Spatial Correlation of Expression in *P. falciparum*

J.B. Christian[1], C. Shaw[3], J. Noyola-Martinez[1], M.C. Gustin[2], D.W. Scott[1], R. Guerra[1,*]
[1]Department of Statistics, [2]Department of Biochemistry and Cell Biology, Rice University
[3]Department of Molecular and Human Genetics, Baylor College of Medicine

ABSTRACT
Malaria is responsible for half a billion infections and two million deaths each year. Understanding the biology of *P. falciparum* is critical if effective vaccines are to be developed to fight against this aggressive parasite. New information about the regulatory mechanisms of *P. falciparum* aids the elucidation of the fundamental metabolic and transcriptional pathways which we must understand to design better treatments and prevention. Of particular importance is the intraerythrocytic development cycle (IDC), the part of its life spent in the blood stream of unwitting mammals, which is responsible for the physical symptoms experienced by infected individuals. The goal of this investigation is to examine spatially dependent co-regulation of gene expression over the 48-hour IDC. Correlation between gene expression and gene location over a few genes demonstrates evidence of co-regulated genes or operons, while correlation over many genes may demonstrate evidence for some other transcriptional regulation mechanism. We develop a visualization and statistical testing methodology to examine expression-location correlation which we apply to a time-course microarray study of the IDC transcriptome. Contrary to the current paucity of evidence, our findings show evidence for spatial correlation. The biological implications of detected blocks of moderate but consistent spatial correlation provide novel insights into the transcriptome of *P. falciparum*.

Keywords
Co-regulation, spatial correlation, DNA sequence data, microarray data, integration and analysis, visualization, permutation tests.

1. INTRODUCTION
Understanding the regulatory mechanisms in *P. falciparum* helps identify new targets for both preventing or stopping malaria infections. Examining regulation of transcription is a key to achieving these goals, and there are many interesting transcriptional phenomena in *Plasmodium*. Protozoa such as *Plasmodium* are capable of regulating gene expression by altering the structure of its chromosomes. For example, expression of the var cell surface protein of *Plasmodium* is regulated by a silencing mechanism [5]. In other eukaryotes, gene silencing and related epigenetic phenomena are typically mediated by covalent modification of histones that can spread along chromosomes, altering the accessibility of genes to the transcription apparatus [7]. Whether this type of regulation extends beyond the var genes to other genetic loci remains to be determined. This investigation explores the basic properties of location dependent transcriptional regulation by searching for both small chromosomal areas with highly correlated gene expressions, as well as searching for larger chromosomal regions with correlated gene expressions. Bozdech et al [2] used a heuristic to perform a limited search for spatial correlation and found a few places among the 14 linear chromosomes of *P. falciparum* where there was evidence of spatial correlation. Other approaches based on a simple Pearson correlation have been proposed [1]. Without a formal measure of statistical significance, however, it will be difficult to apply the approach on a wide-scale basis. We consider an analogous examination of pairwise correlations along each chromosome, with the addition of a permutation test to assess the significance of the result. We also consider a more formal covariogram approach.

Analytical Objective: The overall objective is to develop a statistical framework to examine spatial correlation between gene expression and location along chromosomal regions. To this end, we develop methods to analyze (1) pairwise correlation between adjacent genes without regard for the distance between them, (2) pairwise correlation between adjacent genes with distance restriction between them, (3) correlation through a formal covariogram function. An added benefit of the methods is their use in detecting possible errors in annotation as may occur when one gene is accidentally annotated as multiple separate genes.

This space intentionally left blank.

2. METHODS
2.1 Data Pre-Processing
To perform this analysis, it was necessary to create a data matrix which combined information from the gene expression with the gene locations. The normalized quality-control microarray data produced by Bozdech et al [2] was combined with the *P. falciparum* annotated nucleotide sequence [7] to create a joint dataset. This dataset was created by matching the unique gene identifiers found at http://plasmodb.org from the provided gene expression data with the annotations in the sequence data, creating a data matrix for 4457 unique genes. The start of a gene was defined as the end of the open reading frame closest to the 5' end of its strand. That is, an open reading frame on [100,200] would start at 100 if it is located on the Watson strand and at 200 if it is located on the Crick strand. For this investigation, the Watson strand is the reference strand, with all chromosomal locations listed for that strand.

