Identifying Stage-Specific Genes by Combining Information from two Different Types of Oligonucleotide Arrays

Yin Liu[1], Ning Sun[2], Junfeng Liu[2], Liang Chen[3], Michael McIntosh[4], Liangbiao Zheng[2] and Hongyu Zhao[2,5]

[1]Program of Computational Biology and Bioinformatics, [2]Department of Epidemiology and Public Health, [3]Department of Molecular, Cellular, Developmental Biology, [4]Department of Internal Medicine, [5]Department of Genetics, Yale University

ABSTRACT

The identification of stage-specific genes in the malaria parasite *Plasmodium falciparum* may provide a starting point to identify key elements for the malaria parasite to complete its life cycle. In this study, we address this question through the combined analysis of gene expression data collected from two distinct microarray platforms. Although it is intuitive that a joint analysis is likely to be more informative than that based on a single source, such analysis faces many statistical challenges in addition to different sets of genes may be probed on different platforms. First, the platforms are sufficiently different that it is difficult to correlate expression levels measured on different platforms. Second, the time resolution of the two data sets differs. To address these challenges, we have developed novel statistical methods to integrate these two distinct platforms. Based on our methods, we have identified genes that are either uniquely expressed or differentially expressed at the sporozoite and gametocyte stages. Some of these genes are known to be specific at these two stages and some are novel, providing potential candidates for transmission-blocking vaccine development. We also analyze the functions of the identified genes based on Gene Ontology (GO) classification and investigate the predicted interacting proteins. The detailed results are available at http://bioinformatics.med.yale.edu/CAMDA2004.

Keywords: microarray, sporozoite, gametocyte, nonparametric regression, gene ontology, ortholog

1. INTRODUCTION

DNA microarray technology allows the transcription levels of many genes to be measured simultaneously, and different microarray platforms are commonly used in gene expression studies. For example, in the analysis of *Plasmodium falciparum*, the DeRisi group used microarrays based on long (70-nucleotide) oligonucleotides to quantify the relative mRNA levels of 4,488 predicted *Plasmodium falciparum* genes at 46 time points across the complete asexual intraerythrocytic developmental cycle (IDC) or asexual blood stages at a 1-hour resolution[1]. Independently, the Winzeler group employed the Affymetrix (25-nucleotide) array to examine the gene expression profiles at 6 periodic asexual blood stages, including early ring, late ring, early trophozoite, late trophozoite, early schizogony, and late schizogony stages. The parasite samples were synchronized by two independent methods: a 5% D-sorbitol treatment and a temperature cycling incubator. Besides the asexual blood stages, the gene expression levels were also measured at the gametocyte and sporozoite stages[2]. Our objective in this study is to identify genes either uniquely or differentially expressed in sporozoites and gametocytes. In our study, the genes not expressed at the asexual blood stages but expressed in sporozoites/gametocytes are defined as the genes uniquely expressed at these two stages, while the genes differentially expressed in sporozoites/gametocytes represent the genes constitutively expressed at the blood stages and upregulated in sporozoites/gametocytes. Although the Winzeler data itself can be used alone to address this question, the higher resolution of the DeRisi data may offer additional information on gene expression during the asexual stages. Therefore, we have developed statistical methods to combine information from these two studies to fully exploit the expression data from these two different data sources. Although our methods are developed in the context of analyzing these two specific data sets, the general approach may prove useful for other similar studies in order to discover novel gene regulation patterns, and to validate previous gene expression profiles. The genes identified to be uniquely or differentially expressed at the sporozoite and gametocyte stages may lead researchers to identify potential candidates for transmission-blocking vaccine development because the sporozoites are the infectious form injected to human blood by mosquitoes, and the gametocytes are the form by which the parasite is transmitted from human to mosquitoes.

2. METHODS

2.1 Pre-processing of the Data

For the Winzeler data, the 17 CEL files are processed using Affy R[3]. The intensity levels of the two sporozoite replicates are averaged after normalization. For the DeRisi data, in which the expression values were obtained from two-color microarray experiments with a common reference used on all the arrays, we perform the print-tip group loess normalization method within arrays by using the Limma package [4, 5]. After normalization, the intensity values and log ratio values are averaged for a subset, including 8 time points that had more than one hybridization result.

