Identification of gene expression patterns using planned linear contrasts

Hao Li  
University of Kentucky, lhao@uky.edu

Constance L. Wood  
University of Kentucky, cwood@uky.edu

Yushu Liu  
University of Kentucky, yushu@ms.uky.edu

Thomas V. Getchell  
University of Kentucky, tgetche@uky.edu

Marilyn L. Getchell  
University of Kentucky, mgetch@uky.edu

See next page for additional authors

Follow this and additional works at: http://uknowledge.uky.edu/statistics_facpub

Part of the Statistics and Probability Commons

Repository Citation
Li, Hao; Wood, Constance L.; Liu, Yushu; Getchell, Thomas V.; Getchell, Marilyn L.; and Stromberg, Arnold J., "Identification of gene expression patterns using planned linear contrasts" (2006). Statistics Faculty Publications. 9.  
http://uknowledge.uky.edu/statistics_facpub/9

This Article is brought to you for free and open access by the Statistics at UKnowledge. It has been accepted for inclusion in Statistics Faculty Publications by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
Identification of gene expression patterns using planned linear contrasts

Notes/Citation Information
Published in BMC Bioinformatics, v. 7, 245.

© 2006 Li et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Digital Object Identifier (DOI)
http://dx.doi.org/10.1186/1471-2105-7-245
Identification of gene expression patterns using planned linear contrasts
Hao Li*1, Constance L Wood1, Yushu Liu1, Thomas V Getchell2,4, Marilyn L Getchell3,4 and Arnold J Stromberg1

Address: 1Department of Statistics, University of Kentucky, 817 Patterson Office Tower, Lexington, KY40536-0027, USA, 2Department of Physiology, College of Medicine, Lexington, KY40536-0298, USA, 3Department of Anatomy and Neurobiology, College of Medicine, Lexington, KY40536-0298, USA and 4309 Sanders-Brown Center on Aging, University of Kentucky Medical Center, Lexington, KY40536-0230, USA

Email: Hao Li* - lhao@uky.edu; Constance L Wood - cwood@uky.edu; Yushu Liu - yushu@ms.uky.edu; Thomas V Getchell - tgetche@uky.edu; Marilyn L. Getchell - mgetch@uky.edu; Arnold J Stromberg - astro@ms.uky.edu

* Corresponding author

Abstract

Background: In gene networks, the timing of significant changes in the expression level of each gene may be the most critical information in time course expression profiles. With the same timing of the initial change, genes which share similar patterns of expression for any number of sampling intervals from the beginning should be considered co-expressed at certain level(s) in the gene networks. In addition, multiple testing problems are complicated in experiments with multi-level treatments when thousands of genes are involved.

Results: To address these issues, we first performed an ANOVA F test to identify significantly regulated genes. The Benjamini and Hochberg (BH) procedure of controlling false discovery rate (FDR) at 5% was applied to the P values of the F test. We then categorized the genes with a significant F test into 4 classes based on the timing of their initial responses by sequentially testing a complete set of orthogonal contrasts, the reverse Helmert series. For genes within each class, specific sequences of contrasts were performed to characterize their general ‘fluctuation’ shapes of expression along the subsequent sampling time points. To be consistent with the BH procedure, each contrast was examined using a stepwise Studentized Maximum Modulus test to control the gene based maximum family-wise error rate (MFWER) at the level of \( \alpha_{new} \) determined by the BH procedure. We demonstrated our method on the analysis of microarray data from murine olfactory sensory epithelia at five different time points after target ablation.

Conclusion: In this manuscript, we used planned linear contrasts to analyze time-course microarray experiments. This analysis allowed us to characterize gene expression patterns based on the temporal order in the data, the timing of a gene’s initial response, and the general shapes of gene expression patterns along the subsequent sampling time points. Our method is particularly suitable for analysis of microarray experiments in which it is often difficult to take sufficiently frequent measurements and/or the sampling intervals are non-uniform.
Background
Recent advances in DNA microarray technologies have made it possible to investigate the transcriptional portion of gene networks in a variety of organisms. When microarray experiments are performed to monitor gene expression over time, researchers can address questions concerning the detection of the cellular processes underlying the observed regulatory effects, inference of regulatory networks and, ultimately, assignment of functions to the genes analyzed in the time courses.