2.2 Correlation Analysis

In this article we use Pearson correlation as a measure of distance between two expression profiles. Other measures of distance are possible, for example Euclidean distance.

2.2.1 Adjacent Pairwise Correlation

To analyze the pairwise correlations between adjacent genes we created a vector of pairwise Pearson correlations for each chromosome. This vector was constructed by finding the *n-1* (adjacent) pairwise correlations among the *n* genes on each chromosome. Figure 1 shows the correlations by chromosomal order using chromosome 2 as the example; the color-coding identifies which strand of the chromosome each gene is on. Note the highly correlated clump found previously by [2] containing 6 genes at about the 42nd location; there are 5 clustered correlations in a row, which we denote as a 5-clump. The pairwise correlations of interest to this investigation were 2, 3, 4 and 5-clumps or *k*-clumps in general. Further we limit these *k*-clumps to those with correlations all within a window, *d*, with at least one correlation above some threshold δ. Formally, we define a *k*-clump as $k+1$ consecutive genes $(i_1,\ldots i_{k+1})$ with consecutive pairwise correlations $[r_j \equiv r(i_j,i_{j+1}),\ j=1,\ldots,k]$ satisfying the constraints that $\max(r_j)-\min(r_l)<d$ with $j\neq l$, and $j,\ l$ in $\{1,\ldots,k\}$ and that $r_j \geq \delta$ for at least one j in $\{1,\ldots,k\}$. Throughout our investigation, we searched for clumps of genes with d = 0.1, 0.2 or 0.3, with at least one correlation above a δ of 0.5.

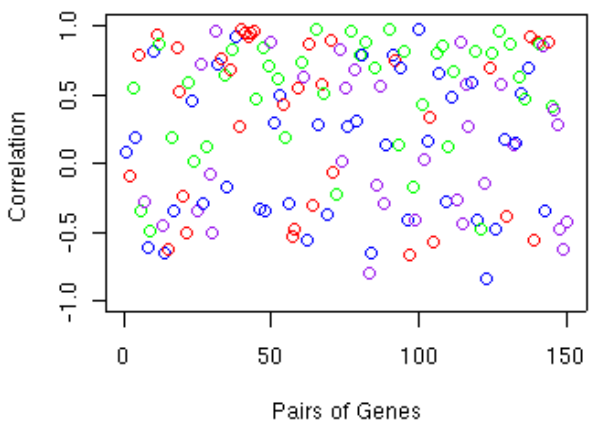

Figure 1: Pairwise correlations between consecutive genes on chromosome 2. Red dots indicate that both genes in the correlation are on the Crick strand, purple indicates both are on Watson, green indicates Crick-Watson and blue, Watson-Crick.

Many such clumps were found using this approach including the said 5-clump on chromosome 2 around gene 42. This raises the question of significance. How likely is it to find *k*-clumps at random if there is no spatial correlation? For example, would we expect to find a 5-clump on chromosome 2 if there were no co-regulation? If so, how many 5-clumps might we expect on a null-hypothesis chromosome? To test the significance of the clumps of genes with high correlation, we developed a permutation test (2.2.3), of which applications are quite natural in various genetic contexts [3, 10].

2.2.2 Distance-based Pairwise Correlation

The correlations based on adjacent neighbors do not take account of the distance between the two genes. As such, the power to detect location dependent correlation may be dampened if many of the gene neighbors are relatively far apart. To account for inter-gene distance we averaged all pairwise correlations among genes found in a moving window of length 50kb. Results for chromosome 6 are shown in Figure 2, and for chromosome 11 in Figure 3. These analyses allowed us to find small clumps with high spatial correlation. The process was repeated for windows of various lengths ranging from 10kb to 100kb.