2.2 Identification of genes uniquely expressed at the sporozoite/gametocyte stages

Our first objective is to identify genes uniquely expressed at the sporozoite/gametocyte stages, i.e. genes that are not expressed across the asexual blood stages but expressed at the sporozoite/gametocyte stages. Because the DeRisi data did not cover the sporozoite/gametocyte stages, it is not informative on its own for the identification of these genes. On the other hand, although the Winzeler data can be used to address this question, some genes that are expressed at the asexual blood stages data may be missed due to the lower resolution throughout the asexual stages in the Winzeler data. Our strategy is to first use the DeRisi data to identify genes not expressed at the asexual stages and then use the Winzeler data to examine, among this set of genes, which genes are expressed at the sporozoite/gametocyte stages. First, we need to define an objective criterion to infer whether a gene is expressed or not across the blood stages based on the DeRisi data. To achieve this goal, we utilize the 281 “EMPTY” spots on the DeRisi arrays as negative controls. For each channel, the intensities of all the “EMPTY” spots are standardized to have a mean of 0 and variance of 1 through linear transformation. The standardized intensities across all the 46 time points are then summarized. The density distributions of the standardized intensity levels for the red channel and the green channel have very similar patterns (Fig.1). Because some of the “EMPTY” spots may hybridize and yield positive signals (as suggested by the long right tails in Fig. 1), we remove the spots corresponding to the upper 10% of the distribution, leaving 252 “EMPTY” spots serving as negative controls in our following analysis.

For each time point t, we calculate the mean *empMean*$_t$ and variance *empVar*$_t$ of the red channel intensities of the 252 “EMPTY” spots, and then we standardize the intensities for all other spots on the arrays by

$$R_{i,t,std} = \frac{R_{i,t} - empMean_t}{\sqrt{empVar_t}},$$

where $R_{i,t}$ represents the intensity value of spot *i* at time point t. The standardized intensities are summarized across all the 46 time points, so we get the value of $R_{i,std} = \sum_{t=1}^{46} R_{i,t,std}$ for each “EMPTY” spot. The 95% percentile of these values was chosen as the expression cutoff. The genes corresponding to the spots that have summarized intensities across all the 46 time points below this cutoff are considered as genes not expressed at the blood stages.

For the Winzeler data, we need to identify genes expressed at the sporozoites/gametocyte stages. Similar to the DeRisi data, we need to choose an intensity value as cutoff to infer whether a gene is expressed or not at a specific stage. Because there are no “EMPTY” spots that we can use to derive an expression cutoff for the Winzeler data, we have to resort to other methods in our analysis. To this end, we assume that the proportion of genes not expressed at the blood stages based on the Winzeler data is the same as that based on the DeRisi data. Our previous analysis on the DeRisi data yield the result that 17% of genes are not expressed at the blood stages. Based on our assumption, we get the maximum value of the 17% percentile of gene expression levels for each of the 6 blood stages obtained from the Winzeler data and increase the value to a certain extent so that 17% of the genes can be identified as not expressed at the blood stages with taking the adjusted value as the cutoff. The genes with intensity values in the sporozoites/gametocytes above the expression cutoff are considered as genes expressed at the sporozoite/gametocyte stages. Among this set of genes, those not expressed at the blood stages are identified as genes uniquely expressed at these two stages.

2.3 Identification of genes up-regulated at the sporozoite/gametocyte stages

In contrast to the identification of genes uniquely expressed at the sporozoite/gametocyte stages where only a cut-off is needed to infer whether a given gene is expressed or not at a given stage, the inference of expression level changes from the combined analysis of two distinct platforms is more difficult. This requires the establishment of correspondence of measured intensity levels between the two platforms. If the two sets of data had been collected at the same time points, such analysis would be relatively straightforward if we assume that the expression level of the same gene is rather similar in the two experiments. However, the DeRisi data and the Winzeler data have rather different resolutions with 46 time points in the DeRisi data and only 6 time points in the Winzeler data across the asexual stages. To address this problem, we first identify a set of "invariant" genes, which are constitutively expressed at the asexual stages and use the measured expression levels of these genes to derive the correspondence of measured expression levels between the two datasets. For the DeRisi data, the variances of the log-ratio values $\log_2(Cy5/Cy3)$ are calculated for each expressed gene and the set of genes with a variance below a specific cutoff, 0.2 in this study, are considered as the "invariant" gene set. Similarly, the “invariant” gene set for the Winzeler data can be identified after the variances of the intensity values at the 6 blood stages are calculated. Genes common in both invariant gene sets are then selected. As the expression levels of these genes were relatively constant across the blood stages in both datasets, we calculate the mean of the gene expression values at the blood stages for each gene both based on the DeRisi data and the Winzeler data. We then apply the local linear regression method to capture the relationship between the gene expression values obtained from the DeRisi data and those obtained from the Winzeler data through $\min_{\alpha,\beta}\sum_{i=1}^{n}\{y_i-\alpha-\beta(x_i-x)\}^2 w(x_i-x;h)$.

