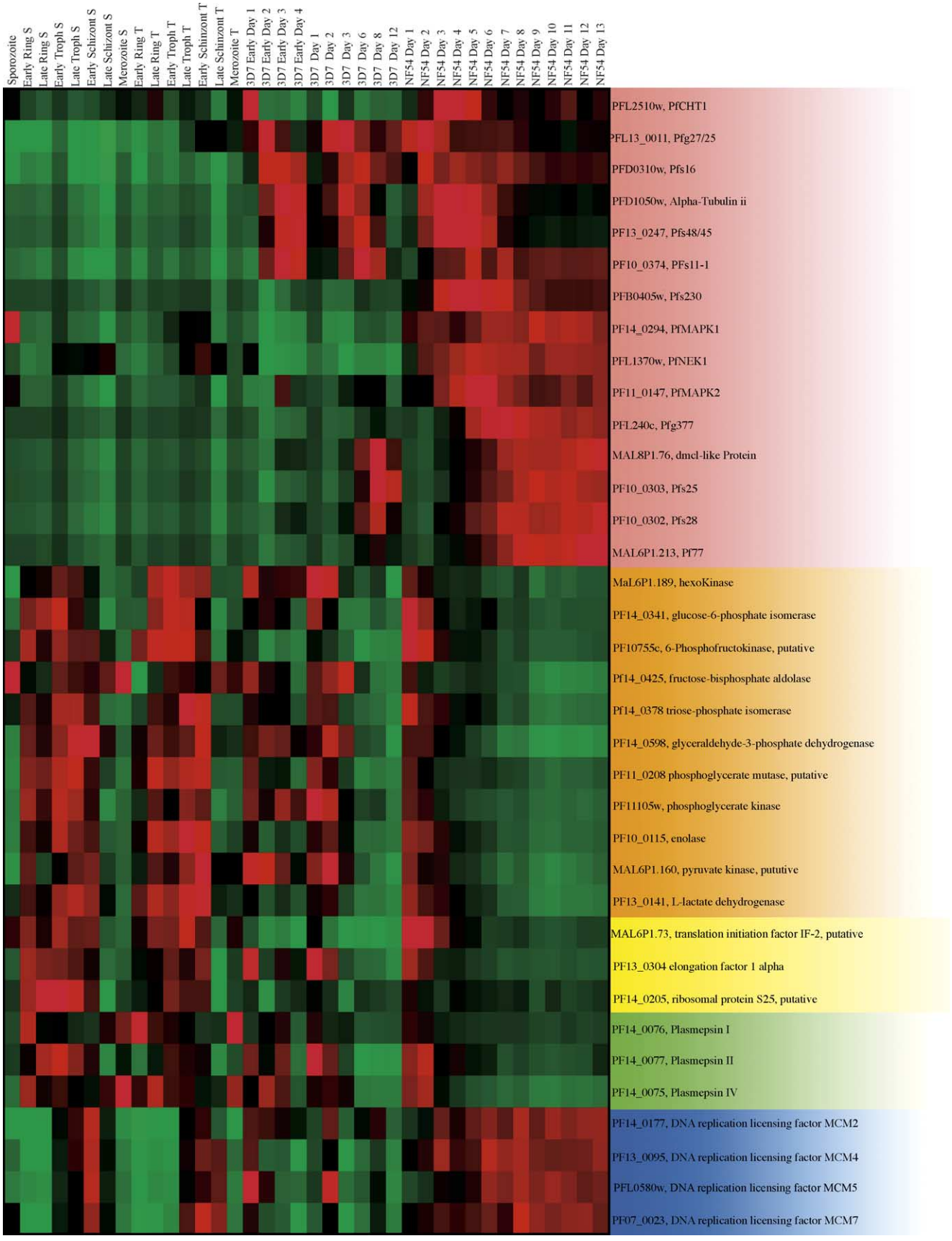
Using expression level and probe signal distribution cutoffs of $E > 10$ and $\log P < -0.5$, respectively, transcripts were found present for an average of 3410 genes per gametocyte microarray time point. These present genes included 15 well-studied sexual development genes we examined (Fig. 3). Further inspection of the gametocyte expression patterns for these characterized sexual stage genes revealed induction for some during the earlier gametocyte stages I–III (*pfg27/25*, *pfs16*, *alpha tubulin ii*, *pfs48/45*), while others exhibited induction during the later gametocyte stages IV–V (*pfs25*, *pfs28*, *pf77*, *dmc1-like protein*) (Fig. 3). In contrast, genes involved in processes such as glycolysis (enolase, hexokinase, triose-phosphate isomerase, etc.), protein biosynthesis (elongation factors, initiation factors and ribosomal proteins), and hemoglobin catabolism (plasmepsins) demonstrated down-regulation in later stage gametocytes (Fig. 3). Other genes, such as those involved in DNA replication (DNA replication licensing factors), showed more constitutive expression across all life cycle stages sampled (Fig. 3). The raw data files for these gametocyte development microarray experiments can be accessed at http://carrier.gnf.org/publications/Gametocyte.[5]

To confirm the sexual-stage-specific expression observed for several hypothetical genes, we conducted semi-quantitative RT-PCR on total RNA isolated from parasites on Day 7 of the NF54 time course and from a mixed asexual NF54 population. In agreement with the microarray data, transcripts for the hypothetical genes PF11_0214, PF14_0067, and MAL8P1.16 were all shown to be present only in gametocytes, whereas the known asexual-stage-specific protein erythrocyte binding antigen *eba-175* (PF07_0128) [19] was present only in the mixed asexual stage population (Fig. 4).

3.3. OPI sexual development analysis

To aid in the identification of hypothetical genes most likely to be involved in sexual development, we applied

[5] The raw data files are also available at PlasmoDB (release 4.4) (www.plasmodb.org) [18].


