BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 1943–1949
doi:10.1093/bioinformatics/bti260

Gene expression

Significance analysis of functional categories in gene expression studies: a structured permutation approach

William T. Barry[1], Andrew B. Nobel[2] and Fred A. Wright[1,*]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina 27599-7420, USA and
[2]Department of Statistics, University of North Carolina at Chapel Hill, North Carolina 27599-3260, USA

Received on October 30, 2004; revised and accepted on December 31, 2004
Advance Access publication January 10, 2005

ABSTRACT

Motivation: In high-throughput genomic and proteomic experiments, investigators monitor expression across a set of experimental conditions. To gain an understanding of broader biological phenomena, researchers have until recently been limited to *post hoc* analyses of significant gene lists.

Method: We describe a general framework, significance analysis of function and expression (SAFE), for conducting valid tests of gene categories *ab initio*. SAFE is a two-stage, permutation-based method that can be applied to various experimental designs, accounts for the unknown correlation among genes and enables permutation-based estimation of error rates.

Results: The utility and flexibility of SAFE is illustrated with a microarray dataset of human lung carcinomas and gene categories based on Gene Ontology and the Protein Family database. Significant gene categories were observed in comparisons of (1) tumor versus normal tissue, (2) multiple tumor subtypes and (3) survival times.

Availability: Code to implement SAFE in the statistical package R is available from the authors.

Contact: fwright@bios.unc.edu; wbarry@bios.unc.edu; nobel@email.unc.edu

Supplementary information: http://www.bios.unc.edu/∼fwright/SAFE

INTRODUCTION

High-throughput biotechnologies such as microarrays and two-dimensional (2D) gel electrophoresis enable researchers to simultaneously measure the expression of much of the genome, at either the transcriptional (Schena *et al*., 1995) or translational level (Honore *et al*., 2004). These technologies have found wide application in many areas of biology and medicine, including identifying genes differentially expressed across groups of samples or experimental conditions (Schena *et al*., 1995), performing classification or discrimination analysis in heterogeneous diseases such as cancer (Bhattacharjee *et al*., 2001; Petricoin *et al*., 2002), and elucidating the relationship between expression and covariates such as survival or tumor grade (Beer *et al*., 2002).

In many applications, the experimenter seeks to identify a statistically significant association between the expression profiles and another variable related with each array, such as a sample group assignment, an experimental factor or survival time. We will refer to this additional variable as the 'response', regardless of whether it is observed or determined by the experimental design. The most common method to analyze expression data proceeds in a gene-specific manner, using a statistical model to relate the response to the expression of each gene. An appropriate test statistic is calculated for each gene and used to assign a parametric or permutation-based p-value (Tusher *et al*., 2001; Dudoit *et al*., 2002b; Newton *et al*., 2004). Once a test statistic has been chosen, the primary statistical obstacle is accounting for multiple comparisons. Ranked lists of genes with small p-values are typically produced and subjected to an appropriate form of error rate control, such as the family-wise error rate (FWER) or the false discovery rate (FDR).

While it is important to identify individual genes that are associated with the response, most biological phenomena and human diseases are thought to occur through the interactions of multiple genes, via signaling pathways or other functional relationships. As the understanding of cellular processes has grown, so too have databases that provide biological annotation for known genes. For example, SWISS-PROT provides a set of keywords for each gene, based on a taxonomy that includes pathways, diseases and general biological processes (Boeckmann *et al*., 2003). The InterPro and Protein Families (Pfam) databases classify genes using homology-based domains in the protein sequence (Sonnhammer *et al*., 1997). More recently, The Gene Ontology Consortium (2000) has developed a comprehensive taxonomy of gene annotation for the separate ontologies viz. Biological Process, Cellular Component and Molecular Function. Each ontology is structured as a directed acyclic graph, with a hierarchy of terms that vary from broad levels of classification (e.g. Metabolism) down to more narrow levels (e.g. GTP biosynthesis).

With the availability of more comprehensive annotations, the focus of many gene-expression studies has shifted from the activity of individual genes to that of broader functional groups. The traditional gene-specific approach to expression analysis, however, does not readily produce an understanding of the biological processes driving the association between expression and response. In many cases, researchers have informally compared lists of significant genes to existing annotation in order to make judgments about the underlying biology (Tusher *et al*., 2001). Frequently, the list of significant genes is too long to develop a parsimonious understanding of the role of biological function.

Recently, a number of publications and software packages have adopted a more systematic approach to understanding the role of biological function by constructing *post hoc* tests for the relative

*To whom correspondence should be addressed.

