

Twenty-one genes show specific time dependent modulation of expression during IDC in *P. falciparum*

Boris Fedorov
Molecular Staging Inc.
Suite 701
300 George street
New Haven, CT 06511
(203)752-7904
borisf@molecularstaging.com

Serguei Lejnine
Molecular Staging Inc.
Suite 701
300 George street
New Haven, CT 06511
(203)752-7932
sergeyl@molecularstaging.com

ABSTRACT

We applied several new approaches to analyze CAMDA 2004 datasets of *P. falciparum* asexual intraerythrocytic developmental cycle. Genes with time-modulated expression were called out using a modified variogram. Wavelet decomposition was used for robust smoothing of expression profiles. Twenty-one genes with unusual signal modulation were identified using a phaseogram. Putative functions were assigned to three genes using homology to known proteins. Biological implications of the findings are discussed.

General Terms

Algorithms, measurements

Keywords

Microarray data analysis, normalization, filtering, variogram, wavelet

1. INTRODUCTION

Parasite *P. falciparum* is main cause of human malaria resulting in the deaths of 1.5 to 2.7 million people annually. [1]. A billion people are at risk of malaria exposure in southeast Asia, with risk in Africa hard to estimate [2]. Approximately 5300 proteins of *P. falciparum* were predicted from the 22.8 Mb genome sequence [3]. Thirty-five percent of these proteins have no identified functions. Attempts to predict the function of the unknown proteins have been made by clustering expression profiles [4].

The asexual intraerythrocytic developmental cycle (IDC) of *P. falciparum* is an important stage that can be more easily studied in vitro than other stages. This stage lasts 48 hours, during which the majority of known genes are expressed only once, each in its own particular development phase. Some genes with distinct expression activity may be expressed in different phases of IDC that can play important biological role in the development of this parasite. Identification of such genes could be important in developing new therapeutic treatments that

target parasite proliferation in the blood cells. In this report, we have developed a robust new approach to identify genes with modulated expression. In this analysis, three thousand nine hundred and two genes were identified as modulated. Wavelets decomposition was used to perform a robust fit of gene profiles. We have identified 21 genes with significant modulation differences observed over a period of 48 hours. The function of six of these putative genes was not known. Eighteen genes were assigned to a putative functional/structural group in PlasmoDB. We have assigned functional groups to 3 proteins based on BLAST similarity.

2. MATERIALS AND METHODS

2.1 Initial data processing

Malaria parasite (*Plasmodium falciparum*) life cycle microarray transcriptome data were obtained from the CAMDA 2004 Conference website, available as a contest dataset [5]. Original Gene Pix Result (GPR) data were used for the analysis.

Data from GPR files were merged into a Microsoft Access table including information about values, replicas and gene names. There were a total of 425920 spots. One hundred and forty spots were flagged as bad (missed) spots and excluded from the analysis.

2.2 Data pre-processing

The goal of the analysis is to call genes with detectable measurements demonstrating modulated expression over time. In general, dose response curves are not available for any of the genes. The assumption is that gene expression is measured with a quasi-linear dynamic range if time dependent modulation of gene expression exists. However, since each slide/time point is scanned at different photomultiplier tube (PMT) settings observed modulation might be due to PMT changes. To avoid this artifact, slides were normalized by aligning total sums of log₂ values of each slide for each scanned channel. The assumption of constant total signal is supported by the fact that at every time point no more than 30% of genes are modulated in the Cy5 channel. Signal from the Cy3 channel should be the same based on the study experimental design.

2.3 Preliminary data analysis

Our gene calling method is based on time dependent modulation of intensity. Variability could be dependent on channel and gene intensity. Logarithm transformation of data is a technique known to stabilize variance in physical measurements. Fig.1 shows a scatter plot of average intensity for each gene measured over multiple slides versus standard deviation. Cy3 channel shows the variance is independent of average intensity. Cy5 channel shows uneven dependence due to time dependent modulations. However, the variance of about 50% of the Cy5 data points (covered by the Cy3 channel data, Figure 1) is independent of intensity and similar to the variance observed in the Cy3 channel. This demonstrates variance of the non-modulated genes is similar between different channels.

