

Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas

Kerby Shedden
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1027
1-734-764-0438

kshedden@umich.edu

Jeremy Taylor
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109-2029
1-734-936-3287

jmgmt@umich.edu

ABSTRACT

We propose a simple data analysis procedure that aims to uncover a form of differential gene expression. Rather than focus on differences in the group means, as is usually done, we search for pairs of genes such that the strength or direction of their interaction is associated with an outcome variable. This more complex form of differential expression may be especially relevant in studying clinical outcomes such as survival and grade, since it has often been difficult to identify marker genes whose mean expression varies directly with such outcomes. In applying our method to two lung cancer microarray data sets, we discovered that a substantially greater number of genes are likely to be associated with clinical outcomes such as tumor stage via differential correlation than are associated via changes in mean expression.

Categories and Subject Descriptors

J.3 [Life and Medical Sciences]

General Terms

Algorithms, Measurement, Experimentation

Keywords

Differential correlation, gene expression, interaction.

1. INTRODUCTION

Important clinical disease characteristics such as survival times and tumor stage often exhibit only weak associations with gene expression. One possible reason for this may be that complex clinical responses are biologically manifested in subtle ways, and hence may not be detected using conventional statistical measures that look for expression shifts in the average levels of single "marker genes".

We propose a simple analysis method for relating gene expression levels to binary response variables that aims to detect a link between the degree of association within a pair of genes and the clinical response. This is a more subtle form of interaction compared to differences in mean expression, which are often used to identify

differential
expression.
Since the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are Conference '00, Month 1-2, 2000, City, State.
Copyright 2000 ACM 1-58113-000-0/00/0000...\$5.00.

products of numerous genes are known to be involved in transcriptional regulation, and such regulation is known to be altered in the progression of certain cancers, the existence of differential correlation is well-supported biologically.

The organization of this article is as follows. Section 2 contains a general description of our proposed differential correlation methodology, and Section 3 contains the results of applying the method to two lung cancer gene expression datasets. In Section 4 we discuss some limitations and possible future directions.

2. DIFFERENTIAL CORRELATION

Complex clinical assessments such as survival or tumor stage generally do not exhibit clear associations with gene expression. In many cases, the number of genes with significant mean difference between two groups of samples is comparable to the number of significant differences found under randomization. This may be due to low statistical power in the study design, but more fundamentally it may be due to a small or vanishing number of genes whose mean expression level varies directly with the levels of the response variable.

Nevertheless, there may be marked effects on expression covariation related to the clinical outcome. Here we investigate the most simple such effect -- a change in the correlation between two genes as the outcome varies. In the case of binary outcomes, this suggests identifying pairs of genes whose correlation differs significantly between the two levels of the response variable.

Ideally, one can propose a mechanistic explanation for any particular instance of differential correlation. For example, a pair of genes whose expression is more tightly correlated for the less severe state of the outcome variable compared to the more severe state may reflect a decoupling of expression associated with disease progression, perhaps resulting from loss of a common regulator. A pair of genes whose co-expression increases with disease progression may reflect genes acting in concert to produce tumor phenotypes such as vascularization or rapid growth.

To identify genes exhibiting differential correlation, for each pair of genes i, j we calculate the robust correlation coefficient (biweight midcorrelation, [1]) between the expression levels of the two genes within each group. Suppose this yields correlation coefficients ρ_1 and ρ_2 . The difference $\Delta_{ij} = \rho_1 - \rho_2$ measures the increase or decrease in correlation between the two groups. We select genes where $\Delta_{ij} >$

0.6 for both datasets, or where $\Delta_{ij} < -0.6$ for both datasets. The statistic could also be constructed using standard Pearson correlations, but we found that with relatively small sample sizes, large shifts in Pearson correlation were often due to a single outlying sample.