Here, the kernel function $w(x_i-x;h)$ ensures that the observations whose covariate values x_i close to the point *x* are given the most weights in determining the estimate, and the smoothing parameter *h* controls the degrees of smoothing applied to the data[6]. The local linear estimator is,

$$\hat{m}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{\{s_2(x;h)-s_1(x;h)(x_i-x)\}w(x_i-x;h)y_i}{s_2(x;h)s_0(x;h)-s_1(x;h)^2},$$

where $s_r(x;h)=\left\{\sum(x_i-x)^r w(x_i-x;h)\right\}/n$. The results are shown in Figure 2. Based on this nonparametric regression model, we may use the gene intensities at the sporozoite/gametocyte stages obtained from the Winzeler data to predict the values that would have been collected through the DeRisi platform. These predicted values are then compared to the measured intensities throughout the blood stages in the DeRisi data to identify genes differentially expresses at these two stages. In our study, the genes with constant expression levels at the blood stages and expression levels increased at least 1.5 fold at the sporozoite/gametocyte stages compared to the blood stages are considered as genes up-regulated at these two stages. Down-regulated genes are not considered at the two stages because we are only interested in identifying the genes directly related to the transmission between human and mosquitoes.

2.4 Gene Ontology Analysis

Gene Ontology (GO) annotations are downloaded from PlasmoDB (http://plasmodb.org). There are 2,199 gene products (about 41% of the whole proteome) that have been assigned GO terms. We map the GO terms to the more generalized or high-level terms (GO slim terms) to gain a high-level view of gene functions. The sporozoite and gametocyte stage-specific genes are compared to the overall genes based on GO annotations using GO slim terms, and the comparisons are performed in the three GO ontologies - "molecular function", "biological process" and "cellular component". As not all the gene products were assigned a GO term, we rescale the percentages of the proteins in each GO category so that the total is 100%.

To assess the statistical significance for the GO term enrichment of the sporozoite and gametocyte stage-specific genes, we investigate whether the list of identified genes have any GO term overrepresented in their annotation compared to what would be expected by chance from the population of all the genes in *P.falciprum.* The p-value is calculated from the hypergeometric distribution as following:

$$p-value=\sum_{x}^{n}\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

where *N* represents the total number of genes in the population, in which *N* have a particular GO term annotation. And *n* and *x* represent the number of genes in the list of interst and the number of genes in the list annotated with the particular GO term, respectively. The p-value is corrected for multiple testing using Bonferroni correction, a conservetive approach. There are 9 GO slim terms tested for both “molecular function” and “biological process” ontology, and 7 GO slim terms tested for “cellular component” ontology. These numbers are used for correcting the p-values. The list of sporozoite/gametocyte stage-specific genes are considered as have a GO term overrepresented compared to the overall genes if the corrected p-value is less than 0.05.

2.5 Protein-protein interaction pairs in *P.falciparum*

To study whether proteins coded by genes uniquely/differentially expressed at the sporozoite/gametocyte stages interact with each other, we utilize the interaction data from yeast because there is a lack of data for *P. falciparum*. More specifically, we perform “all-against-all” BLASTP comparisons of sequences of the *Sacchromycces cerevisiae* and *P. falciparum* proteomes, and the program INPARANOID[7] is applied on the BLASTP results to identify orthologous groups. Sequence pairs with reciprocal best hits are identified as putative ortholog pairs, and the sequences from the same species that are more similar to the putative orthologs than to any other sequences are considered as “paralogs”, belonging to the same group of orthologs. Based on the concept of “interolog’[8], we assume that if protein A and protein B interact in *S. cerevisiae* and have corresponding orthologs A’ and B’ in *P. falciparum*, then A’ and B’ would form an interacting protein pair in *P. falciparum*. We use the interaction dataset for *S. cerevisiae* in the MIPS[9] database to predict interacting protein pairs in *P. falciparum* by transferring the protein interaction information between the two species.

3. RESULTS

3.1 Genes uniquely or differentially expressed at the sporozoite and gametocyte stages

The Winzeler data includes results generated from two different procedures to synchronize *P. falciparum*. We identify sporozoite/gametocyte stage-specific genes using data generated from both synchronization procedures. Table 1 summarizes the results of our study. As shown in Table 1, both synchronizations yield similar results with an almost complete overlap between different synchronizations.

A total of 408 genes are found to be expressed at the sporozoite stage but not expressed at the asexual blood stages, and 118 genes are constitutively expressed at the asexual blood stages and up-regulated at the sporozoite stage. Among these genes, some of them are experimentally known to be sporozoite specific. For example, the sporozoite surface protein 2 and the circumsporozoite surface protein are well-known markers of the sporozoite stage and are included in our identified gene set.

Similarly, a total of 124 genes constitutively expressed at the asexual blood stages are up-regulated at the gametocyte stage. An additional set of 335 genes is identified as expressed at the gametocyte stage but not at the asexual stages. Included in this list are well-known gametocyte-specific genes, such as those encoding meiotic recombination protein DMC1 and 25kDa ookinate surface antigen. Compared with the results in the Winzeler study, our gene set includes 76% of 61 genes identified as sporozoite specific and 69% of 210 genes identified as gametocyte specific in the Winzeler study respectively.

