Funct Integr Genomics (2003) 3:180–188
DOI 10.1007/s10142-003-0087-5

ORIGINAL PAPER

Debashis Ghosh · Terrence R. Barette · Dan Rhodes · Arul M. Chinnaiyan

Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer

Received: 24 September 2002 / Accepted: 31 March 2003 / Published online: 22 July 2003

Abstract With the proliferation of related microarray studies by independent groups, a natural step in the analysis of these gene expression data is to combine the results across these studies. However, this raises a variety of issues in the analysis of such data. In this article, we discuss the statistical issues of combining data from multiple gene expression studies. This leads to more complications than those in standard meta-analyses, including different experimental platforms, duplicate spots and complex data structures. We illustrate these ideas using data from four prostate cancer profiling studies. In addition, we develop a simple approach for assessing differential expression using the LASSO method. A combination of the results and the pathway databases are then used to generate candidate biological pathways for cancer.

Keywords Bioinformatics · Differential expression · Gene expression · LASSO · Multiple comparisons

Introduction

Microarray technology has been used with great success for molecular profiling studies in a variety of scientific areas, such as in cancer experiments. Use of these global high-throughput assays have allowed researchers to discover novel biomarkers and differentially expressed genes. Typically there have been two components in most published microarray studies. At the first stage, the samples are interrogated using microarray technology. Based on the resulting gene expression profiles, candidate genes are selected and validated using other techniques, examples of which include quantitative RT-PCR and western blots. The goal of the second stage is to determine which genes are truly differentially expressed and to exclude the possibility that the first stage of the experiment yielded a false positive. Thus, we are reduced from analyzing genes on a high-throughput basis to a one gene at a time approach.

While finding new biomarkers represents one goal of microarray profiling studies in cancer settings, a more ambitious task would be to find oncogenic pathways. Related work has begun in combining gene expression and sequence data to find regulatory networks in organisms such as *Sacchromyces cerivisiae* (Tavazoie et al. 1999; Bussemaker et al. 2000), but such an approach would be much harder in cancer because of problems such as inaccurate exon prediction and the complex composition of tissue.

Because of the explosion in the use of microarray technology, many research groups are conducting gene expression profiling studies in similar scientific areas. If one group were to find a gene differentially expressed using microarrays, there is a chance that this is simply a false positive. However, if two groups independently find the same gene to be differentially expressed, then the chance of this error is reduced. It becomes smaller as the number of studies in which the gene is reported to be differentially expressed increases. Thus, by combining the results across several microarray experiments, we can gain significant increases in power of detecting differentially expressed genes. This is one of the guiding factors behind the development of public microarray databases such as the Stanford Microarray Database and Gene Expression Omnibus; for more information, their URL locations are http://genome-www5.stanford.edu/MicroArray/SMD/ and http://www.ncbi.nlm.nih.gov/geo/, respectively. Because there is no widely accepted standard format and ontology regarding the public databases, there has also been a movement in the scientific community to

D. Ghosh (✉)
Department of Biostatistics, School of Public Health,
University of Michigan,
1420 Washington Heights, Ann Arbor, MI 48109-2029, USA
e-mail: ghoshd@umich.edu
Tel.: +1-734-6159824
Fax: +1-734-7632215

T. R. Barette · D. Rhodes · A. M. Chinnaiyan
Department of Pathology,
University of Michigan,
Ann Arbor, MI 48109-2029, USA

develop a minimum set of information publicly reported about each microarray experiment (Brazma et al. 2001).

This topic of combining data across studies is known as meta-analysis and has been well-studied in the statistical literature (Normand 1999). However, there is a slew of issues that arises when attempting to perform meta-analysis with microarray data. These include combining gene expression data across diverse experimental platforms (e.g., Affymetrix and cDNA microarray technologies), complicated data structures, multiple comparisons and the presence of duplicate spots. One goal of this article is to describe the relevant statistical issues in attempting to analyze data from multiple microarray studies. Our motivating example involves determining which genes are differentially expressed between locally advanced prostate cancer and benign tissue. In this article, we also develop some new statistical methods for analyzing such data and combine them with existing biological databases to incorporate external knowledge into the analysis. In this paper we describe the data collected, discuss the relevant statistical issues and describe the statistical techniques for assessing differential expression across multiple studies. We propose methods for generating candidate cancer pathways using a combination of statistical and bioinformatic techniques and finally make some concluding remarks.

Materials and methods

Microarray data were collected from four publicly available prostate cancer gene expression datasets that were generated by four independent laboratories (Dhanasekaran et al. 2001; Luo et al. 2001; Magee et al. 2001; Welsh et al. 2001). In all four studies, comparisons were made between the gene expression profiles for clinically localized prostate cancer and benign prostate tissue specimens. One of the goals in these studies was the identification of differentially expressed genes between the two tissue types. Summaries of the experiments are provided in Table 1. There are several points to note about the studies. Affymetrix technology (Lipshutz et al. 1999) was utilized in two studies, while in the Dhanasekaran and Luo studies, spotted cDNA microarrays (Schena 2000) were used. In addition, the number of samples varied widely between the studies. To increase the power, the metastatic and locally advanced prostate cancer cases from the Dhanasekaran, Magee and Welsh studies were combined. The smallest study of prostate cancer was that of Magee et al. (2001), while the largest was that of Dhanasekaran et al. (2001). We now describe some of the relevant issues involved in consideration of these data.

