BMC Genomics

Relationship between gene co-expression and probe localization on microarray slides

BMC Genomics 2003, 4:49

Yuval Kluger (yuval.kluger@med.nyu.edu)
Haiyuan Yu (haiyuan.yu@yale.edu)
Jiang Qian (jiang.qian@jhmi.edu)
Mark B Gerstein (mark.gerstein@yale.edu)

ISSN 1471-2164

Article type Research article

Submission date 13 Nov 2003

Acceptance date 10 Dec 2003

Publication date 10 Dec 2003

Article URL http://www.biomedcentral.com/1471-2164/4/49

Relationship between gene co-expression and probe localization on microarray slides

Yuval Kluger*[1,2], Haiyuan Yu[1]*, Jiang Qian*[1,3], Mark Gerstein[1]‡

[1]Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, CT 06520, USA

[2] Current address: Cell Biology, New York University School of Medicine, New York, NY 10016, USA

[3] Current address: Wilmer Institute, Johns Hopkins School of Medicine, Baltimore, MD, 21287, USA

* These authors contributed equally to this work.

‡ To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861;

Email: Mark.Gerstein@yale.edu

ABSTRACT

BACKGROUND: Microarray technology allows simultaneous measurement of thousands of genes in a single experiment. This is a potentially useful tool for evaluating co-expression of genes and extraction of useful functional and chromosomal structural information about genes.

RESULTS: In this work we studied the association between the co-expression of genes, their location on the chromosome and their location on the microarray slides by analyzing a number of eukaryotic expression datasets, derived from the *S. cerevisiae*, *C. elegans*, and *D. melanogaster*. We find that in several different yeast microarray experiments the distribution of the number of gene pairs with correlated expression profiles as a function of chromosomal spacing is peaked at short separations and has two superimposed periodicities. The longer periodicity has a spacing of 22 genes (~42 Kb), and the shorter periodicity is 2 genes (~4Kb).

CONCLUSION: The relative positioning of DNA probes on microarray slides and source plates introduces subtle but significant correlations between pairs of genes. Careful consideration of this spatial artifact is important for analysis of microarray expression data. It is particularly relevant to recent microarray analyses that suggest that co-expressed genes cluster along chromosomes or are spaced by multiples of a fixed number of genes along the chromosome.

BACKGROUND

Since the discovery of the DNA double-helix structure, chromosomal configuration has been the focus of intense research. Although it is well known that in the higher eukaryotes, chromosomal structure plays a role in gene expression, the precise mechanism remains unclear [1-3].

Microarray technology has enabled us to simultaneously measure expression levels of tens of thousands of genes. Many prior gene expression analyses have focused on studying gene co-expression and inferring functional relationships from expression relationships [4, 5]. Recently researchers have been looking at the association between chromosomal gene organization and gene expression [6-8]. These analyses suggest that chromosomal spatial organization affects gene expression in a very systematic way.

There are numerous methods of performing gene expression experiments, including cDNA microarrays, oligonucleotide arrays and Affymetrix microarray chips. These different technologies could potentially affect the gene co-expression results. Given that in many microarray chips DNA spots are printed in an order related to the gene order on the chromosomes, the systematic relationship between gene co-expression and chromosomal location raises the suspicion that part of the gene pair correlations are associated with inherent chip artifacts rather than true biological co-expression. In this work we further investigated the association between gene co-expression and chromosomal location with attention to the impact of the location of the genes on the microarray slides, looking at datasets obtained with different microarray technologies.

RESULTS

After calculating correlation coefficients between gene-expression profiles for all of the pairs of genes on each chromosome, we selected out gene pairs with a correlation coefficient of 0.7 or

greater. We constructed a histogram of the number of gene pairs as a function of their relative chromosomal distance. The blue curve in Figure 1 represents the distribution of highly correlated gene pairs as a function of the relative pair chromosomal distance using the Spellman et al alpha-factor arrested cell cycle dataset. It shows that adjacent gene pairs tend to be co-expressed. This phenomenon was observed in all the datasets we examined (see supplementary information at http://bioinfo.mbb.yale.edu/~kluger/artifact/GB_CHROMOSOME_DISTANCE.ppt). In some of these datasets [8-11] in which genes were spotted on the arrays according to their chromosomal order, it appears that genes that are spaced by multiples of fixed number of ORFs along the chromosome are more likely to co-express, as seen by the long and short-range periodicities of the blue curve of Figure 1. These chromosomal co-expression regularities can also be revealed by inspecting the correlation map (or its Fourier transform) similar to the ones shown in [6, 12] (see supplementary information at http://bioinfo.mbb.yale.edu/~kluger/artifact/correlation_maps.ppt).