There is a natural connection between gene function and gene expression. Based on our understanding of cellular processes, genes that are contained in a particular pathway, or respond to a common internal or external stimulus, should be co-regulated and consequently, should show similar patterns of expression. Therefore, identifying patterns of gene expression and grouping genes into expression classes may provide much greater insight into their biological functions. A large group of statistical methods, generally referred to as "cluster analysis", have been developed to identify genes that behave similarly across a range of experimental conditions, including time courses. These statistical algorithms can be divided into two classes, depending on whether they are based on 'similarity' measures or not. Methods based on 'similarity' measures rely on defining a distance (or 'dissimilarity') between gene expression vectors; Euclidean distance and/or the Pearson correlation coefficient are the two most commonly used distance measures. Examples of similarity measures-based methods are hierarchical clustering [1], k-means [2], self-organization maps (SOM) [3,4], and support vector machine (SVM) [5]. These methods do not consider the temporal structure of the data when used to analyze time-course experiments. In addition, some methods could confuse the clusters because the actual expression patterns of the genes themselves become less relevant as clusters grow in size [6].

The clustering methods in the second class are based on statistical models, without defining a 'similarity' measure. Using statistical models to represent clusters changes the question from how close two data points are to how likely a given data point is under the model. Such clustering methods are more commonly used to analyze time-course microarray experiments. Examples of such methods are based on cubic spline [7], ANOVA model [8], autoregressive curves [9], first-order kinetics [10], Hidden Markov Models [11,12], Bayesian model average [13], order-restricted inference methodology [14], and Gaussian Mixture Models [15-19]. Such approaches may be restricted either by the rigorous assumptions of the stochastic models [9,11,12], or by the small number of time points and non-uniform sampling intervals in gene expression data [7,9,10].

In gene networks, the level of expression of individual genes changes based on their functional position in the network. Therefore, the most critical information in time course expression profiles is the timing of the changes in expression level for each gene [10], and secondarily is the general shape of its expression pattern [20,21]. In addition, different genes will be activated or inactivated at each level of a gene network. Therefore it may not be reasonable to expect that the expression levels of those co-expressed genes will go up and down concordantly all the way through the entire sampling period. With the same timing of initial change, genes which share similar pattern of expression for any number of sampling intervals from the beginning should be considered co-expressed at certain level(s) in the gene network. However, statistical methods to analyze these patterns have not yet been reported.

Attention to the multiplicity problem in gene expression analysis has been increasing. Numerous methods are available for controlling the family-wise type I error rate (FWER). Since microarray experiments are frequently exploratory in nature and the sample sizes are usually small, Benjamini and Hochberg [22] suggested a potentially more powerful procedure, the false discovery rate (FDR), to control the expected proportion of errors among the identified differentially expressed genes. A number of studies for controlling FDR have followed [23-29]. In microarray experiments with multi-level treatments, the multiple testing problems are two dimensional. Not only are thousands of genes involved, but for each gene, either pre-selected contrasts or post-hoc comparisons may be needed to characterize its expression pattern. There are very few studies that have investigated how to deal with such multiple-testing problems in the microarray literature [30].

In this manuscript, we propose a different strategy based on planned linear contrasts (pre-selected contrasts) for the analysis of time-course microarray experiments. Specifically, our approach takes into consideration the temporal order in the data, including the timing of a gene's initial response and the general shapes of gene expression patterns along the subsequent sampling time points. Our methods are particularly suitable for analysis of microarray experiments in which it is often difficult to take sufficiently frequent measurements and/or the sampling intervals are non-uniform. We demonstrated our method on the analysis of microarray data from murine olfactory sensory epithelia at five different time points after target ablation.

Results
Olfactory sensory neurons (OSNs) detect odors in the ambient environment and transmit the sensory informa-
The death of OSNs can be induced experimentally by microsurgical removal of their axonal targets in the brain (olfactory bulbectomy, OBX). The temporal regulation of genes associated with the death of OSNs and other cellular processes as a result of OBX can be systematically investigated at 2 hr, 8 hr, 16 hr and 48 hr post-OBX. Based on the statistical methods described (see Methods), 1234 genes were considered to be significant by the procedure of controlling FDR at 5% for multiple testing across genes. The largest P-value considered to be significant was 0.009545 as determined by the FDR procedure. The temporal regulation of these 1234 genes fell into four distinct classes based on the first significant change in their temporal profile that occurred at either 2 hr (Class I), 8 hr (Class II), 16 hr (Class III), or 48 hr (Class IV) post-OBX. Among the 1234 genes (Figure 1), 212 were grouped into Class I in which the differential expression of these genes was detected as early as 2 hours post-OBX.