Results

Effect of experimental platform

As previously mentioned, two of the studies involved Affymetrix technology, while the other two utilized cDNA technology. While both of these microarray platforms measure relative mRNA measurements, the actual technologies for doing so are quite different. We now provide a brief description of each type of technology.

The Affymetrix GeneChip technology consists of sets of 25-mer long oligonucleotides, which are called probes. The probesets usually consist of 20 perfect match (PM) oligonucleotides and 20 mismatch (MM) oligonucleotides. Each set of PM oligos ostensibly encodes for a unique region of a gene. The MM oligos are the same as the PM oligos, except for a one-base change in the center of the oligonucleotide. The role of the MM oligos is to serve as a control measurement for the PM oligos by accounting for non-specific hybridization. The mRNA sample is converted to cDNA by a reverse transcription reaction and then prepared for hybridization to the GeneChip. After the hybridization reaction occurs, the material is stained twice and scanned. What is used for analysis is the average difference between the PM and MM measurements.

The spotted cDNA microarray is a glass slide where spots correspond to genetic material that encodes a gene. Two types of mRNA samples are then hybridized to the microarray: the test sample (sample of interest) and the reference sample. Both samples are fluorescently labelled; typically the test sample is labelled with red dye, while a green dye is attached to the reference sample. The point of the reference sample is to serve as an internal control on the chip. After the hybridization occurs on the microarray, the slide is scanned at two intensities which are referred to as the red and green intensities. If a gene is overexpressed in the test sample relative to the reference, its spot will have a large value in the green channel. If the converse holds, then the red channel intensity of the spot will be large. The typical unit of analysis with cDNA microarrays is the ratio of the gene expression measurement in the red channel to that of the green channel.

One major issue is whether or not we can directly combine the raw measurements from the two different technologies. In a recent study performed by Kuo et al.

Table 1 Description of prostate cancer profiling studies

Author	Array[a]	Reported clones	Number of samples		
			Benign prostate	Localized PCA	Metastatic PCA
Dhanasekaran et al.	cDNA	9,984	19	14	20
Luo et al.	cDNA	6,500	9	16	0
Magee et al.	Oligo	7,068	4	8	3
Welsh et al.	Oligo	8,900	9	23	1

[a] *cDNA* means that spotted microarray technology was utilized; *Oligo* means that Affymetrix GeneChip technology was utilized

(2002), they compared Affymetrix and spotted cDNA gene expression measurements from a large-scale study involving 60 cell lines from the National Cancer Institute. They found that the correlation between the actual gene measurements from the two technologies was fairly low. They concluded that "data from spotted cDNA microarrays could not be directly combined with data from synthesized oligonucleotide arrays.'' Furthermore, they concluded that it was unlikely that the two types of data could be transformed or normalized into a common standardized index. Thus, we avoid directly combining the original gene expression data from the two types of technologies. Our approach is to instead use as our "data" the *t*-statistic comparing prostate cancer to benign tissue for each gene in each of the studies. We did this because we felt that this statistic might be more robust to the choice of technology than the actual raw data.

We explored the relationship between the *t*-statistics between the different studies based on genes that were common to each pair of studies. This is summarized in Table 2. Here, we find that the correlation between the studies is between 0.25 and 0.46. What is interesting is that the agreement between the experiments using the same technology does not appear to be better than the correlation between experiments of differing technologies. While the Magee study, utilizing Affymetrix arrays, has maximum correlation with the Welsh study, also an Affymetrix study, the Dhanasekaran study, in which spotted cDNA microarrays were used, has maximum correlation with the Welsh study. This seems to confirm the assumption that the *t*-statistic is less dependent on the technology used. We also reran the same analysis with the raw data; these data are presented in Table 3. We find that the data exhibit poorer correlation at the raw measurement level, with the exception of the Welsh and Magee studies.

Table 2 Correlations between *t*-statistics for the four studies. The Pearson correlation coefficient was utilized; correlation was computed over genes common to each pair of studies

Author	Dhanasekaran	Luo	Magee	Welsh
Dhanasekaran	1	0.376	0.283	0.447
Luo	0.376	1	0.250	0.459
Magee	0.283	0.25	1	0.318
Welsh	0.447	0.459	0.318	1

Table 3 Correlations between raw data of the four studies. The Pearson correlation coefficient was utilized; correlation was computed over genes common to each pair of studies, averaged across all samples

Author	Dhanasekaran	Luo	Magee	Welsh
Dhanasekaran	1	0.068	–0.078	0.003
Luo	0.068	1	0.068	0.047
Magee	–0.078	0.068	1	0.79
Welsh	0.003	0.047	0.79	1

Table 4 Number of common clones in the four studies

Author	Dhanasekaran	Luo	Magee	Welsh
Dhanasekaran	9,984	5,106	1,919	2,906
Luo	5,106	6,500	1,560	2,132
Magee	1,919	1,560	3,350	2,221
Welsh	2,906	2,132	2,221	6,812

Data structures

Ideally, the microarray platforms for each of the four studies would contain the same set of genes, but unfortunately, this is not the case. In Table 4, we have tabulated the number of clones common to each pair of studies. Some clarification needs to be mentioned regarding the table. The reason the number on the diagonals do not correspond exactly to the totals reported in Table 1 is because Table 3 represents the clones that were actually publicly available. Thus, there is an issue of gene selection from the Welsh and Magee studies which complicates the analysis. However, we will not pursue this issue further here.