The red curve in Figure 1 shows the distribution of the subset of all pairs separated by short chip distance (not only those that are highly correlated) as a function of the pair chromosomal distance. The blue and red distributions share common characteristics, i.e., enrichment in the number of gene pairs at short chromosomal distance as well as at specific chromosomal distances determined by the long and short-range periodicities. This commonality indicates that it is more likely that a pair of genes will co-express if its relative distance on the chip is short.

In some experiments (including those done with Affymetrix chips), the order of genes on the chips is not simply related to their chromosomal order [8]. In this case, periodicities such as those seen in Figure 1 are not observed. However, inspection of the correlation map and its

Fourier transform reveals unexpected regularities (see supplementary information at http://bioinfo.mbb.yale.edu/~kluger/artifact/correlation_maps.ppt).

To see whether the commonalities and regularities mentioned above are spurious we examined the relationship between gene pair co-expression and chip co-localization. Figure 2 demonstrates the relationship between the average gene pair correlation and their relative location on the array. Ideally we would expect the value of the average correlation to be spatially independent and around zero. As can be seen in Figure 2 the average correlation peaks at short chip distances and decays as the distance grows.

Evaluating the average correlation of all gene pairs as a function of their relative chip distance is a useful and simple tool to assess the extent and magnitude of spatial artifacts. This can be used with any type of chip, regardless of whether the genes were arranged on the array according to chromosomal order or not. For instance, in microrray analysis of muscle-expressed genes in *C. elegans* [7] genes were not printed on the array according to their chromosomal order and they were spotted on 8 by 4 different blocks. In this case we still observe the decay of the average co-expression as a function of chip distance. Furthermore, the curve is modulated with seven nodes separated by distances equal to the block spacing as can be seen in Figure 3. This enrichment in co-expression of gene pairs is a systematic artifact linked to the fact that adjacent genes on the DNA source plates are printed into the same position on different blocks of the microarray slide.

DISCUSSION AND CONCLUSIONS

In this work we demonstrate that adjacent gene pairs on a chromosome tend to be co-expressed. This is consistent with similar findings based on analysis of mRNA microarray experiments by Cohen et al [6], Roy et al [7] and Spellman et al [8], as well as with proteomic data [13]. In many microarray chips DNA spots were printed in an order related to the gene order on the chromosomes. This systematic relationship raises the suspicion that part of the above-mentioned short-range enrichments and short and long-range periodicities are associated with inherent chip artifacts.

Figure 2 clearly shows that there is an artifact in the data - the closer the gene pairs are on the microarray chips, the higher the average correlation coefficient is. We call this trend a local chip artifact. Naively one would expect that local chip artifacts in microarray experiments be canceled considering the ratio of the sample and reference cells. This is certainly not the case if the noise at each microarray spot is not a multiplicative one. Thus, the biased enrichment of co-expression at short chip distances substantially contributes to a magnification of co-expression at short or periodic chromosomal distances, if genes are organized on the chip in an order related to their order on the chromosomes.

We have demonstrated that the relative chip and source plate distances between genes have a noticeable effect on their measured co-expression. Systematic artifacts such as print tip effects (on spotted microarrays), and random artifacts such as scratches, blotches and cross hybridization lead to enhancement of multi-experimental correlations between pairs of genes located in close proximity on the chip or source plate. The multi-array correlation of any pair of genes is increased if one of these artifacts occurs even on a single array. Experimental or

computational corrections of these artifacts are necessary for characterizing the relationships between gene co-expression and gene function or chromosomal co-localization.

Local normalization procedures have been proposed in preprocessing of microarray data [14]. However, we note that applying local normalization corrections prior to the evaluation of multi-experimental correlations tends to decorrelate genes separated by large chip distances, and leave adjacent genes correlated. Therefore, we propose that the multi-experimental correlation for any pair of genes will exclude the experiments where one (or both) of the genes is located in an array surrounding that has unusual features such as scratches or blotches of very high intensity signal. Finally, this artifact is present in a wide variety of microarray experiments (see supplementary information at http://bioinfo.mbb.yale.edu/~kluger/artifact/all.ppt).