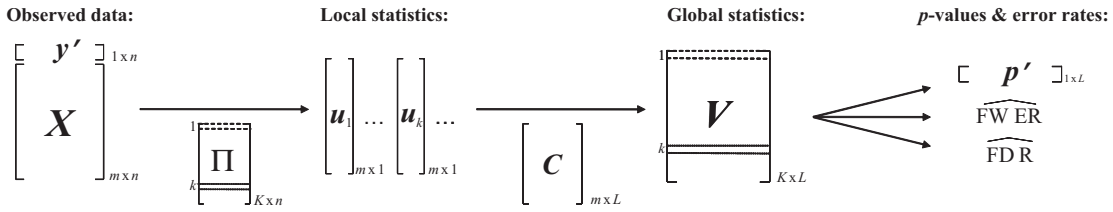


Fig. 1. Schematic for the SAFE procedure. The necessary components are illustrated above. The observed data consist of a matrix of normalized expression estimates, $\boldsymbol{X}$ and a response vector, $\boldsymbol{y}$. Permissible permutations of the response vector are specified in the matrix $\boldsymbol{\Pi}$. For each permutation, a vector of local statistics, illustrated here for the first and k-th permutation, measures the association between the permuted response and the expression of each gene. Gene category assignments are defined a priori and specified in matrix $\boldsymbol{C}$. For each permutation, global statistics are computed from the local statistics and $\boldsymbol{C}$. Based on the matrix of global statistics across all permutations, empirical p-values are obtained for each category along with estimated error rates.

enrichment of a gene category, or keyword, within the list of significant genes (Kim and Falkow, 2003; Draghici *et al*., 2003; Beißbarth and Speed, 2004; Hosack *et al*., 2003; Al-Shahrour *et al*., 2004; Zhong *et al*., 2004; Zeeberg *et al*., 2003; Berriz *et al*., 2003). A drawback to such gene-list approaches is that they rely on the initial gene list in a fundamental way and are sensitive to the choice of both significance criterion and error-control procedure. Moreover, they do not consider a gene's relative position in the ranked list. If genes belonging to a functional category show a modest downward shift in p-values compared to the remaining genes, this effect might not be noticeable when examining only the category membership in the list of significant genes. Indeed, after appropriate correction for multiple testing, there might be no significant genes at all, so that gene-list approaches utterly fail. Examples of such situations are presented further below.

As currently implemented, the gene-list methods rely on standard sampling theory, using the incorrect assumption that the genes are uncorrelated. For categories with highly correlated genes, the true Type I error rate may thus be substantially higher than the presumed rate. Finally, we note that the current gene-list methods use conservative error-control procedures when assessing multiple (possibly overlapping) keyword categories. To overcome these drawbacks, we propose a different, *ab initio*, approach to perform inference about gene categories that incorporates the entire set of p-values or their associated test statistics. Permutation is an integral part of the approach, and is used both to control the Type I error for individual categories and to provide refined estimates of multiple-comparison error rates.

METHODS

A New Approach: the SAFE Framework

In order to assess the significance of multiple gene categories, we propose a flexible, permutation-based framework, termed SAFE (for significance analysis of function and expression). SAFE extends and builds on an approach first employed in Virtaneva *et al*. (2001) for a two-sample microarray comparison of cancer subtypes. More recently, a method similar to Virtaneva *et al*. (2001) was proposed by Mootha *et al*. (2003) for a comparison of diabetes subtypes. In both methods, a two-stage approach is employed to assess the significance of a gene category. First, gene-specific statistics are calculated that measure the association between expression and the response of interest. Hereafter, we will refer to these as 'local' statistics. Then a larger-scale 'global' statistic is constructed as a function of the local statistics, with the goal of detecting a shift in the local statistics within a gene category to more extreme values, as compared to the remaining genes. The significance of the global statistics is assessed by repeatedly permuting the response values and recomputing local and global statistics. In this manner, the correlation of local statistics within each category is preserved, as is the correlation of global statistics across categories that contain overlapping genes.

The SAFE procedure is described in detail below and presented in Figure 1. It generalizes and extends the methods of Virtaneva *et al*. (2001) and Mootha *et al*. (2003) in critical respects. SAFE encompasses a wide variety of experimental designs and response vectors, and incorporates appropriate methods of error rate estimation directly into the permutation scheme. The procedure also leads naturally to informative plots for visualizing gene category significance.

Observed data Consider an experiment in which the expression of m genes is measured in each of n samples. The available data are in the form of an $m \times n$ matrix $\boldsymbol{X}$, where the element x_{ij} is the normalized expression estimate of gene i in sample j. The i-th row of $\boldsymbol{X}$, denoted $\boldsymbol{x}_i$, is the expression profile of gene i. The term gene is used here to generically identify a row of $\boldsymbol{X}$, although for some platforms expression estimates for a single gene might appear in multiple rows. We assume that suitable normalization and other preprocessing of the data (*cf.* Dudoit *et al*., 2002a; Li and Wong, 2001) has been performed. Each sample j is associated with an additional response variable y_j. As noted above, y_j may be a treatment assignment or a numerical response such as tumor grade or survival time; $\boldsymbol{y} = (y_1, \ldots, y_n)$ will be referred to as the response vector.

Prior to SAFE analysis, a collection of functional categories of interest is identified. For the l-th category, let $c_{il} = 1$ if gene i belongs to category l, and $c_{il} = 0$ if gene i falls outside the category. Let $\mathbf{c}_l$ be the $m \times 1$ vector of these indicators. The set of all such indicators is represented as an $m \times L$ matrix $\boldsymbol{C}$, where L is the number of categories under examination.

Statistics and permutation SAFE requires the user to specify two statistics. The first is a local statistic $U(\boldsymbol{x}_i, \boldsymbol{y})$, which measures the association between the expression profile of gene i and the response vector. In a study where $y_j \in \{0, 1\}$ denotes one of two experimental conditions, U might be an ordinary t-statistic for comparing $\{x_{ij}{:}y_j = 0\}$ and $\{x_{ij}{:}y_j = 1\}$. As genes in the same category might exhibit changes in either direction, a two-sided local statistic such as $|t|$ is also a natural choice.