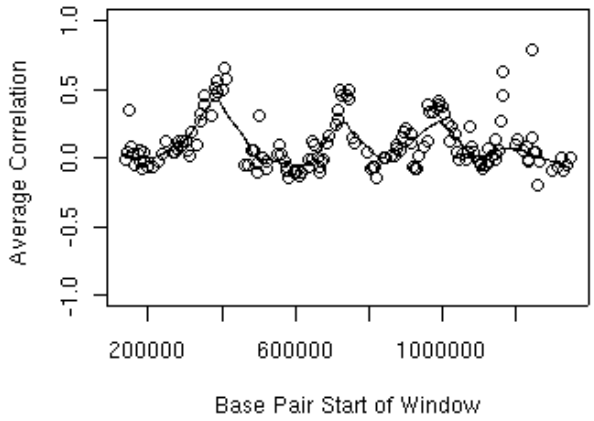

Figure 2: The distance-based pairwise correlation plot for chromosome 6, over a window size of 50kb, with a lowess smoothed line using a span of 0.15.

We performed an ANOVA test for linearity [9] on the distance-based correlation data to determine if there is evidence that "blocks" of non-zero correlation occur along the chromosomes. The ANOVA test for linearity requires repeated observations, and we used two methods of binning to form repeated measurements. For the first binning, each bin contained 10 correlations, and the mean of the locations for the genes in these 10 correlations was used as the location of the bin. In the second method, each chromosome was divided into 10 bins, and the location of each bin was determined by the mean of the locations of the genes in that bin.

2.2.3 Permutation Test Algorithm

To test the null hypothesis that there is no spatial correlation, the gene orderings were kept constant, while the gene expression profiles were permuted. Repeating this process B=1000 times generated a null distribution for finding *k* genes in a row with $\max(r_i)-\min(r_j)<d$ for $i,\ j$ in $\{1,\ldots,k\}$. We also allowed for a parameter δ that would restrict the pairwise correlations in a clump to be at least as great as δ. We performed this test for $k=2,\ldots,6$, for $d=.1,.2,.3$ and $\delta=0.5$.

For a given k, d, δ and B the permutation test was performed this way:

For i=1 to B

a. Permute gene expression profiles

b. Compute pairwise correlations

c. Count number of k-clumps

The permutation test can be applied to both the adjacent and distance-based pairwise correlation approaches. The choice defines step (b) in the above algorithm. The number of possible permutations is too large to perform an exhaustive count. Therefore, a random number (B) of permutations are performed.

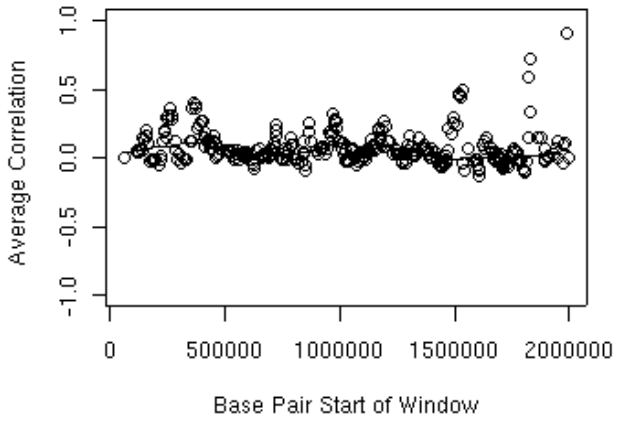

Figure 3: The distance-based pairwise correlation plot for chromosome 11, over a window size of 50kb, with a lowess line using a window size of 0.15

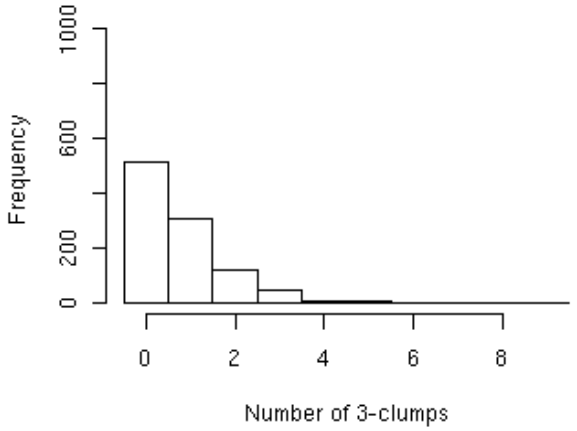

Figure 4: Reference distribution used in perumtation test. This is for chromosome 9, for k=3, d=.1, δ=.5, and 4 3-clumps were observed yielding an estimated p-value of 0.014.

