A Significance Method for Time Course Microarray Experiments Applied to Two Human Studies

John D. Storey[†‡*], Jeffrey T. Leek[†], Wenzhong Xiao[§], James Y. Dai[†], Ron W. Davis[§]

August 2004

Abstract: A common goal in microarray experiments is to identify genes that are differentially expressed among two or more biological conditions. There is currently no standard methodology for detecting differential expression in time course studies. However, it is clear that monitoring the behavior of gene expression over time is important and will be a common experimental design in the future. Here we present a general statistical significance method for detecting temporal differential expression that can be applied to the typical types of comparisons and sampling schemes. We apply this method to two studies that we have carried out on humans. The goal of one study is to identify genes showing temporal differential expression between controls and endotoxin-treated individuals, and the other is to identify genes that show aging effects in the kidney. Genes identified in both studies corroborate previous findings and also provide novel insights. This methodology has been implemented in the freely distributed EDGE **software package.**

Keywords: aging, differential expression, endotoxin, expression arrays, time series, q-values

† Department of Biostatistics and ‡ Department of Genome Sciences, University of Washington, Seattle WA 98195.

§ Stanford Genome Technology Center and Department of Biochemistry, Stanford University, Stanford CA 94305.

∗ To whom correspondence should be sent: `jstorey@u.washington.edu`

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 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 mixture model 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. However, the K-means clustering method [20, 21] has been modified to compare expression over time between two groups [22]. This method can only be applied to a few hundred genes because of the computational cost of fitting a single model to all genes simultaneously. Moreover, the statistical significance is calculated under the assumption that the clustering fit to one of the groups is true, which is nonstandard and potentially problematic.

The method that we present and apply here draws on principles from the extensive statistical literature on time course data analysis [23,24] and from established microarray significance analysis techniques [15]. 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 observed at more than one time point, 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 studies that involve about forty thousand genes each.

Existing significance methods for microarrays that are all versions of analysis of variance (ANOVA) [3, 9–11] could be adapted to time course data if strong restrictions were placed on the observed time points, requiring the sampling of the microarrays to be uniform across groups and individuals. The method presented here is applicable to much more general cases, where groups or individuals may be sampled at different time points and where time points may be missing.

As opposed to adaptations of ANOVA, our method borrows information across time in a flexible manner so that general sampling schemes and missing data can be accommodated. This flexible modeling also allows the behavior over time from gene to gene to differ substantially without losing statistical power. The approach permits fewer degrees of freedom to be used over adaptations of ANOVA, also resulting in a gain in power.

The 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. The goal was to identify genes whose expression shows significantly different behavior between the treated and untreated individuals over time. In a second study, we examined the effect of age on gene expression in the kidney, where samples were obtained from humans ranging in age from 27 to 92 years. Although genome-wide transcriptional changes associated with age have been studied in several model organisms [25–27], age-dependent expression in humans is unclear. The goal here was to identify genes whose expression changes significantly with respect to age. Results are obtained in both studies that corroborate previous findings and provide novel insights into these problems.

Materials and Methods

Sample preparations. Details on the protocols are described elsewhere [28, 29]. *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 [28]. Total RNA was extracted using an RNeasy kit (Qiagen, Inc). Samples from hours 4 and 6 were unavailable for one of the controls due to a sample collection error. *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 or renal transplant biopsy from patients ranging in age from 27 to 92 years. These samples were classified 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.

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 [30]. Expression values were then transformed

by taking $\log_2(\text{Data} + c)$ where c is such that the data pooled across the entire study were approximately symmetric. In both studies, this required using $c = 10$.

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

Removal of genes in the kidney aging study. The individuals obtained 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) latent variables. An initial analysis showed that genes affected by latent variables were disproportionately called significant. However, information on tissue type (cortex or medulla) and sex was available for each individual. Therefore, genes whose expression is well explained by these variables, irrespective of any age dependent behavior, can unequivocally be evaluated for temporal differential expression. We applied a novel filtering technique that sought to identify these genes, while at the same time guarding against anti-conservatively biasing the subsequent significance analysis. 34,061 such genes were identified, reducing the total number of genes included in the analysis by 10,863. Through an extensive simulation study, we were able to show that the filtering method effectively removed genes confounded by latent variables while not inflating the significance of the remaining genes (see Supplementary Information). This method is potentially applicable to other observational genome-wide expression studies. In the subsequent significance analysis, sex and tissue type indicator variables (taking values 0 or 1) were included in the parametrization of the average time curve: $\mu_i(t) = \text{sex} + \text{tissue} + \alpha_i + \boldsymbol{\beta}_i^T \mathbf{s}(t)$. This allows for different intercepts for each sex/tissue type combination.

Results

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 133 humans ranging in age from 27 to 92 years. Only one array was obtained per individual and their age at that time was recorded (see Materials and Methods). Figure 1a displays a simulated example of expression measured on a single gene in this type of study. The solid line is the population average time curve for the gene, 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 observation represents an independently sampled individual.

The goal of the kidney aging study was to determine whether each gene has expression that changes with age. According to our method, this 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: LAPTM5, a Lysosomal associated multi-spanning membrane 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, which minimizes the sum of squares among a general class of curves. 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 nonparametric data re-sampling technique.

In another study, eight human volunteers were randomly divided into endotoxin-treated and control groups of equal size (see 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.

The goal of the endotoxin study was 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 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 is fit to the combined groups,

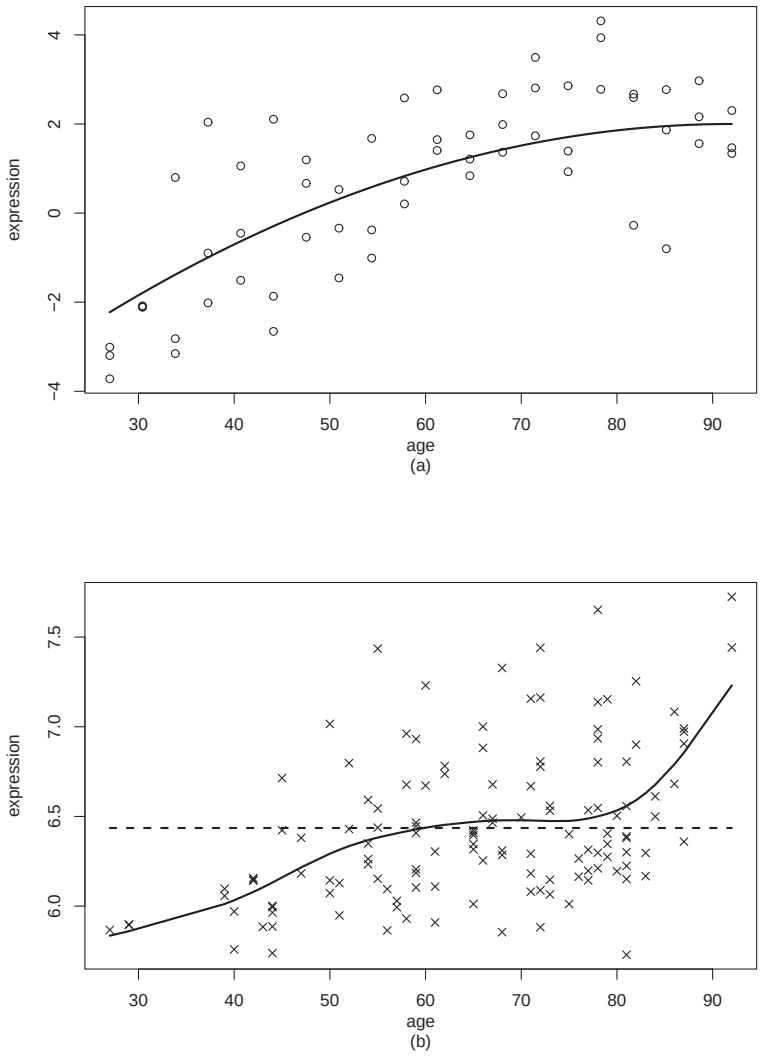

Figure 1: (a) A simulated example of expression data obtained by independent sampling. The solid line is the population average time curve and the o's are observed expression values. (b) The expression values of the most significant gene in the kidney aging study. The solid line is the curve fitted under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

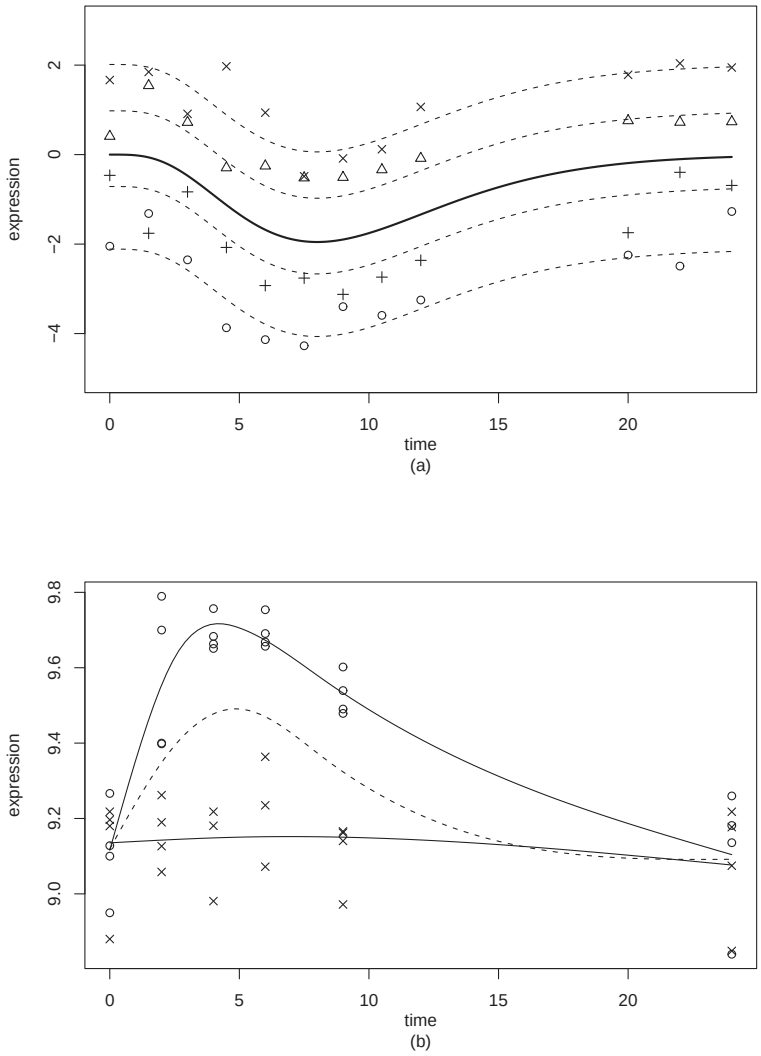

Figure 2: (a) An example of longitudinal sampling of four individuals. The solid line is the population average time curve. The dashed lines are the average time curves *for the individuals*. The points of a common shape correspond to one of the individuals. (b) The expression values of a significant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

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 its 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 much 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 substantially 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 data.

The methodology we present here 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 (see 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 $i = 1, 2, \ldots, 34061$ genes and $j = 1, 2, \ldots, 133$ 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.

It is usually straightforward to determine whether independent or longitudinal sampling has been implemented (see Supplementary Information). This distinction is relevant as it affects important details in the model and significance method. 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 $i = 1, 2, \ldots, 44924$ genes 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 also modeled as a curved will be dealt with elsewhere; however, the endotoxin study did not contain enough time points to warrant this extra level of complexity.

In both cases, the models are fit by minimizing the sum of squares between the curve and the observed expression values. (The procedure is slightly more complicated for longitudinal data – see the Supplementary Information.) The model fitting technique requires the population average time curve $\mu_i(t)$ to be 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 straightforward basis is the polynomial basis defined by $s_1(t) = t, s_2(t) = t^2, \ldots, s_p(t) = t^p$. 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. Specifically, suppose that the goal is to minimize the sum of squares from among all continuous curves rather than from an arbitrary basis. This is obviously an ill-posed problem because there are an infinite number of interpolating curves that yield a sum of squares equal to zero for a given set of data. However, if one limits the set of curves to those that are not 'too curvy', then remarkably there is a unique one that minimizes the sum of squares [31]. This curve is a natural cubic spline and can be constructed using a B-spline basis (see Supplementary Information). We developed a method to automatically choose the dimension of the basis p. The basic idea of the method 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 the endotoxin study and $p = 5$ for the kidney aging study. In both cases, the value of p that yielded the most power across all significance cut-offs was chosen even though this was not the criterion used (see Supplementary Information).

The hypothesis tests for differential expression can be written in terms of the $\mu_i(t)$ for each gene. 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 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 $\mu_i(t)$, and the alternative 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 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. 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 (see Supplementary Information).

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. Rigorous justification and exact formulas for all cases can be found in the Supplementary Information.

The null distribution of these statistics is calculated through a nonparametric method called the bootstrap [32]. 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 (see 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, 33–35]. For a given significance threshold, the false discovery rate is the proportion of false positives among all genes called significant:

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

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 (see Supplementary Information).

By applying reasonable q-value cut-offs in both studies, many relevant genes were called significant. Several of these are listed in Table 1, and many more are included in the Supplementary Information. At q-value cut-offs of 0.1%, 1%, and 5%, there are 4163, 7409, and 12720 significant genes, respectively, in the endotoxin study. This indicates that an endotoxin injection causes profound gene expression changes in blood cells. This high degree of measured significance is expected for a number of reasons. First, individuals treated with endotoxin are quickly and strongly affected, so their comparison to normal individuals may be considered a drastic one. Second, some differential expression may be due to changing distributions of cell populations in the blood in addition to transcriptional changes. Third, measuring differential expression over time is a more sensitive study design than the typical static design; *any* change over time qualifies as differential expression.