Sporozoite
Early Ring S
Late Ring S
Early Troph S
Late Troph S
Early Schizont S
Late Schizont S
Merozoite S
Early Ring T
Late Ring T
Early Troph T
Late Troph T
Early Schinzont T
Late Schinzont T
Merozoite T
3D7 Early Day 1
3D7 Early Day 2
3D7 Early Day 3
3D7 Early Day 4
3D7 Day 1
3D7 Day 2
3D7 Day 3
3D7 Day 6
3D7 Day 8
3D7 Day 12
NF54 Day 1
NF54 Day 2
NF54 Day 3
NF54 Day 4
NF54 Day 5
NF54 Day 6
NF54 Day 7
NF54 Day 8
NF54 Day 9
NF54 Day 10
NF54 Day 11
NF54 Day 12
NF54 Day 13
PFL2510w, PfCHT1
PFL13_0011, Pfg27/25
PFD0310w, Pfs16
PFD1050w, Alpha-Tubulin ii
PF13_0247, Pfs48/45
PF10_0374, PFs11-1
PFB0405w, Pfs230
PF14_0294, PfMAPK1
PFL1370w, PfNEK1
PF11_0147, PfMAPK2
PFL240c, Pfg377
MAL8P1.76, dmcl-like Protein
PF10_0303, Pfs25
PF10_0302, Pfs28
MAL6P1.213, Pf77
MaL6P1.189, hexoKinase
PF14_0341, glucose-6-phosphate isomerase
PF10755c, 6-Phosphofructokinase, putative
Pf14_0425, fructose-bisphosphate aldolase
Pf14_0378 triose-phosphate isomerase
PF14_0598, glyceraldehyde-3-phosphate dehydrogenase
PF11_0208 phosphoglycerate mutase, putative
PF11105w, phosphoglycerate kinase
PF10_0115, enolase
MAL6P1.160, pyruvate kinase, pututive
PF13_0141, L-lactate dehydrogenase
MAL6P1.73, translation initiation factor IF-2, putative
PF13_0304 elongation factor 1 alpha
PF14_0205, ribosomal protein S25, putative
PF14_0076, Plasmepsin I
PF14_0077, Plasmepsin II
PF14_0075, Plasmepsin IV
PF14_0177, DNA replication licensing factor MCM2
PF13_0095, DNA replication licensing factor MCM4
PFL0580w, DNA replication licensing factor MCM5
PF07_0023, DNA replication licensing factor MCM7

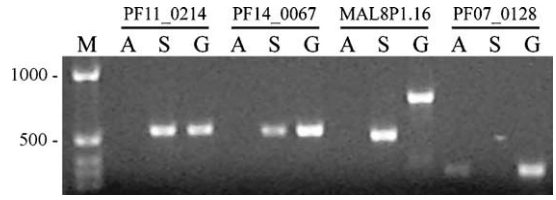


Fig. 4. Validation of microarray data with semi-quantitative RT-PCR. Lanes: M: DNA marker; A: asexual stage cDNA template (from NF54 mixed asexual population); S: sexual stage cDNA template (from NF54 Day 7 gametocytes); G: NF54 genomic DNA template (positive control). Products in lanes S and absent from lanes A for the hypothetical genes PF11_0214, PF14_0067 and MAL8P1.16 demonstrate sexual stage specific expression, as shown by microarray analysis. Likewise, product in lane A and absent in lane S for *eba-175* (PF07_0128) demonstrates asexual-specific expression as previously reported. For MAL8P1.16, primers were designed such that they spanned an intron to allow for the detection of any possible contaminating genomic DNA template, of which none was observed. All product sizes were as expected.

a clustering algorithm recently developed in our laboratory called ontology-based pattern identification [11]. Like other clustering approaches to understanding gene function from microarray data, OPI utilizes the principle of guilt-by-association (GBA), which is based on the observation that for many organisms, genes involved in similar biological processes and functions tend to share similar mRNA expression profiles [20]. Hence, if one can identify uncharacterized genes that possess expression patterns similar to genes for which function is known, one can rapidly make predictions that these uncharacterized genes may also be involved in a similar biological process. However, unlike other clustering approaches, OPI uses existing classifications of gene functions in the form of gene annotations to guide the clustering process. By doing this, OPI is able to generate empirically-optimized gene clusters whose annotated and unannotated gene members, because they share highly correlated expression patterns, have a high likelihood of being involved in the same biological process based on GBA.

To guide OPI in identifying genes most likely to be only involved in sexual development, we created a list of 15 known sexual development genes based on a review of the *P. falciparum* sexual development literature (Table 1) [9,21,22]. The resulting analysis produced a statistically-significant cluster (p-value 3.8×10^{-10}) containing 246 genes that exhibited highly correlated expression patterns (Pearson correlation 0.87) specific for the gametocyte life cycle stages sampled. This cluster, called "Sexual Development" (GO:GNF0004), represents those genes most likely to be specifically involved in the sexual development process and can be accessed as part of the web-accessible public OPI database http://carrier.gnf.org/publications/Gametocyte. Notably, a large proportion of the genes present in this cluster were categorized as hypothetical (∼75%), reflecting the paucity of current knowledge regarding gene function in sexual development.

Of the 15 sexual development genes used to seed the OPI cluster generation, 11 were included in the final sexual development cluster. The four annotated genes absent from the cluster due to a poor correlation in their life cycle expression profiles with other genes in the group were the kinases *pfmapk1* (PF14_0294) and *pfnek1* (PFL1370w), the early gametocyte gene *pfg27/25* (PF13_0011), and the chitinase *pfcht1* (PFL2510w) (Table 1). In the case of *pfcht1*, the poor correlation was because transcript levels for this gene were in some cases below the level of detection in the 3D7 time course. However, for *pfmapk1*, *pfnek1* and *pfg27/25*, the poor correlation was because in addition to the gametocyte stages, these genes also showed significant expression in the asexual stages, as has been reported previously using more traditional methods [9,23–25]. However, since the goal was to identify with high confidence hypothetical genes most likely to be involved only in the process of sexual development, the exclusion of a few known gametocyte genes also exhibiting asexual expression was to be expected, if not desired.

