BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2903–2909
doi:10.1093/bioinformatics/btm482

Gene expression

Extracting three-way gene interactions from microarray data

Jiexin Zhang, Yuan Ji and Li Zhang*

Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Unit 237, Houston, TX 77030-4009, USA

Received on April 28, 2007; revised on September 16, 2007; accepted on September 23, 2007

Advance Access publication October 5, 2007

Associate Editor: David Rocke

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023

ABSTRACT

Motivation: It is an important and difficult task to extract gene network information from high-throughput genomic data. A common approach is to cluster genes using pairwise correlation as a distance metric. However, pairwise correlation is clearly too simplistic to describe the complex relationships among real genes since co-expression relationships are often restricted to a specific set of biological conditions/processes. In this study, we described a three-way gene interaction model that captures the dynamic nature of co-expression relationship between a gene pair through the introduction of a controller gene.

Results: We surveyed 0.4 billion possible three-way interactions among 1000 genes in a microarray dataset containing 678 human cancer samples. To test the reproducibility and statistical significance of our results, we randomly split the samples into a training set and a testing set. We found that the gene triplets with the strongest interactions (i.e. with the smallest P-values from appropriate statistical tests) in the training set also had the strongest interactions in the testing set. A distinctive pattern of three-way interaction emerged from these gene triplets: depending on the third gene being expressed or not, the remaining two genes can be either co-expressed or mutually exclusive (i.e. expression of either one of them would repress the other). Such three-way interactions can exist without apparent pairwise correlations. The identified three-way interactions may constitute candidates for further experimentation using techniques such as RNA interference, so that novel gene network or pathways could be identified.

Contact: lzhangli@mdanderson.org

Supplementary information: http://odin.mdacc.tmc.edu/~zhangli/ThreeWay

1 INTRODUCTION

High-throughput genomic data are a rich resource for elucidating how genes are interconnected (Alm and Arkin, 2003; Lander, 1999; Quackenbush, 2003; Zhang, 2002). Many computational tools have been developed to assess gene interactions using microarray gene expression profiling data (Alm and Arkin, 2003). A common approach is to cluster genes using pairwise correlation as a distance metric (Eisen *et al.*, 1998; Ji *et al.*, 2005; Tavazoie *et al.*, 1999). However, pairwise correlation is clearly too simplistic to describe the complex relationships among real genes since it is rare to find gene pairs that are constitutively co-expressed. Typically, co-expression relationships are often restricted to a specific set of biological conditions and/or processes (Rao and Arkin, 2001). For example, some co-expression relationships were found to exist only in cancer but not in normal tissue (Choi *et al.*, 2005). A gene pair may be co-expressed only in a specific organ, at a specific developmental stage, after a drug treatment or in a particular disease state. To deal with these complications, new methods have been developed to identify co-expressed gene groups in subsets of biological conditions (Dettling *et al.*, 2005; Shedden and Taylor, 2004) or modules (Getz *et al.*, 2000; Ihmels *et al.*, 2002; Jornsten and Yu 2003; Segal *et al.*, 2003; Wu *et al.*, 2004). Using such methods, it is possible to build a dynamic network of co-expressed groups form a microarray dataset. Each link (edge) in the network represents a significant pairwise correlation between two genes evaluated from expression profiles of a subset of the microarray samples.

In this study, we took an alternative approach to assess co-expression gene network. Instead of assuming that certain biological conditions/modules may affect the co-expression relationship between a gene pair, we assume that there is a third gene (named the 'controller gene' hereinafter) associated with the biological conditions/modules that can affect the co-expression relationship. With this approach, we are looking for three-way gene interactions instead of two-way interactions. The three-way model captures the dynamic nature of co-expression relationship through the introduction of the third gene. In reality, more than one gene may affect the co-expression of a gene pair. However, models that involve more than three genes are much less tractable because the number of combinations of four or more genes is much larger than that of three genes. Thus, the three-way model represents an appealing compromise between realism and tractability. Moreover, because it is believed that the entire human gene network contains mostly nodes with sparse connections and only a small number of hubs that have a large number of connections (Luscombe *et al.*, 2002; Rzhetsky and Gomez, 2001; Thieffry *et al.*, 1998; Wagner, 2002), it is reasonable to expect that the three-way interaction model should be a good approximation to many cases in the real network.