Based on Table 4, we find that approximately 80% of the clones in the Luo study are also represented in the Dhanasekaran study. The Magee study has the most clones in common with the Welsh study. However, what we find is a highly non-nested pattern of gene expression data across studies. Each study has clones in common with the other three studies, but there are genes that are unique to each study. We are thus faced with a complicated missing data mechanism.

We will make the assumption that the data are missing at random (MAR; Little and Rubin 1987) based on the nature of how the microarrays are generated. With cDNA microarrays, the researchers usually obtain a set of sequence-verified cDNAs from a company such as Research Genetics (http://www.resgen.com/). Since the scientists typically place all available cDNAs on the slide, there is no a priori reason to believe that the spots that are used depend on their underlying gene expression level. The Affymetrix arrays are typically standard arrays (e.g., Hu6800 arrays has approximately 7,000 full-length human mRNA transcripts available), so again assuming MAR appears to be reasonable here as well.

A major issue involves matching the spots from the cDNA microarray to those representing the probesets on the oligonucleotide microarrays. The identifiers for the spots on the cDNA microarray are GenBank accession numbers, while on the oligonucleotide microarrays, the probesets have their own unique numbering system. To match the cDNA microarray spot to the corresponding probeset involves matching the gene sequence for the GenBank accession number for the spot to that of the probeset. To accomplish such a task, we use BLAST (http://www3.ncbi.nlm.nih.gov/BLAST). Thus, we find that some bioinformatic manipulations are necessary in order to match the identifiers from the two types of technologies.

Duplicate spots

Based on the BLAST search previously described, this leads to the presence of duplicate spots in each of the studies. We define a duplicate spot on a microarray to be genetic material from at least two different locations on the slide that correspond to the same UniGene cluster identification number (e.g., Hs. 3196). In Table 5, we summarize the number of duplicate spots per study. Because these duplicate spots have the same UniGene cluster identification number, and each UniGene cluster supposedly corresponds to a unique gene, we would expect the behavior of duplicates to be that of biological replicates. However if we look at Fig. 1, we find that there is substantial variability in the duplicate spots for the four studies. For certain duplicate spots, the differential expression ranges from negative to positive. There are several potential reasons to explain the variation. It might be due to experiment-specific artifacts. One biological reason is that the duplicate spots represent alternatively spliced forms of transcript mRNA. Another possible reason is that the classification for the spot is incorrect due to errors in the UniGene database. Some authors have estimated that the misclassification rate of sequences in this database might be on the order of 35–40% (Irizarry et al. 2000). The implication of these facts is that we need to take into account the variability present because of the duplicate spot in our analyses. For the analyses that we perform later, we treat the duplicate spots as biological replicates.

Table 5 Number of duplicate spots for the four studies

Dhanasekaran	Luo	Magee	Welsh
1,119	293	121	757

Multiple comparisons

This issue also exists for most microarray experiments because of the large number of genes that are tested for differential expression. Because of the complicated data structure used here, the number of comparisons for multiple studies is much larger than that for any one study. As a result, control of the traditional familywise type I error level does not seem appropriate here. A popular quantity to control in the recent literature has been the false discovery rate (FDR; Benjamini and Hochberg 1995). In the next section, we define these quantities in the context of multiple testing. We then present some methods for assessing differential expression in genes, primarily focusing on the false discovery rate.