The second, global statistic V assesses how the distribution of local statistics within a category differs from local statistics outside the category. For a given category l and local statistics $u_1, \ldots, u_m$, $V(\{u_i\}, \boldsymbol{c}_l)$ measures the difference between the local statistics of genes in category l, namely $\{u_i{:}c_{il} = 1\}$, and the local statistics of genes in the complement of the category, namely $\{u_i{:}c_{il} = 0\}$. Typically little is known about the joint density of the local statistics. For this reason we favor rank-invariant choices for V, such as the Wilcoxon rank sum (Virtaneva *et al*., 2001) or Kolmogorov–Smirnov statistic (Mootha *et al*., 2003), as likely to retain reasonable power under a variety of circumstances.

The significance of the global statistic for each functional category is assessed through a group $\Pi = \{\pi_1, \ldots, \pi_K\}$ of permissible permutations of the response vector. The permutations in Π reflect the underlying experimental design, including pairing of samples, blocking or other

sampling-based constraints. For many experimental designs, all $n!$ permutations are permissible, although fewer distinct permutations of the response vector may exist (as in the two-sample problem). For datasets of even modest size, it may not be computationally feasible to use all permutations, and the elements of Π are chosen as a random sample from all permissible permutations. The elements of Π are represented as permutations of the integers $\{1, \ldots, n\}$, so that Π is an $n \times K$ matrix. Let π_1 be the identity permutation, corresponding to the observed order of the response vector.

For each gene and each permutation $\pi_k \in \Pi$, let $u_{ik} = U(\boldsymbol{x}_i, \boldsymbol{y} * \pi_k)$ be the value of U when the response is permuted according to π_k. Here $\boldsymbol{y} * \pi = (y_{\pi(1)}, \ldots, y_{\pi(n)})$ is a reordering of the components of $\boldsymbol{y}$ according to π. Let $\boldsymbol{V}$ be the $K \times L$ matrix with entries $v_{kl} = V(\{u_{ik}\}, \boldsymbol{c}_l)$, the global statistic for the l-th functional category under permutation π_k. Empirical p-values are computed for each category as $p_l = K^{-1} \sum_{k=1}^{K} I\{v_{kl} \geq v_{1l}\}$, with $I\{\cdot\}$ denoting the indicator function.

Error estimation and plots Except where noted, we report the nominal empirical p-value for a significant category, uncorrected for multiple comparisons. To account for multiplicity in gene category tests, standard estimates of the FWER (Westfall and Young, 1989) or the FDR (Yekutieli and Benjamini, 1999) are computed for the set of categories that fall within a given rejection region. The matrix of global statistics is converted into a $K \times L$ matrix of permuted p-values with elements

$$p_{kl} = \frac{1}{K} \sum_{h=1}^{K} I\{v_{hl} \geq v_{kl}\}.$$

For a rejection region, $[0, p]$, the Westfall–Young estimate of the FWER is

$$\widehat{\mathrm{FWER}}_{WY}(p) = \max_{l:p_l \leq p} \left[\frac{1}{K} \sum_{k=1}^{K} I\left(\min_{h:p_h \geq p_l} p_{kh} \leq p_l \right) \right]$$

and the Yekutieli–Benjamini estimate of the FDR is

$$\widehat{\mathrm{FDR}}_{YB}(p) = \min_{l:p_l \geq p} \left[\frac{1}{K-1} \sum_{k=2}^{K} \left(\frac{\hat{V}_k(p_l)}{\hat{V}_k(p_l) + \hat{S}(p_l)} \right) \right]$$

where $\hat{V}_k(p) = \sum_{l=1}^{L} I(p_{kl} \leq p)$ and $\hat{S}(p) = \hat{V}_1(p) - \left[[1/(K-1)] \sum_{k=2}^{K} \sum_{l=1}^{L} I(p_{kl} \leq p) \right]$. Non-resampling based error estimates, such as q-values (Storey and Tibshirani, 2003), and the FDR step-up procedure (Benjamini and Hochberg, 1995) can be readily applied to $\{p_l\}$. Note that the permutation approach ensures that the Type I error is controlled for individual categories, even if the component genes exhibit highly correlated expression. Furthermore, permutation enables control of multiple-testing error rates without the need to adopt overly conservative procedures. For example, permutation-based control of the FWER exploits positive correlation among the global statistics for categories with overlapping genes, while a Bonferroni threshold in this case will be highly conservative. In our examples using the GO ontologies, the dependence between some categories (nodes) is very strong, as these categories may contain identical or nearly identical sets of genes.

The association of gene expression to the response can be presented across a category in the form of a SAFE-plot. For category l, the SAFE-plot displays the empirical cumulative distribution function of the ranked local statistics $\{u_i : c_{il} = 1\}$ against that of all genes. SAFE-plots thus display the realtive magnitude and direction of the differential expression for each gene in the given category. For hierarchically structured annotations such as GO, it is also useful to display SAFE results across a directed acyclic graph of the ontology thereby, revealing the relationships among significant categories.