We applied a simple statistical method to screen gene triplets and selected those in which the correlation of two genes can be

*To whom correspondence should be addressed.

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

significantly associated with the expression level of a controller gene. Using a microarray dataset containing 678 samples collected from human cancer tissues and cell lines, we examined about 0.4 billion gene triplets based on 1000 genes. By splitting the data randomly into a training set and a testing set of equal sample sizes, we demonstrated that the most statistically significant triplets identified using the training set are validated in the independent testing set.

2 METHODS

2.1 Microarray data source and pre-processing

The raw microarray data (CEL files) were downloaded from Gene Expression Omnibus database at ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE2109. The data were generated by International Genomic Consortium (IGC) in its Expression Project for Oncology using Affymetrix human genome array HG-U133 plus 2.0. We used data in IGC batches 1, 2, 3, 5 and 6, but excluded batch 4, because we found that the probe signal distributions of the samples in batch 4 showed a marked difference from the samples in other batches. Most of the samples were extracted from cancer tissues and a few from cell lines. A detailed description of sample information can be found in Supplementary Table S1.

We used the quantile normalization method (Bolstad *et al.*, 2003) to normalize probe level data (PM probes only) and used the PDNN model (Zhang *et al.*, 2003) to extract the gene expression values. The gene expression data were quantile normalized again to reduce systematic and technical biases between samples. Then, we randomly split the samples into a training set and a testing set, each with 339 samples. The training set and the testing set were processed separately in subsequent analyses.

2.2 Identification of genes with bimodal expression profiles

A model-based clustering algorithm, MCLUST (Fraley and Raftery, 2002), was used to determine whether the distribution of log-expression values of a gene is a single normal or a mixture of two normal distributions. We used the Bayesian information criteria to select between the two models. When the mixture of two normal distributions was found to be a better fit, we then applied the MCLUST to obtain a threshold value T that splits the samples into two subgroups. To ensure adequate sample size for our subsequent evaluations, only subgroups with at least 60 samples were kept for subsequent analysis.

2.3 Evaluating interactions in a gene triplet

Consider a gene triplet A, B and C. Without loss of generality, suppose gene C is the controller gene. Based on the threshold obtained from MCLUST, we divided the 339 samples in the training set into a low expression group of n_1 samples and a high expression group of n_2 samples according to the expression levels of gene C. Then, we computed r_1, the Pearson correlation coefficient of the log-expression values between gene A and gene B from the n_1 samples, and r_2 from the n_2 samples. Because correlation values can be unstable when the variances of the expression levels of genes A and B are small, we discarded the triplets in which either variance of gene A or that of gene B was <0.1 in either n_1 or n_2 samples. We used Fisher's z-transformation (Fisher and Belle, 1993) to transform the correlation coefficients to a test statistic z:

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \quad \text{NE} \tag{1}$$

where

$$z_1 = 0.5 \times \ln\left[\frac{1+r_1}{1-r_1}\right] \quad \text{NE} \tag{2}$$

$$z_2 = 0.5 \times \ln\left[\frac{1+r_1}{1-r_1}\right] \quad \text{NE} \tag{3}$$

3 RESULTS

3.1 Identification of significant triplets

First, we applied a gene filtering process to our microarray dataset to reduce computational cost of our survey and ensure quality of the expression measurements. Using the training set, we filtered out genes that had small variation across samples, poor annotation or redundant probe design. Of the ~57 000 probe sets on the array, it is common that multiple probe sets represent a gene. For any two probe sets representing the same gene, we removed the one with less variance. Then, we removed probe sets that correspond to no entries in RefSeq database (Pruitt *et al.*, 2003) or have no gene symbols. Subsequently, we selected the top 1000 probe sets that have the largest variances. Among these probe sets, the minimum variance of log-expression values was 0.37, which was much higher than the median variance of all probe sets on the array, which was 0.033. Consequently, the remaining genes should have large changes in expression levels that cannot be explained by merely random noise. The process resulted in a training set composed of 1000 probe sets (genes) and 339 samples. The testing set was assembled using the same probe sets and the remaining samples. The training set and testing set can be found in Supplementary Tables S2 and S3, respectively.

Next, we examined expression profiles of the 1000 genes and found 796 of them have bimodal expression distributions in the training set using a model-based clustering algorithm. For each of the 796 genes, the samples were dichotomized into a low expression group and a high expressed group (see Methods section for details). To assess the statistical significance of each triplet interaction, we applied Fisher's z-transformation (Fisher and Belle, 1993) to convert Pearson correlation coefficients into z values (see Methods section for details). We evaluated all 0.40 billion triplets that can potentially be formed by these 1000 genes with the 796 control genes.

The main results of our survey are summarized in Figure 1. The central peak of the distribution of the 0.40 billion z values resembles a normal distribution (Fig. 1a and b). In principle, the distribution of z values asymptotically approaches the standard normal distribution (mean 0 and variance 1), if the correlations are evaluated from independent random data. However, we found the variance of z values to be 1.88 instead of 1.0. The inflated variance was supposed to be caused by the

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023

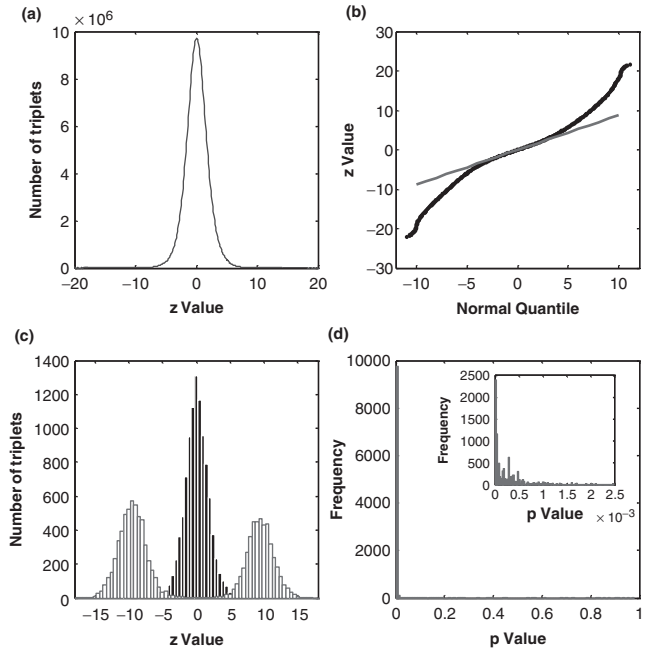

Fig. 1. Survey statistics. (**a**) Distribution of z values for 0.4 billion triplets calculated from the training set; (**b**) Quantile-quantile plot: z values versus normal quantile, which was generated from $N(0, 1.8)$; (**c**) The distribution of z_t values calculated from testing set for the top 10 000 triplets with the highest $|z|$ values in training set (red) and the distribution of z values from 10 000 randomly assembled triplets (black); (**d**) The distribution of P-values calculated from testing set for the most significant 10 000 triplets evaluated in the training set. The inset shows zoomed figure for P-value < 0.0025.

fact that the z values obtained from our survey are not independent from each other, i.e. the z values cannot be regarded as resulted from null data. Recent research suggests that such inflated variance of test statistics is a common phenomenon due to effects of pairwise correlations between test statistics. This widened distribution of the null test statistics has a substantial impact in large scale testing (Efron, 2007). The heavy tails (Fig. 1b) suggested that the significant interactions were prevalent.

