

Research article

Open Access

Detecting intergene correlation changes in microarray analysis: a new approach to gene selection

Rui Hu¹, Xing Qiu*¹, Galina Glazko¹, Lev Klebanov² and Andrei Yakovlev¹

Address: ¹Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, New York 14642, USA and ²Department of Probability and Statistics, Charles University, Sokolovska 83, Prague 18675, Czech Republic

Email: Rui Hu - Rui_Hu@urmc.rochester.edu; Xing Qiu* - Xing_Qiu@urmc.rochester.edu; Galina Glazko - Galina_Glazko@urmc.rochester.edu; Lev Klebanov - levbkl@gmail.com; Andrei Yakovlev - rhu@bst.rochester.edu

* Corresponding author

Published: 15 January 2009

Received: 14 October 2008

BMC Bioinformatics 2009, 10:20 doi:10.1186/1471-2105-10-20

Accepted: 15 January 2009

This article is available from: <http://www.biomedcentral.com/1471-2105/10/20>

© 2009 Hu et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Microarray technology is commonly used as a simple screening tool with a focus on selecting genes that exhibit extremely large differential expressions between different phenotypes. It lacks the ability to select genes that change their relationships with other genes in different biological conditions (differentially correlated genes). We intend to enrich the above procedure by proposing a nonparametric selection procedure that selects differentially correlated genes.

Results: Using both simulations and resampling techniques, we found that our procedure correctly detected genes that were not differentially expressed but differentially correlated. We also applied our procedure to a set of biological data and found some potentially important genes that were not selected by the traditional method.

Discussion and Conclusion: Microarray technology yields multidimensional information on the function of the whole genome. Rather than treating intergene correlation as a nuisance to the traditional gene selection procedures which are essentially univariate, our method utilizes the rich information contained in the correlation as a new selection criterion. It can provide additional useful candidate genes for the biologists.

Background

It has become common practice to use microarray technology to find "interesting" genes by comparing two or more different phenotypes. Modern methods of microarray data analysis typically employ two-sample statistical tests to test differential expression of genes, combined with multiple testing procedures to guard against Type 1 errors (see [1,2] for reviews). Such methods are biased towards selecting those genes that display the most pronounced differential expression. Once the list of genes showing statistically significant differential expression has been generated, these genes are often ranked using purely

statistical criteria and this ranking is thought to reflect their relative importance. Quite typically, a certain number of genes with the smallest p -values are finally selected from the list of all "significant" genes. While most biologists recognize that the magnitude of differential expression does not necessarily indicate biological significance, in the absence of better methods, this remains the dominant means to initially prioritize candidate genes. From a biological perspective, the above-described paradigm is far from a perfectly valid approach, because genes are not independent entities – they can interact with each other in many ways. As an example, a "chain reaction"

type of a dependence structure of gene expressions was documented in the literature [3]. In such situation, even a very small change in expression of a particular gene may have dramatic physiological consequences if the protein encoded by this gene plays a *catalytic* role in a specific cell function. Many other downstream genes may amplify the signal produced by this truly interesting gene, thereby increasing their chance to be selected by formal statistical methods. For an upstream regulatory gene, however, the chance of being selected by such methods may diminish as one keeps hunting for downstream genes that tend to show bigger changes in their expression. As a result, the initial list of candidates may be inflated with many effector genes that do little to elucidate the fundamental mechanisms of biological processes.

There are two natural ways to remedy this situation. One way is to use bioinformatics tools that utilize prior biological knowledge, such as partially known pathways, to prioritize candidate genes. This approach is now routinely used in biological studies and there are ongoing efforts to enrich it with specially designed algorithms [4]. The main weakness of the above approach is that the current biological knowledge is still quite limited and sometimes inaccurate. Another way is to extract additional information on the changes of the *relationships* between different genes from microarray data by using pertinent statistical methods.

For example, if an upstream gene ceases to be catalytic in one phenotype, or this gene is active in two *different* biological pathways in two phenotypes, a carefully designed statistical test based on the intergene dependence structure should be able to detect this change. In more general situations, intergene dependence structure alone may be insufficient to pick up those upstream genes directly, but knowing the relationship changes across conditions points out possible directions for searching the interesting genes.

Notwithstanding the importance of testing for differential expression of genes, we suggest approaching the problem of microarray data analysis from a different angle. We designed a new method to select those genes that are likely to change their relationships with other genes. More specifically, we suggest selecting candidate genes using a statistical test that detects changes in the whole correlation vector associated with each gene. This additional information will be instrumental in making the final selection of candidate genes more meaningful.

