

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

Yin Liu

Program of Computational Biology and
Bioinformatics

Yale University
300 George St. Suite 511
New Haven, CT 06520
(203)785-3711

Yin.liu@yale.edu

Ning Sun

Departments of Epidemiology and
Public Health

Yale University
300 George St. Suite 511
New Haven, CT 06520
(203)785-3695

Ning.sun@yale.edu

Michael T. Mcintosh

Department of Internal Medicine
Yale University School of Medicine

60 College St., LEPH 606
New Haven, CT 06520
(203)785-3458

Michael.mcintosh@yale.edu

Lianbiao Zheng

Departments of Epidemiology and
Public Health

Yale University
60 College St., LEPH 507
New Haven, CT 06520
(203)785-2908

Liangbiao.zheng@yale.edu

Hongyu Zhao

Departments of Epidemiology and
Public Health & Genetics

Yale University
60 College St., LEPH 201
New Haven, CT 06520
(203)785-6271

Hongyu.zhao@yale.edu

ABSTRACT

Gene expression profiling in the malaria parasite *Plasmodium falciparum* has been examined with long or short oligonucleotide microarrays. We have developed a powerful method to integrate results generated from these two distinct platforms. Our method was validated by the correct identification of genes known to be differentially expressed in sporozoite and gametocyte stages. In addition, novel genes highly expressed in the two stages were discovered, providing potential candidates for transmission blocking vaccine development. We also analyzed gene functions based on Gene Ontology (GO) classification and investigated the relationship between predicted protein-protein interaction pairs and gene expression profiles.

Keywords

microarray, sporozoite, gametocyte, linear regression model, ortholog.

1. INTRODUCTION

DNA microarray technology allows the transcription levels of many genes to be measured simultaneously. In this study, we combined information from two studies that use different microarray technologies. The data generated by the Derisi group used long (70-nucleotide) oligonucleotides and measured the relative mRNA levels of 4,488 predicted *Plasmodium falciparum* genes across the complete asexual intraerythrocytic developmental cycle or asexual blood stages¹. Another study by the Winzeler group used the Affymetrix (25-nucleotide) array to examine the gene expression profiles in the asexual blood stages as well as in gametocyte and sporozoite stages². Our objective in this analysis was to identify additional genes differentially expressed in sporozoite and gametocyte stages by fully exploiting the expression data from these two different sources. This new approach will be useful for researchers who wish to combine data sets from distinct microarray platforms in order to gain larger sample sizes, discover novel gene regulation patterns, and/or validate previous gene expression profiles.

2. METHODS

2.1 Pre-processing of the Data

For the Winzeler data, the 17 CEL files were read into the Affy

package and normalized using “rma” as the background correction method³. The intensity data of the two sporozoite replicates were averaged after normalization. For the Derisi data, expression values were obtained from two-color microarray experiments with a common reference used on all the arrays. We performed 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 were averaged for the 8 time points represented by more than one array hybridization.

There are 281 “EMPTY” spots in the Derisi’s array. The intensity of these spots were standardized to have a mean of zero and unit length by location and scale transformation. The standardized intensities were then summarized across all the 46 time points for each “EMPTY” spot. We found that 10 percent of these values are not within a normal distribution, possibly due to dye contamination on the array, hence these were removed which left 252 true “EMPTY” spots. After the mean $empMean_t$ and variance $empVar_t$ of the standardized intensities of the true “EMPTY” spots at each time point t were calculated, the intensities for all other spots on the arrays were standardized by

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

where $R_{i,t}$ represents the intensity value of spot i at time point t .

Finally, the standardized intensities were summarized across all 46 time points for each spot and 95% quantile of these values was chosen as the expression cutoff. Spots that had a summarized intensity below this cutoff were considered as “non-expressed” spots. By exploring the correspondence between genes and spots, we could identify the genes that were not expressed in the asexual blood stages.