Due to the enormous size of the survey, false positives are very difficult to quantify accurately. For example, with 0.4 billion triplets, we had 0.4 billion tests and P-values. To control the false positive rate at a small level, we need to compute the P-values with extremely small precision (e.g. as small as 10^{-12}), which is practically impossible. Our approach to the problem is to use the testing set for validation. We computed the z values using the testing set (z_t) for the 10 000 triplets that have the largest $|z|$ values ($|z| > 11.2$) evaluated from the training set. Figure 1c shows the distribution of these z_t values, which are mostly concentrated on the tails of the normal distribution (histogram in red). We also randomly selected 10 000 triplets in the testing set and computed their z values (z_{random}, Fig. 1c, histogram in black). The distribution of z_t shares little overlap with the distribution of z_{random} values, implying that these triplets are indeed statistically significant. To obtain a P-value for each z_t value, we treated the distribution of these 10 000 z_{random} values as an empirical null for P-value estimation. As shown in Figure 1d, most of the resulting P-values are very small (97.8% of them are < 0.01).

3.2 The 'L'-shaped relationships in top significant triplets

We observed an interesting reoccurring pattern from the top triplets with the highest $|z|$ values. As shown in Figure 2, depending on the expression of the controller gene, there are two distinct types of interactions for the remaining two genes: they are linearly correlated in one type while form an 'L'-shaped pattern in the other. The 'L'-shape reflects a mutually exclusive relationship within the gene pair; only one, but not both, of the genes could be highly expressed in a sample.

In Figures 3 and 4, we presented two examples in detail. The first example involves genes CFTR (cystic fibrosis transmembrane conductance regulator, ATP-binding cassette), MYB (v-myb myeloblastosis viral oncogene homolog) and USH1C (Usher syndrome 1C). Figure 3a contains the log expression values of MYB and USH1C in the training set. There is no apparent relationship between the expression values of these two genes (correlation coefficient is 0.09). The distribution of the log-expression values of CFTR appear to be bimodal (Fig. 3b), which is composed of a narrow but high peak at left side and a broad low peak on the right side. By separating samples into two subgroups according to the bimodality of CFTR, the relationship between MYB and USH1C emerges. When the expression level of CFTR is high (>0.98), the correlation between MYB and USH1C is positive ($r = 0.7$) (Fig. 3d); when the expression level of CFTR is low ($\leq$0.98), not only the correlation becomes negative ($r = -0.434$), but also emerges an L-shaped relationship (Fig. 3c). Previous studies indicate that both MYB and CFTR are regulated by NF-kappa B (Brouillard *et al.*, 2001; Suhasini and Pilz, 1999). However, it remains an intriguing problem to find out what causes the mutually exclusive relationship between USH1C and MYB.

In the second example (Fig. 4), the genes involved are Kruppel-like factor 5 (KLF5), cadherin 6 type 2 (CDH6) and UDP glucuronosyltransferase 1 family, polypeptide A (UGT1A). When KLF5 is expressed low, CDH6 and UGT1A show an 'L'-shaped relationship ($r = -0.313$); when KLF5 is expressed high, CDH6 and UGT1A are co-expressed ($r = 0.893$).

We investigated if the 'L'-shaped relationships shown in Figures 2–4 were caused by expressions of tissue specific genes. To explore this issue, we used Figure 5 to show the tissue compositions of the 'L'-shaped relationships seen in Figures 3 and 4. In Figure 5a, most of the samples in the vertical arm came from kidney, while most of the samples in the horizontal arm came from breast. These observations imply that when the control gene (CFTR) is expressed low, USH1C is kidney specific and MYB is breast specific. In Figure 5b, colon samples dominated the vertical arm, but no tissue type appeared to dominate the horizontal arm. More examples are shown in

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023

Fig. 2. Top 96 triplets. These triplets have P-value < 0.0003 and $|z| > 14.83$. Horizontally, every pair of scatter plots show log-expression values of gene A and gene B in two groups of samples that were partitioned according to bimodality of gene C. Data shown in red represent samples in which gene C is expressed at high levels; other samples are shown in black. Expression values of the gene C are not shown. '*' labels the L-shaped expression profile between gene A and gene B. Annotations of the triplets can be found in Supplementary Material S4.

