8-31-2004

Significance Analysis of Time Course Microarray Experiments

John D. Storey
University of Washington, jstorey@u.washington.edu

Wenzhong Xiao
Stanford University

Jeffrey T. Leek
University of Washington

Ronald G. Tompkins
Massachusetts General Hospital

Ron W. Davis
Stanford University

Introduction

The identification of genes that show changes in expression between varying biological conditions is a common goal in microarray experiments [1]. Differential expression can be studied from a static or temporal viewpoint. In a static experiment the arrays are obtained irrespective of time, essentially taking a snapshot of gene expression. In a temporal experiment the arrays are collected over a time course, allowing one to capture the dynamic behavior of gene expression. A large amount of work has been done on the problem of identifying differentially expressed genes in static experiments [2–15]. Since the regulation of gene expression is a dynamic process, it is also important to identify and characterize changes in gene expression over time. Here we present a general statistical method that identifies genes that are differentially expressed over time. We apply the method to two recent studies that we have carried out on humans.

A number of clustering methods have been applied to time course microarray data, including hierarchical clustering [16, 17], principal components based clustering [18], Bayesian model-based clustering [19], and K-means clustering of curves [20, 21]. None of these clustering methods is directly applicable to identifying genes that show statistically significant changes in expression over time. The K-means clustering method has been modified to compare expression over time between two groups [22], but this method can only be applied to a few hundred genes at a time because of the computational cost of fitting a single model to all genes simultaneously [20,21]. This approach also requires that the statistical significance be calculated under the assumption that the clustering fit to one of the groups is true, which is nonstandard and potentially problematic. The method that is proposed here draws on ideas from the extensive statistical literature on time course data analysis [23, 24], particularly spline based methods [25–33]. It is applicable to both detecting changes in expression over time within a single biological group, and to detecting differences in the behavior of expression over time between two or more groups. Individuals may be sampled at multiple time points, or each time point may represent an independently sampled individual. The computational cost is not substantially greater than methods for static experiments, so there is no impeding limit on the number of genes that may be tested. We apply the method here to microarray studies that each measures about forty thousand genes.

A number of methods already exist [2, 3, 8–12] that are appropriate for identifying genes that are differentially expressed between two or more static biological conditions. These methods are designed to compare unordered categorical conditions, such as three different cancer tumor types [34] or two different treatments [9]. Therefore, one could apply these to a time course study by treating each time point as a different "biological condition", even though the inherent ordering and spacing provided by time points would be ignored. However, adapting these methods requires

the study to be perfectly balanced, that is, each time point must have the same number and type of observations. Also, they require that more than one observation be obtained at each time point; otherwise, a "biological condition" would have only one observation, which renders the existing methods useless. The method proposed here does not place either of these restrictions on the study design, and it takes into account the ordering and spacing information provided by the time points. It is even possible to conduct an analysis when every sample has been obtained at a unique time point.

No existing method for static differential expression could be straightforwardly applied to the studies we analyze here, even when treating the time points as categorical biological conditions. In one study, a control individual is missing two time points, thereby yielding an unbalanced design. One could try to impute these missing time points for every gene, but this would require a lot of modeling assumptions and unnecessary work. Imputing missing time points is not necessary in our proposed method. In the other study, there are many time points where only one array has been obtained. This unavoidable feature is not a problem since our proposed method borrows information across the time points by directly utilizing the time structure. However, one cannot adapt currently available methods to this study unless the data are binned in an arbitrary way so that a balanced experiment with repeated observations is fabricated.

As it turns out, when a balanced design exists with repeated observations at each time point, then several existing approaches can be formulated a special case of our procedure. That is, if one applies our method so that it uses every available degree of freedom, then the procedure is a type of classical ANOVA similar to those proposed for microarrays [3, 10, 11], which is also similar to modified versions of classical ANOVA such as SAM [9, 15]. However, if every degree of freedom is used, then the time structure has essentially been ignored. As long as a model using less degrees of freedom appropriately captures the signal across time, then standard statistical theory says that this model will be more powerful. Therefore, our method is more powerful whenever it utilizes the time structure thereby saving degrees of freedom. Since the proposed procedure can be viewed as a generalization of the existing methods for static experiments, it should be noted that many of the modifications to standard t-tests and F-tests that have been proposed (such as shrinking variance estimates [9]) can be applied to our method as well. However, most of these modifications are not yet completely justified or understood, so we defer such fine tuning.

The proposed method is applied to two recent studies that we have carried out on humans. These studies encompass both types of sampling (longitudinal and independent, discussed in detail below) and both types of differential expression over time (between groups and within a single group). In one study, gene expression was monitored over time in controls and in individuals treated with endotoxin, which is widely used to study acute inflammatory and immune response.

Our goal is to understand the mechanism of endotoxin response by identifying genes with expression that is different over time between the treated and untreated groups [35]. In a second study, we examined the effect of age on gene expression in the kidney, where samples were obtained from human subjects ranging in age from 27 to 92 years. Although genome-wide transcriptional changes associated with age have been studied in several model organisms [36–38], age-dependent expression in humans is unclear. The goal here was to identify genes whose expression changes significantly with respect to age in human kidney cortex tissue [39]. Genes are identified in both studies that corroborate previous findings and provide insights into these problems.

Materials and Methods

Sample preparations. Details on the protocols are described elsewhere [35, 39]. *Endotoxin study*: In order to monitor gene expression responses to bacterial endotoxin in blood leukocytes, eight adult volunteers were recruited by the Clinical Research Center at UMDNJ-Robert Wood Johnson Medical School. Four subjects were administered endotoxin and four were administered a placebo. Blood samples were collected before endotoxin infusion and at 2, 4, 6, 9, and 24 hours after infusion. The leukocytes were isolated from the blood samples using a modified lysis protocol [35]. Total RNA was extracted using an RNeasy kit (Qiagen, Inc). Samples from hours 4 and 6 were unavailable for one of the controls. *Kidney aging study*: To investigate changes in gene expression in the human kidney across different ages, samples were obtained from normal kidney tissues removed at nephrectomy for various medical reasons or renal transplant biopsy from 74 patients ranging in age from 27 to 92 years, and dissected into cortex and medulla sections based on histological evaluation. Each frozen tissue section was homogenized and total cellular RNA was isolated according to the TRIzol Reagent protocol. Comprehensive evaluations were performed and reported elsewhere on various medical factors of the patients [39], and it is unlikely that these factors have confounded age-regulated changes in gene expression [39, 40]. Only cortex samples (72 in total) were used in this study.

Microarray analysis. Total RNA was extracted, and messenger RNA amplified and hybridized onto Human U133A and U133B GeneChips according to the protocols recommended by Affymetrix (Santa Clara, CA). 44,924 probe sets on the arrays were analyzed. Normalization was performed using dChip, and expression levels were calculated using the perfect match only model [41]. Expression values were then transformed by taking $\log_2(\text{Data} + 10)$ where the relatively negligible number 10 was added to stabilize the variance of values close to zero.

Statistical and computational details. Exhaustive details (including formula derivations, algorithms and statistical justifications) can be found in the Supplementary Information.

