Neurochemical Research, Vol. 29, No. 6, June 2004 (© 2004), pp. 1213–1222

Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex*

Paul Pavlidis,[1,2,6] Jie Qin,[1] Victoria Arango,[3,4,5] John J. Mann,[3,5] and Etienne Sibille[3,5]

(Accepted August 14, 2003)

One of the challenges in the analysis of gene expression data is placing the results in the context of other data available about genes and their relationships to each other. Here, we approach this problem in the study of gene expression changes associated with age in two areas of the human prefrontal cortex, comparing two computational methods. The first method, "overrepresentation analysis" (ORA), is based on statistically evaluating the fraction of genes in a particular gene ontology class found among the set of genes showing age-related changes in expression. The second method, "functional class scoring" (FCS), examines the statistical distribution of individual gene scores among all genes in the gene ontology class and does not involve an initial gene selection step. We find that FCS yields more consistent results than ORA, and the results of ORA depended strongly on the gene selection threshold. Our findings highlight the utility of functional class scoring for the analysis of complex expression data sets and emphasize the advantage of considering all available genomic information rather than sets of genes that pass a predetermined "threshold of significance."

KEY WORDS: Age; brain; gene expression; microarray; prefrontal cortex; statistics.

INTRODUCTION

Many studies of gene expression, if not most, seek genes that are "differentially expressed," namely, those that show changes in expression associated with an experimental variable of interest. Such genes can be rare or plentiful and fall into a continuum of varying levels of statistical confidence. In most studies, genes are eventually divided into two groups, those that show differential expression and those that do not, often based on essentially arbitrary statistical or *ad hoc* criteria. For example, all genes with t test P values less than 0.0001 might be selected or all those with more than two-fold changes in expression. These criteria are arbitrary in the sense that they are guided by considerations of sensitivity–specificity tradeoffs, which reflect the relative costs of false positives and negatives rather than an underlying biological reality.

Whereas identifying lists of individual genes that show expression changes is important, there is an increasing need to move beyond this level of analysis. Instead of simply enumerating a list of genes, we want to know how they interact as parts of complexes, pathways, and

* Special issue on Expression Profiling Within the Central Nervous System II.

[1] Columbia Genome Center, Columbia University, New York, New York.
[2] Department of Biomedical Informatics, Columbia University, New York, New York.
[3] Department of Psychiatry, Columbia University, New York, New York.
[4] Department of Anatomy & Cell Biology, New York State Psychiatric Institute, New York, New York.
[5] Department of Neuroscience, New York State Psychiatric Institute, New York, New York.
[6] Address reprint requests to: Paul Pavlidis, Department of Biomedical Informatics and Columbia Genome Center, Columbia University, 1150 St. Nicholas Ave, New York, New York 10032. Tel: 212-851-5141; Fax: 212-851-8149; E-mail: pp175@columbia.edu

0364-3190/04/0600–1213/0

networks. For instance, it is often not known whether robust changes in individual genes will have more biological relevance compared to weaker but coordinated effects in sets of genes that act synergistically within pathways. The promise of genomics technologies will not be realized until we are able to integrate disparate data sources in order to make predictions about normal cellular processes as well as mutation, disease, and drug effects.

A simple way of partly addressing this type of question is to categorize the set of selected genes on the basis of known functionally related groups. Such groups can be defined in a number of ways, most popularly using gene ontology annotations (1). This type of data organization can provide insight into biological "themes" in the selected genes. However, most studies use a small number of very broadly defined groupings such as "metabolism" or "cell signaling" because a finer analysis would result in hundreds, if not thousands of gene categories coming under consideration.

Identifying a small number of relevant gene classes among a large set of candidate classes requires statistical tools. An increasingly popular mode of analysis is "overrepresentation analysis" (ORA). Given a list of selected differentially expressed genes, one asks whether particular functionally defined groups of genes are overrepresented. The most common way of statistically testing this is by the use of the cumulative hypergeometric distribution (or its binomial approximation), which measures the probability of observing at least a particular number of genes in a class among the selected genes. This approach has been used by a number of groups for the interpretation of microarray data (2–5) and has been applied in at least one study of the nervous system (6).

There are a number of interesting issues raised by this type of analysis. First, we note that ORA requires dividing the data into two groups of genes: "selected" and "non-selected." As mentioned above, this division is typically arbitrary, and there are several reasons why it might be desirable to avoid this division. First, it would be interesting to analyze the entire set of data for meaningful patterns among related genes, not just the selected genes. A gene that shows subtle changes in expression, and thus fails to be "selected," might be highly relevant when viewed in a larger context of genes that interact with it. Another reason is the choice of the threshold used to select genes may have an effect on the results. In other words, the gene classes given high overrepresentation scores may change if the gene selection threshold is changed. Finally, once genes are "selected," any information contained in the actual *P* values (or other scores) for the genes is not used by overrepresentation analysis. If not all selected genes are considered equal (they are usually rated on a continuous scale), it makes sense to treat genes differently based on the strength of their individual scores.

An alternative to overrepresentation analysis, "functional class scoring" (FCS) (7), addresses these problems. Two key features distinguish FCS from ORA. First, no gene selection step is used. Second, FCS provides a score based on all the genes in the class. This allows genes that might not have been selected on their own merits to still have a positive impact and preserves information contained in the gene-wise scores. Like ORA, FCS uses as its input a set of gene-wise scores [typically *P* values from a *t* test, analysis of variance (ANOVA), or regression analysis, but fold-change or other scores can be used], which makes it readily applicable to most data sets where ORA is applied. A related approach has been described (8) in which gene groups are analyzed using ANOVA to compare expression in a gene group to the expression of all genes.

Previously, we have described an extensive and continuous effect of aging on gene expression in two areas of the human prefrontal cortex (Erraji-Benchekroun et al., submitted). In this paper, we compared FCS to ORA to analyze these age-related changes in gene expression in human brain. To our knowledge, ORA has not been critically evaluated for its ability to detect patterns in gene expression data. Our results show how FCS can rapidly identify interesting patterns in the data, some of which would be difficult to detect without considering genes in related groups. We also found that FCS was more consistent than ORA between the two brain regions studied and that ORA results varied depending on the gene selection criteria used. FCS is a powerful alternative to ORA and can be applied in most situations where ORA is used.

EXPERIMENTAL PROCEDURE

Microarray Data. The human brain gene expression data set we studied is described fully elsewhere (Erraji-Benchekroun et al., submitted). Briefly, samples from 41 subjects, ranging from 13 to 79 years of age (45.2 ± 20.8 years, mean ± SD), were obtained from the brain collection of the Human Neurobiology Core, Conte Center for the Neuroscience of Mental Disorders, at the New York State Psychiatric Institute. All samples were collected and analyzed at this same site and were processed by the same experimenter. Gray matter from Brodmann areas (BA) 9 and 47 was dissected from frozen sections that had been transferred from −80°C to −20°C 2 h prior to sampling. Microarray samples were prepared according to the Affymetrix protocol (http://www.affymetrix.com/support/) and analyzed on HG-U133A GeneChips. One RNA sample was run for each brain region from each subject, and one array was run for each sample. Data were required to

pass strict quality control measures for RNA and hybridization quality (Erraji-Benchekroun et al., submitted). The microarray quality control parameters, measured with the Affymetrix Microarray Suite (version 5) software (http://www.affymetrix.com/support/), were as follows: noise (RawQ) less than 5, background signal less than 100, consistent number of genes detected as present across arrays, consistent scale factors, actin and GAPDPH 5′/3′ signal ratios less than 3, and consistent detection of BioB and BioC hybridization spiked controls. Of the original 82 samples, 75 microarrays met inclusion criteria. Final probe set signal intensities were extracted with the robust multiarray average (RMA) algorithm using the Bioconductor software suite (9,10).

Gene-wise Statistical Analysis. To score genes by their tendency to show changes in expression with age, we use the Pearson correlation of the expression level of each gene with the age of the subjects. *P* values were calculated for each correlation assuming the data are normally distributed (11). A more detailed analysis, which also includes selection of genes using ANOVA and nonparametric methods, is presented elsewhere (Erraji-Benchekroun et al., submitted). Genes detected as expressed in 10% of the samples or less or with a coefficient of variation below 2% (based on $\log_2$ scale) were screened out, leaving 11,838 probes for further testing.

Gene Ontology Annotations. All gene ontology annotations were obtained from publicly available sources through http://www.geneontology.org. For the experiments described here, we used only the "biological process" and "molecular function" ontologies. Gene ontology (GO) classes with fewer than 8 or more than 150 genes represented in the data were not considered. This choice is partly pragmatic: very small or very large classes are unlikely to be as informative because they may be too specific or too general. Another consideration is reducing the number of classes to mitigate multiple testing concerns. Classes that had identical gene members (i.e., they differ only in name) were considered as one. This resulted in 965 different classes for consideration. In part due to the hierarchical nature of the gene ontology, many of these classes overlap extensively, and most genes are members of multiple classes.

Functional Class Scoring. As described previously (7), in FCS all genes in a particular class are examined, and the class is given an aggregate "raw score." Specifically, the arithmetic mean of the $-\log(P \text{ value})$ from the correlation analysis for all genes in the class is used as an aggregate score. This procedure is repeated for each GO term. This method is referred to as the "experiment score" method (7). To convert the raw score for each class into a *P* value, we compared the raw score to an empirically derived distribution of raw scores for randomly selected classes of the same size using a statistical resampling approach (12). Specifically, for a class of size k, we repeatedly draw a random set of genes of size k from the data and calculate the raw score. This procedure is repeated q times ($q = 200{,}000$ in the experiments described here), and the resulting score distribution is stored. The *P* value for a class with a raw score r is calculated as the fraction of random trials resulting in a score higher than r. If this value was zero, the *P* value was arbitrarily set to $1/(2q)$. Thus, for the current study, the "best" *P* value reported is 2.5×10^{-6}. Note that the null distribution is different for classes of different sizes. In particular, the distribution for small class sizes is broader. Thus, we calculate null distributions for each class size under consideration. The software was implemented in the Java language (version 2), and pseudorandom numbers were generated using the Sun Microsystems implementation of the java.util.random class (http://java.sun.com/).

Overrepresentation Analysis. We used the binomial approximation to the cumulative hypergeometric distribution to calculate *P* values for each gene ontology class. Genes were selected using gene-wise *P* value (P_{gene}) thresholds varying from 10^{-3} to 10^{-6}.

Correcting for Multiple Occurrences of a Gene. In many data sets, including ours, the same gene can be represented multiple times. If not corrected for, this effectively results in some genes getting "multiple votes" in the scoring procedures. Thus, the multiple votes must be combined into a single representative value. A conservative approach, used by Pavlidis et al. (7), is to use the mean of the (log-transformed) *P* values for the repeated occurrences. However, we have since found that this approach gives too much weight to apparent negative results (high *P* values) yielded by some probe sets that can be explained by poor probe performance, rather than true negative results. For this reason, here we combine all occurrences of a gene into one using the minimum (best) *P* value for the occurrences as the aggregate score, rather than the mean. To calculate the size of a class, we count only the unique genes. Thus, each class has an "effective size" that is less than or equal to the number of probes in the class. This procedure is followed for both ORA and FCS.

Comparing GO Terms. Two GO terms were considered "related" if one was the immediate child or parent of the other in the GO hierarchy. In some cases, we also considered two terms related if they reside in different aspects of the GO but contain many of the same genes. For example, "calcium ion transport" is a "biological process" that is similar to (though not the same as) "voltage gated calcium channel activity" in the "molecular function" tree. Similarly, "complement activity" was considered similar to "complement activation." Our aim was to not overly penalize a method in a comparison if the basic information recovered was equivalent.

Multiple Testing Correction. Because we are testing many classes, we incur an accumulated risk of false-positive findings. This problem is identical for the overrepresentation and FCS methods. For this study, we test the method of Benjamini and Hochberg (13), controlling the false-discovery rate (FDR) at 0.05. We also used Bonferroni *P* value correction, in which *P* values are adjusted to min $\{P^* m, 1.0\}$, where m is the number of classes tested. We note that both of these *P*-value correction methods assume that the classes are independent, which is clearly not the case. Thus, they are conservative.

RESULTS

We began by analyzing the linear correlation of expression of each gene with age. There are numerous genes that show apparently continuous changes in expression associated with age. For example, using a stringent *P*-value threshold of 10^{-6}, we select 32 genes in BA9. At a less stringent threshold of 10^{-3}, we find 435 genes (at this threshold, we expect about 10 false positives). The complete results of our search for gene expression patterns associated with aging, including additional analyses that do not assume a linear relationship of expression to age, are presented elsewhere (Erraji-Benchekroun et al., submitted).

FUNCTIONAL CLASS SCORING

Our FCS method uses as input all 11,838 *P* values from the analysis (not just those with *P* values below a threshold). We tested 965 GO classes for the analysis

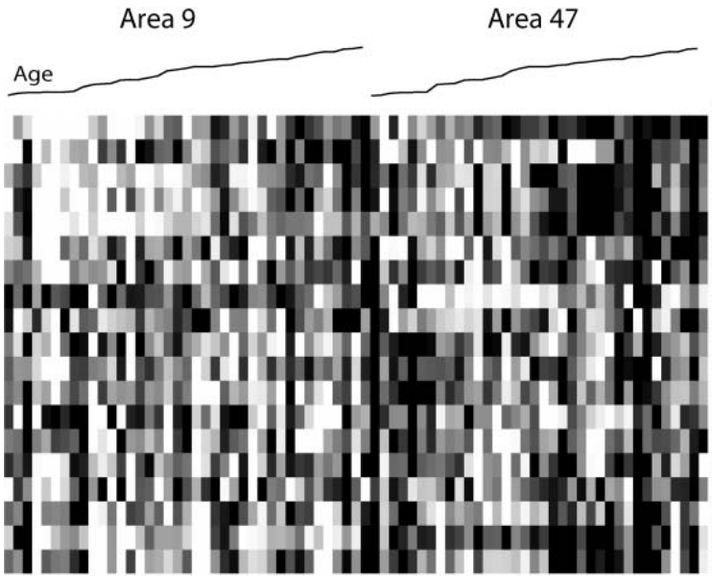
we present here. The top GO classes identified in each of the two prefrontal cortical regions by FCS are shown in Table I. Whereas the two regions (BA9 and BA47) are cytoarchitectonically and connectionally distinct, our gene-by-gene analysis shows that the gene expression patterns in the two regions are very similar (Erraji-Benchekroun et al., submitted). For this reason, we expected the class scoring methods to yield similar results in both regions. In agreement with this hypothesis, they yielded similar results by FCS, with five classes in common in the top 10 (4 of the top 5 in common) and several other classes representing related classes [e.g., "calcium–calmodulin binding activity" in Brodmann area 9 (BA9), related to "calcium/calmodulin-dependent protein kinase activity," found in both regions]. The fifth-ranked class in BA9 (neuropeptide hormone) is also ranked relatively high in BA47 (13th). When similar classes are considered along with identical classes, the agreement of BA9 to classes found in BA47 (among the top 10) is 6 classes, whereas similar or identical counterparts of 8 classes selected in BA9 could be found in the top 10 list for BA47 (see Tables I and III). Our shorthand notation for these results is "6\8." The details of one class that receives a high score in both regions ("calcium-calmodulin binding activity") are shown in Fig. 1. Figure 1 also illustrates that in order to get a high FCS score, not all of the genes in the class need to show changes in expression.

A few classes are highly ranked in only one brain region (Table I). For example, in BA47, "heavy metal sensitivity/resistance" is ranked 10th (class $P = 0.0012$), whereas in BA9, this class has a P of >0.5. This class of 13 genes primarily consists of metallothionein genes, a number of which show a tendency to increase expression with age (in particular, metallothioneins 1F, 1L, 1X, and 1H). Expression of the genes in this class was highly correlated between the two brain regions in the same subject ($R > 0.95$). Though there was also some tendency for these genes to increase with age in BA9 (as might be expected given the correlation of the gene expression patterns between the two areas), this effect was more pronounced in BA47. Interestingly, the effect of age on expression of these genes in BA47 was statistically quite weak, and using linear correlation, none has a P value better than 6.5×10^{-4} (uncorrected

Table I. The Top 10 Classes Identified by FCS

BA47		BA9	
Class*	P value†	Class*	P value†
Complement activation GO:0006956	2.5×10^{-6}	**Voltage-gated ion channel activity GO:0005244**	2.5×10^{-6}
Voltage-gated ion channel activity GO:0005244	0.000065	Calmodulin binding activity GO:0005516	0.00009
Complement activation, classical pathway GO:0006958	0.00013	**Cation channel activity GO:0005261**	0.000125
Cation channel activity GO:0005261	0.00049	**Complement activation GO:0006956**	0.000185
Neuropeptide hormone activity GO:0005184	0.000545	**Complement activation, classical pathway GO:0006958**	0.000205
Potassium channel regulator activity GO:0015459	0.00058	*Humoral immune response GO:0006959*	0.00022
Complement activation, alternative pathway GO:0006957	0.000745	*Voltage-gated calcium channel activity GO:0005245*	0.000255
Calcium/calmodulin-dependent protein kinase activity GO:0004685	0.000785	**Calcium/calmodulin-dependent protein kinase activity GO:0004685**	0.000405
Transmembrane receptor protein tyrosine kinase signaling pathway GO:0007169	0.001315	Muscle development GO:0007517	0.000515
Heavy metal sensitivity/resistance GO:0009634	0.001455	*Calcium ion transport GO:0006816*	0.00058

Note: FCS, functional class scoring; GO, gene ontology.
*The class name and GO identification number are given in order of decreasing score. Classes that are found in both regions are in **bold** (five classes). Classes that have related counterparts in the other area are in *italic*. Six of the classes in BA47 have identical or similar counterparts in the BA9 results, and nine of the classes listed for BA9 have identical or similar counterparts in the BA47 results.
†The P value for the class.



Probe	Gene P value Area 9	Area 47	Gene
210349_at	**3.33E-07**	**8.71E-07**	CAMK4
204392_at	**5.99E-05**	**0.039**	CAMK1
207613_s_at	**0.00165**	0.114	CAMK2A
210787_s_at	**0.00422**	**0.0169**	CAMKKB
34846_at	**0.00767**	**0.00881**	CAMK2B
212252_at	**0.0354**	**0.00194**	CAMKK2
211483_x_at	0.0691	0.771	CAMK2B
208802_at	0.0964	0.829	SRP72
213108_at	0.112	0.253	CAMK2A
212669_at	0.116	**0.00479**	CAMK2G
210404_x_at	0.127	0.509	CAMK2B
212757_s_at	0.157	**0.0036**	CAMK2G
208095_s_at	0.173	0.0814	CAMK2G
214322_at	0.199	0.533	CAMK2G
209956_s_at	0.285	0.905	CAMK2B
213812_s_at	0.296	0.0851	CAMKK2
208801_at	0.391	0.953	SRP72
208803_s_at	0.797	0.59	SRP72
208800_at	0.932	0.274	SRP72

Fig. 1. One of the high-scoring classes identified by FCS, "calcium/calmodulin-dependent protein kinase activity." Each row represents one gene, each column one sample. Darker shades of gray indicate lower expression levels. Some genes occur more than once in the data set, but only the best score is used for the group statistics. The samples in the image are arranged by area and then by increasing age within each area. Ages range from 13 to 79 years. The genes are identified at right, with the *P* values derived from the correlation coefficient in each area shown. Genes with *P* values <0.05 are in bold. (CAMK, calmodulin-dependent protein kinase; CAMKK, CAMK kinase; SRP72, signal recognition particle.) The inclusion of SRP72 in this class appears to be due to a report describing a naturally occurring CAMK–SRP fusion in the pancreas (26) and is thus questionable. This gene does not greatly influence the results here as the *P* values for SRP are rather high (best value >0.09 in BA9, >0.2 in BA47). Presumably, removing SRP from this class would result in an even higher score.

for multiple testing), and in our gene-by-gene analysis (27), which placed the significance threshold in the 10^{-5} range, only one gene in this class was identified as "significantly changed" in BA47. The profile of this top scoring gene, metallothionein 1F (MT1F), is shown in Fig. 2A. The reason this class receives a good FCS *P* value can be understood by observing that 10 of the 13 genes in this class have unusually good *P* values (e.g., <0.02), giving the class a mean $-\log P$ value of 1.79, which is unlikely to be observed by chance for a class

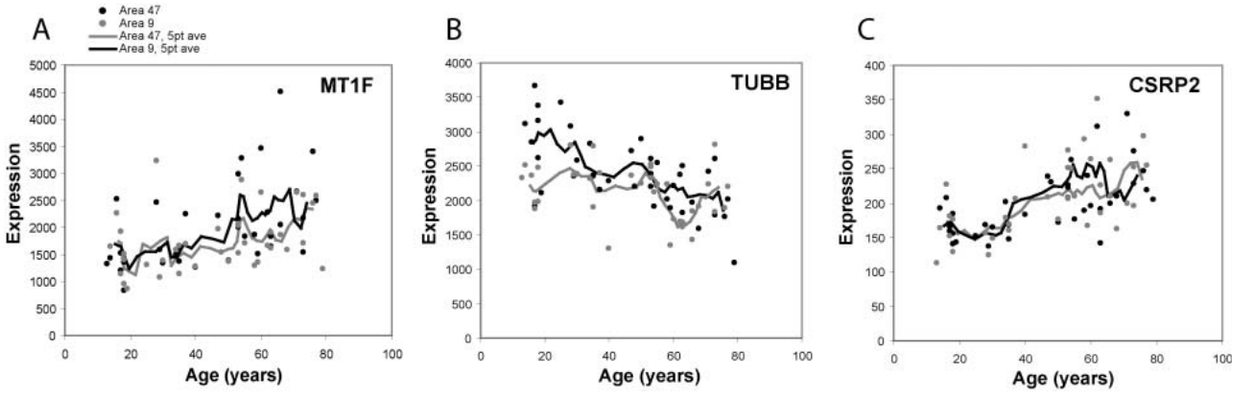


Fig. 2. Examples of expression profiles for genes from two classes identified by FCS. The legend applies to all three graphs. In each graph, the raw expression levels are plotted (points) along with five-point running averages (lines). Areas 47 and 9 are plotted in black and gray, respectively. (A) Metallothionein 1F, the highest scoring gene in the "heavy metal sensitivity/resistance." (B) and (C) The top two genes from the "muscle development class," beta-tubulin (TUBB) and cysteine and glycine-rich protein 2. (CSRP2). CSRP2 shows a consistent effect in the two regions, whereas the data for TUBB is in less agreement.

this size. The effects of noise on a weak signal helps explain why we fail to select this class in BA9, though genuine differences between brain regions could also be at play.

Another class found in only one region was "muscle development," in BA9 (this class was also identified by ORA at one of the gene selection thresholds tested). This class also stands out, as the source tissue is cerebral cortex and the identification of a muscle-related class at first glance appears counterintuitive, and we address this in our discussion. Here, the disagreement between brain regions also appears at least partly attributable to noisy data superimposed on a weak signal. For example, the highest scoring gene in this class (beta tubulin, TUBB), shows much less clear evidence of a change in expression with age in BA47 (Fig. 2B), whereas the second-highest scoring gene (cysteine and glycine-rich protein 2, CSRP2) looks quite similar in the two brain regions (Fig. 2C).

COMPARISON OF FCS TO OVERREPRESENTATION ANALYSIS

We compared our FCS method to an overrepresentation analysis (ORA), which requires setting a threshold to divide the data into "selected" and "nonselected" genes. Several criteria were used for the comparison. First, as it has been reported that complement genes increase expression in the rodent brain with age (14–17), and this effect is corroborated by our data (27), recovery of complement-related GO classes is a good test of the effectiveness of the methods. Second, we examined how well ORA agrees between the brain regions as compared to FCS, which showed high agreement between regions. Third, we examined how much the ORA results change as the gene selection P value threshold is changed, testing 4 different P_{gene} levels (10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}). For our initial comparisons, we simply compare the top 10 classes listed by each method.

ORA gives high scores (top 10) to several GO classes related to complement function high scores in both regions at only two P value thresholds (10^{-3}, the highest value tested, and 10^{-4}; Table II). Complement-related classes did not appear in the top 10 at other thresholds. The agreement between brain regions was also lower for ORA, with only 2 or 3 classes in common among the top 10, compared to 5 for FCS. When similar classes were considered, the maximum for ORA was 6|4 (for $P_{\text{gene}} < 10^{-3}$) and as little as 4|3 (using $P_{\text{gene}} < 10^{-6}$) compared to 6|8 for FCS (Table III).

In addition to having more disagreements between brain regions, the results for ORA vary depending on the P value threshold used (Table IV). For example, in BA47, going from a P_{gene} threshold of 10^{-3} to 10^{-4} results in 7 of the top 10 ORA classes remaining the same in BA47 and 5 in BA9. When P_{gene} is lowered to 10^{-5}, the agreement is 3 classes in BA47 and none in BA9.

For this data set, ORA and FCS are most similar when the gene selection threshold for ORA is 10^{-3}, with 4 and 6 classes among the top 10 matching those in the top 10 FCS results in areas 47 and 9, respectively. We note that the group of metallothionein genes discussed above is given a rank no higher than 34th by the ORA method (uncorrected class $P = 0.007$ at $P_{\text{gene}} = 0.001$, BA47). ORA and FCS appear generally comparable in terms of statistical power (Tables I and II), though the power of ORA varied depending on P_{gene} (not shown). Another caveat is that the P values we estimate for FCS are limited in precision by the number of random trials run, and for this reason ORA can in principle produce much lower P values. Using Benjamini–Hochberg correction (controlling the false-discovery rate at 0.05) and considering the ORA with $P_{\text{gene}} < 0.001$, in BA9 both methods select 9 classes. In BA47, ORA selects 12 and FCS selects 3. Using the more conservative Bonferroni criterion (controlling the family-wise error rate at 0.05), both methods select only one class in each region.

DISCUSSION

Our results demonstrate the importance of using prior knowledge of gene function in the interpretation of gene expression data. By using FCS, we were able to identify rapidly a selected set of functional gene classes that are relevant to the effects of age on gene expression in the aging human forebrain. These classes, including several relating to calcium signaling, electrical excitability, neuropeptide hormones, and protein tyrosine kinase signaling, may be particularly relevant to age-related cognitive decline. Though many of these classes could have been identified by "hand-annotating" the data set, FCS put the results in a statistical framework and allowed the detection of patterns that were only apparent when the genes were viewed in their functional context. A detailed discussion of the biological implications of our findings is presented elsewhere (27); here we focus our comments on issues surrounding the computational methods that we used.

We found that the ORA results agreed less between the two brain regions than FCS. We feel this is significant because we expected the two brain regions to

Table II. Top 10 Classes Selected by Overrepresentation Analysis Results, $P_{gene} < 0.001$*

BA47, ORA		BA9, ORA	
Class†	*P* value‡	Class†	*P* value‡
[a]**Complement activation GO:0006956**	7×10^{-8}	[a]*Muscle development GO:0007517*	2.45×10^{-5}
[a]**Complement activation, classical pathway GO:0006958**	7.3×10^{-7}	Regulation of cell proliferation GO:0042127	6.02×10^{-5}
Humoral immune response GO:0006959	1.07×10^{-5}	[a]**Complement activation, classical pathway GO:0006958**	0.0001
Regulation of muscle contraction GO:0006937	0.000141	[a]**Humoral immune response GO:0006959**	0.000125
Intramolecular isomerase activity, interconverting aldoses and ketoses GO:0016861	0.000177	[a]**Complement activation GO:0006956**	0.000126
Circadian rhythm GO:0007623	0.000264	Di-, trivalent inorganic cation transport GO:0015674	0.000143
[a]*Complement activation, alternative pathway GO:0006957*	0.000264	[a]Calmodulin binding activity GO:0005516	0.000187
Complement activity GO:0003811	0.000389	[a]Calcium ion transport GO:0006816	0.000528
Negative regulation of adenylate cyclase activity GO:0007194	0.000523	Nutritional response pathway GO:0007584	0.000794
*Neuropeptide hormone activity GO:0005184	0.000523	Tubulin binding activity GO:0015631	0.000845

Note: ORA, overrepresentation analysis; GO, gene ontology.
*These results are for a gene selection *P* value threshold of 0.001, which had the greatest similarity to the FCS results and also the best agreement between brain regions.
†The class name and GO identification number are given. Classes that are found in both regions are in **bold.** Classes that have related counterparts in the other area are in *italic*. Classes that were also identified by FCS in the same area are starred ([a]; four in BA47 and six in BA9). A number of other classes are similar to classes identified by FCS in the same or other region.
‡The *P* value for the class.

"behave" similarly in the analysis based on the gene-by-gene analysis (Erraji-Benchekroun et al., submitted). Perhaps more importantly, the ORA results were not consistent when the *P* value threshold was changed. Because the *P* value threshold for gene selection is determined by pragmatic rather than biological criteria, there is no "right answer" for overrepresentation analysis.

In many cases, ORA and FCS can capture the same information, and for this data set at a *P* value threshold of 10^{-3}, there was good agreement between them.

Table III. Summary of Data Comparing Brain Regions*

Method	BA47†	BA9†
FCS	6(5)	8
ORA, 0.001	6(3)	4
ORA, 0.0001	5(2)	2
ORA, 0.00001	2(2)	6
ORA, 0.000001	4(3)	3

Note: FCS, functional class scoring. ORA, overrepresentation analysis. The P_{gene} used is given for the ORA results.
*For each method tested, the value is the number of classes that were the same or similar between brain regions.
†The number of classes that are identical between the brain regions is in parentheses in the second column.

Table IV. Consistency of ORA Results as Gene Selection Threshold is Varied*

Method	BA47†	BA9†
ORA, 0.0001	7(7)	5(4)
ORA, 0.00001	3(3)	2(0)
ORA, 0.000001	2(2)	2(0)

Note: ORA, overrepresentation analysis; FCS, functional class scoring.
*We compared how many classes in the top 10 for ORA using $P_{gene} < 0.001$ are preserved as we decrease the P_{gene}.
†The value is how many classes are the same or similar to those found at $P_{gene} < 0.001$. The number of classes that are identical to those found at 0.001 are in parentheses.

However, at thresholds below 10^{-4}, ORA fails to give as high rankings to the (relatively) well-characterized effects of age on complement pathway gene expression in brain. Our conclusion from this comparison is that FCS captures similar information as ORA might, without the problem of selecting an appropriate threshold. FCS also has an intuitive appeal in that it allows all genes in the class to contribute to the result based on the individual scores for the genes, rather than simply setting a threshold. Thus, it is more "exploratory" than ORA, and may be more sensitive.

The identification of the "heavy metal sensitivity/resistance" class is a case where the FCS method was particularly illuminating, because our extensive screen for genes showing age-related changes identified just one of the genes in this class (27). However, we have reason to believe that these metallothionein genes are indeed upregulated with age (18,19), and a closer examination of the data revealed that several of the genes do appear to show age-related changes in our data, but were not easily detected by our gene-wise statistical methods, perhaps due to high interindividual variability. This demonstrates how the FCS method can pick up weak signals that are distributed among multiple genes. The ORA method does worse in this situation, as it can only give high scores to classes containing genes that make the chosen statistical criteria. In other words, if we rely on ORA alone, this class of genes is not as readily identified. Of course, FCS is not unlimited in its sensitivity and fails to give "heavy metal sensitivity/resistance" a high score in BA9, even though many of the genes show some weak evidence of increases with age in that region.

In situations where one is forced to divide genes into "selected" and "nonselected," some type of over-representation analysis is the only available approach. One example is the analysis of the output of a clustering algorithm for clusters enriched in particular types of genes (20). However, if a division of the data would be based on an arbitrary threshold, the FCS method seems more appropriate. FCS seems particularly helpful when graded changes in expression are expected; we can envision scenarios where it is easier to find true sets of "changed" and "unchanged" genes, but in our data the genes lie along a continuum of strong to weak changes with age. We note that the question of "cluster enrichment" can also be addressed by FCS, using a variant that does not require clustering (7).

One drawback of FCS is the need to calculate a background distribution via random sampling, which increases the time it takes to calculate the class scores. As this process takes only approximately 1 min (for 200,000 trials) on a modern PC, this has not hindered our use of the methods. The software we have developed produces both FCS and ORA statistics, which facilitates comparison of the methods and is available on request from the authors.

It is important to note that a high FCS or ORA score for a class does not imply that all genes in the class are relevant to the process under study. For FCS it indicates that the overall pattern of gene statistics in the class (as reflected by the aggregate score) is different than what one expects by chance. For ORA, it simply says that more genes in the class are recovered by the gene selection method than would be expected by chance. This means that FCS (and ORA) can only bring a functional class to our attention, and a closer examination of the data for the genes in the class is required to interpret the result.

A problem that we have not fully addressed is multiple testing, which affects FCS and ORA equally. We usually test hundreds of different classes, which reduces our power to detect statistically significant changes. This problem could be reduced by greatly narrowing the number of classes we consider. However, we wish to take as unbiased an approach as possible, as it is in practice difficult to decide *a priori* which classes to examine. Because the classes overlap (in large part due to the hierarchical nature of the GO), standard methods of *P* value correction do not work very well, as they assume that the tests are independent and thus are too conservative. Thus, neither ORA nor FCS select many classes when standard multiple testing correction methods are used, and we currently tend to pay more attention to the ranking of classes rather than corrected *P* values. This approach appears to be in accordance with the practice of many groups using ORA (3–6) but is not entirely satisfying. We are currently investigating a combination of approaches to this problem, including the use of *P* value correction methods that can handle a high degree of dependency (21,22) as well as various methods for limiting the "universe" of classes that are examined without overly biasing the approach.

Another general problem with gene class analysis is that the available annotations tend to be quite broad and are also incomplete. There are also errors or questionable annotations; a possible example is described in the legend to Fig. 1. We anticipate that as gene annotations mature, gene class analysis will increase in value and power. The gene ontology, in particular, does not yet reflect many types of "functions" that would be of interest to neuroscientists, and many known genes still lack good annotations in the publicly available databases. Thus, many genes have annotations at only fairly

high levels in the GO hierarchy. One way this problem manifests itself is that the best-annotated classes tend to be rather large, which can dilute the impact of patterns affecting a small number of genes. Another problem is that there are biases in the annotations. We initially found it curious that "muscle development" was one of the classes identified in our analysis, but a closer inspection reveals that this class contains a diverse set of genes including cytoskeletal proteins, potassium channels, and growth factors. It is possible that this finding reflects some change in vascular smooth muscle development in the brain associated with age (23). However, we also noted that many of these genes have known or plausible roles in neurons, and many have diverse roles in cell function and development that are only partly reflected in their GO annotations. This could be due to limited specific knowledge or to bias in the annotations. Thus, whereas there is a "neural development" GO class, it has only 20 members in our data compared to the 81 for muscle development. Our identification of a "muscle development" class may in fact reflect a set of genes that also have functions in *brain* development (or another process such as plasticity), and which are affected by age. A specific example of a gap in the annotations is the second-highest-scoring gene in the muscle development class: CSRP2. CSRP2 has a well-described role in muscle cell differentiation (24) but appears to have been little studied in brain. It is interesting that CSRP2 expression is decreased in rat forebrain after cocaine self-administration (25). Thus, this gene presumably hasd function in the brain, but there is no way this could be reflected in the GO annotations without further experimental data. Our results reinforce the potential role of CSRP2 (and many other genes) in brain function and suggest a possible link of this gene to development or plasticity in the brain.

Finally, we emphasize the extreme simplicity of the concepts behind FCS. Though it goes beyond viewing genes as isolated entities, it falls short of capturing the full range of information that could be brought to bear to identify biologically meaningful patterns. More sophisticated ways are needed for incorporating other data and prior knowledge about gene relationships such as transcriptional regulatory and protein–protein interactions.

ACKNOWLEDGMENTS

We thank Shahmil Merchant and Edward Chen for programming support and Gordon Barr, Michael Barnes, and Puhong Gao for valuable discussion. We are also grateful to Loubna Errazi-Benchekroun, Peggy Smyrniotopoulos, and Hanga Galfalvy for related work on aspects of the genomic analysis of human prefrontal cortex. This work was supported by a NIMH Conte Center for the Neuroscience of Mental Disorders grant (No. MH62185 to J. J. M. and V. A.), a NIMH grant (No. F32MH63559 to E. S., No. K01MH067721 to ES, No. MH40210 to V. A.), and by the American Foundation for Suicide Prevention (ES).

REFERENCES

1. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25:25–29.
2. Kim, C. C. and Falkow, S. 2003. Significance analysis of lexical bias in microarray data. BMC Bioinformatics 4:12.
3. 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., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4:R28.
4. 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.
5. Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. and Conklin, B. R. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 4:R7.
6. Blalock, E. M., Chen, K. C., Sharrow, K., Herman, J. P., Porter, N. M., Foster, T. C., and Landfield, P. W. 2003. Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. J. Neurosci. 23:3807–3819.
7. Pavlidis, P., Lewis, D. P., and Noble, W. S. 2002. Exploring gene expression data with class scores. Pac. Symp. Biocomput. 474–485.
8. Middleton, F. A., Mirnics, K., Pierri, J. N., Lewis, D. A. and Levitt, P. 2002. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. J. Neurosci. 22:2718–2729.
9. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31:e15.
10. Dudoit, S., Gentleman, R. C., and Quackenbush, J. 2003. Open source software for the analysis of microarray data. Biotechniques (Suppl.):45–51.
11. Zar, J. H. 1999. Biostatistical analysis. Prentice Hall, Upper Saddle River, NJ.
12. Efron, B. and Tibshirani, R. 1993. An introduction to the bootstrap. Chapman & Hall, New York.
13. Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57:289–300.
14. Tanaka, M. and Cyong, J. C. 1985. The development of complement activating ability as an age related factor in murine brains. Microbiol. Immunol. 29:1219–1227.
15. Terao, A., Apte-Deshpande, A., Dousman, L., Morairty, S., Eynon, B. P., Kilduff, T. S., and Freund, Y. R. 2002. Immune response gene expression increases in the aging murine hippocampus. J. Neuroimmunol. 132:99–112.
16. Lee, C. K., Weindruch, R. and Prolla, T. A. 2000. Gene-expression profile of the ageing brain in mice. Nat. Genet. 25:294–297.
17. Jiang, C. H., Tsien, J. Z., Schultz, P. G., and Hu, Y. 2001. The effects of aging on gene expression in the hypothalamus and cortex of mice. Proc. Natl. Acad. Sci. USA 98:1930–1934.
18. Mocchegiani, E., Giacconi, R., Cipriano, C., Muzzioli, M., Fattoretti, P., Bertoni-Freddari, C., Isani, G., Zambenedetti, P., and Zatta, P.

2001. Zinc-bound metallothioneins as potential biological markers of ageing. Brain Res. Bull. 55:147–153.
19. Miyazaki, I., Asanuma, M., Higashi, Y., Sogawa, C. A., Tanaka, K., and Ogawa, N. 2002. Age-related changes in expression of metallothionein-III in rat brain. Neurosci. Res. 43:323–333.
20. Jakt, L. M., Cao, L., Cheah, K. S., and Smith, D. K. 2001. Assessing clusters and motifs from gene expression data. Genome Res. 11:112–123.
21. Westfall, P. H. and Young, S. S. 1993. Resampling-based multiple testing. John Wiley & Sons, New York.
22. Yekutieli, D. and Benjamini, Y. 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 82: 171–196.
23. Marin-Padilla, M. 1988. Embrionic vascularization of the cerebral cortex. Pages 479–510, in Peters, A. and Jones, E. G. (eds.), Cerebral cortex, vol. 7. Plenum, New York.
24. Chang, D. F., Belaguli, N. S., Iyer, D., Roberts, W. B., Wu, S. P., Dong, X. R., Marx, J. G., Moore, M. S., Beckerle, M. C., Majesky, M. W., and Schwartz, R. J. 2003. Cysteine-rich LIM-only proteins CRP1 and CRP2 are potent smooth muscle differentiation cofactors. Dev. Cell. 4:107–118.
25. Freeman, W. M., Brebner, K., Patel, K. M., Lynch, W. J., Roberts, D. C., and Vrana, K. E. 2002. Repeated cocaine self-administration causes multiple changes in rat frontal cortex gene expression. Neurochem. Res. 27:1181–1192.
26. Breen, M. A. and Ashcroft, S. J. 1997. A truncated isoform of Ca2+/calmodulin-dependent protein kinase II expressed in human islets of Langerhans may result from trans-splicing. FEBS Lett. 409:375–379.
27. Erraji-BenChekroun, L., Underwood, M. D., Arango, V., Galfalvy, H., Pavlidis, P., Smyrniotopoulos, P., Mann, J. J., and Sibille, E. (submitted) Active, continuous and extensive molecular aging in human prefrontal cortex.