Table 1: Example of genes from different classes

| Class | Gene  | P_F     | P_t     | FC   | Gene Function |
|-------|-------|---------|---------|------|---------------|
| I     | Pdcd5 | 5.40E-03| 6.30E-04| 1.2  | apoptosis     |
|       | Cetn3 | 7.40E-05| 3.10E-04| 1.2  | Ca binding    |
|       | Kit   | 3.40E-03| 5.30E-04| 1.4  | growth factor |
| II    | Ccl2  | 3.20E-05| 3.70E-04| 4.1  | chemotaxis    |
|       | Csf3  | 6.30E-03| 5.80E-04| 3.2  | growth factor |
|       | Bub3  | 2.10E-04| 1.20E-04| 1.2  | cell cycle    |
| III   | Omp   | 1.40E-04| 1.60E-05| -1.6 | marker protein|
|       | Ptdss2| 9.00E-05| 8.00E-04| -1.4 | enzymatic activity |
|       | Tfps  | 2.10E-03| 3.60E-04| 2.1  | endocytosis   |
| IV    | Casp6 | 4.50E-03| 6.60E-04| -1.4 | Apoptosis     |
|       | C66B  | 4.40E-04| 2.90E-05| 2    | macrophage marker |
|       | Sfmi-4| 1.50E-04| 7.30E-06| 4    | cell cycle    |
The expression level of the gene for olfactory marker protein (Omp), which is expressed in mature OSNs, was unchanged at 2 hr and 8 hr following OBX. Its first significant change in expression was a down-regulation at 16 hr post-OBX. Its expression continued to decrease at 48 hr. The fluctuation pattern of Omp expression was indexed by (0 0 0 -1 -1), shown in the upper right corner of the graph. Diagrams in the following figures were plotted similarly.

The expression level of the gene for olfactory marker protein Omp, which is expressed in mature OSNs, was unchanged at 2 hr and 8 hr following OBX. The initial change, a down-regulation at 16 hr post-OBX, indicated that degeneration was evident between 8 hr and 16 hr post-OBX (Figure 2). The significant down-regulation of Omp (p = 1.6E-5, Table 1) continued to the 48 hr time-point that was accompanied by a -1.6 FC in OMP mRNA, indicating degenerative changes in OSNs accompanying their cell death.

The genes for programmed cell death 5 (Pdcd5), centrin 3 (Cetn3), and Kit are examples of Class I genes that showed their first significant change in temporal expression at 2 hr post-OBX, with Pdcd5 and Cetn3 being up-regulated and Kit being down-regulated (Figure 3). In contrast, Class II genes showed their first significant change in temporal expression at 16 hr post-OBX (Figure 4); they included the genes for chemokine (C-C motif) ligand 2 (Ccl2), colony stimulating factor 3 (Csf3), and budding uninhibited by benzimidazoles 3 homolog (Bub3) that were up-regulated simultaneously. The genes for phosphatidylinerse synthase 2 (Ptdss2) and the transferrin receptor (Tfrc) are examples of Class III genes that showed their first significant change in temporal expression at 16 hr post-OBX, with Ptdss2 and Tfrc down-regulated and up-regulated respectively (Figure 5). The genes identified statistically as Class IV genes were initially quiescent until their first significant change in expression at 48 hr post-OBX (Figure 6) as shown by the genes for caspase 6 (Casp6), CD68 antigen (Cd68), and schlafen 4 (Slf4). From a functional perspective, the regulation of the genes for Pdcd5, Ptdss2, and Casp6 at 2 hr, 16 hr, and 48 hr respectively suggested that the molecular mechanisms associated with OSN degeneration and cell death occurred over a 2d time frame that is consistent with the systematic down-regulation of the gene for Omp. The up-regulation of the genes for Ccl2 and Cd68 at 8 hr and 48 hr respectively suggested the expression of macrophage chemoattractant protein-1 (CCL2) by resident and recruited macrophages identified phenotypically with CD68 antibody that indicate the delivery of bioactive molecules associated with the earliest regeneration of the sensory epithelium. The genes for Kit, Csf3, Bub3, and Slfn4 are broadly defined as having growth factor activity, which suggested that molecular mechanisms associated with the transformation of progenitor cells into mature OSNs through the proliferative stages of the cell cycle was initiated within 2 hr of OBX and continued throughout the following 48 hr. The results of our statistical and bioinformatics analyses clearly indicate that the categorization of genes into four Classes based on their first significant temporal regulatory event has biological relevance at the cellular level in this neurosensory tissue.