To get a broad picture of the behavior of differentially expressed genes, a singular value decomposition was performed on the 4163 most significant genes. Figure 3 shows the top two eigen-genes that explain 66% and 16% of the variance, respectively. The relevant information is the shape of

Table 1: A partial list of significant genes in each study. See the Supplementary Information for more extensive lists.

rank	gene name	gene annotation
		endotoxin study (up-regulated genes)
1	ORM1	orosomucoid 1
2	TNFAIP3	tumor necrosis factor, alpha-induced protein 3
3	GYG	glycogenin
4	CD59	CD59 antigen p18-20
5	IER3	immediate early response 3
6	ATP9A	ATPase, Class II, type 9A
7	IL1RN	interleukin 1 receptor antagonist (IL1RN)
8	TNFAIP6	tumor necrosis factor, alpha-induced protein 6
9	IL1B	interleukin 1
10	CD59	CD59 antigen p18-20
		endotoxin study (down-regulated genes)
1	RPL27A	ribosomal protein L27a
2	RPS7	ribosomal protein S7
3	RPL5	ribosomal protein L5
4	RPL13A	ribosomal protein L13a
5	SLC25A6	solute carrier family 25, member 6
6	RPS4X	ribosomal protein S4, X-linked
7	PECAM1	platelet/endothelial cell adhesion molecule
8	RPL30	ribosomal protein L30
9	EFE2	eukaryotic translation elongation factor 2
10	RPL34	ribosomal protein L34
		kidney aging study
1	LAPTM5	Lysosomal-associated multispanning membrane protein-5
2	NNMT	nicotinamide N-methyltransferase
3	TYROBP	TYRO protein tyrosine kinase binding protein
4	HLA-F	major histocompatibility complex, class I, F
5	CRABP1	cellular retinoic acid binding protein 1
6	HLA-B	major histocompatibility complex, class I
7	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1
8	C1R	complement component 1, r subcomponent
9	C1QA	complement component 1, q subcomponent, alpha polypeptide
10	HLA-F	major histocompatibility complex, class I, F
11	PIGR	polymeric immunoglobulin receptor
12	AMPD3	adenosine monophosphate deaminase (isoform E)
13	GPNMB	glycoprotein (transmembrane) nmb
14	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1
15	DKFZp761P0423	hypothetical protein DKFZp761P0423
16	C1QR1	complement component 1, q subcomponent, receptor 1
17	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1
18	MS4A7	membrane-spanning 4-domains, subfamily A, member 7
19	RGS1	regulator of G-protein signalling 1
20	KIAA1268	KIAA1268 protein

each eigen-gene, not their magnitudes or direction of differential expression. The most significant genes follow the shape of the first eigen-gene closely. Among the 4163 most significant genes 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 the gene changes in expression in both directions over the time course. Genes that are significant but not among the very top show behavior more similar to this eigen-gene.

Most of the top 50 significantly down-regulated genes are ribosomal proteins (see Table 1 and the Supplementary Information). It is well known that genes encoding ribosomal proteins are under tight transcriptional regulation. For example, a reduction in transcription has been detected in yeast after heat shock [36]. A similar pattern occurs in this endotoxin shock scenario in our study. There are number of other statistically significant genes involved in protein synthesis that are down-regulated, such as translation initiation factor (EIF3), translation elongation factor (EEF1), and poly-A binding protein. Correspondingly, genes involved in energy production, such as ATP synthase, cytochrome c oxidase, and lactate dehydrogenase are also found to be significantly down-regulated. Previous studies have shown that *in vitro* priming of peripheral blood mononuclear cells with LPS endotoxin down-regulates major histocompatibility complex class II (MHC II) molecules [37]. Here we found both MHC II DP alpha 1 and MHC II DP beta 1 to be among the most significantly down-regulated genes. This may reflect an impaired antigen presentation during endotoxin shock.

Consistent with an induced inflammatory response by endotoxin, the majority of significantly up-regulated genes are involved in the host immune defense system. The most significantly up-regulated gene is ORM1, a key acute phase plasma protein, which has been shown to be strongly induced in endotoxin stimulated monocytes [38]. Another early response gene is IER3, a gene protecting cells from induced apoptosis by Fas or tumor necrosis factor type alpha [39]. Many crucial players in cytokine production and regulation were identified, such as IL-1B, IL1RAP [40], IL1RN, CCL4 [41], NFKBIB [42], TNF-alphaIP6 [43], IL-18R, IL-8, and Toll-like receptor 5 [44]. In particular, the Toll-like receptor (TLR) is the key receptor recognizing LPS and triggering inflammatory and immune responses against pathogens [45]. These findings confirm the intensive production of pro-inflammatory cytokines, one of the hallmarks of the immune response to endotoxin shock.

In addition, other cellular and humoral immune mechanisms are up-regulated during the initial stages of the time course. CD59 is physically and functionally associated with natural cytotoxicity receptors and activates human NK cell-mediated cytotoxicity [46]. BCL6 over-expression prevents increases in reactive oxygen species and inhibits apoptosis in B-cell lymphoma cells [47]. FCAR encodes a receptor for the Fc region of IgA, and its expression triggers phagocytosis, antibody-dependent cell-mediated cytotoxicity, and stimulation of the release of inflammatory mediators

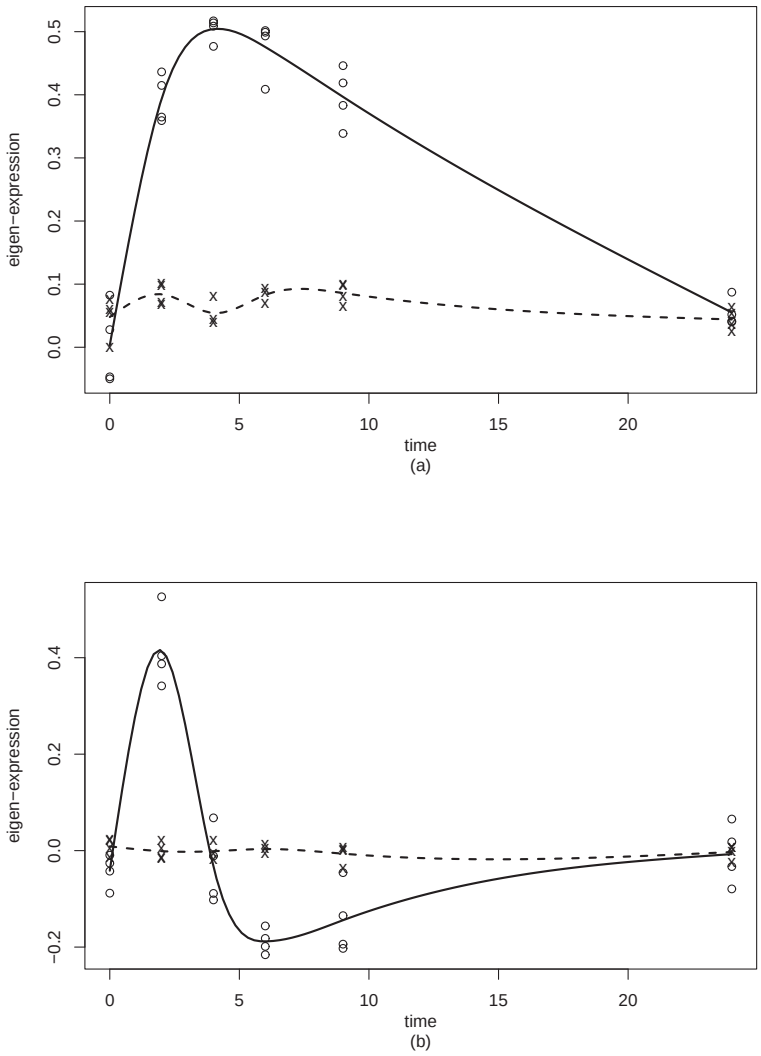

Figure 3: The top two eigen-genes obtained from genes significant at a q-value cut-off of 0.1% in the endotoxin study. The control individuals are represented by `x`'s and the treated individuals by `o`'s. The dashed curve is the model fit to the control group and the solid curve is the model fit to the treated group. (a) The first eigen-gene explains 66% of the variance. (b) The second eigen-gene explains 16% of the variance.

[48]. Furthermore, SOD2, a gene involved in destroying free superoxide radicals, was found to be significantly over-expressed. This may be involved in protecting cells from reactive oxygen species (ROS) induced by endotoxin [49]. Taken together these results suggest that in response to endotoxin shock, blood cells reallocate resources and up-regulate transcription of genes involved in various types of defensive mechanisms.

There are a number of relevant genes that were identified with q-values around 1%. Since these genes are ranked only among the top 7000, they would have easily been missed without the statistical precision of the method. For example, two relevant immune-related genes were found to be up-regulated. Interferon regulatory factor 1 (IRF1), a versatile regulatory molecule, has been shown to serve as an activator of interferons alpha and beta transcription [50] as a transcription activator of genes induced by interferons alpha, beta, and gamma [51]. Furthermore, IRF1 plays roles in regulating apoptosis and tumor-suppression [52]. Leukocyte Ig-like receptor (LIR) subfamily B, member 3 belongs to a family of inhibitory immunoreceptors expressed on many different types of immune cells [53]. Several other relevant genes involved in cell growth and apoptosis were found to be significantly up-regulated at the 1% false discovery rate level. These include an inhibitor of growth family gene (ING4), which is a tumor suppressor gene that may interact with TP53, inhibit cell growth, and induce apoptosis [54], and p53DINP1, a p53-inducible gene, which regulates p53-dependent apoptosis [55].

The genes found to be significant in the kidney aging study also provide evidence that our method identifies relevant differentially expressed genes. 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 1%, 5% and 10%, there are 189, 786 and 1821 significant genes, respectively. The top two eigengenes based on the 786 most significant genes are shown in Figure 4. (Again the shape of each eigen-gene is the meaningful piece of information.) The first eigen-gene gives a picture of the type of increasing or decreasing behavior that is present in many of the significant genes. The second eigen-gene indicates that there may be some non-monotone component to changes in expression among these significant genes.

The aging process has previously been shown to involve a widespread and complicated alteration in gene expression [25,26]. In our study, significant genes were identified that have a variety of cellular functions. Among the 100 most significant genes, 90% of them are increasing with age, and 10% are decreasing. The most significant gene is LAPTM5, a lysosomal-associated multispanning membrane protein-5. This gene has been hypothesized to have a special functional role during embryogenesis and in adult hematopoietic cells [56]. There are number of genes involved in signal transduction that show increased expression with age, including tyrosine kinase binding protein (TYROBP), regulator of G-protein signaling (RGS1), tyrosine phosphatase receptor 3 (PTPRC),

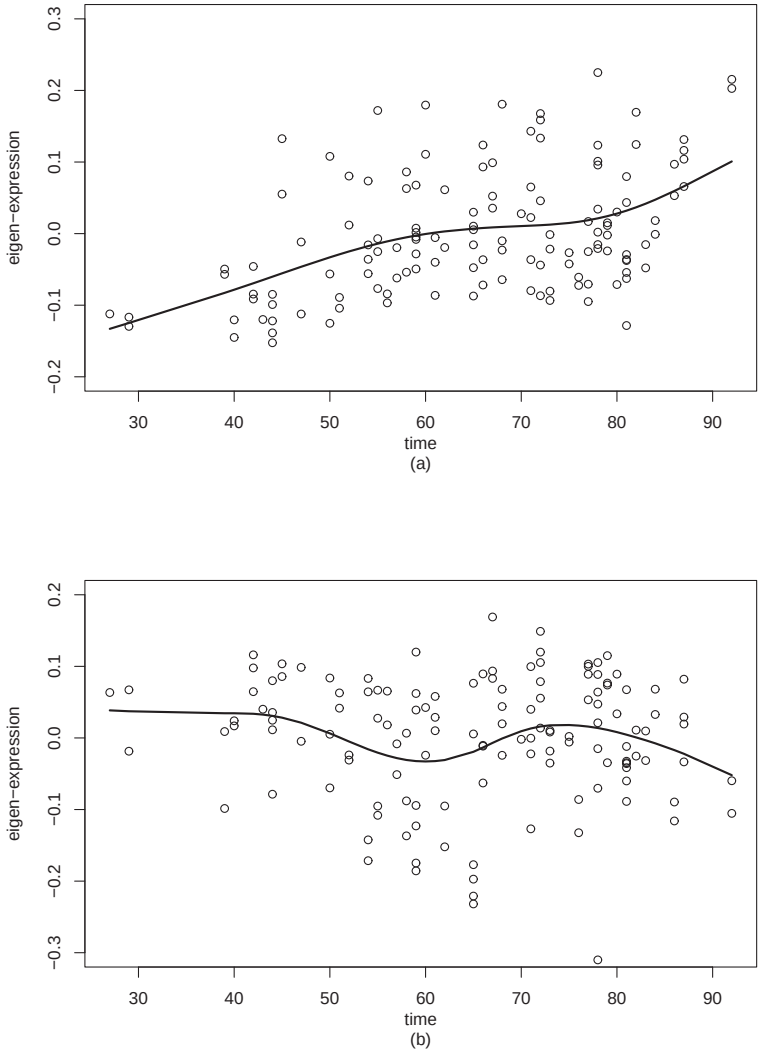

Figure 4: The top two eigen-genes obtained from genes significant at a q-value cut-off of 5% in the kidney aging study. (a) The first eigen-gene explains 44% of the variance. (b) The second eigen-gene explains 12% of the variance.

and Rho GDP dissociation inhibitor (RHGDIB). This does not contradict the existence of accelerated cellular functions during the aging process. Note that it has been shown that caloric restriction (CR) can increase life span in diverse organisms by decreasing the expression of a number of age related genes [25].

Several genes involved in immune responses show increasing expression with age, such as MHC I, MHC II, complement 1, TNF alpha -induced protein 3 (TNFAIP3), chemokine ligand 2 (CCL2), and interferon induced transmember protein (IFITM1). Previous studies have shown that changes in MHC II expression on the T-cell surface are associated with aging [57]. Complement 1 is part of the innate immune system and underlies one of the main effector mechanisms of antibody-mediated immunity, and the protein levels associated with this gene have been reported to increase with age [58]. A complex combination of physiological changes in older individuals may account for the increased expression of immune system related genes. For example, accumulated DNA damage drives normal cells to cancer cells, which activates the immune surveillance system. Also a deteriorated health condition in older individuals may make them more susceptible to bacterial infection, which in turn activates the host immune system. TINF2, another gene related to aging, is found to have increasing expression with age. This is a novel human gene essential for proper regulation of telomere length [59]. The increased expression of TINF2 coincides with the telomerase theory of aging, a theory hypothesizing that the loss of telomere in each cell division is involved in the aging process. Increased expression of TINF2 may play a role in maintaining the chromosome stability.