We propose to enrich the statistical inference from microarray gene expression data by testing the following hypothesis: *the i th gene does not change its relationships with all other genes across the two phenotypes (conditions) under*

study. This can be accomplished by comparing the joint distribution of the correlation coefficients between this gene and other genes in different conditions.

We conducted a series of simulations with different configurations. The results obtained by our method were compared with those of a similarly designed univariate selection procedure. We observed that our method correctly selected those genes which change correlations with other genes but retained the same marginal distributions.

We also conducted various experiments with biological data. A large set of childhood leukemia data available from St. Jude Children's Research Hospital [5] were used. Our method selected some genes which were not selected by a comparable univariate approach.

Biological data

The biological dataset used in this study was from the St. Jude Children's Research Hospital (SJCRH) Database on childhood leukemia [5]. Two groups of data were employed: patients ($n = 88$) with hyperdiploid acute lymphoblastic leukemia (HYPERDIP) and patients ($n = 79$) with a special translocation type of acute lymphoblastic leukemia (TEL). To make two data groups more comparable, only the first 79 patients in HYPERDIP were used.

Since the original probe set definitions in Affymetrix GeneChip data were known to be inaccurate [6], we updated them by using a custom CDF file to produce values of gene expressions. The CDF file was obtained from <http://brainarray.mbni.med.umich.edu>. After that, each patient was represented by an array reporting the logarithm (base 2) of expression level on the set of 7084 genes. For both datasets, each gene was standardized so that it had zero mean and unit standard deviation. This was to avoid introducing false correlation coefficients when doing permutations.

Methods

Correlation vectors

Let us denote m as the number of genes. For the i th gene, we computed the $(m - 1)$ -dimensional random vector $\mathbf{r}_i = (r_{i1}, \dots, r_{i, i-1}, r_{i, i+1}, \dots, r_{im})$. Here r_{ij} is the sample correlation coefficient between the i th and the j th gene. This vector represents the relationships between the i th gene and all other genes. Denote the $(m - 1)$ -dimensional joint distribution functions of \mathbf{r}_i in two different conditions by $F_{\mathbf{r}_i(A)}(x)$ and $F_{\mathbf{r}_i(B)}(x)$. A pertinent statistical test can be designed to test the basic null hypothesis

$$\mathbf{H}_i : F_{\mathbf{r}_i(A)}(x) = F_{\mathbf{r}_i(B)}(x). \quad (1)$$

To increase the sensitivity of our test to departures from H_{i_0} , especially when the correlation coefficients are very high, we applied the Fisher transformation to the sample correlation coefficients:

$$w_{ik} = \frac{1}{2} \log \frac{1+r_{ik}}{1-r_{ik}}, \tag{2}$$

where $k = 1, \dots, i - 1, i + 1, \dots, m$. The power improvement was confirmed by our simulation (see Table 1). We denote the correlation vectors in two conditions by $w_i(A)$ and $w_i(B)$, respectively.

Instead of testing H_{i_0} , we tested

$$H'_i = F_{w_i(A)}(x) = F_{w_i(B)}(x), \tag{3}$$

where $F_{w_i(A)}(x)$ and $F_{w_i(B)}(x)$ are the joint distribution functions of $w_i(A)$ and $w_i(B)$, respectively. If H'_i was rejected, we declared the i th gene to be a differentially correlated gene.

In order to test the hypotheses based on the joint distribution functions of correlation vectors, we needed to create samples of correlation vectors. The following two methods were employed for this purpose:

- Group method: Divide each dataset into 8 subgroups, each containing 10 slides (9 slides for the last subgroup of the biological data). By computing correlation vectors from each subgroup, we obtained a sample of size 8.
- Resampling method: Randomly select 60 slides to calculate the correlation vector. Repeat 20 times to get 20 correlation vectors in each group, respectively.

Through the simulations, we found that these two methods were comparable, with the resampling method being slightly better in terms of testing power (see Table 1, 2, 3 and 4). However, the resampling method was much more computationally demanding. As an example, it took approximately 30 hours to analyze the biological data (7084 genes and 79 slides in both conditions, 10,000 permutations) with the group method.

For the resampling method, the computation time was 216 hours instead. All computations were done in a Saturn cluster computer which includes 2 nodes each with 8× AMD Opteron dual-core processors 2 GHz (16 processor cores), 16×2GB SDRAM.

Throughout this paper, we use the group method unless otherwise noted.