Randomization is used to assess whether all candidate gene pairs exhibiting differential correlation can be explained by random variation. Specifically, the outcome variable levels are uniformly permuted across the samples, and the Δ_{ij} values are recomputed for each pair of genes in the permuted data. If the number of Δ_{ij} values in the actual data exceeding a given threshold (we use 0.6) is greater than the 95th or 99th percentile under randomization, it is highly likely that at least some of the genes pairs exhibiting differential correlation are genuine.

3. ANALYSIS OF THE LUNG TUMORS

3.1 Data Integration

We analyzed data collected using two Affymetrix microarrays. The University of Michigan data (originally reported in [2]) were obtained using the full length (HuFL) array that has 7,129 probe sets. The Harvard data (originally reported in [3]) were obtained using the U95A array that has 12,625 probesets. Our analysis focuses on the adenocarcinoma samples, of which there are 79 in the Michigan dataset and 84 in the Harvard dataset (after averaging duplicates and removing samples with low tumor cellularity). For both datasets, we used the trimmed PM-MM difference as the numerical summary for each probeset. A detailed discussion of the data processing methods can be found online: <http://dot.ped.med.umich.edu:2000/pub/Ovary/index.html>.

We mapped each probeset to a Unigene accession number using the array annotation files available from the Affymetrix web site. There were 5,141 distinct Unigene accession numbers that mapped to at least one probeset on both arrays. The majority of the Unigene numbers mapped to a single probeset on each array, but the remainder mapped to as many as 8 probesets.

To construct a summary for each Unigene accession number, we averaged the probeset summaries within each sample across the probesets that map to a common Unigene accession number. These averages (henceforth referred to as gene expression levels) were then truncated at zero, \log_2 transformed, and quantile normalized (see above reference to our data processing methods). The gene expression levels exhibited reproducible aggregate characteristics between the two datasets -- within-dataset means had correlation 0.52 across genes, and within-dataset standard deviations had correlation 0.56 across genes.

The most variable genes were considered for subsequent analysis. We selected the 1,102 genes having standard deviation greater than 0.5 in both data sets. This is a rather high standard deviation threshold, leaving only highly variable genes. For purposes of illustrating our methodology, we select genes according to this strict rule, but a more complete biological investigation might use a lower threshold.

Clinical outcome variables that were measured in similar ways in the two studies and that could easily be dichotomized were selected for analysis. These variables are survival (24 months survival vs. death before 24 months, omitting censored cases), stage (I vs. III), grade (well and moderate vs. poor), smoking status (less than 10 pack years vs. 10 or more pack years), and K-

Ras mutation status (wild type vs. mutant). Table 1 contains the number of samples at each level, for each variable, in the two data sets.

3.2 Baseline Analysis

We began by carrying out a baseline analysis using standard methods. For each clinical response variable, the samples from each dataset were stratified into two groups, which were subsequently compared at each gene using two sample t-tests and fold changes. For each dataset, three sets of genes were identified: (i) genes having a t-test p-value smaller than 0.05, (ii) genes having a two-fold or greater change in mean expression, and (iii) genes having a 1.5-fold or greater change in mean expression. Next the genes satisfying (i) for both datasets were selected, and from these only the genes with consistent direction of expression change in the two datasets were retained. Similarly, genes satisfying (ii) or (iii) and having consistent direction of expression change in the two datasets were considered. The numbers of such genes for each outcome are given in Table 2. Next to each observed number are the 95th and 99th percentiles under randomization, estimated from 300 randomizations.