Besides genes that are known to be stage-specific, we have also identified some genes that have not previously been shown as sporozoite- or gametocyte-specific in the Winzeler study. For example, the protein encoded by MAL13P1.304 is a potential malaria surface antigen and was identified as up-regulated at the sporozoite stage in our results. In addition, MAL6P1.195, encoding a RNA-binding protein MEI2, has been found to be specifically expressed in gametocytes in our analysis. Although the proteins encoded by these genes have been identified as sporozoite/gametocyte specific in the proteomics study based on mass spectrometry data[10], these genes were not identified as sporozoite- or gametocyte-stage specific in the Winzeler study. Therefore, our methods may provide a more comprehensive list of stage-specific genes that are worthy of further investigation and may represent potential candidate targets for the development of transmission-blocking vaccines.

3.2 Gene Ontology Classification

The comparisons of GO annotations with high-level GO terms between the sporozoite/gametocyte stage-specific genes and the overall genes are shown in Fig. 3a-3c, and the list of GO terms associated with a significant p-value are provided in Table 2.

In the “molecular function” ontology, a higher percentage of proteins encoded by the sporozoite/gametocyte uniquely expressed genes are assigned to the “defense/immunity protein" and "cell adhesion" categories compared to the overall gene products. And the statistical analysis provides the evidence that the sporozoite/gametocyte uniquely expressed genes have these two GO terms overrepresented (Table 2). This result is reasonable as the genes specific in sporozoites/gametocytes are involved in the evasion of the host immune system and the cell communication process.

Among all the categories in the "biological process" ontology, the identified stage-specific genes, including 34% of sporozoite specifically expressed genes and 24% of gametocyte specifically expressed genes are over-represented in the "cell communication" category with p-values of 2.60E-13 and 5.05E-7 respectively. These cell communication related genes are known to be involved in "host-pathogen interactions" or "cell-cell adhesion" processes, which may reflect the specific processes relevant to the sporozoite and gametocyte stages[11].

In the "cellular component" ontology, a higher percentage of sporozoite specific gene products belong to the "extracellular" category (with p-value of 9.46E-13). More detailed analyses reveal that this is mainly due to the large number of erythrocyte membrane protein 1 and rifin genes in our identified gene set, and these genes have been shown as sporozoite/gametocyte specific in previous studies[2, 10].

We also compare the GO enrichment of our identified genes with the results from the Winzeler study. We select the genes identified as gametocyte specific in the Winzeler results but are not included in our identified gene set and perform GO analysis on these genes (Fig. 4). According to the "molecular function" and "biological process" ontologies, these genes do not show different GO term enrichment compared to the overall gene products (with p-values larger than 0.05, supplementary data online). This suggests that these genes as a group are different from the genes identified from our set.

3.3 Correlate protein interaction with gene expression

Based on comparative study, only 935 *P. falciparum* proteins have corresponding *S. cerevisiae* orthologs, and a total of 646 interacting protein pairs among these 935 proteins are predicted based on the ortholog list. There may be correlation between expression patterns among the interacting protein partners because the functionality of the interacting pairs depends on the presence of two proteins participating the interaction. To test our hypothesis, we study the number of interacting protein pairs among the sporozoite and gametocyte stage-specific genes and the results are summarized in Table 3.

Because there are 15 proteins having more than 5 interacting partners, we evaluate the statistical significance of the observed number of interacting pairs through simulations after removing these so-called “hub” proteins. More specifically, we randomly select the same number of proteins from the ortholog list (e.g. 62) and record the number of interactions among these randomly selected proteins. This procedure is repeated 10,000 times and the statistical significance of the observed number of interacting pairs can be estimated based on the 10,000 simulated results. As shown in Table 2, there is marginal evidence suggesting that the proteins in the list are more likely to interact with each other than expected by chance.

Only a small number of genes with orthologs in *S. cerevisiae* are found to be sporozoite/gametocyte stage-specific, resulting in the identification of only a few protein-

protein interactions at these two stages. Of the 5 interacting protein pairs found in gametocytes or sporozoites, 4 were found to be common at both gametocyte and sporozoite stages (MAL7P1.50 and PF07_0139, PF00_0002 and MAL7P1.145, PF10_0258 and PF11_0481, PF10_0258 and PF14_0030). Among them, both PF00_0002 and MAL7P1.145 play a role in DNA mismatch repair, an important process for *P. falciparum* reproduction at the gametocyte stages and perhaps required in sporozoites in preparation for extensive DNA replication and schizogony, which occurs following the invasion of hepatocytes.