N-statistic

We chose a multivariate nonparametric test based on N -statistic with Euclidean kernel for testing the hypothesis H'_i . This statistic has been successfully used to select differentially expressed genes and gene combinations in microarray data analysis [7-10]. The N -statistic is defined as follows:

$$N_i = \frac{2}{n_s^2} \sum_{k=1}^{n_s} \sum_{l=1}^{n_s} L(w_i(A, k), w_i(B, l)) - \frac{1}{n_s^2} \sum_{k=1}^{n_s} \sum_{l=1}^{n_s} L(w_i(A, k), w_i(A, l)) - \frac{1}{n_s^2} \sum_{k=1}^{n_s} \sum_{l=1}^{n_s} L(w_i(B, k), w_i(B, l)), \tag{4}$$

where n_s is the number of correlation vector samples in each group, $w_i(\cdot, k)$ indicates the k th replication of the correlation vector (using the group or the resampling method), and $L(x, y) = \|x - y\| = \sqrt{\sum_{s=1}^d (x_s - y_s)^2}$ is the kernel defined by Euclidean distance.

Table 1: SIMU2, true positives (TP) and false positives (FP) in simulations with dependent base.

Effect Size	CV Method		WRS Method	
	FP mean(STD)	TP mean(STD)	FP mean(STD)	TP mean(STD)
0.1	0.25(0.7)	0.1(0.3)	0.05(0.22)	0.0(0.0)
0.2	0.9(2.39)	1.2(4.79)	0.15(0.48)	0.05(0.22)
0.3	1.1(3.51)	4.9(10.77)	0.15(0.48)	0.2(0.87)
0.4	0.9(3.48)	80.5(25.79)	0.65(1.35)	0.0(0.0)

In CV method with effect size 0.4, the TP drops to 2.9(6.82) without Fisher transformation. Total number of genes: 708. Number of differentially correlated genes: 100. Method: group method. Extended Bonferroni threshold: 1.0.

Table 2: SIMUI, true positives (TP) and false positives (FP) in simulations with independent base.

Effect Size	CV Method		WRS Method	
	FP mean(STD)	TP mean(STD)	FP mean(STD)	TP mean(STD)
0.1	1.0(1.0)	4.1(4.15)	0.65(0.85)	0.25(0.54)
0.2	0.6(0.73)	37.4(14.47)	0.75(0.83)	0.1(0.3)
0.3	1.1(1.04)	85.45(11.1)	0.9(1.09)	0.1(0.3)
0.4	0.9(0.77)	97.95(3.25)	0.85(0.91)	0.05(0.22)

Total number of genes: 708. Number of differentially correlated genes: 100. Method: group method. Extended Bonferroni threshold: 1.0.

After this step, the *i*th gene was assigned a non-negative number N_i , a measurement of how much intergene correlation structure had changed from condition *A* to condition *B*.

0.1 Resampling based *p*-values

We used the following algorithm to obtain resampling based *p*-values for each gene:

1. Randomly shuffle the slides in two different conditions, then split them into two groups.
2. Compute correlation vectors for each gene by using the group method or the resampling method.
3. Compute *N*-statistic for each gene based on the correlation vectors.
4. Repeat the above steps for $K = 10,000$ times, record the resampling based *N*-statistics as N_{ik} , $i = 1, \dots, m$, $k = 1, \dots, K$. They can be used to construct the (resampling based) null distribution for each index *i*.
5. Compute N_i , the *N*-statistic for each gene (without random shuffles).

6. Obtain the resampling based *p*-value, p_i , by comparing N_i with the null distribution constructed from N_{ik} . Specifically, p_i is defined to be $\frac{\#(N_{ik} \geq N_i)}{K}$, the proportion of N_{ik} which is greater than or equal to N_i .

Finally, we applied the extended Bonferroni adjustment [11] with threshold 1.0 to control PFER (per-family error rate). Extended Bonferroni adjustment is less conservative than the FWER (familywise error rate) controlling procedures and more stable than FDR (false discovery rate) controlling procedures in the context of microarray analysis. More details about this multiple testing adjustment procedure can be found in [11].

From the computational perspective, it was very tempting to reduce the number of permutations by pooling all N_{ik} to construct one grand null distribution. However, we noticed that the null distributions for different genes can be very different. Based on our biological data, the density functions for the significant genes tended to shift to the left compared to those associated with the non-significant genes (see Figure 1).

Table 3: SIMUI, true positives (TP) and false positives (FP) in simulations with independent base.