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023

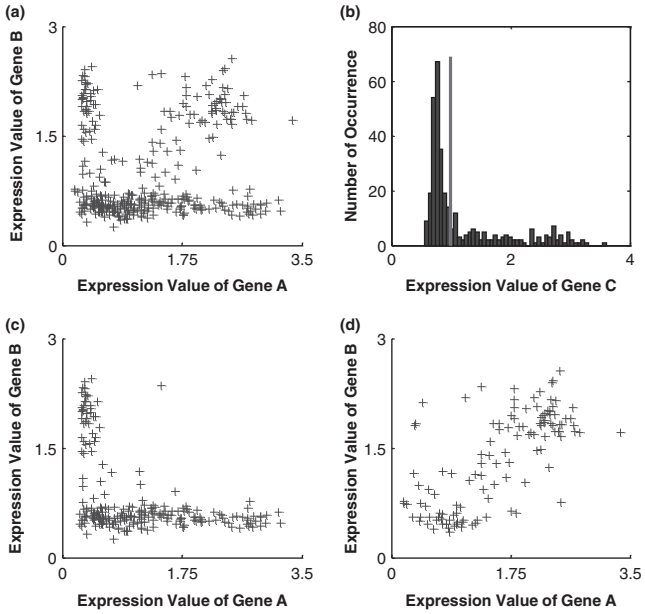

Fig. 3. An example of a significant triplet. The three genes are CFTR (gene C), MYB (gene A) and USH1C (gene B). The $|z|$ value of this triplet is 11.6. (**a**) Log-transformed expression value of MYB and USH1C in training set; (**b**) The histogram of log-transformed expression value of CFTR with the red bar indicating separation of the bimodal distribution; (**c**) Log-transformed expression value of MYB and USH1C when $\ln(\text{CFTR}) \leq 0.98$; (**d**) Log-transformed expression value of MYB and USH1C when $\ln(\text{CFTR}) > 0.98$.

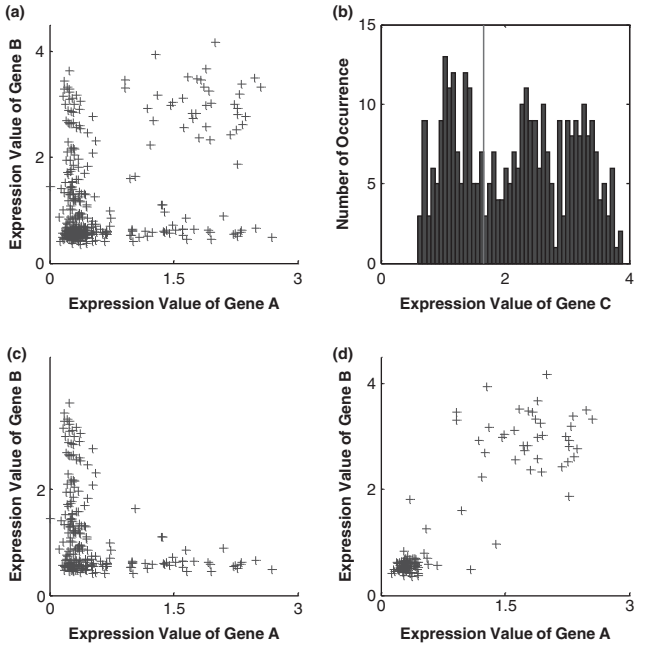

Fig. 4. An example of another significant triplet. The three genes are: KLF5 (gene C), CDH6 (gene A) and UGT1A (gene B). The $|z|$ value of this triplet is 15.46. (**a**) Log-transformed expression value of CDH6 and UGT1A; (**b**) The histogram of log-transformed expression value of KLF5, red bar indicating separation of the bimodal distribution; (**c**) Log-transformed expression value of CDH6 and UGT1A when $\ln(\text{KLF5}) \leq 1.66$; (**d**) Log-transformed expression value of CDH6 and UGT1A when $\ln(\text{KLF5}) > 1.66$.