Inspection of genes included in the sexual development cluster beyond the 11 used to seed its generation yielded several interesting observations. Five putative proteases were present [26] including one plasmepsin-like aspartic protease (PF14_0623), two cysteine proteases (PFL2290w, PF11_0298), and two serine proteases (PF14_0067, MAL8P1.16). As for kinases, two putative serine/threonine-protein Nek1 kinases (PFL0080c, PFE1290w), three putative protein kinases (MAL13P1.84, PFC0485w, PFI1290w), and a putative serine/threonine protein kinase 2 (MAL7P1.100) were included. Metabolic genes present included six members of the type II fatty acid pathway, lipoate synthase (MAL13P1.220), pyruvate dehydrogenase α subunit (PF11_0256), β-ketoacyl-ACP synthase III (FabH) (PFB0505c), β-ketoacyl-ACP reductase (FabG) (PFI1125c), enoyl-ACP reductase (FabI) (MAL6P1.275), and malonyl-CoA:ACP transacylase (FabD) (PF13_0066); two members of the heme biosynthesis pathway, δ-aminolevulinic acid synthetase (PFL2210w) and porphobilinogen deaminase (PFL0480w); and two members of the mitochondrial TCA cycle, pyruvate dehydrogenase E1 component α-subunit (PF11_0256) and malate dehydrogenase (MAL6P1.242). Furthermore, insights into the possible functions of the many hypothetical genes in this cluster were gained through

Fig. 3. Colormetric expression map for genes involved in various biological processes (y-axis) across all the life cycle stages sampled (x-axis). Bright red indicates highest expression; bright green indicates lowest expression (gene and array normalized). Asexual stages ending with S and T indicate respective use of sorbitol and temperature for synchronization as previously described [9]. All of the 15 known sexual development genes examined (red) exhibited up regulation in gametocytes. In contrast, examples of genes involved in glycolysis (orange), protein biosynthesis (yellow) and hemoglobin catabolism (green) demonstrated down-regulation during gametocytogenesis, while examples of genes involved in DNA replication (blue) showed more constitutive expression across all life cycle stages sampled. The expression map was generated using the publicly-available software Cluster [67].

Table 1
Fifteen known sexual stage genes manually given the annotation "Sexual Development" (GO:GNF0004) for OPI analysis

Gene ID	Name	Biological role	References
PFD1050w	*alpha-tubulin ii*	Male gametocyte specific	[55]
PFL2510w*	*pfcht1*	Ookinete invasion of mosquito midgut	[56,57]
MAL8P1.76	*dmc1-like protein*	Involved in meiosis	[9]
PF14_0294*	*pfmapk1*	Cell signaling	[58]
PF11_0147	*pfmapk2*	Cell signaling	[24]
PF10_0374	*pf11-1*	Host cell rupture	[59,60]
PF13_0011*	*pfg27/25*	Early sexual differentiation	[25,53]
PFL2405c	*pfg377*	Female gametocyte-specific	[61]
PFL1370w*	*pfnek1*	Cell signaling	[23]
PFD0310w	*pfs16*	Early sexual differentiation	[52]
PFB0405w	*pfs230*	Immune evasion/gamete development	[62]
PF10_0303	*pfs25*	Zygote/ookinete surface protein	[63,64]
PF10_0302	*pfs28*	Zygote/ookinete surface protein	[64]
PF13_0247	*pfs48/45*	Gametocyte/gamete surface protein	[65]
MAL6P1.213	*pf77*	Female gametocyte-specific	[66]

Genes marked with an asterisk were not included in the final 246 gene OPI sexual development cluster.

review of gene information provided by PlasmoDB as of December 2004 [18] (Table S2).

3.4. Gametocyte transcriptional regulation

Relatively little is known about how *P. falciparum* regulates the transcription of genes important for its pathogenesis and development, as no specific transcription factors and very few DNA regulatory elements have been identified in the parasite to date [27,28]. In yeast, analysis of the upstream promoter regions of co-expressed groups of genes has proven to be a powerful method for the identification of novel sequence elements controlling transcription [29]. With the OPI sexual development cluster on hand, representing 246 genes exhibiting highly correlated gametocyte-specific expression patterns, it was possible to take this type of approach to search for conserved *cis*-regulatory elements potentially involved in sexual development transcriptional control mechanisms using a pattern enumeration algorithm called GBSSR [17].

GBSSR analysis of the upstream 1000 bases relative to the start codon for all genes in the sexual development cluster revealed the presence of a palindromic sequence, TGTANNTACA, in 65 of the 246 genes. Overall, this motif is found in 301 genes in the genome. However, the probability that 65 of them would be included in the sexual development cluster by chance is 7.6×10^{-26}. This represents a highly significant degree of enrichment for this motif in promoters of genes exhibiting gametocyte-specific mRNA expression patterns.

Two of the genes whose upstream regions contained the motif were the ookinete surface proteins *pfs25* and *pfs28*. Previously, Dechering et al. performed deletion mapping on the *pfs25* promoter and demonstrated that elimination of the region containing the motif resulted in a marked reduction of luciferase reporter activity [30] (Fig. 5a). Although