Functional analysis of significant genes. Probe sets on Affymetrix U133 GeneChips were mapped to gene IDs (`http://www.ncbi.nlm.nih.gov/entrez/`, `http://www.affymetrix.com/analysis/netaffx/`). *Endotoxin study*: Among the 4163 probe sets significant at 0.1%, 2914 unique genes were identified from

3892 probe sets having mapped gene IDs (271 were unmapped). *Kidney study*: Among the 417 probe sets significant at 10%, 300 unique genes were identified from 364 probe sets with mapped gene IDs and 53 unmapped. In both studies, the Ingenuity^{TM} Pathways Knowledge Base (IPKB) was used for functional analysis of genes. Briefly, the IPKB consists of more than a million individually-modelled relationships into an ontology of more than 550,000 biological concepts. Relationships between genes, proteins, small molecules, complexes, cells, processes, and diseases were manually extracted by Ph.D.-level scientists from over 200,000 peer-reviewed articles [35].

Removal of probe sets in the kidney aging study. The human subjects observed in the kidney aging study did not represent a purely random sample. Because of this, any observed temporal differential expression could be confounded with unobserved variables. An initial analysis showed that a number of probe sets contained outliers possibly due to unobserved variables. (These probe sets will disproportionately be called significant because they violate the statistical assumptions needed to perform a valid significance analysis.) The outliers showed no systematic patterns with respect to probe sets or arrays (e.g., no particular array was infested with outliers), although outliers tended to appear in clumps with respect to age within a probe set. We attempted to alleviate this issue by selecting probe sets whose expression is well explained by the available sex variable, irrespective of any age dependent behavior. We applied a filtering technique that sought to identify these probe sets, while at the same time guarding against anti-conservatively biasing the subsequent significance analysis. 35,068 such probe sets were identified when considering the clinical variable identifying sex, reducing the total number of probe sets included in the analysis by 9856. Through an extensive simulation study, we were able to show that the filtering method effectively removed probe sets confounded by unobserved variables while not inflating the significance of the remaining probe sets (Supplementary Information). This method is potentially applicable to other observational genome-wide expression studies. In the subsequent significance analysis, sex indicator variables (taking values 0 or 1) were included in the parametrization of the average time curve: $\mu_i(t) = \alpha_i + \delta_i \times \text{sex} + \boldsymbol{\beta}_i^T \mathbf{s}(t)$. This allows for different intercepts for each sex.

Results

Experimental objectives and statistical formulations. We developed a general statistical method that identifies genes showing temporal differential expression. This method was applied to two human studies encompassing both types of temporal differential expression.

In one study, kidney samples were obtained from 72 human subjects ranging in age from 27 to 92 years. Only one array was obtained per sample and the age of the subject was recorded (Materials and Methods). Figure 1a displays a simulated example of expression measured for this type of study on a single probe set. The solid line is the population average time curve for the

probe set, which is its true average expression over time with all sources of variation removed. The points are the observed expression values, one per each individual, and these can be thought of as independent random deviations from the solid line due to biological and measurement variation. 'Independent sampling' was performed in this study because each sample of cortex tissue represents an independently sampled individual.

To determine whether each gene has expression that changes with age, our method involves performing a hypothesis test on each gene of whether its population average time curve is flat or not. We call this type of differential expression 'within class temporal differential expression.' Figure 1b shows the expression measurements from one of the genes in the study. (This is a highly significant gene: CRABP1, a cellular retinoic acid binding protein.) The gene is tested by first fitting a model under the null hypothesis that there is no differential expression, and then under the alternative hypothesis that there is differential expression. The null model is the dashed flat line, which minimizes the sum of squares among all possible flat lines. The alternative model is the solid curve that minimizes the sum of squares among a general class of curves, namely natural cubic splines. A statistic is calculated that compares the goodness of fit of these two models. This statistic is a quantification of evidence for differential expression, and the larger it is the more differentially expressed the gene appears to be. For every gene a statistic is calculated in this way, and a significance cut-off is applied to them using a nonparametric false discovery rate criterion [15]. This process involves calculating the null distribution of the statistics when there is no differential expression, which is accomplished through a data re-sampling technique.

In another study, eight human volunteers were randomly divided into endotoxin-treated and control groups of equal size (Materials and Methods). Figure 2a is an artificial example of expression measurements from a single gene in a group of four individuals. The solid line is the population average time curve for this gene. The dashed lines are the average time curves *for the individuals*, meaning that these are the true underlying time curves for each individual with the sources of variation removed up to their individual variation. The deviation of an expression value from its corresponding 'individual average time curve' can be thought of as an independent random event. The deviation of an individual average time curve from the population average can also be thought of as an independent random event. However, this implies that the deviations of expression values from the population average time curve are correlated within individuals. The sampling performed in this study is called 'longitudinal sampling' because each individual is observed at more than one time point.

Here we want to identify genes that show significantly different expression between the endotoxin-treated and control groups across time. We call this type of differential expression 'between class temporal differential expression.' The methodology is applied similarly here as in the kidney study.

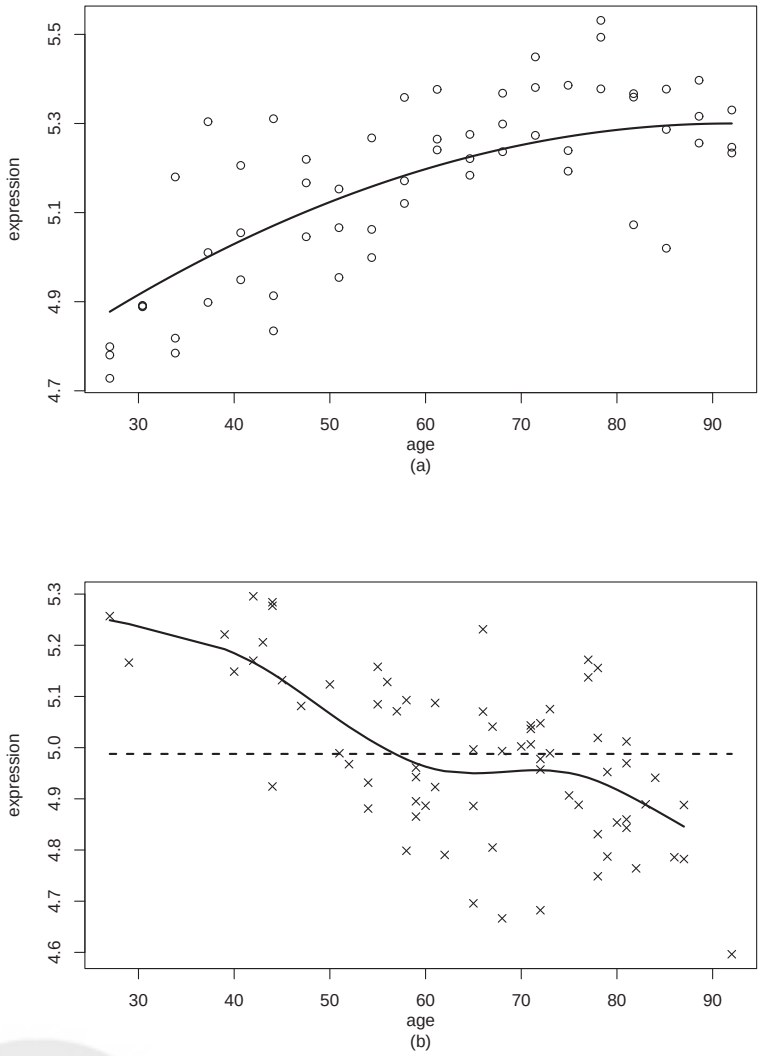

Figure 1: (a) A simulated example of log-transformed expression measurements obtained by independent sampling. The solid line is the population average time curve and the o's are observed expression values. (b) The $\log_2$-transformed expression values of the most significant gene in the kidney aging study. The solid line is the curve fitted under the alternative hypothesis of differential expression. The dashed line is the model fitted under the null hypothesis of no differential expression.