Genes in each class share the same timing of their earliest significant change in expression. The expression pattern of each gene at subsequent time points may vary. We therefore can further cluster genes in each class into subgroups based on their subsequent expression patterns or ‘fluctuation patterns’. For genes in Class I (Figure 1), there theoretically may as many as 34 fluctuation patterns. In our example study, we found 23 different patterns in this class. There were 10, 6, and 2 patterns for genes in Class II, III, and IV respectively. We can use simple diagrams to illustrate these patterns and a series of characters (1, 0, -1) to index their expression patterns as described in the Methods. For example, the fluctuation pattern of Omp expression (Figure 2) can be represented by (0 0 0 1 -1), and the expression of Pdcd5 can be indicated by (0 0 0 0 0) (Figure 3). Genes with the same fluctuation patterns will be finally grouped into the same group (Figure 7). For example, gene Cetn2 and Cetn3 shared the same expression pattern and can be grouped together. This pattern was indicated by (0 1 -1 0 -1). Genes Csf3 and Pdcd8 had their initial responses at hr16 and shared the same fluctuation pattern. These two genes can be classified into another group indicated as (0 0 1 -1 0). Genes CD68 and Slfn4 (Figure 6) can also form a cluster for the same reason.
Discussion

In this study, we adopted linear models to describe our data and used planned linear contrasts to analyze time-course microarray experiments. We identified 1234 genes with significant changes in expression in a microarray study of murine olfactory epithelium, and 1105 of them were grouped into 4 classes based on the timing of their initial changes. We further categorized these 1105 genes into 41 fluctuation patterns. We also used simple diagrams to illustrate these fluctuation patterns and a series of characters (1, 0, -1) to index these patterns. Although the ANOVA F tests were significant, 129 genes cannot be grouped into any of these 4 classes based on our criteria. A significant ANOVA F test among a group of means indicates that the largest contrast among all possible contrasts is significant. Therefore, a gene with a significant F test does not necessarily have a significant selected contrast. Therefore the expression patterns of these genes should be interpreted carefully.

The critical value $|M_{n,m-2,p}|$ used to select significant contrasts is the uniform upper bound for testing a complete set of contrasts regardless of the correlation structure among these contrasts. It is a conservative approach. For planned linear contrasts, the most powerful bound can be found based on the correlation structure of these contrasts [31,32]. In general, the most powerful bound can’t be obtained without knowing the correlation structure among the contrasts [33]. The uniform bound, however, can be obtained from testing a complete set of orthogonal contrasts using the Studentized Maximum Modulus Distribution [34]. In practice, although a little bit conservative, it is straightforward to use this uniform bound to test all contrasts especially when the number of different combinations of contrasts is large.

Our methods emphasized the relative differences between adjacent sampling time points and the direction of the dif-
ferences. The information about exact magnitudes of gene expressed at each time point was not included in our methods. For example, two genes may have the same pattern index \(0\ 1\ -1\ 0\ 0\), but the magnitude of changes for the two genes may be dramatically different. Therefore, even for genes in the same index groups, their expression patterns should be examined with care.

The temporal order in the data was considered in our methods by the selection sequence but was not parameterized in our model. The information about the differences among sampling intervals were also ignored in our analysis. With small sample sizes and non-uniform sampling intervals, which are very common in biomedical research, our methods may be more straightforward and robust than those commonly in use. With large sample sizes and relative uniform sampling intervals, other methods, such as regression analysis, mixture models, or autoregressive models can be applied.

**Conclusion**

Linear models were adopted to describe microarray data, and sequences of planned linear contrasts were used to group genes into different expression patterns based on their initial and subsequent changes in expression. Our methods are particularly suitable for analysis of microarray experiments in which it is often difficult to take sufficiently frequent measurements and/or the sampling intervals are non-uniform. Our methods can also be extended to designs with more than one factor.

**Methods**

**Microarray experiments**

The goal of this study was to investigate the induction of gene regulation at short time intervals (2, 8, 16, and 48 hrs) following deafferentation of olfactory sensory neurons by target ablation (olfactory bullectomy, OBX) compared with sham controls [35]. Total RNA was isolated from the olfactory epithelium of 3 male mice per time point (1 GeneChip/mouse). Following hybridization with Affymetrix GeneChips MG U74Av2, 3 chips per time point (a total of 15 GeneChips), the signal intensities were generated by Affymetrix Microarray Suite v5.0.

In our study, all positive control genes and genes that resulted in "absent" calls for all chips across all time points were removed from further analysis. If there was no evidence that these genes were expressed in any of the samples, then these genes can be removed to reduce prob-
lems associated with multiple comparisons. Other methods of removing low intensity points were also suggested by Bolstad et al., 2003[36]. All ESTs were also removed from the analysis because the primary aim of these experiments was to identify known genes that were differentially regulated; eliminating ESTs further reduced problems with multiple comparisons. After data filtering steps, 6464 genes remained, and the background-corrected intensities of these genes were subjected to further statistical analyses.