4. CONCLUSIONS AND DISCUSSION

The identification of stage-specific genes provides a starting point to identify key regulatory elements essential for the malaria parasite to complete its life cycle. In this article, we have developed statistical methods to combine information from two datasets generated under distinct microarray platforms to identify genes either uniquely expressed or differentially expressed at the sporozoite/gametocyte stages compared to the asexual stages. Our identified genes show significant enrichment for certain Gene Ontology categories related to the functions and processes involved in the sporozoite/gametocyte stages. Although the genes identified in our study have a high degree of overlap with those from the Winzeler study, we did not observe any functional enrichment for those genes identified in the Winzeler study but not in our analysis, suggesting that our methods have a higher degree of specificity. By combining information for two different sources, we were able to take advantage the higher resolution of the DeRisi data (as compared to the Winzeler data) to study gene expression patterns at the two stages that were only collected in the latter study. This combined analysis allowed us to identify a larger number of genes that are up-regulated at the gametocyte and sporozoite stages than that based on one data source where the time resolution is low. It is conceivable that even more information can be extracted from other data sources, if they become available, to better understand the mechanisms responsible for the transmission of this protozoan malaria.

Only a small number of genes with orthologs in *S. cerevisiae* were found to have different expression patterns in gametocytes and sporozoites, resulting in the identifications of only a small number of protein-protein interactions. Our simulation results indicated marginal evidence of increased likelihood of interactions among stage specific proteins. These interacting proteins may serve as effective targets for blocking transmission by antimalaria drug or vaccine development, as they are likely to be involved in both sexual stage development as well as invasion of the human host.

5. ACKNOWLDEGMENTS

We thank Nianjun Liu and Naixin Li for help discussion regarding this study. This work was supported in part by NSF grant DMS-0241160.

6. REFERENCES

[1] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J and DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol*. 1(1): E5, 2003.

[2] Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ and Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301(5639): 1503-8, 2003.

[3] http://biosun01.biostat.jhsph.edu/Eririzarr/Raffy/

[4] Yang YH, Dudoit S, Luu P and Speed TP. Normalization for cDNA microarray data. *SPIE BiOS* 2001.

[5] http://bioinf.wehi.edu.au/limma

[6] Bowman A and Azzalini A. Applied smoothing techniques for data analysis, Clarendon Press, Oxford, 1997.

[7] Remm M, Storm CE and Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* 314(5): 1041-52, 2001.

[8] Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M and Gerstein M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* 14(6):1107-18, 2004.

[9] Mewes HW, et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research.* 32 Database issue: D41-4, 2004.

[10] Florens L, Washburn M, *et al.*A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419: 520-526, 2002.

[11] Gardner M, *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature* 419: 498-511, 2002.

7. FIGURES and TABLES

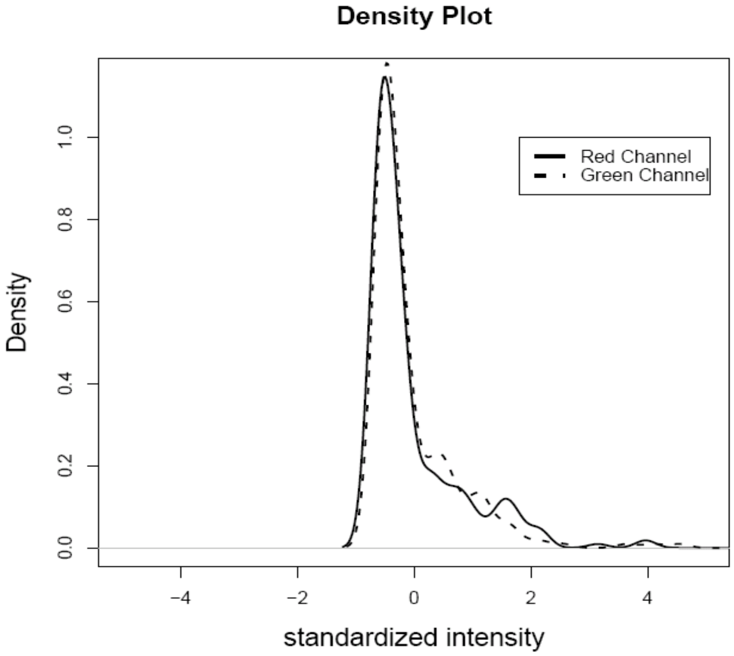


Figure 1. Density plot of the intensities in red and green channels of the "EMPTY" spots. The intensities of the 281 "EMPTY" spots have been transformed to a common distribution with mean 0 and variance 1 and summarized across all the 46 time points.

