

Chromosomal Clustering of Stage-Specific Periodically Expressed Genes in Plasmodium Falciparum

Pingzhao Hu

Celia Greenwood, Cyr Emile M'lan and Joseph Beyene*

Hospital for Sick Children Research Institute
and University of Toronto

**The Fifth International Conference for
the Critical Assessment of Microarray Data Analysis (CAMDA 2004)**

**Duke University
Durham, NC, U.S.A**

November 10-12, 2004

*Contact: joseph@utstat.toronto.edu

Outline

1. Background and Objectives
2. Data Set and Preprocessing
3. Methods & Results
 - 3.1 -- Identification of Periodically Expressed Oligonucleotides
 - 3.2 – Classification of Periodically Expressed Oligonucleotides to Cell-Cycle Stages
 - 3.3 – Chromosomal Clustering of Stage-Specific Periodically Expressed Genes and Brief Functional Analysis
4. Conclusions

1. Background & objective

- **Plasmodium Falciparum is responsible for the vast majority of episodes of malaria worldwide**
 - ❖ *Genomic research on this organism will have far reaching public health implications*
- **Periodic nature of genes expressed in asexual intraerythrocytic development cycle (IDC) of Plasmodium Falciparum has been studied by Bozdech et al., 2003**
- **Our objective is to investigate association between chromosomal location and stage-specific periodical expression of genes expressed in IDC**

2. Data Set and Preprocessing

- Three datasets were provided by CAMDA 2004. We used the quality controlled data set (to facilitate comparison with work by other groups)
- This dataset was previously normalized using NOMAD (NOrmalization of MicroArray Data) system and contains 5080 Oligonucleotides measured at 46 time points spanning 48 hours
- 243 of the Oligonucleotides had a missing value at one or more time points
 - ❖ *We imputed missing data using a 10-nearest neighbor weighting method (Hastie et al. 1999 and Troyanskaya et al. 2003)*
- The Oligonucleotides are scattered over the 14 chromosomes of the *P.falciparum* genome

3.1. Identification of Periodically Expressed Oligonucleotides -- *Model*

- We applied a *multiple linear regression model* to quantify the periodicity for the expression profiles of each oligonucleotide (Booth 2003)

$$y_j = b_0 + b_1 \cos(2\pi j / T) + b_2 \sin(2\pi j / T) + e_j$$

T is the periodicity of the expression profile and b_0, b_1 and b_2 are oligonucleotide-specific parameters to be estimated from the data

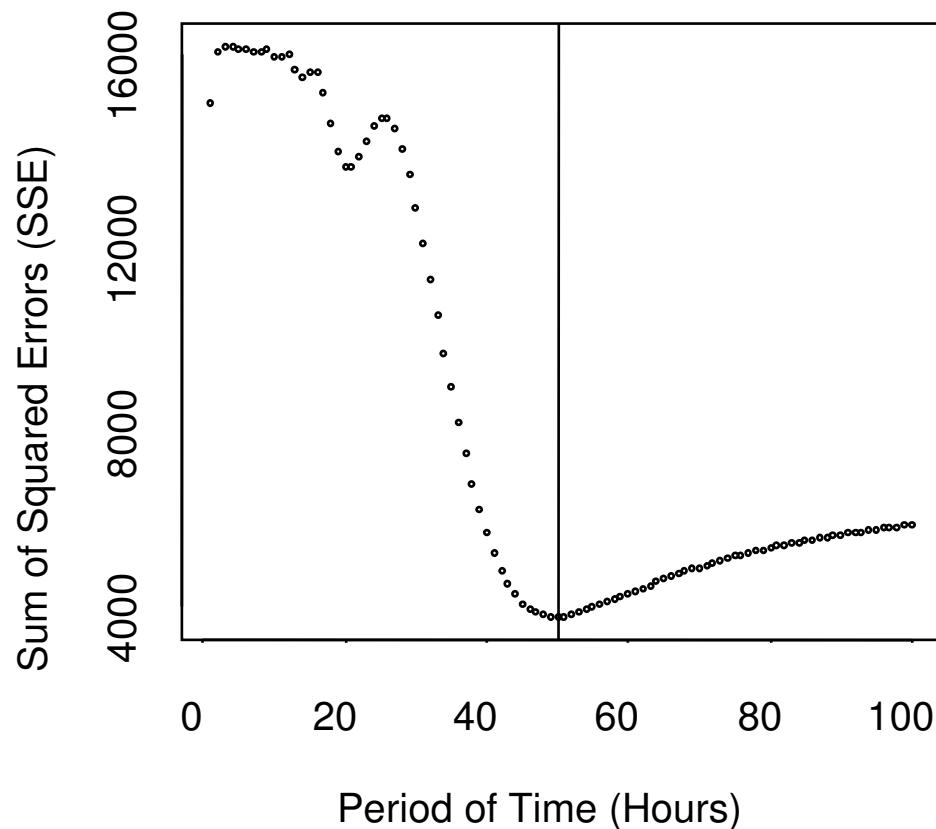
- Estimates of the oligonucleotide specific parameters can be obtained by a *least squares fit*; The period T is first estimated separately
- Goodness-of-fit of the model to each oligonucleotide's expression profile is measured by R^2 , the *proportion of variance explained (PVE)* by the periodicity