There are fewer genes whose expression decreases significantly with age. The most significantly decreasing gene is CRABP1, cellular retinoic acid binding protein 1, which is involved in retinoic acid-mediated differentiation and proliferation. Several other metabolic and energy production genes are down-regulated, such as adenosine monophosphate deaminase, lipoprotein lipase, dihydrolipoamide branched chain transacylase. These may be related to a slowing metabolism rate in older individuals. The results obtained by our method in this kidney aging study suggest that during the aging process a large proportion of genes have increasing expression, including those involved in signal transduction, cell growth, defense mechanisms and genome stability, while genes related to metabolism were shown to be significantly decreasing.

Discussion

We have presented a general significance method that identifies differentially expressed genes in time course microarray experiments. It was applied here to studies involving both types of sampling and both types of temporal differential expression. This method is easily applied in more complicated

situations, where three or more groups are compared, for example (see Supplementary Information). 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 we have worked out to thoroughly and broadly define this method. Moreover, we have rigorously justified 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. These findings confirm that the method developed here produces biologically meaningful results. An analysis of the the endotoxin study suggest that a reallocation of resources takes place in the blood during endotoxin shock, while genes involved in various types of defense mechanisms are up-regulated. 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, defense mechanisms and genome stability. Genes related to metabolism were found to be decreasing with age.

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/`.

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**, 98349839.

[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**, 8899.

[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.

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

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

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

[28] Inflammation and the Host Response to Injury Consortium. (2004) CR1 experiment. Data and details available at http://www.gluegrant.org/.

[29] 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) A transcriptional profile of aging in the human kidney. Unpublished manuscript. See `http://cmgm.stanford.edu/~kimlab/index_publications.html`.

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

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

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

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

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

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

[36] Zengel, J. & Lindahl, L. (1985) *J. Bacteriol.* **163**, 140–147.

[37] Wolk, K., Dcke, W. D., von Baehr, V., Volk, H. D., & Sabat, R. (2000) *Blood* **96**, 218–223.

[38] Dente, L., Pizza, M., Metspalu, A., & Cortese, R. (1987) *EMBO J.* **6**, 2289–96.

[39] Wu, M. (2003) *Apoptosis* **8**, 11–18.

[40] Radons, J., Gabler, S., Wesche, H., Korherr, C., Hofmeister, R., & Falk, W. (2002) *J. Biol. Chem.* **277**, 16456–16463.

[41] Nakajima, T., Inagaki, N., Tanaka, H., Tanaka, A., Yoshikawa, M., Tamari, M., Hasegawa, K., Matsumoto, K., Tachimoto, H., Ebisawa, M., Tsujimoto, G., Matsuda, H., Nagai, H., & Saito, H. (2002) *Blood* **100**, 3861–3868.

[42] Castro-Alcaraz, S., Miskolci, V., Kalasapudi, B., Davidson, D., & Vancurova, I. (2002) *J. Immunol.* **169**, 3947–3953.

[43] Klampfer, L., Lee, T., Hsu, W., Vilcek, J., & Chen-Kiang, S. (1994) *Mol. Cell. Biol.* **14**, 6561–6569.

[44] Yu, Y., Zeng, H., Lyons, S., Carlson, A., Merlin, D., Neish, A., & Gewirtz, A. (2003) *Am. J. Physiol. Gastrointest. Liver Physiol.* **285**, G282–G290.

[45] Yamamoto, M., Yamazaki, S., Uematsu, S., Sato, S., Hemmi, H., Hoshino, K., Kaisho, T., Kuwata, H., Takeuchi, O., Takeshige, K., Saitoh, T., Yamaoka, S., Yamamoto, N., Yamamoto, S., Muta, T., Takeda, K., & S., A. (2004) *Nature* **8**, 218–22.

[46] Marcenaro, E., Augugliaro, R., Falco, M., Castriconi, R., Parolini, S., Sivori, S., Romeo, E., Millo, R., Moretta, L., Bottino, C., & Moretta, A. (2003) *Eur. J. Immunol.* **33**, 3367–3376.

[47] Kurosu, T., Fukuda, T., Miki, T., & Miura, O. (2003) *Oncogene* **22**, 4459–4468.

[48] Herr, A., Ballister, E., & Bjorkman, P. (2003) *Nature* **423**, 614–620.

[49] Spolarics, Z. (1996) *Am J Physiol.* **270**, G660–6.

[50] Harada, H., Fujita, T., Miyamoto, M., Kimura, Y., & Maruyama, M. (1989) *Cell* **58**, 729–739.

[51] Storm van's Gravesande, K., Layne, M., Ye, Q., Le, L., & Baron, R. (2002) *J. Immunol.* **168**, 4488–4494.

[52] Sanceau, J., Boyd, D., Seiki, M., & Bauvois, B. (2002) *J. Biol. Chem.* **277**, 35766–35775.

[53] Borges, L., Hsu, M., Fanger, N., Kubin, M., & Cosman, D. (1997) *J. Immunol.* **159**, 5192–5196.

[54] Shiseki, M., Nagashima, M., Pedeux, R., & Kitahama-Shiseki, M. (2003) *Cancer Res.* **63**, 2373–2378.

[55] Tomasini, R., Samir, A., Carrier, A., Isnardon, D., & Cecchinelli, B. (2003) *J. Biol. Chem.* **278**, 37722–37729.

[56] Adra, C., Zhu, S., Ko, J., Guillemot, J., Cuervo, A., Kobayashi, H., Horiuchi, T., Lelias, J., Rowley, J., & Lim, B. (1996) *Genomics* **15**, 328–37.

[57] Malinowski, K. & Rapaport, F. (1995) *Cell Immunol* **162**, 68–73.

[58] Nagaki, K., Hiramatsu, S., Inai, S., & Sasaki, A. (1980) *J Clin Lab Immunol.* **3**, 45–50.

[59] Kim, S., Kaminker, P., & Campisi, J. (1999) *Nat. Genet.* **23**, 405–412.

SUPPLEMENTARY INFORMATION:

A Significance Method for Time Course Microarray Experiments Applied to Two Human Studies

John D. Storey[†‡*], Jeffrey T. Leek[†], Wenzhong Xiao[§], James Y. Dai[†], Ron W. Davis[§]

August 2004

† Department of Biostatistics and ‡ Department of Genome Sciences, University of Washington, Seattle WA 98195.

§ Stanford Genome Technology Center and Department of Biochemistry, Stanford University, Stanford CA 94305.

∗ To whom correspondence should be sent: jstorey@u.washington.edu

Contents

1 Data and notation

Let $y_{ij}(t)$ be the relative expression level of gene i on individual j at time t, where there are $i = 1, 2, \ldots, M$ genes on each array and $j = 1, 2, \ldots, N$ different individuals sampled. The data are modeled as

$$y_{ij}(t) = \mu_i(t) + \gamma_{ij}(t) + \epsilon_{ij}(t).$$

The population average time curve for gene i is $\mu_i(t)$. (If the individuals consist of more than one biological group, a further index has to be placed on $\mu_i(t)$ to let each group have a different population average time curve.) Individuals deviate from the population average time curve by $\gamma_{ij}(t)$, implying that $\mu_i(t) + \gamma_{ij}(t)$ is the individual average time curve for individual j. The measurement error and remaining sources of random variation are modeled by $\epsilon_{ij}(t)$. There are T_j time points (and arrays) observed on individual j. Let t_{jk} be the time at which the kth array for individual j is obtained, $k = 1, 2, \ldots, T_j$. Let $y_{ijk} \equiv y_{ij}(t_{jk})$ be the observed relative expression value of gene i on the kth array obtained from individual j, and similarly denote $\epsilon_{ijk} \equiv \epsilon_{ij}(t_{jk})$. We model these ϵ_{ijk} as independent random variables with mean zero and variance σ_i^2. In this work we model the individual-specific random effects as independent scalar random variables γ_{ij} that have mean zero and variance τ_i^2. A more general treatment of $\gamma_{ij}(t)$ will be given elsewhere.

The observed expression level of gene i on individual j at time t_{jk} is written as

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

Independent sampling can be written as a special case of longitudinal sampling by noting that it implies $T_j = 1$ for all $j = 1, \ldots, N$. Therefore, the k index can be dropped and an independently sampled observation can be written as

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

In this case, one can equivalently write

$$y_{ij} = \mu_i(t_j) + \epsilon'_{ij},$$

where $\epsilon'_{ij} = \gamma_{ij} + \epsilon_{ij}$ are independent random variables with mean zero and variance $\tau_i^2 + \sigma_i^2$. Thus, the independent sampling model is essentially the longitudinal model with the individual average curve $\gamma_{ij}(t)$ removed, and with an error term that captures a broader spectrum of variation.

As an example, Figure 1 shows a gene found to be significant in the endotoxin study. A gray line connects the observations for a single control individual. It can be seen that these are among the largest values in the controls across the entire time course, giving evidence that there is in fact a random effect for individuals. Moreover the elevation appears to be constant across the time course, which supports the use of a scalar shift for individual random effects.

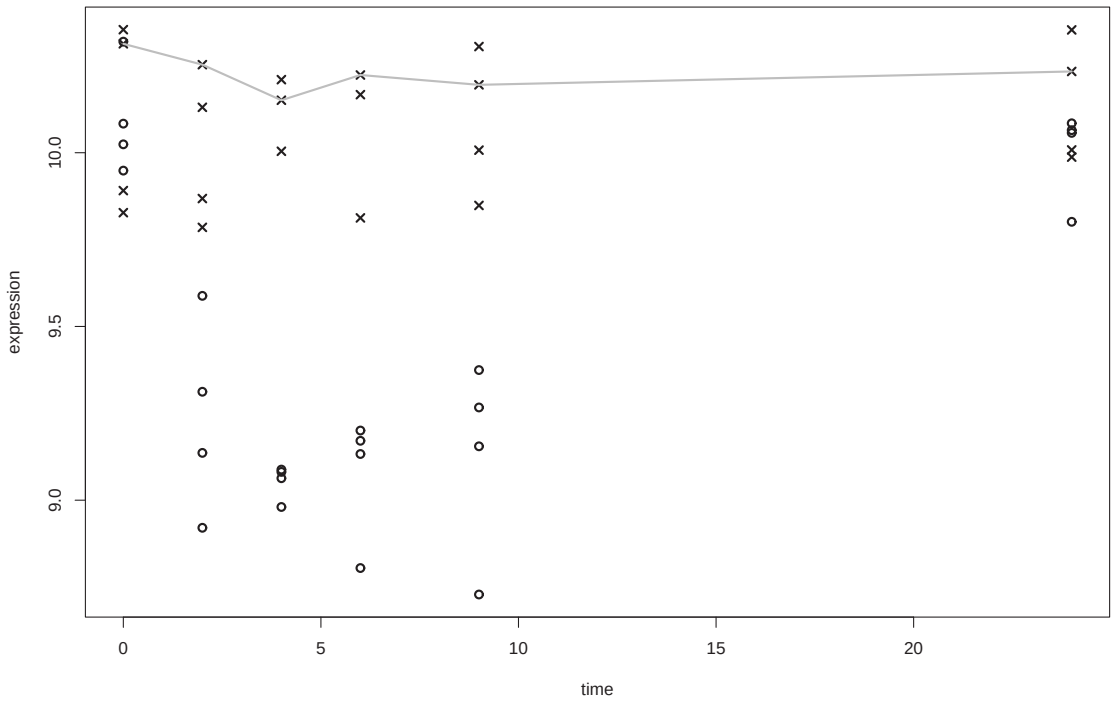

Figure 1: A highly significant gene in the endotoxin study. The controls are represented by `x`'s and the treated individuals by `o`'s. The control and treated groups do not show significant differential expression at times 0 and 24. However, there is a substantial down regulation of expression in between these hours for endotoxin treated individuals. The observations corresponding to one of the controls are connected with a solid grey line throughout the time course to clearly shows their temporal correlation.

2 Basis representations of population average time curves

The population average time curves $\mu_i(t)$ are written in terms of an intercept plus a p-dimensional linear basis:

$$\mu_i(t) = \beta_{i0} + \beta_{i1}s_1(t) + \beta_{i2}s_2(t) + \cdots + \beta_{ip}s_p(t)$$

where $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_p(t)]^T$ is a pre-specified p-dimensional basis and $\boldsymbol{\beta}_i = [\beta_{i0}, \beta_{i1}, \beta_{i2}, \ldots, \beta_{ip}]^T$ is a $(p+1)$-dimensional vector of unknown gene-specific parameters. (Note that in the main article we denoted the intercept by α_i – here we write it as β_{i0} for notational convenience.)

A straightforward basis is the polynomial basis: $s_1(t) = t, s_2(t) = t^2, \ldots, s_p(t) = t^p$. That is, $\mu_i(t)$ is modeled as a polynomial of degree p,

$$\mu_i(t) = \beta_{0i} + \beta_{i1}t + \beta_{i2}t^2 + \cdots + \beta_{ip}t^p.$$

Using this basis is a restrictive assumption that one may not want to make. Instead, a more flexible approach is possible that also allows $\mu_i(t)$ to be written in terms of a p-dimensional basis. Suppose that a curve is fit that minimizes the sum of squares among all curves $\mu_i(t)$ with continuous second derivatives. Unfortunately there are an infinite number of such curves that will interpolate the data, thereby over-fitting the data. Therefore, one can penalize for the "curviness" of the $\mu_i(t)$ by making the restriction that

$$\int \left|\mu_i''(t)\right|^2 dt \leq c$$

for some c. The parameter c allows one to balance the trade-off between minimizing the sum of squares and preventing over-fitting. As $c \to \infty$ the fitted curve interpolates the data, and as $c \to 0$ the fitted curve becomes the linear least squares fit.