Table 1. Sample sizes

Outcome	Level	U. Mich.	Harvard
Early Death	≤ 24 months	17	30
	> 24 months	60	53
Stage	I	60	62
	III	19	8
Grade	Well/Moderate	58	29
	Poor	20	14
Smoking	< 10 pack years	14	12
	≥ 10 pack years	63	72
K-Ras	Wild type	42	39
	Mutant	37	24

Table 2. Results of the statistical analysis

Outcome	t-test	2-fold	1.5 fold	Diff. Corr.
Early death	13(5,16)	0(0,1)	2(5,16)	62(115,165)
Stage	10(6,14)	0(1,3)	9(6,14)	1444(1328,1361)
Grade	75(6,12)	2(0,1)	22(6,12)	920(641,889)
Smoking	19(5,8)	16(5,8)	16(5,8)	98(82,100)
K-Ras	21(5,12)	0(0,0)	6(5,12)	210(190,255)

The results of the baseline analysis (Table 2, columns 2-4) indicate that a small number of genes are differentially expressed for each outcome, except for grade, which produces a moderate level of differential expression. An even smaller number of genes exhibit differences that are large in magnitude. Nevertheless, based on the randomization analysis, the t-test results are statistically significant for all 5 outcomes, and it is unlikely that more than half of the genes that are identified are false positives.

Many of the genes identified in the baseline analysis as being associated with the clinical outcomes do not have known biological functions that are easy to relate to the biological nature

of the outcome. However several genes associated with proliferation exhibit significant association with tumor grade. PCNA, Cyclin B1, TOPIIA, and BOP1 are upregulated in poorly differentiated tumors, reflecting the likely faster growth rate of poorly differentiated cancer cells.

3.3 Results of the differential correlation analysis

3.3.1 Randomization analysis and global significance

We identified pairs of genes with differential correlation greater than 0.6 in both datasets, or smaller than -0.6 in both datasets. These pairs were identified from among the $\approx 6 \times 10^5$ distinct pairs that can be formed from the 1,102 genes meeting the variability conditions. The fourth column of Table 3 shows the results of this analysis. Compared to the randomized results, stage, grade, smoking, and K-Ras show an excess of differentially correlated pairs, while early death does not.

The biological significance of this finding is that it suggests that some of the clinical outcomes have a much broader relationship with gene expression than is indicated by consideration of mean differences. For example, while only 10 genes exhibit strong evidence of differential mean expression with stage, the 1,444 pairs showing significant differential correlation with stage include 858 distinct genes (60% of all genes considered). While the randomization analysis suggests that many of the 1,444 pairs may be false positives, even if only 100 pairs are truly differentially correlated (taking a very conservative view of the randomization results), these pairs are likely to contain far more than 10 distinct genes.

3.3.2 An example – negative interaction of BENE and Hs.143288 is specific to poorly differentiated tumors

Focusing now on a specific example, Figure 1 shows a pair of genes that are differentially correlated with respect to grade in both data sets. Figure 1a shows that the genes BENE and Hs.143288 have little association in well or moderately-differentiated samples (perhaps there is a positive trend in the Harvard data, but this is quite weak). On the other hand, there is a strong negative trend between the two genes in the poorly differentiated samples. For both data sets, high levels of Hs.143288 expression are associated with low levels of BENE expression. Adding to the potential biological significance of this relationship is that both genes vary widely across the tumors in both datasets – BENE undergoes five doublings between the least and greatest expression, and Hs.143288 undergoes more than two doublings.

This pair of genes also illustrates the value of looking at differential correlation in addition to inspecting genes that are differentially expressed in mean expression. Neither BENE nor Hs.143288 is significantly differentially expressed in mean between the two classes of samples, thus they would not be considered to be relevant to grade based on usual measure of differential association such as t-tests or fold-change statistics.

The BENE gene codes for a membrane-bound protein with unknown molecular function, and the Hs.143288 gene codes for a hypothetical protein with sequence similarity to mouse, rat, and *C.elegans* collagen. With such little biological information, it is difficult to propose a mechanistic explanation for this relationship. One hypothetical explanation might be that advanced tumors

segregate into two distinct clusters – one exhibiting high BENE expression and low Hs.143288 expression, and the other exhibiting high Hs.143288 expression and low BENE expression. This would suggest a permanent silencing of either BENE or Hs.143288 expression in all advanced tumors (but not a silencing of both genes in any one tumor). An alternative hypothetical explanation would be that both genes are transiently expressed in advanced tumor cells, but the expression is coordinated so that the two genes are never expressed simultaneously. This coordination may be associated with phenotypes such as proliferation, invasiveness, or vascularization that are more prominent in advanced tumors.