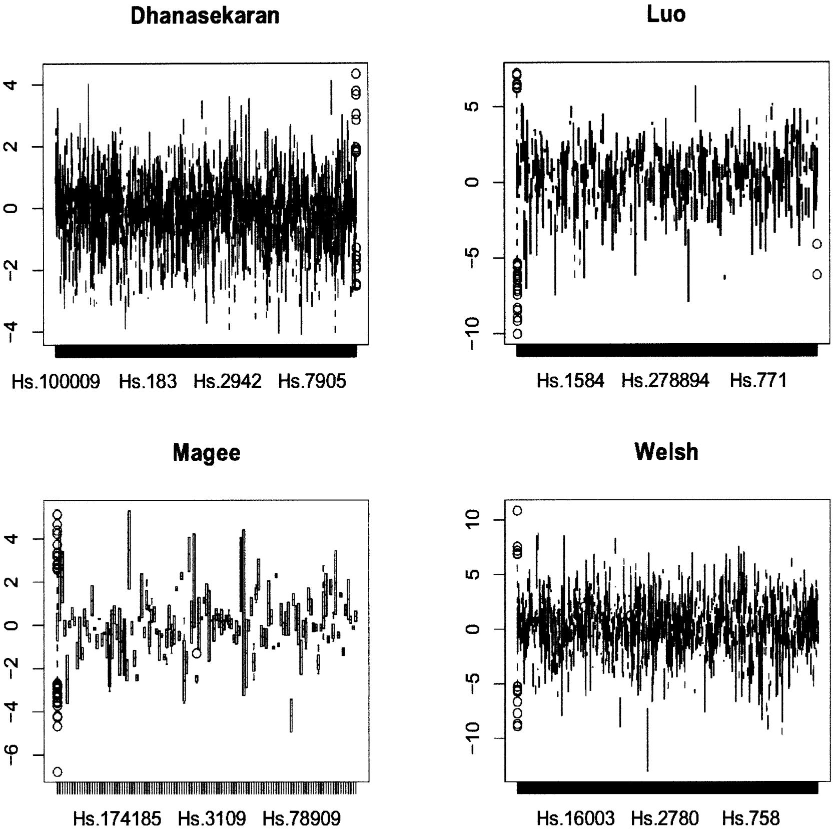

Fig. 1 Plots of t-statistics for duplicate spots for each of four prostate cancer profiling studies

Definitions of FWER, FDR, pFDR and q-value

Suppose we are interested in assessing differential expression between two conditions for a set of m genes. Of these m genes, suppose that m_0 truly show no differential expression. Because we are testing several thousand hypotheses, controlling the type I error rate for each gene fails to control the type I error rate for the entire experiment. To guard against making too many type I errors, the familywise error rate (FWER) has typically been controlled. A review of methods for controlling this quantity can be found in Shaffer (1995). To better understand the quantities being considered in this paper, we consider a 2×2 contingency table (Table 6). Based on Table 6, the FWER is defined as $P(V \geq 1)$, which is the probability that the number of false positives is greater than 1. The definition of FDR as put forward by Benjamini and Hochberg (1995) is

$$\mathrm{FDR} \equiv E\left[\frac{V}{Q}|Q > 0\right] P(Q > 0).$$

The conditioning on the event [Q >0] is needed because the fraction V/Q is not well-defined when Q =0. Storey (2002) points out the problems with controlling this quantity and suggests use of the positive false discovery rate (pFDR), defined as

$$\mathrm{pFDR} \equiv E\left[\frac{V}{Q}|Q > 0\right].$$

Conditional on rejecting at least one hypothesis, the pFDR is defined to be the fraction of rejected hypotheses that are in truth null hypotheses. This quantity is analogous to type I error rates in single hypothesis testing problems.

In classical hypothesis testing problems, a strength of evidence measure is provided by the p-value. In the "Significance Analysis of Microarrays'' (SAM) software (Tusher et al. 2001), an analogous measure, known as the q-value, has been developed. We define it later in the paper.

Statistical methods for differential expression

Based on the considerations made in the previous section, we now propose some procedures for assessing differential expression across studies.

Let Y_{gjk} denote the t-statistic comparing gene expression between prostate cancer and healthy tissue for the jth duplicate spot for the gth gene in the kth study, g =1,..., G, j =1,...,m_k, k =1,...,n_g. In the prostate cancer example presented here, the maximum value for n_g is 4.

In our first algorithm, we will use the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) for assessing differential expression of genes. If μ_g denotes the mean normalized difference in gene expression between prostate cancer and healthy tissue for the gth gene, then the relevant optimization problem is to minimize

$$\sum_{k=1}^{n_g} \sum_{j=1}^{m_k} \left(Y_{gjk} - \mu_g\right)^2$$

subject to the constraint that $\sum_{g=1}^{G} |\mu_g| \leq \lambda$. One of the advantages of the LASSO is that some of the estimated values of μ_g will be exactly zero. Biologically, the genes with non-zero values of μ_g will constitute a list of candidate genes. The parameter λ plays the role of a shrinkage parameter here. The estimates of μ_g will have the form $\widehat{\mu}_g = \mathrm{sign}\left(t_g\right)\left(|t_g| - c\right)I\left(t_g \geq c\right)$ where t_g is the average of the t-statistics for the gth gene across duplicates and studies, and c is a constant involving λ and the sorted values of t_g.

Table 6 Outcomes of tests of hypotheses for m genes (*DE* differential expression)

	Conclude no DE	Conclude DE	Total
No DE	U	V	m_0
DE	T	S	m_1
	W	Q	m

Table 7 Effect of λ on number of estimated non-zero genes $\left(\widehat{k}\right)$

λ	$\left(\widehat{k}\right)$
1	435
400	350
1,000	242
2,000	96
2,300	62
2,400	51
2,800	13

To study the effect of λ on the number of genes with non-zero coefficients, we fit the procedure using several values of λ. We did not pursue cross-validation because the dependence structure of the genes is unknown, and estimation of the cross-validation under dependency is not straightforward. The numbers are summarized in Table 7.

The UniGene accession numbers for the top 25 genes are listed in Table 8. Based on the table, we find that most of the top scoring genes are expressed sequence tags (ESTs) that code for genes of unknown functions. Many of these ESTs only appeared in one study, so a more conservative criteria would be to take the top ranked genes based on μ_g that appeared in at least two studies (i.e. n_g >2). These are provided in Table 9. What we find on this list are genes that are involved in basic cell signaling and protein-protein interaction pathways. Presumably, the differential expression of these genes in tumor relative to normal tissue suggests that these normal cell functions have been altered, which has led to tumorigenesis.