Surprisingly, for each fixed c there is a unique and easily constructed solution to this optimization criterion. The solution is called a natural cubic spline, and it can be parameterized by a B-spline basis (Green & Silverman 1994). If a B-spline basis is used to parameterize $\mu_i(t)$, then finding the values of $\boldsymbol{\beta}_i$ to minimize the sum of squares is equivalent to achieving the above optimality criterion for some c. That is, this parametrization finds the best continuous curve among all of those that meet the "curviness" restriction. A fit obtained from a B-spline basis of dimension p corresponds to the optimal solution for some c, where the larger p is, the larger its corresponding c is. A p-dimension B-spline basis is equivalent to a p-dimensional polynomial basis in terms of degrees of freedom of the model. The natural cubic spline fit is more flexible but less interpretable; however, interpretability of fitted parameters does not play a role in identifying differentially expressed genes.

Figure 2 shows an example of B-spline basis functions when $p = 4$. It is straightforward to construct such a basis, and most statistical software packages automatically will do so. However, one has to choose where to place the p knots that anchor these basis functions. A common approach

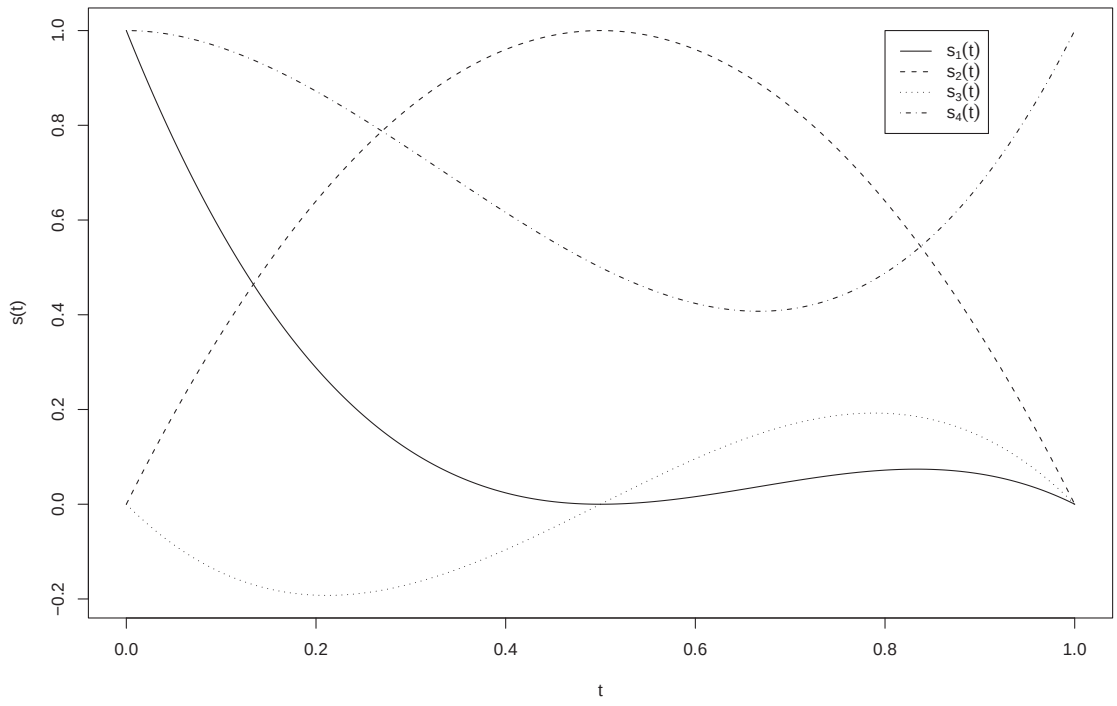

Figure 2: An example of a 4-dimensional B-spline basis. Each function is defined over the entire range of the time course, although they are scale here to lie between 0 and 1.

is to place the knots at equal intervals over the range of time. This approach does not always work well when there is irregular sampling as in the endotoxin study. Therefore, our method places the knots at the $0, \frac{1}{p-1}, \frac{2}{p-1}, \ldots, \frac{p-1}{p-1}$ quantiles of the observed time points. Note that if the time points are evenly spaced then the two approaches are essentially equivalent.

We applied both bases to the endotoxin and kidney aging studies, and in each case a p-value for each gene was calculated by the method outlined in Section 7. We applied a Kolmogorov-Smirnov test to compare the two sets of p-values. Over a range of reasonable basis dimensions p, the p-values resulting from the B-spline basis were highly statistically significant for being stochastically smaller than the p-values from the polynomial basis. In other words, the pooled p-values were uniformly smaller when using the B-spline basis, which results in an increase in realized power. This comparison is of course done under the assumption that the p-values calculated under either basis are valid (i.e., null genes have uniformly distributed p-values). However, as discussed in Section 6, we performed extensive simulations to verify this.

3 Model fitting

The population average time curve $\mu_i(t)$ can be estimated by minimizing a sum of squares criterion, regardless of whether the polynomial or B-spline basis is used. Therefore, it is possible to provide a unified framework for model fitting that is applicable to either basis. The least squares criterion is common in model fitting, and it can be justified by the Gauss-Markov theorem (Weisberg 1985, Rice & Wu 2001). Also, under the assumption that the random variation follows the Normal distribution this criterion produces maximum likelihood estimates, although this assumption is not necessary for our procedure.

Recall that null and alternative models are fit to each gene. For within class temporal differential expression, the null model is estimated under the constraint that $\mu_i(t) = \beta_{i0}$, and the alternative under the general parametrization $\mu_i(t) = \beta_{i1}s_1(t) + \beta_{i2}s_2(t) + \cdots + \beta_{ip}s_p(t)$. In the case of between class temporal differential expression, the null model is estimated under the constraint that $\mu_i^1(t) = \cdots = \mu_i^G(t)$, which is accomplished by fitting one model to all G groups combined. The alternative model is estimated by fitting a separate curve to each group. The model fitting procedure depends on which type of sampling has been used to obtain the arrays. In the case of independent sampling, the standard least squares technique can always be used. In the case of longitudinal sampling, an easily modified version of this can be used when testing for within class temporal differential expression, or when testing for between class temporal differential expression under the assumption that the intercepts are the same among all groups (as in the endotoxin study). If the intercepts are included in the analysis, then a more computationally intensive iterative least squares technique must be used. These different cases are summarized in Table 1, and we explain each one in the remainder of this section.

Remark. In Section 1, we defined N to be the total number of observed individuals. However, for notational convenience in this section, the number of individuals that a curve is being fitted to is written as N. For example, suppose that two groups of 10 individuals are examined for between class differential expression. Under the null hypothesis there is one curve fitted to $N = 20$ individuals, but under the alternative hypothesis there are two curves fitted to $N = 10$ individuals each.

Independent sampling.

In the case of independent sampling, each individual is observed at only one time point, which implies that $T_j = 1$ for $j = 1, 2, \ldots, N$. Therefore, to discuss this case we drop the k subscript. A

Table 1: The model fitting techniques used for different sampling schemes and types of comparisons.

Sampling	Comparison	Intercept included?	Fitting method
Independent	Within class	Yes or No	Least squares
Independent	Between class	Yes	Least squares
Independent	Between class	No	Least squares
Longitudinal	Within class	Yes or No	Individual centered least squares
Longitudinal	Between class	Yes	EM least squares
Longitudinal	Between class	No	Individual centered least squares

curve is fitted by finding the values of $\boldsymbol{\beta}_i$ that minimize the sum of squares

$$\sum_{j=1}^{N} \left(y_{ij} - [\beta_{i0} + \beta_{i1} s_1(t) + \beta_{i2} s_2(t) + \cdots + \beta_{ip} s_p(t)]\right)^2 .$$

These are easily calculated by using standard least squares regression techniques. First, an $N \times (p+1)$ matrix $\mathbf{X}$ is formed such that the jth row is equal to $[1, s_1(t_j), s_2(t_j), \ldots, s_p(t_j)]$. The fitted parameters are then

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}_i, \tag{1}$$

where $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iN}]^T$. The estimated curve can be evaluated at any time point by

$$\widehat{\mu}_i(t) = \widehat{\beta}_{i0} + \widehat{\beta}_{i1} s_1(t) + \widehat{\beta}_{i2} s_2(t) + \cdots + \widehat{\beta}_{ip} s_p(t),$$

which implies that the fitted value corresponding to observation y_{ij} is $\widehat{\mu}_i(t_j)$. Whereas the above is the estimator used for fitting the general curve, the null model for within class differential expression is a simpler special case: $\widehat{\mu}_i(t) = \overline{y}_{i\cdot} = \sum_{j=1}^{N} y_{ij}/N$.

In the case that between class differential expression is tested and the intercepts are assumed to be equal among all groups, a similar procedure is followed. Define $\mathbf{S}$ to be the $N \times p$ matrix such that the jth row is equal to $[s_1(t_j), s_2(t_j), \ldots, s_p(t_j)]$. Also, let $\overline{y}_i$ be the average expression value of gene i across *all* groups. Then we remove β_{i0} from $\boldsymbol{\beta}_i$ and calculate

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{S}^T\mathbf{S}\right)^{-1} \mathbf{S}^T \left(\mathbf{y}_i - \overline{y}_i \mathbf{1}_N\right), \tag{2}$$

where $\mathbf{1}_N$ is a vector of N 1's.

Longitudinal sampling.

Longitudinal sampling presents the additional challenge that expression measurements for a given individual are dependent. This is empirically seen in the endotoxin study in Figure 1. Recall that $y_{ijk} = \mu_i(t_{jk}) + \gamma_{ij} + \epsilon_{ijk}$, where γ_{ij} and ϵ_{ijk} are identically and independently distributed random variables with means equal to zero and variances equal to τ_i^2 and σ_i^2, respectively. Therefore, for $k \neq k'$, it follows that $\mathbf{Cov}\left(y_{ijk}, y_{ijk'}\right) = \mathbf{Cov}\left(\gamma_{ij}, \gamma_{ij}\right) = \tau_i^2$. Moreover, τ_i^2 cannot be directly estimated because the γ_{ij} are unobserved. There are two approaches to deal with this complication. The first is to mean center the expression measurements for each individual, which removes the random effect γ_{ij} but also removes the intercept β_{i0}, and then apply standard least squares estimation. This approach can only be used when the intercepts are also not necessary to include in the significance analysis (see Table 1). The second approach is to treat the γ_{ij} as missing data, and then apply least squares estimation via the EM algorithm. The estimated covariance structure can then be used to find the least squares estimates.

Longitudinal sampling: Notation. We first introduce some detailed notation used in describing the two longitudinal fitting procedures. Let $\mathbf{y}_{ij} = [y_{ij1}, y_{ij2}, \cdots, y_{ijT_j}]^T$ be the vector of observations for gene i and individual j. The vector of observed measurements for gene i is denoted by

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \\ \vdots \\ \mathbf{y}_{iN} \end{pmatrix}.$$

Recall that this vector includes only the individuals that are involved in the current curve fitting. The vectors $\boldsymbol{\epsilon}_{ij}$ and $\boldsymbol{\epsilon}_i$ are analogously defined from the ϵ_{ijk}, and $\boldsymbol{\gamma}_i$ from the individual random effects γ_{ij}. Also, define $\mathbf{S}_j$ to be the $T_j \times p$ matrix $\mathbf{S}_j$ where row k is equal to $[s_1(t_{jk}), s_2(t_{jk}), \cdots, s_p(t_{jk})]$, $k = 1, \ldots, T_j$. The individual's $\mathbf{S}_j$ are combined to obtain an overall matrix of the evaluated basis functions:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_N \end{pmatrix}.$$

In order to be able to include the intercept, we define

$$\mathbf{X}_j = \left(\mathbf{1}_{T_j}\ \mathbf{S}_j\right),$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{pmatrix}.$$

Finally, define $\mathbf{R}$ to be the $\sum_{j=1}^{N} T_j \times N$ block diagonal matrix composed of the $\mathbf{1}_{T_j}$, where $\mathbf{1}_{T_j}$ is a vector of T_j 1's. These collapsed versions of the data then have the relationship that

$$\mathbf{y}_{ij} = \mathbf{X}_j \boldsymbol{\beta}_i + \gamma_{ij} \mathbf{1}_{T_j} + \boldsymbol{\epsilon}_{ij}$$

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{R}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i,$$

which greatly simplifies the following algorithms.

Longitudinal sampling: Individual centered least squares. The average expression value of individual j is $\overline{y}_{ij\cdot} = \sum_{k=1}^{T_j} y_{ijk}/T_j$. Define $y^c_{ijk} = y_{ijk} - \overline{y}_{ij\cdot}$ to be the mean-centered observation for gene i, individual j, and time point k. We can similarly center the errors $\epsilon^c_{ijk} = \epsilon_{ijk} - \overline{\epsilon}_{ij\cdot}$ and the basis $\mathbf{s}^c_{jk} = \mathbf{s}(t_{jk}) - \sum_{k'=1}^{T_j} \mathbf{s}(t_{jk'})/T_j$. Also, removing β_{i0} from $\boldsymbol{\beta}_i$, it is easily seen that $y^c_{ijk} = \boldsymbol{\beta}_i^T \mathbf{s}^c_{jk} + \epsilon^c_{ijk}$. The intercept β_{i0} and individual random effect γ_{jk} have been subtracted out while the parameters of interest $\boldsymbol{\beta}_i$ have remained intact. This relationship can be written similarly in vector format,

$$\mathbf{y}^c_i = \mathbf{S}^c \boldsymbol{\beta}_i + \boldsymbol{\epsilon}^c_i.$$

It turns out that $\boldsymbol{\beta}_i$ is estimated by

$$\widehat{\boldsymbol{\beta}}_i = \left[(\mathbf{S}^c)^T \mathbf{S}^c\right]^{-1} (\mathbf{S}^c)^T \mathbf{y}^c_i, \tag{3}$$

which is completely analogous to the independent sampling estimate in equation (1).