The null reference distribution is used to calculate a p-value for the number of *k*-clumps observed. An example of a reference distribution is given in Figure 4. Shown for chromosome 9 is the histogram for the number of times (over B permutation) a 3-clump occurs under the null hypothesis of no spatial correlation. For example, about 125 of the B permutations has a 3-clump appearing two times. To compute the p-value, we have an observed number of *k*-clumps, k_{obs}, as well as null distribution, the histogram of the recorded counts of *k*-clumps from the permutation test. Let x_i, i=(1,…,B) denote the number of *k*-clumps found in permutation *i*. Then the estimated p-value is

$$\text{p-value}=\#\{\ x_i \geq k_{obs}\}/\text{B}.$$

2.2.4 Covariogram

A covariogram is a function that relates correlation as a function of distance [4]. In this case we want to know if gene expression correlation depends on the distance between the two genes. The first approach of adjacent correlation does not account for inter-gene distance, while the second distance-based approach depends on a moving window of contiguous sets of genes. Here a true covariogram function (γ) would give the correlation between genes *x* and *y* given that they are d_0 base-pairs apart:

$$\gamma\ (x,\ y;\ d_0) = \ r(x,\ y \mid \text{distance}(x,y)=d_0)$$

The assumption is that γ is homogeneous with respect to location; that is, we assume a constant correlation between genes x and y that are $d_0$base-pairs apart, no matter where the genes may be located on the chromosome, the only important factor being that they are d_0 bp apart. A consequence of this assumption is that it makes sense to pool all correlations across the entire chromosome (or entire genome for that matter) that are based on pairs of genes d_0 bp apart; by construction the moving window of the distance-based approach does not have such an assumption. A covariogram was created for each chromosome in the following way. First, correlations were calculated for all pairs of genes within 10kb of each other. Then the average correlation was calculated and plotted against the midpoint of each interval, as seen in Figure 5. Note that the assumption of homogeneity may be violated, which will require some modifications to the method.

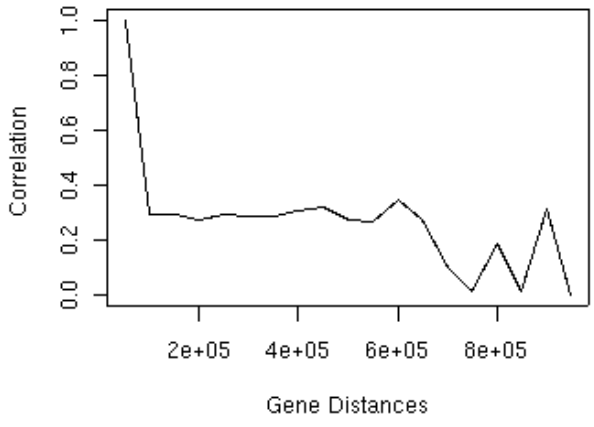

Figure 5: Covariogram for chromosome 2 using a 50kb window size

3. RESULTS

3.1 Correlation Analysis

A pairwise correlation plot for chromosome 2 is shown in figure 1. By signifying the strand orientation, we can identify features such as candidate operons or other co-regulated regions, as all genes in clumps of the same color have the same orientation. Results for chromosomes with significant numbers of *k*-clumps are shown in table 1. The most interesting feature in the correlation plots were the 6 highly correlated genes located at about position 42 in the first third of chromosome two, also noted by Bozdech et al [2]. Based on the annotation as putative cysteine protease, these appear to be tandem repeats of cysteine protease. Also of interest were significant clusters on chromosomes 4, 9, 13 and 14 that need to be investigated further. For most of the *k*-clumps, the annotation information is not currently available. Figure 6 gives an example from chromosome 13. We found six 3-clumps on chromosome 13, yielding a p-value of 0.05. Shown are the four profiles which correspond to two hypothetical proteins, synaptobrevin-like protein (putative) and elongation factor tu (putative).