Supplementary Material (Fig. S5). These examples suggest that in some of the cases the 'L'-shaped relationships involve tissue specific genes while in others the 'L'-shaped relationships are not related to tissue types. It should be noted, however, such tissue specificity is conditional on the control gene. For example, besides expressed in kidney, USH1C is also highly expressed in spinal cord (Su *et al.*, 2002; URL: http://symatlas.gnf/SymAtlas). Thus, the triplet interactions identified in our study cannot be explained simply by tissue specific genes.

It is important to note that many of the three-way interactions revealed by these triplets could not have been predicted by merely pairwise correlation. We calculated pairwise correlation for the 9780 triplets that have large $|z|$ values and small *P*-values. The results showed that 52.8% (5164), 38.37% (3753) and 21.84% (2136) of these triplets have pairwise correlations all <0.6, 0.5 and 0.4, respectively. Therefore, it seems that our method is particularly useful in identifying potential gene triplets where there is no obvious correlation for any gene pair, but more complex interactions exist.

4 DISCUSSION

Our survey of three-way gene interactions finds significant statistical associations among expression profiles of three genes. A significant gene triplet (A, B and C) is supposed to exhibit the following traits: (1) Gene C's expression profile displays bimodality, which enables us to partition the samples (arrays) into a high expression group and a low expression group; (2) Correlation of expression between genes A and B in the high expression group is significantly different from that in the low expression group.

It will require further experimentation to identify the underlying molecular interactions that drive the observed statistical associations found in the significant triplets. The statistical associations may also be caused by technical factors, such as cross-hybridization, RNA degradation and biases induced in normalization procedures. The technical factors may play a dominant role when the signal-to-noise ratio in the profiling data is low, which is why we have selected only the highly variable genes in our survey. Our survey was designed with the following molecular mechanisms in mind: a transcription factor X, which regulates both gene A and B, is conditional on the expression level of gene C. However, the X is unknown, and its relationship to C is not determined by our method. Gene C may encode X. Alternatively, gene C may regulate X's activity through some intermediate agents. There are also other molecular mechanisms that can explain changes in co-expression relationship between a gene pair. For example, Tomlins *et al.* (2005) found that there is a recurrent chromosomal aberration in prostate cancer tissues that led to fusion of two genes (TMPRSS2 and ETS transcription factor). In normal tissues, the expression of these two genes