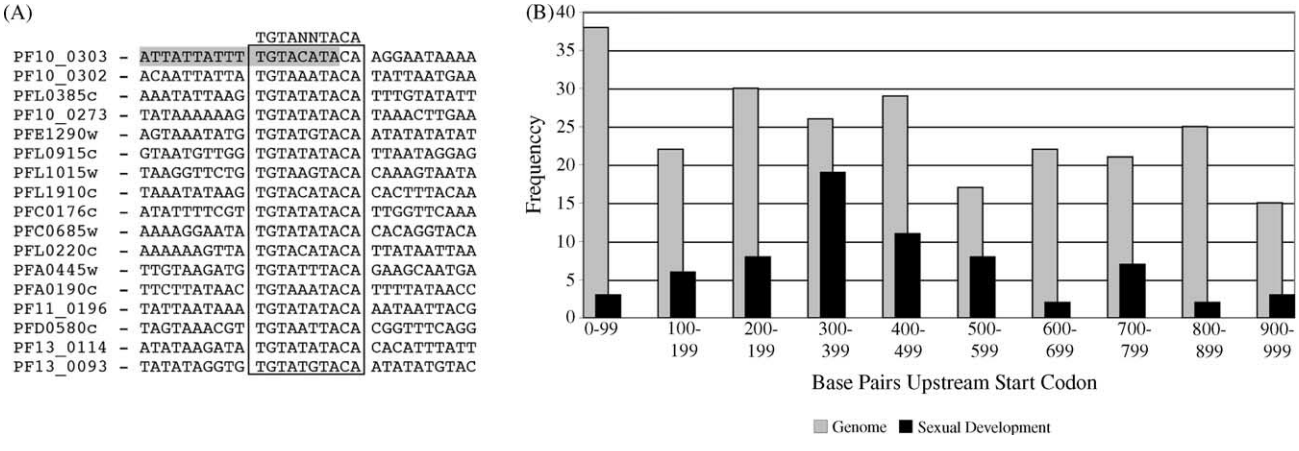


Fig. 5. (A) Alignment of examples of sequences from genes in the OPI sexual development cluster containing the TGTANNTACA motif in their promoter regions. Gray highlight indicates the 3′-most end of the promoter region upstream of *pfs25* (PF10_0303) deleted by Dechering et al. (exo5 construct) that resulted in a reduction of luciferase reporter activity to ∼5% that of the full-length promoter construct [30]. B) Distribution of the TGTANNTACA motif in the 1000 base pair upstream regions relative to the start codon of genes in the OPI sexual development cluster (black) vs. genes in the remainder of the genome (gray). The motif frequency is highest between 200 and 600 bases for genes in the sexual development cluster, but relatively even across the upstream 1000 bases for genes not in the sexual development cluster.

the authors of this study did not identify the palindromic sequence cited here, their work strongly suggests that this motif may indeed have functional relevance. Further evidence suggesting a biological function for this motif comes from an examination of its positional distribution relative to the start codons of genes it is upstream. For genes in the sexual development cluster, the motif appeared most frequently approximately 200–600 bases upstream, whereas for genes not included in this cluster, its position was essentially evenly distributed across the 1000 upstream bases (Fig. 5b). This suggests that the mere presence of the motif ahead of a gene is likely not sufficient for gametocyte-specific expression, but that context-dependent factors also play an important role. It should also be noted that some proteins encoded by genes containing the motif in their promoters do not appear until the later gamete, zygote and ookinete stages (i.e. *pfs25* and *pfs28*). This observation is consistent with the hypothesis that translational silencing may exist for some sexual development genes [31,32].

Another sequence for which there was enrichment in the 1000 bases upstream of the start codon of 187 of the 246 sexual development cluster genes was TCCTT. Although this motif is found fairly common throughout the genome, the probability of finding this amount of enrichment in the sexual development cluster by chance is 1.2×10^{-6}. On average, genes in the sexual development cluster contained 1.8 copies of the motif with 18% of the second copies being located within 50 bases of the first. It is possible that this redundancy confers specificity. This sequence is also the partial complement of the core sequence of the PAF-1 transcription factor binding site, AAGGAATA, that was identified previously as important for *pfs25* promotor activity in the sexual stages [30].

3.5. Comparison of P. falciparum and P. berghei gametocyte transcriptomes

Recently, Hall et al. published a comprehensive survey of the life cycles of rodent malaria model organisms that included an analysis of the transcriptome of immature and mature *Plasmodium berghei* gametocytes using a genome survey sequence amplicon DNA microarray [31]. In all, transcripts for 977 genes were reported induced two-fold or more in *P. berghei* gametocytes compared to pure *P. berghei* asexual parasites. Of these 977 genes, 504 possessed *P. falciparum* orthologues, which allowed a direct comparison to the 246 genes exhibiting sexual-stage-specific expression in the OPI sexual development cluster. This comparison found 64 genes that were common to both lists, including the well-characterized sexual stage genes encoding *alpha-tubulin ii*, *dmc1-like protein*, and *pfs25*.

Aside from biological differences between these two related species and the fact that many genes did not possess orthologues for inter-species comparison, one likely reason for the apparently low overlap between these two gene lists is differences in the criteria set by each group to define gametocyte-specific genes. By using a two-fold change threshold, Hall and coworkers likely included many genes in their list of 977 genes that, although present in gametocytes, were not necessarily specific to this stage of development. This was evidenced by the presence of many DNA replication genes in the *P. berghei* gametocyte transcriptome list that are also present in *P. berghei* asexual stages. OPI analysis of the *P. falciparum* gametocyte transcriptome used in this study sets much more stringent criteria for defining genes as gametocyte-specific. Because we primarily used genes with sexual-stage-specific expression profiles to guide OPI sexual development cluster generation, genes expressed at significant levels in the asexual stages as well as in gametocytes were excluded from the OPI sexual development cluster, as demonstrated by the absence of DNA replication genes such as DNA licensing factors and DNA polymerases. However, despite the differences between the studies, each provides important insights into the biology of these related malarial parasites. Moreover, the list of 64 genes common to each gametocyte transcriptome represents those genes that we can say with higher confidence play important roles in sexual development.

3.6. GO annotation OPI analysis

In addition to using OPI to systematically identify genes most likely to be involved in sexual development using manually assigned annotations, OPI was extended to predict genes involved in a vast array of other parasite processes using GO annotations as a guide. The GO consortium classifies genes with respect to biological processes, cellular components, and molecular functions using a controlled vocabulary [12]. The October 2004 annotations for the 3059 differentially regulated *P. falciparum* genes identified in this study were obtained from the GO consortium website (http://www.geneontology.org). OPI analysis using these annotations as a guide resulted in the generation of 380 statistically significant gene clusters representing 202 biological processes, 96 molecular functions, and 82 cellular components. The *p*-values for these clusters ranged from 1.7×10^{-4} to 1.0×10^{-75}, which is several orders of magnitude lower than those previously reported using a *k*-means clustering approach [9]. Some of the most statistically significant and biologically interesting of these clusters are listed in Table 2. The complete results of this analysis can be accessed alongside the sexual development cluster at the web-accessible OPI database http://carrier.gnf.org/publications/Gametocyte.