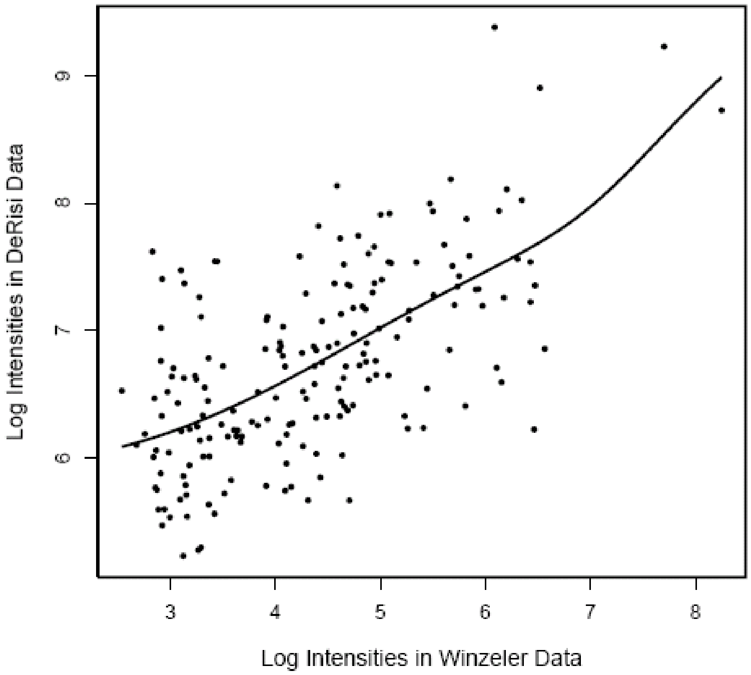
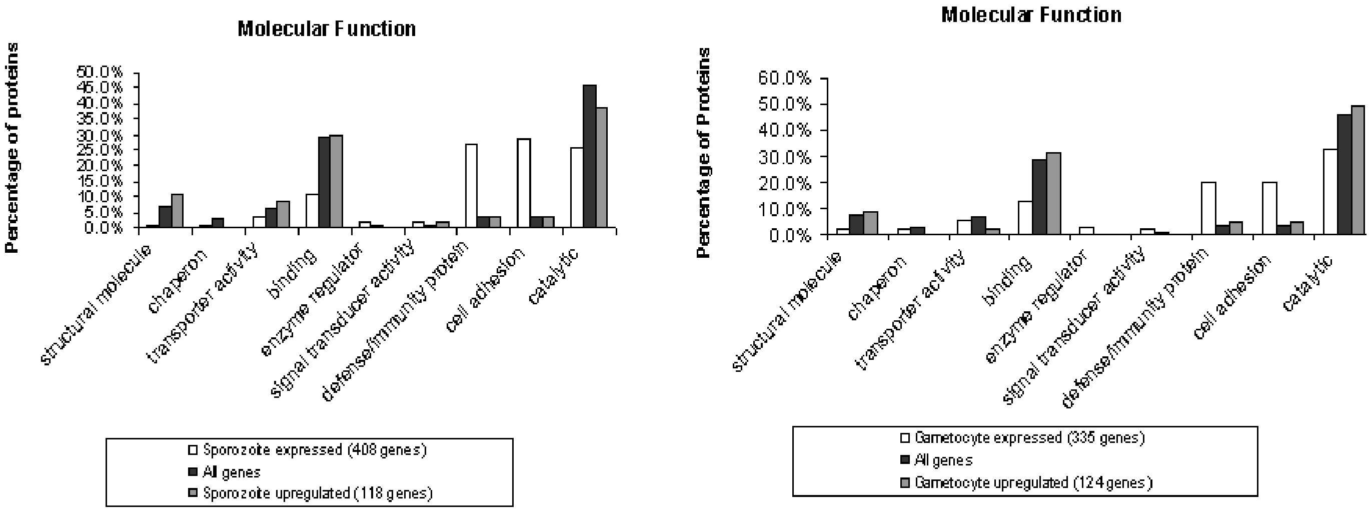
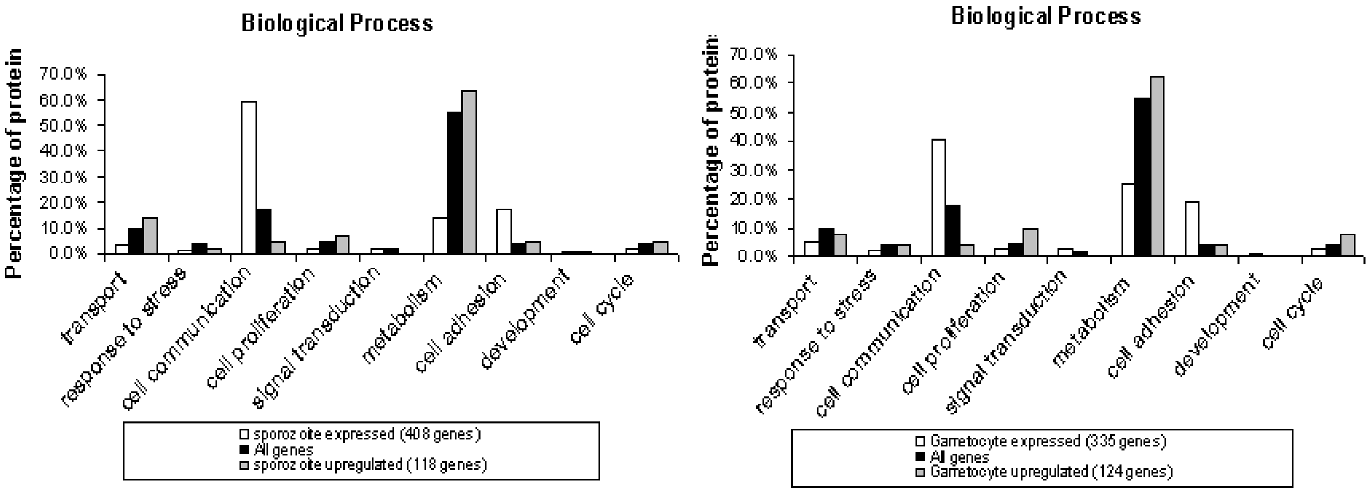