MATERIALS AND METHODS

We analyzed various microarray datasets from different organisms and different microarray platforms (cDNA, Affymetrix). These include datasets from studies of the yeast cell cycle, diauxic shift and gene knockouts [9-11, 15, 16] , muscle-expressed genes in *C. elegans* [7], *Drosophila* [8] and *E. coli* [17, 18]. The location of the probes on the slides of the cDNA array experiments are available on the web. The probes for the yeast cell cycle Affymetrix experiment (Cho et al) were generously supplied by the authors.

In order to study whether the chromosomal spatial organization affects gene expression, we calculated the multi-experimental correlation coefficients between gene-expression profiles for every pair of genes on a chromosome. This correlation is equivalent to the scalar product of the standardized gene pair expression profiles. We then selected pairs with correlation coefficients greater than 0.7, and constructed a distribution (histogram) of these pairs as a function of their relative chromosomal distance. The distance between each pair of genes was measured in terms of open reading frames (ORFs). Each bin of the histogram represents the percentage of highly correlated pairs that have a given pair-distance. Subsequently, we normalized this histogram of observations by dividing it by a corresponding random histogram of expected distances.

The random histogram is generated by following procedure. First, we obtain the pairs of genes that have high correlation coefficient (>0.7) on each chromosome. Then, the gene pairs were randomly placed on the chromosome and the distances between them were measured in units of ORFs. The procedure was repeated 100 times and an average histogram was obtained as reference histogram.

1. Orphanides G. and Reinberg, D: RNA polymerase II elongation through chromatin, Nature. *407:* 471-5, 2000.
2. Manuelidis LA: view of interphase chromosomes, Science. *250:* 1533-40, 1990.
3. Cremer T and Cremer C: Chromosome territories, nuclear architecture and gene regulation in mammalian cells, Nature Reviews Genetics. *2:* 292 -301, 2001.
4. Brown PO and Botstein D: Exploring the new world of the genome with DNA microarrays, Nat Genet. *21:* 33-7, 1999.
5. Eisen MB, Spellman PT, Brown PO, and Botstein D: Cluster analysis and display of genome-wide expression patterns, Proc Natl Acad Sci U S A. *95:* 14863-8, 1998.
6. Cohen BA, Mitra RD, Hughes JD, and Church GM: A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression, Nat Genet. *26:* 183-6., 2000.

7. Roy PJ, Stuart JM, Lund J, and Kim SK: Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans, Nature. *418:* 975-9., 2002.
8. Spellman PT and Rubin GM: Evidence for large domains of similarly expressed genes in the Drosophila genome, Journal of Biology. *1:* 1, 2002.
9. DeRisi JL, Iyer VR, and Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale, Science. *278:* 680-6., 1997.
10. Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, and Futcher B: Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth, Nature. *406:* 90-4., 2000.
11. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, and Davis RW: A genome-wide transcriptional analysis of the mitotic cell cycle, Mol Cell. *2:* 65-73., 1998.
12. Qian J, Kluger Y, Yu H, and Gerstein M: Identification and correction of spurious spatial correlations in microarray data, Biotechniques. *35:* 42-4, 46, 48, 2003.
13. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, and Carucci DJ: A proteomic view of the Plasmodium falciparum life cycle, Nature. *419:* 520-6., 2002.
14. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. *30:* 15, 2002.
15. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol Biol Cell. *9:* 3273-97., 1998.
16. Hughes TR., Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, and Friend, SH: Functional discovery via a compendium of expression profiles, Cell. *102:* 109-26, 2000.
17. Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, and Yanofsky C: DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli, Proc Natl Acad Sci U S A. *97:* 12170-5, 2000.
18. Courcelle J, Khodursky A, Peter B, Brown PO, and Hanawalt PC: Comparative gene expression profiles following UV exposure in wild-type and SOS deficient Escherichia coli, Genetics. *158:* 41-64, 2001.

Captions

Figure 1: Pair Correlation as a Function of Chromosomal Separation

Co-expressed and co-localized gene pair distributions as a function of the pair chromosomal distance for the alpha factor arrested cell cycle dataset [15]. We defined co-expression by selecting gene pairs that have a correlation coefficient > 0.7 across the time course experiments. The pair distance was measured in terms of the number of ORFs separating its individual genes (blue curve). The distribution of the chip co-localized pairs (red curve) was constructed using pairs constrained to chip distances smaller than 12.5% of the maximal possible pair distance. Both distributions are enriched at a small chromosomal distance and share short and long range periodicities of 2 and 22 ORFs. These periodicities are pure artifacts, whereas the enrichment is partially biological. The chip design is such that nearest neighbor genes on the chromosome are printed in different blocks on the microarray chips whereas next nearest neighbor genes are printed in the same vicinity on the chip. This is correlated with the short-range periodicity as reflected in the staggered pattern shown in Fig 1a.