This solution is surprisingly simple even though the components of $\mathbf{y}^c_{ij}$ are linearly dependent. However, the result straightforwardly follows from a conditional likelihood property shown in Verbeke & Molenberghs (2000), for example. It is shown there that minimizing the sum of squares

$$\left(\mathbf{y}^c_i - \mathbf{S}^c \boldsymbol{\beta}_i\right)^T \left(\mathbf{y}^c_i - \mathbf{S}^c \boldsymbol{\beta}_i\right)$$

(i.e., minimizing the original sum of squares given the $\overline{y}_{ij\cdot}$, $j = 1, 2, \ldots, N$) is equivalent to minimizing

$$\left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i\right)^T \mathbf{A} \left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i\right), \tag{4}$$

where $\mathbf{A}$ is the block diagonal matrix composed of

$$\mathbf{A}_j = \mathbf{I}_{T_j} - \mathbf{1}_{T_j} \left(\mathbf{1}^T_{T_j} \mathbf{1}_{T_j}\right)^{-1} \mathbf{1}^T_{T_j},$$

for $j = 1, 2, \ldots, N$. By standard results, the least squares solution to (4) is

$$\widehat{\boldsymbol{\beta}}_i = \left[\mathbf{X}^T\mathbf{A}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}_i.$$

Since $\mathbf{A}^T\mathbf{A} = \mathbf{A}$, $\mathbf{A}\mathbf{y}_i = \mathbf{y}_i^c$, and $\mathbf{A}\mathbf{X} = (\mathbf{0}\ \mathbf{S}^c)$, it follows that

$$\left[\mathbf{X}^T\mathbf{A}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y}_i = \left[(\mathbf{S}^c)^T\mathbf{S}^c\right]^{-1}(\mathbf{S}^c)^T\mathbf{y}_i^c,$$

which is exactly the solution in (3).

Longitudinal sampling: EM algorithm least squares. Here we present the model fitting algorithm when the intercepts are included in the significance analysis. If the γ_{ij} were observed, then the goal would be to find $\boldsymbol{\beta}_i$ that minimizes the sum of squares

$$\sum_{j=1}^{N}\sum_{k=1}^{T_j}\left(y_{ijk} - \boldsymbol{\beta}_i^T\mathbf{s}(t_{jk}) - \gamma_{ij}\right)^2 = \left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{R}\boldsymbol{\gamma}_i\right)^T\left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{R}\boldsymbol{\gamma}_i\right),$$

which has the solution

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y}_i - \mathbf{R}\boldsymbol{\gamma}_i\right)$$

However, the γ_{ij} are not observed. Thus, taking the dependence into account the goal is to minimize

$$\left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{R}\boldsymbol{\gamma}_i\right)^T\Sigma_i^{-1}\left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{R}\boldsymbol{\gamma}_i\right),$$

where Σ_i is the covariance matrix of the y_{ijk}. Σ_i is a block diagonal matrix composed of $T_j \times T_j$ matrices Σ_{ij}, where Σ_{ij} is constructed so that its diagonal elements are equal to $\sigma_i^2 + \tau_i^2$ and the off-diagonal elements equal to τ_i^2. Since Σ_i is unknown, it has to be estimated simultaneously with $\boldsymbol{\beta}_i$. In order to do this, we employ the EM algorithm based on the Normal likelihood (although the Normal distribution is not an assumption that is made for the significance analysis).

For the maximization (M) step of the algorithm, the γ_{ij} are assumed to be known, which leads to the following maximum likelihood estimates:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_i &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y}_i - \mathbf{R}\boldsymbol{\gamma}_i\right) \\
\widehat{\tau}_i^2 &= \frac{1}{N}\sum_{j=1}^{N}\gamma_{ij}^2 \\
\widehat{\sigma}_i^2 &= \frac{1}{\sum_{j=1}^{N}T_j}\left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{R}\boldsymbol{\gamma}_i\right)^T\left(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i - \mathbf{R}\boldsymbol{\gamma}_i\right)
\end{aligned}$$

For the expectation (E) step of the algorithm, the expectation of γ_{ij} is computed conditional on $\boldsymbol{\beta}_i$, τ_i^2, σ_i^2 and the data:

$$\mathbf{E}\left[\gamma_{ij}|\mathbf{y}_i, \boldsymbol{\beta}_i, \tau_i^2, \sigma_i^2\right] = \frac{\tau_i^2}{\tau_i^2 + \sigma_i^2/T_j} \times \frac{\sum_{k=1}^{T_j}\left[y_{ijk} - \mu_i(t_{jk})\right]}{T_j}.$$

The algorithm proceeds according to four steps.

1. Set an initial value for the individual random effects $\widehat{\boldsymbol{\gamma}}_i^{(0)}$.
2. Evaluate the estimates given the current value $\widehat{\boldsymbol{\gamma}}_i^{(m)}$ to produce the estimates $\widehat{\boldsymbol{\beta}}_i^{(m)}, \widehat{\tau}_i^{2\ (m)}, \widehat{\sigma}_i^{2\ (m)}$.
3. Set $\widehat{\gamma}_{ij}^{(m+1)} = \mathbf{E}\left[\gamma_{ij} | \mathbf{y}_i, \widehat{\boldsymbol{\beta}}_i^{(m)}, \widehat{\tau}_i^{2\ (m)}, \widehat{\sigma}_i^{2\ (m)}\right]$ for $j = 1, \ldots, N$.
4. Repeat steps 1-3 until convergence of all estimates in terms of the overall likelihood.

Note that an EM algorithm is performed on all M genes simultaneously. This is obviously more computationally intensive that the previous model fitting procedures. However, as each gene's EM algorithm converges, that gene is removed from further iterations. This saves a substantial amount of computational time since, in our experience with the endotoxin study, most of the genes converge in a few steps. Unfortunately, for any given gene an infinite Normal likelihood is obtained if $\widehat{\tau}_i^2 \to 0$, causing the EM algorithm to produce nonsensical estimates. Therefore, one has to take the usual precautions to prevent this from happening (Little & Rubin 2002). Future work will attempt to borrow strength across genes to decrease the convergence times and stabilize the EM algorithms.

4 Determining the dimension of the basis

Recall that the population average time curves $\mu_i(t)$ are parameterized by

$$\mu_i(t) = \beta_{i0} + \beta_{i1} s_1(t) + \beta_{i2} s_2(t) + \cdots + \beta_{ip} s_p(t),$$

where $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_p(t)]^T$ is a pre-specified p-dimensional basis. Here we describe a method for automatically choosing the dimension of the basis p. The basic idea of the approach is to take a singular value decomposition of the data and extract the top few "eigen-genes", which are the eigen-vectors in the gene space (Alter et al. 2000). 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 on all of these. The following algorithm is used to implement the procedure.

1. Apply a singular value decomposition to the expression data, and extract the top L eigen-genes.
2. Within each biological group (suppose there are G groups), apply the appropriate model fitting technique to each of the top L eigen-genes over a range of p.
3. For each eigen-gene and biological group, determine the optimal p by generalized cross-validation (Green & Silverman 1994). Call these $\widehat{p}_{ij}$ for $i = 1, \ldots, L$ and $j = 1, \ldots, G$.

4. Use the dimension $\widehat{p} = \max_{i,j}\{\widehat{p}_{ij}\}$ to fit curves to each gene.

In the endotoxin study where there are $G = 2$ biological groups, we considered all possible choices $p = 1, \ldots, 5$ and the top $L = 5$ eigen-genes. In the kidney aging study where there is $G = 1$ biological group, we considered the range $p = 1, 2, \ldots, 9$ and the top $L = 5$ eigen-genes. (Note that the intercept is not included in the basis, so $p = 1$ corresponds to a line.) The algorithm chose $\widehat{p} = 4$ in the endotoxin study and $\widehat{p} = 5$ in the kidney aging study. For each study, we tested whether this value of p provided the most powerful choice in an overall sense. In both studies, the p-values resulting from the automatically chosen p were highly significant for being stochastically less than the p-values from other choices of p, according to a Kolmogorov-Smirnov test (Lehmann 1975). All comparisons in both studies were significant at levels $< 10^{-5}$, except for the comparison with $p = 5$ in the endotoxin study, which was significant at level < 0.04. Since the p-values resulting from the automatically chosen $\widehat{p}$ were stochastically less than other choices of p, this means that more p-values are significant at each p-value cut-off, which is empirical evidence for having the highest power.

The motivation for setting $\widehat{p} = \max_{i,j}\{\widehat{p}_{ij}\}$ is so that the algorithm chooses a single p that is aimed at being of sufficient size for all genes. If $\widehat{p}$ is taken too large, then one loses power because degrees of freedom in the model are wasted. If $\widehat{p}$ is taken too small, then power is also lost because the truly differentially expressed genes are not adequately modeled. Therefore, p would ideally be chosen on a gene-by-gene basis since the best value may be different from gene to gene. However, we found that applying the algorithm to each gene resulted in over-fitting the data and artificially inflating the significance. Therefore, we decided to apply one value of p to all genes. We also chose to err on the side of taking p too large. The bootstrap method for simulating null statistics requires that the observed residuals be accurate approximations to the true residuals. If p is taken too small, then this may not be the case. Simulations were performed to verify that using $\widehat{p}$ in the model fitting yields valid null statistics – see Section 6.

5 Constructing the statistics

A statistic is formed for each gene that is an analogue to the t- and F-statistics, which are commonly used in the static differential expression case. The statistic compares the goodness of fit of the model under the null hypothesis to that under the alternative hypothesis. Suppose that for the gene i, SS_i^0 is the sum of squares of the residuals obtained from the model fit under the null hypothesis, and SS_i^1 is the analogous quantity under the alternative hypothesis. The statistic for gene i is defined to be

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

This is proportional to the typical F-statistic used in comparing two linear models. Sometimes the above F_i is multiplied by $\mathrm{df}_1/(\mathrm{df}_0 - \mathrm{df}_1)$ where df_0 and df_1 are the degrees of freedom of the model under the null and alternative hypotheses, respectively. This adjustment is made when the random deviations are assumed to follow the Normal distribution so that null statistics can be assumed to follow the F-distribution. However, we do not make that assumption here and the above simpler version is used. The intuition behind this statistic is that $\mathrm{SS}_i^0 - \mathrm{SS}_i^1$ quantifies in the increase in goodness of fit, and dividing this difference by SS_i^1 provides exchangeability of the F_i among the genes.

Let $\widehat{y}_{ijk}^0$ and $\widehat{y}_{ijk}^1$ be the fitted values under the null and alternative models, respectively, corresponding to observed expression value y_{ijk}. In the vector format defined in Section 3, we write all fitted values for gene i as $\widehat{\mathbf{y}}_i^0$ and $\widehat{\mathbf{y}}_i^1$ where these again correspond to $\mathbf{y}_i$. Under longitudinal sampling where the intercept is not of interest, and under independent sampling, the sums of squares are

$$\mathrm{SS}_i^0 = \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^0\right)^T \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^0\right),$$
$$\mathrm{SS}_i^1 = \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^1\right)^T \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^1\right).$$

More explicitly, under independent sampling we have

$$\mathrm{SS}_i^0 = \sum_{j=1}^{N} \left[y_{ij} - \widehat{y}_{ij}^0\right]^2, \ \mathrm{SS}_i^1 = \sum_{j=1}^{N} \left[y_{ij} - \widehat{y}_{ij}^1\right]^2,$$

where the k subscript has been dropped as before. Under longitudinal sampling, where the intercept is not of interest,

$$\mathrm{SS}_i^0 = \sum_{j=1}^{N} \sum_{k=1}^{T_j} \left[y_{ijk} - \widehat{y}_{ijk}^0\right]^2, \ \mathrm{SS}_i^1 = \sum_{j=1}^{N} \sum_{k=1}^{T_j} \left[y_{ijk} - \widehat{y}_{ijk}^1\right]^2.$$

Note that the centered data or the original data produce the same sums of squares.

For the longitudinal case, where the intercepts are included in the analysis, the covariance structure of the data are taken into account when forming the statistic. We use a statistic that is proportional to the generalized likelihood ratio statistic that follows from the Normal likelihood (Mardia et al. 1980). As defined in Section 3, let Σ_i be the covariance matrix of the y_{ijk}. Further, let $\widehat{\Sigma}_i$ be the estimate produced by plugging in the $\widehat{\sigma}_i^2$ and $\widehat{\gamma}_i^2$ that were formed under the alternative hypothesis model. Then

$$F_i = \left[\left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^0\right) - \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^1\right)\right]^T \widehat{\Sigma}_i^{-1} \left[\left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^0\right) - \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i^1\right)\right].$$

This statistic is more complicated than the simple form $F_i = [\mathrm{SS}_i^0 - \mathrm{SS}_i^1]/\mathrm{SS}_i^1$ given earlier, although it is essentially the same formula in matrix format. In fact, it can be shown that under zero covariances (i.e., where Σ_i is a diagonal matrix), the formulas exactly coincide.

6 Simulating null statistics

The null distribution of the statistics is calculated through a method called the bootstrap (Efron & Tibshirani 1993). 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. Using these null data, statistics are formed exactly as before that simulate the case where there is no differential expression. The null data are generated by re-sampling the residuals obtained under the alternative model fit and adding them back to the null model fit. The alternative model residuals are used because they come from the unconstrained model, and therefore the observed residuals approximate the residual distribution regardless of whether the null or alternative hypothesis is true. The residuals are added back to the null model because this fit comes from the class of models that represents a true null hypothesis. The algorithms appropriate for each type of model fitting/sampling scheme are detailed below. They are implementations of the method that is presented in Efron & Tibshirani (1993) for bootstrap hypothesis testing.

Independent sampling. The residuals are first calculated from the alternative model fit by $\widehat{\epsilon}_{ij} = y_{ij} - \widehat{y}^1_{ij}$. Then $y^*_{ij} = \widehat{y}^0_{ij} + \epsilon^*_i$ are formed, where ϵ^*_i is randomly sampled with replacement from among the $\widehat{\epsilon}_{ij}$ for fixed i. Note that the randomly sampled residuals are added back to the fitted values under the *null* model. This occurs across all genes over B iterations to produce null statistics $F^{0b}_1, F^{0b}_2, \ldots, F^{0b}_M$, for $b = 1, 2, \ldots, B$.