3.1. Identification of Periodically Expressed Oligonucleotides – *Estimation of Periodicity T*

- We estimated the periodicity T by **minimizing the sum of squared errors (SSE)** of the linear regression model over a range of T (Booth 2003)
 - ❖ Bozdech et al. (2003) found that most expression profiles exhibited an overall expression period of 0.75-1.5 cycles per 48 h
 - ❖ We varied T from 1 to 100 and fit the multiple linear regression model (shown in the previous slide) based on 472 Oligonucleotides that have known stages
 - Table S2 and Figure 2 of Bozdech et al.'s paper show the 472 periodically expressed oligonucleotides and their stages

3.1. Identification of Periodically Expressed Oligonucleotides—*Results*

Estimation of the periodicity T



- ❖ The sum of squared errors (SSE) is minimized at 50 hours

3.1. Identification of Periodically Expressed Oligonucleotides – *Ranking Criterion*

- For $T=50$, we ranked genes by their corresponding R-squared values
- The statistical significance of each R-squared value was determined using the *F-statistic*

$$F = (J - p)R^2 / (p - 1)(1 - R^2)$$

J : no. of time points (46); p : no. of parameters (3)

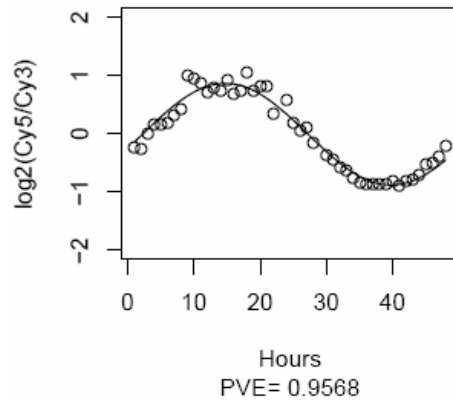
- We applied a permutation-based FDR (False Discovery Rate) procedure to evaluate the significance of the F-statistic (Taylor et al. 2004)
 - ❖ **We permuted the times (columns) in the data**
 - ❖ **Statistically significant oligonucleotide were chosen by comparing the F-statistic with a given cutpoint at the estimated FDR**

3.1. Identification of Periodically Expressed Oligonucleotides – *Results*

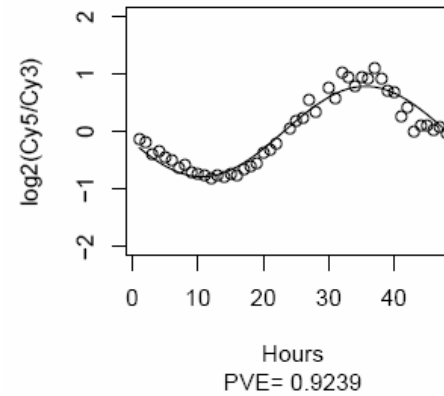
- Using a cutoff value of $PVE \geq 0.7$, which corresponds to F -statistic=50.2, we selected 2949 oligonucleotides (out of the total 5080 oligonucleotides)
- After 10,000 permutations of the time points, the estimated FDR is $3 * 10^{-5}$, suggesting the randomized datasets do not demonstrate periodicity