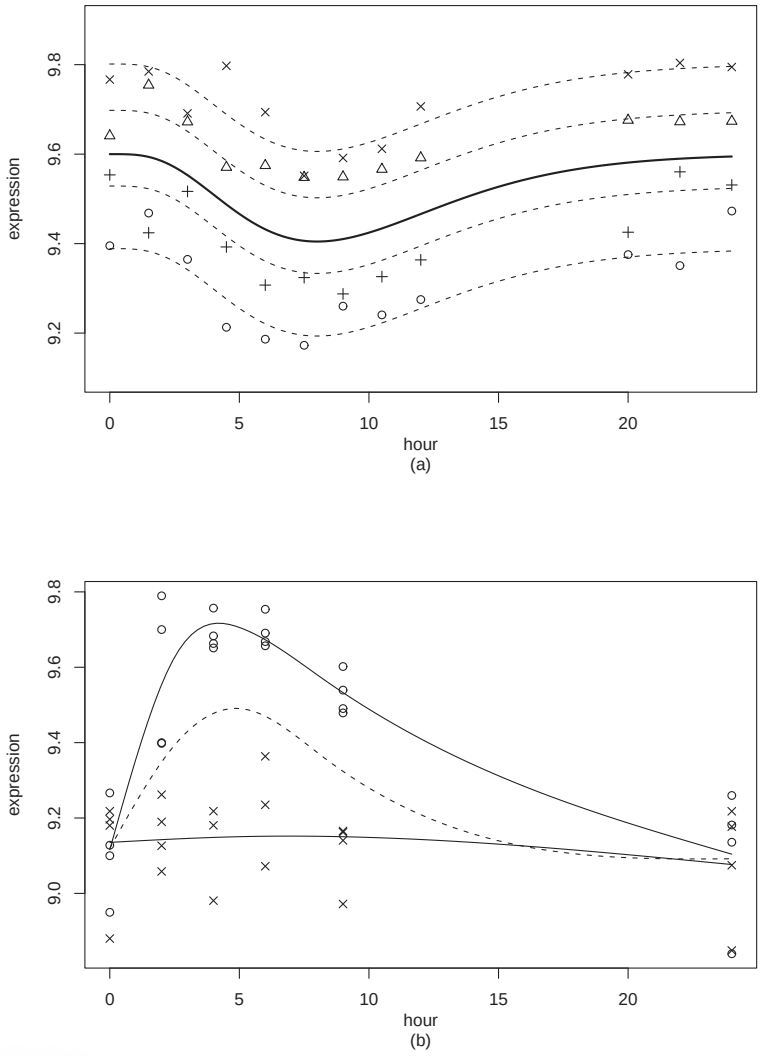

Figure 2: (a) A simulated example of log-transformed expression measurements obtained from longitudinal sampling of four individuals. The solid line is the population average time curve. The dashed lines are the average time curves *for the individuals*. The points of a common shape correspond to one of the individuals. (b) The log-transformed expression values of a significant gene from the endotoxin study. The solid lines are the curves fitted under the alternative hypothesis of differential expression. The dashed line is the curve fitted under the null hypothesis of no differential expression.

Figure 2b shows the expression measurements from one of the genes in the study. (This gene is interferon regulatory factor 1, which is significant at a false discovery rate of 1%.) Under the null hypothesis of no differential expression, the treated and control groups have the same population average time curve. Therefore, a single curve (a natural cubic spline) is fit to the combined groups, which is represented by the dashed curve in Figure 2b. The alternative model is formed by fitting a separate curve to each group, as is shown by the solid lines in Figure 2b. A statistic is computed based on the improvement in goodness of fit in going from a single curve to the separate curves for each group. As before this statistic is a quantification of evidence for differential expression, and the larger it is the more differentially expressed the gene appears to be. A significance cut-off is applied to these statistics in the same fashion as in the kidney aging study.

In contrast to a static experiment, it is more difficult in the time course setting to form statistics that accurately quantify differential expression. Determining the distribution of the statistics when there is no differential expression is also more challenging. The behavior of expression over time may vary greatly from gene to gene, so a flexible modeling approach must be taken in forming statistics. Under certain study designs (e.g., the endotoxin study) there may be dependence between the expression measurements within a single individual, which complicates the formation of statistics and the simulation of the null distribution. Finally, simple permutation methods cannot be used to simulate null statistics because of the complex structure of time course measurements.

Proposed statistical method. Methodology was developed to address these issues in a statistically rigorous fashion. In doing so, a general model for gene expression over time within a single biological group was first carefully formulated. Even though a single model can be applied to both studies (Supplementary Information), a simplified version is possible for the kidney study because of its independent sampling scheme. Let y_{ij} be the relative expression level of gene i in individual j, where there are $i = 1, 2, \ldots, 35068$ probe sets and $j = 1, 2, \ldots, 72$ individuals. Individual j is observed at age t_j, which lies somewhere between 27 and 92 years. The expression values are modeled by

$$y_{ij} = \mu_i(t_j) + \epsilon_{ij},$$

where $\mu_i(t_j)$ is the population average time curve for gene i evaluated at time t_j, and ϵ_{ij} is the random deviation from this curve. In terms of Figure 1a, $\mu_i(t)$ is shown by the solid line. The distance between this curve and an observed expression value is ϵ_{ij}. The ϵ_{ij} are assumed to be independent random variables with mean zero and gene-dependent variance σ_i^2.

The following model was developed for the endotoxin study, and for longitudinal sampling in general. Let y_{ijk} be the relative expression level of gene i on individual j at the kth time point, where there are $i = 1, 2, \ldots, 44924$ probe sets and $j = 1, 2, \ldots, 8$ different individuals sampled (4

within each group). For each individual, there were $k = 1, 2, \ldots, 6$ time points observed at times t_{jk}, except for the one control who is missing two time points. The expression values are modeled by

$$y_{ijk} = \mu_i(t_{jk}) + \gamma_{ij} + \epsilon_{ijk}.$$

The population average time curve for gene i is again $\mu_i(t)$. Individual j deviates from $\mu_i(t)$ by γ_{ij}, implying that $\mu_i(t) + \gamma_{ij}$ is the individual average time curve for individual j. The measurement error and remaining sources of random variation are modeled by ϵ_{ijk}. In Figure 2a the solid line is represented by $\mu_i(t)$, and the dashed lines by $\mu_i(t) + \gamma_{ij}$. Each expression value deviates from its corresponding dashed line by ϵ_{ijk}. The γ_{ij} and ϵ_{ijk} are assumed to be independent random variables with means equal to zero and gene-dependent variances τ_i^2 and σ_i^2, respectively. The case where γ_{ij} is modeled as a curve will be dealt with elsewhere; however, the endotoxin study did not contain enough observations to permit this extra level of complexity.