There are also 54 pairs of genes which are possibly co-regulated via a shared promoter region. To identify these pairs, we used the criteria that each of these pairs consist of a Crick oriented gene followed by a Watson oriented gene, that the beginnings of the two open reading frames were less than or equal to 2 kb apart, and that the correlation between the two gene expressions was at least 0.7. This list of genes, their annotation information and their gene ontology information is provided by request. A smoother version of this correlation analysis is found by using the distance-based correlations for distances of 10 and 50kb, and averaging all pairwise correlations in that window. Figures 2 and 3 show these results. Note the significant reduction in noise by increasing the size of the window, which suggests "blocks" of correlated genes on some chromosomes. Finally, for all of the chromosomes, the distance-based correlation was binned in two ways for the ANOVA test of linearity. For each chromosome, both binning methods yielded p-values ≤ 0.01, with many p-values $\leq$ 1e-10.

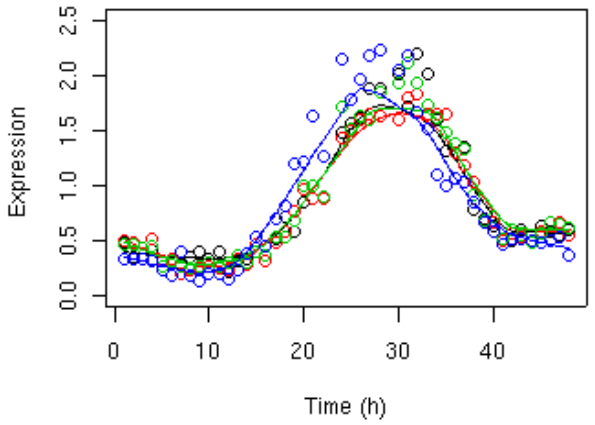

Figure 6: A 3-clump on chromosome 13, among the four genes shown, there are two unannotated genes (red, green), synaptobrevin-like protein (putative, black curve) and elongation factor tu (putative, blue curve).

3.2 Covariograms

A covariogram for chromosome 2 is shown in Figure 5. Note the small spatial correlation that is present in genes from 100 kb up to 600 kb apart. A partial explanation of this can be seen in the distance-based correlation plot shown in Figure 2, where there are three regions of nonzero correlations, from 25-150 kb, 275-375 kb and 450-575 kb. The first block is rifin rich with several erythrocyte membrane binding proteins and many hypothetical proteins. The second block appears to contain tandem repeats of cysteine protease, while the third block has mainly hypothetical proteins.

Covariograms for chromosomes 2, 4, 5, 9 and 10 also suggest correlation over large portions of their chromosomes. Examining the distance-based correlation plots with a large window size showed several of these blocks of interest.

4. CONCLUSIONS

Spatial correlation between gene expression profiles and chromosomal location may be defined in several ways. Considering adjacent pairwise correlations ignores inter-gene distance and thus may result in a loss of power to detect spatial correlation. Accounting for distance by restricting adjacent genes to be within a certain distance (bp) or through a formal covariogram function should increase the power. We have considered and compared the three approaches in the context of the *P. falciparum* time-course array study of Bozdech et al. [2]. Unlike previously reported findings we do find evidence of spatial correlation after accounting for inter-gene distance. Critical to the findings is a measure of statistical inference which we have implemented with a permutation approach.

The analysis of the pairwise correlations shows that there are some statistically significant clusters of genes which are of interest. More specifically, the number of clumps of a certain size (*k*) are more than expected under the null hypothesis of no spatial correlation. Indicating the orientation of genes by color-coding provides a useful visual examination of potentially interesting areas of spatial correlation along a chromosome. Of particular interest are Crick-Watson alignments, which potentially share a promoter region. Several such significant findings were observed, such as two genes from chromosome 6, glutaredoxin and translation initiation factor IF-2, both of which are involved in stress response [5]. Having annotation is crucial in assessing the biological significance of these findings. As such, we have (Shaw lab) generated a gene ontology (GO) database for *P. falciparum* that allows us to annotate our spatial correlation findings with function. Figure 6 shows the expression profiles of a 3-clump. This type of result can help assess which clumps are worth pursuing for further investigation. Although there are many annotations, many more of these potentially interesting pairs still lack annotation. Also interesting are clumps that have genes residing on the same DNA strand as they may provide clues to polycistronic regions, and several of these have also been detected.

