Johns Hopkins University, Dept. of Biostatistics Working Papers

3-25-2005

Searching for Differentially Expressed Gene Combinations

Marcel Dettling
Department of Oncology, Johns Hopkins Medical Institute, mdettli1@jhmi.edu

Edward Gabrielson
Departments Oncology and Pathology, Johns Hopkins Medical Institute, egabriel@jhmi.edu

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

Suggested Citation

Dettling, Marcel; Gabrielson, Edward; and Parmigiani, Giovanni, "Searching for Differentially Expressed Gene Combinations" (March 2005). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 77.
http://biostats.bepress.com/jhubiostat/paper77

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Searching for Differentially Expressed Gene Combinations

Marcel Dettling[(1)], Edward Gabrielson[(1,2)]
and Giovanni Parmigiani[(1,2,3)]

Departments of Oncology (1),
Pathology (2) and Biostatistics (3),
Johns Hopkins Medical Institutions
Baltimore MD 21205, USA

March 25, 2005

Abstract

Background: Comparison of mRNA expression levels across biological samples is a widely used approach in genomics. Available data-analytic tools for deriving comprehensive lists of differentially expressed genes rely on data summaries formed using each gene in isolation from others. These approaches ignore biological relationships among genes and may miss important biological insight provided by genomics data.

Methods: We propose a fast, easily interpretable and scalable approach for identifying pairs of genes that are differentially expressed across phenotypes or experimental conditions. These are defined as pairs for which there is detectable phenotype discrimination using the joint distribution, but not from either of the the marginal distributions of two genes. Our approach is based on comparing the phenotype-specific gene correlation matrices to the overall gene correlation matrix.

Results: Application of our approach to two cancer datasets demonstrates that these experiments include gene pairs that show a detectable relationship with phenotype only when considered jointly. Also, the gene pairs

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

identified by our method have a tendency to share biological relationships, as evidenced by further investigation of available information on gene function.

Conclusions: Important information on gene function, phenotype-related dependencies, and interactions among genes can be gleaned by systematic searches that compare the joint distributions of all possible gene pairs across conditions.

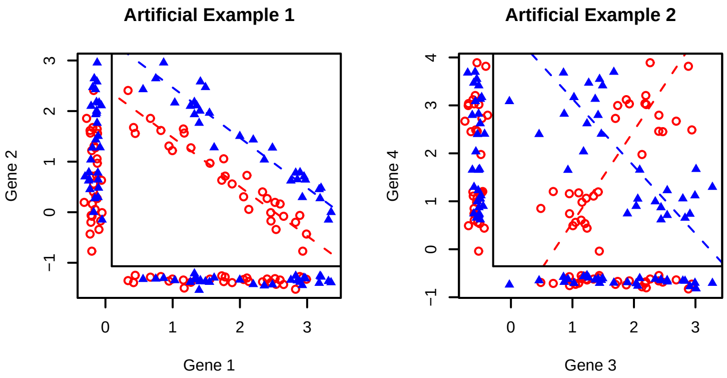
1 Background

Gene expression monitoring by microarray technologies has become an important approach in biological and medical research over the past decade. A common experimental design is the comparison of two sets of samples from different phenotypes (i.e. tumorous and normal tissue), with the goal of searching for genes showing differential expression. This is usually done via statistical testing procedures and often, subsequent multiple testing corrections. Prominent examples include *t*-testing, Significance Analysis of Microarrays [1] and Empirical Bayes analysis [2]. A comprehensive review of such approaches can be found in Pan [3]. All these methods employ a one gene at-a-time strategy, only considering the association between single genes and the phenotype.

Many approaches for classification of phenotypes using microarrays do consider multiple genes simultaneously, but they address a different question, and their goal is to produce parsimonious sets of differentially expressed genes [4]. While interesting, these approaches have the limitation that they cannot be applied comprehensively to all possible pairs, i.e. there currently are no practical tools for exploring phenotype-related dependencies and interactions among all gene pairs in large datasets. In this paper we present a methodology for addressing this issue, and we show that it can find interesting biological relationships that would be missed by existing approaches.

We are interested in searching for two types of gene pairs, illustrated in Figure 1 by artificial examples. In the left panel, the two genes show a pronounced joint association on the phenotype: if the sum of their expression levels exceeds 3 units, we solely observe the blue triangle phenotype. A biological mechanism leading to this occurs when the two genes are substitutes in a molecular process which is closely linked to the phenotype. Therefore,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive


Artificial Example 1

Artificial Example 2

Figure 1: Two artificial examples of joint differential expression. The unit of the x- and y-axis is gene expression, blue triangles and red circles correspond to samples of two different phenotypes. The inner panels reflect the joint distribution, the outer margins display the univariate marginal distributions. The dashed lines represent the first principal components, conditional on the phenotype.

we denote this situation as the *substitution case*. Note that neither of the two genes shows a strong association with the phenotype in the univariate marginal distribution, and thus would have been highly unlikely to appear in a gene list produced by a one gene at-a-time testing approach. A complementary case occurs when two genes cluster around two positively sloped axes: then, the phenotype is associated with a difference in expression, a situation we refer to as the *gap case*.