The population average time curve $\mu_i(t)$ is parameterized in terms of an intercept plus a p-dimensional linear basis:

$$\begin{aligned} \mu_i(t) &= \alpha_i + \boldsymbol{\beta}_i^T \mathbf{s}(t) \\ &= \alpha_i + \beta_{i1} s_1(t) + \beta_{i2} s_2(t) + \cdots + \beta_{ip} s_p(t), \end{aligned}$$

where $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_p(t)]^T$ is a pre-specified p-dimensional basis, α_i is the unknown gene-specific intercept, and $\boldsymbol{\beta}_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip}]^T$ is a p-dimensional vector of unknown gene-specific parameters. A common basis is defined by $s_1(t) = t, s_2(t) = t^2, \ldots, s_p(t) = t^p$ [23]. That is, the average curve $\mu_i(t)$ is modeled as a polynomial of degree p: $\mu_i(t) = \alpha_i + \beta_{i1} t + \beta_{i2} t^2 + \cdots + \beta_{ip} t^p$. The polynomial basis was effective in both studies, but a more flexible basis resulted in an increase in power while making fewer assumptions. If the goal is to minimize the sum of squares from among all continuous curves that are not 'too curvy', then there exists a unique solution called a natural cubic spline, which can be parameterized by the B-spline basis [42]. Natural cubic spline models have been developed and applied in a variety of other scenarios [25–33]. For either type of basis, the curve was fit by minimizing the sum of squares between the curve and the observed expression values, which reduces to fitting the α_i and $\boldsymbol{\beta}_i$ by standard least squares regression methods. The model fitting procedure for longitudinal sampling is slightly more complicated in that it also takes into account the dependence of measurements within an individual (Supplementary Information).

Not only does the basis representation of $\mu_i(t)$ facilitate model fitting, but it also greatly simplifies the hypothesis tests for differential expression. They can now be written in terms of the α_i and $\boldsymbol{\beta}_i$ for each gene, which is a source of flexibility in the types of designs that our method can consider. Specifically, the tests do not depend on specific time points, so general sampling schemes may be analyzed. In the kidney aging study, the null hypothesis of no differential expression is equivalent

to restricting $\mu_i(t)$ to be a constant, and the alternative hypothesis of differential expression allows $\mu_i(t)$ to be a curve. The null hypothesis model is fit under the constraint that $\mu_i(t) = \alpha_i$ and $\boldsymbol{\beta}_i = \mathbf{0}$, and the alternative hypothesis model under the general parametrization of $\mu_i(t)$. For the endotoxin study, the null hypothesis is that the treated and control groups have equal $\mu_i(t)$ (i.e., equal α_i and $\boldsymbol{\beta}_i$), and the alternative hypothesis is that they are not equal. The null hypothesis model is obtained by fitting a curve to the two groups combined, and the alternative hypothesis model by fitting a separate curve to each group. In this particular study, we were not interested in a difference in the intercepts α_i because all individuals started out as untreated at time 0. Therefore the intercept was implicitly assumed to be equal between the two groups under both hypotheses, which comes down to a test for equality of the $\boldsymbol{\beta}_i$ between the two groups. This will not always be the case when testing for between class temporal differential expression, so we have developed model fitting methods for both scenarios (Supplementary Information).

One of the difficulties in employing a basis representation of the time curves is that the dimension p must be specified. This aspect is important because setting p too small may result in decreased ability to capture the signal, while setting p too large wastes degrees of freedom. Either type of error leads to a loss in power. We developed a method to automatically choose a single value of p for all genes. The basic idea is to take a singular value decomposition of the expression data and extract the top few 'eigen-genes', which are the eigen-vectors in the gene space [18]. The top eigen-genes represent directions in the gene space that explain the most variance. Curves are fit to these eigen-genes and p is chosen to be of sufficient size through a cross-validation technique. The method chose $p = 4$ for both studies (although the fact that they are equal is a coincidence). In both cases, this value yielded significance results that showed empirical evidence for being the most powerful choice (Supplementary Information). Another option is to perform the significance analysis over a range of p, although this could result in an anti-conservative bias. Also, when allowing p to vary from gene to gene, a substantial inflation of significance was detected because p was chosen to work well under the alternative hypothesis. One of the main motivations for specifying only one p for all genes and using a singular value decomposition, which captures global trends in the data, was to avoid over-fitting any particular gene.

A statistic for each gene was then formed that quantifies differential expression. The statistic was defined to be an analogue of the t- and F-statistics that are commonly used in static differential expression methods. The statistic compares the goodness of fit of the model under the null hypothesis to that under the alternative hypothesis. First, fitted values resulting from the null and alternative models are calculated that correspond to each observed value. The residuals of the model fits are then obtained by subtracting the fitted values from the observed values. Calculating SS_i^0 to be the sum of squares of the residuals obtained from the null model, and SS_i^1 from the

alternative model, the statistic for gene i was constructed as

$$F_i = \frac{\mathrm{SS}_i^0 - \mathrm{SS}_i^1}{\mathrm{SS}_i^1}.$$

This is proportional to the typical F-statistic used in comparing two nested models. The intuition behind the formula is that $\mathrm{SS}_i^0 - \mathrm{SS}_i^1$ quantifies the increase in goodness of fit, and dividing this difference by SS_i^1 provides exchangeability of the F_i across the genes. Justification and exact formulas for all cases can be found in the Supplementary Information.

The null distribution of these statistics is calculated through a data re-sampling method called the bootstrap [43]. The basic idea is that the data are re-sampled in such a way that new versions of null data are randomly generated for each gene. More specifically, residuals from the alternative model are first calculated. These residuals are randomly sampled with replacement and added back to the null fitted values. Using these simulated null data, models are fit and statistics are formed exactly as was done before to produce a set of null statistics. The re-sampling was performed on each study for 500 bootstrap iterations. In the longitudinal sampling case, extra steps have to be taken to deal with the dependence of residuals within an individual, but the basic idea is the same (Supplementary Information).

The larger F_i is, the better the alternative model fit is over the null model fit. Therefore it is reasonable to rank the genes for differential expression by the size of the F_i. This is equivalent to calling genes significant that have $F_i \geq c$ for some positive cut-point c. It is straightforward to form a p-value for each gene by calculating the frequency by which the null statistics exceed the observed statistic. Even though the p-value is a useful measure of significance for an individual gene, it is difficult to interpret a p-value threshold applied to thousands of genes simultaneously. We instead estimated 'q-values' as the measure of significance for each gene.

The q-value is a false discovery rate (FDR) specific measure of significance. Estimation of q-values in genomics studies, including for identifying differentially expressed genes, has previously been proposed and studied [15, 44–46]. For a given significance threshold, the false discovery rate is the proportion of false positives among all genes called significant:

$$\mathrm{FDR} \approx \frac{\#\text{false positive genes}}{\#\text{significant genes}}.$$

The q-value of a particular gene measures the false discovery rate that is incurred when calling that gene (and every gene more extreme) significant. For example, if a gene has a q-value of 1%, then drawing the cut-off for significance at this gene leads to at most 1% false positives among all significant genes. The observed statistics, null statistics, and significance rule were used to estimate q-values for the genes as previously described (Supplementary Information). By applying reasonable q-value cut-offs in both studies, many apparently relevant genes were called significant.

Table 1: A comparison of the proposed method to a standard t-test and a "SAM" test (which is a t-test with adaptive asymmetric thresholding). The ratio of the number of probe sets called significant is an estimate of the increase in power that our procedure provides. For the static methods, 8 subjects and 16 arrays were used to compared two time points (0h vs 2h and 0h vs 4h). The proposed method was compared to these in two ways. First, all 8 subjects and 46 arrays were used, potentially giving our method an advantage due to a decrease in technical replication variance. Second, only 4 subjects were used at times 0h, 4h, 6h, 24h for a total of 16 arrays, giving the static methods a substantial advantage in that they have twice as many biological replicates. The fairest comparison lies in between these two, even though both outperform the static methods substantially.