Examples

To demonstrate the applicability and flexibility of SAFE, gene category analyses were conducted for several responses in a study of human lung carcinomas by Bhattacharjee *et al.* (2001). A total of 202 lung specimens were assayed with U95Av2 oligonucleotide arrays (Affymetrix, Santa Clara, CA). The data consisted of 16 normal tissues and 186 tumors, subclassified as adenocarcinomas ($n = 139$), pulmonary carcinoids ($n = 20$), small-cell lung carcinomas ($n = 6$) and squamous cell lung carcinoma ($n = 21$). Additional clinical information, including survival times, were available for 125 of the adenocarcinomas. Our significance analyses focused on three comparisons: (1) a two-sample comparison of normal versus cancerous samples, (2) an ANOVA model comparing cancer subtypes and (3) a survival analysis within the adenocarcinoma subgroup. The functional categories were derived from Pfam and GO. We highlight a few instances in which the results are supported by previous biological findings. These results are intended to serve as a proof of principle for SAFE, rather than a comprehensive reanalysis of the data.

Table 1. A list of significant gene categories for each response

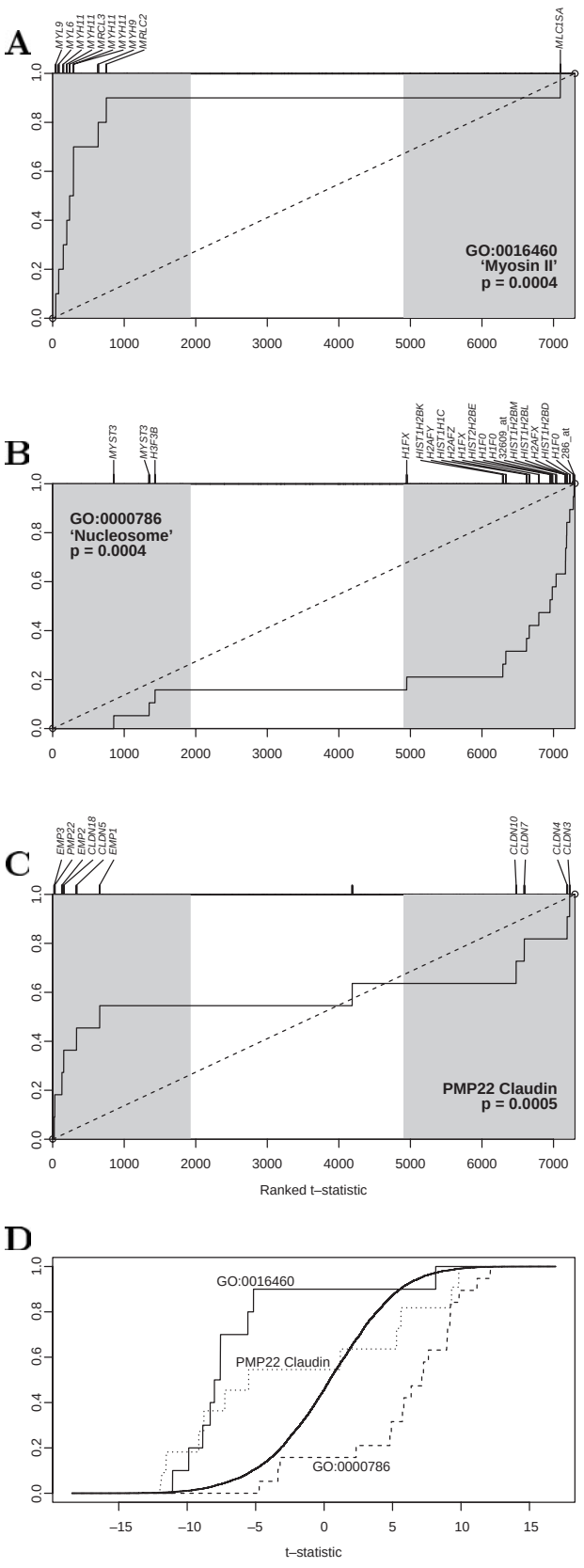
Category ID and name	Size	p-value	$\widehat{\mathrm{FDR}}$
Normal versus cancer			
GO:0016460, 'Myosin II'	10	0.0004	0.066
GO:0000786, 'Nucleosome'	19	0.0004	0.066
Pfam:PMP22_Claudin	11	0.0005	0.066
ANOVA among subtypes			
GO:0007010, 'Cytoskeleton org. and biogen.'	128	0.0003	0.064
GO:0007017, 'Microtubule-based process'	67	0.0005	0.064
GO:0006996, 'Organelle org. and biogen.'	153	0.0005	0.064
GO:0016043, 'Cell org. and biogenesis'	283	0.0007	0.064
GO:0009117, 'Nucleotide metabolism'	82	0.0008	0.064
GO:0007028, 'Cytoplasm org. and biogen.'	175	0.0011	0.087
GO:0006164, 'Purine nucleotide biosynth.'	45	0.0016	0.099
Survival of adenocarcinomas			
GO:0005643, 'Nuclear pore'	30	0.0002	0.034
GO:0046930, 'Pore complex'	30	0.0002	0.034

The FDR is estimated for all 635 gene categories.

Raw CEL files for the 202 U95Av2 arrays were obtained from http://www.pnas.org. Expression estimates and absent/present calls were obtained from dChip v1.3 software (http://www.dchip.org) using the PM-MM model from Li and Wong (2001). In keeping with the terminology above, each U95Av2 probeset is referred to as a gene. Arrays were normalized by quadratic scaling to an artificial array of median expressions for each gene (Yoon *et al.*, 2002). Genes were filtered out when called absent in more than half the samples of every tissue type, resulting in 7299 expressed genes.