A more complex case is shown in our second artificial example in the right panel of Figure 1. There is no obvious demarcation in space, and again, neither of the two genes carries information on its own. However, together they do. Biologically speaking, this example could reflect an *on/off-situation*. If both genes are off (expression values below 1.5 units), or both genes are on (expression value above 1.5 units), we observe the red circle phenotype. In contrast, if only one of the genes is turned on, the blue triangle phenotype is predominant.

Statistically, we define joint differential expression as good phenotype discrimination by the joint distribution, but not from either of the univariate marginal distributions of two genes. From a functional genomics perspec-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

tive, such pairs could represent interesting novel biological interactions, as for example genes that are in the same pathway.

The identification of gene pairs with joint differential expression is ambitious for several reasons. First, gene pair identification is subject to the curse of dimensionality. While the usual number p of genes is in the tens of thousands, the number of gene pairs is $p(p-1)/2$, usually in the millions. Second, there are no existing and quickly computable test statistics that exactly address our notion of joint differential expression. Existing bivariate tests such as Hotelling's T^2 [5] only screen for differences in the bivariate mean vectors and will thus favor pairs that consist of genes with strong marginal effects. Third, identifying joint differential expression based on comparing predictive models for pairs and single genes is conceptually sound but unattractive due to its prohibitive computational burden.

Here we propose a novel, efficient and scalable approach for searching gene pairs with joint differential expression. It relies on calculating an appropriately defined test statistic from the unconditional, as well as both the class-conditional correlation matrices. Hence, we name our method *CorScor*, as a shorthand for correlation scoring. Its biggest advantages are its straightforward interpretation and the fact that it can be calculated very quickly, which allows for an exhaustive search among the millions of pairs even in large gene expression datasets. On the basis of two gene expression datasets from the literature, we illustrate our method, and collect empirical evidence that it yields gene pairs that are more than random artifacts and also have a tendency to share biological relationships.

2 Results

2.1 Data Preparation

We illustrate the power and utility of our method with two datasets. The first is a colon cancer dataset from Alon et al. [6], which is publicly available from `http://microarray.princeton.edu/oncology`. It originated from Affymetrix Hum6000 arrays and contains the expression values of the 2000 genes with highest minimal intensity across 62 colon tissues, 40 of which were tumorous and 22 of which were normal. We transformed the data by a base 10 log-transformation and standardized each array to zero mean and unit

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

variance across genes. The second is a breast cancer dataset from Hedenfalk et al. [7], publicly available at `http://research.nhgri.nih.gov/microarray/NEJM_Supplement`. The data were obtained from Stanford-type cDNA microarrays. They monitor 2654 genes across 22 breast cancer samples, 7 of which were found to carry germline BRCA1 mutations. Besides a base 10 log-transformation of the intensity ratios, no further normalization steps were taken. Our selection of data illustrates that *CorScor* works independently of the platform. We require accurately preprocessed expression data from n samples and p genes, stored in a $(n \times p)$-matrix denoted by (x_{ig}). In what follows, we will encode the phenotype information generically as 0 and 1, and store it in the n-dimensional response variable y.

2.2 The Gap/Substitution Cases

Our method for revealing genes with joint differential expression relies on computing a correlation-based score function. Given a pair consisting of genes g and g', we determine the correlation coefficient $\rho(g, g')$ amongst their expression vectors. Next, by restricting in turn to just the samples from each phenotype, we obtain both the class-conditional correlation coefficients $\rho_0(g, g')$ and $\rho_1(g, g')$.

For revealing gene pairs that jointly discriminate the two phenotypes according to a gap or substitution mechanism as shown by the artificial example in the left panel of Figure 1, we recommend to compute the scoring function

$$S(\rho, \rho_0, \rho_1) = |\rho_0 + \rho_1 - \alpha\rho| \tag{1}$$

for all gene pairs (g, g'). Note that the operations in (1) can be done for all gene pairs simultaneously by element-wise operations on the three $(p \times p)$-correlation matrices. As illustrated in Figure 2, gene pairs with high scores indeed show good joint differential expression on the Colon and BRCA1 data, i.e. accurate phenotype discrimination and comparably uninformative marginals.