Longitudinal sampling & EM least squares. Under longitudinal sampling with intercepts included, two sets of residuals are calculated: the within-individual residuals $\widehat{\epsilon}_{ijk} = y_{ijk} - \widehat{y}^1_{ijk}$ and the between-individual residuals $\widehat{\gamma}^1_{ij}$. (The $\widehat{\gamma}^1_{ij}$ are those obtained as part of the EM algorithm presented in Section 3.) Then $y^*_{ijk} = (\widehat{y}^0_{ijk} - \widehat{\gamma}^0_{ij}) + \epsilon^*_i + \gamma^*_i$ are formed where ϵ^*_i is randomly sampled with replacement from among the $\widehat{\epsilon}_{ij}$, and γ^*_i is randomly sampled with replacement from among the $\widehat{\gamma}^1_{ij}$ for fixed i. For each fixed j, a newly sampled ϵ^*_i is added to each $k = 1, 2, \ldots, T_j$, however the same γ^*_i is used for all of these (i.e., only one γ^*_i is obtained per individual). Again, this is performed across all genes over B iterations to produce null statistics $F^{0b}_1, F^{0b}_2, \ldots, F^{0b}_M$, for $b = 1, 2, \ldots, B$.

Longitudinal sampling & individual centered least squares. Under longitudinal sampling, where the intercepts are not included in the analysis, the residuals are calculated by $\widehat{\epsilon}^c_{ijk} = y_{ijk} - \widehat{y}^1_{ijk}$, which have a known correlation structure (beyond that of the negligible correlation induced by the fact that residuals add to zero). The correlation structure exists because the model was fit to individual centered data, which are dependent. A simple approach to dealing with the correlation structure is to bootstrap from among the entire set of residuals corresponding to an individual. That is, $\mathbf{y}^*_{ij} = \widehat{\mathbf{y}}^0_{ij} + \boldsymbol{\epsilon}^*_i$ are formed, where $\boldsymbol{\epsilon}^*_i$ is a vector of length T_j randomly sampled from among

the $\widehat{\boldsymbol{\epsilon}}_{ij}^{c}$ for fixed i. However, this requires a balanced experiment which means that each individual has been sampled at the exact same time points. In the endotoxin study, one of the controls is missing two time points, so this simple approach is not possible. Also, in many cases the number of observed individuals N will be small, and bootstrapping at the individual level may not be accurate enough.

We implemented an approach to deal with the endotoxin study and any other study where it is not feasible to apply the bootstrap at the individual level. The basic idea is that the residuals are transformed to be uncorrelated; these uncorrelated residuals are then bootstrapped and transformed back to a set of residuals having the original correlation structure. Specifically, let Γ_j be the $(T_j-1)\times(T_j-1)$ matrix such that the diagonal elements are equal to $1-1/T_j$ and the off diagonal elements equal to $-1/T_j$. It can easily be shown that Γ_j is the correlation matrix of the $\widehat{\epsilon}_{ijk}^{c}$ for $k=1,\ldots,T_j-1$. (We do not include the last residual because it is equal to the sum of the other residuals.) Then calculate

$$\begin{pmatrix} \widehat{\epsilon}_{ij1} \\ \widehat{\epsilon}_{ij2} \\ \vdots \\ \widehat{\epsilon}_{ij,T_j-1} \end{pmatrix} = \Gamma_j^{-1/2} \begin{pmatrix} \widehat{\epsilon}_{ij1}^{c} \\ \widehat{\epsilon}_{ij2}^{c} \\ \vdots \\ \widehat{\epsilon}_{ij,T_j-1}^{c} \end{pmatrix}$$

and $\widehat{\epsilon}_{ijT_j} = -\sum_{k=1}^{T_j-1}\widehat{\epsilon}_{ijk}$. For each fixed gene i, randomly sample with replacement from among the $\widehat{\epsilon}_{ijk}$ over $j=1,\ldots,N$ and $k=1,\ldots,T_j$ to obtain ϵ_{ijk}^{*}. These ϵ_{ijk}^{*} are transformed to have the original correlation structure of the centered residuals:

$$\begin{pmatrix} \epsilon_{ij1}^{c*} \\ \epsilon_{ij2}^{c*} \\ \vdots \\ \epsilon_{ij,T_j-1}^{c*} \end{pmatrix} = \Gamma_j^{1/2} \begin{pmatrix} \epsilon_{ij1}^{*} \\ \epsilon_{ij2}^{*} \\ \vdots \\ \epsilon_{ij,T_j-1}^{*} \end{pmatrix}$$

and $\epsilon_{ijT_j}^{c*} = -\sum_{k=1}^{T_j-1}\epsilon_{ijk}^{c*}$. Finally, form $y_{ijk}^{c*} = \widehat{y}_{ijk}^{0} + \epsilon_{ijk}^{*}$ as a null version of the data. This process is repeated over B iterations to produce null statistics $F_1^{0b}, F_2^{0b}, \ldots, F_M^{0b}$, for $b=1,2,\ldots,B$.

Even though bootstrap methodology has been studied thoroughly and has been shown to produce valid results under the right conditions (Efron & Tibshirani 1993), we investigated the operating characteristics of our procedures via simulations. The basic approach was to simulate data under different scenarios by varying the numbers of genes, individuals and arrays; the distributions of the random deviations; the variances of the random deviations; the complexity of the curves; and the presence of outliers. The method was performed exactly as we would in practice, including automatically choosing the dimension of the basis p from the data (see Section 4). For each setting, 500 data sets were generated and the p-values for the null genes were calculated (see our p-value

formula in the next section). Then each set of p-values was tested against the uniform distribution via the Kolmogorov-Smirnov test, where the null hypothesis was that the observed p-values come from a distribution that is stochastically greater than or equal to the uniform.

This produced 500 Kolmogorov-Smirnov p-values that were then similarly tested against the uniform distribution. If each simulated data set produced valid null statistics, then we would expect these Kolmogorov-Smirnov p-values to also be stochastically greater than or equal to the uniform distribution because the null hypothesis in each case would be true (Lehmann 1986). Therefore, if the last test produced a non-significant p-value, then the conclusion was that the null statistics resulting from our procedure are valid in that setting. Using this evaluation, the bootstrap procedure appeared to be robust under a variety of distributions, variances, and curve complexities. No setting failed except for those cases where we included substantial outliers. The bootstrap (and permutation methods) are often invalid when the re-sampled residuals are not independent and identically distributed. The outliers affected the accuracy of the bootstrap and they also biased the curve fitting. Therefore, when applying this procedure, one must be careful to screen the expression values for outliers. Most commonly used pre-processing software does this automatically.

7 Estimating q-values

A common measure of significance, the p-value, can be formed for each gene by calculating the probability that a null gene has a statistic as or more extreme than the observed statistic. Numerically this can be accomplished by measuring the frequency by which the bootstrap null statistics exceed each observed statistic, implying that the p-value for gene i can be calculated by

$$p_i = \sum_{b=1}^{B} \frac{\#\{j : F_j^{0b} \geq F_i, j = 1, \ldots, M\}}{M \cdot B}. \tag{5}$$

A more conservative approach to calculating p-values is to only consider the null statistics generated from that gene:

$$p_i = \frac{\#\{b : F_i^{0b} \geq F_i, b = 1, \ldots, B\}}{B}.$$

However, we use the former definition given in equation (5). This formula is justified by weaker conditions than assuming approximate Normality of the random deviations. Results indicating that each gene does not even have to follow the same null distribution in order for these p-values to be valid for use in q-value estimation can be found in Storey et al. (2004). Moreover, one can always test whether the conditions of Storey et al. (2004) are met or not. Basically one has to verify that the distribution of the pooled p-values does not depend on the magnitudes of the observed statistics, which can be straightforwardly accomplished by a Kolmogorov-Smirnov test.

In the main text, we motivated the use of q-values as a measure of significance rather than p-values. Background on estimating q-values and their use in genomics can be found in Storey (2002) and Storey & Tibshirani (2003). Note that π_0, which is the proportion of genes that are not differentially expressed, is estimated as a part of q-value estimation. This quantity is useful in itself since $1 - \pi_0$ gives the proportion of differentially expressed genes, even though all of these cannot usually be identified with certainty. Following is the general algorithm for estimating q-values from their corresponding p-values (Storey & Tibshirani 2003).

1. Let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$ be the ordered p-values. This also denotes the ordering of the genes in terms of their evidence for differential expression.

2. For a range of λ, say $\mathcal{R} = \{0, 0.01, 0.02, \ldots, 0.95\}$, calculate

$$\widehat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda; j = 1, \ldots, M\}}{M(1 - \lambda)}.$$

   Let $\hat{f}$ be the natural cubic spline of $\widehat{\pi}_0(\lambda)$ on λ. (We use a smoother with 3 degrees of freedom.) Set the estimate of π_0 to be

$$\widehat{\pi}_0 = \hat{f}(\max \mathcal{R}).$$

3. Calculate

$$\widehat{\mathrm{q}}(p_{(M)}) = \min_{t \geq p_{(M)}} \frac{\widehat{\pi}_0 M \cdot t}{\#\{p_j \leq t; j = 1, \ldots, M\}} = \widehat{\pi}_0 \cdot p_{(M)}.$$

4. For $i = M - 1, M - 2, \ldots, 1$, calculate

$$\widehat{\mathrm{q}}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\widehat{\pi}_0 M \cdot t}{\#\{p_j \leq t; j = 1, \ldots, M\}} = \min \left(\frac{\widehat{\pi}_0 M \cdot p_{(i)}}{i}, \widehat{\mathrm{q}}(p_{(i+1)}) \right).$$

5. The estimated q-value for the i^{th} most significant gene is $\widehat{\mathrm{q}}(p_{(i)})$.

The following estimate of the false discovery rate when calling all p-values $\leq t$ significant is implicit in the algorithm:

$$\widehat{\mathrm{FDR}}(t) = \frac{\widehat{\pi}_0 M \cdot t}{\#\{p_i \leq t; i = 1, \ldots, M\}}.$$

Some articles have distinguished re-sampling based false discovery rates estimation from p-value based estimation. In particular, Tusher et al. (2001) do not point out that their false discovery rate method uses the estimate implicit in the Benjamini & Hochberg (1995) procedure. Also, Dudoit et al. (2003) incorrectly criticize the approach of Tusher et al. (2001), failing to recognize that it employs the Benjamini & Hochberg (1995) estimate by way of gene non-specific p-values. It can

easily be shown that re-sampling based false discovery rate estimation is equivalent to p-value based estimation, where the re-sampling method simply determines how p-values are calculated.

For a fixed significance cut-off c, the re-sampling based false discovery rate estimate is

$$\widehat{\text{FDR}}(c) = \frac{\widehat{\pi}_0 \sum_{b=1}^{B} \#\{F_i^{0b} \geq c; i = 1, \ldots, M\}/B}{\#\{F_i \geq c; i = 1, \ldots, M\}},$$

where $\widehat{\pi}_0$ is derived from the

$$\widehat{\pi}_0(c') = \frac{\#\{F_i < c'; i = 1, \ldots, M\}}{\sum_{b=1}^{B} \#\{F_i^{0b} < c'; i = 1, \ldots, M\}/B}$$

over some range of c' exactly as in the above algorithm. As was stated in Storey & Tibshirani (2003), the estimate in Storey (2002) is easily shown to be equivalent to the above formula, where the only difference is that it is written in terms of thresholding the statistics directly. The key observation is that one can equivalently define the type I error rate of a given cut-off by $\sum_{b=1}^{B} \#\{F_i^{0b} \geq c\}/(M \cdot B)$ rather than the p-value threshold t. In fact, if we define

$$c(t) \equiv \min\{F_i : p_i \leq t\}$$

then it can be shown that

$$\frac{\widehat{\pi}_0 \sum_{b=1}^{B} \#\{F_i^{0b} \geq c(t); i = 1, \ldots, M\}/B}{\#\{F_i \geq c(t); i = 1, \ldots, M\}} = \frac{\widehat{\pi}_0 M \cdot t}{\#\{p_i \leq t; i = 1, \ldots, M\}}$$

making the two false discovery rate estimates equal. Therefore, q-values derived from either method are equal as long as the p-values are calculated from the gene non-specific empirical distribution of the simulated null statistics.

8 Distinguishing independent and longitudinal sampling

The fundamental question to ask when distinguishing between independent and longitudinal sampling is: does the deviation from the population average curve at time t depend on the observed deviation at previous time points? If the answer is yes, then the longitudinal sampling methodology should be applied. In some cases, the distinction is directly apparent. For example, if gene expression is measured on the same individual every hour for 24 hours, it is clear that the random deviation at hour 22 will depend on the random deviation at hour 21, (see Figure 1). However, if gene expression is measured on a different individual from a sample population each hour over the same time period, it is not likely that the random deviations at successive hours will depend on the random deviations at previous time points.

Unfortunately, the decision is not always as clear as in these two cases; take for example, gene expression measurements from different individuals over time who are exposed to ever increasing

temperature. Even though the measurements may not be correlated due to individual specific parameters, the measurements on individual j at 23° may be more closely related to individual $(j+1)$ at 24° then to individual one at 10°. For more subtle cases such as this, where serial correlation is postulated but is not obvious, it is better to err toward overestimation and apply the longitudinal sampling methodology. The loss of power and increase in computation time is more attractive than the erroneous estimates that result from applying independent sampling methodology to longitudinal samples.

9 Gene filtering method used on the kidney study

The individuals obtained 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) latent variables. An initial analysis showed that genes affected by latent variables were disproportionately called significant. However, information on tissue type (cortex or medulla) and sex was available for each individual. Therefore, genes whose expression is well explained by these variables, irrespective of any age dependent behavior, can unequivocally be evaluated for temporal differential expression. We applied a filtering technique that sought to identify these genes, while at the same time guarding against anti-conservatively biasing the subsequent significance analysis. The basic approach is to fit a high dimensional curve to each gene including the sex and tissue type covariates in the parametrization of the curve (see Materials and Methods in the main article). This fit removes any effects based on time, sex or tissue type. If there are any outlying residuals, then this indicates that there is an unobserved covariate affecting this gene. For example, HLA-F was removed by the filtering method. The residuals obtained from fitting the above curve to this gene are plotted in Figure 3. Several large outliers indicate the presence of an unobserved variable affecting that particular gene. It is easily verified that the outliers are not due to a systematic array-wide trend. The precise procedure to filter the genes involves the following steps:

1. On gene i, fit a curve to all arrays except one using a $p = 10$ dimensional basis, taking into account tissue type and sex.