As an exploration of the data, each SAFE analysis involved functional categories derived from GO and Pfam. Annotations for the U95Av2 array are available from http://www.affymetrix.com. GO gene categories sets were generated from the hierarchical structure of an ontology (Zeeberg *et al.*, 2003). Every GO term is represented by a node in a directed acyclic graph, and the functional category is defined as containing genes annotated either directly to that node or to any descendant node in the ontology. Perl scripts were used to extract the structure of the ontologies from the GO website (http://www.geneontology.org/) and assign categories. A total of 3860 GO nodes and 1811 Pfam domains were linked to the 7299 expressed genes. In order to retain power after correcting for multiple hypotheses, only categories of a sufficient size were considered. GO biological process and molecular function nodes containing at least 40 expressed genes were tested, along with cellular component nodes and Pfam domains annotated to at least 10 expressed genes (resulting in 207, 132, 120 and 176 categories, respectively).

For each response vector, an appropriate local statistic was chosen, the Wilcoxon rank sum were used as the global statistic and $K = 10\,000$ permutations was randomly generated. For each category, the empirical p-value was computed along with the Benjamini–Yekutieli FDR and Westfall–Young FWER estimates of the corresponding rejection region (Westfall and Young, 1989; Yekutieli and Benjamini, 1999). All categories with an estimated FDR ≤ 0.1 are reported as significant in Table 1 (complete results for all 635



categories are provided in Supplementary Table 1). It should be noted that because categories from three GO ontologies and Pfam were considered simultaniously in this exploratory analysis, the total number of categories is far greater than previous studies have reported (Zeeberg *et al*., 2003; Zhong *et al*., 2004; Berriz *et al*., 2003; Mootha *et al*., 2003); thus a stricter penalty for multiple testing is exacted. To demonstrate the results that would be achieved by examining only a single category set, separate error rate estimates are provided in Supplementary Table 1 for each ontology and Pfam.

RESULTS

Two-sample comparison

Differential expression was examined across normal and tumor samples using the absolute value of the Welch t-statistic as the local statistic. Observed values of the local statistic ranged from 0 to 18.4. In a gene-specific analysis based on 10 000 permutations of the array assignments, 1235 genes achieved the minimum gene-specific empirical p-value 0.0001 and 4319 had $p \leq 0.05$ ($|t| \geq 2.26$). With such dramatic differences between normal and tumor tissues generating a long gene list, obtaining useful biological conclusions require a broader perspective.

Among the four sets of functional categories applied in SAFE, three categories had $p \leq 0.0005$ and met the significance criterion for inclusion in Table 1: the cellular component nodes 'Myosin II' (GO:0016460), 'Nucleosome' (GO:0000786) and the Pfam domain 'PMP22 Claudin'. SAFE-plots display the relative extent and direction of differential expression observed for the sets of genes in these categories (Fig. 2). Of the ten expressed genes annotated to 'Myosin II,' nine were substantially underexpressed in the tumor samples compared to normal ($p = 0.0004$). In contrast, the GO term Nucleosome had 16 of 19 genes overexpressed in the tumor samples ($p = 0.0004$). Of the 11 genes annotated to 'PMP22 claudin,' 4 were substantially overexpressed in cancer and 6 were substantially underexpressed, ($p = 0.0005$). These results demonstrate the various directions of differential expression that can be detected in a two-sample SAFE analysis. Further, since no overlap in genes was observed among the three categories, we consider these to be separate findings (Supplementary Figure 1A).

The roles of myosin-related and cell-motility genes have long been studied in cancer and metastasis. A novel myosin family gene, *MYO18B*, was recently shown to be inactivated in ∼50% of lung cancers (Nishioka *et al*., 2002). The nucleosome genes we observed to be overexpressed in cancer were primarily histone family genes; acetylation of histones has been linked to *MYO18B* inactivation and lung cancer (Tani *et al*., 2004). Overexpression of claudin-4, as observed here, has been linked to metastatic breast and pancreatic cancers (Michl *et al*., 2003; Nichols *et al*., 2004). By examining

Fig. 2. SAFE-plots for significant categories in normal versus tumor. Welch t-statistics were computed for all expressed genes. The shaded region represents the range of local statistics that fall in the 5% tail area of the empirically derived null distribution ($|t| \geq 2.26$). The empirical cumulative distribution function for a gene category is plotted (solid line) against the ranks of all genes (dashed line). Tick marks above each plot display the location of genes within a category. Several genes are represented by more than one U95Av2 probeset. Probesets are labeled by their gene symbol when known and by the probeset ID otherwise. Significant gene categories can show consistent (**A**) underexpression, (**B**) overexpression in tumor versus normal or (**C**) bidirectional differential expression. (**D**) The unranked empirical cumulative distribution functions (labeled) are plotted against that of all genes (solid curve).