The rationale for the success of scoring function (1) is as follows. High conditional correlations arise if the data points are tightly aligned along a straight line, which can be represented by the first principal component, shown in Figure 2 by the dashed lines. Good joint differential expression requires such tight clustering and hence, high conditional correlations with concordant sign, but it also requires a shift between the alignment axes.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

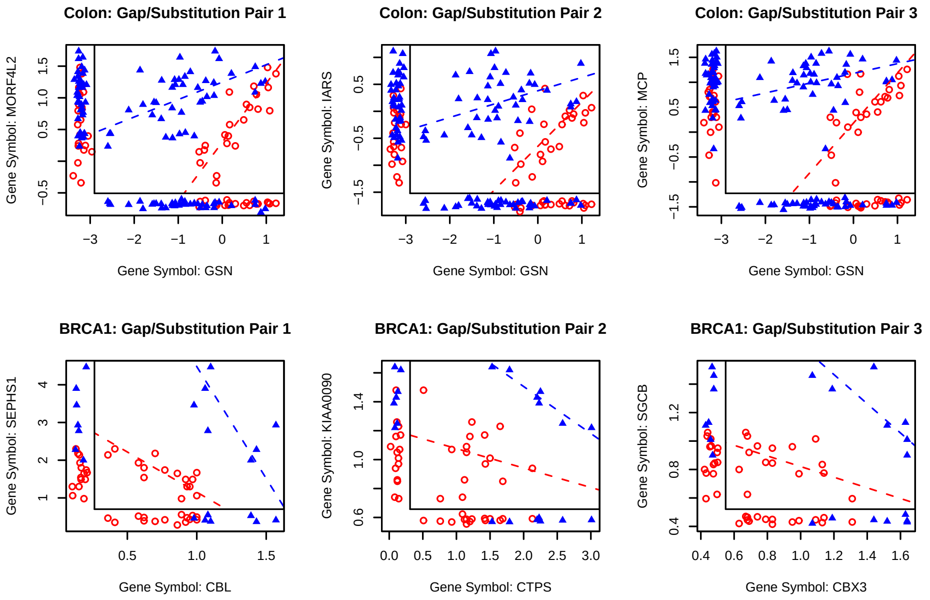


Figure 2: Six examples of joint differential expression of the gap/substitution type, obtained from the Colon and BRCA1 datasets. The inner panels show the joint distribution, the outer margins display the univariate distributions. Blue triangles stand for cancers in Colon and BRCA1 mutants in Breast, the red circles stand for normal samples in Colon and sporadic cancers in Breast. The dashed lines represent the conditional first principal components.

The bigger this shift, and thus the clearer the joint separation, the lower the unconditional correlation ρ gets. Hence, we diminish the sum of ρ_0 and ρ_1 by $\alpha\rho$. By taking the absolute value, we achieve symmetric treatment of positively and negatively sloped alignment axes, i.e. we can capture the gap and the substitution case together. The scalar tuning parameter α governs the balance between separation and parallel alignment. We observed empirically good results with $\alpha \in [1, 2]$, and use $\alpha = 1.5$ throughout the paper.

The first three columns in Table 1 show the values of ρ, ρ_0, ρ_1 and S for the three highest scoring gene pairs according to the scoring function presented in Equation (1). As expected, the class-conditional correlations ρ_0 and ρ_1 tend to be high in absolute value and concordant in their signs, whereas the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

	Colon			BRCA1		
	Pair 1	Pair 2	Pair 3	Pair 1	Pair 2	Pair 3
ρ	0.19	-0.01	0.02	0.27	0.32	0.31
ρ_0	0.84	0.65	0.67	-0.79	-0.20	-0.38
ρ_1	0.53	0.33	0.34	-0.63	-0.96	-0.78
$S(\rho, \rho_0, \rho_1)$	1.09	0.99	0.98	1.82	1.64	1.62

Table 1: Conditional and unconditional correlations coefficients, as well as the value of the scoring functions from Equation (1) with $\alpha = 1.5$, for the top 3 gene pairs in both the Colon and the Breast data.

overall correlation is low, and sometimes even has a discordant sign.

An effective means of visualization for the structure among gene pairs with joint differential expression is a heatmap, as shown in Figure 3. We select the first 50 genes involved in the top ranked gene pairs and color code the score for all $50^2/2 = 1250$ gene pairs from black (low value) over shaded grey to white (high value, excellent joint differential expression). Rows and columns of this symmetric matrix are rearranged according to a hierarchical clustering, such that genes that share common joint differential expression properties lie adjacent. We hypothesize that clustered genes tend to share biological relationship. An exploratory analysis on the Colon data supports this: a fairly tight cluster can be found at positions 39-45 of the matrix. It consist of the genes with HUGO symbols

GSN, ACTN1, SPARCL1, ITGA7, TPM1 and COL6A2.