4. Discussion

Although there are well-characterized differences in drug sensitivity between asexual and sexual stage parasites [33], the specific metabolic characteristics unique to each stage remain poorly understood, especially with regard to the mitochondrion and apicoplast. The mitochondria of asexual and

Table 2
Eighteen examples of statistically significant gene clusters generated using OPI with GO annotations as a guide

GO #	Type	Description	Total	*p*-value	TPR	Corr
GO:0020033	BP	Antigenic variation	288	1.0E−75	82.6% (86/104)	0.72
GO:0007154	BP	Cell communication	290	3.2E−44	50.0% (84/168)	0.73
GO:0007155	BP	Cell adhesion	50	3.3E−23	60.0% (17/28)	0.83
GO:0009266	BP	Response to temperature	70	4.0E−6	42.9% (6/14)	0.71
GO:0030260	BP	Cell invasion	87	6.1E−16	100% (10/10)	0.69
GO:0016052	BP	Carbohydrate catabolism	82	3.1E−10	50.0% (11/22)	0.79
GO:0020020	CC	Food vacuole (sensu apicomplexa)	213	1.5E−7	100% (7/7)	0.73
GO:0020030	CC	Infected host cell surface knob	46	1.3E−24	76.2% (16/21)	0.83
GO:0000502	CC	Proteasome complex (sensu eukarya)	75	3.5E−21	76.2% (16/21)	0.80
GO:0020008	CC	Rhoptry	4	3.3E−12	100% (4/4)	0.97
GO:0020036	CC	Maurer's cleft	246	1.4E−08	77.0% (10/13)	0.79
GO:0020007	CC	Apical complex	4	5.0E−11	66.7% (4/6)	0.95
GO:0005198	MF	Structural molecule activity	446	1.0E−11	52.0% (39/75)	0.60
GO:0008092	MF	Cytoskeletal protein binding	277	6.2E−6	57.1% (8/14)	0.80
GO:0003774	MF	Motor activity	302	7.5E−8	66.7% (10/15)	0.82
GO:0008233	MF	Peptidase activity	93	3.2E−12	32.1% (18/56)	0.72
GO:0004672	MF	Protein kinase activity	242	2.6E−5	26.0% (13/50)	0.81
GO:0004536	MF	Deoxyribonuclease activity	32	9.0E−9	100% (4/4)	0.89

BP: biological process, CC: cellular component, MF: molecular function. The column "Total" represents the total number of genes in each cluster. The *p*-values for each cluster were assigned using the hypergeometric distribution representing the probability that the number of genes possessing the GO annotation used to define the cluster would fall into the cluster of size "Total." The column true positive rate (TPR) represents the number of genes in the cluster that possess the GO annotation used to define the cluster divided by the total number of genes in the data set that possess the GO annotation, multiplied by 100 for a percentage. The correlation for a cluster is defined as Pearson's correlation coefficient between the life cycle expression patterns of the two least correlated genes.

sexual blood stage parasites are remarkably distinct morphologically, with ultrastructural studies showing that *P. falciparum* gametocytes contain multiple (4–8) mitochondria with a relatively high density of cristae, whereas the single mitochondrion of asexual parasites have relatively few cristae [34,35]. These sexual stage mitochondria are thought to be functionally active, since anti-malarials that appear to target the parasite electron transport chain show some efficacy against gametocytes [34]. Moreover, while the asexual stages of the parasite life cycle in the human host rely heavily on glucose for the production of energy by fermentation, it has been proposed that a switch to aerobic mitochondria-driven energy production occurs during the more oxygen-rich mosquito sexual stages, coinciding with the appearance of more cristate mitochondria during these stages [35–37]. Our data support this hypothesis, as we show that while transcripts for glycolytic enzymes are down regulated in gametocytes, transcripts for 15 of the 16 mitochondrial TCA cycle enzymes are present in gametocytes, with transcripts for pyruvate dehydrogenase and malate dehydrogenase in particular showing highly gametocyte-specific expression patterns as indicated by their inclusion in the OPI sexual development cluster. Furthermore, evidence for the importance of organelle function during sexual development is provided by our data, as we showed significant up regulation of transcripts during gametocytogenesis for three of the six genes currently identified as part of the heme biosynthesis pathway in *P. falciparum* for which our microarray had probes. These included the mitochrondion-localized enzyme, δ-aminolevulinate synthetase (1st step) [38,39], as well as the enzymes porphobilinogen deaminase (3rd step) and uroporphyrinogen decarboxylase (5th step), both thought to be localized to the apicoplast [39]. De novo heme biosynthesis may be important during the sexual stages as the parasite leaves the hemoglobin-rich blood environment of the human host. Our data also supports this notion by showing a simultaneous reduction in hemoglobin catabolism gene expression during sexual development as compared to the asexual stages.

Despite the apicoplast of sexual stage parasites receiving relatively scarce attention in the literature, there is some evidence that the plastid-like organelle plays a role in sexual development. Thiostrepton is an antibiotic that targets the large subunit ribosomal RNA encoded by the apicoplast genome [40]. Administration of this drug to mice infected with *P. berghei* leads to reduction of parasite transmission to mosquitoes, suggesting that the apicoplast is involved in sexual development [41]. However, the exact processes and enzymes that may be important for this putative sexual role of the apicoplast remain largely unexplored. Inspection of the genes involved in apicoplast-related processes revealed that five of the six genes belonging to the apicoplast-localized portion of type II fatty acid biosynthesis pathway were significantly up regulated during the gametocyte stages. Moreover, gametocyte-specific expression was also observed for the genes encoding the enzymes lipoate synthase and pyruvate dehydrogenase α subunit, both thought to be cytosolically localized. These enzymes are responsible for generating the fatty-acid precursor molecules lipoate and acetyl-CoA. It is unclear why fatty acid synthesis would be important to sexual stage parasites, although one possible explanation is the likely requirement for a substantial increase in synthesis of membrane precursors for growth of the gametocyte and the process of exflagellation. Regardless, because many of these genes have a prokaryotic origin, they

provide attractive targets for the development of potentially low toxicity drugs with transmission-blocking properties.