Table 8 Top 25 genes based on $\widehat{\mu}_g$

Acccession number	Gene name
Hs.75432	IMPDH2 IMP (inosine monophosphate) dehydrogenase 2
Hs.189869	ESTs
Hs.192052	ESTs
Hs.171939	*Homo sapiens* mRNA; cDNA DKFZp761L1121
Hs.93485	*Homo sapiens* mRNA; cDNA DKFZp761D191
Hs.289104	Alu-binding protein with zinc finger domain
Hs.23578	ESTs
Hs.106671	Cleft lip and palate associated transmembrane protein 1
Hs.268575	ESTs
Hs.57787	ESTs
Hs.6566	Thyroid hormone receptor interactor 13
Hs.45033	Proline-rich 4 (lacrimal)
Hs.8832&	ESTs
Hs.267182	T-box 3 (ulnar mammary syndrome)
Hs.16359	ESTs
Hs.301206	ESTs
Hs.122843	CASP8-associated protein 2
Hs.189713	ESTs
Hs.314319	*Homo sapiens* hsr1 mRNA (partial)
Hs.22015	*Homo sapiens*, similar to RIKEN cDNA 1810054O13 gene
Hs.24176	Heart alpha-kinase
Hs.192966	KIAA0265 protein
Hs.326816	ESTs
Hs.221396	ESTs
Hs.334338	Hypothetical protein MGC12837

Table 9 Top 25 genes based on $\widehat{\mu}_g$ with $n_g > 2$

Acccession number	Gene name
Hs.3196	Surfeit 1
Hs.153880	Polymerase (RNA) mitochondrial (DNA directed)
Hs.86859	Growth factor receptor-bound protein 7
Hs.194329	Hypothetical protein FLJ21174
Hs.177543	EST, clone IMAGE:117491, 3′ end
Hs.1211	Acid phosphatase 5, tartrate resistant
Hs.119206	Insulin-like growth-factor-binding protein 7
Hs.79380	Periodic tryptophan protein homolog (yeast)
Hs.75216	Protein tyrosine phosphatase, receptor type, F
Hs.76989	KIAA0097 gene product
Hs.169378	Multiple PDZ domain protein
Hs.18747	POP7 (processing of precursor, *Saccharomyces cerevisiae*) homolog
Hs.153639	Hypothetical SBBI03 protein
Hs.7314	KIAA0614 protein
Hs.238126	CGI-49 protein
Hs.171814	Parathymosin
Hs.169900	Poly(A) binding protein, cytoplasmic 4 (inducible form)
Hs.26468	Amyloid beta (A4) precursor protein-binding, family A, member 2
Hs.108660	ATP-binding cassette, sub-family C (CFTR/MRP), member 5
Hs.194772	Oligodendrocyte myelin glycoprotein
Hs.29189	ATPase, Class VI, type 11A
Hs.105584	Ribosomal protein S6 kinase, 90 kDa, polypeptide 4
Hs.268530	G protein pathway suppressor 1
Hs.88474	Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)
Hs.77889	Friedreich ataxia region gene X123

We now present methods for assessing differential expression based on direct estimation of the FDR. In our first algorithm, we stratify the analysis of gene expression data by the study. For each study, we fit the model

$$E[Y_{ijk}] = \beta_{0ik} + \beta_{1ik}X_j; \qquad (1)$$

where X_j is a covariate for tissue type of the jth sample. For example, if there are two tissue types to be compared (normal tissue coded 0 and cancerous tissue coded 1), then the interpretation of β_{1ik} is as the difference in average gene expression between cancer and normal tissue for the ith gene in the kth study. Consequently, fitting Eq. 1 is equivalent to computing t-statistics for each gene within each study. Model 1 (Eq. 1) can be fit using ordinary least squares (OLS), yielding a set of statistics T_{1k},..., T_{mk}, where T_{ik} is the least squares estimator of β_{1ik} divided by its standard error. We then compute a global statistic of differential expression for each gene

$$T_{i.} = n^{-1}\sum_{k=1}^{K} n_k T_{ik}, \qquad (2)$$

where $n = \sum_{k=1}^{K} n_k$. Suppose we have defined the rejection regions R so that the test statistic has type I error α, i.e. $P(T_i \in R|H = 0) = \alpha$. We also define another region A that will be useful in estimation of π_0. Intuitively, A is the region where we expect to capture most of the genes for which we have no true differential expression. The next step involves permuting the tissue labels (the Xj's) within each study for B permutations and refitting model 1 (Eq. 1) to the permuted dataset. This yields a set of simulated null statistics $T_{1k}^{0b}, ..., T_{mk}^{0b}$. Analogous to Eq. 2, we can compute the global statistic $T_{i.}^{0b}$ for each permuted dataset (b =1,…,B; i =1,,…,m). We can then estimate the pFDR as

$$\widehat{pFDR}(R) = \frac{\widehat{\pi}_0 (mB)^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m} P\left(T_{i.}^{0b} \in R\right)}{m^{-1}\sum_{i=1}^{m} I(T_{i.} \in R)},$$

where

$$\widehat{\pi}_0 = \left\{m^{-1}\sum_{i=1}^{m} I(T_{i.} \in A)\right\} / \left\{(mB)^{-1}\sum_{b=1}^{B}\sum_{i=1}^{m} I\left(T_{i.}^{0b} \in A\right)\right\}.$$

Note that pFDR depends heavily on the set R but we have suppressed the dependence in the notation.