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023

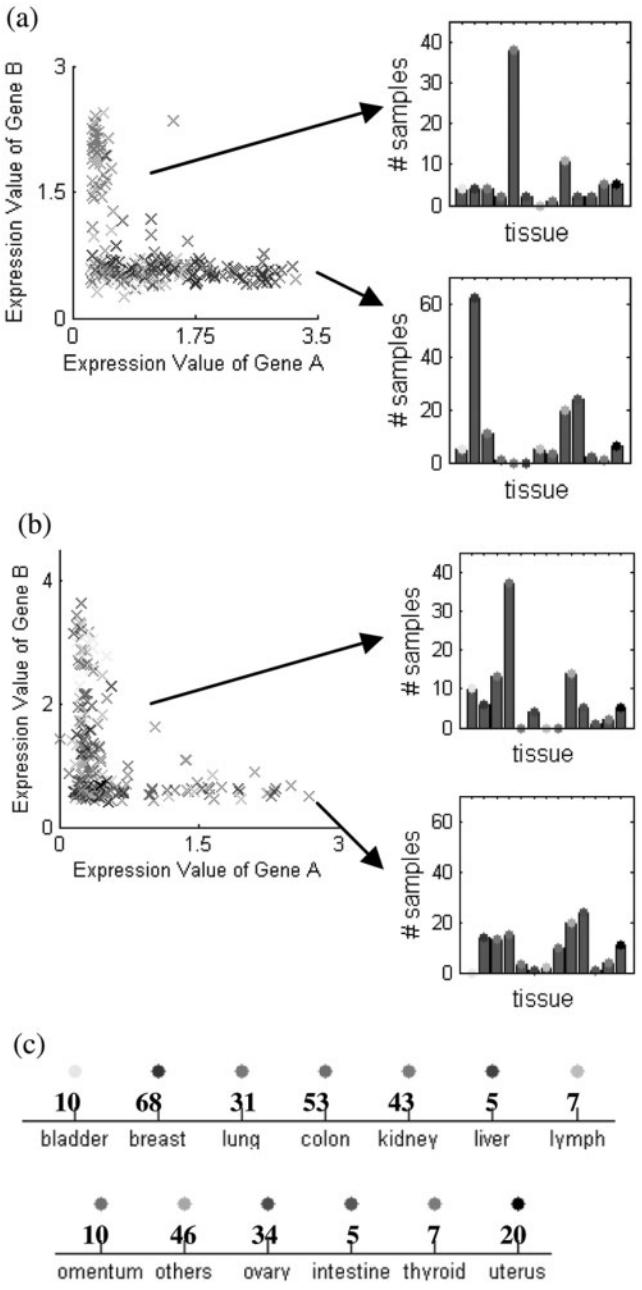

Fig. 5. Tissue composition in L-shaped relationships (**a**) The 'L'-shaped pattern corresponds to that in Figure 3c; (**b**) The 'L'-shaped pattern corresponds to that in Figure 4c. The histograms on the left show the tissue type compositions of each arm in the 'L'. The colors represent tissues types as indicated in **c**.

are mutually exclusive (i.e. 'L'-shaped). In cancer tissues, because the two genes are fused together, they become strictly co-expressed. There is no known controller gene in this example to affect TMPRSS2 and ETS.

We have no direct experimental evidence to support the interactions of the top triplets from our survey. However, it is possible to test these triplet interactions experimentally. For example, we may manipulate the expression level of a control gene using RNA interference and examine the co-expression pattern of two other genes.

In recent years, there have been several algorithms developed for inferring three-way interactions based on fuzzy logic (Woolf and Wang, 2000), mutual information (Bowers *et al*., 2004) and liquid association (LA) (Li, 2002). The mutual information method was applied to phylogenetic profile data to identify interacting protein triplets (Bowers *et al*., 2004). This method has not been tested with gene expression profiling data. In LA method, all genes in a three-way interaction are supposed to have unimodal distributions. The controller gene acts as a continuous modulator rather a qualitative switch as in our method. To ensure unimodal distributions, LA method transforms all gene expression values to their normal quantiles. We evaluated the triplets shown in Figure 2 with LA method and found all of them are highly significant (P-value $< 10^{-4}$). However, when LA method was used to conduct our large-scale survey, we encountered some difficulties. We found that LA method yielded P-values < 0.01 in 20% cases of randomly assembled triplets from our data, i.e. small P-values occurred frequently. Thus, to distinguish the top significant triplets from ~0.5 billion triplets, we had to compute the small P-values with very high precision. Because LA method relies on permutation to compute P-values, it is impractical to compute a large number of extremely small P-values. Besides methodological differences, these studies mentioned above are also different from ours because they used much smaller sample sizes and did not show validation. These studies (Bowers *et al*., 2004; Li, 2002; Woolf and Wang, 2000) did describe a few examples of triplet interactions with plausible biological mechanisms.