Effect Size	CV Method		WRS Method	
	FP mean(STD)	TP mean(STD)	FP mean(STD)	TP mean(STD)
0.1	0.95(0.92)	4.1(2.68)	0.65(0.85)	0.25(0.54)
0.2	0.65(0.85)	42.9(13.86)	0.75(0.83)	0.1(0.3)
0.3	0.65(0.73)	92.1(8.28)	0.9(1.09)	0.1(0.3)
0.4	1.45(1.07)	99.3(1.82)	0.85(0.91)	0.05(0.22)

Total number of genes: 708. Number of differentially correlated genes: 100. Method: resampling method. Extended Bonferroni threshold: 1.0.

Table 4: SIMU2, true positives (TP) and false positives (FP) in simulations with dependent base.

Effect Size	CV Method		WRS Method	
	FP mean(STD)	TP mean(STD)	FP mean(STD)	TP mean(STD)
0.1	1.05(2.77)	0.25(0.7)	0.05(0.22)	0.0(0.0)
0.2	0.85(2.87)	1.35(4.99)	0.15(0.48)	0.05(0.22)
0.3	0.55(1.56)	6.25(12.32)	0.15(0.48)	0.2(0.87)
0.4	1.1(4.57)	86.7(21.86)	0.65(1.35)	0.0(0.0)

Total number of genes: 708. Number of differentially correlated genes: 100. Method: resampling method. Extended Bonferroni threshold: 1.0.

Univariate gene selection method

We would like to emphasize that our method (henceforward denoted as the CV method) is nonparametric. Because of this, we decided to compare the CV method to a nonparametric univariate gene selection method: Wilcoxon rank-sum test with the same extended Bonferroni adjustment (henceforward denoted as the WRS method).

Results

Simulations

To gain better insight into the performance of the CV method, we simulated several sets of data. All sets had two groups of 80 arrays representing two different biological conditions (condition A and condition B). Each array had $m = 708$ genes. Denote the genes in the first condition by x_i , $1 \leq i \leq m$ and genes in the second condition by y_i , $1 \leq i \leq m$. For both groups, all genes were identically distributed with marginal distribution $N(0, 1)$. With different baseline correlation structure, we had the following two classes of simulated datasets:

- **SIMU1:** Any two distinct genes that were both in the set of the first 100 genes were correlated with coefficient ρ_d in condition A, 0.0 in condition B. Otherwise the correlation coefficient was 0.0. Here ρ_d was a constant taking value in $\{0.1, 0.2, 0.3, 0.4\}$. Condition B can be thought of as the control condition where genes were independent of each other. We called this dataset the independent base data.
- **SIMU2:** Any two distinct genes that were both in the set of the first 100 genes were correlated with coefficient $0.5 + \rho_d$ in condition A, 0.5 in condition B. Otherwise, the correlation coefficient was 0.5. Again, ρ_d was a constant taking value in $\{0.1, 0.2, 0.3, 0.4\}$. Unlike SIMU1, the baseline intergene correlation was 0.5. We called this dataset the dependent base data.

By this design, the differentially correlated genes were the first 100 genes for both SIMU1 and SIMU2. ρ_d can be seen

as a parameter indicating how much correlation structure had changed across two conditions.

For SIMU1 and SIMU2 and every ρ_d we applied both the CV method and WRS method, and recorded the true/false positives. We also repeated this process 20 times with different random seeds to get the mean and standard deviation of the true/false positives. The results are shown in Table 1 and Table 2. As expected, the CV method detected differentially correlated genes while the WRS method did not. The power of the CV method clearly increased as the effect size gets larger. Also, it was easier to detect differentially correlated genes in the independent base data than in the dependent base data. This means that high baseline correlation structure deteriorated the power of the CV method.

Simulations with biological data

The difference between SIMU1 and SIMU2 was that the baseline intergene correlation was much higher in SIMU2. This was an attempt to model the intergene dependence structure in biological data. In some sense, a better way of modeling the actual dependence structure is through resampling from the biological data.

First, we combined HYPERDIP and TEL data and randomly permuted the slides. We then divided them into two groups of an equal number of slides, mimicking two biological conditions without differentially correlated genes. For both conditions, genes were standardized so that the sample means equaled zero and the sample standard deviations were one. We denoted the entries in two groups by x_{ij} and y_{ij} , $1 \leq i \leq 7084$ and $1 \leq j \leq 79$, and the correlation matrix of these two groups by $\{\rho_{ik}\}$, $1 \leq i, k \leq 7084$.