Three out of these six genes (GSN, ACTN1 and SPARCL1) share a common annotation in the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG, [8]). They are all involved in the *regulation of actin cytoskeleton*. The remaining three genes lack of a pathway annotation in KEGG, but an analysis of their Gene Ontology terms (GO, [9]) still reveals a functional connection: TPM1 has the GO terms *actin binding* and *cytoskeleton*. SPARCL1 is involved in *calcium ion binding*, a term it shares with GSN and ACTN1.

The heatmap of the BRCA1 data, shown in the right panel of Figure 3, does not show an equally pronounced block structure. The absence of KEGG-annotation for a large portion of the genes makes it challenging to carry out the same type of validation. However, consistent with the known

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive
Hosted by The Berkeley Electronic Press

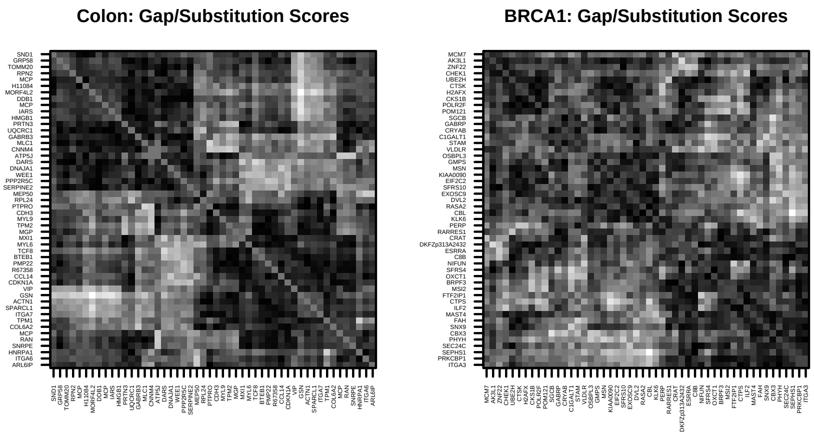


Figure 3: Symmetric heatmap of *CorScor* values from Equation (1), for the Colon and BRCA1 data. Columns and rows are rearranged according to a hierarchical clustering. Displayed are the 50 genes which are involved in the pairs with the highest scores. Black stands for very low, grey for intermediate and white for very high score.

DNA-binding function of the BRCA1 gene [10], many of the genes are related to binding activities. For a full overview of the genes involved in the heatmaps, we refer to our supplementary webpage at `http://stat.ethz.ch/~dettling/jde.html`.

Our findings on the Colon data illustrate that *CorScor* has the potential to bring up gene pairs with functional relationship, and that our heatmaps are a helpful visualization tool to group and detect the most important ones among them. The major benefit of *CorScor*, compared to established clustering techniques based on single genes' expression values, is that we are able to capture genes without strong marginal effects. The genes involved in our pairs do not show pronounced fold-changes across the phenotypes, but nevertheless seem to be key in molecular processes closely linked to the phenotype.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

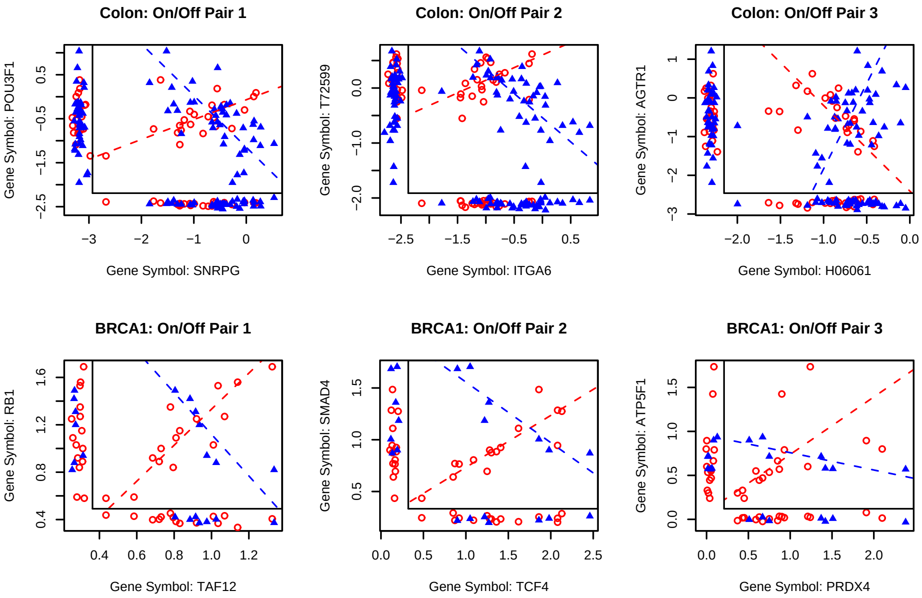


Figure 4: Six examples of joint differential expression according to the on/off-scenario, obtained from the Colon and BRCA1 data. The inner panels show the joint distribution, the outer margins display the univariate distributions. Blue triangles stand for cancerous and BRCA1 mutant, the red circles for normal and BRCA1 wild-types, respectively. The dashed lines represent the direction of the conditional first principal components.