	Number of Probe Sets Called Significant					
Q-value	Proposed Method	Proposed Method	T-test		SAM	
Cut-off	(8 subjects, 46 arrays)	(4 subjects, 16 arrays)	0h vs 2h	0h vs 4h	0h vs 2h	0h vs 4h
1%	7409	548	0	0	0	0
2%	9188	2683	226	91	65	0
3%	10,467	4392	1678	837	1756	695
4%	11,642	5859	2524	1826	2535	1718
5%	12,720	7229	3202	2686	3101	2694

Analysis of systemic inflammatory response induced by LPS. In the endotoxin study there are 4163, 7409, and 12720 significant probe sets at respective q-value cut-offs of 0.1%, 1%, and 5% (Supplementary Table 1). Figure 3A shows a hierarchical clustering [16] of the 4136 probe sets significant at the 0.1% false discovery rate level. These results indicate that an endotoxin injection causes a profound gene expression response in blood leukocytes, which is expected for a number of reasons. First, in vivo administration of LPS invokes an acute systemic inflammation which dramatically perturbs body's physiology. Second, some differential expression may be due to changing distributions of cell populations in the blood in addition to transcriptional changes [35]. Also, measuring differential expression over time is a more sensitive study design than the typical static design; *any* change over time qualifies as differential expression. A t-test [15] and the SAM software [9, 15] were used to test for differential expression between 0 and 2 hours, and between 0 and 4 hours. The significance results are displayed in Table 1, where it can be seen that an actual time course analysis offers a sizeable increase in statistical power over a static design analysis.

To get a broad picture of the behavior of differentially expressed genes, a singular value decomposition was performed on the 4163 most significant probe sets. Figure 4 shows the top two eigen-genes that explain 66% and 16% of the variance, respectively. The relevant information is

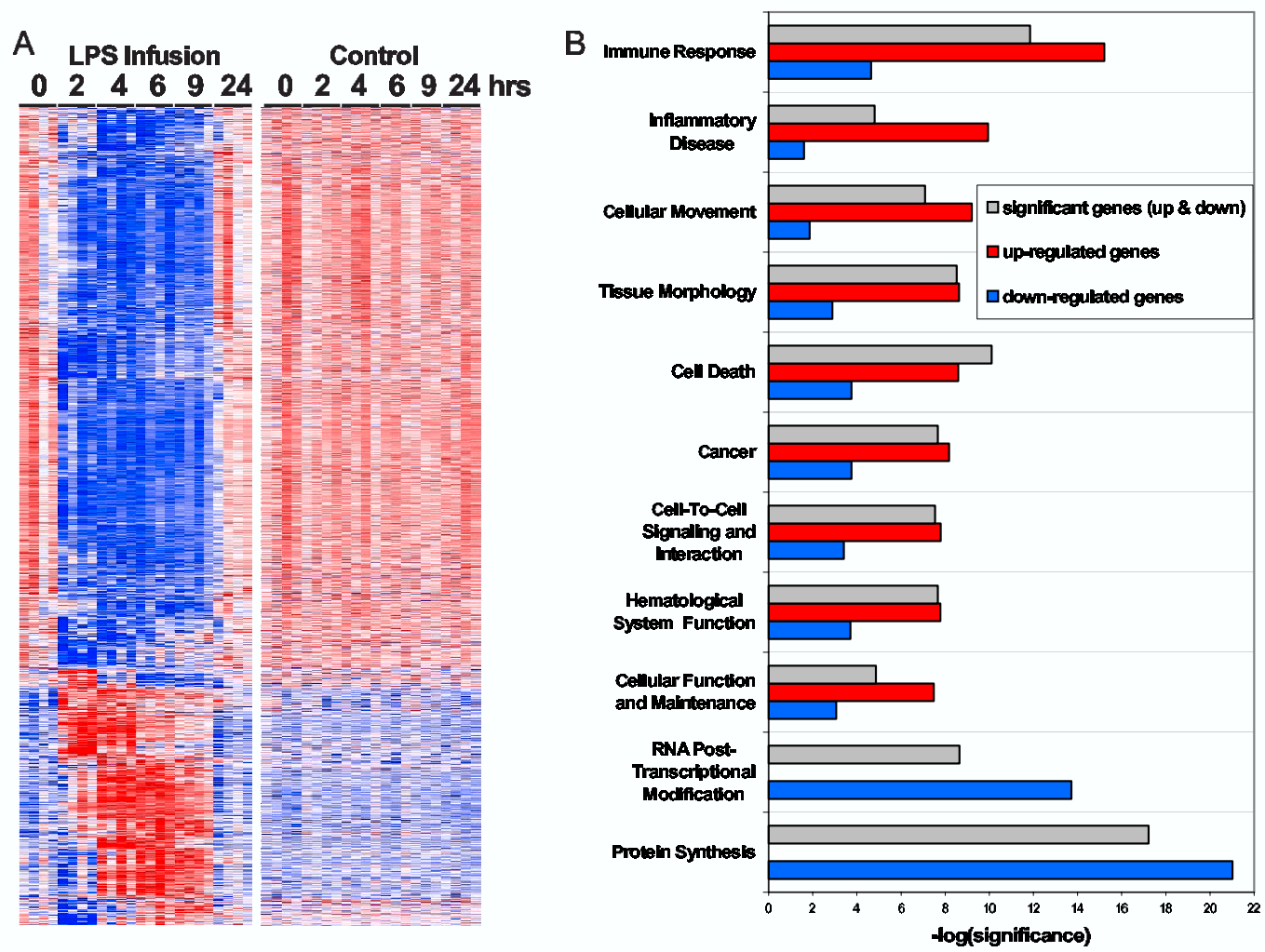

Figure 3: **A.** Hierarchical clustering of the probe sets found to be significant at a q-value cut-off of 0.1% in the endotoxin study. The columns are labeled according to group and time in hours. The heat map is such that red intensity indicates high expression and blue intensity low expression **B.** Functional analysis of genes. Shown are representative functions differentially enriched (difference in log[significance] > 4) between the the groups of up-regulated (red) and down-regulated (blue) genes. The significance values of these functions in the combined group are shown in gray.

the shape of each eigen-gene, not their magnitudes or direction of differential expression. Among the 4163 most significant probe sets 27% are up-regulated as the eigen-gene shows; 73% are down-regulated, which is simply the reflection of the eigen-gene's shape across the time axis. The second eigen-gene shows more complex behavior, where changes in expression occur both directions over the time course. Among these 4136 probe sets, 2914 unique genes (756 up, 2158 down)were identified from 3892 probe sets with mapped gene IDs (271 unmapped). Global functional analysis was performed using gene ontology built from experimental evidence compiled in Ingenuity Pathways Knowledge Base (IPKB) (Figure 3B and Supplementary Table 1).