Legend for this figure:

blue curve: pairs with correlation > 0.7

red curve: pairs with slide spacing < 12.5% maximal spacing

Figure 2: Average Pair Correlation as Function of Chip Distance

Average co-expression as a function of the distance of gene pairs on the chip for the alpha factor arrested cell cycle dataset [15]. For pairs used to generate the red curve, the individual genes in each pair are from the same chromosome. For the green curve members of each pair are from different chromosomes.

Legend for this figure:

red curve: intra

green curve: inter

Figure 3: Average Rank Co-expression as a Function of the Separation of Gene Pairs on the Chip for Worm Muscle-expression Experiment [7].

Co-expression is defined by the score $\overline{r}_i \cdot \overline{r}_j - |\overline{r}_i - \overline{r}_j|$, where $\overline{r}_i$ is the average rank percentile of expression of gene i across several replicated array experiments. For pairs used to generate the blue curve, the individual genes in each pair are from the same chromosome. For the red curve members of each pair are from different chromosomes. The green curve is a pair distribution similar to the blue curve distribution but with randomized pair-wise chip distances.

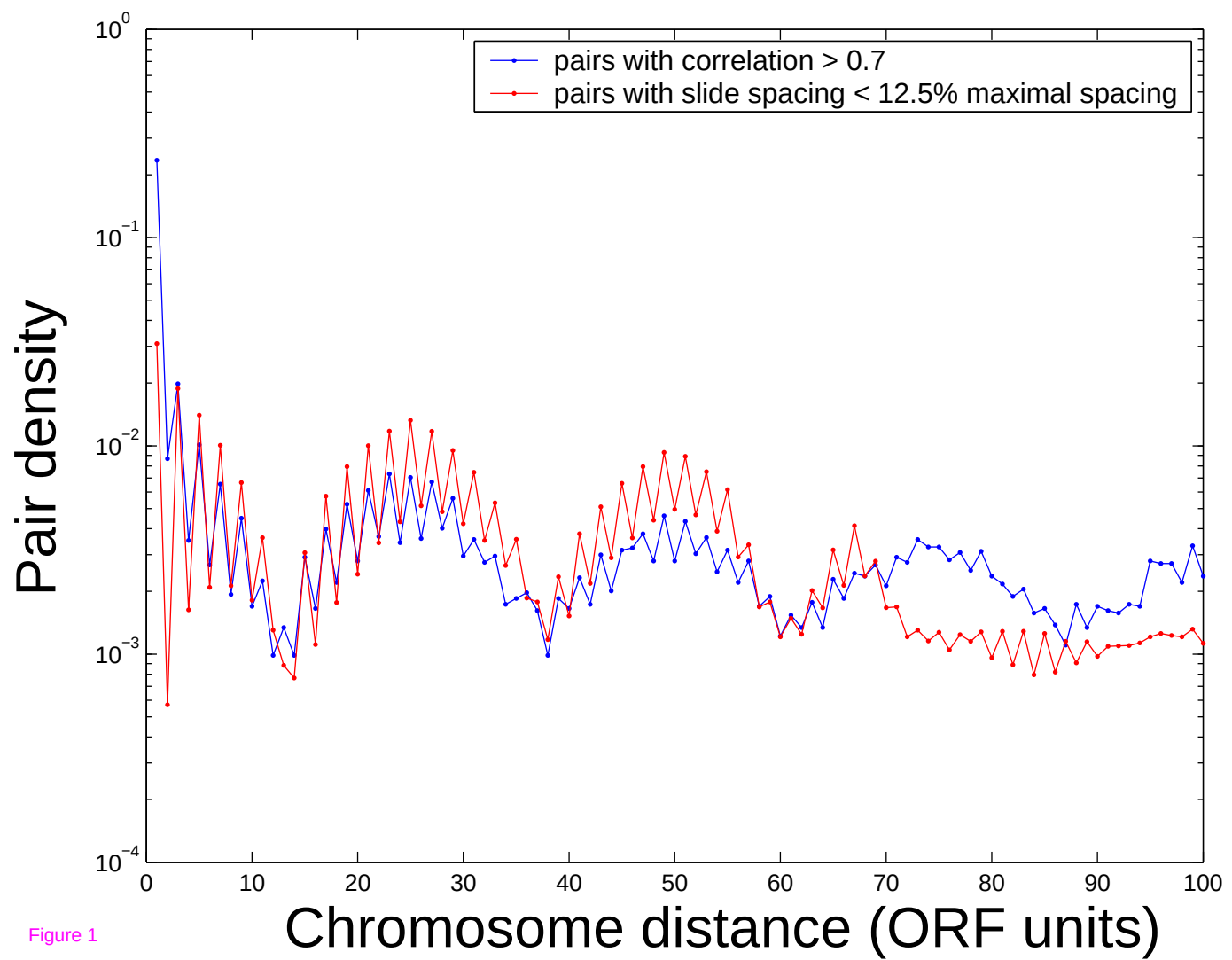
pairs with correlation > 0.7
pairs with slide spacing < 12.5% maximal spacing
Pair density
Chromosome distance (ORF units)
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
10
20
30
40
50
60
70
80
90
100