Algorithm and analysis
Statistical model
We use a linear model to describe the experiment. Let $Y_g$ be the vector of observed expression levels for gene $g$, $g = 1, ..., 6464$ then

$$Y_g = X \beta_g + \epsilon_g$$

where $X$ is the matrix of known constants, $\beta_g = (\mu_{g1}, \mu_{g2}, ..., \mu_{gm})$, and $m$ is the number of time points ($m = 5$ in this study). $\epsilon_g$ is the random error, and we assume $\epsilon_g \sim \text{MVN}(0, \sigma_g^2I)$.

Reverse Helmert series
A contrast is a linear combination of parameters for which the coefficients sum to zero. A complete set of orthogonal contrasts is a set of k-1 contrasts in k treatments (or treatment combinations) which provides a complete partitioning of the variability among parameters into mutually exclusive and exhaustive parts. Each contrast in such a set is orthogonal to every other remaining one [37]. One commonly used complete set of orthogonal contrasts is the reverse Helmert series, in which one treatment group is compared with the average of all remaining treatment groups. Subsequent contrasts eliminate the first group and then proceed by comparing one of the remaining groups to the average of the other remaining groups, as shown below:

$$L\beta = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & 2 & 2 & -1 \\ . & . & . & . \\ 1 & m-1 & m-1 & m-1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_m \end{pmatrix}$$

One of the advantages of the Reverse Helmert contrasts is that these contrasts are orthogonal and, hence, contrasts among the sample means are uncorrelated. Basing tests on uncorrelated contrasts avoids the problems inherent in interpreting conditional tests. Adjacent Differences(AD) are sometimes used to identify the point at which initial gene expression occurs. However, these contrasts are not orthogonal and consecutive contrasts have a correlation of 0.5. Consequently, the probability of identifying the correct threshold is lower for AD than for the Reverse Helmert Contrasts.
**Clustering genes based on the timing of their initial responses**

The reverse Helmert series can test the following \( m-1 \) hypotheses sequentially:

\[
H_{10} : \mu_1 - \mu_2 = 0 \\
H_{20} : \frac{\mu_1 + \mu_2}{2} - \mu_3 = 0 \\
\vdots \\
H_{m0} : \frac{\mu_1 + \mu_2 + \ldots + \mu_{m-1}}{m-1} - \mu_m = 0
\]

Genes will be partitioned into \( m-1 \) classes based on the testing results of \( H_{10} \sim H_{m0} \), where \( s = m-1 \). Class 1 contains genes that reject \( H_{10} \) genes that reject \( H_{20} \) from the remaining list are grouped into class 2, and so on. Class \( s \) includes genes that reject \( H_{m0} \) without rejecting the previous \( s-1 \) hypotheses. Therefore Genes in Class 1 are considered to be early responding genes whose expression levels are significantly altered during the first sampling interval, that is, at the 2nd sampling time point. Genes that do not change their expression levels until the 3rd sampling time point are collected in Class 2, and so on. As indicated by the described partition process, genes within a class share the same timing of onset or cessation of expression.

**Clustering genes within a class**

Genes in each of these above \( m-1 \) classes can be further classified based on their ‘fluctuation’ shapes at the subsequent sampling points. For gene \( g \) in class \( j \), where \( j = 1, 2, \ldots, s \), the following \( s-j \) contrasts are needed,

\[
H_{g}^{j1} : \mu_{j+1} - \mu_{j+2} = 0 \\
H_{g}^{jk} : \begin{cases} 
\mu_{j+k} - \mu_{j+k+1} = 0 & \text{if reject } H_{g}^{j+k-1} \\
\frac{\mu_{j+k-1} + \mu_{j+k}}{2} - \mu_{j+k+1} = 0 & \text{if fail to reject } H_{g}^{j+k-1} 
\end{cases} \quad k = 2, \ldots, s-j
\]

Therefore, a specific sequence of \( m-1 \) hypotheses will be performed for each gene to determine its expression pattern. Let \( \tilde{\lambda}_{g}^{jk} \) be the unbiased estimate of the contrast corresponding to the hypothesis \( H_{g}^{jk} \), in a balanced experiment with sample size \( n \) in each treatment group, the statistic

\[
T_{g}^{jk} = \frac{\tilde{\lambda}_{g}^{jk}}{\sqrt{\frac{\sum_{i=1}^{m} \sigma_{g}^{2}}{n}} \sqrt{\text{MSE}_{g}} \sqrt{\text{MSE}_{g}^{2} + \text{MSE}_{g}^{2}}} \sim t_{v}
\]

where \( c_{i} \) is the \( i \)th coefficient for the contrast, and \( i = 1, \ldots, m \). \( \text{MSE}_{g} \) is the usual unbiased estimate of \( \sigma_{g}^{2} \), and \( v = N - m \) is the error degree of freedom (df), where \( N = mn \).