2. Apply this fit to the removed array and obtain its residual.

3. Repeat Steps 1 and 2 for all arrays $j = 1, 2, \ldots, 133$.

4. If any residual is greater than 3 times the interquartile range from median residual value, eliminate gene i.

5. Apply Steps 1–4 on all genes $i = 1, 2, \ldots, 44,924$.

In applying this algorithm 10,863 genes were removed, reducing the total number of genes included in the analysis to 34,061. There are two potential sources of error induced by the filtering method. The first is if it does not effectively remove genes affected by latent variables. The second is if the method preferentially includes null genes that have some curvature by chance. In either case, the significance may become artificially inflated. However, we show that these can be controlled by using a well chosen p in Step 1 and a well chosen residual cut-off in Step 4.

To determine if the filter created a bias toward more significant p-values, and to determine reasonable values to use in Steps 1 and 4, we performed a simulation study. The simulated data consisted of 2000 null (flat) time courses with normally distributed random deviations reflecting the variances observed in the kidney aging data. For 250 of these genes a signal was added to simulate a latent variable. In addition, random affects for both age and sex were included. The filter was applied to the simulated data for several combinations of p and multiples of the interquartile range. Then our method for identifying differential expression was applied with 500 bootstrap iterations, $p = 5$, and the natural cubic spline basis.

For each combination, the p-values corresponding to "valid" genes (i.e., genes not affected by any latent variables) were compared to the uniform distribution by a one sided Kolmogorov-Smirnov test. The one sided Kolmogorov-Smirnov test was used to determine if the observed distribution of null p-values were stochastically greater than the uniform distribution or not. This tests whether the filter inflates significance or not. Figure 4 shows the p-values from the Kolmogorov-Smirnov test for a range of filters. The horizontal line is the Kolmogorov-Smirnov p-value applied to all valid genes. Note that for many of the filter settings, the Kolmogorov-Smirnov p-values fall above the horizontal line, indicating that the filtering process has induced a conservative bias in the estimated p-values. Further, irrespective of which filter we used there is little evidence that anti-conservative bias is induced.

The left panel of Figure 5 shows the total number of "invalid" genes (i.e., genes affected by at least one latent variables) that remained after the same filter settings were applied. On the right is the number of genes out of the 250 simulated invalid genes that were removed by each filter setting. Note that there is a steep drop off in the number of genes retained when a filter with greater than $p = 10$ was used. It is also clear that the filter with cut-point 2 times the interquartile range eliminates too many valid genes. In an effort to balance this against the need for conservative p-values we selected the filter using the cut-point of 3 times the interquartile range and a $p = 10$ dimensional basis.

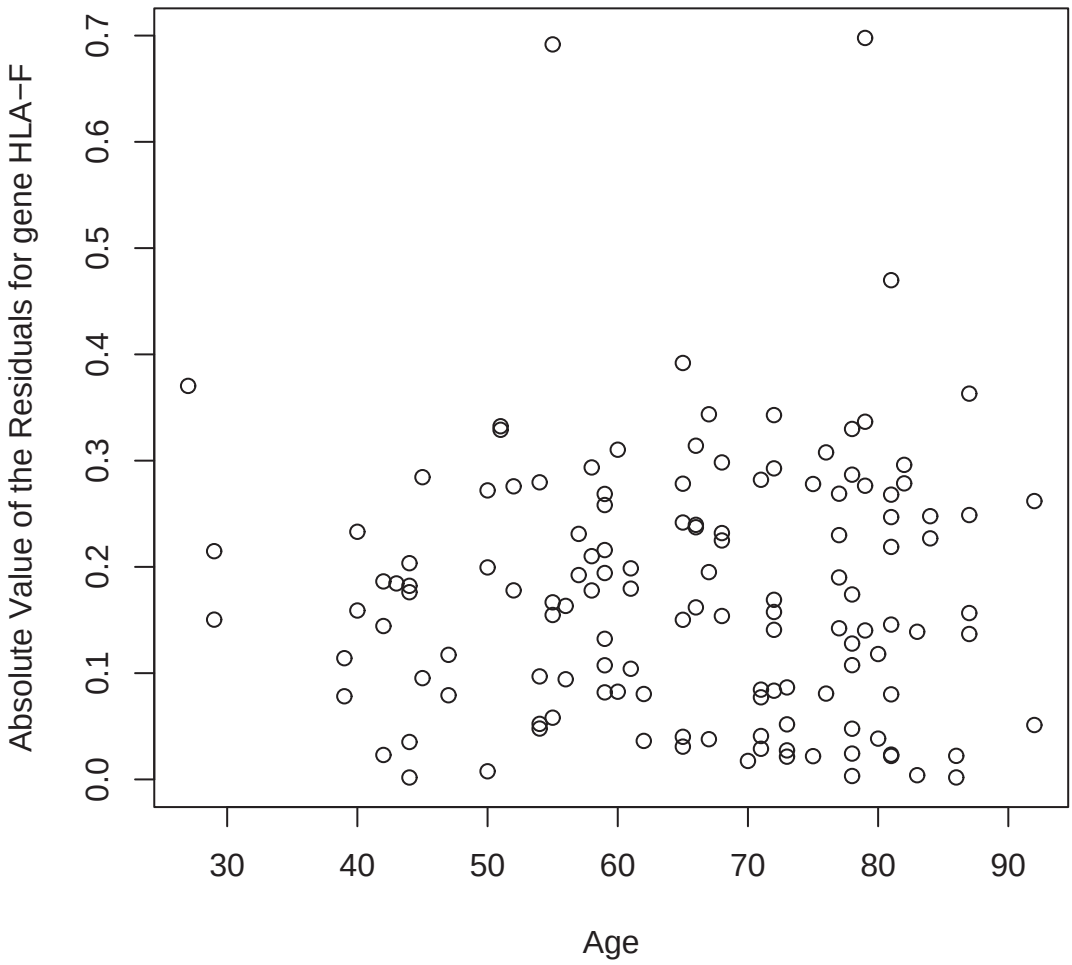

Figure 3: A plot of the residuals obtained from fitting a curve to gene HLA-F, demonstrating the presence of gene-specific outliers and a possible underlying latent variable. This gene was one of 10,867 filtered from the kidney aging study.

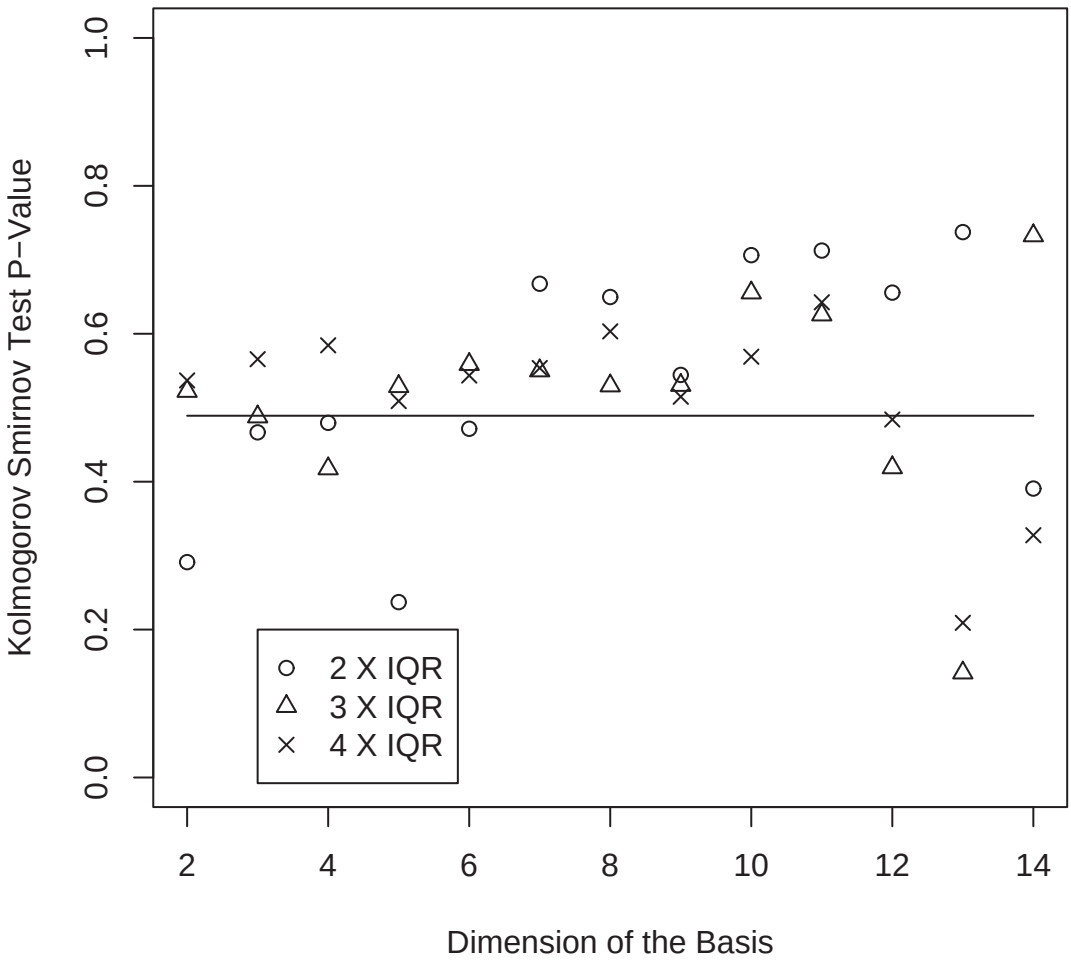

Figure 4: A plot of significance values from the Kolmogorov-Smirnov test that the filtering method does not cause anti-conservative bias over various filter settings. The horizontal line represents the significance level of the performed on the genes before filtering.

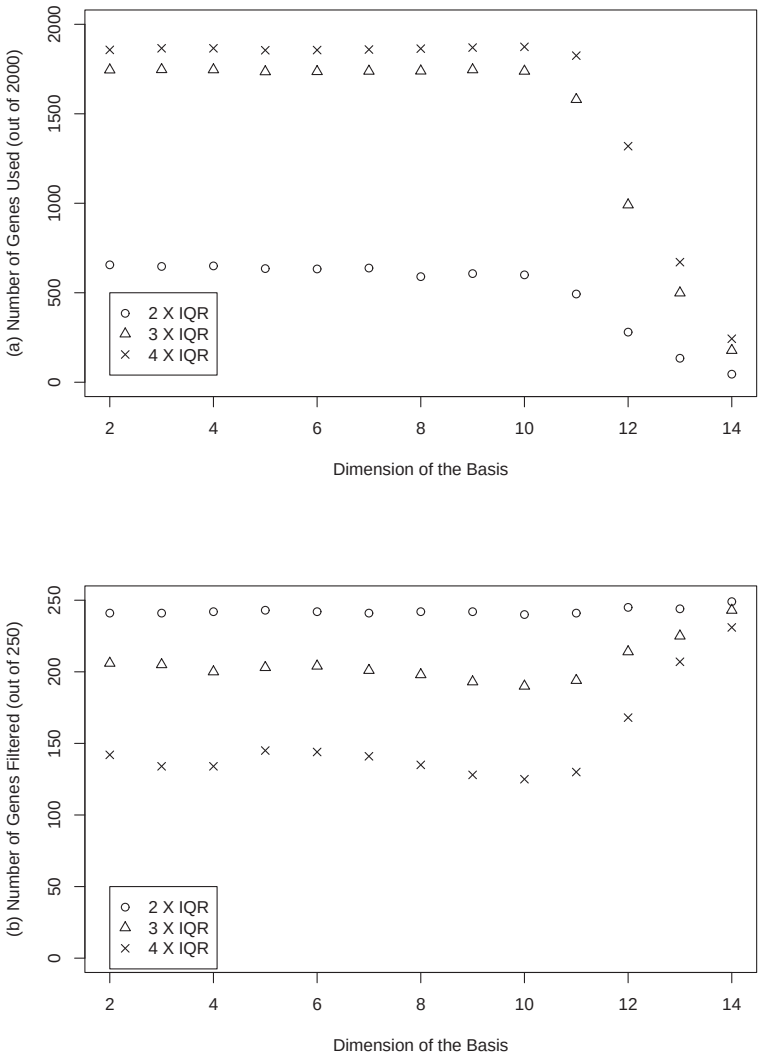

Figure 5: (a) A plot of the total number of the 2000 simulated genes that were not filtered for different filter settings. Note the steep decline in retention for $p > 10$. (b) A plot of the total number of the 250 genes with simulated outliers that were removed by each filter. Since the outliers were random, the total number of genes that can be classified as having outliers may be less than 250.

10 Additional results from the endotoxin study

Table 2 lists the 50 most significant genes that show down-regulation over the time course. Table 3 lists the 50 most significant genes that show up-regulation over the time course. For completeness, we further considered the question of whether the controls show any non-flat expression (i.e., no gene shows within class temporal differential expression). If this were the case, one could test the endotoxin group directly for within class temporal differential expression rather than comparing them to the control group. Our method was applied to each group individually, testing for within class temporal differential expression. The significance analysis indicates that at least 4% of the genes are temporally differentially expressed in the control group ($\widehat{\pi}_0 = 0.96$). No particular gene shows a strong enough signal to be called differentially expressed with statistical confidence. The top eigen-gene corresponding to the control group (shown in Figure 3a of the main text) displays the nonlinear trend that is easily seen in a number of genes in the control group. This trend may be due to systematic data collection effects, physical stress induced by the study, or natural circadian changes. When testing the treated group for within class temporal differential expression, the results become much more significant compared to the original analysis. For example, 22,325 genes are found to be significant at a q-value cut-off of 5%. A careful examination of the data indicates that this is unreasonable. Therefore, the controls play an important role in detecting temporal differential expression in the endotoxin study. In an uncontrolled clinical setting, it is likely important to compare treated individuals to controls when testing for genes that show a response over time.