2.3 The On/Off-Case

Another scenario where joint differential expression is important is illustrated with the artificial example in the right panel of Figure 1. While the marginal distributions are non-informative, the joint distribution clearly is: one phenotype is prevalent when either both genes' expression is turned on/turned off, whereas the other phenotype is predominant when only one of the genes is expressed. An effective correlation-based scoring function to capture these gene pairs is

$$S(\rho, \rho_0, \rho_1) = |\rho_1 - \rho_0|, \tag{2}$$

the difference of the class-conditional correlations ρ_0 and ρ_1. Table 2 shows the values of ρ_0, ρ_1 and S for the top scoring gene pairs in the Colon and

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

	Colon			BRCA1		
	Pair 1	Pair 2	Pair 3	Pair 1	Pair 2	Pair 3
ρ_0	0.54	0.48	-0.72	0.86	0.93	0.89
ρ_1	-0.67	-0.68	0.42	-1.00	-0.93	-0.95
$S(\rho, \rho_0, \rho_1)$	1.21	1.17	1.13	1.86	1.86	1.84

Table 2: Conditional and unconditional correlations coefficients, as well as the value of the scoring functions from Equation (2) for the top 3 gene pairs in both the Colon and the BRCA1 data.

BRCA1 data. We observe fairly high conditional correlations here, partly due to the fact that we use Spearman's rank correlation in the scoring function (2).

Figure 4 shows scatterplots of the highest scoring gene pairs on the Colon and BRCA1 data. Joint differential expression is clearly present and an interesting biological interpretation can be derived from these scatterplots. As an example, we discuss the best scoring gene pair from the BRCA1 data: for the wild-type samples (represented by red circles), there is a high positive correlation between TAF12, a gene that is related to transcription initiation, and RB1, a transcription inhibitor. For the BRCA1 mutant samples, the situation is reversed and the two genes show a strong negative correlation. This observation suggests a specific nuclear pathway that may be distorted as a result of BRCA1 mutations.

We emphasize again, that due to the very different scope, such findings could not be made with one-at-a-time gene selection and/or hierarchical clustering based on gene expression values. Again for this on/off-scenario, the full information and annotation of the genes which are involved in the most promising gene pairs is available from our website at `http://stat.ethz.ch/~dettling/jde.html`.

2.4 False Discovery Rate Analysis

Next, we address the question of whether and how many gene pairs achieve promising score values just by chance alone, or in other words, whether there could be any false discoveries among our gene pairs. To analyze this, we generated 100 noise gene expression datasets by scrambling the phenotype

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

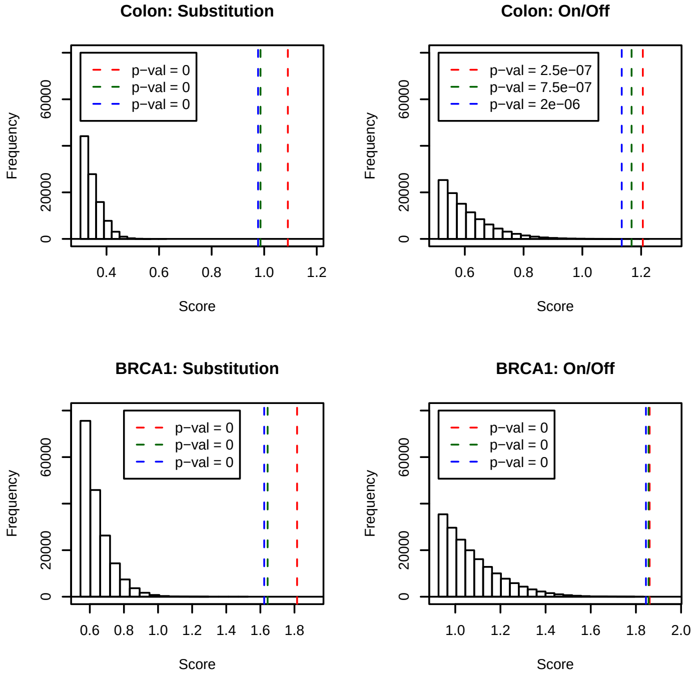


Figure 5: Histograms, displaying the right tail of the permutation distribution of *CorScor* in the Colon and BRCA1 data. The dashed vertical lines indicate the score values of the top three gene pairs from Figures 2 and 4.

labels. We then run *CorScor* and rank the resulting values. By averaging the scores within rank over the 100 permutations, we obtain a null distribution.