Next, we generated a 79-dimensional random vector with *i.i.d.* standard normal components. Denote this vector by $\mathbf{a} = \{a_j\}$, $1 \leq j \leq 79$. We added \mathbf{a} to the first 300 row vectors in the first condition with a tuning parameter α as follows:

Estimated Null Distribution Densities

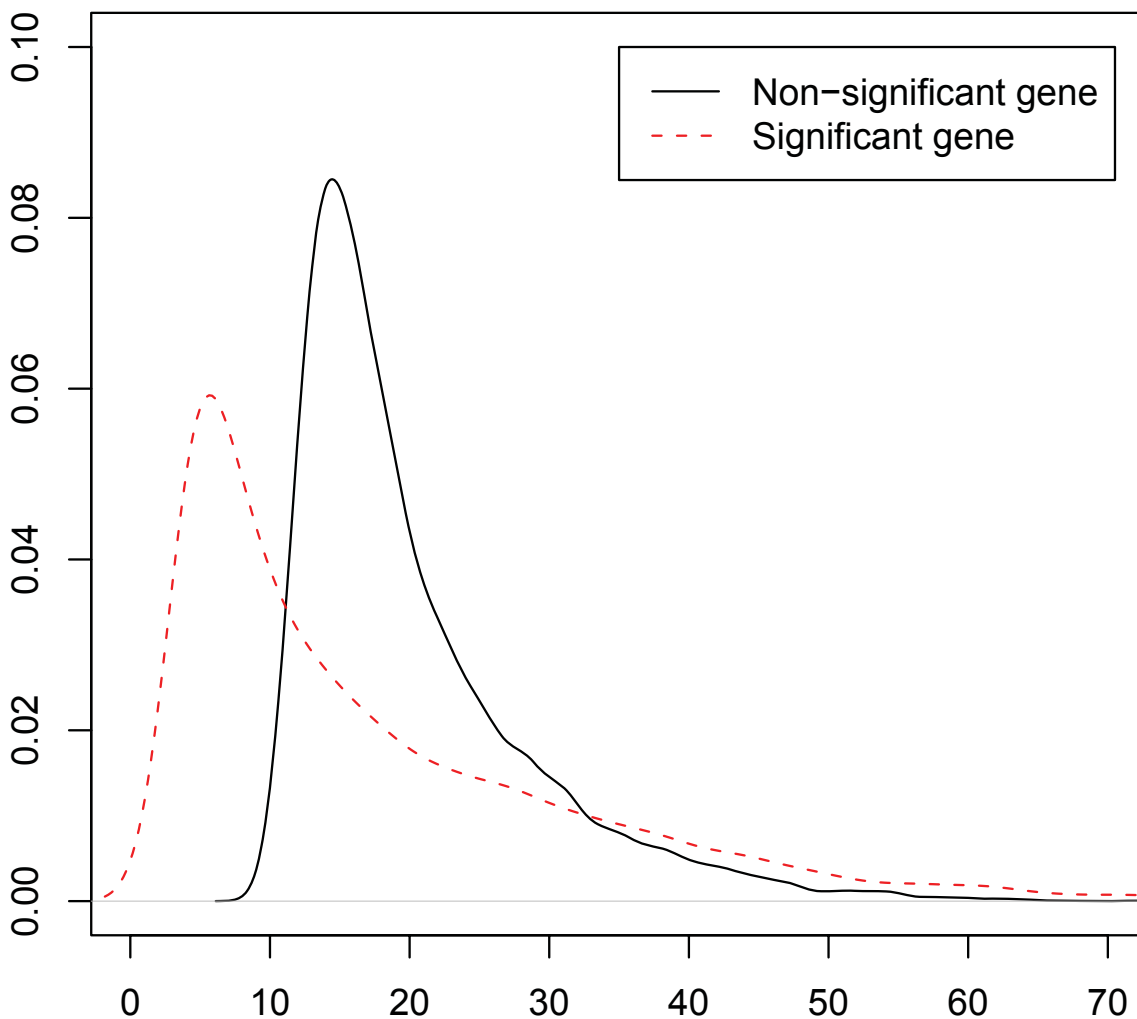


Figure 1
Estimated Null Density Functions (resampling method with Bonferroni threshold 0.05) without shuffling slides, the estimated N-statistic is 173.83 for the significant gene and 25.74 for the non-significant gene. The estimated density functions of most other genes follow the same pattern.