11 Additional results from the kidney aging study

Table 4 lists the 50 most significant genes in the kidney aging study. In addition to the results discussed in the main text, we investigated the importance of the dimension of the basis p in detecting temporal differential expression. The motivation is that the data are noisy enough that it is difficult to detect nonlinear behavior by eye. Figure 1b of the main text displays this phenomenon; it can be seen that the alternative model fit is highly nonlinear, although the nonlinearity of the data is not obvious. As mentioned in Section 4, our automatic choice of $p = 5$ was superior to other reasonable choices. Therefore, capturing the full complexity of the data is important in gaining power. In particular, we performed a Kolmogorov-Smirnov test to compare the p-values from $p = 5$ to $p = 1$, which fits a line rather than a curve. The null hypothesis is that the p-values from $p = 1$ are stochastically less than or equal to those from $p = 5$. The test was significant at level 5.53×10^{-52}, meaning that the p-values from $p = 1$ are substantially stochastically greater, and therefore less powerful, than those from $p = 5$. When using $p = 1$ we find 161, 504, and 1004

significant genes at q-value cut-offs of 1%, 5% and 10%, respectively. In contrast, for $p = 5$ we find 189, 786 and 1821 significant genes, respectively, at these same cut-offs.

References

Alter, O., Brown, P. O. & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences* **97**: 10101–10106.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **85**: 289–300.

Dudoit, S., Shaffer, J. P. & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science* **18**: 71–103.

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

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

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edn, Springer-Verlag.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edn, Wiley-Interscience.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1980). *Multivariate Analysis*, Academic Press.

Rice, J. & Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves, *Biometrics* **57**: 253–259.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B* **64**: 479–498.

Storey, J. D., Taylor, J. E. & Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach, *Journal of the Royal Statistical Society, Series B* **66**: 187–205.

Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* **100**: 9440–9445.

Tusher, V., Tibshirani, R. & Chu, C. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proceedings of the National Academy of Sciences* **98**: 5116–5121.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer-Verlag.

Weisberg, S. (1985). *Applied Linear Regression*, 2nd edn, John Wiley & Sons.

Table 2: The 50 most significant down-regulated genes in the endotoxin study.

rank	gene name	gene annotation	gene function
1	RPL27A	ribosomal protein L27a	protein synthesis
2	RPS7	ribosomal protein S7	protein synthesis
3	RPL5	ribosomal protein L5	protein synthesis
4	RPL13A	ribosomal protein L13a	protein synthesis
5	SLC25A6	solute carrier family 25, member 6	transporter activity
6	RPS4X	ribosomal protein S4, X-linked	protein synthesis
7	PECAM1	platelet/endothelial cell adhesion molecule	cell adhesion
8	RPL30	ribosomal protein L30	protein synthesis
9	EFE2	eukaryotic translation elongation factor 2	translation
10	RPL34	ribosomal protein L34	protein synthesis
11	RPS15A	ribosomal protein S15a	protein synthesis
12	RPL7A	ribosomal protein L7a	protein synthesis
13	GLTSCR2	glioma tumor suppressor candidate region gene 2	tumor suppressor candidate gene
14	RPL9	ribosomal protein L9	protein synthesis
15	HLA-DPB1	major histocompatibility complex, class II, DP beta 1	peptide present in immune system
16	RPL35	ribosomal protein L35	protein synthesis
17	LAMR1	laminin receptor 1 (ribosomal protein SA)	protein synthesis
18	RPL3	ribosomal protein L3	protein synthesis
19	RPS11	ribosomal protein S11	protein synthesis
20	RPL14	ribosomal protein L14	protein synthesis
21	RPL32	ribosomal protein L32	protein synthesis
22	RPL6	ribosomal protein L6	protein synthesis
23	RPS6	ribosomal protein S6	protein synthesis
24	RPL12	ribosomal protein L12	protein synthesis
25	RPS28	ribosomal protein S28	protein synthesis
26	RPL37	ribosomal protein L37	protein synthesis
27	RPS20	ribosomal protein S20	protein synthesis
28	ASAHL	N-acylsphingosine amidohydrolase	hydrolase activity
29	RPS18	ribosomal protein S18	protein synthesis
30	RPS6	ribosomal protein S6	protein synthesis
31	RPS19	ribosomal protein S19	protein synthesis
32	RPS10	ribosomal protein S10	protein synthesis
33	RPS3A	ribosomal protein S3A	protein synthesis
34	RPL21	ribosomal protein L21	protein synthesis
35	RPL4	ribosomal protein L4	protein synthesis
36	EEF1G	eukaryotic translation elongation factor 1 gamma	translation
37	RPL27	ribosomal protein L27	protein synthesis
38	RPL18	ribosomal protein L18	protein synthesis
39	RPL36	ribosomal protein L36	protein synthesis
40	RPL5	ribosomal protein L5	protein synthesis
41	CDKN1C	cyclin-dependent kinase inhibitor 1C	inhibitor of G1 cyclin/Cdk complexes
42	RPL10A	ribosomal protein L10a	protein synthesis
43	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	peptide present
44	RPL19	ribosomal protein L19	protein synthesis
45	ITGA4	integrin, alpha 4	cell adhesion
46	PCNA	proliferating cell nuclear antigen (PCNA)	increase the processivity of leading strand synthesis during DNA replication
47	RPL12	ribosomal protein L17	protein synthesis
48	RPS12	ribosomal protein S12	protein synthesis
49	AMPD2	adenosine monophosphate deaminase 2	hydrolase activity
50	RPS2	ribosomal protein S2	protein synthesis

Table 3: The 50 most significant up-regulated genes in the endotoxin study.

rank	gene name	gene annotation	gene function
1	ORM1	orosomucoid 1	key acute phase plasma protein,immunosuppression
2	TNFAIP3	tumor necrosis factor, alpha-induced protein 3	inhibit NF-kappa B activation as well as TNF-mediated apoptosis
3	GYG	glycogenin	transferase activity
4	CD59	CD59 antigen p18-20	activates human NK cell-mediated cytotoxicity
5	IER3	immediate early response 3	protection of cells from Fas-or tumor necrosis factor type alpha-induced apoptosis
6	ATP9A	ATPase, Class II, type 9A	energy production and conversion, by automatic sequence prediction
7	IL1RN	interleukin 1 receptor antagonist (IL1RN)	modulate a variety of interleukin 1 related immune and inflammatory responses
8	TNFAIP6	tumor necrosis factor, alpha-induced protein 6	signal transduction, inflammatory response
9	IL1B	interleukin 1	proinflammatory cytokine
10	CD59	CD59 antigen p18-20	activates human NK cell-mediated cytotoxicity
11	FCAR	Fc fragment of IgA, receptor	interacts with IgA-opsonized targets and triggers immunologic defense processes
12	IL1RAP	interleukin 1 receptor accessory protein	Interleukin 1 induced acute phase and proinflammatory response
13	NFKBIA	NF kappa light polypeptide gene enhancer in B-cells inhibitor	Signal transduction in NF-kappaB pathway
14	PTX3	pentaxin-related gene, rapidly induced by IL-1 beta	inflammatory response
15	ALOX5AP	rachidonate 5-lipoxygenase-activating protein	leukotriene synthesis
16	HP	haptoglobin	plasma glycoprotein
17	CRISP3	cysteine-rich secretory protein 3	undetermined
18	MAP4K4	mitogen-activated protein kinase kinase kinase kinase 4	signal transduction
19	HSP	heat shock protein 70kDa protein 1A	help protein folding and maintain stability
20	DAF	decay accelerating factor for complement	regulators of complement activation
21	PRV1	polycythemia rubra vera 1	cell surface glycoprotein
22	TLR5	toll-like receptor 8	mediate the production of cytokines
23	TPK1	thiamin pyrophosphokinase 1	regulation of thiamine metabolism
24	GAS7	growth arrest-specific 7	cell development
25	G0S2	putative lymphocyte G0/G1 switch gene	lymphocyte cell cycle
26	PGLYRP1	peptidoglycan recognition protein 1	bacterial binding, peptidoglycan receptor activity
27	CKAP4	cytoskeleton-associated protein 4	membrane protein
28	S100P	S100 calcium binding protein P	cell cycle progression and differentiation
29	DYSF	dysferlin, limb girdle muscular dystrophy 2B	membrane regeneration and repair
30	CASP5	caspase 5, apoptosis-related cysteine protease	central role in the execution-phase of cell apoptosis
31	PADI4	peptidyl arginine deiminase, type IV	granulocyte and macrophage development leading to inflammation response
32	BCL2A1	BCL2-related protein A1	target of NF-kappa B in response to inflammatory mediators, anti- and pro-apoptotic regulators
33	SRPK1	SFRS protein kinase 1	regulation of both constitutive and alternative splicing
34	IL18R1	interleukin 18 receptor 1	proinflammatory cytokine receptor
35	RAB27A	member RAS oncogene family	protein transport and small GTPase mediated signal transduction
36	C20orf3	chromosome 20 open reading frame 3	strictosidine synthase activity
37	CCL4	chemokine	signal transduction to mediate inflammatory response
38	ANXA3	annexin A3	signal transduction
39	WDFY3	WD repeat and FYVE domain containing 3	undetermined
40	ALPL	alkaline phosphatase, liver/bone/kidney	metabolism, alkaline phosphatase activity
41	SLC2A3	solute carrier family 2, member 3	glucose transporter
42	FOSL2	FOS-like antigen 2	regulators of cell proliferation
43	OSM	oncostatin M	regulates cytokine production
44	MMP9	matrix metalloproteinase 9	breakdown of extracellular matrix, mobilization of hematopoietic progenitor cells
45	STK3	serine/threonine kinase 3	protein kinase
46	MAPK14	mitogen-activated protein kinase 14	signal transduction for proliferation, differentiation, transcription regulation et al.
47	LAMP3	lysosomal-associated membrane protein 3	oncogenesis, cell proliferation
48	EHD1	EH-domain containing 1	protein-protein interactions and intracellular sorting
49	CAMP	cathelicidin antimicrobial peptide	antimicrobial protein found in leukocytes
50	ACSL1	acyl-CoA synthetase long-chain family member 1	lipid biosynthesis and fatty acid degradation

Table 4: The 50 most significant genes in kidney aging study.

rank	gene name	gene annotation	gene function	increase/decrease
1	LAPTM5	Lysosomal-associated multispanning membrane protein-5	cellular differentiation and apoptosis	up
2	NNMT	nicotinamide N-methyltransferase	catalyzes drug and xenobiotic metabolism	up
3	TYROBP	TYRO protein tyrosine kinase binding protein	signal transduction	up
4	HLA-F	major histocompatibility complex, class I, F	present peptide in immune system	up
5	CRABP1	cellular retinoic acid binding protein 1	retinoic acid-mediated differentiation and proliferation	down
6	HLA-B	major histocompatibility complex, class I	present peptide in immune system	up
7	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	present peptide in immune system	up
8	C1R	complement component 1, r subcomponent	effector of antibody-mediated immunity	up
9	C1QA	complement component 1, q subcomponent, alpha polypeptide	effector of antibody-mediated immunity	up
10	HLA-F	major histocompatibility complex, class I, F	present peptide in immune system	up
11	PIGR	polymeric immunoglobulin receptor	transport of immunoglobulin into external secretion	up
12	AMPD3	adenosine monophosphate deaminase (isoform E)	nucleotide metabolism	down
13	GPNMB	glycoprotein (transmembrane) nmb	growth delay and reduction of metastatic potential	up
14	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	present peptide in immune system	up
15	DKFZp761P0423	hypothetical protein DKFZp761P0423	undetermined	up
16	C1QR1	complement component 1, q subcomponent, receptor 1	phagocytosis	up
17	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	present peptide in immune system	up
18	MS4A7	membrane-spanning 4-domains, subfamily A, member 7	signal transduction	up
19	RGS1	regulator of G-protein signalling 1	signal transduction	up
20	KIAA1268	KIAA1268 protein	undetermined	up
21	HLA-DMA	major histocompatibility complex, class II, DM alpha	present peptide in immune system	up
22	TNFAIP3	tumor necrosis factor, alpha-induced protein 3	inhibit NF-kappa B activation, TNF-mediated apoptosis	up
23	MMP7	matrix metalloproteinase 7	breakdown of extracellular matrix	up
24	LYZ	lysozyme (renal amyloidosis)	anti-microbial agents	up
25	LST1	leukocyte specific transcript 1	immune response	up
26	TINF2	TERF1 (TRF1)-interacting nuclear factor 2	regulator of telomere length in human	up
27	LOC374969	hypothetical protein LOC374969	undetermined	up
28	LPL	lipoprotein lipase	lipoprotein metabolism	down
29	PTPRC	protein tyrosine phosphatase, receptor type, C	signal transduction	up
30	HLA-DRB1	major histocompatibility complex, class II, DR beta 1	present peptide in immune system	up
31	CCL2	chemokine (C-C motif) ligand 2	immunoregulatory and inflammatory processes	up
32	TMSB10	thymosin, beta 10	actin modulating activity	up
33	ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1	transcription regulation	up
34	IFITM1	interferon induced transmembrane protein 1	immune response	up
35	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	present peptide in immune system	up
36	LOC388025	similar to Ig alpha-2 chain C region	undetermined	up
37	RGS1	regulator of G-protein signalling 1	signal transduction	up
38	LOC399810	similar to ribosomal protein L13a	undetermined	up
39	ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta	signal transduction	up
40	PRG1	proteoglycan 1, secretory granule	neutralizing hydrolytic enzymes	up
41	WFDC2	WAP four-disulfide core domain 2	protease inhibitor	up
42	DBT	dihydrolipoamide branched chain transacylase	metabolism	down
43	MSR1	macrophage scavenger receptor 1	macrophage-associated processes	up
44	KCTD10	potassium channel tetramerisation domain containing 10	potassium ion transport	up
45	MPEG1	macrophage expressed gene 1	undertermined	up
46	A2M	alpha-2-macroglobulin	protease inhibitor and cytokine transporter	up
47	BIRC3	baculoviral IAP repeat-containing 3	inhibits apoptosis	up
48	TNFAIP3	tumor necrosis factor, alpha-induced protein 3	limiting inflammation	up
49	GALNTL1	N-acetylgalactosaminyltransferase-like 1	undetermined	up
50	UGCG	UDP-glucose ceramide glucosyltransferase	glycosphingolipid biosynthesis	up