The histograms in Figure 5 display the right tail of the permutation distribution to the right of the 95% quantile. The dashed vertical lines mark the score value of the top three gene pairs (shown in Figures 2 and 4) on both the gap/substitution and the on/off situation, and for both datasets. The top gene pairs reach a highly significant status and exceed the permutation scores by a clear margin. The permutation distribution has a somewhat heavier tail and slower decay for the on/off-situation. Furthermore, when comparing the Colon and BRCA1 permutation scores, we observe that the latter are at a markedly higher value. This is caused by the difference in sample size. When arbitrarily restricting the Colon dataset to the same size as the BRCA1, the score values were in the same range (data not shown).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

	0	10^{-6}	10^{-5}	10^{-4}	10^{-3}
Colon (1)	1446	2990	7552	16579	37796
Colon (2)	0	2	28	354	3901
BRCA1 (1)	14	186	1560	8002	41780
BRCA1 (2)	8	28	240	2282	19658

Table 3: The number of genes that exceed a given quantile of the permutation distribution in the Colon and BRCA1 data. The numbers in parentheses refer to the scoring function, i.e. the equation where it was defined: (1) is the gap/substitution scenario, and (2) the on/off scoring situation.

Table 3 shows the number of gene pairs that exceed a given quantile of the permutation distribution. Again here, we observe that in the gap/substitution scenario, more gene pairs reach very high significance levels. In general, our results confirm that gene pairs with joint differential expression are far beyond the noise level in the microarray datasets we analyzed.

2.5 Comparison with Predictive Modeling

Next, we contrast the results of searching for jointly differentially expressed gene pairs by *CorScor* to an alternative search based on predictive modeling, implemented with logistic regression. This is also a new method, and it is presented here because it seemed a natural alternative. This approach is far more computation intensive and currently not applicable to arrays with tens of thousands of features. We chose the following procedure for our predictive modeling search: in the gap/substitution situation and for each gene pair (g, g'), we fitted three logistic regression models: a multivariate model with both genes as additive inputs, and two univariate models with each gene as input. This generates conditional probability estimates $p_i(x_g, x_{g'})$, $p_i(x_g)$ and $p_i(x_{g'})$ for each observation i. We then compute three log-likelihoods on the basis of these probabilities,

$$\ell(y, p(\cdot)) = \sum_{i=1}^{n} y_i \cdot \log(p_i(\cdot)) + (1 - y_i) \cdot \log(1 - p_i(\cdot)). \tag{3}$$

The log-likelihood is a very natural measure for the amount of discrimination in binary problems. A gene pair with good joint differential expression reflecting a gap or substitution should show good discrimination for the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

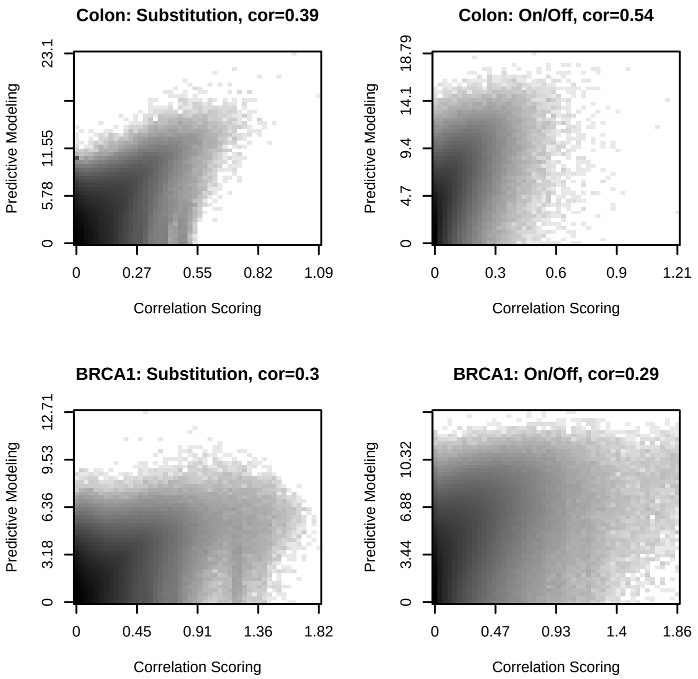


Figure 6: Density plots for a comparison of the gap/substitution scoring function from correlation scoring defined in (1) and predictive modeling (4), as well as the on/off objective measures defined in (2) and (5). Each panel is divided into a 50 × 50 cell grid. The darker the color of a cell, the more instances are therein.

multivariate model, but comparably poor discrimination for the single gene models. Hence, we can define a scoring function based on predictive modeling as

$$R(g,g') = \ell(y, p(x_g, x_{g'})) - \frac{1}{2}\Big(\ell(y, p(x_g)) + \ell(y, p(x_{g'}))\Big). \tag{4}$$

The left two panels in Figure 6 show scatterplots of *CorScors* outcome versus predictive modeling scores in the gap/substitution situation. The correlation between the two measures is 0.39 for the Colon data, and 0.30 for the BRCA1 data.