3.3.3 Genes participating in widespread differential correlation

Although many genes are differentially correlated with at least one other gene, we found that a few genes dominate all others, in that they participate in widespread differential correlation with many other genes. These genes may potentially play a more global role in reporting, or causing, widespread alterations in the interaction of gene expression levels. At a more practical level, they may serve as biomarkers for detecting dramatic shifts in correlation structure associated with a clinical endpoint.

For example, using tumor stage as the outcome, five genes engage in differential correlation with at least 20 other genes. Two of these genes are implicated in other epithelial adenocarcinomas, specifically, disease of ovary (WFDC2/HE4; Hs.2719) and colon (galectin-4; Hs.5302). Among the remaining three genes are a gene associated with female fertility (NRIP1; Hs.155017), a widely expressed enzyme (MTHFD2; Hs.154672), and a gene of unknown function (Hs.380833).

Although galectin-4 is generally reported as being expressed only in colon, many of the lung tumors exhibit moderate expression of this transcript. While cross-hybridization of a different transcript is a likely explanation for this, given that we identified galectin-4 based on its differential correlation with respect to stage, it is notable that galectin-4 has been specifically noted as being associated with stage in colon cancer [4].

The WFDC2/HE4 gene has been reported to be a biomarker for ovarian cancer [5]. Expression in other tissues has been observed as well. Notably, high expression of WFDC2/HE4 is primarily found in malignant ovarian tumors, and it is expressed at much lower levels in non-malignant tumors. Since degree of malignancy is roughly associated with tumor stage, the specific expression of WFDC2/HE4 in malignant ovarian tumors may possibly be related to the fact that we found WFDC2/HE4 to be differentially correlated with stage in lung tumors.

Figure 1a: Moderately and well-differentiated samples show no association between Hs.143288 and BENE expression

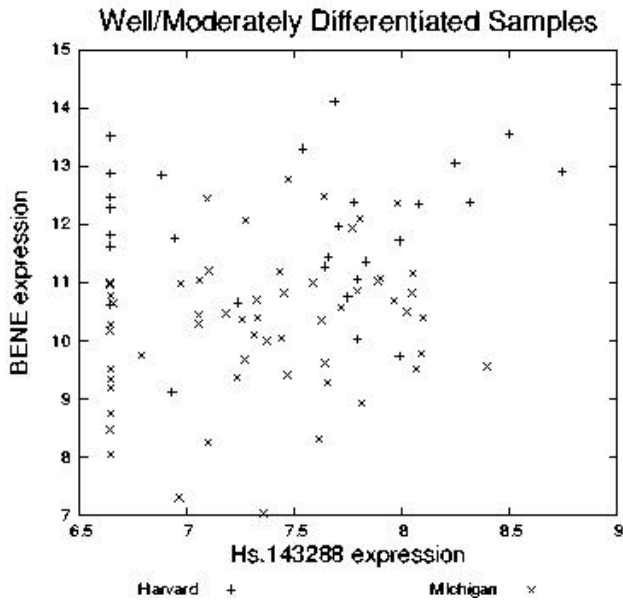
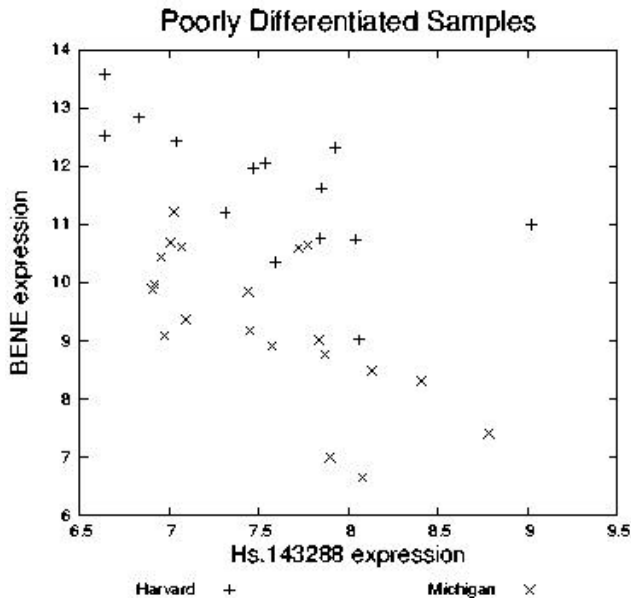