The proteases of Apicomplexan parasites have long been considered attractive drug targets due to the many important biological roles they play and the technical feasibility of designing specific chemical inhibitors against them [42]. Although protease activities have been implicated in schizont rupture during the intraerythrocytic cycle of *Plasmodium* parasites [43,44], it is not known if similar activities are involved in the sexual stages. The five putative proteases [26] identified as having gametocyte-specific expression from the OPI sexual development cluster make attractive targets for the development of transmission-blocking drugs. One of these genes, MAL8P1.16, encodes a putative member of the rhomboid protease family. Rhomboids are intramembrane serine proteases that have substrate cleavage sites within their transmembrane domains. In *Drosophila melanogaster*, they have been shown to cleave proteins containing epidermal growth factor (EGF) domains, allowing the EGF ligands to bind to EGF receptors [45]. Recent evidence suggests that these proteases also play a role in activating signals for intercellular communication in the pathogenic Gram-negative bacterium *Providencia stuartii* [46,47]. Potential cleavage of surface proteins by this putative rhomboid protease during sexual development in *P. falciparum* has yet to be investigated, despite evidence that rhomboid proteases are expressed in the oocyst stage of the related Apicomplexan parasite in *Toxoplasma gondii* [48].

The intracellular signaling pathways responsible for the triggering of parasite sexual development in the vertebrate host remain largely uncharacterized. Despite several environmental factors being shown to induce gametocytogenesis both in vitro and in vivo [49], chromosomal deletions being associated with a loss of the ability to produce gametocytes [50,51], and the identification of early gametocyte development markers such as Pfs16 [52] and Pfg27/25 [25,53], the details regarding the specific signaling mechanisms responsible for the switch from a asexual to sexual mode of replication remain elusive. Recently, Gardiner et al. identified a gene, *pfgig* (PFI1720w), that when genetically manipulated through silencing or over-expression, resulted in a respective reduction or induction of gametocyte production [54]. As the authors of this study noted, our previous life cycle microarray data showed no expression for this gene in late-stage gametocytes [9]. However, the early-stage gametocyte microarray data presented herein shows significant expression for this gene, thus supporting the notion that it is involved in early sexual development events. Furthermore, the six kinases demonstrated here to have gametocyte-specific expression patterns, as well as the sequence motifs enriched in upstream regions of genes in the sexual development cluster, both represent promising leads for further study toward elucidating the components of sexual development signaling pathways.

Perhaps the most interesting and exciting aspect of the sexual development cluster identified in this study is the large abundance of hypothetical genes it possesses, and it is in this regard that analyses such as the one presented here hold the most promise. Understanding the biological role hypothetical genes play is of particular importance with regard to pathogenic organisms such as *P. falciparum*, where over 60% of the predicted ORFs possess no putative function and therefore are ignored by rational drug and vaccine design efforts. Sequence-based high-throughput genomic and proteomic research can act as powerful complementary approaches where more traditional methods of understanding gene function have been slow to yield results. However, it remains a daunting challenge to analyze the large amounts of data obtained by these genomics-based studies in order to generate insightful hypotheses that give rise to biologically relevant conclusions. Of the 246 genes in the OPI sexual development cluster described herein, 197 were part of a *k*-cluster with partially gametocyte-like expression patterns in our previous life cycle microarray work using a single late-stage gametocyte time point [9]. However, due to limitations in the *k*-means clustering approach employed in this past study, these 197 genes were scattered across six of fifteen clusters and intermingled with genes possessing non-gametocyte specific expression patterns. The sexual development cluster generated here using a more stringent OPI approach represents a consolidation of gametocyte-specific genes into one highly correlated gene cluster. This consolidation not only allowed us to say with higher confidence that any particular gene contained within the cluster is involved specifically in sexual development, but also identified 49 genes, some of which have earlier gametocyte-specific expression patterns, that were previously overlooked using *k*-means clustering.

In its current state, the generic nature of OPI has both strengths and weaknesses. One shortcoming of OPI is that it has a tendency to generate some clusters that are statistically significant but may not be biologically meaningful. For example, clusters possessing more than 500 genes can be challenging to interpret biologically and will contain many false positives. Additionally, many of the clusters defined by molecular function annotation such as proteases and kinases that operate in various stages and locations of the parasite may also be irrelevant if only life cycle expression data is used as the basis for analysis. Likewise, genes involved in negative regulation of a process are unlikely to share an expression profile with the genes that they regulate and therefore may be missed in the OPI clustering process. Another concern with using OPI to study organisms such as *P. falciparum* whose genome is poorly annotated is that many important insights will be missed. However, the primary goal of OPI is not to comprehensively describe every detail of the parasite biology, but rather to use current information to make studies on individual gene functions using traditional methods more focused and efficient. As gene annotations for *P. falciparum* improve over time and as the repertoire of malaria life cycle stages profiled using microarrays increases to include later sexual stage and liver stage parasites, the specificity and quality of gene clusters obtained via OPI analysis will undoubtedly be further refined. Moreover, the generic nature of OPI makes

it amenable to the analysis of other types of high-throughput genome-wide data generated by methods such as quantitative proteomics and large-scale yeast-two hybrid experiments, which can address the functional states of proteins in a more direct manner than is possible using solely transcriptional analysis. Integration of data from these varying approaches will help lay the groundwork for a malaria systems biology aimed at generating more robust hypotheses that will be critical to advancing the understanding of this important disease organism.

Acknowledgements

We would like to thank Irwin W. Sherman and the anonymous referees for critical reading of the manuscript and helpful suggestions. We would like to thank Ivan R. Baxter for assistance with GBSSR. This work was supported by a grant to E.A.W. from The Ellison Foundation and the W.M. Keck Foundation, grants to D.A.B. from The Wellcome (Trust Thematic program grant (Ref 066742) and University Award (Ref 058038)), and funding to D.R.C. from the NMRC and the Military Infectious Disease Research Program (MIDRP). This research was performed while J.A.Y. held a National Science Foundation Graduate Research Fellowship and P.L.B. held a National Research Council Research Associateship Award at the Naval Medical Research Center (NMRC). The opinions expressed are those of the authors and do not reflect the official policy of the Department of the Navy, Department of Defense, or the US government.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molbiopara.2005.05.007.

