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

Jason A. Young, Quinton L. Fivelman, Peter L. Blair, Patricia de la Vega, Karine G. Le Roch, Yingyao Zhou, Daniel J. Carucci, David A. Baker, Elizabeth A. Winzeler

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 approximately 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.
1. REPORT DATE
2005

2. REPORT TYPE
N/A

3. DATES COVERED
-

4. TITLE AND SUBTITLE
The Plasmodium falciparum sexual development transcriptome: A microarray analysis using ontology-based pattern identification

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)

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:

<table>
<thead>
<tr>
<th>a. REPORT</th>
<th>b. ABSTRACT</th>
<th>c. THIS PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>unclassified</td>
<td>unclassified</td>
<td>unclassified</td>
</tr>
</tbody>
</table>

17. LIMITATION OF ABSTRACT
SAR

18. NUMBER OF PAGES
13

19a. NAME OF RESPONSIBLE PERSON

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. This is the 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 (Milltenyi 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 and early gametocytes were passed through a CS column. 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 (Milltenyi 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 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 stages were 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-Pf16 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-Pf16 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 purifed 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 G3′; reverse primer 5′-GGG TAT TAC CAT AAC TTT TCA ATT GTC CTC G3′; PF14_0067: forward primer 5′-GCA TAT TGT AAA GAA TGG TGT AAA GCC ACG-3′; reverse primer 5′-CCA GCA GTA AAT CCT TCA TGT AAT ACT TTA G-3′; MAL8P1.16: forward primer 5′-GTT GTT GTA TCA TCT GAA AAT ATT AAG CTC C-3′, reverse primer 5′-CTT AGC CTT TGT TAA ATT CT TCA AGC ACG-3′) and the known asexual-stage-specific gene ebu-175 (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 GCT C-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. falciparum* 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:0000693), 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 hypergeometric *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 permuted 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
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. RNA was isolated on each day for hybridization to the microarray. (A) 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).

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 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 colometric 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 cut-offs 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 (PF072725, pf16, alphatubulin ii, pf68/45), while others exhibited induction during the later gametocyte stages IV–V (pf72, pf28, pf77, amsc1-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.3

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
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 function 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 0.85) 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 pfef2727/25 (PF13_0011), and the citrate synthetase pfeg27/25 (PFL2510w) (Table 1). In the case of pfef27, 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 pfef27/25, 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 plasmin-like aspartic protease (PF11_0623), two cysteine proteases (PFL2209w, PF11_0298), and two serine proteases (PF14_0294, PF11_0294). Notably, a large proportion of the genes present in this cluster, 207 (84%) of the 246, were in some cases below the level of detection in the 3D7 time course (with a median expression level of 0.007). However, more than half of the putative proteases present (114, 46%) were expressed at levels above the threshold.

Of the 15 known sexual development genes used to seed the OPI database, 14 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 pfef2727/25 (PF13_0011), and the citrate synthetase pfeg27/25 (PFL2510w) (Table 1). In the case of pfef27, 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 pfef27/25, 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 plasmin-like aspartic protease (PF11_0623), two cysteine proteases (PFL2209w, PF11_0298), and two serine proteases (PF14_0294, PF11_0294). Notably, a large proportion of the genes present in this cluster, 207 (84%) of the 246, were in some cases below the level of detection in the 3D7 time course (with a median expression level of 0.007). However, more than half of the putative proteases present (114, 46%) were expressed at levels above the threshold.