The on/off-scenario requires a different approach. For each gene pair (g, g'), we chose to measure the improvement in predictive accuracy when com-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

paring a full two-gene interaction model versus a two-gene additive model. This requires generating conditional probability estimates $p_i(x_g, x_{g'}, x_{gg'})$ and $p_i(x_g, x_{g'})$ using logistic regression for each observation i. These are then inserted into the log-likelihood from (3). From these, we can obtain a predictive modeling based scoring function for the on/off scenario via

$$T(g, g') = \ell(y, p(x_g, x_{g'}, x_{gg'})) - \ell(y, p(x_g, x_{g'})). \tag{5}$$

The concordance of this measure with *CorScors* output is illustrated in the right two panels of Figure 6. We observe a correlation of 0.54 in the colon data and 0.29 in the BRCA1 data, but many of *CorScors* top scoring gene pairs are not identified by predictive modeling.

2.6 Software

All our computations were implemented in the statistical programming language R [11]. Via its function `cor`, it provides a very convenient and efficient routine for estimating Pearson and Spearman gene pair correlation coefficients from an expression matrix. In the Colon and BRCA1 data, an exhaustive search across all gene pairs with *CorScor* takes about 5 seconds on an 1.5GHz Intel Pentium powered personal computer with 512Mb or RAM.

All our code for identifying gene pairs with joint differential expression, as well as for their visualization by scatterplots and heatmaps, is available a documented package named `corscor`, and will be submitted to the Bioconductor project [12]. Links and updates can also be found on our website at `http://stat.ethz.ch/~dettling/jde.html`.

3 Discussion

In a recent paper, Xiao and colleagues [13] considered multivariate searches for differentially expressed gene combinations. Their goal is to uncover subsets of predefined size k that are such that the multivariate distribution of expression in the two phenotypes differ. Similar ideas were used by the same group in the context of data exploration and variable selection [14, 15]. The goal of this approach is that of uncovering sets that potentially consist of combinations of joint and marginally differentially expressed genes. This is

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

a different target from that considered here. For example in Figure 4, vertically shifting all the blue points would increase multivariate difference but leave *CorScors* output unchanged. Here, we emphasize the search for interactions per se, because of the clearer functional genomics implications, but high multivariate distance can also be of interest. A weakness of the Xiao et al. approach is that computational demands make it impossible to apply it in a comprehensive fashion and one has to rely on stochastic exploration of the set of subsets. In the proposed implementation each set is evaluated by an additional cross-validation.

In section 2.5, we presented an approach to screen for joint differential expression based on predictive modeling. While this shares the scope of *CorScor*, it is not scalable to the current dimensions of gene expression data. A full search with predictive modeling on the Colon or the BRCA1 data with less than 3000 genes each requires about two weeks of CPU time, whereas *CorScor* needs about 5 seconds only. Since the number of gene pairs and thus the computing time grows quadratically with the number of genes, the analysis of a roughly quintuple sized Affymetrix HGU133 array with more than 12'000 genes would increase the computing time by a factor of roughly 25, making the predictive modeling approach prohibitive for practical application. We also observed that the gene pairs found by *CorScor* and the predictive modeling approach differ. To develop a better sense for the nature of the differences, we visually compared a large number of gene pairs from both methods (not shown). The scatterplots of the top gene pairs according to the gap/substitution predictive modeling scoring function in Equation (4) reveal that the predictive approach is very sensitive to outliers, whereas *CorScor* is much more robust in this regard. Additionally, the joint separation is often more pronounced with *CorScor*. In the on/off search, visual scatterplot inspection and examination of gene annotations favor *CorScor* even more clearly. The predictive modeling objective function in (5) does not seem to exactly match the scope of its correlation based counterpart and generally did not yield any gene pairs that could serve as indicators for distorted molecular processes.