2.2. Integration of Results Generated by Two Different Array Technologies

We first identified the genes with unchanged expression levels across the blood stages in both datasets. For the Derisi data, the variances of the ratio values $\log_2(Cy5/Cy3)$ were calculated for each expressed gene and the set of genes with a variance below the cutoff, (arbitrarily set as 0.2 in this study), were considered as the “invariant” gene set. Similarly, the “invariant” gene set for the Winzeler data could be identified after the variances of the intensity values at the six blood stages were calculated. Genes common to both invariant gene sets were selected. To obtain comparable gene expression levels across different oligonucleotide arrays, we perform a linear regression analysis on the gene expression values of each gene in the common “invariant” gene set. The linear model used is of the form:

The linear model is of the form: $\log Y = a + b \log X$,

The two microarray datasets were generated with samples collected from blood stages in different manners: six time points

for the Winzeler data and 48 time points at 1h intervals for the Derisi data. To compensate for the differences between the Winzeler and Derisi data collection time points, we took the median expression values for each gene at time 1-3h, 12-15h, 16-19h, 24-28h, 30-33h and 40-45h as their expression values at the corresponding stages: early ring stage, late ring stage, early trophozoite stage, late trophozoite stage, early schizont stage and late schizont stages, respectively to match the Winzeler data set. Hence, the vectors Y and X represent the mean expression values of each gene at six blood stages in the Derisi data set and the Winzeler data set, respectively. The result of linear regression was taken as:

$$\log Y = 5.280 + 0.3648 * \log X$$

Figure 1a shows the regression plot and Figure 1b the normal QQ plot that indicates the residues are normally distributed. Given the regression results and with the expression level in the Winzeler dataset, we predicted the gene expression values for sporozoite and gametocyte stages that are missing in the Derisi dataset. These values were compared with the gene expression levels from the blood stages in the Derisi data allowing differentially expressed genes in these two stages to be identified.

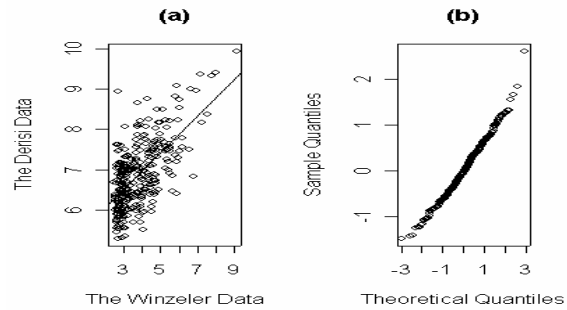


Figure 1. (a) The regression fit for predicting the Derisi data from the Winzeler data. (b) Normal QQ plot of residues.

2.3 Identify Protein Interaction Pairs in *P.falciparum*

We carried out “all-against-all” BLASTP comparisons of sequences of the *Saccharomyces cerevisiae* and *P. falciparum* proteomes, and the program INPARANOID⁶ was applied on the BLASTP results to identify orthologous groups. Sequence pairs with reciprocal best hits were identified as putative ortholog pairs, and the sequence from the same species that are more similar to the putative orthologs than to any other sequences were considered as “paralogs”, belonging to the same group of orthologs. Based on the concept of “interolog”⁷, we assumed that if protein A and protein B interact in *S. cerevisiae* and have the corresponding orthologs A’ and B’ in *P. falciparum*, then A’ and B’ are interacting protein pairs in *P. falciparum*. We used the reliable interaction dataset in the MIPS⁸ database and by transferring the protein interaction information between the two

species; we predicted the protein interaction pairs in *P. falciparum*.

3. RESULTS

3.1 Genes in Sporozoite and Gametocyte Stages

The Derisi dataset includes results generated from two different procedures to synchronize *P. falciparum*. We identified genes that were strongly up-regulated in the sporozoite and gametocyte stages using data generated from both synchronization procedures. As shown in Tables 1 and 2, both synchronizations yielded similar results. Furthermore, the majority of genes were identified in both synchronization datasets.

Table 1 shows the number of genes that were up-regulated in the sporozoites. Many of the genes we identified are experimentally known to be sporozoite specific. For example, the sporozoite surface protein 2 and the circumsporozoite protein are well-characterized sporozoite-specific proteins and are included in the identified gene set. These genes were identified in our list as sporozoite-specific and not expressed in the asexual stage. Winzeler's report identified 108 genes, (cluster one), that were highly expressed in sporozoites. Among these genes, 16 were not included in the Derisi study. With these genes excluded, our identified set of sporozoite-expressed genes has a concordance rate of 76%. In total, we found 751 genes that were up-regulated or only expressed in sporozoites. This compares to 108 highly expressed sporozoite genes identified by the Winzeler group.