3.1. Examples of Expression Profile of 4 Periodically Expressed Genes – *Results*

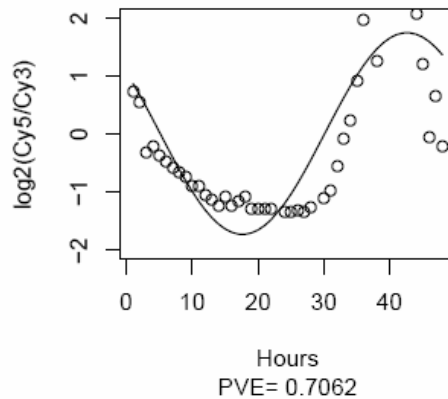
PFL2355w



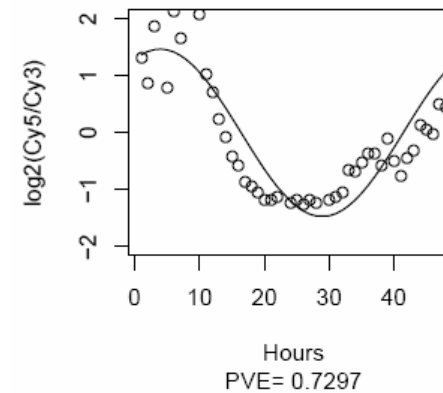
PFA0285c



PFC0185w



PF11_0231



3.2. Classification of the Periodically Expressed Oligonucleotides - *Background*

- Previous studies on classifying periodically expressed genes into cell-cycle stages were mainly focused on clustering methods (Spellman et al. 1998; Whitfield et al. 2002, Lu et al. 2004)
- Limitations of these methods include:
 - (1) *hard to use prior stage information;*
 - (2) *Can not assign a confidence level for the classification*
- We applied a supervised classification method.

3.2. Classification of the Periodically Expressed Oligonucleotides– *Data*

➤ **Training Data** (based on Bozdech et al. 2003)

Stages	Gene Functions	No. of Oligonucleotides
Ring/Early Trophozoite	Transcription machinery (23)	214
	Cytoplasmic Translation machinery (159)	
	Glycolytic Pathway (14)	
	Ribonucleotide Synthesis (18)	
Trophozoite/ Early Schizont	Deoxynucleotide Synthesis (7)	93
	DNA Replication Machinery (40)	
	TCA Cycle (11)	
	Proteasome (35)	
Schizont	Plastid Genome (27)	131
	Merozoite Invasion (87)	
	Actin Myosin Motility (17)	
Early Ring	Early Ring Transcripts (34)	34

➤ **Testing Data:** All periodically expressed oligonucleotides which have not been used in the “training” step

3.2. Classification of the Periodically Expressed Oligonucleotides– *Approach*

- Here we have a **multi-class classification problem**, with the 4 classes corresponding to the four stages
- Two general approaches for a multi-class classification problem:
 - ❖ **One vs. One** – pair-wise comparisons leading to $k*(k-1)/2$ possible comparisons. For our data, $k=4$, so there are **6 possible classifiers**.
 - ❖ **One vs. All** – requires k comparisons
- Since our data is very unbalanced (“Early Ring” stage consisting of only 7.2% of all data), we applied the **one vs. one approach**
- **Support Vector Machine** (SVM) was applied to train the 6 classifiers.
- **10 fold cross-validation** was used on training data to evaluate the performance of the classifiers
- Assignment of a stage-unknown periodically expressed oligonucleotide to a stage is based on a **cutoff probability** (confidence level)