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] Stowers A, Carter R. Current developments in malaria transmission-blocking vaccines. Expert Opin Biol Ther 2001;1(4):619–28.
[3] Kaslow DC. Transmission-blocking vaccines: uses and current status of development. Int J Parasitol 1997;27(2):183–9.
[4] Hawking F, Wilson ME, Gammage K. Evidence for cyclic development and short-lived maturity in the gametocytes of *Plasmodium falciparum*. Trans R Soc Trop Med Hyg 1971;65(5):549–59.
[5] Ponnudurai T, Lensen AH, Meis JF, Meuwissen JH. Synchronization of *Plasmodium falciparum* gametocytes using an automated suspension culture system. Parasitology 1986;93(Pt 2):263–74.
[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] Young JA, Winzeler EA. Using expression information to discover new drug and vaccine targets in the malaria parasite *Plasmodium falciparum*. Pharmacogenomics 2005;6(1):17–26.
[8] Johnson JR, Florens L, Carucci DJ, Yates JR. Third proteomics in malaria. J Proteome Res 2004;3(2):296–306.
[9] Le Roch KG, Zhou Y, Blair PL, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 2003;301(5639):1503–8.
[10] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 2003;1(1):5.
[11] Zhou Y, Young JA, Santrosyan A, Chen K, Yan FS, Winzeler EA. In silico gene function prediction using ontology-based pattern identification. Bioinformatics 2004.
[12] Creating the gene ontology resource: design and implementation. Genome Res 2001;11(8):1425–33.
[13] Trager W, Jensen JB. Human malaria parasites in continuous culture. Science 1976;193(4254):673–5.
[14] Lambros C, Vanderberg JP. Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. J Parasitol 1979;65(3):418–20.
[15] Le Roch KG, Zhou Y, Batalov S, Winzeler EA. Monitoring the chromosome 2 intraerythrocytic transcriptome of *Plasmodium falciparum* using oligonucleotide arrays. Am J Trop Med Hyg 2002;67(3):233–43.
[16] Zhou Y, Abagyan R. Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. BMC Bioinformatics 2002;3(1):3.
[17] Kreps J, Budworth P, Goff S, Wang R. Identification of putative plant cold responsive regulatory elements by gene expression profiling and a pattern enumeration algorithm. Plant Biotechnol J 2003;1(5):345–52.
[18] Bahl A, Brunk B, Coppel RL, et al. PlasmoDB: the *Plasmodium* genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). Nucleic Acids Res 2002;30(1):87–90.
[19] Adams JH, Blair PL, Kaneko O, Peterson DS. An expanding ebl family of *Plasmodium falciparum*. Trends Parasitol 2001;17(6):297–9.
[20] Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. Genome Res 1999;9(12):1198–203.
[21] Hayward RE, Derisi JL, Alfadhli S, Kaslow DC, Brown PO, Rathod PK. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. Mol Microbiol 2000;35(1):6–14.
[22] Kongkasuriyachai D, Kumar N. Functional characterisation of sexual stage specific proteins in *Plasmodium falciparum*. Int J Parasitol 2002;32(13):1559–66.
[23] Dorin D, Le Roch K, Sallicandro P, et al. Pfnek-1, a NIMA-related kinase from the human malaria parasite *Plasmodium falciparum* Biochemical properties and possible involvement in MAPK regulation. Eur J Biochem 2001;268(9):2600–8.
[24] Dorin D, Alano P, Boccaccio I, et al. An atypical mitogen-activated protein kinase (MAPK) homologue expressed in gametocytes of the human malaria parasite *Plasmodium falciparum*. Identification of a MAPK signature. J Biol Chem 1999;274(42):29912–20.
[25] Pologe LG. Aberrant transcription and the failure of *Plasmodium falciparum* to differentiate into gametocytes. Mol Biochem Parasitol 1994;68(1):35–43.
[26] Wu Y, Wang X, Liu X, Wang Y. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. Genome Res 2003;13(4):601–16.
[27] Coulson RM, Hall N, Ouzounis CA. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. Genome Res 2004;14(8):1548–54.
[28] Aravind L, Iyer LM, Wellems TE, Miller LH. *Plasmodium* biology: genomic gleanings. Cell 2003;115(7):771–85.
[29] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet 1999;22(3):281–5.
[30] Dechering KJ, Kaan AM, Mbacham W, et al. Isolation and functional characterization of two distinct sexual-stage-specific promoters of