Fig. 3. Colormetric expression map for genes involved in various biological processes (x-axis) across all the life cycle stages sampled (y-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>
<thead>
<tr>
<th>Gene ID</th>
<th>Name</th>
<th>Biological role</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFD1050w</td>
<td>alpha-tubulin ii</td>
<td>Male gametocyte specific</td>
<td>[55]</td>
</tr>
<tr>
<td>PFL2510w</td>
<td>pfcht1</td>
<td>Ookinet invasion of mosquito midgut</td>
<td>[56,57]</td>
</tr>
<tr>
<td>MALAPI.76</td>
<td>dhc1-like protein</td>
<td>Involved in meiosis</td>
<td>[9]</td>
</tr>
<tr>
<td>PF14.2094</td>
<td>pfmapk1</td>
<td>Cell signaling</td>
<td>[58]</td>
</tr>
<tr>
<td>PF14.2047</td>
<td>pfmapk2</td>
<td>Cell signaling</td>
<td>[24]</td>
</tr>
<tr>
<td>PF12.0734</td>
<td>pf111</td>
<td>Host cell rupture</td>
<td>[59,60]</td>
</tr>
<tr>
<td>PF13.0011</td>
<td>pf127025</td>
<td>Early sexual differentiation</td>
<td>[25,53]</td>
</tr>
<tr>
<td>PFL2656c</td>
<td>pf377</td>
<td>Female gametocyte-specific</td>
<td>[61]</td>
</tr>
<tr>
<td>PFL3730w</td>
<td>pf109</td>
<td>Cell signaling</td>
<td>[23]</td>
</tr>
<tr>
<td>PFD0106w</td>
<td>pf116</td>
<td>Early sexual differentiation</td>
<td>[52]</td>
</tr>
<tr>
<td>PFBI086w</td>
<td>pf1239</td>
<td>Immune evasion/gamete development</td>
<td>[62]</td>
</tr>
<tr>
<td>PF10.0235</td>
<td>pf25</td>
<td>Zygote/ookinete surface protein</td>
<td>[63,64]</td>
</tr>
<tr>
<td>PF10.0202</td>
<td>pf26</td>
<td>Zygote/ookinete surface protein</td>
<td>[64]</td>
</tr>
<tr>
<td>PFL1247</td>
<td>pf604045</td>
<td>Gametocyte/germ surface protein</td>
<td>[65]</td>
</tr>
<tr>
<td>MALAPI.213</td>
<td>pf277</td>
<td>Female gametocyte-specific</td>
<td>[66]</td>
</tr>
</tbody>
</table>

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

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 pf25 and pf28. Previously, Dechering et al. performed deletion mapping on the pf25 promoter and demonstrated that elimination of the region containing the motif resulted in a marked reduction of luciferase reporter activity [30] (Fig. 5a). Although
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 genes was TCCTT. Although this motif 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].

### 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 $1.7 \times 10^{-4}$ to $1.0 \times 10^{-4}$, 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...
Isosexual 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 are thought to be functionally active, since anti-malarials that appear to target the parasite electron transport chain show thought to be functionally active, since anti-malarials that appear to target the parasite electron transport chain show 91