Figure 1

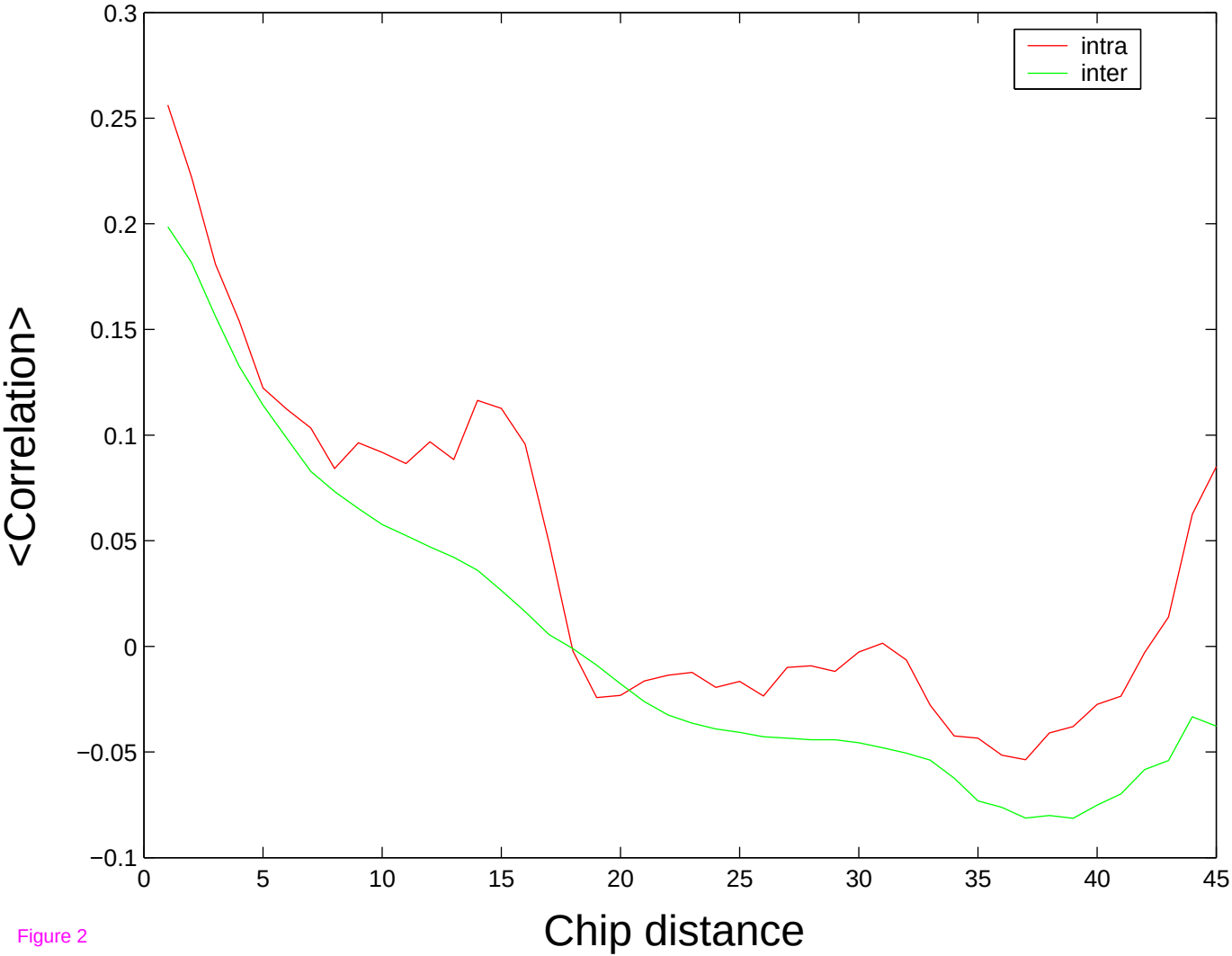
intra
inter
<Correlation>
Chip distance
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
0
5
10
15
20
25
30
35
40
45