the human malaria parasite *Plasmodium falciparum*. Mol Cell Biol 1999;19(2):967–78.
[31] 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(5706):82–6.
[32] Le Roch KG, Johnson JR, Florens L, et al. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res 2004;14(11):2308–18.
[33] Lang-Unnasch N, Murphy AD. Metabolic changes of the malaria parasite during the transition from the human to the mosquito host. Annu Rev Microbiol 1998;52:561–90.
[34] Krungkrai J, Prapunwattana P, Krungkrai SR. Ultrastructure and function of mitochondria in gametocytic stage of *Plasmodium falciparum*. Parasite 2000;7(1):19–26.
[35] Langreth SG, Jensen JB, Reese RT, Trager W. Fine structure of human malaria in vitro. J Protozool 1978;25(4):443–52.
[36] Howells RE. Mitochondrial changes during the life cycle of Plasmodium berghei. Ann Trop Med Parasitol 1970;64(2):181–7.
[37] Bongaerts G. Do the mitochondria of malaria parasites behave like the phoenix after return in the mosquito? Regeneration of degenerated mitochondria is required for successful Plasmodium infection. Med Hypotheses 2005;64(3):599–601.
[38] Varadharajan S, Dhanasekaran S, Bonday ZQ, Rangarajan PN, Padmanaban G. Involvement of delta-aminolaevulinate synthase encoded by the parasite gene in de novo haem synthesis by *Plasmodium falciparum*. Biochem J 2002;367(Pt 2):321–7.
[39] Sato S, Clough B, Coates L, Wilson RJ. Enzymes for heme biosynthesis are found in both the mitochondrion and plastid of the malaria parasite *Plasmodium falciparum*. Protist 2004;155(1):117–25.
[40] McConkey GA, Rogers MJ, McCutchan TF. Inhibition of *Plasmodium falciparum* protein synthesis. Targeting the plastid-like organelle with thiostrepton. J Biol Chem 1997;272(4):2046–9.
[41] Sullivan M, Li J, Kumar S, Rogers MJ, McCutchan TF. Effects of interruption of apicoplast function on malaria infection, development, and transmission. Mol Biochem Parasitol 2000;109(1):17–23.
[42] Rosenthal PJ. Proteases of malaria parasites: new targets for chemotherapy. Emerg Infect Dis 1998;4(1):49–57.
[43] Wickham ME, Culvenor JG, Cowman AF. Selective inhibition of a two-step egress of malaria parasites from the host erythrocyte. J Biol Chem 2003;278(39):37658–63.
[44] Salmon BL, Oksman A, Goldberg DE. Malaria parasite exit from the host erythrocyte: a two-step process requiring extraerythrocytic proteolysis. Proc Natl Acad Sci USA 2001;98(1):271–6.
[45] Urban S, Lee JR, Freeman M. Drosophila rhomboid-1 defines a family of putative intramembrane serine proteases. Cell 2001;107(2):173–82.
[46] Gallio M, Sturgill G, Rather P, Kylsten P. A conserved mechanism for extracellular signaling in eukaryotes and prokaryotes. Proc Natl Acad Sci USA 2002;99(19):12208–13.
[47] Urban S, Schlieper D, Freeman M. Conservation of intramembrane proteolytic activity and substrate specificity in prokaryotic and eukaryotic rhomboids. Curr Biol 2002;12(17):1507–12.
[48] Brossier F, Jewett TJ, Sibley LD, Urban S. A spatially localized rhomboid protease cleaves cell surface adhesins essential for invasion by Toxoplasma. Proc Natl Acad Sci USA 2005;102(11):4146–51.
[49] Talman AM, Domarle O, McKenzie FE, Ariey F, Robert V. Gametocytogenesis: the puberty of *Plasmodium falciparum*. Malar J 2004;3(1):24.
[50] Day KP, Karamalis F, Thompson J, et al. Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. Proc Natl Acad Sci USA 1993;90(17):8292–6.
[51] Alano P, Roca L, Smith D, Read D, Carter R, Day K. *Plasmodium falciparum*: parasites defective in early stages of gametocytogenesis. Exp Parasitol 1995;81(2):227–35.
[52] Dechering KJ, Thompson J, Dodemont HJ, Eling W, Konings RN. Developmentally regulated expression of pfs16, a marker for sexual differentiation of the human malaria parasite *Plasmodium falciparum*. Mol Biochem Parasitol 1997;89(2):235–44.
[53] Carter R, Graves PM, Creasey A, et al. *Plasmodium falciparum*: an abundant stage-specific protein expressed during early gametocyte development. Exp Parasitol 1989;69(2):140–9.
[54] Gardiner DL, Dixon MW, Spielmann T, et al. Implication of a *Plasmodium falciparum* gene in the switch between asexual reproduction and gametocytogenesis. Mol Biochem Parasitol 2005;140(2):153–60.
[55] Rawlings DJ, Fujioka H, Fried M, Keister DB, Aikawa M, Kaslow DC. Alpha-tubulin II is a male-specific protein in *Plasmodium falciparum*. Mol Biochem Parasitol 1992;56(2):239–50.
[56] Vinetz JM, Dave SK, Specht CA, et al. The chitinase PfCHT1 from the human malaria parasite *Plasmodium falciparum* lacks proenzyme and chitin-binding domains and displays unique substrate preferences. Proc Natl Acad Sci USA 1999;96(24):14061–6.
[57] Tsai YL, Hayward RE, Langer RC, Fidock DA, Vinetz JM. Disruption of *Plasmodium falciparum* chitinase markedly impairs parasite invasion of mosquito midgut. Infect Immun 2001;69(6):4048–54.
[58] Doerig CM, Parzy D, Langsley G, Horrocks P, Carter R, Doerig CD. A MAP kinase homologue from the human malaria parasite. *Plasmodium falciparum*. Gene 1996;177(1-2):1–6.
[59] Feng Z, Hoffmann RN, Nussenzweig RS, et al. Pfs2400 can mediate antibody-dependent malaria transmission inhibition and may be the *Plasmodium falciparum* 11.1 gene product. J Exp Med 1993;177(2):273–81.
[60] Scherf A, Carter R, Petersen C, et al. Gene inactivation of Pf11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametogenesis. Embo J 1992;11(6):2293–301.
[61] Severini C, Silvestrini F, Sannella A, Barca S, Gradoni L, Alano P. The production of the osmiophilic body protein Pfg377 is associated with stage of maturation and sex in *Plasmodium falciparum* gametocytes. Mol Biochem Parasitol 1999;100(2):247–52.
[62] Williamson KC, Fujioka H, Aikawa M, Kaslow DC. Stage-specific processing of Pfs230, a *Plasmodium falciparum* transmission-blocking vaccine candidate. Mol Biochem Parasitol 1996;78(1–2):161–9.
[63] Barr PJ, Green KM, Gibson HL, Bathurst IC, Quakyi IA, Kaslow DC. Recombinant Pfs25 protein of *Plasmodium falciparum* elicits malaria transmission-blocking immunity in experimental animals. J Exp Med 1991;174(5):1203–8.
[64] Duffy PE, Kaslow DC. A novel malaria protein, Pfs28, and Pfs25 are genetically linked and synergistic as falciparum malaria transmission-blocking vaccines. Infect Immun 1997;65(3):1109–13.
[65] Kocken CH, Jansen J, Kaan AM, et al. Cloning and expression of the gene coding for the transmission blocking target antigen Pfs48/45 of *Plasmodium falciparum*. Mol Biochem Parasitol 1993;61(1):59–68.
[66] Baker DA, Thompson J, Daramola OO, Carlton JM, Targett GA. Sexual-stage-specific RNA expression of a new *Plasmodium falciparum* gene detected by in situ hybridisation. Mol Biochem Parasitol 1995;72(1–2):193–201.
[67] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95(25):14863–8.