Our method is also closely related to another recently developed method called differential correlation (Shedden and Taylor, 2004), which seeks for significant changes in correlation of expression between a gene pair that can be associated with a dichotomized clinical factor. With our method, the clinical factor is replaced by expression of another gene.

Experimentally or computationally, we note that there is a lack of well-studied gene triplet interactions. However, we believe this is not due to lack of three-way gene interactions in the cells, but due to the fact that most studies focus on characterizing gene relationships involving only two genes at a time. However, as we shown in our study here, three-way interaction cannot be decomposed as sum of three pairwise interactions; strong three-way interactions can be identified when there is no sign of obvious pairwise correlations. We hope this study will attract more experimentalists to examine triplet interactions among genes.

ACKNOWLEDGEMENTS

We thank Margaret Newell for editorial assistance and Haitao Zhao for managing and processing the microarray database. L.Z.' s research was partially funded by M. D. Anderson Cancer Center Institutional Research Grant and NIH grants (CA108558-01; CA016672-28). Y.J.' s research was partially supported by the CML P01 grant, CA049639.

Conflict of Interest: none declared.

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023

REFERENCES

Alm,E. and Arkin,A.P. (2003) Biological networks. *Curr. Opin. Struct. Biol.*, **13**, 202.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bowers,P.M. *et al.* (2004) Use of logic relationships to decipher protein network organization. *Science*, **306**, 2246–2249.

Brouillard,F. *et al.* (2001) NF-kappa B mediates up-regulation of CFTR gene expression in Calu-3 cells by interleukin-1beta. *J. Biol. Chem.*, **276**, 9486–9491.

Choi,J.K. *et al.* (2005) Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, **21**, 4348–4355.

Dettling,M. *et al.* (2005) Searching for differentially expressed gene combinations. *Genome Biol.*, **6**, R88.

Efron,B. (2007) Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.*, **102**, 93–103.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fisher,L.D. and Belle,G.v. (1993) *Biostatistics*. John Wiley & Sons Inc, New York, NY.

Fraley,C. and Raftery,A. (2002) Model based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Getz,G. *et al.* (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Ihmels,J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Ji,Y. *et al.* (2005) Applications of beta-mixture models in bioinformatics. *Bioinformatics*, **21**, 2118–2122.

Jornsten,R. and Yu,B. (2003) Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics*, **19**, 1100–1109.

Lander,E.S. (1999) Array of hope. *Nat. Genet.*, **21**, 3–4.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Luscombe,N.M. *et al.* (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol.*, **3**, 0040.

Pruitt,K.D. *et al.* (2003) NCBI reference sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.

Quackenbush,J. (2003) Genomics. Microarrays—guilt by association. *Science*, **302**, 240–241.

Rao,C.V. and Arkin,A.P. (2001) Control motifs for intracellular regulatory networks. *Annu. Rev. Biomed. Eng.*, **3**, 391–419.

Rzhetsky,A. and Gomez,S.M. (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, **17**, 988–996.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Shedden,K. and Taylor,J. (2004) Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. In (ed.) *Methods of Microarray Data Analysis*. Kluwer Academic Publishers, Boston.

Su,A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4.

Suhasini,M. and Pilz,R.B. (1999) Transcriptional elongation of c-myb is regulated by NF-kappaB (p50/RelB). *Oncogene*, **18**, 7360–7369.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Thieffry,D. *et al.* (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays*, **20**, 433–440.

Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Wagner,A. (2002) Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Res.*, **12**, 309–315.

Woolf,P.J. and Wang,Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics*, **3**, 9–15.

Wu,C.J. *et al.* (2004) Gene expression module discovery using gibbs sampling. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 239–248.

Zhang,L. *et al.* (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.

Zhang,M.Q. (2002) Extracting functional information from microarrays: a challenge for functional genomics. *Proc. Natl Acad. Sci. USA*, **99**, 12509–12511.

Downloaded from https://academic.oup.com/bioinformatics/article/23/21/2903/374314 by guest on 26 May 2023