Figure 2. The nonparametric regression curve for the log intensities of all the "invariant" genes at the blood stages obtained from the DeRisi data and the Winzeler data. The smoothing parameter *h* used for control the degrees of smoothing is 1. The log intensities of "invariant" genes obtained from the Winzeler data are based on one synchronization method that uses a 5% D-sorbitol treatment.

(a)


Molecular Function
Percentage of proteins
50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
structural molecule
chaperon
transporter activity
binding
enzyme regulator
signal transducer activity
defense/immunity protein
cell adhesion
catalytic
□ Sporozoite expressed (408 genes)
■ All genes
▨ Sporozoite upregulated (118 genes)
Molecular Function
Percentage of Proteins
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
structural molecule
chaperon
transporter activity
binding
enzyme regulator
signal transducer activity
defense/immunity protein
cell adhesion
catalytic
□ Gametocyte expressed (335 genes)
■ All genes
▨ Gametocyte upregulated (124 genes)

(b)


Biological Process
Percentage of protein
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
transport
response to stress
cell communication
cell proliferation
signal transduction
metabolism
cell adhesion
development
cell cycle
□ sporozoite expressed (408 genes)
■ All genes
▨ sporozoite upregulated (118 genes)
Biological Process
Percentage of protein
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
transport
response to stress
cell communication
cell proliferation
signal transduction
metabolism
cell adhesion
development
cell cycle
□ Gametocyte expressed (335 genes)
■ All genes
▨ Gametocyte upregulated (124 genes)

(c)

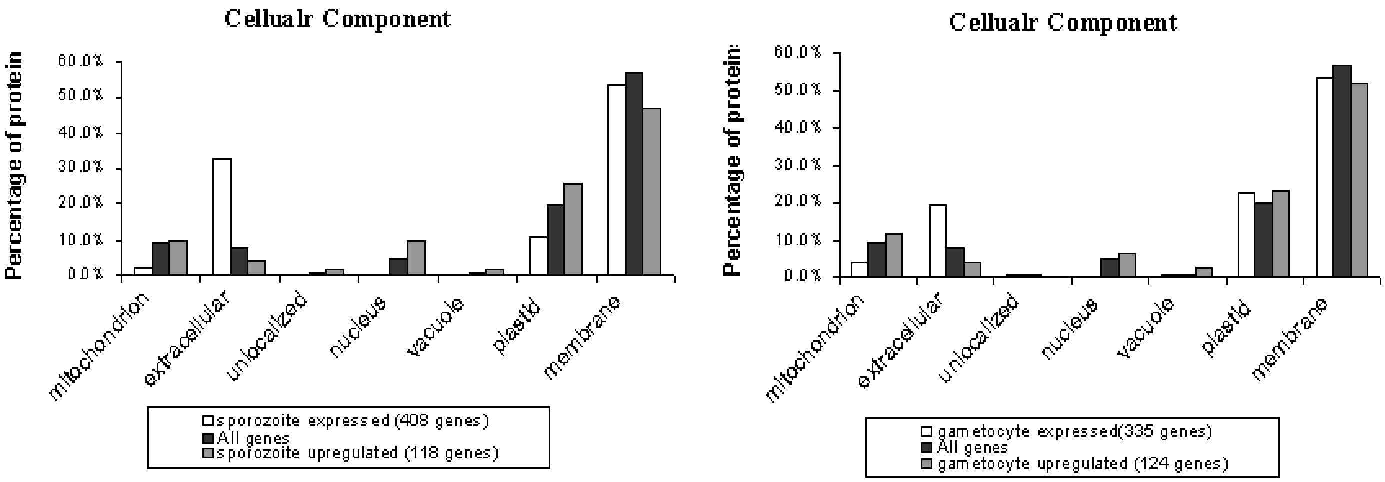


Figure 3. Gene Ontology classifications of *P.falciparum* sporozoite and gametocyte stage-specific genes according to the "molecular function" (a), "biological process" (b) and "cellular component" (c) ontologies of the GO system. The percentages of the proteins encoded by the stage-specific genes in each of the high-level GO categories are compared with that of all the *P. falciprum* genes.