$\sqrt{1-\rho}x_{ij} + \sqrt{\rho}a_j$. These transformed entries are denoted as x_{ij} again, and we name this dataset SIMU3.

The first condition had the correlation coefficients as follows:

$$\text{corr}(x_i, x_k) = \begin{cases} \rho_{ik} + \rho(1 - \rho_{ik}) & \text{for } 1 \leq i, k \leq 300 \text{ and } i \neq k, \\ \rho_{ik} \sqrt{1 - \rho} & \text{for } 1 \leq i \leq 300 \text{ and } 301 \leq k \leq 7084, \\ \rho_{ik} & \text{for } 301 \leq i, k \leq 7084 \text{ and } i \neq k, \end{cases}$$

Noticeably, the correlation coefficients between any two of the first 300 genes of the first group differed substan-

tially from those of the second group, and these were considered as differentially correlated genes. The correlation coefficients between any two of the remaining 6784 genes of the first group were the same as the corresponding correlation coefficients between those of the second group, so they were not considered as differentially correlated genes.

There was one caveat in this approach. Even if x_i was one of the 6784 genes that were not differentially correlated, $corr(x_i, x_k)$ was still different in two conditions if x_k happened to be a gene from the first 300 genes. In other words, the first 300 differentially correlated genes induced some small changes for those genes that were not differentially correlated. In practice however, when summarized through the N -statistic, these differences for the latter 6784 genes were negligible and it was reasonable to view them as random fluctuations.

We set $\alpha = 0.5$ in SIMU3. As before, we applied both the CV method and the WRS method. The above procedure was repeated 10 times, the results were summarized in Table 5.

The CV method selected most of differentially correlated genes while WRS method did not. The CV method also produced fewer false positives than the WRS method.

Data analysis based on biological data

Using the WRS method, we found 102 differentially expressed genes. Using the CV method (resampling method was used here to gain more power), we detected 16 differentially correlated genes. Out of these, 11 were differentially expressed and 5 were not. These 5 genes were: CD1C (antigen precursor), HDHD1A (haloacid dehalogenase-like hydrolase domain containing 1A, enzyme involved in many catalytic activities), BASP1 (brain acid-soluble protein 1), CYB5A (cytochrome b5 type A, microsomal) and TFPI (tissue factor pathway inhibitor, which helps to regulate the extrinsic blood coagulation cascade). In the original study, two of these genes were found as differentiating among two leukemia subtypes (BASP1 and CYB5A) and the other three were never mentioned. Differential correlation of these genes

in two leukemia subtypes might provide some valuable information for better understanding the underlying subtypes' differences; however, these genes could not be captured by conventional tests. The results of this study (with multiple thresholds before and after the Bonferroni adjustments) can be found in Table 6.

Discussion and Conclusion

Our method represents a radical conceptual change from current approaches focused solely on differentially expressed genes. However, this method is not intended to replace the existing methodology but rather to provide biologists with an *additional source of information* for decision making. As an example, the univariate method failed to detect Gene CYB5A, which had a modest unadjusted p -value 0.168 based on Wilcoxon ranksum statistic. Yet CYB5A was detected as a differentially correlated gene with an unadjusted p -value 0.0 (its observed N -statistic was larger than all permutation N -statistics). To get a rough idea of how many genes had different correlation coefficients with CYB5A across two conditions, we looked at the marginal distributions of CYB5A' correlation vector. 252 genes were detected to have changed correlations with CYB5A dramatically across conditions. The selection procedure of these 252 genes can be summarized as follows: First, we splitted the slides in two conditions into 8 subgroups, respectively, as in the Group method. Second, we calculated 8 correlation vector samples of gene CYB5A in each condition. Finally, for all 7083 correlations (with 8 samples in each condition), we applied Wilcoxon ranksum test to each of them to obtain an unadjusted p -value. After extended Bonferroni adjustment, 252 genes were selected at significant level 2.0. According to BioCarta, about one third of all pathways associated with these 252 genes are related to cell cycle progression, cell division and control of centrosome duplication. HYPERDIP phenotype is characterized by the presence of more than 50 chromosomes. As a consequence, all pathways working for cellular maintenance and proliferation should be highly activated in HYPERDIP phenotype. The differential correlation of genes involved in pathways, related to cell proliferation between HYPERDIP and TEL seemed reasonable and might deserve future studies.