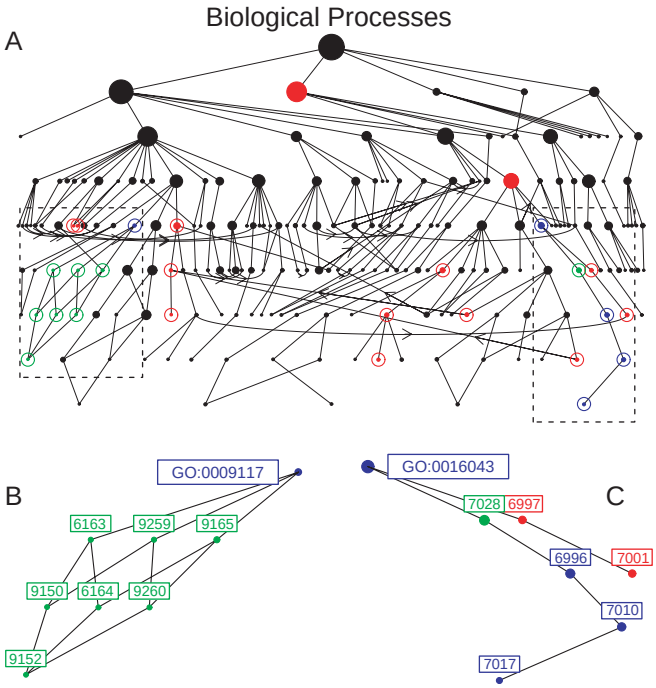


Fig. 3. SAFE results for ANOVA comparisons of tumor subtypes are displayed for (**A**) 207 GO biological processes containing at least 40 expressed genes. The graph displays the inheritance structure of GO terms. Nodes can have multiple parents, and for lateral or upward edges, arrows are drawn to indicate the child node. The area of each node is proportional to the number of genes belonging to the node. Nodes are colored by statistical significance of SAFE analysis: blue ($p < 0.001$) green ($0.001 \leq p < 0.01$), or red ($0.01 \leq p < 0.1$). Two distinct subgraphs containing all significant nodes (blue or green) are expanded in: (**B**) nodes under 'Nucleotide metabolism' (GO:0009117) and (**C**) nodes under 'Cell organization' (GO:0016043) and biogenesis). A larger version of the full directed graph is shown in Supplementary Figure 4.

entire gene categories instead of individual genes, we are able to identify a manageable number of gene categories warranting further hypothesis and study.

ANOVA

The standard ANOVA F-statistic was used as a local statistic to compare gene expression among the four cancer subtype; F ranged from 0 to 421. A total of 2689 genes achieved the minimum possible empirical p-value ($p = 0.0001$). The substantial differences in expression profiles between cancer subtypes provided the basis for successful discrimination in the original report (Bhattacharjee *et al.*, 2001). Here we employ SAFE to establish which functional categories consistently differ in expression across cancer subtypes.

Seven biological process nodes (having p-values ≤ 0.0016) met the criterion of FDR ≤ 0.1 for inclusion in Table 1. It is apparent from examining the SAFE results across the hierarchical structure of the ontology (Fig. 3) that significant categories fall into two distinct families: 'Cell organization and biogenesis' (GO:0009117), and 'Nucleotide metabolism' (GO:0016043). Figure 3B also illustrates that a broader category can be more significant than any of the nodes beneath it, due to the aggregation of gene effects across different descendants. These results add biological interpretability to the cluster analyses and gene-specific analyses from the original report.

Survival analysis

Censored survival data were available for 125 subjects with adenocarcinomas, with 71 observed deaths and 54 censored observations. The association between a gene's expression and survival was assessed with a univariate Cox proportional hazard model. The local statistic was the absolute value of the gene's regression coefficient divided by its standard error, $|\hat{\beta}|/\mathrm{SE}(\hat{\beta})$. The resulting Z-like statistics ranged from 0 to 3.98. The data provide an example where standard gene-specific approaches fail to provide useful conclusions. While 496 expressed genes had a gene-specific p-value <0.05 ($|z| \geq 1.96$), none was significant after multiple-testing correction (all FDR and FWER estimates were >0.2). We then applied the SAFE approach, which is sensitive to the aggregate effect of genes with related biological functions.

After accounting for multiple testing, two related GO cellular component nodes were significant (Table 1): 'Nuclear pore' (GO:0005643) and 'Pore complex' (GO:0046930). However, the two nodes contain an identical set of 30 genes and should be considered a single finding ($p = 0.0002$). Supplementary Figure 2A displays the set of linked genes and their respective direction of association with survival. Likewise, the parental node, Nuclear membrane, was marginally significant ($p = 0.0012$, $\widehat{\mathrm{FDR}} = 0.106$) but shared 30 of 51 genes with the other nodes. An additional SAFE-plot for the genes unique to 'Nuclear membrane' (GO: Supplementary Figure 2B) indicates that only the nuclear pore genes are associated with survival.

Although the original report (Bhattacharjee *et al.*, 2001) found a relationship between survival and a cluster-defined adenocarcinoma subclass ($p = 0.005$), our result is stronger, remarkably specific in its biological implications and offers new directions for exploration. We note that the role of nuclear transport in cancer (Kau *et al.*, 2004) and cancer aggressiveness (Agudo *et al.*, 2004) has been the subject of recent attention.

DISCUSSION

High-throughput biotechnologies have generated great interest in the elucidation of biological pathways and regulatory gene networks, and much effort has been made toward understanding these phenomena using gene-expression data. The various approaches can be depicted in a broad spectrum with two extremes: one computational and statistical, the other experimental. At the computational extreme, data from multiple gene-expression studies may be used to automatically construct networks of genes with highly correlated expression patterns. These patterns can in turn be used to annotate genes and predict function in a probabilistic manner (Zhou *et al.*, 2002; Troyanskaya *et al.*, 2003). However, these approaches are unable to test the association of functional categories with new experimental conditions or disease states. At the other extreme, detailed pathways are being developed for the direct interaction of genes at the translational level (Kanehisa, 1997), based on careful study of model organisms. These pathways may provide useful functional annotation, but currently cannot be used to model or predict disease states or the response of a tissue to experimental perturbation. Thus there remains a great need for approaches that efficiently test a large number of hypotheses for the relationship between gene expression and biological function in

the context of a given experiment. We propose SAFE as such a procedure, to be used repeatedly as a workhorse to investigate possible functional relationships.

The implementation of SAFE considered here differs considerably from standard gene-list methods. However, the gene-list methods may be thought of as essentially special cases of the SAFE framework. Both approaches compare the genes within a category against those in its complement, in contrast to other methods that assess significance by directly combining gene-specific effects (Goeman *et al*., 2004). The presence or absence of a gene on a list can be viewed as the outcome of a binary local statistic. The global statistic for the gene-list methods is the sum of these local statistics, i.e. the total number of genes in a category appearing on the list. The SAFE framework (Fig. 1) makes it clear that other choices of local and global statistics are possible, with possible improvements in power. Using rank-based global statistics, SAFE can detect gene categories with a high proportion of marginally significant genes that fail to appear on the significant gene list.

To illustrate this point, consider the association of the node 'Nuclear pore' (GO:0005643) to survival among adenocarcinomas (Table 1, SAFE p-value 0.0002). The result is highly significant and the collective shift in ranked statistics quite obvious (Supplementary Figure 2A). However, among 496 genes with parametric gene-specific p-values <0.05, there are only four genes belonging to Nuclear pore. Even using the anticonservative hypergeometric test for list membership gives $p = 0.1431$, illustrating the improved power of SAFE over gene-list methods for this category.

Gene-list methods typically rely on standard sampling theory to test the significance of a functional category and assume that the local statistics are independent. Thus the null distributions for gene-list global statistics are assumed to be hypergeometric (Draghici *et al*., 2003; Beißbarth and Speed, 2004; Al-Shahrour *et al*., 2004; Hosack *et al*., 2003; Zeeberg *et al*., 2003; Zhong *et al*., 2004; Berriz *et al*., 2003) or approximations thereof (Kim and Falkow, 2003). These latter differences are minor when one considers that the independence assumption is clearly violated for some categories. For example, we note that the 67 genes in the GO cellular component node 'Cytosolic ribosome' (GO:0005830) had an average pairwise correlation of 0.406 across the adenocarcinoma samples. For these samples, a randomly chosen set of 67 genes is very unlikely to have such a high correlation ($p = 0.0001$ for $|r| > 0.406$ in 10 000 randomly sampled gene sets). The p-values used by SAFE are based on permutations of the response vector that keep the gene-expression values intact, thereby preserving the correlation among genes. Permutation-based p-values have been proposed in some of the gene-list methods (Al-Shahrour *et al*., 2004; Zhong *et al*., 2004; Berriz *et al*., 2003), but in a completely different manner. These approaches are equivalent to permuting the rows of the category matrix $\boldsymbol{C}$, while maintaining the observed test statistics and significant gene-list. As intended, these methods account for the correlations among overlapping categories, but fail to address the possibility of inflated Type I errors resulting from dependent local statistics.

As implemented above, SAFE calculates permutation-based p-values using a separate null permutation distribution for each category (i.e. column of $\boldsymbol{V}$), rather than pooling all the values in $\boldsymbol{V}$ into a single null distribution. In contrast, Mootha *et al*. (2003) used pooling to compute a FWER-adjusted p-value for the largest Kolmogorov–Smirnov statistic, after scaling the statistics based upon differing category sizes. However, such standardization methods ignore the unknown correlation among local statistics and can therefore produce unequal null distributions among the categories. The inadequate standardization of global statistics provides a strong rationale against pooling in SAFE. Indeed, examining the permutation distributions of Wilcoxon statistic standardized for category size (Supplementary Figure 3) reveals many instances in the example data where the global statistics remain improperly scaled. In this circumstance, a p-value generated from the pooled null distribution will not control the Type I error of a given category properly, and can differ from the nominal p-value by a factor ≥ 10 (Supplementary Figure 3). Although pooling within SAFE meets the technical requirements for weak control of the FWER (Westfall and Young, 1989), inadequate standardization will reduce power for most categories.

In the examples we have used 'hard' category assignments: a gene belongs to a category or it does not. Recent publications have promoted more probabilistic approaches to gene function (Fraser and Marcotte, 2004; Troyanskaya *et al*., 2003) that are appropriate for unknown genes or those with less certain annotation. This suggests an extension of SAFE to 'soft' categories, in which the degree of membership in a category is reflected by a score on the interval [0, 1]. Soft categories can be easily incorporated into the SAFE framework, requiring that only an appropriate global statistic be chosen to weight the local statistics by their respective scores. Soft categories would also allow one to appropriately downweight the local statistics of a gene represented in multiple rows of $\boldsymbol{X}$, whereas current methods cannot flexibly handle such redundancy. Extensions such as these will further extend the power and potential of SAFE as a generic method to test for relationships between biological function and gene expression.

ACKNOWLEDGEMENTS

We thank the reviewers for their helpful suggestions. Karl Strohmaier provided assistance with Perl scripts to incorporate the GO inheritance structure. Jonathan Gelfond assisted with C programming to fit the Cox regression model. Support provided in part by NIH grants P30 HD003110, T32 ES07018 and P30 ES10126.

REFERENCES

Agudo,D., Gómez-Esquer,F., Martínez-Arribas,F., Núñez-Villar,M.J., Pollán,M. and Schneider,J. (2004) Predictive makers and cancer prevention nup88 mRNA overexpression is associated with high aggressiveness of breast cancer. *Int. J. Cancer*, **109**, 717–720.

Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Beißbarth,T. and Speed,T.P. (2004) Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

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

Berriz,G.F., King,O.D., Bryant,B., Sander,C. and Roth,F.P. (2003) Characterizing gene sets with funcassociate. *Bioinformatics*, **19**, 2502–2504.

Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The Swiss-Prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Dudoit,S., Yang,Y.H., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002a) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Dudoit,S., Yang,Y.H., Speed,T.P. and Callow,M.J. (2002b) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Fraser,A.G. and Marcotte,E.M. (2004) A probabilistic view of gene function. *Nat. Genet.*, **36**, 559–564.

Goeman,J.J., van de Geer,S.A., de Kort,F. and van Houwelingen,H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Honore,B., Ostergaard,M. and Vorum,H. (2004) Functional genomics studied by proteomics. *Bioessays*, **26**, 901–915.

Hosack,D.A., Dennis,G.Jr., Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with ease. *Genome Biol.*, **4**, P4.

Kanehisa,M. (1997) A database for post-genome analysis. *Trends Genet.*, **13**, 375–376.

Kau,T.R., Way,J.C. and Silver,P.A. (2004) Cell adhesion and signalling by cadherins and ig-cams in cancer. *Nat. Rev. Cancer*, **4**, 106–117.

Kim,C.C. and Falkow,S. (2003) Significance analysis of lexical bias in microarray data. *BMC Bioinformatics*, **4**, 12.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index imputation and outlier detection. *Proc. Natl Acad. Sci., USA*, **98**, 31–36.

Michl,P., Barth,C., Buchholz,M., Lerch,M.M., Rolke,M., Holzmann,K.H., Menke,A., Fensterer,H., Giehl,K., Lohr,M. *et al*. (2003) Claudin-4 expression decreases invasiveness and metastatic potential of pancreatic cancer. *Cancer Res.*, **63**, 6265–6271.

Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al*. (2003) Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Newton,M.A., Noueiry,A., Sarkar,D. and Ahlquist,P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Nichols,L.S., Ashfaq,R. and Iacobuzio-Donahue,C.A. (2004) Claudin 4 protein expression in primary and metastatic pancreatic cancer support for use as a therapeutic target. *Amer. J. Clin. Pathol.*, **121**, 226–230.

Nishioka,M., Kohno,T., Tani,M., Yanaihara,N., Tomizawa,Y., Otsuka,A., Sasaki,S., Kobayashi,K., Niki,T., Maeshima,A. *et al*. (2002) Myo18b, a candidate tumor suppressor gene at chromosome 22q12.1, deleted, mutated and methylated in human lung cancer. *Proc. Natl Acad. Sci., USA*, **99**, 12269–12274.

Petricoin,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,C., Fishman,D.A., Kohn,E.C. and Liotta,L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, **270**, 467–470.

Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tani,M., Ito,J., Nishioka,M., Kohno,T., Tachibana,K., Shiraishi,M., Takenoshita,S. and Yokota,J. (2004) Correlation between histone acetylation and expression of the *myo18b* gene in human lung cancer cells. *Genes Chromosomes Cancer*, **40**, 146–151.

The Gene Ontology Consortium (2000) Gene ontology:tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Troyanskaya,O.G., Dolinski,K., Owen,A.B., Altman,R.B. and Botstein,D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.

Virtaneva,K.I., Wright,F.A., Tanner,S.M., Yuan,B., Lemon,W.J., Caligiuri,M.A., Bloomfield,C.D., de la Chapelle,A. and Krahe,R. (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl Acad. Sci. USA*, **98**, 1124–1129.

Westfall,P.H. and Young,S.S. (1989) *P*-value adjustment for multiple tests in multivariate binomial models. *J. Amer. Statist. Assoc.*, **84**, 780–786.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference*, **82**, 171–196.

Yoon,H., Liyanarachchi,S., Wright,F.A., Davuluri,R., Lockman,J.C., de la Chapelle,A. and Pellegata,N.S. (2002) Gene expression profiling of isogenic cells with different tp53 gene dosage reveals numerous genes that are affected by tp53 dosage and identifies *cspg2* as a direct target of p53. *Proc. Natl Acad. Sci., USA*, **99** (24), 15632–15637.

Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al*. (2003) Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.

Zhong,S., Tian,L., Li,C., Storch,F.K. and Wong,W.H. (2004) Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework. *Proc. IEEE Comput. Syst. Bioinformatics*, in press.

Zhou,X., Kao,M.C. and Wong,W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci., USA*, **99**, 12783–12788.