With this approach, we have implicitly assumed that the effect of tissue type is the same across all the studies. A more general model we could fit is

$$E[Y_{ijk}] = \gamma_{0ik} + \gamma_{1ik}X_j + \gamma_{2ik}Z_k + \gamma_{3ik}X_jZ_k. \qquad (3)$$

where X_j has the same definition as before and Z_k is a study indicator. Note that if we set $\gamma_{2ik} = \gamma_{3ik} = 0$, then we are reduced to model 1 (Eq. 1). Model 3 (Eq. 3) can be fit for each gene using OLS as well. Now, we calculate a test statistic using the likelihood ratio test for testing H_0: $\gamma_{ik} = \gamma_{3ik} = 0$ for each gene, and this yields a set of statistics

$\widetilde{T}_1, ..., \widetilde{T}_m$ for assessing differential expression for the m genes. We repeat the permutation procedure as before, shuffling the tissue labels within the study and repeating the model fitting procedure. This yields simulated null statistics $\widetilde{T}_1^{0b}, ..., \widetilde{T}_m^{0b}$. Using the definitions of R and R'' from the previous paragraph, we estimate the pFDR as

$$pF\widetilde{D}R(R) = \frac{\widehat{\pi}_0 (mB)^{-1} \sum_{b=1}^{B} \sum_{i=1}^{m} P\left(\widetilde{T}_{i.}^{0b} \in R\right)}{m^{-1} \sum_{i=1}^{m} I\left(\widetilde{T}_{i.} \in R\right)},$$

where

$$\widehat{\pi}_0 = \left\{ m^{-1} \sum_{i=1}^{m} I\left(\widetilde{T}_{i.} \in R\right) \right\} /$$

$$\left\{ (mB)^{-1} \sum_{b=1}^{B} \sum_{i=1}^{m} I\left(\widetilde{T}_{i.}^{0b} \in A\right) \right\}.$$

These algorithms are summarized in Boxes 1 and 2.

Box 1. Proposed algorithm 1 for estimating false discovery rate (FDR) and positive FDR (pFDR)

(a) For each study and for each gene, fit a *t*-statistic comparing average gene expression in cancerous tissue with that in healthy tissue; this yields $(T_{1k}, \ldots, T_{mk})$.

(b) Calculate $T_i = n^{-1} \sum_{k=1}^{K} n_k T_{ik},\ i = 1, ..., m.$

(c) Permute the tissue labels within the study, and calculate permuted *t*-statistics $T_{1k}^{0b}, ..., T_{mk}^{0b}$.

(d) Calculate for *b*th permutation,

$$T_{i.}^{0b} = n^{-1} \sum_{k=1}^{K} n_k T_{ik}^{0b}.$$

(e) Estimate $\hat{\pi}_0$ as

$$\hat{\pi}_0 = \frac{m^{-1} \sum\limits_{i=1}^{m} I(T_{i.} \in A)}{(mB)^{-1} \sum\limits_{b=1}^{B} \sum\limits_{i=1}^{m} I(T_{i.}^{0b} \in A)}.$$

(f) Estimate pFDR as

$$p\widehat{\mathrm{FDR}}(R) = \frac{\widehat{\pi}_0 (mB)^{-1} \sum_{b=1}^{B} \sum_{i=1}^{m} P\left(T_{i.}^{0b} \in R\right)}{\widehat{P}(V > 0) \max\left[m^{-1} \sum_{i=1}^{m} I(T_{i.} \in R)\right]},$$

where

$$\widehat{P}(Vgt; 0) = B^{-1} \sum_{b=1}^{B} I\left(\widehat{R}^b > 0\right)$$

and $\widehat{R}^b$= number of genes called significant for b th permutation.

(g) Estimate FDR as

$$\widehat{\mathrm{FDR}}(R) = \frac{\widehat{\pi}_0 (mB)^{-1} \sum_{b=1}^{B} \sum_{i=1}^{m} P\left(T_{i.}^{0b} \in R\right)}{\max\left[m^{-1} \sum_{i=1}^{m} I(T_{i.} \in R), 1\right]},$$

Box 2. Proposed algorithm 2 for estimating false discovery rate (FDR) and positive FDR (pFDR)

(a) For each study and for each gene, fit the following model using least squares:

$$E\left[Y_{ijk}\right] = \gamma_{0ik} + \gamma_{1ik} X_j + \gamma_{2ik} Z_k + \gamma_{3ik} X_j Z_k.$$

(b) Calculate a likelihood ratio test statistic of $H_0{:}\gamma_{1ik} = \gamma_{3ik} = 0$ for each gene, yielding statistics $\widetilde{T}_i,\ i = 1, ..., m.$

(c) Permute the tissue labels within the study, repeat steps (a) and (b) to get $\widetilde{T}_1^{0b}, ..., T_m^{0b}$.