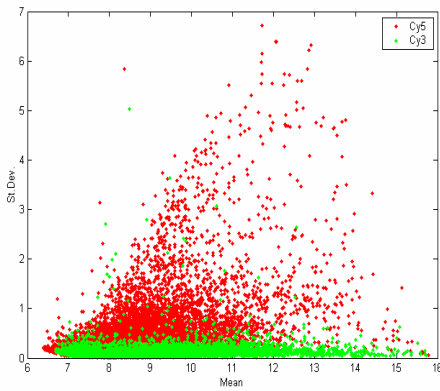


Figure 1. Plot of standard deviation for Cy5 (red) and Cy3 (green) channels vs. average values of log2 intensity.

2.4 Filtering

2.4.1 Variogram

To identify genes with modulated intensity, variability of gene intensity in the Cy5 channel is compared to variability of the gene in the Cy3 channel, which represents gene specific noise. The measure of the variability should be robust to outliers and reflect the cyclical nature of the data. Our goal was to find genes with biologically interesting properties. In order to devise such a measure, we have developed a method similar to that of the variogram [6].

Our method of the variogram consists of an analysis of variance of each gene as a function of time interval. As an example, take a gene with modulated expression (Figure 2).

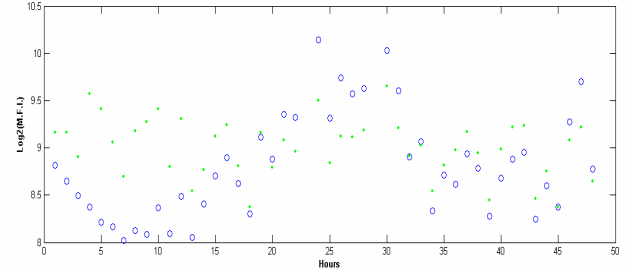


Figure 2. Scatter plot of intensity of gene A11025_2 versus time points for both channels (green is Cy3 and blue is Cy5 channel).

The variogram is defined as

$$V(h) = \frac{\sum_{i,j} \text{stdev}(y_i, y_j)}{N},$$

where i, j – all possible combinations of data points, and $|i-j|=h$,

N – the number of such combinations and $|y_i, y_j|$ – data distribution in interval between i and j points.

The averaged estimate is robust to sharp fluctuations in intensity and shows real tendency of signal.

Typical variograms of a modulated gene are shown in Figure 3 for each channel. The Cy3 (reference channel) variogram is a flat line, that reflects non-modulated intensity. The Cy5 variogram of a modulated gene monotonically increases until sill value.

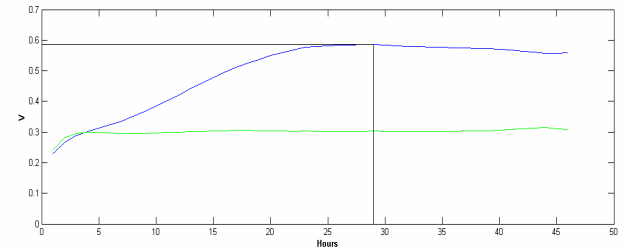


Figure 3. Variograms of Cy5 (blue line) and Cy3 (green line) for gene A11025_2. Intersection of black lines is a critical point of Cy5 variogram or sill value.

Sill value is marked as the intersection of black lines in Figure 3. It characterized by the time interval and peak of the variance. If the peak of the variance in the Cy5 channel is higher than in the Cy3 (reference) variogram, then amplitude of changes in Cy5 channel is higher than in Cy3 channel.

A gene is considered modulated if the difference in sill values between Cy5 and Cy3 channels is three times higher than a sum of sill value of the gene and sill value of an EMPTY feature measured in the Cy3 channel. The EMPTY feature was added to reduce false positives due to extremely low sill values in present in both channels. Three thousand nine hundred and two genes were identified as time modulated genes using this method.

2.4.2. Wavelet Analysis

A time dependent profile of each called gene was fitted using wavelet-based decomposition. Wavelets are superior to methods combining running average and Fourier transform due to improvements in robustness to outliers and sensitivity to irregular signal shapes. Wavelet analysis consists of breaking up a signal into shifted and scaled versions of the original predefined wavelet ([7], [8]).

Wavelet decomposition was used to fit time dependence of the logarithm ratio of Cy5/Cy3 for each called gene. Typical results of wavelet-based fitting are shown on Figure 4.