**Indexing gene expression patterns**

Let the state of the first observation be 0, for gene \( g \), its expression profile can be transformed into a sequence of expression fluctuation as follows:

\[
S_{g}^{jk} = \begin{cases} 
1 & \text{if } k \geq j, \text{ reject } H_{g}^{j+k}, \text{ and } \lambda_{g}^{j+k} < 0 \\
0 & \text{if } k < j \text{ or fail to reject } H_{g}^{j+k} \\
-1 & \text{if } k \geq j, \text{ reject } H_{g}^{j+k}, \text{ and } \lambda_{g}^{j+k} > 0 
\end{cases} \quad j = 1, \ldots, s, \quad k = 1, \ldots, s
\]

where \( k = 1, 2, \ldots, s \) is an index, where \( k = r \) if it is the \( r \)th contrast for gene \( g \). \( S \) is the transformed value of the gene expression profiles. Thus an \( m \)-time-point expression profile is transformed into an \( m-1 \)-state sequence of expression fluctuation consisting of a character set \( \{1, 0, -1\} \). Each character in the sequence indicates whether the mean expression level of the gene is significantly up-regulated \( (1) \), not altered \( (0) \), or significantly down-regulated \( (-1) \) at the next time point, while the whole sequence represents the fluctuation pattern of the gene expression. Besides the pattern in which the gene’s expression level is unchanged throughout the entire sampling period, there are at most \( 2 \times 3^{m-1} \) fluctuation patterns for genes in Class \( k \). There are no more than \( 3^{m-1} \) patterns of expression in total in an \( m \)-time-point microarray experiment.

**Multiple testing control**

An ANOVA F test was performed for each gene to identify the differentially expressed genes. This F test is testing the hypothesis \( H_{g}^{1} = \mu_{g1}, \ldots, \mu_{gm} \), which is equivalent to test the composite hypothesis \( L\beta = 0 \). The BH procedure of controlling FDR at 5% was applied to the P values of the F test. A cutoff point \( \alpha_{\text{new}} \) which is equal to the largest P value considered to be significant, was determined by the above BH procedure. By this procedure, each test for the gene that rejected the F test is at least \( \alpha_{\text{new}} \) level test.

For each selected gene, a specific sequence of \( m-1 \) contrasts was tested to determine its expression pattern. To be consistent with the BH procedure performed, the family-wise error rates (FWER) for these genes have to be controlled at least at the level of \( \alpha_{\text{new}} \). In this study, we controlled the maximum family-wise error rates (MFWER) by using the Studentized Maximum Modulus distribution [34]. The following theorem in the Appendix outlined the concern of gene-based controlling MFWER at the level of \( \alpha_{\text{new}} \).

**Outline of the analysis**

A short summary of the statistical methods used in this study follows:
1. Linear models were used to describe the data based on the experimental design. For each gene, an ANOVA F test was performed based on the described model, and the corresponding P-value was obtained.

2. To adjust for multiple tests based on the large number of genes, the BH method of controlling FDR [22] at 5% was applied to the P-values obtained above, providing a list of genes (list I) that exhibit significant differences among the means of the 5 sampling points.

3. Using $\alpha_{\text{new}}$, which equals the largest P-value determined to be significant in step 2 as the cut-off point, we grouped genes in list I into 4 classes based on the timing of their initial responses by testing the reverse Helmert contrasts sequentially. The Studentized Maximum Modulus distribution parameter $m - 2 = 3$ and $v = 10$ were used in this example study, where $\alpha_{\text{new}} = 0.009545$ and $M_{\text{new},m-2,v} = 3.8651$.

4. Using the same critical value $M_{\text{new},m-2,v}$, we further clustered genes in each of the above classes by testing appropriate contrasts for the subsequent sampling time points.

5. Based on the results of the $m-1$ contrasts for each gene, we also can select genes which share similar pattern of expression for any number of sampling intervals from the beginning.

Statistical software
We used the SAS (version 9.0) proc GLM procedure to do model fitting and significance analysis. The SAS program implementing linear models for the olfactory sensory epithelia data is available [38].

Authors’ contributions
HL carried out the statistical analyses and formulation of statistical methods. YL automized the statistical analysis using Splus/R. TVG and MLG carried out the molecular genetics studies. CLW and AJS supervised the study. All authors contributed to the writing of this manuscript. All authors read and approved the final manuscript.