(d) Estimate $\hat{\pi}_0$ as

$$\hat{\pi}_0 = \frac{m^{-1} \sum\limits_{i=1}^{m} I\left(\hat{T}_i \in A\right)}{(mB)^{-1} \sum\limits_{b=1}^{B} \sum\limits_{i=1}^{m} I\left(\tilde{T}_i^{0b} \in A\right)}.$$

(e) Estimate pFDR as

$$p\mathrm{F\widetilde{D}R}(R) = \frac{\hat{\pi}_0 (mB)^{-1} \sum\limits_{b=1}^{B} \sum\limits_{i=1}^{m} P\left(\tilde{T}_i^{0b} \in R\right)}{\tilde{T}(V > 0) \max\left[m^{-1} \sum\limits_{i=1}^{m} I\left(\tilde{T}_{i.} \in R\right), 1\right]},$$

where $\widetilde{P}(V > 0) = B^{-1} \sum_{b=1}^{B} I\left(\widetilde{R}^b g > 0\right)$ and $\widetilde{R}^b$= number of genes called significant for *b*th permutation.

(f) Estimate FDR as

$$\mathrm{F\widetilde{D}R}(R) = \frac{\widehat{\pi}_0 (mB)^{-1} \sum_{b=1}^{B} \sum_{i=1}^{m} P\left(\widetilde{T}_{i.}^{0b} \in R\right)}{\max\left[m^{-1} \sum_{i=1}^{m} I\left(\widetilde{T}_i \in R\right), 1\right]}.$$

One assumption that has been implicitly used throughout this paper is that the expression measurements are independent across genes. However, this assumption is not necessary to the validation calculation of the pFDR using either of the two methods described. In particular, the estimation of pFDR is valid under the dependence conditions on the genes which are described in Storey and Tibshirani (submitted for publication).

Calculation of q-values

Suppose that the sets R are of the form $R = \{t{:}|t|{>}c\}$ for some value c. Based on this set definition, we can define the q-value of an observed statistic x to be

$$q - value(x) = \min_{c:x \in R} p\mathrm{FDR}.$$

In other words, it is the minimum pFDR possible when rejecting a statistic with value x. The q-value will take values between 0 and 1, with smaller values indicative of stronger evidence for differential expression. In the work of Efron and Tibshirani (2002) and Storey (2002), connections were drawn between the FDR estimation approach described here with the original method of Benjamini and Hochberg (1995). We briefly discuss this technique to explain how one can estimate q-values for each gene.

Table 10 Estimated positive false discovery rate (pFDR) and false discover rate (FDR) for various choices of *A* and *R* with prostate cancer data

R	*A*	pFDR	FDR	No. genes called significant
$\{t:\vert t\vert >1.5\backslash\}$	$\{t:\vert t\vert <0.15\backslash\}$	0.06	0.06	976
$\{t:\vert t\vert >1.5\backslash\}$	$\{t:\vert t\vert <0.5\backslash\}$	0.11	0.11	1,133
$\{t:\vert t\vert >2\backslash\}$	$\{t:\vert t\vert <0.15\backslash\}$	0.04	0.04	829
$\{t:\vert t\vert >2\backslash\}$	$\{t:\vert t\vert <0.5\backslash\}$	0.08	0.08	1,021
$\{t:\vert t\vert >3\backslash\}$	$\{t:\vert t\vert <0.15\backslash\}$	0.03	0.03	777
$\{t:\vert t\vert >3\backslash\}$	$\{t:\vert t\vert <0.5\backslash\}$	0.06	0.06	976

First, consider the approach described in Table 1. Based on the null statistics $T_{1.}^{0b}, \ldots, T_{m.}^{0b}$, one can construct *p*-values $P_1, \ldots, P_m$ by

$$P_i = B^{-1}\sum_{b=1}^{B} I\left(|T_{i.}^{0b}| \geq |T_{i.}^{0}|\right)$$

i=1,…,m. In the Benjamini-Hochberg procedure, we order the p-values in increasing order from $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$. If we wish to control the FDR (and hence the pFDR) at a threshold level q^*, we estimate

$$k = \max\left\{i : mP_{(i)}/i \leq q^*\right\}$$

and conclude that the genes corresponding to $P_{(1)}, \ldots, P_{(k)}$ are differentially expressed. The way to estimate a q-value for each gene would be to find the smallest threshold level q^* such that it is in the set of genes that are differentially expressed. The smaller the q-value, the stronger the evidence that the gene is differentially expressed between the two conditions. A similar approach works with the method in Table 2. We first apply the methods for estimating FDR. We used the methods summarized in Boxes 1 and 2. Since the results were similar for the two, we only show the results using Box 1. We took B =2,000 in our analyses. The estimates of FDR for various choices of *R* and *A* are given in Table 10, along with the number of genes declared significant at this cutoff value for FDR.

Next, we used the Benjamini-Hochberg sequential procedure for determining which genes were differentially expressed using various values of q^*. Based on Table 11, we see that there are about 100 genes that express high statistical significance for differential expression. In fact, there are 123 genes with q-values less than or equal to 0.005 and so on. The number of significant genes using this FDR controlling procedure is much larger than using an adjustment such as the Bonferroni correction or Benjamini and Yekutieli (2002), which gave zero genes as significant.