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

<table>
<thead>
<tr>
<th>GO #</th>
<th>Type</th>
<th>Description</th>
<th>Total</th>
<th>p-value</th>
<th>TPR</th>
<th>Cor</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0020033</td>
<td>BP</td>
<td>Antigenic variation</td>
<td>288</td>
<td>1.0E−75</td>
<td>82.6% (84/104)</td>
<td>0.72</td>
</tr>
<tr>
<td>GO:0007154</td>
<td>BP</td>
<td>Cell communication</td>
<td>290</td>
<td>3.2E−44</td>
<td>50.0% (84/168)</td>
<td>0.73</td>
</tr>
<tr>
<td>GO:0007155</td>
<td>BP</td>
<td>Cell adhesion</td>
<td>50</td>
<td>3.3E−23</td>
<td>60.0% (17/28)</td>
<td>0.83</td>
</tr>
<tr>
<td>GO:0009266</td>
<td>BP</td>
<td>Response to temperature</td>
<td>70</td>
<td>4.0E−6</td>
<td>42.9% (6/14)</td>
<td>0.71</td>
</tr>
<tr>
<td>GO:0016200</td>
<td>BP</td>
<td>Cell invasion</td>
<td>87</td>
<td>6.1E−16</td>
<td>100% (10/10)</td>
<td>0.69</td>
</tr>
<tr>
<td>GO:0016052</td>
<td>BP</td>
<td>Carbohydrate catabolism</td>
<td>82</td>
<td>3.1E−10</td>
<td>50.0% (1/2)</td>
<td>0.79</td>
</tr>
<tr>
<td>GO:0020020</td>
<td>CC</td>
<td>Food vacuole (sensu apicoplasta)</td>
<td>213</td>
<td>1.5E−7</td>
<td>100% (7/7)</td>
<td>0.73</td>
</tr>
<tr>
<td>GO:0020030</td>
<td>CC</td>
<td>Infected host cell surface knob</td>
<td>46</td>
<td>1.3E−24</td>
<td>76.2% (16/21)</td>
<td>0.83</td>
</tr>
<tr>
<td>GO:0003052</td>
<td>CC</td>
<td>Protoplasma complex (sensu eukarya)</td>
<td>75</td>
<td>3.5E−21</td>
<td>76.2% (16/21)</td>
<td>0.80</td>
</tr>
<tr>
<td>GO:0020036</td>
<td>CC</td>
<td>Rhopty</td>
<td>4</td>
<td>3.3E−12</td>
<td>100% (4/4)</td>
<td>0.97</td>
</tr>
<tr>
<td>GO:0020036</td>
<td>CC</td>
<td>Motor’s cleft</td>
<td>246</td>
<td>1.4E−08</td>
<td>77.0% (10/13)</td>
<td>0.79</td>
</tr>
<tr>
<td>GO:0020007</td>
<td>CC</td>
<td>Apical complex</td>
<td>4</td>
<td>5.0E−11</td>
<td>66.7% (4/6)</td>
<td>0.95</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>MF</td>
<td>Structural molecule activity</td>
<td>446</td>
<td>1.0E−11</td>
<td>52.0% (7/13)</td>
<td>0.60</td>
</tr>
<tr>
<td>GO:0008002</td>
<td>MF</td>
<td>Cytochrome c protein binding</td>
<td>277</td>
<td>6.2E−9</td>
<td>57.1% (8/14)</td>
<td>0.80</td>
</tr>
<tr>
<td>GO:0033774</td>
<td>MF</td>
<td>Motor activity</td>
<td>302</td>
<td>7.5E−8</td>
<td>66.7% (10/15)</td>
<td>0.82</td>
</tr>
<tr>
<td>GO:0028233</td>
<td>MF</td>
<td>Polypeptide activity</td>
<td>93</td>
<td>3.2E−12</td>
<td>32.1% (18/56)</td>
<td>0.72</td>
</tr>
<tr>
<td>GO:004672</td>
<td>MF</td>
<td>Protein kinase activity</td>
<td>242</td>
<td>2.6E−5</td>
<td>26.0% (13/50)</td>
<td>0.81</td>
</tr>
<tr>
<td>GO:004536</td>
<td>MF</td>
<td>Deyxosylamine activity</td>
<td>32</td>
<td>9.0E−9</td>
<td>100% (4/4)</td>
<td>0.89</td>
</tr>
</tbody>
</table>

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.

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.
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 Pf616 [52] and PfE27/25 [25,53], the details regarding the specific signaling mechanisms responsible for the switch from asexual to sexual mode of replication remain elusive. Recently, Gardiner et al. identified a gene, *pfxis* (PF11720w), that when genetically manipulated through silencing or over-expression, resulted in a respective 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 the systematic clustering of genes and inter-tissue expression analysis 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 Pf616 [52] and PfE27/25 [25,53], the details regarding the specific signaling mechanisms responsible for the switch from asexual to sexual mode of replication remain elusive. Recently, Gardiner et al. identified a gene, *pfxis* (PF11720w), that when genetically manipulated through silencing or over-expression, resulted in a respective 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, gifts to D.A.B. from The Wellcome (Trust Thematic program grant 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 Research Council Research Associateship (Ref 058038), and funding to D.R.C. from the NMRC and 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