The most significant functional groups from the 756 up-regulated genes take part in Immune Response (121 genes), Inflammatory Disease (41 genes), Cellular Movement (84 genes), Tissue Morphology (72 genes), and Cell Death (144 genes). These are consistent with an intense response of leukocytes to LPS. The apparent expression of many elements of inflammation are up-regulated, including secretive cytokines, chemokines and associated proteins (CCL20, CCL3, CCL4, CXCL16, CXCL2, IL1A, IL1B, IL8, TNF, TNFSF13B, IL1RAP, IL1RN, TNFAIP3, TNFAIP6), their membrane receptors (CCR1, IL10RB, IL18R1, IL18RAP, IL1R1, IL1R2, IL4R, IL8RA, IL9R, TNFRSF6), toll-like receptors (TLR4, TLR5, TLR6, TLR8), Fc receptors (FCAR, FCER1G, FCGR1A, FCGR2A, FCGR2B), interferon receptors (IFNGR1, IFNGR2, IFNAR1), protein tyrosine phosphatases (PTP4A1, PTPN2, PTPN22, PTPNS1, PTPRC, PTPRJ, PTPRN2, PTPRO), JAK/STAT (JAK2, JAK3, STAT2, STAT3, STAT5A, STAT5B) and NFKB/IKB proteins (NFKB2, NFKBIA, NFKBIE). These and other elevated genes participate in the activation of the extracellular and intracellular signaling pathways in innate immune response, as well as many functions of leukocytes, including cell movement of leukocytes, infiltration, migration, phagocytosis, activation, chemotaxis, proliferation and recruitment. The most significantly up-regulated gene is ORM1, a key acute phase plasma protein.

The most significantly down-regulated functions are Protein Synthesis (83 genes) including genes of elongation initiation factors (EIF), ribosomal proteins (RPL/RPS), mitochondrial ribosomal protein (MRP); and RNA Post-Transcriptional Modification (60 genes), including genes such as heterogeneous nuclear ribonucleoproteins (hnRNP), small nuclear ribonucleoproteins (SNRP), splicing factors(SFRS). Members of RNA polymerase II (POLR2) are also decreased. Correspondingly, genes involved in oxidative phosphorylation, such as mitochondrial complexes I (NDUF), II (SDH), III (UQCR), and IV (COX) are also found to be significantly down-regulated. This concerted suppression of the cells protein syntheses, transcription programs and energy productions may reflect a generalized stress response in blood leukocytes. Interestingly, expression of members associated with the major histocompatibility complex class II (MHC II) complex is also suppressed (CD1D, CD74, HLA-DMA, HLA-DMB, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-DRB1), con-

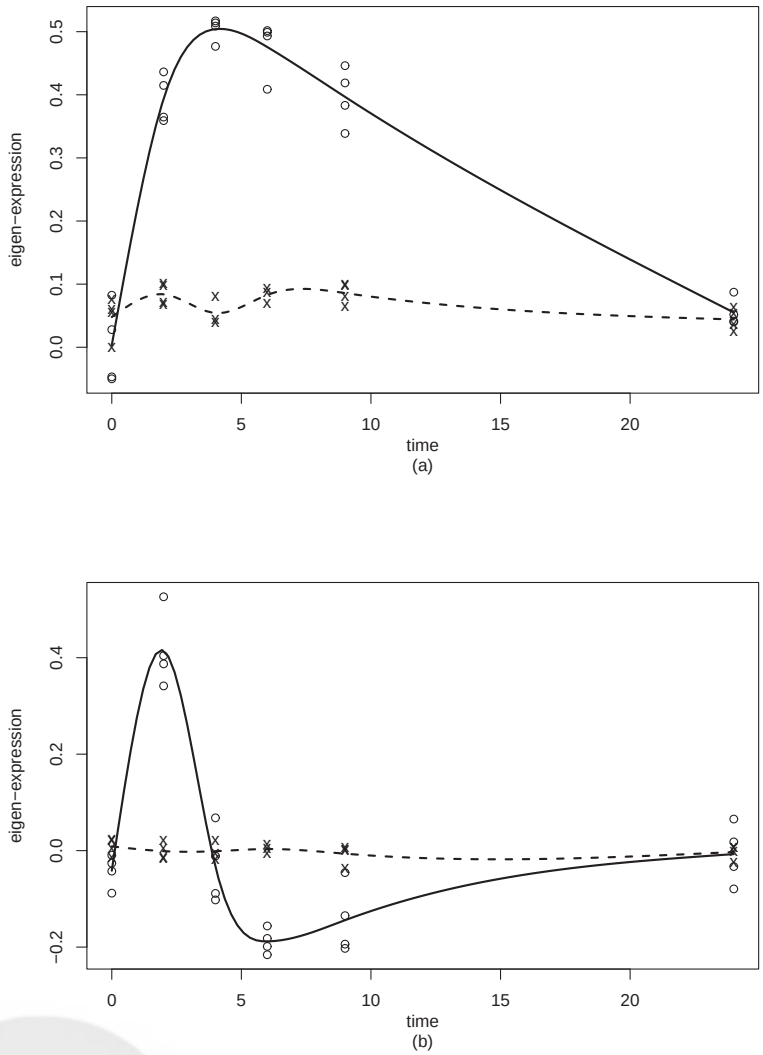

Figure 4: The top two eigen-genes obtained from probe sets significant at a q-value cut-off of 0.1% in the endotoxin study. The control individuals are represented by `x`'s and the treated individuals by `o`'s. The dashed curve is the model fit to the control group and the solid curve is the model fit to the treated group. (a) The first eigen-gene explains 66% of the variance. (b) The second eigen-gene explains 16% of the variance.

sistent with the reduced antigen presenting capability after LPS shock.

Taking these results together, we observed a comprehensive transcriptional response to LPS stimulation, where leukocyte cells reallocate resources and up-regulate transcription of genes involved in innate immune and defense mechanisms.

Under the assumption that the control individuals have flat expression over time, it would have sufficed to simply test the endotoxin individuals for within class temporal differential expression. However, we found that among the controls *at least* 6% of the probe sets show within class temporal differential expression (Supplementary Table 3). [The 6% estimate is directly obtained from the false discovery rate procedure [15,44].] This indicates that the controls do in fact play an important role in the analysis, which is not entirely surprising when considering that the study took place over a 24 hour period where a number of potentially influential events may have taken place. We also tested the endotoxin treated individuals for within class temporal differential expression and found 5781, 16523, and 21789 significant probe sets at respective q-value cut-offs of 0.1%, 1%, and 5%. At the 1% and 5% q-value cut-offs there are over twice as many significant genes compared to those obtained perviously. Given this and the fact that the controls do show within class temporal differential expression, our conclusion is that it is necessary to use the treated and controls individuals in a test for between class temporal differential expression in order to identify genes differentially expressed due to LPS injection.

Analysis of age related genes in the kidney cortex. There appears to be substantial differential expression in this study, although not nearly as much as in the previous study. At q-value cut-offs of 5% and 10%, there are 187 and 436 significant probe sets, respectively. The top eigen-gene based on the 187 most significant probe sets shows a steady increase/decrease in expression, and the second eigen-gene is flat (Supplementary Information). Among the 436 probe sets significant at 10%, 320 unique genes were identified from 389 probe sets with mapped gene IDs and 47 unmapped. Of these 320 genes, 240 have increased expression in kidneys of older age and 80 have decreased expression.

Functional analysis was performed on these age-regulated genes (the complete information is shown in Supplementary Table 2). The most significant functional group is Immune Response (69 genes), including genes such as associated with the complement component (C1QA, C1QB, C1QG, C1QR1, C1R, C1S, C4A), cytokines, chemokines and receptors (CCL2, CCL20, CCR1, CX3CR1, CXCL9, CXCR4, IL10RB, CSF1R, CSF2RB), and MHC II complex (FCER1G, HLA-DMA, HLA-DPA1, HLA-DQB1, HLA-DRB1, HLA-DRB3). All but three of these genes show up-regulated expression with age. This suggests an increased abundance of immune response genes in kidney cortex, either as a result of low level of inflammation over life or as a compensation for the decreased immune function in older age. Renal infiltration with immunocompetent cells which is known to

correlate with renal diseases and increase over age can also account for some of these observed changes.