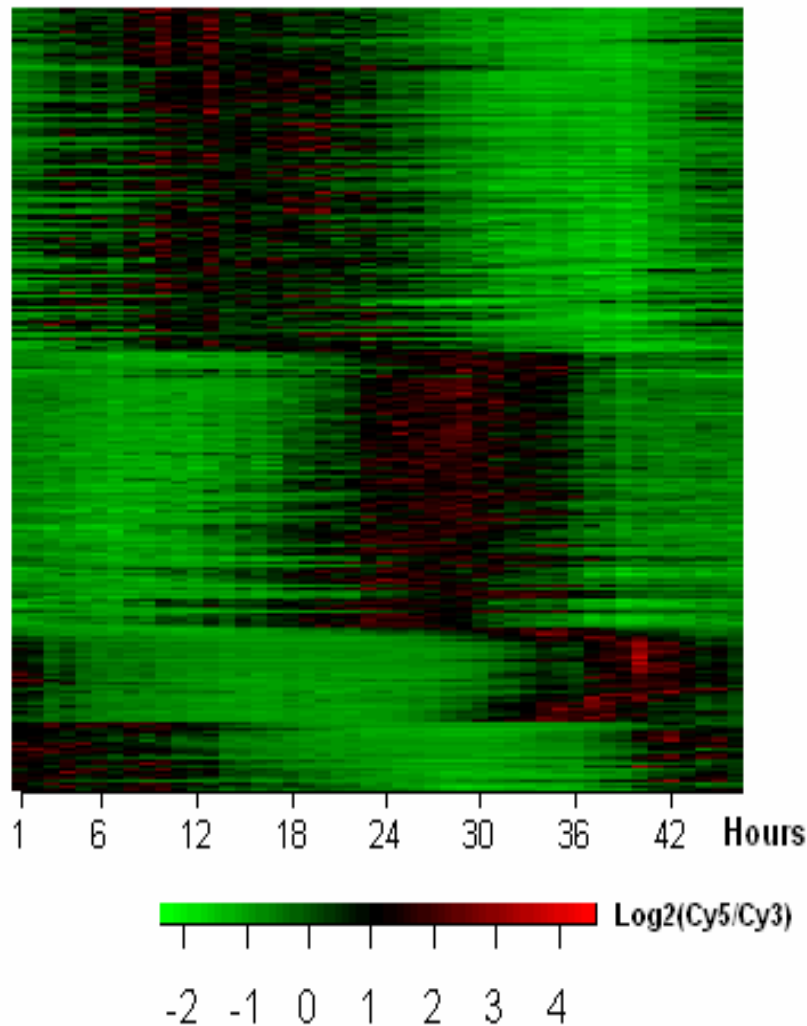
3.2. Classification of the Periodically Expressed Oligonucleotides— *Stage assignment based on a confidence level*

- Computation of confidence level of assigning an oligonucleotide x to a specific stage y ($y \in \{1,2,3,4\}$) involves three steps:
 - ❖ *Obtain 6 decision values from the 6 pair-wise SVM classifiers*
 - ❖ *Transform these values to 6 pair-wise class probabilities using a logistic function (Platt, 2000) and then to 4 stage-specific probabilities using a coupling algorithm (Hastie and Tibshirani, 1998)*
 - ❖ *And finally, we obtain the maximum probability over the 4 stages and assign the oligonucleotide x to stage y if this maximum probability is 0.8 or greater.*

3.2 Classification of the Periodically Expressed Oligonucleotides to Get Stage-Specific Periodically Expressed Genes – *Results*

- 472 oligonucleotides (351 genes) were used as “training” set and 2545 oligonucleotides (1918 genes) as “test” set
- The overall 10 fold cross-validation error was 3.4%
- Given a confidence level of 80% (estimated probability =0.8), we assigned
 - ❖ 718 genes to stage 1 (ring/early trophozoite)
 - ❖ 624 genes to stage 2 (trophozoite/early schizont)
 - ❖ 141 genes to stage 3 (schizont)
 - ❖ 167 genes to stage 4 (early ring)
 - ❖ 268 periodically expressed genes with estimated probability less than 0.8 were not assigned to any of the four stages