The covariograms also provide interesting information on spatial correlation. Several chromosomes indicated long-range moderate but consistent correlation ($r \sim 0.3$). This observation led us to examine larger window widths in our distance-based correlation,

resulting in several relatively long "blocks" of spatial correlation. This block structure does not appear to be a random artifact as indicated by an ANOVA test for linearity on the moving window correlation data, with all chromosomes having significant p-values. This indicates that there may be some related function in these regions, or perhaps that there are silenced regions [5] along the chromosome. More work is needed to determine the function of the non-annotated proteins in these regions.

This investigation has shown that there are several types of spatial correlation present in the *P. falciparum* data. To the best of our knowledge there appear to be no other formal inferential methods for dealing with spatial correlation dealing with sequence and expression data. There appear to be two types of clumps, those of size 2-6 with highly correlated genes and those of larger blocks of up to 100 kbs with moderate correlation.

5. REFERENCES

[1] Aburatani, S. et al. Statistical Analysis of the Relationship between Gene Expression and Location. *Genome Informatics, 14,* (2003), 306-7.

[2] Bozdech, Z. et al. The Transcriptome of the Intraerythrocytic Development Cycle of *Plasmodium falciparum*. *PLoS Biology,1,* (18 Aug. 2003), 1-16.

[3] Churchill, G.A., and Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics, 138*, (1994) 963-971.

[4] Cressie, N.A.C. *Statistics for Spatial Data.* J. Wiley, New York, NY, 1993.

[5] Calderwood, M.S. et al. *Plasmodium falciparum var* Genes Are Regulated by Two Regions with Separate Promoters, One Upstream of the Coding Region and a Second within the Intron. *J. Biol. Chem., 278*, (Sep 2003), 36:34125-34132.

[6] Dever, T.E. Translation Initiation: Adept at Adapting, *Trends Biochem Sci, 10*, (Oct 1999), 398-403.

[7] Gardner, M.J., Hall, N., et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature, 419* (2002), 6906:498-511.

[8] Grewal, S.I.S and Moazed, D., Heterochromatin and Epigenetic Control of Gene Expression. *Science*, *301,* (Aug 2003) 5634:798-802 .

[9] Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. *Applied Linear Statistical Models*. The McGraw-Hill Companies, Inc., Boston, MA, 1996.

[10] Wan, Y., Cohen, J., and Guerra, R. A permutation test for the robust sib pair linkage method, *Annals of Human Genetics*, *61*, (1997), 79-87.

Table 1a. Results of permutation tests, chromosomes with p-values ≤ 0.1, and for a correlation threshold of δ=0.5, and for correlations within d=0.1.

Chr	k	Observed	p-value
2	2	14	<0.001
2	3	4	0.004
2	4	2	0.006
2	5	1	0.003
4	2	17	<0.001
4	3	2	0.069
7	2	12	0.073
8	2	13	0.004
9	2	14	0.029
9	3	4	0.014
9	4	1	0.076
10	2	19	<0.001
13	3	6	0.050
14	2	32	0.014
14	3	6	0.043

Table 2b. Results of permutation tests, chromosomes with p-values ≤ 0.1, and for a correlation threshold of δ=0.5, and for correlations within d=0.2.

Chr	k	Observed	p-value
2	2	24	<0.001
2	3	8	0.002
2	4	2	0.038
2	5	1	0.031
4	2	20	<0.001
4	3	4	0.052
6	2	14	0.073
7	2	17	0.100
8	2	16	0.054
9	2	19	0.069
9	3	7	0.023
9	4	2	0.097
9	5	1	0.056
10	2	28	0.001
10	3	7	0.038
11	2	25	0.072
11	3	7	0.067