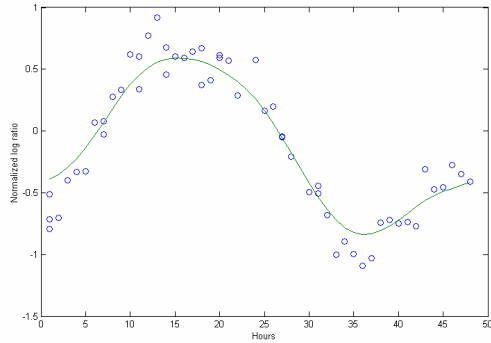


Figure 4. Scatter plot of $\log_2(\text{Cy5}/\text{Cy3})$ as a function of time (blue circles). Solid green line represents wavelet based fitting to a scattered data.

3. RESULTS

Phase of time dependent expression profiles were calculated for each gene using FFT algorithm. Profiles were sorted by phase and plotted on a 3-D surface (Figure 5). Two main maximums correspond to expression profiles with a 48 hour time period and reflect time in IDC phase (Figure 5.). There are several expression profiles with extremes of expression on the diagonal between main maximums (Figure 6.). These genes show multiple expression peaks, which correspond to a cycle with a periodicity different from 48-hour.

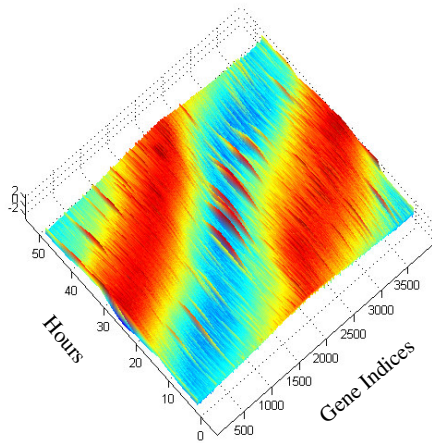


Figure 5. A 3-dimension surface view of expression structure of selected genes sorted by phase.

Twenty-one genes with extreme expression on the diagonal between main maximums were identified (Figure 6.). Time dependent expression profiles are shown in Figure 7.

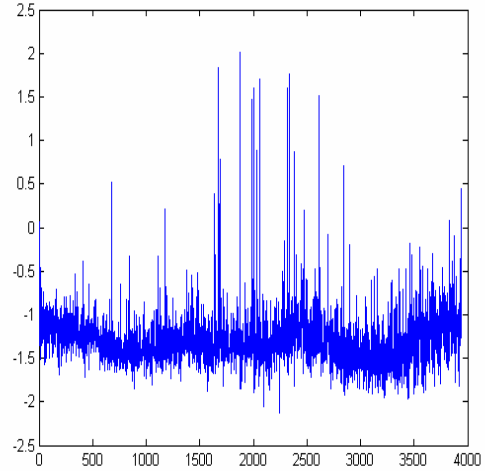


Figure 6. Diagonal section of expression surface.

Each of these genes has at least two peaks in different IDC phases (Figure 7).

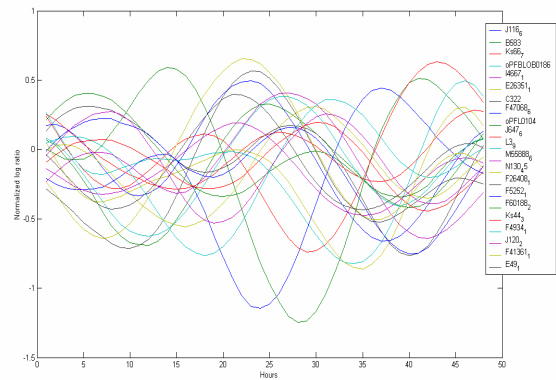


Figure 7. Time dependent profiles of $\log_2(\text{Cy5}/\text{Cy3})$ for 21 genes with extreme point on the diagonal .

We have assigned putative function to three genes PFE0460c, PF10_0097 and MAL8P1.111 (Table 1). Annotation was done using analysis of BLAST results provided by PlasmoDB [9].