Table 11 Number of genes called significant $\left(\hat{k}\right)$ for various values of q* based on the Benjamini-Hochberg procedure

q*	$\left(\hat{k}\right)$
0.2	1,584
0.1	1,204
0.05	949
0.02	732
0.01	449
0.005	123

Bioinformatics investigations

Based on the differentially expressed genes found using the LASSO, we can probe existing bioinformatic databases to determine potential transcriptional pathways. One ideal database for performing this search is the Kyoto Encyclopedia of Genes and Genomes (KEGG), which is located at http://www.genome.ad.jp/kegg/. KEGG is a knowledge base that allows one to systematically analyze gene functions at many levels. It consists of subdatabases containing information regarding the genomes of organisms, cellular processes and enzymatic reactions. By inputting the identification numbers for groups of genes that are differentially expressed, we are able to find candidate pathways that may represent potential therapeutic targets in prostate cancer.

Based on the lists of genes generated using the method proposed in the “Statistical methods for differential expression” section, we interrogated the KEGG database. KEGG outputs of these analyses can be found at http://www.sph.umich.edu/~ghoshd/COMPBIO/Meta/PathwaySearchResult1.htm and http://www.sph.umich.edu/~ghoshd/COMPBIO/Meta/PathwaySearchResult2.htm. The lists of pathways generated by these analyses can be then validated experimentally by cancer biologists. Further results using other methods for differential expression can be found in Rhodes et al. (2002).

Discussion

In this paper, we have outlined the issues involved in combining results from several microarray experiments. These considerations led to the development of a simple method for determining differential expression in prostate cancer versus benign prostate tissue across multiple studies. We can then interrogate existing databases for potential therapeutic targets and candidate biological pathways. This approach thus avoids having to utilize traditional laboratory methods for validation of genes, which are often expensive and time-consuming.

While differential expression has been used for single microarray studies (Efron et al. 2001; Ibrahim et al. 2002; Lonnestedt and Speed 2002), less work has been done in the multiple study setting. We have also developed a q-value-based method for assessing differential expression (Rhodes et al. 2002).

There are certain limitations of our approach. First, we are attempting to generate candidate biological pathways using gene expression profiles taken from tissue samples at one point. There is a fundamental confounding of longitudinal and cross-sectional effects here because of the cross-sectional study design. In particular, if two genes are differentially expressed in prostate cancer tissue relative to benign tissue, it does not necessarily mean that one gene regulates the other or that they are co-regulated. By utilizing the bioinformatics databases, we are attempting to bring in external biological knowledge into the analysis as well. While we are able to generate candidate pathways, it should be pointed out that they need to be studied further experimentally to scientifically validate them.

In the analysis plan we have proposed, we have assumed that the *t*-statistics for genes are independent. While these assumptions will not literally hold true, they are used to derive a relatively simple measure of differential expression. The other purpose of this approach is as a means of ranking genes that would be useful for further follow-up study. One such analysis is the KEGG search we mentioned in the Bioinformatics investigation.

It should be emphasized that the methods presented here involve both statistical modeling procedures as well as bioinformatics-based methods. As new high-throughput technologies are developed for proteomic and eventually metabolomic data, extracting maximum information from them will require a combination of these two approaches.

Acknowledgements The first author would like to thank Sharon Sun and Xihong Lin for useful discussions. The work of D.G. has been partially supported by a Prostate Cancer Seed Grant and MUNN Idea Grant from the University of Michigan Cancer Center.

References

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Benjamini Y, Yekutieli D (2002) The control of the false discovery rate in multiple testing under dependency. Ann Stat 30:1165–1188

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 29:365–371

Bussemaker H, Li H, Siggia ES (2001) Regulatory element detection using correlation with expression. Nat Genet 27(2):167–171

Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412:822–826

Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol 23:70–86

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Ibrahim JG, Chen M-H, Gray RJ (2002) Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc 97:88–99

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, Lee CJ (2000) Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat Genet 26:233–236

Kuo WP, Jenssen T-K, Butte AJ, Ohno-Machado L, Kohane IS (2002) Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18:405–412

Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet S21:20–24

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

Lonnestedt I, Speed TP (2002) Replicated microarray data. Stat Sin 12:31–46

Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM, Isaacs WB (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. Cancer Res 61:4683–4688

Magee JA, Araki T, Patil S, Ehrig T, True L, Humphrey PA, Catalona WJ, Watson MA, Milbrandt J (2001) Expression profiling reveals hepsin overexpression in prostate cancer. Cancer Res 61:5692–5696

Normand SL (1999) Meta-analysis: formulating, evaluating, combining and reporting. Stat Med 18:321–359

Rhodes D, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 62:4427–4433

Schena M (2000) Microarray biochip technology. Eaton, Sunnyvale, Calif.

Shaffer J (1995) Multiple hypothesis testing. Annu Rev Psychol 46:561–584

Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc B 64:479–495

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nature Genet 22:281–285

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to ionization radiation response. Proc Natl Acad Sci 98:5116–5121

Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF Jr, Hampton GM (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res 61:5974–5978