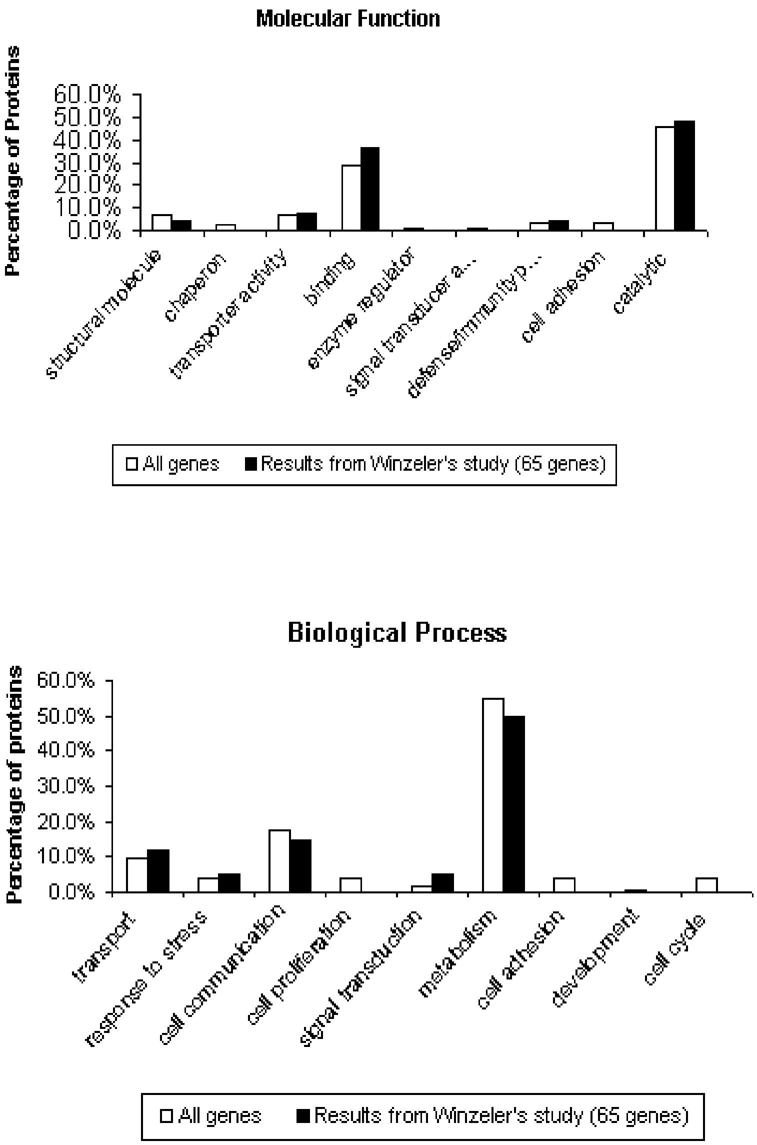


Figure 4. Comparisons between the genes identified as gametocyte specific in the Winzeler results but not included in our identified gene set and all genes according to the "molecular function" and "biological process" ontologies of the GO system.

Table 1. The Number of sporozoite and gametocyte stage-specific genes. In the category of "Constituively expressed", the genes upregulated at the sporozoites/gametocyte stages are listed. In the category of "Not expressed", the genes uniquely expressed at the sporozoite/gametocyte stages are listed.

Expression pattern at the asexual blood stage	Sporozoite			Gametocyte		
	Sync1	Sync2	Overlap	Sync1	Sync2	Overlap
Constitutively expressed	120	139	118	124	140	124
Not expressed	418	411	408	346	339	335
Total	538	550	526	470	479	459

Table 2. The list of GO terms overrepresented by sporozoite and gametocyte stage-specific genes. The p-values are calculated from hypergeometric distributions and corrected for multiple testing using Bonferroni correction. The GO terms associated with a corrected p-value less than 0.05 along with the corresponding gene set are listed. The full list of GO terms associated with their p-values is available online.

GO Term		Gene Set	Corrected p-values
Molecular Function	Defense/immunity protein	Sporozoite Expressed	1.24E-11
		Gametocyte Expressed	1.40E-9
	Cell adhesion	Sporozoite Expressed	5.91E-12

		Gametocyte Expressed	5.75E-10
Biological Process	Cell communication	Sporozoite Expressed	2.60E-13
		Gametocyte Expressed	5.05E-7
	Cell adhesion	Sporozoite Expressed	5.91E-12
		Gametocyte Expressed	5.75E-10
Cellular Component	Extracellular	Sporozoite Expressed	9.46E-13

Table 3. Interacting protein pairs in sporozoites and gametocytes. The empirical statistical significance is calculated as the fraction of the 10,000 permutations having a larger number of protein pairs than that based on the observed data.

	Number of proteins having yeast orthologs	Number of protein pairs within the gene set	Empirical statistical significance
Sporozoites	62	5	0.0405
Gametocytes	54	5	0.0396
Whole Proteome	935	646	