Appendix
Theorem For any balanced one-way model with $m$ treatment groups and assuming normality and equal variance $\sigma^2$, let $\lambda_1, \lambda_2, ..., \lambda_k$ be an arbitrary complete set of contrasts such that

$$
\lambda_i = \sum_{i=1}^{m} c_i \mu_i, i = 1, 2, ..., K
$$

Under the null hypothesis, let $P$ be the distribution of the vector $\lambda = [\lambda_1, \lambda_2, ..., \lambda_k]$ with mean 0 and covariance matrix $\Sigma$, let $P_K$ be the distribution of the vector of a complete set of contrasts $\lambda^* = [\lambda_1^*, \lambda_2^*, ..., \lambda_K^*]$ with the covariance matrix $\Sigma_K = I \sigma^2$ is the diagonal of $\Sigma$ then the gene based maximum family-wised error rate (MFWER) at any level of $\alpha$ of testing a specific sequence of contrasts (list in the Methods) after rejecting the overall F test is achieved by comparing $|T_i|$ with $|M_{a,b,1,v}|$, where $v$ is the df of error,

$$
T_i = \sqrt{\frac{\sum_{i=1}^{m} c_i^2}{MSE(n - 1)}} M_{a,b,1,v}
$$

and $|M_{a,b,1,v}|$ is the $100(1 - \alpha)$ percentile from the Studentized Maximum Modulus distribution.

Proof Let $\lambda_1, \lambda_2, ..., \lambda_k$ be any complete set of contrasts, then let

$$
V_0 = \{0 \leq i \leq K: \lambda_i = 0\} \text{ and } V_1 = \{0 \leq j \leq K: \lambda_j \neq 0\},
$$

let test function

$$
\phi(\lambda_i) = \begin{cases} 1 & \text{if reject the } H_{0i} : \lambda_i = 0 \\ 0 & \text{if fail to reject the } H_{0i} : \lambda_i = 0 \end{cases}
$$

(1) Suppose $V_1$ is empty such that $\lambda_i = 0 \forall i$, then MFWER is

$$
\text{MFWER} = \Pr\left\{ \text{at least one false reject} \right\}
= \Pr\left\{ \text{F test rejected}| \bigcup_{i \in V_0} \phi(\lambda_i) = 1 \right\}
\leq \Pr\{ \text{F test rejected} \}
\leq \alpha
$$

(2) $V_0$ is empty such that $\lambda_i \neq 0 \forall i$, then MFWER is 0.
(3) Suppose that neither $V_0$ nor $V_1$ is empty, then MFWER is

$$MFWER = P_{K_1} \left[ \left\{ \begin{array}{l}
\text{at least one false rejection} \\
\end{array} \right. \right]$$

$$\leq P \left[ \bigcup_{v_0} \{ \theta(v_0) = 1 \} \right]$$

$$\leq P \left[ \bigcup_{v_0} \{ \theta(v_0) = 1 \} \right]$$

where $v_0$ is the number of elements in $V_0$ and $\max \{ v_0 \} = K - 1$

$$= 1 - P \left[ \bigcap_{v_0} \{ \theta(v_0) = 0 \} \right]$$

$$\leq - P \left[ \bigcap_{v_0} \{ \theta(v_0) = 0 \} \right]$$

$$\leq - P \left[ \bigcup_{v_0} \{ \theta(v_0) = 1 \} \right]$$

$$\leq - P \left[ \bigcup_{v_0} \{ \theta(v_0) = 1 \} \right]$$

$$= p_{K_0} \left[ \bigcup_{v_0} \{ \theta(v_0) = 1 \} \right]$$

under $p_{K_0}$ based on Sidak’s inequality (8) [39],

$$\max \left| T_i \right|$$

has a Studentized Maximum Modulus distribution [34] with parameter $K - 1$ and $v$, let

$$MFWER = \alpha^* (\alpha, m - 1, v) = \max \{ \alpha(\alpha, m - 1, v) \} = \alpha,$$

then $q = |M_{K_0,1,1}|$ is the 100(1-$\alpha$) percentile from above distribution.

Acknowledgements

This work was supported by NIH AG-016824 (TVG) and NIH-P20RR16481 and NSF-EPS-0132295 (AS). We also wish to thank Donna Wall, Microarray Core Facility, and Radhika Vaishnav, M.S., Department of Physiology, for their expertise.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. PNAS 1998, 95(25):14863-14868.

2. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. 1999, 22(3):281-288.

3. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS 1999, 96(6):2907-2912.

4. Garrity GM, Liburn TG: Self-organizing and self-correcting classifications of biological data. Bioinformatics 2005, 21(10):2309-2314.

5. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 2000, 97(1):262-267.

6. Quackenbush J: Computational analysis of microarray data. Nat Rev Genet 2001, 2(6):418-427.

7. Bar-Joseph Z, Gerber G, Giord DK, Jaakkola TS, Simon I: A new approach to analyzing gene expression time series data. Proceedings of RECOMB, Washington DC, USA 2002:39-48.

8. Park T, Yi S-G, Lee S, Lee SY, Yoo D-H, Ahn J-1, Lee Y-S: Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics 2003, 19(6):694-703.

9. Ramoni MF, Sebastiani P, Kohane IS: From the Cover: Cluster analysis of gene expression dynamics. PNAS 2002, 99(14):9121-9126.

10. Sasik R, Iranfar N, Hwa T, Loomis WF: Extracting transcriptional events from temporal gene expression patterns during Dictyostelium development. Bioinformatics 2002, 18(1):61-66.

11. Ji X, Le-Ling J, Sun Z: Mining gene expression data using a novel approach based on hidden Markov models. FEBS Letters 2003, 542(1-3):125-131.

12. Schlep A, Schonhuth A, Steinhoff C: Using hidden Markov models to analyze gene expression time course data. Bioinformatics 2003, 19(90001):225-263.

13. Yeung KY, Bumgarner RE, Raftery AE: Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 2005, 21(10):2394-2402.

14. Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, Umbach DM: Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 2003, 19(7):834-841.

15. Bensmail H, Celeux G, Raftery AE, Robert CP: Inference in model-based cluster analysis. Statistics and Computing 1997, 7:1-10.

16. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: Model-based clustering and data transformations for gene expression data. Bioinformatics 2001, 17(10):977-987.

17. Fraley C, Raftery AE: How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal 1998, 41:578-588.

18. Medvedovic M, Sivaganesan S: Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 2002, 18(9):1194-1206.

19. Pan W, Lin J, Le C: Model-based cluster analysis of microarray gene-expression data. Genome Biology 2002, 3(2):research0009-0001-research0009-0008.

20. Wester X, Fuhrman S, Caron GB, Smith S, Barker JL, Somogyi R: Large-scale temporal gene expression mapping of central nervous system development. PNAS 1998, 95(1):334-339.

21. Muller-Levent CS, Cho KH, Wolkenhauer O: Microarray data clustering on temporal variation: FCV with TSD preclustering. Appl Bioinformatics 2003, 2(1):35-45.

22. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc 1995, B(75):289-300.

23. Benjamini Y, Yekutieli D: The control of the false discovery rate under dependency. Ann Stat 2001, 29:1165-1188.

24. Benjamini Y, Yekutieli D: Quantitative Trait Loci Analysis using the False Discovery Rate. Genetics 2005, genetics.104.036699.

25. Efron B, Tibshirani R, Storey JD, Tusher V: Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001, 96:1151-1160.

26. Storey JD: The positive false discovery rate: A Bayesian interpretation and the Q-Value. Technical Report 2001-12, Department of Statistics, Stanford University. 2001.

27. Benjamini Y, Yekutieli D: Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003, 19(3):348-357.

28. Grant GR, Liu J, Stoekert CJ Jr: A practical false discovery rate approach to identifying patterns of differential expression in microarray data. Bioinformatics 2005, 21(1):2684-2690.

29. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005, 21(13):3017-3024.

30. Li H, Wood C, Getchell T, Getchell M, Stromberg A: Analysis of oligonucleotide array experiments with repeated measures using mixed models. BMC Bioinformatics 2004, 5(1):209.

31. Nelson PR: Multivariate normal and t distributions with $P = \alpha_{1-k}$, Commun Stat Simulation & computation 1982, 11:239-248.
32. Kirk RE: Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Brooks/Cole; 1982:92.
33. Wilcoxon RR: New designs in analysis of variance. Ann Rev Psychol 1987, 38:29-60.
34. Bechhofer RE, Dunnett CW: Multiple comparisons for orthogonal contrasts: example and table. Technometrics 1982, 24:213-222.
35. Getchell TV, Liu H, Vaishnav RA, Kwong K, Stromberg AJ, Getchell ML: Temporal profiling of gene expression during neurogenesis and remodeling in the olfactory epithelium at short intervals after target ablation. Journal of Neuroscience Research 2005, 80(3):309-329.
36. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003, 19(2):185-193.
37. Klockars AJ, Hancock GR: Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. Psychological Bulletin 1992, 111(3):505-510.
38. Contrast [http://www.mc.uky.edu/UKMicroArray/contrast.txt]
39. Sidak Z: Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Am Stat Asso 1967, 62:426-433.