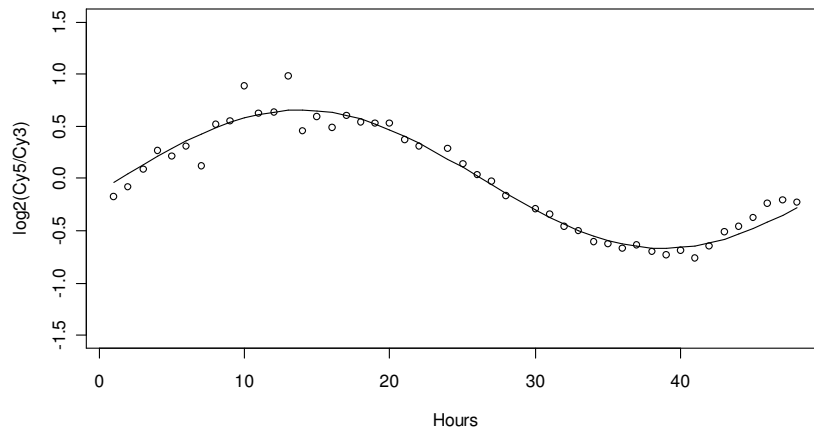
3.2. Heat Map of the Stage-Specific Periodically Expressed Genes in 4 IDC Stages – *Results*



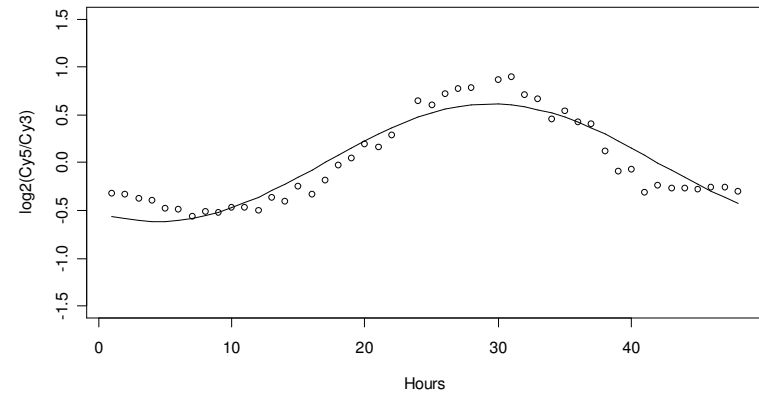
- ❖ The genes (included training and testing data) were ordered by stages (from top to bottom, stage 1-4)
- ❖ Within each stage, genes were sorted by the estimated probability in decreasing order (Genes in training data have probability 1)

3.2. Stage-Specific Meta-Gene Expression Profiles – *Results*

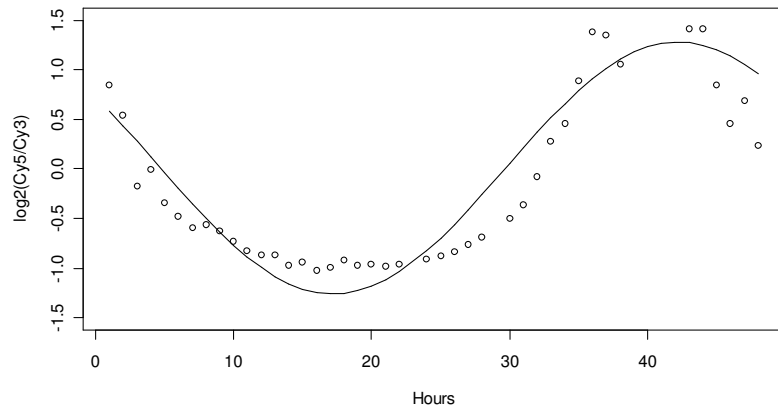
Average Gene Expression Profile of Ring/Early trophozoite Stage



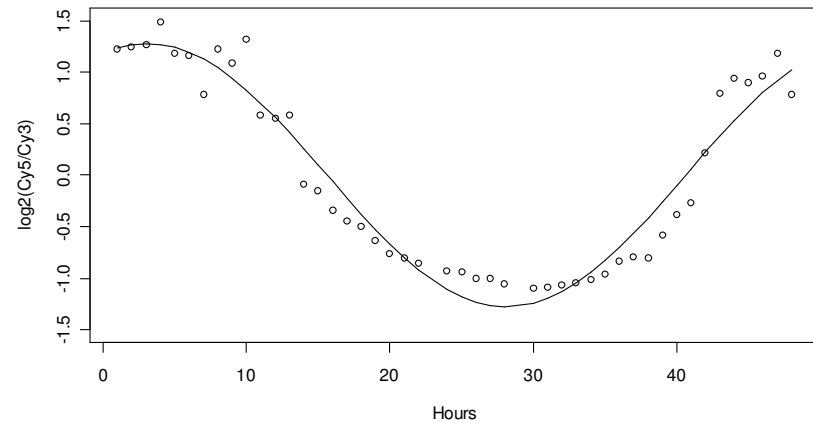
Average Gene Expression Profile of Trophozoite/Early Schizont Stage