Figure 1b: Poorly differentiated samples show negative association between Hs.143288 and BENE expression



4.DISCUSSION

We have proposed a statistical analysis strategy for gene expression data that aims to uncover subtle effects that may be attributable to clinically-important factors. Our investigation of two lung-cancer data sets [2,3] suggests that widespread statistically-significant effects can be discovered based on differential correlation that would be missed using an analysis strategy that focuses exclusively on group means. In particular, numerous genes may be associated with stage, grade, smoking status, and K-Ras mutation status that would not be detected using conventional measures of differential expression.

Differential correlation may be considered either in the context of a single data set, or, as we have done here, when looking at multiple data sets. In the later case, we identified pairs of genes that are consistently differentially correlated in all the available data. As with any statistical analysis, genes identified by differential correlation will be less likely to be attributable to experimental bias or noise if multiple data sets are used. In addition, there may be a specific reason that differential correlation may be especially suitable for use with multiple data sets. Since differential correlation depends only on relationships between the expression levels of two genes, and not on absolute levels of gene expression, it is likely to be less sensitive to certain types of artifacts that may produce systematic differences between results obtained in different laboratories. Any such artifact that results in a monotone relationship between the measurements in two datasets will generally not have as strong an influence on differential correlation as it would exert on fold changes or t-test statistics.

We note that we have not had success in one of our primary goals, which was to enlarge the set of genes associated with survival using the differential correlation technique. Our baseline analysis suggests that only a small number of genes are associated with early death in both data sets, and none of these genes has a magnitude change greater than 1.5. No pair of genes shows significant differential correlation associated with early death.

One practical drawback of our method is that the response variable must be dichotomous, so that it can be used to stratify the samples into two classes. In some cases, such as when the response variable is survival time, this requires coarsening the resolution of the measurement, perhaps leading to a loss of relevant information. We note that a similar, but more general methodology called *Liquid Association* [6] has recently been developed that has similar goals as our method, but that is formulated so that a continuous rather than a qualitative characteristic controls the changes in correlation.

5.ACKNOWLEDGMENTS

Our thanks to Rork Kuick for preparing the probeset-level data summaries, and for reviewing the manuscript.

6.REFERENCES

- [1] Wilcox, Rand (1997). Introduction to Robust Estimation and Hypothesis Testing. Academic Press, New York.
- [2] Beer, D, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 9 (816), 2002.
- [3] Bhattacharjee, A, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*. 98 (24), 13790-13795, November 2001.
- [4] Nagy N, Legendre H, Engels O, Andre S, Kaltner H, Wasano K, Zick Y, Pector JC, Decaestecker C, Gabius HJ, Salmon I, Kiss R. Refined prognostic evaluation in colon carcinoma using immunohistochemical galectin fingerprinting. *Cancer*. 2003 Apr 15;97(8):1849-58.
- [5] Hellstrom I, Raycraft J, Hayden-Ledbetter M, Ledbetter JA, Schummer M, McIntosh M, Drescher C, Urban N, Hellstrom KE. The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. *Cancer Res*. 2003 Jul 1;63(13):3695-700.
- [6] Li, K.C. Genome-wide coexpression dynamics: Theory and application. *PNAS* 2002 99 16875-16880.