Figure 2

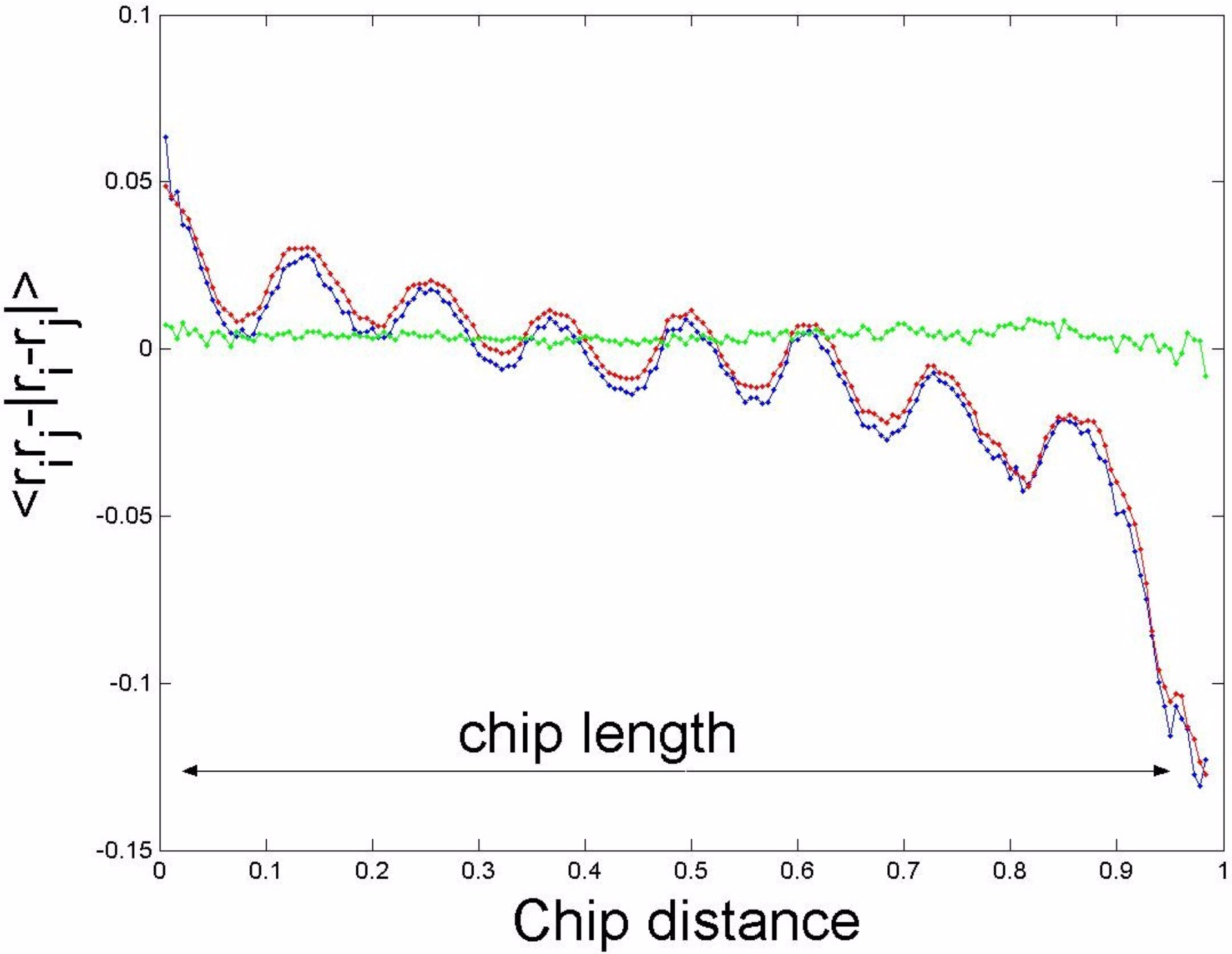
0.1
0.05
0
-0.05
-0.1
-0.15
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
<r_i r_j-|r_i-r_j|>
chip length
Chip distance

Figure 3