The microarray technology yields unique multidimensional information on the functioning of the whole genome machinery at the level of transcription so that much can be learned about relationships between genes and mechanisms by which the cell assigns tasks to different genes to maintain a specific function. It is unfortunate that such an advanced technology continues to be used as a simplistic screening tool with a focus on big differences between mean values of expression measurements. The true potential of microarray technology has yet to be unveiled. It is noteworthy that recent years have seen a

Table 5: SIMU3, true positives (TP) and false positives (FP) in simulations of biological data with tuning parameter $\alpha = 0.5$.

CV Method		WRS Method	
FP mean(STD)	TP mean(STD)	FP mean(STD)	TP mean(STD)
0.0(0.0)	270.6(11.09)	0.2(0.6)	0.0(0.0)

Total number of genes: 7084. Number of differentially correlated genes: 300. Method: group method. Extended Bonferroni threshold: 1.0.

Table 6: Numbers of differentially expressed (DE) and differentially correlated (DC) genes from biological data before and after Bonferroni adjustment with variant significant levels.

	Before Adjustment		After Adjustment	
	level = 0.05	level = 0.05	level = 0.5	level = 1.0
DC Genes	275	10	10	16
DE Genes	421	68	93	102
Both DC and DE Genes	140	8	8	11
DC But Not DE Genes	135	2	2	5
DE But Not DC Genes	281	60	85	91

Total number of genes: 7084. Method: resampling method.

growing interest in correlations between gene expression levels in statistical methodologies for microarray analysis [9,12-25]. The correlation coefficient has been used extensively as a measure of similarity in gene clustering since a seminal paper by Eisen et al. [26]

However, very few studies have examined the possibility of using the intergene correlation structure to find important genes that are linked to disease. One obstacle lies in the fact that there are m different sample means but $\frac{m(m-1)}{2}$ different sample correlation coefficients. It is much harder to catch the differences hidden in the correlation matrix that has much higher degrees of freedom (25, 087, 986 in our study). Furthermore, it is much more computationally intensive to compute the sample correlation coefficients than the sample means. Consequently, we could not afford to use more than 10, 000 permutations to get finer p -value estimation, and we are thus reluctant to recommend the slower permutation based resampling method to the biologists, despite the fact that we know this method is more powerful than the group method.

As illustrated by our study on biological data (with 79 slides in each condition) we were able to identify 102 genes, which changes the medians of their (univariate) distributions, yet only 16 genes were reported as differentially correlated (see Table 6). This seeming inadequacy of power was also shown in the simulation studies (Table 1, Table 2). This phenomenon might be subject to a number of explanations. It might be caused by the small sample size. Due to the nature of the CV method, its statistical perspective is to compare the distribution of *sample correlation coefficients* instead of expression levels in different biological conditions. With the group method, we splitted 79 slides into subgroups and computed the sample corre-

lation vectors from each group. As a result, we only had eight sample correlation vectors for each condition. We could divide the data into more subgroups, and then there would be fewer slides per subgroup so that the sample correlation coefficient computed from each subgroup would be less accurate. This was a trade-off. With the resampling method, we had 20 correlation vectors. Having more than 20 resamplings would enhance the accuracy of the estimated N -statistics and improve the power; meanwhile, it would demand more computing time, making this another trade-off.

The choice of the Euclidean distance as the kernel for the N -statistics might be another culprit. The Euclidean distance kernel is a *generic* kernel that is invariant under any orthogonal transformation. In other words, it is symmetric and indifferent to all departure from the null distribution. A specifically designed kernel that is sensitive to the likely departure from the null distribution caused by the changes of correlation might significantly increase the selection power of the CV method. Last, it may have been that indeed fewer genes were differentially correlated than were differentially expressed in the biological data. The hypotheses that the CV method was testing were entirely different from those tested by the univariate selection methods, such as the WRS method. It is absolutely possible that one gene is differentially expressed but not differentially correlated, or vice versa. The very fact that 11 out of 16 differentially correlated genes were differentially expressed in our study is an interesting phenomenon that is worth further investigation.

We believe many improvements can be made to enhance the selecting power of the CV method. We also firmly believe, as larger sets of microarray gene expression data become readily available, quantitative insights into dependencies between gene expression levels will gain increasing importance.

Authors' contributions

The basic idea was first proposed by AY, LK and XQ. The detailed study design was developed by all members in the research team. RH carried out the needed computations and simulations and the majority of the software development. GG conducted the pathway analysis for differentially expressed and differentially correlated genes. RH and XQ were responsible for most of the write-up of the findings.