Table 1. Numbers of genes up-regulated in the sporozoite stage.

Expression pattern in the asexual stage	Sync 1	Syn2	Total
Constitutively expressed	172	117	172
Fluctuated	135	90	135
Not expressed	437	440	444
Total	744	647	751

Similarly, we found 166 genes constitutively expressed in asexual blood stages and up-regulated in gametocyte stages (Table 2). An additional 369 genes were identified as exclusively expressed in the gametocyte stages (Table 2). Included in this list are well-known gametocyte-specific genes such as those encoding meiotic recombination protein dmc1 and 25kDa ookinete surface antigen. When compared to clusters 2 and 3 in the Winzeler data set that contains a total of 370 gametocyte-specific genes, our set exhibited a concordance rate of 78%.

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, protein MAL13P1.304 is a potential malaria surface antigen and was identified up-regulated in the sporozoites in our results. The protein *P. falciparum* myosin has been shown to appear on the erythrocyte plasma membrane⁹, and is also differentially expressed in both sporozoites and gametocytes as shown in our analysis. These genes may be worthy of further investigation and may represent potential candidate targets for the development of transmission blocking vaccines.

Table 2. Numbers of genes up-regulated in the gametocyte stage.

Expression pattern in the asexual stage	Sync 1	Syn2	Total
Constitutively expressed	166	124	166
Fluctuated	177	130	177
Not expressed	362	363	369
Total	705	617	712

3.2 Gene Ontology Classification

Gene Ontology (GO) classification is used to describe the roles of genes in organisms. We assigned the GO slim terms (high level terms) to 2199 gene products (about 41% of the whole proteome). Genes that are expressed in sporozoites were compared with those that are expressed in blood stages but not in sporozoites based on GO annotation at high level GO terms (Figure 2a). Similarly, the comparison was performed for genes differentially expressed in gametocytes and the results are shown in Figure 2b.

In almost all categories in the “biological process” ontologies, genes differentially expressed in sporozoites or gametocytes don't show different biological process enrichment compared to the genes expressed in asexual blood stages. However, higher values can be seen for the identified stage-specific genes (18% of 751 sporozoite-specific genes and 8% of 712 gametocyte-specific genes) in the “cell communication” category (Figure 2). The enrichment of cell communication related genes was also found to be true for the sporozoite-specific genes identified by Winzeler (data not shown). Many of these genes were found to be involved in “host-pathogen interactions” or “cell-cell adhesion” processes, which may reflect the specific processes relevant to sporozoite and gametocyte stages¹⁰.

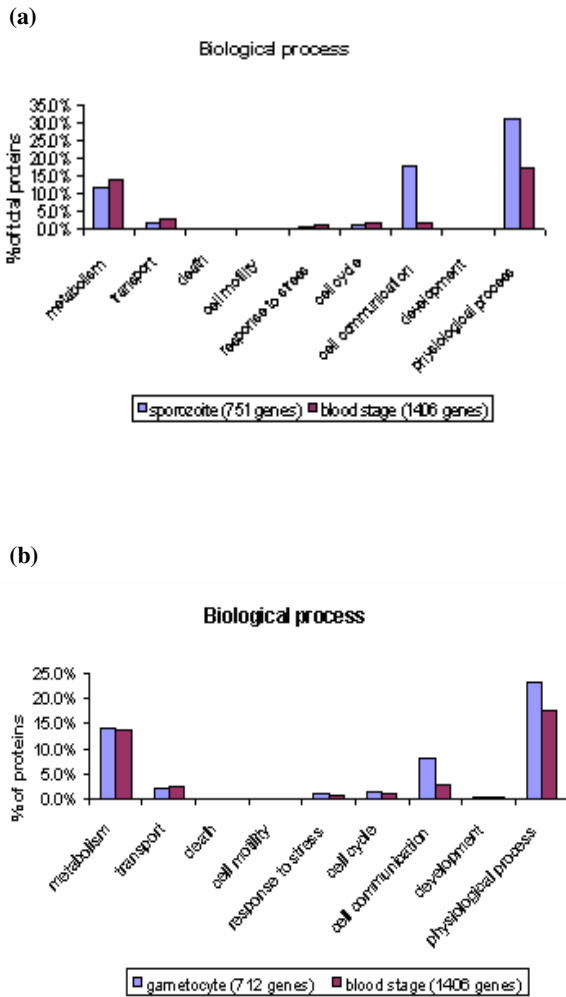


Figure 2. Gene Ontology classifications of *P.falciparum* sporozoite and gametocyte differentially expressed genes according to the “biological process” ontologies of the GO system.