The aging process has previously been shown to involve a widespread and complicated alternation in gene expression [36, 37, 47, 48]. In our analysis, significant genes were identified that have a variety of cell-to-cell signaling and interaction (68 genes). For example, 34 genes are involved in the quantity or mobilization of calcium and 14 genes in tyrosine phosphorylation. 59 genes were identified as involved in apoptosis and cell death. The expression levels of apoptosis enhancer proteins MYC (C-MYC), SP1 (Sp1 transcription factor), CASP1 (caspase1), UBD (ubiquitin D) increase with age in the kidney, while the anti-apoptosis protein HSPA9B (mortalin-2) decreases with age. However, the expression of BCL2A1, an inhibitor of the intrinsic apoptosis pathway, is observed to be increasing with age [49].

Gene expression may change with age in response to declining kidney function and increased susceptibility to kidney disease. Interestingly, a group of 12 genes known to be localized in mitochondrion (HSPA9B, COX8A, COX7C, AKAP1, BCKDHA, CLPP, FDX1, AMT, DBT, ATP5G3, AK2, NME4) are identified as significant and their expression level uniformly declines in older age. A number of age-regulated genes are also known to be involved in renal diseases, including C1QA, CCR1, LYN, PTPRC, and TNFSF13B which all have positive correlation over age [50–54]. The expression of TRPM6 (a transient receptor potential cation channel) negatively correlates with age, and mutations of this gene cause hypomagnesemia with secondary hypocalcemia [55]. These genes could therefore provide valuable information or as potential biomarkers on clinical course of kidney aging.

We employed a 4-dimensional B-spline basis (p=4) in the significance analysis, although other dimensions are possible. A 1-dimensional basis is equivalent to identifying differential expression by fitting a linear regression and testing whether the slope is zero or not. Applied to log-transformed expression data, a linear regression is only able to identify genes that show simple exponential behavior in their differential expression. By testing the null hypothesis of linear or flat log-transformed expression, we found that at least 25% of the 35,068 probe sets are differentially expressed in a manner more complex than an exponential function of age. Among the 417 most significant probe sets in the kidney study, at least 93% of these are more complex than simple exponential behavior. However, this may be a substantially exaggerated estimate due to the 417 probe sets being preferentially sampled for nonlinear behavior. Thus, among the 417 most significant probe sets when using the 1-dimensional basis, it is estimated that at least 47% of these are more complex than simple exponential behavior. At a q-value cut-off of 5%, just over half of the genes found to be significant using the 4-dimensional basis are also significant when using the 1-dimensional basis. Therefore, there is potentially a functional class of genes showing simple exponential temporal

differential expression and another functional class showing substantially more complex differential expression.

Discussion

We have proposed a significance method to identify differentially expressed genes in time course microarray experiments and applied it to studies involving two types of sampling and both types of temporal differential expression. The method may also be applied to more complicated situations, where three or more groups are compared, for example. The four basic steps that are required for a complete significance analysis method are included in the method: a statistic is formed for each gene, the null distribution is calculated, the statistics are used to rank the genes, and a measure of significance is applied to each gene that takes into account the multiple comparisons. There is much detail that has been worked out to thoroughly and broadly define the method. Moreover, we have attempted to justify the method in terms of well established statistical concepts. Extensive simulations have been performed that test the robustness of the method under a variety of models and sources of random variation. See the Supplementary Information for these details.

Temporally differentially expressed genes were identified in two time course microarray studies on humans. The findings provide evidence that the method produces biologically meaningful results. The human *in vivo* model of bacterial endotoxin represents a unique opportunity to examine the onset of systemic inflammation in blood leukocytes. Our analysis suggests that a genome-wide transcriptional response takes place where blood leukocytes reallocate resources and up-regulate transcription of genes involved in innate immune and defense mechanisms, even though the response to LPS is almost exclusively signaled by the toll-like receptor TLR4 pathway [56]. It was shown that the inclusion of controls in this study played an important role, where the controls themselves showed differential expression over time. In the clinical setting, comparing a treatment group to "time 0" does not provide a proper control over the entire time course. The significance analysis of the kidney aging study indicates that a large proportion of genes have expression that increases with age, including those involved in signal transduction, cell growth, and genome stability. Genes related to metabolism were found to be decreasing with age. We attempted to reduce the effect of unobserved variables on our results, although this is not guaranteed to eliminate all confounders in an observational study.

We have developed a freely available point-and-click software package called EDGE that includes this methodology, as well as new methodology for static experiments. EDGE can be downloaded at `http://faculty.washington.edu/~jstorey/edge/`.

Supplementary Materials

The Supplementary Information and Tables can be found at `http://faculty.washington.edu/jstorey/time/`.

Acknowledgments

This research was supported by the NIH Large Scale Collaborative Research Grant[1] U54 GM2119-03 (PI: Ronald Tompkins, `http://www.gluegrant.org/`). Thanks to Stuart Kim, Steve Calvano and Stephen Lowry for their generosity in sharing data.

References

[1] Slonim, D. K. (2002) *Nature Genetics* **32**, 502–508 (supplement).

[2] Ideker, T., Thorsson, V., Siegel, A., & Hood, L. (2000) *Journal of Computational Biology* **7**, 805–817.

[3] Kerr, M. K., Martin, M., & Churchill, G. A. (2000) *Journal of Computational Biology* **7**, 819–837.

[4] Lee, M. T., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000) *Proceedings of the National Academy of Sciences* **18**, 9834 – 9839.

[5] Baldi, P. & Long, A. D. (2001) *Bioinformatics* **17**, 509–519.

[6] Newton, M., Kendziorski, C., Richmond, C., Blatter, F., & Tsui, K. (2001) *Journal of Computational Biology* **8**, 37–52.

[7] Thomas, J. G., Olson, J. M., Tapscott, S. J., & Zhao, L. P. (2001) *Genome Research* **11**, 1227–1236.

[8] Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001) *Journal of the American Statistical Association* **96**, 1151–1160.

[9] Tusher, V., Tibshirani, R., & Chu, C. (2001) *Proceedings of the National Academy of Sciences* **98**, 5116–5121.

[10] Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., & Paules, R. S. (2001) *Journal of Computational Biology* **8**, 625–637.

[11] Dudoit, S., Yang, Y., Callow, M., & Speed, T. (2002) *Statistica Sinica* **12**, 111–139.

[12] Olshen, A. B. & Jain, A. N. (2002) *Bioinformatics* **18**, 961–970.

[13] Ibrahim, J. G., Chen, M. H., & Gray, R. J. (2002) *Journal of the American Statistical Association* **97**, 88 – 99.

[14] Storey, J. D. & Tibshirani, R. (2003) *Methods in Molecular Biology* **224**, 149–157.

[15] Storey, J. D. & Tibshirani, R. (2003) *Proceedings of the National Academy of Sciences* **100**, 9440–9445.