Acknowledgements

This research is supported by NIH Grant GM079259 (X. Qiu), NIH/NIGMS grants RO1 GM075299 (A. Almudevar), project IAA101 I20801 and grant MSM 002 I60839 from the Ministry of Education, Czech Republic (L. Klebanov). We are grateful for insightful comments from Dr. David Oakes. We would also like to thank Dr. Sung Yong Park and Ms. Christine Brower for their technical assistance with computing in general and parallel computing in particular. In addition, we also appreciate Ms. Cheryl Cicero's proofreading efforts.

References

- Dudoit S, Shaffer J, Boldrick J: **Multiple hypothesis testing in microarray experiments.** *Statistical Science* 2003, **18**:71-103.
- Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y: *Design and Analysis of DNA Microarray Investigations* Springer Verlag; 2003.
- Klebanov L, Jordan C, Yakovlev A: **A new type of stochastic dependence revealed in gene expression data.** *Stat Appl Genet Mol Biol* 2006, **5**:Article7.
- Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102**(38):13544-13549.
- Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.** *Cancer Cell* 2002, **1**(2):133-143.
- Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33**(20):e175.
- Szabo A, Boucher K, Carroll W, Klebanov L, Tsodikov A, Yakovlev A: **Variable selection and pattern recognition with gene expression data generated by the microarray technology.** *Mathematical Biosciences* 2002, **176**:71-98.
- Szabo A, Boucher K, Jones D, Tsodikov AD, Klebanov LB, Yakovlev AY: **Multivariate exploratory tools for microarray data analysis.** *Biostatistics* 2003, **4**(4):555-567.
- Xiao Y, Frisina R, Gordon A, Klebanov L, Yakovlev A: **Multivariate search for differentially expressed gene combinations.** *BMC Bioinformatics* 2004, **5**:164.
- Klebanov L, Gordon A, Xiao Y, Land H, Yakovlev A: **A permutation test motivated by microarray data analysis.** *Computational Statistics and Data Analysis* 2005.
- Gordon A, Glazko G, Qiu X, Yakovlev A: **Control of the Mean Number of False Discoveries, Bonferroni, and Stability of Multiple Testing.** *The Annals of Applied Statistics* 2007, **1**:179-190.
- Jaeger J, Sengupta R, Ruzzo WL: **Improved gene selection for classification of microarrays.** *Pac Symp Biocomput* 2003:53-64.
- Goeman JJ, Geer SA van de, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20**:93-99.
- Geman D, d'Avignon C, Naiman DQ, Winslow RL: **Classifying gene expression profiles from pairwise mRNA comparisons.** *Stat Appl Genet Mol Biol* 2004, **3**:Article19.
- Lai Y, Wu B, Chen L, Zhao H: **A statistical method for identifying differential gene-gene co-expression patterns.** *Bioinformatics* 2004, **20**(17):3146-3155.
- Shedden K, Taylor J: **Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas.** *Methods of Microarray Data Analysis IV* 2005:121-131.
- Laan MJ van der, Birkner MD, Hubbard AE: **Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives.** *Stat Appl Genet Mol Biol* 2005, **4**:Article29.
- Lu Y, Liu P, Xiao P, Deng H: **Hotelling's T² multivariate profiling for detecting differential expression in microarrays.** *Bioinformatics* 2005, **21**(14):3105-3113.
- Qiu X, Brooks AI, Klebanov L, Yakovlev A: **The effects of normalization on the correlation structure of microarray data.** *BMC Bioinformatics* 2005, **6**:120.
- Qiu X, Klebanov L, Yakovlev A: **Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes.** *Statistical Applications in Genetics and Molecular Biology* 2005, **4**:34.
- Almudevar A, Klebanov LB, Qiu X, Salzman P, Yakovlev AY: **Utility of correlation measures in analysis of gene expression.** *NeuroRx* 2006, **3**(3):384-395.
- Qiu X, Yakovlev A: **Some comments on instability of false discovery rate estimation.** *J Bioinform Comput Biol* 2006, **4**(5):1057-1068.
- Klebanov L, Qiu X, Yakovlev A: **Testing differential expression in non-overlapping gene pairs: A new perspective for the empirical Bayes method.** *Journal of Bioinformatics and Computational Biology* 2008, **6**:301-316.
- Klebanov L, Yakovlev A: **Diverse correlation structures in gene expression data and their utility in improving statistical inference.** *Annals of Applied Statistics* 2008, **1**(2):538-559.
- Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y: **A multivariate extension of the gene set enrichment analysis.** *J Bioinform Comput Biol* 2007, **5**(5):1139-1153.
- Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