Average Gene Expression Profile of Schizont Stage



Average Gene Expression Profile of Early Ring Stage



3.3. Chromosomal clustering of the Stage-Specific Periodically Expressed Genes-- *Approach*

- Previous studies included two approaches:
 - ❖ *Correlation-based clustering (Cohen et al.2000; and Bosdech et al. 2003)*
 - ❖ *Stage-specific clustering (Florens et al. 2002)*
- We used the second approach
 - ❖ *We mapped the periodically expressed gene assigned to any of the four stages (confidence level $\geq 80\%$) to the 14 chromosomes*
 - ❖ *A chromosomal cluster is defined as two or more adjacent genes whose expression patterns were matched to the same stage*

3.3 Number of Stage-Specific Clusters in each Chromosome with Different Cluster Size – *Results*

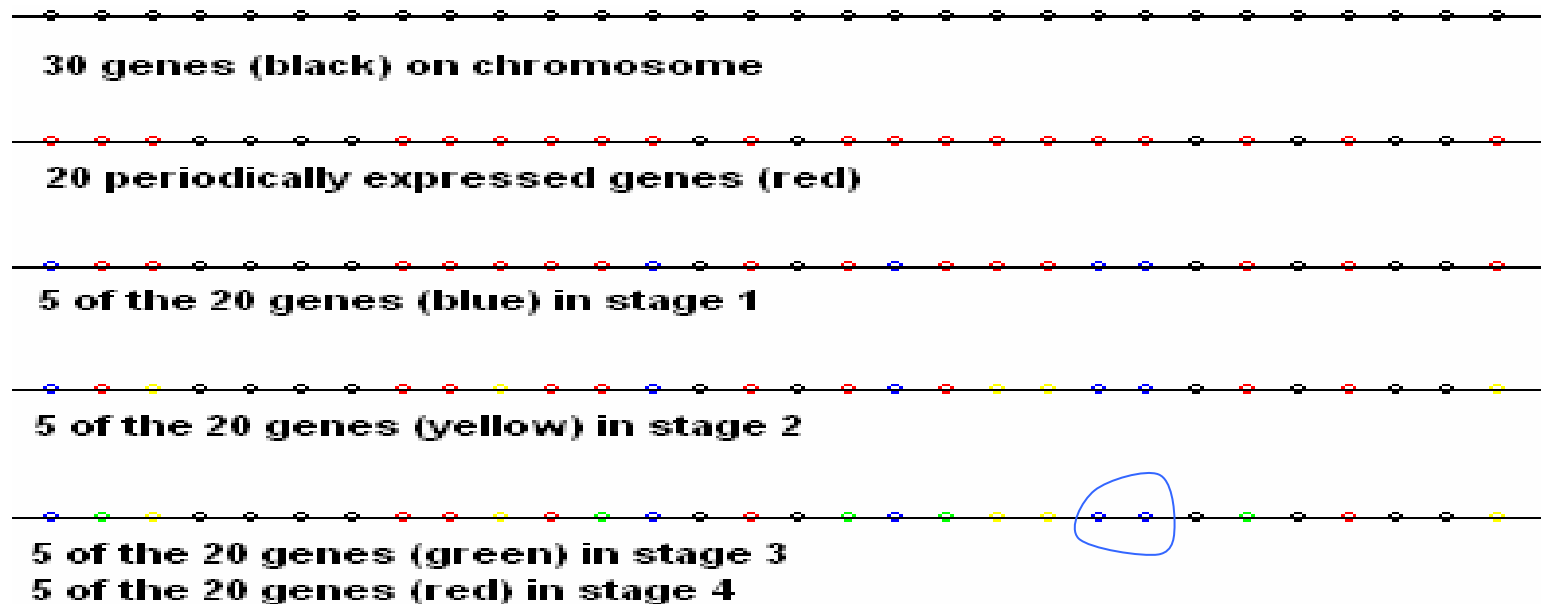
Chromosome	# of adjacent loci predicted to belong to the same stage			
	2	3	4	5
Chr-1	4	1		
Chr-2	15	2	1	1
Chr-3	14	2	2	1
Chr-4	9	3	2	1
Chr-5	19	1	1	
Chr-6	13			
Chr-7	15	5	1	
Chr-8	14	1		
Chr-9	16	2		
Chr-10	12	5	1	1
Chr-11	13	7	1	
Chr-12	18	3	1	
Chr-13	33	15	2	
Chr-14	43	8	3	
total	238	55	15	4
Total number of clusters: 312				