Table 1. Proteins with assigned putative function at present

ID	Locus	Description
f47068_6	PFE0460c	<i>P. yoelii</i> lipase precursor gehm. 6/101 (60% similarity)

j647_6	PF10_0097	<i>H. sapiens, M. musculus</i> MITOCHONDRIAL RIBOSOMAL PROTEIN
f60188_2	MAL8P1.111	CCAAT-box DNA binding protein subunit B [Plasmodium falciparum 3D7]

One of the genes PFE0460c showed 60% similarity with lipase precursor from the rodent malaria parasite. Lipase is important for fat conversion and could affect energy utilization processes. MAL8P1.111 is similar to transcriptional factors. It is known that this class of transcription regulators is involved in metabolic regulation by affecting expression levels of growth hormone. Twelve proteins are characterized in PlasmoDB (Table 2). The role and function of 6 proteins are unknown (Table 2).

Table 2. Proteins with known and unknown function

ID	Locus	Description	Putative Function
j116_6	PF10_0347	(AJ252286) merozoite surface protein 3 [Plasmodium reichenowi] 0.31	Known
oPFBLOB0186	PFE0855c	DNA-directed RNA polymerase (EC 2.7.7.6) beta'-2 chain - Plasmodium falciparum plastid 0.27	Known
E26351_1	Unknown	DNA-directed RNA polymerase (EC 2.7.7.6) beta'-2 chain - Plasmodium falciparum plastid 0.27	Known
l3_9	PFL0835w	GTP-binding protein, putative	Known
m55888_6	MAL13P1.216	DNA helicase, putative:: (AB006699) DNA repair protein RAD5 protein [Arabidopsis thaliana] 0.36	Known
f4934_1	MAL8P1.54	DNA-directed RNA polymerase (EC 2.7.7.6) beta'-2 chain - Plasmodium falciparum plastid 0.29	Known
f41361_1	MAL6P1.192	ATP-dependent DEAD box helicase, putative:: (AJ132843) mitochondrial RNA helicase [Arabidopsis thaliana] 0.4	Known
e49_1	MAL6P1.148	6-pyruvoyl tetrahydropterin synthase, putative	Known
OPFL0104	PFL2390c	asparagine-rich protein (clone 25C4) - <i>S. cerevisiae</i> nucleic acid binding activity	Known
c322	PFC0490w	similarity to oligosaccharyl transferase essential subunit	Known
f26408_1	MAL8P1.65	<i>Arabidopsis thaliana</i> putative u3 small nucleolar ribonucleoprotein protein t12p18.20	Known
f5252_1	MAL8P1.65	Similar to <i>Arabidopsis thaliana</i> putative u3 small nucleolar ribonucleoprotein protein t12p18.20	Known
ks66_7	PF11_0473	No NR protein Similarities	Unknown

n130_45	PF14_0547	No description	Unknown
ks44_3	PF11_0350	No description	Unknown
j120_2	PF10_0033	No description	Unknown
B583	Unknown	Unknown	Unknown
I4667_1	Unknown	Unknown	Unknown

4. CONCLUSION

Twenty-one genes were identified with a cycle modulation with a periodicity different from 48 hours. Putative function was assigned to three genes based on BLAST similarity. It is intriguing that there was a surprisingly small number of identified genes (21 of 5300 totally estimated proteins) using this method. Therefore, we recommend additional investigation into better understanding the role these genes might play in the development of *Plasmodium falciparum*.

5. REFERENCES

- [1]. Breman JG, Egan A, Keusch GT. The intolerable burden of malaria: a new look at the numbers. *Am J Trop Med Hyg.* 2001 Jan-Feb;64(1-2 Suppl):iv-vii.
- [2]. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Dis.* 2004 Jun;4(6):327-36.
- [3]. Nirmalan N, Sims PF, Hyde JE. Quantitative proteomics of the human malaria parasite *Plasmodium falciparum* and its application to studies of development and inhibition. *Mol Microbiol.* 2004 May;52(4):1187-99.
- [4]. 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.* 2003 Sep 12;301(5639):1503-8.
- [5]. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol.* 2003 Oct;1(1):E5. Epub 2003 Aug 18.
- [6]. Cressie, N. (1993). *Statistics for Spatial Data*, New York: Wiley, Chapter 2.
- [7]. Daubechies, I. (1992), *Ten lectures on wavelets*, SIAM.
- [8]. Mallat, S. (1998), *A wavelet tour of signal processing*, Academic Press.
- [9]. Kissinger, J.C., et. al. 2002. The *Plasmodium* Genome Database. *Nature* 419: 490-492