In the on/off search in particular, a critical difference is in the fact that pairs can show strong evidence of a reversal in the sign of the conditional correlations, while still having a substantial overlap of the two conditional distributions (see for example the top left and top right pairs in Figure 4). This can lead to a high *CorScor* values, but only leads to a moderate predictive score, and a small multivariate distance. These cases however can be

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

highly relevant biologically, and it is important to be able to identify them. In conclusion, of the two approaches that we are proposing and investigating here, *CorScor* is simplest and more efficient computationally, and it also appears to identify gene pairs that are more promising candidates for a detailed biological analysis.

Relevance networks [16] examine interactions among genes by thresholding covariance matrices and graphically displaying the connections among the genes whose correlations exceed the threshold. We investigate a different type of gene interactions here, and namely interactions that are altered as a result of the comparison of interest. However, the type of visualization implemented in relevance networks could also be employed to represent the findings of our algorithm. Moreover, our approach was illustrated here using Pearson's and Spearman's correlation, but the general idea can be extended straightforwardly to any easily computed measure of pairwise association among gene expression levels.

4 Conclusions

In summary, this paper presents a novel approach for finding gene pairs with joint differential expression. This represents a complement to the widely used one gene at-a-time testing approaches and the associated list enrichment tests. The idea behind joint differential expression is to find genes that only in pairs, but not individually, discriminate two given phenotypes. These pairs allow to explore dependence and interaction among genes, as well as to screen for molecular processes which are linked to disease. Since the usual number of gene pairs is in the millions, there is a need for a quickly computable criterion. We propose two scoring functions based on conditional and unconditional correlation coefficients. We show that this measure has the ability to uncover gene pairs that show promising scatterplot patterns, tend to share biological relationship at the molecular level, and are clearly beyond the false discovery level.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Authors contributions

Dettling: planning, writing, software implementation, data analysis; Gabrielson: planning, critical reading; Parmigiani: planning, writing.

Acknowledgment

Work supported by NSF grant NSF034211, by the Johns Hopkins SPORE in breast P50CA88843 and gi cancer P50CA62924 and by grant P30CA06973.

References

[1] Tusher V.G., Tibshirani R. and Chu G. (2001), *Significance Analysis of Microarrays Applied to the Ionizing Radiation Response*. PNAS, **98**, 5116–5121.

[2] Efron B., Tibshirani R., Storey J. and Tusher V. (2001), *Empirical Bayes Analysis of a Microarray Experiment*. JASA, **96**, 1151–1160.

[3] Pan W. (2002), *A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments*. Bioinformatics, **18**, 546–554.

[4] Parmigiani G., Garrett E., Irizarry R. and Zeger S. (2003), *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York.

[5] Hotelling H. (1947), *Techniques of Statistical Analysis*, chapter Multivariate Quality Control. McGraw-Hill, New York.

[6] Alon U., Barkai N., Notterdam D., Gish K., Ybarra S., Mack D. and Levine A. (1999), *Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays*. Proceedings of the National Academy of Science, **96**, 6745–6750.

[7] Hedenfalk I., Duggan D., Chen Y. *et al.* (2001), *Gene-Expression Profiles in Hereditary Breast Cancer*. New England Journal of Medicine, **344**, 539–548.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

[8] Kanehisa M. and Goto S. (2000), *KEGG: Kyoto Encyclopedia of Gene and Genomes.* Nucleic Acids Research, **28**, 27–30.

[9] Consortium T.G.O. (2000), *Gene Ontology: Tool for the Unification of Biology.* Nature Genetics, **25**, 25–29.

[10] Paull T., Cortez D., Bowers B., Elledge S. and Gellert M. (2001), *From the Cover: Direct DNA Binding by BRCA1.* PNAS, **98**, 6086–6091.

[11] R Development Core Team, Vienna, Austria (2004), *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-00-3.

[12] Gentleman R., Carey V., Bates D. *et al.* (2004), *Bioconductor: Open Software Development for Computational Biology and Bioinformatics.* Genome Biology, **5**, R80.

[13] Xiao Y., Frisina R., Gordon A., Klebanov L. and Yakovlev A. (2004), *Multivariate Search for Differentially Expressed Gene Combinations.* BMC Bioinformatics, **5**, 1–10.

[14] Szabo A., Boucher K., Carroll W., Klebanov L., Tsodikov A. and Yakovlev A. (2002), *Variable Selection and Pattern Recognition with Gene Expression Data Generated by the Microarray Technology.* Mathematical Biosciences, **176**, 71–98.

[15] Szabo A., Boucher K., Jones D., Klebanov L., Tsodikov A. and Yakovlev A. (2003), *Multivariate Exploratory Tools for Microarray Data Analysis.* Biostatistics, **4**, 555–567.

[16] Butte A.J., Tamayo P., Slonim D., Golub T.R. and Kohane I.S. (2000), *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.* **97**, 12182–12186.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper77