Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables

Matthew A. Zapala* and Nicholas J. Schork*†‡

*Biomedical Sciences Graduate Program and the Polymorphism Research Laboratory, Department of Psychiatry, and †Division of Biostatistics, Department of Family and Preventive Medicine, Moores UCSD Cancer Center, Center for Human Genetics and Genomics, and the California Institute of Telecommunications and Information Technology, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Communicated by Dennis A. Carson, University of California at San Diego School of Medicine, La Jolla, CA, October 25, 2006 (received for review July 25, 2006)

A fundamental step in the analysis of gene expression and other high-dimensional genomic data is the calculation of the similarity or distance between pairs of individual samples in a study. If one has collected *N* total samples and assayed the expression level of *G* genes on those samples, then an $N \times N$ similarity matrix can be formed that reflects the correlation or similarity of the samples with respect to the expression values over the *G* genes. This matrix can then be examined for patterns via standard data reduction and cluster analysis techniques. We consider an alternative to conventional data reduction and cluster analyses of similarity matrices that is rooted in traditional linear models. This analysis method allows predictor variables collected on the samples to be related to variation in the pairwise similarity/distance values reflected in the matrix. The proposed multivariate method avoids the need for reducing the dimensions of a similarity matrix, can be used to assess relationships between the genes used to construct the matrix and additional information collected on the samples under study, and can be used to analyze individual genes or groups of genes identified in different ways. The technique can be used with any high-dimensional assay or data type and is ideally suited for testing subsets of genes defined by their participation in a biochemical pathway or other *a priori* grouping. We showcase the methodology using three published gene expression data sets.

analysis of variance | high-dimensional data

The introduction of high-throughput technologies, such as DNA microarrays and proteomics platforms, has provided researchers with a set of assays that are unprecedented in their sophistication. These technologies allow researchers to interrogate the expression levels of thousands to tens-of-thousands of genes or proteins simultaneously (1, 2). Although of tremendous importance, the use of these technologies is plagued by the fact that they generate enormous amounts of data, whose significance, both statistically and biologically, can be difficult to fathom in any single experiment (3). In essence, the collection of expression levels on thousands of genes on relatively few individuals or other units of observation, such as cells or cell types, creates enormous potential for false positive results when each gene is analyzed in isolation (4).

Many clever and useful data analysis strategies for the assessment of gene expression and related high-dimensional genomic data have been proposed (5). The vast majority of these strategies rely on either some form of data reduction, such as cluster analysis (6), or eigenstructure analysis (7, 8), which raises a number of questions about the appropriateness of the cluster method used, the number of clusters or eigenvalue/eigenvector pairs seen as "optimal," appropriate or statistically significant, as well as the biological meaning of the clusters or eigenvectors that emerge (9). Despite this fact, one common and appropriate strategy exploited by a number of data analysis approaches, which is in fact a precursor and fundamental construct to many contemporary gene expression analysis methods, involves the construction of a similarity or distance matrix, which reflects the similarity/dissimilarity of each pair of individuals with respect to the gene expression values obtained on them. This strategy was outlined in many of the earliest proposed gene expression analysis methods (6, 10–12), has become a standard tool for gene expression data analysis and visualization tools (13, 14), and is, in fact, even a typical ingredient in cluster and eigenstructure analyses.

We describe a method for testing the relationship between variation in a distance matrix and predictor information collected on the samples whose gene expression levels have been used to construct the matrix. The method provides a formal test of the organization of a similarity or distance matrix as it relates to predictor variable information collected on the individual samples, such as clinical parameters on subjects whose tumors have been evaluated for gene expression or genotype data of different inbred mouse strains assayed for gene expression. As a result, the method is the perfect companion for heat map and tree-based representations of high-dimensional data organized by some feature or *a priori* grouping factor meant to graphically represent and reveal a relationship between the genes used to construct the heat map or tree and these features or groupings. By testing more global hypotheses about the patterns within a similarity or distance matrix, the procedure avoids the need for cluster analysis and is very appropriate for situations where the number of data points collected is much larger than the number of samples or individuals. We first describe the derivation of the method, and then showcase its application to three publicly available data sets. We also want to emphasize that the procedure can be used to study any number of groups of genes, including single genes or all of the genes in a data set, making it very flexible and a method that only adds to existing univariate approaches.

Methods

Basic Model. In describing the proposed analysis methodologies, we follow the notation in McArdle and Anderson (15). We do not focus on many of the alternative methodologies for distance-based analyses developed by Krzanowski (16), Gower and Krzanowski (17), Legendre and Anderson (18), and Gower and Legendre (19), although many of these techniques may have some merit in the analysis of genomic data. Note that we used boldface to indicate matrices or vectors in our notation. Let $\mathbf{Y}$ be

Author contributions: M.A.Z. and N.J.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

‡To whom correspondence should be addressed. E-mail: nschork@ucsd.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0609333103/DC1.

an $N \times P$ matrix harboring gene expression values on N subjects for P genes. Let $\mathbf{X}$ be an $N \times M$ matrix harboring information on M predictor or regressor variables whose relationship to the gene expression values is of interest, where the first column contains a column vector whose every element is 1, and reflects an intercept term, as in standard regression contexts. These predictor variables could include the ages of individuals assayed, clinical diagnoses, strain memberships, cell line types, or genotype information. A standard multivariate multiple regression model for this situation would be (20, 21)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad [1]$$

where $\boldsymbol{\beta}$ is an $M \times P$ matrix of regression coefficients and ε is an error term, often thought be distributed as a (multivariate) normal vector. The least-squares solution for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, with the matrix of residual errors for the model being

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}_{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}, \qquad [2]$$

where $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and is the traditional "hat" matrix. Unfortunately, If $N \ll P$, as is often the case with gene expression and other genomic data types, then this model is problematic. An alternative would consider how the M predictor variables relate to the similarity or dissimilarity of the subjects under study with respect to the P gene expression values as a whole or as a series of unique subsets of the data.

Let $\mathbf{D}$ be an $N \times N$ distance matrix, whose elements, d_{ij}, reflect the distance (or dissimilarity) of subjects i and j with respect to the P gene expression values. For example, d_{ij} could be calculated as the Euclidean distance or as a function of the correlation coefficient (see *Forming the Distance Matrix* below). Let $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$. One can form Gower's centered matrix $\mathbf{G}$ from $\mathbf{A}$ by calculating

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)A\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right), \qquad [3]$$

where $\mathbf{1}$ is a N-dimensional column vector whose every element is 1 and $\mathbf{I}$ is an $N \times N$ identity matrix. An appropriate F statistic for assessing the relationship between the M predictor variables and variation in the dissimilarities among the N subjects with respect to the P variables is

$$F = \frac{tr(\mathbf{HGH})/(M-1)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(N-M)}, \qquad [4]$$

where $\mathbf{H}$ is a hat matrix, $\mathbf{G}$ is Gower's centered matrix, and $\mathbf{I}$ is the identity matrix, formed as above. M is scalar and reflects the number of predictors and N is the number of subjects. If $P = 1$ (i.e., a univariate analysis) and the distance matrix is computed through the use of the standard Euclidean distance measure, then F in Eq. **4** is the standard F statistic and possesses the typical properties associated with F statistics in ANOVA contexts. This result is due to the fact that the inner product matrix $(Y'Y)$ used in standard univariate analysis of variance and regression contexts contains the same information, in terms of total sums-of-squares, as the outer product matrix (YY'), which reflects interpoint squared differences or distances $(tr(Y'Y) = tr(YY'))$ (15). When different distance measures are used, the properties of F are more complicated, suggesting the use of alternative methods for assessing statistical significance (see *Assessing Statistical Significance* below).

Forming the Distance Matrix. The formation of the distance matrix is an important step in the use of the proposed procedure. There is a bewildering array of potential distance measures one could use with the proposed method (22). The correlation coefficient, r, is often used to assess the similarity between two individuals based on gene expression values (14). A correlation matrix with elements r_{ij} can be converted to a distance matrix with elements d_{ij} easily enough through a simple transformation

$$d_{ij} = \sqrt{2(1 - r_{ij})}. \qquad [5]$$

This transformation leads to a distance matrix with metric properties, although distance measures with nonmetric properties can be used in the analysis method described as well (17). We discuss aspects of the choice of a distance measure in *Results*, but more work in this area is needed. One additional aspect of the formation of a distance matrix that deserves attention involves handling missing data. Intuitively, if one has collected thousands of gene expression values to be used to create distance profiles, then a few missing observations are not likely to have much of an influence. For example, one could simply not use these genes in the formation of the distance matrix, ignore the missing values only when assessing pair-wise distance for a pair of observations with missing data, or assign individuals with missing data imputed values that are then used to compute distance. However, the delineation of a threshold beyond which the number of missing values creates problems for a distance-based analysis is important and worthy of further research.

Assessing Statistical Significance. The distribution of the F statistic defined in Eq. **4** is complicated and its derivation for any particular distance matrix is unlikely to generalize to other distance matrices, especially with small sample sizes. Therefore, one can rely on permutation tests to evaluate the probabilistic significance of an observed F statistic computed from Eq. **4** (23–25). Permutations can either involve permuting the raw data or simultaneously permuting the rows and columns of the $\mathbf{G}$ matrix, as is done in Mantel's matrix correspondence test (15). In addition, if permutation tests are used, the degrees-of-freedom terms in the numerator $(M - 1)$ and the denominator $(N - M)$ are not required in the formulation of the statistic presented in Eq. **4**. Finally, given that different predictor variables, or subsets of variables, can be tested for association with variation in a distance matrix, one can pursue step-wise or variable selection procedures with the technique, identical to univariate standard multiple regression analysis (26). Beyond a P value, an estimation of the proportion of variation within the matrix that is explained by a particular set of M predictor variables can be calculated by dividing $tr(\mathbf{HGH})$ (i.e., the sum of the diagonal elements of a matrix) by $tr(\mathbf{G})$. In our analyses, independent variables are tested both individually and in a forward stepwise manner. The independent variables selected for the model in a stepwise manner are based on the highest cumulative proportion of variance that is explained by the inclusion of an additional variable in the regression model. An F statistic and P value are calculated for the addition of this variable to the model.

Assessing Level Accuracy and Power of the Proposed Hypothesis Test. To examine properties of the proposed analysis procedure, a series of studies investigating the level accuracy and power of the proposed test statistics were performed. To examine the level accuracy of the test, we simulated 30 samples each measured on 100 variables. The variables were assumed to follow a standard normal distribution; hence, there was no structure to the data. We assumed that the first 15 samples had a different origin than the second 15. We then tested the relationship between this grouping factor (coded as 0 for the first 15 samples and 1 for the second 15 samples) and the distance between the samples calculated with different distance measures using the proposed procedure with 1,000 random permutations of the data. We

Table 1. Level accuracy of the proposed permutation test

Distance metric	% of tests $P \leq 0.01$	% of tests $P \leq 0.05$	% of tests $P \leq 0.25$	% of tests $P \leq 0.50$
Pear.	1.3	4.8	26.5	51.4
Spear.	1.5	4.6	27.4	52.9
Kend.	1.3	4.9	23.9	47.9
Conc.	1.3	4.8	26.7	52
Eucl.	1.2	6.1	24.7	49.1
Cheb.	1.3	5.9	25.4	48.6

P values were calculated using the proposed method based on 1,000 permutations of the data. The simulations were repeated 1,000 times and the percentage of *P* values below a certain threshold is reported for each of the following metrics: Pearson correlation (Pear.), Spearman rank (Spear.), Kendall Tau (Kend.), Lin's concordance correlation (Conc.), Euclidean distance (Eucl.), and Chebychev distance (Cheb).

repeated this process 1,000 times. Table 1 describes the results and clearly shows that the nominal level of the test matches closely with the simulation results, suggesting that the proposed test procedure is nonbiased.

We also considered the power of the proposed test. We simulated data for 30 samples and 100 variables in which 15 samples were assigned to a hypothetical control group (independent variable = 0) and 15 samples were assigned to a hypothetical experimental group (independent variable = 1). Data in the control group were generated as standard normal variables with a mean of 0 and variance 1. Data in the experimental group were generated as standard normal variables with variance = 1 and means that took on values of 0 to 1.5 in increments of 0.001. The power of the proposed permutation-based statistical test was then investigated in these settings. In this context we also generated different simulated data sets for which 100%, 50%, 25%, 10%, or 5% of the variables used in the construction of the distance matrix had means adjusted from 0 (in the appropriate increments) in the experimental group. Fig. 1 describes the results. Note that the gray line in Fig. 1 represents the power curve obtained based on a *t* test with the Bonferroni correction, corrected for 100 multiple comparisons. Fig. 1 clearly shows that the proposed procedure can detect "signals" in the data as long as the number of variables contributing to that signal used in the construction of the distance matrix is moderate.

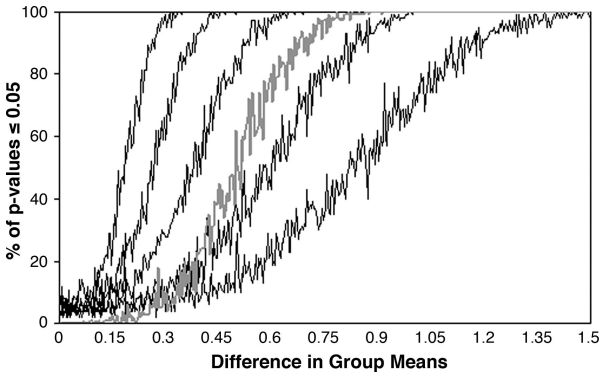

Fig. 1. Power of the proposed distance matrix-based regression procedure as a function of both increasing differences in control vs. experimental settings as well as overall signal-to-noise ratio-based simulated data sets (see text for details).

Results

The proposed method was tested on three different published data sets to display its utility. We briefly consider some of the implementation details and properties of the proposed technique, such as the need for evaluating the distance between the observations, and the dependence of the test statistic on subsets of genes among all those used to derive a distance measure. We note that for the following applications we used the correlation coefficient to derive the distance measure, as this measure has been the standard for gene expression data (14). In addition, we used 1,000 permutations to compute *P* values.

Embryonic Imprint of the Adult Mouse Brain. The first data set involved gene expression data from multiple brain regions and multiple inbred mouse strains (27). The normalized data can be downloaded from the Gene Expression Omnibus (GEO) using record number GSE3594. The authors had three hypotheses about the relationships of the gene expression patterns between the different brain regions in the adult mouse. The gene expression patterns of these brain regions could be related to each other based on adult anatomy, evolutionary relationships, or embryonic origin. The authors performed hierarchical cluster analysis and created a Pearson correlation heat map matrix where they hypothesized that the gene expression patterns of the adult mouse brain bear an imprint based on the adult tissue's embryological origin. The heat map and the hierarchical tree constructed based on the similarity of gene expression patterns across all of the genes suggest that adult structures are related to each other based on the classic five vesicle embryonic neural regions [telencephalon, diencephalon, mesencephalon, metencephalon, and myelencephalon; see supporting information (SI) Fig. 3). Using the proposed regression analysis procedure, we provide statistical evidence that embryonic origin is the most likely hypothesis of the three because it explains the largest proportion of variation in the similarity of the overall gene expression profile of the brain regions ($P < 0.001$ and proportion of variation in pair-wise distances explained by embryological origin = 0.33, adult anatomy = 0.26, and evolutionary relationships = 0.19). The authors also suggested that anterior-to-posterior (A/P) position along the neural tube could dictate expression patterns in the adult neural structures. The position along the neural tube was tested individually and in combination with embryological origin. Importantly, A/P position added a significant proportion of explained variation in brain region gene expression similarity over and above embryological origin ($P < 0.001$ and proportion of variation explained above embryological origin = 0.032).

Aging Human Brain. The second data set examined gene expression patterns in the human frontal cortex among individuals who died at various ages (26–106 years) (28). The normalized data can be downloaded from GEO using record number GSE1572. The authors performed Spearman rank correlations to determine 463 genes that correlated with age ($P < 0.005$) and a Pearson correlation-based heat map matrix was then calculated that covered all pair-wise comparisons of individuals. We analyzed the distance matrix based on the pair wise correlations (Eq. **5**) using the proposed regression method to quantify the effect that age and sex may have on the gene expression patterns for the 463 genes found to be correlated with age (age $P < 0.001$ and proportion of variation explained = 0.35; sex $P = 0.224$). Although sex was not a significant predictor of the gene expression patterns in the frontal cortex, age appeared to explain ≈35% of the variation in the similarity in the gene expression patterns among the individuals based on the age-related genes (see SI Fig. 4*A*). Moreover, the association with age was not only apparent in age-related genes (as identified from Spearman rank correlations), but was also evident in the correlation matrix created using all of the genes scored as "Present" in at least one of the samples (age $P < 0.001$ and proportion of variation explained = 0.16; sex $P = 0.78$) (see SI Fig. 4*B*). Therefore, it

appears that the age effect is pronounced enough to be significant even when including all detectable genes on the array (8,507 probe sets).

As emphasized, the proposed regression technique can be applied to each gene in a univariate manner. Univariate analysis can be used to identify a set of genes that, when considered in the construction of a distance matrix, are strongly related to a specific predictor variable. For example, we wanted to identify a set of age-associated genes that would explain a larger proportion of variation than the set that Lu *et al.* (28) identified by using Spearman rank correlations. Using the matrix regression technique for each individual gene (using the Euclidean distance), we calculated an F statistic that, as noted when Euclidean distances are used with a single variable, is identical to an ANOVA statistic (15). The resulting F statistics were then used to rank genes and identify those which demonstrated the largest age association (Fig. 2 *A* and *B*). These ranked genes were then serially used to construct matrices tested with our proposed procedure to identify an optimal set of genes that resulted in the largest proportion of variation explained by age effects (Fig. 2 *C* and *D*). An optimal set was found to occur with the top 100 ranking genes, as age explained 52% of the variation in dissimilarity within the matrix constructed with these genes, which is much higher than the 35% explained with the age-related genes chosen by Lu *et al.* (28). Of the 100 genes identified as an optimal set, 80 were within the Lu *et al.* set. Even using an equivalent number of genes as Lu *et al.* used, the proportion of variation explained by our highest ranked 463 genes is 0.42 compared with the Lu *et al.* proportion of 0.35. Of our 463 highest ranked genes, only 256 are also present in the Lu *et al.* study. Interestingly, among the 20 genes within the top 100 genes identified by our analysis that were not identified in the Lu *et al.* study were two genes involved specifically with neurological function, mitogen-activated protein kinase kinase 1 (*MAP2K1*) and amyloid beta (A4) precursor protein-binding, family A, member 1 (*APBA1*). *MAP2K1* is known to control apoptosis signaling specifically in astrocytes (29) and to regulate *MAPK1*, which was found in the original Lu *et al.* study and is also involved in synaptic transmission. *APBA1* is a neuronal adaptor protein that interacts with, stabilizes and inhibits the proteolysis of the Alzheimer's disease amyloid precursor protein (*APP*) (30). A large number of genes involved in neurological disorders are present in the 207 genes of our expanded gene list which were not identified in the original 463 aging genes identified by Lu *et al.* These include interesting genes implicated in age-specific neurological diseases, such as *CDK5*, *CDK5R1*, *BCL2L1*, and *PLAT* (30–33). The full gene list is found in SI Table 2.

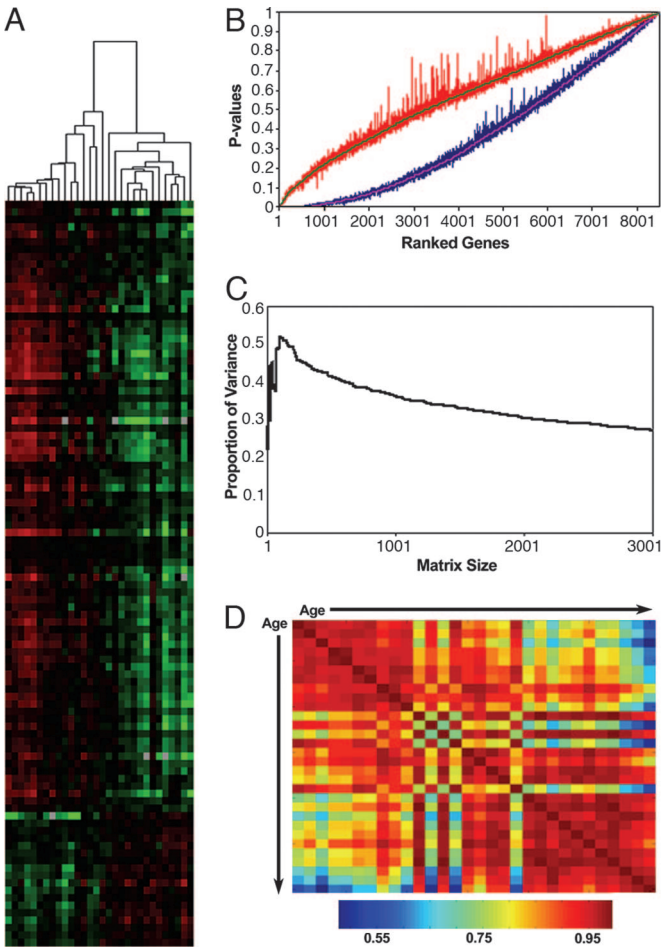

Fig. 2. Use of univariate regressions to identify optimal set of genes for multivariate matrix regression and hierarchical clustering. (*A*) Hierarchical cluster of top 100 genes identified from univariate regression of single genes. The dendrogram above the heat map shows the samples ordered from youngest to oldest (from left to right) with the leftmost branches of the tree connecting data on individuals with an average age of 43, whereas the rightmost branches connect data on individuals with an average age of 80 (*t* test, $P < 0.000001$). Hierarchical clustering was performed by using CLUSTER with the average linkage metric and displayed with JAVATREEVIEW. (*B*) Plot of permuted *P* values of age effects (purple) and sex effects (red) as well as the *P* values obtained from standard ANOVA *F* statistic for age (pink) and sex (green) effects. (*C*) Genes were ranked by the age *F* statistic and then grouped together to identify the set of highly significant age-related genes (here found to be 100 genes) whose expression levels would produce a sample-based distance matrix such that the variation in its elements could be explained maximally by age effects. (*D*) A heat map matrix of the optimal set of 100 highly significant age-related genes for which age explains ≈52% of the variation in dissimilarity across the individuals based on these genes' expression values.

Aging Human Kidney. The third data set considered gene expression in human kidney tissue across two regions (cortex and medulla) in multiple patients of different ages (34). The normalized data can be downloaded from the Stanford Microarray Database. In addition to the gene expression data, there were numerous clinical parameters available on a majority of the patients. The clinical parameters included indicators of renal pathology, such as the degree of glomerular sclerosis or arterial intimal hyalinosis (AIH). There was also information about creatinine levels, medical history, and systolic and diastolic blood pressures. All of this information was used as independent variables to predict gene expression patterns. We restricted our analysis to patient samples for which there was clinical data available for all of the parameters. This limited the analysis to 63 samples. First, we analyzed the Pearson correlation derived distance matrix (Eq. **5**) for genes that had to be scored as "Present" in at least one of the samples. Three variables were found to significantly contribute to the variation in the gene expression similarity/dissimilarity: tissue type (cortex or medulla $P < 0.001$; proportion = 0.18), AIH ($P < 0.001$, proportion = 0.05), and past medical history ($P = 0.033$, proportion = 0.03). Past medical history included a history of hypertension, diabetes mellitus, chronic renal insufficiency, hepatitis B virus, hepatitis C virus, or combinations of those diseases. Next, we analyzed the distance matrix computed from only the age associated genes that the authors had identified by using a linear regression model (985 genes). Interestingly, age was not a predictor of the gene expression pattern in the age associated gene expression correlation matrix. The most significant predictors were AIH ($P < 0.001$; proportion = 0.15), tissue type ($P < 0.002$; proportion = 0.08), and race ($P < 0.001$ proportion = 0.07) (see SI Fig. 5). Age was insignificant with a P value of 0.63 and a contribution of

<0.01. However, the chronicity index, which was an index developed by the authors that scores the morphological and physiological state of the kidney and was designed to give a physiological age to the kidney, was almost significant with a P value of 0.055 and proportion of variation explained of 0.02. Thus, collinearity among chronicity index and other independent variables may have prevented it from entering into the final model as a significant predictor. When we tested the independent variables individually, the chronicity index was significant ($P < 0.001$; proportion = 0.15).

Beyond testing whole sets of genes, the method can test specific subsets of genes for which it may be hypothesized that gene expression is specifically altered. For example, one may be interested in whether genes involved in the Pharm-GKB derived ACE-inhibitor pathway show altered gene expression patterns consistent with a specific form of renal pathology. Testing all of the ACE-inhibitor pathway genes using the proposed procedure, we discovered that not only are there large tissue differences between the cortex and medulla of the kidney in the ACE-inhibitor pathway ($P < 0.001$, proportion = 0.12), but there is a significant association above tissue differences in regards to the patient's level of tubular atrophy/interstitial fibrosis ($P < 0.007$, proportion = 0.08, cumulative proportion = 0.20).

Evaluating Different Distance Measures. We considered the effect of the use of different distance measures on the tests for association. Although not an exhaustive study, we present this to showcase the importance of choosing a distance matrix. The choice of a distance matrix is important in a number of related contexts, such as the choice of a distance matrix for graphically representing data in heat maps or tree diagrams or in cluster analysis settings (35–37). We reevaluated the associations involving the above data sets using the Pearson correlation coefficient, the Spearman correlation, the Kendall Tau correlation, Lin's concordance correlation, the Euclidean distance, and the Chebychev distance to derive the distance matrix (see SI Table 3). This analysis considered the distance matrices constructed from the same genes used in the analyses above. Lin's concordance correlation, the Euclidean distance and the Chebychev distance emphasize the actual proximity of the numerical values of the genes used to compute the distance matrix, and hence stand in contrast to the correlation coefficient which merely considers the linear relationship of the values across the genes used (38, 39). The choice of a distance measure influences the proportion of variation in the distance matrix explained but not necessarily the significance of the relationship between the predictor variables and the distance matrix entries. A more thorough evaluation of this issue is required.

Signal Strength and Distance Matrix. Because it is unlikely that all of the genes considered in a study will be related to a particular predictor variable, the formation of a distance matrix with all of the genes may not show a signal, or as strong a signal, with the predictor variable as a distance matrix constructed with only those genes that are relevant to the predictor variable. Unfortunately, it will be difficult to know *a priori* which genes should go into the construction of the distance matrix. Although our procedure can be used to test each gene individually, or subsets of genes, as noted, we have also considered the more "omnibus" hypothesis testing situation in which one is interested in knowing whether there is any relationship between a predictor variable and gene expression patterns as a whole or across all genes assayed in a study. We were therefore interested in determining how strong the relationship between gene expression similarity and predictor variables considered in our examples is as a function of the number of genes considered in the construction of the distance matrix. This would provide us with insight into the amount of "noise" that could be tolerated and still allow the "signal" relating the gene expression values and the predictor variable to appear. We therefore considered the inclusion of random, simulated gene expression values in the construction of the distance matrix, knowing that these random simulated gene expression values would saturate the signal if enough were added. SI Fig. 6 shows the relationship between the F statistic, the proportion of variation in similarity/dissimilarity explained, and the permutation P value as a function of the number of extraneous gene expression values that are added in the construction of the distance matrix for all data sets tested. Large amounts of noise reduce the overall proportion of variation in similarity/dissimilarity explained as well as the F statistic, as one would expect; however, the permutation test derived P values remain significant. Thus, it takes the addition of $\approx$98% noise to saturate the signal to the point of statistical insignificance.

Discussion

Our proposed method of analysis can easily complement many traditional and alternative methods of analysis for high-dimensional data. In fact, because the proposed procedure can be used to analyze each gene in a univariate manner, it extends traditional univariate procedures. In addition, unlike other approaches, the proposed approach does not require a reduction of the data via principal components (40), cluster (6), factor (41), or multidimensional scaling analysis (42). The proposed analysis procedure also differs from related procedures, such as GSEA and globalTest (43, 44), in that it can be used to emphasize the multivariate nature of the expression values of many genes in the same pathway and treats the system being interrogated as a whole and does not consider each individual gene in a univariate analysis which then considers the result of the univariate analyses in aggregate. The exploitation of this fact can have disadvantages, obviously, because one may be interested in knowing which particular sets of genes are the most perturbed in a particular setting. However, it is arguable that physiological perturbations and variations are likely to "re-set" the coordinated expression patterns of many genes to reach biochemical or physiological homeostasis or equilibrium. Thus, the assessment of the similarity of global gene expression profiles of multiple samples with different features or exposures is appropriate.

Depending on the number of data points that are selected for analysis, it is possible to over fit the regression and identify significant predictor variables whose effect could be assigned to a large number of data points, when in fact only a smaller subset of the data points is truly associated with the predictor variable. However, because the multivariate regression technique can be reduced to a univariate analysis that focuses on single data points, it is possible to identify specific subsets of the data within a larger group for which the predictor variable is having the largest significant effect. The method we have proposed is, in fact, flexible enough to be used in settings for which insight into the effects of single genes or subsets of genes is the goal. Alternatively, one could test subsets or groups of genes based on some (*a priori*) grouping factor, such as participation in a biochemical pathway or genetic network. One could also combine the proposed approach with standard non-distance-based univariate and/or cluster analysis methods to assess the significance of groups of genes identified with these methods with respect to a predictor variable. Finally, beyond testing the relevance of specific clinical or phenotypic predictor variables, the method can be used as a quality control measure to identify potential sources of nonbiological error, such as technician, chip lot or dissection error. These sources of nonbiological variation can be included in the multiple regression as additional independent variables and thus these factors can be controlled for in the analysis.

There are some limitations to the proposed method that go beyond the choice of a distance metric or the manner in which

individual genes or groups of genes are tested. For example, it may be the case that the actual correlation patterns among genes differ across particular groups or levels of a quantitative predictor variable. In fact, differences in correlations among the expression levels of genes across, e.g., individuals treated with different drugs, different strains of mice, older and younger individuals, etc., may reflect actual perturbations in genetic networks, possibly more so than simple differences in the achieved levels of gene expression themselves. The assumption that a single distance metric, and hence distance matrix, characterizes similarities among individuals with respect to gene expression values may be obscuring important differences among the individuals, although the degree to which this is the case is an open question. There are methods for assessing "heteroscedasticity," differences in the covariances among groups of genes, across groups of individuals (refs. 20 and 45–47; see also http://polymorphism.ucsd.edu/mama), but their application to high-dimensional data has not been pursued to the degree that analyses considering differences in average levels of expression have.

The source code for this statistical method is freely available at the Biopython script central page (http://biopython.org/wiki/Scriptcentral) and is being incorporated into the Biopython library. Also, the source code and a user-friendly web application are being incorporated and maintained on the Schork Laboratory web site (http://polymorphism.ucsd.edu/programs.html).

We thank Ondrej Libiger for assistance with coding the program, Charles Abney for assistance in web development, and Dr. Marti Anderson for advice and encouragement. N.J.S. is supported in part by the National Heart, Lung, and Blood Institute Family Blood Pressure Program (FBPP; U01 HL064777-06); the National Institute on Aging Longevity Consortium (U19 AG023122-01); the National Institute on Mental Health Consortium on the Genetics of Schizophrenia (COGS; 5 R01 HLMH065571-02); National Institutes of Health Grants R01 HL074730-02 and HL070137-01; the UCSD Moores Cancer Center, and the Donald W. Reynolds Foundation (Helen Hobbs, Principal Investigator).

1. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) *Nat Biotechnol* 14:1675–1680.
2. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) *Nat Biotechnol* 17:994–999.
3. Bassett DE, Jr, Eisen MB, Boguski MS (1999) *Nat Genet* 21:51–55.
4. Storey JD, Tibshirani R (2003) *Proc Natl Acad Sci USA* 100:9440–9445.
5. Allison DB, Cui X, Page GP, Sabripour M (2006) *Nat Rev Genet* 7:55–65.
6. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) *Proc Natl Acad Sci USA* 95:14863–14868.
7. Alter O, Brown PO, Botstein D (2000) *Proc Natl Acad Sci USA* 97:10101–10106.
8. Alter O, Brown PO, Botstein D (2003) *Proc Natl Acad Sci USA* 100:3351–3356.
9. Bryan J (2004) *J Multivariate Analysis* 90:44–66.
10. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) *Mol Biol Cell* 9:3273–3297.
11. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) *Proc Natl Acad Sci USA* 96:6745–6750.
12. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) *Science* 286:531–537.
13. Slonim DK (2002) *Nat Genet* 32 (Suppl):502–508.
14. D'Haeseleer P (2005) *Nat Biotechnol* 23:1499–1501.
15. McArdle BH, Anderson MJ (2001) *Ecology* 82:290–297.
16. Krzanowski WJ (2002) *J Agric Biol Environ Stat* 7:222–232.
17. Gower JC, Krzanowski WJ (1999) *Appl Stat* 48:505–519.
18. Legendre P, Anderson MJ (1999) *Ecol Monogr* 69:1–24.
19. Gower JC, Legendre P (1986) *J Classification* 3:1–48.
20. Johnson RA, Wichern DW (1992) *Applied Multivariate Statistical Analysis* (Prentice–Hall, Princeton).
21. Anderson MJ (2001) *Aust Ecol* 26:32–46.
22. Webb AR (2002) *Statistical Pattern Recognition* (Wiley, Chichester, UK).
23. Edgington ES (1995) *Randomization Tests* (Marcel Dekker, New York).
24. Manly B (1997) *Randomization, Bootstrap, and Monte Carlo Methods in Biology* (Chapman and Hall, London).
25. Good PI (2000) *Permutation Tests* (Springer, New York).
26. Neter J, Wasserman W, Kutner MH (1985) *Applied Linear Statistical Models* (Richard D. Irwin, Inc., Chicago).
27. Zapala MA, Hovatta I, Ellison JA, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, *et al.* (2005) *Proc Natl Acad Sci USA* 102:10357–10362.
28. Lu T, Pan Y, Kao SY, Li C, Kohane I, Chan J, Yankner BA (2004) *Nature* 429:883–891.
29. Gomez Del Pulgar T, De Ceballos ML, Guzman M, Velasco G (2002) *J Biol Chem* 277:36527–36533.
30. Jacobs EH, Williams RJ, Francis PT (2006) *Neuroscience* 138:511–522.
31. Rademakers R, Sleegers K, Theuns J, Van den Broeck M, Bel Kacem S, Nilsson LG, Adolfsson R, van Duijn CM, Van Broeckhoven C, Cruts M (2005) *Neurobiol Aging* 26:1145–1151.
32. Luetjens CM, Lankiewicz S, Bui NT, Krohn AJ, Poppe M, Prehn JH (2001) *Neuroscience* 102:139–150.
33. Melchor JP, Pawlak R, Strickland S (2003) *J Neurosci* 23:8867–8871.
34. Rodwell GE, Sonu R, Zahn JM, Lund J, Wilhelmy J, Wang L, Xiao W, Mindrinos M, Crane E, Segal E, *et al.* (2004) *PLoS Biol* 2:e427.
35. Hughes T, Hyun Y, Liberles DA (2004) *BMC Bioinformatics* 5:48.
36. Kibbey C, Calvet A (2005) *J Chem Inf Model* 45:523–532.
37. Trooskens G, De Beule D, Decouttere F, Van Criekinge W (2005) *Bioinformatics* 21:3801–3802.
38. Lin LI (1989) *Biometrics* 45:255–268.
39. Lin LI (2000) *Biometrics* 56:324–325.
40. Yeung KY, Ruzzo WL (2001) *Bioinformatics* 17:763–774.
41. Kustra R, Shioda R, Zhu M (2006) *BMC Bioinformatics* 7:216.
42. Taguchi YH, Oono Y (2005) *Bioinformatics* 21:730–740.
43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) *Proc Natl Acad Sci USA* 102:15545–15550.
44. Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC (2005) *Bioinformatics* 21:1950–1957.
45. Anderson TW (1984) *An Introduction to Multivariate Analysis* (Wiley, New York).
46. Krzanowski WJ (1993) *Statistics and Computing* 3:37–44.
47. Anderson MJ (2006) *Biometrics* 62:245–253.