[1] Additional participating investigators in the Large Scale Collaborative Research Program entitled *Inflammation and the Host Response to Injury*: Henry V. Baker, Paul E. Bankey, Timothy R. Billiar, Bernard H. Brownstein, Steve E. Calvano, Irshad H. Chaudry, J. Perren Cobb, Chuck Cooper, Bradley Freeman, Richard L. Gamelli, Nicole S. Gibran, Brian G. Harbrecht, Wyrta Heagy, David M. Heimbach, David N. Herndon, Jureta W. Horton, John Lee Hunt, Jeffrey Johnson, James A. Lederer, Tanya Logvinenko, Stephen F. Lowry, John A. Mannick, Bruce A. McKinley, Carol Miller-Graziano, Michael N. Mindrinos, Joseph P. Minei, Lyle L. Moldawer, Ernest E. Moore, Frederick A. Moore, Avery B. Nathens, Grant E. O'Keefe, Laurence G. Rahme, Daniel G. Remick, Jr., Michael B. Shapiro, Robert L. Sheridan, Geoffrey M. Silver, Richard D. Smith, Scott Somers, Gregory Stephanopoulos, Mehmet Toner, H. Shaw Warren, Michael A. West, Steven E. Wolf, Vernon R. Young.

[16] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998) *Proceedings of the National Academy of Sciences* **95**, 14863–14868.

[17] Zhu, G., Spellman, P. T., Volpe, T., Brown, P. O., Botstein, D., Davis, T. N., & Futcher, B. (2000) *Nature* **406**, 90–94.

[18] Alter, O., Brown, P. O., & Botstein, D. (2000) *Proceedings of the National Academy of Sciences* **97**, 10101–10106.

[19] Wakefield, J., Zhou, C., & Self, S. (2003) *Statistics 7, Proceedings of the Seventh Valencia International Meeting*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, Smith, A. F. M., & West, M. (Oxford University Press), pp. 721–732.

[20] Bar-Joseph, Z., Gerber, G., Jaakkola, T. S., Gifford, D. K., & Simon, I. (2003) *Journal of Computational Biology* **10**, 341–356.

[21] Luan, Y. & Li, H. (2003) *Bioinformatics* **19**, 474–482.

[22] Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D. K., & Jaakkola, T. S. (2003) *Proceedings of the National Academy of Sciences* **100**, 10146–10151.

[23] Diggle, P., Heagerty, P., Liang, K. Y., & Zeger, S. (2002) *Analysis of Longitudinal Data.* (Oxford University Press), 2nd edition.

[24] Ramsay, J. O. & Silverman, B. W. (1997) *Functional Data Analysis.* (Springer).

[25] Wang, Y. (1998) *Journal of the Royal Statistical Society, Series B* **60**, 159–174.

[26] Zhang, D., Lin, X., Raz, J., & Sowers, M. (1998) *Journal of the American Statistical Association* **93**, 710–719.

[27] Brumback, B. & Rice, J. (1998) *Journal of the American Statistical Association* **93**, 961–976.

[28] Zhang, D., Lin, X., & Sowers, M. (2000) *Biometrics* **56**, 31–39.

[29] James, G. & Hastie, T. (2001) *Journal of the Royal Statistical Society, Series B* **63**, 533–550.

[30] Rice, J. & Wu, C. (2001) *Biometrics* **57**, 253–259.

[31] Irizarry, R. A., Tankersley, C. G., Frank, R., & Flanders, S. E. (2001) *Biometrics* **57**, 1228–1237.

[32] James, G. & Sugar, C. (2003) *Journal of the American Statistical Association* **98**, 397–408.

[33] Crainiceanu, C. & Ruppert, D. (2004) *Statistica Sinica* **14**, 713–729.

[34] Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., & Trent, J. (2001) *New England Journal of Medicine* **344**, 539–548.

[35] Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., & et al. (2004) A network-based analysis of systemic inflammation in humans. Submitted.

[36] Pletcher, S., Macdonald, S., Marguerie, R., Certa, U., Stearns, S., Goldstein, D., & L., P. (2002) *Curr Biol* **12**, 712–723.

[37] Zou, S., Meadows, S., Sharp, L. Jan, L., & Jan, Y. (2000) *Proc Natl Acad Sci USA* **97**, 13726–13731.

[38] Bodyak, N., Kang, P., Hiromura, M., Sulijoadikusumo, I., Horikoshi, N., Khrapko, K., & Usheva, A. (2002) *Nucleic Acids Res.* **30**, 3788–94.

[39] Rodwell, G. E. J., Sonu, R., Zahn, J. M., Lund, J., Wilhelmy, J., Wang, L., Xiao, W., Mindrinos, M., Crane, E., Segal, E., Myers, B. D., Davis, R. W., Higgins, J., Owen, A. B., & Kim, S. K. (2004) *PLoS Biology* **2**, e427.

[40] Higgins, J., Wang, L., Kambham, N., Montgomery, K., Maxon, V., Vogelmann, S., Lemley, K., Brown, P., Brooks, J., & van de Rijn, M. (2004) *Mol Biol Cell* **15**, 649–656.

[41] Li, C. & Wong, W. H. (2001) *Proceedings of the National Academy of Sciences* **98**, 31–36.

[42] Green, P. J. & Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. (Chapman & Hall).

[43] Efron, B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. (Chapman & Hall).

[44] Storey, J. D. (2002) *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

[45] Storey, J. D. (2003) *Annals of Statistics* **31**, 2013–2035.

[46] Storey, J. D., Taylor, J. E., & Siegmund, D. (2004) *Journal of the Royal Statistical Society, Series B* **66**, 187–205.

[47] Guarente, L. & Kenyon, C. (2000) *Nature* **408**, 255–262.

[48] Tissenbaum, H. & Guarente, L. (2002) *Dev Cell* **2**, 9–19.

[49] Hipfner, D. & Cohen, S. (2004) *Nature Reviews Molecular Cell Biology* **5**, 805–816.

[50] Yu, C., Yen, T., Lowell, C., & DeFranco, A. (2001) *Curr Biol* **11**, 34–38.

[51] Majeti, R., Xu, Z., Parslow, T., Olson, J., Daikh, D., Killeen, N., & Weiss, A. (2000) *Cell* **103**, 1059–1070.

[52] Mitchell, D., Pickering, M., Warren, J., Fossati-Jimack, L., Cortes-Hernandez, J., Cook, H., Botto, M., & Walport, M. (2002) *J Immunol* **168**, 2538–2543.

[53] Topham, P., Csizmadia, V., Soler, D., Hines, D., Gerard, C., Salant, D., & Hancock, W. (1999) *J Clin Invest* **104**, 1549–1557.

[54] Gross, J., Johston, J., Mudri, S., Enselman, R., Dillon, S., Madden, K., Xu, W., Parrish-Novak, J., Foster, D., Lofton-Day, C., Moore, M., Littau, A., Grossman, A., Haugen, H., Foley, K., Blumberg, H., Harrison, K., Kindsvogel, W., & Clegg, C. (2000) *Nature* **404**, 995–999.

[55] Schilingmann, K., Weber, S., Peters, M., Niemann Nejsum, L., Vitzthum, H., Klingel, K., Kratz, M., Haddard, e., Ristoff, E., Dinour, D., Syrrou, M., Nielson, S., Sassen, M., Waldegger, S., Seyberth, H., & Konrad, M. (2002) *Nature Genetics* **31**, 166–170.

[56] Poltorak, A., He, X., Smirnova, I., Liu, M. Y., Van Huffel, C., Du, X., Birdwell, D., Alejos, E., Silva, M., Galanos, C., Freudenberg, M., Ricciardi-Castagnoli, P., Layton, B., & Beutler, B. (1998) *Science* **282**, 2085–2088.