3.3 Correlate Protein Interaction with Gene Expression

We found only 925 *P. falciparum* proteins which have corresponding *S. cerevisiae* orthologs. Looking at these orthologs, we further explored the number of protein-protein interaction pairs in the set of genes differentially expressed in sporozoites and gametocytes to investigate the relationship between predicted protein-protein interactions and gene expression (Table 3).

We found that the predicted protein-protein interaction pairs do not show significant gene coexpression based on our analysis in the sporozoite and gametocytes. Furthermore, the protein interaction pairs existing in the set of genes that are coexpressed

in sporozoites or gametocytes are much less as compared to the total number of protein-protein interaction pairs in the whole proteome (Table 3). Of the 5 coexpressed protein-protein interaction pairs found in gametocytes or sporozoites, four of sporozoite stages (MAL13P1.148 and MAL7P1.162, PF00_0002 and MAL7P1.145, PF11_0097 and MAL13P1.279, PFE0880c and PFI255w). Among them, both PF00_0002 and MAL7P1.145 play a role in DNA mismatch repair, an important process for *P. falciparum* reproduction in gametocyte stages and perhaps required in sporozoites in preparation for extensive DNA replication and schizogony which occurs following the invasion of hepatocytes.

Table 3. Protein interaction pairs in sporozoites and gametocytes

	Number of proteins having yeast orthologs	Number of protein pairs within the gene set
Sporozoites	80	5
Gametocytes	82	5
Whole proteome	925	625

4. DISCUSSION AND CONCLUSION

The identification of stage-specific genes provides a starting point to identify key regulatory elements and transcriptional regulators essential for the malaria parasite to complete its life cycle. It can provide for a better understanding of mechanisms responsible for the pathology or transmission of malaria. Our work has focused on designing a method for combining information from two quite distinct datasets by performing a linear model on the expression values of the overlapped “invariant” gene sets from the two data sources. We examined the sporozoite and gametocyte stages and identified well-known stage-specific genes. The high degree of overlaps in genes identified in our study and those of the Winzeler’s study indicate our integrated approach is reliable. More importantly, we have identified a large number of genes that are up-regulated in the gametocyte and sporozoites, which would be difficult to obtain from one data source alone and further illustrates the power of combinatorial microarray analysis.

Only a small number of genes with orthologs in *S. cerevisiae* were found to be expressed in the gametocytes and sporozoites, resulting in the identification of only a few protein-protein interactions. In any case, these proteins or processes may serve as effective targets for blocking transmission by antimalarial drug or vaccine development as they are likely be involved in both sexual stage development as well as invasion of the human host.

5. ACKNOWLEDGMENTS

We would like to thank Liang Chen and Junfeng Liu for help discussion regarding this study.

6. REFERENCES

- [1] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, 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, 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/Eririzarra/Raffy/>
- [4] Y. H. Yang, S. Dudoit, P. Luu and T. P. Speed. Normalization for cDNA Microarray Data. SPIE BIOS 2001.
- [5] <http://bioinf.wehi.edu.au/limma>
- [6] Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* 314(5): 1041-52, 2001.
- [7] Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* 14(6):1107-18, 2004.
- [8] Mewes HW, et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research.* 32 Database issue:D41-4, 2004.
- [9] Taraschi TF, O'Donnell M, Martinez S, Schneider T, Trelka D, Fowler VM, Tilley L, Moriyama Y. Generation of an erythrocyte vesicle transport system by *Plasmodium falciparum* malaria parasites. *Blood.* 102(9):3420-6, 2003.
- [10] Gardner M, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498-511, 2002.