Stage	1	2	3	4
# of clusters	0	2	1	1

- ❖ The cluster size in the data ranged from 2 to 5
- ❖ Approximately 76% clusters are small clusters (cluster size is 2)
- ❖ 34 of 51 **large clusters (larger than 2)** identified by Bozdech et al. are also found in the 74 large clusters in our study, suggesting that genes in a stage-specific cluster have high correlation
- ❖ Approximately 33% clusters are identified in Chromosomes 13 and 14 – the 2 longest chromosomes

3.3. Chromosomal clustering of the Stage-Specific Periodically Expressed Genes--- *Illustration using randomly generated data*

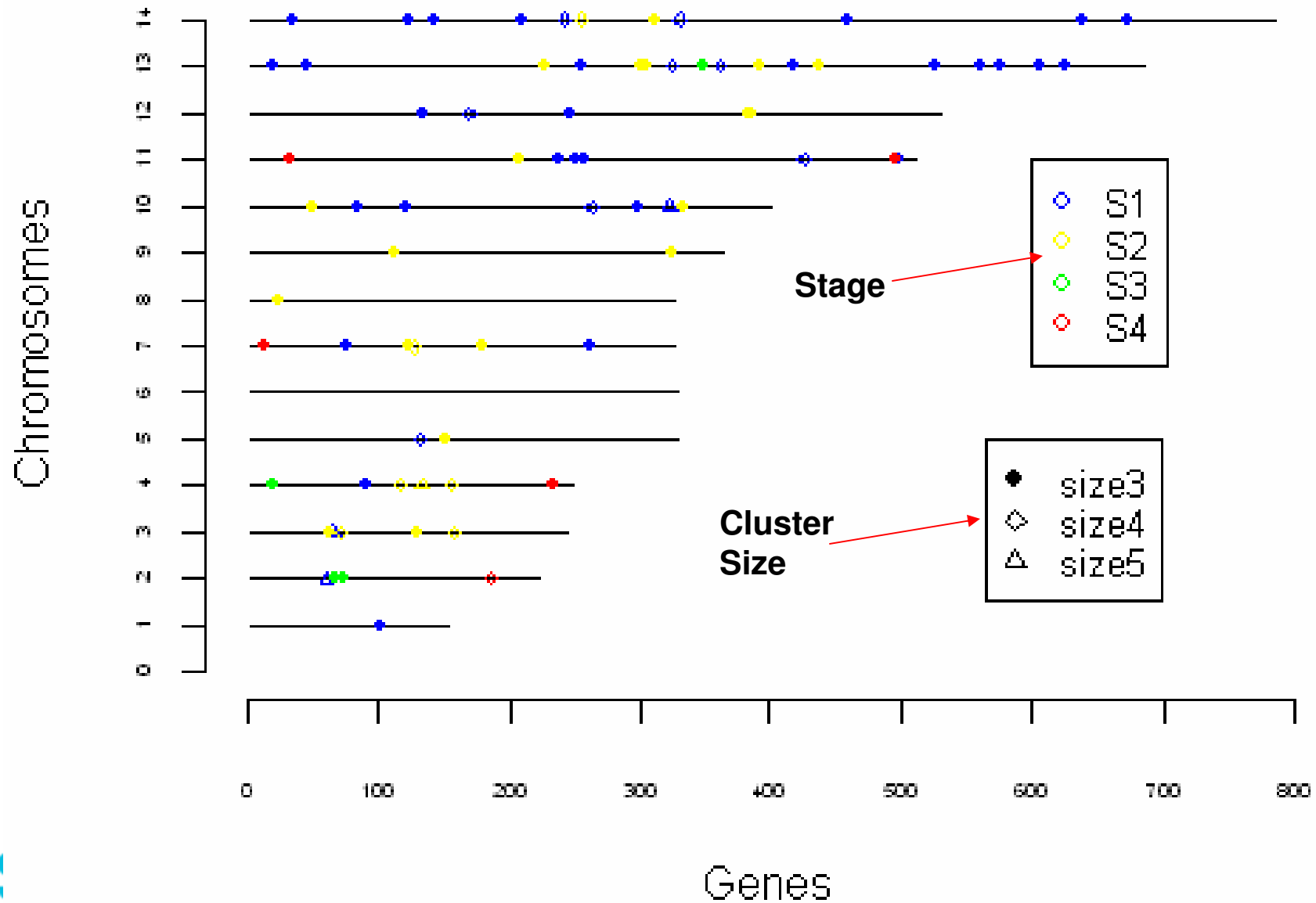
- Suppose the first permuted data on chromosome **c** looks like ($B=1$)



- ❖ *For a given cluster size 2, we found one cluster on chromosome **c** and stage 1. Therefore, $n_{c12} = 1$*
- A preliminary analysis, using permutations and a given cluster size, suggested that the occurrence of the stage-specific clusters is quite small for randomly generated data

3.3. Whole Chromosome View of 74 Large

Stage-Specific Clusters Distributed on 14 Chromosomes – *Results*



3.4. Functional Analysis of 74 Large Cluster Stage-Specific Chromosomal Clusters

- According to Bozdech et al., 2003, only genes in two of the 51 larger clusters were shown to have functional relationship (within cluster)
 - ❖ **the SERA gene cluster and ribosomal protein gene cluster**
- For the 74 large clusters we found, 11 clusters (including the above two) contain at least two loci whose annotation clearly indicates that the genes are functionally related.

3.4. Functional Analysis of 74 Large Cluster-- Three Stage-Specific Chromosomal Clusters

Chromosome	Stage	Locus	Description
10	1	PF10_0121	Hypoxanthine phosphoribosyltransferase
		PF10_0122	phosphoglucomutase
		PF10_0123	GMP synthetase
13	1	MAL13P1.322	RNA processing
		MAL13P1.323	Unknown
		PF13_0340	RNA processing
13	1	PF13_0177	Nucleic acid binding, ATP binding
		PF13_0178	Nucleic acid binding
		PF13_0179	Nucleotide binding, ATP binding
		PF13_0180	Chaperone

4. Conclusions

- Applying a multiple linear regression sinusoidal model, we identified 2949 periodically expressed oligonucleotides
- We used a supervised classification method to assign these oligonucleotides into 4 IDC stages with confidence level of 80% or more
- We detected 312 chromosomal clusters based on stage-specific periodically expressed genes distributed on the 14 chromosomes of the Plasmodium Falciparum genome
 - ❖ *Our findings revealed that the expression of periodically regulated genes is coordinated locally on chromosomes where small clusters of genes within same stage are regulated jointly*
- Some of the chromosomal clusters we identified contain genes that are functionally related

Key References

- **Bozdech, et al. The Transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PloS Biology*, 1, 1-16, 2003.**
- **Cohen, B.A., Mitra, R.D., Hughes, J.D., & Church, G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genetics*, 26, 183-186, 2000.**
- **Spellman, P.T., et al. Comprehensive identification of cell-cycle-regulated genes of the Yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9, 3723-3297, 1998.**
- **Booth, J.G., et al. *Clustering periodically expressed genes using microarray data: a statistical analysis of the yeast cell cycle data*. University of Florida, Statistics Department Technical Report. 2003.**
- **Hastie, T. & Tibshirani, R. Classification by pairwise coupling. *The Annals of Statistics*, 26, 451–471, 1998.**