Funct Integr Genomics
DOI 10.1007/s10142-008-0096-5

ORIGINAL PAPER

Identifying differentially expressed genes in human acute leukemia and mouse brain microarray datasets utilizing QTModel

Jian Yang · Yangyun Zou · Jun Zhu

Received: 15 July 2008 / Revised: 13 August 2008 / Accepted: 17 August 2008

Abstract One of the essential issues in microarray data analysis is to identify differentially expressed genes (DEGs) under different experimental treatments. In this article, a statistical procedure was proposed to identify the DEGs for gene expression data with or without missing observations from microarray experiment with one- or two-treatment factors. An *F* statistic based on Henderson method III was constructed to test the significance of differential expression for each gene under different treatment(s) levels. The cutoff *P* value was adjusted to control the experimental-wise false discovery rate. A human acute leukemia dataset corrected from 38 leukemia patients was reanalyzed by the proposed method. In comparison to the results from significant analysis of microarray (SAM) and microarray analysis of variance (MAANOVA), it was indicated that the proposed method has similar performance with MAANOVA for data with one-treatment factor, but MAANOVA cannot directly handle missing data. In addition, a mouse brain dataset collected from six brain regions of two inbred strains (two-treatment factors) was reanalyzed to identify genes with distinct regional-specific expression patterns. The results showed that the proposed method could identify more distinct regional-specific expression patterns than the previous analysis of the same dataset. Moreover, a computer program was developed and incorporated in the software QTModel, which is freely available at http://ibi.zju.edu.cn/software/qtmodel.

Keywords Microarray · Two-treatment factors · Differentially expressed gene (DEG)

Introduction

Microarray experiments usually involve a large number of complex procedures, including RNA preparation, hybridization, scanning, etc., which may bring various random variations into the raw intensities (Brown and Botstein 1999). Moreover, microarray experiments generate large and complex multivariate datasets with thousands of interrelated variables (genes), small sample sizes (sometimes less than ten arrays) from unbalanced design, little or no replication, and missing values. The essential problem of gene expression microarray data analysis is to identify differentially expressed genes (DEGs) under different treatment levels. Various statistical methods have been proposed and applied for detecting genes with differential expression, of which, MAANOVA (microarray analysis of variance; Kerr et al. 2000) and SAM (significance analysis of microarray; Tusher et al. 2001) are two of the most prevalent methods. The MAANOVA method is the first to employ the fixed-effects ANOVA (analysis of variance) model to identify DEGs, and SAM uses a modified *t* test and permutation test to assess significance. The mixed model is a general and powerful approach for microarray experiments with several sources of random variation. The most commonly used method for mixed model analysis of microarray is the gene-specific modeling first proposed by Wolfinger et al. (2001). This method uses *t* test to assess the

J. Yang · J. Zhu (✉)
Institute of Bioinformatics,
College of Agriculture and Biotechnology,
Zhejiang University,
Hangzhou, Zhejiang 310029, China
e-mail: jzhu@zju.edu.cn

Y. Zou
Institute of Biomedical Sciences,
Fudan University,
Shanghai 200032, China

significance of each gene based on the variance component estimates obtained by restricted maximum likelihood (REML) algorithm. Lu et al. (2005) modified the method of Wolfinger et al. using minimum norm quadratic unbiased estimation (MINQUE) algorithm for variance component estimation and Jackknife resampling technique for significane test. In addition, Cui et al. (2005) extended the ANOVA approach of Kerr et al. by employing the gene-specific model using mixed model.

Most of the previous studies focused on monitoring gene expression changes across a single-treatment factor of interest. The treatment factor can be of two levels such as normal vs. disease samples, wild-type vs. mutant organism, or multiple levels such as different types of tissues or tumors, different time points in a biological process, or different drug treatments. However, many experimenters are increasingly interested in studying complex systems and generate large datasets with more than one treatment factor (Jin et al. 2001), e.g. the experiment investigating gene expression profile in multiple tissues of different mouse strains (Sandberg et al. 2000) or investigating the gene expression changes for different tumor subtypes under different drug treatments (Lamb et al. 2006). However, most of the prevalent software packages are not able to handle data with two-treatment factors. Although, the MAANOVA package can deal with such kind of data, but it cannot directly handle missing observations.

In the present study, a statistical procedure was introduced for identification of DEGs as biomarkers for classifying treatment subtypes (Zou et al. 2006). Moreover, the method was extended to identify DEGs for microarray data with two-treatment factors. In addition, two real datasets were reanalyzed to illustrate the efficacy and reliability of the proposed method.

Models and statistical methods

Currently, different types of expression microarray systems are available, such as the spotted two-color complementary DNA (cDNA) microarray (DeRisi et al. 1996) and the single-color oligonucleotide microarray (Gunderson et al. 2004; Lockhart et al. 1996). The two-color cDNA microarray has two different dyes and relatively large variability between arrays; therefore, four basic experimental effects, i.e., treatment of primary interest, genes, dyes, and arrays, should be considered (Kerr et al. 2000). For single-color microarray platforms (e.g., Affymetrix and Illumina), as to only one dye designed in technology scenario, and high reproducibility of different arrays as well, the effects due to arrays and dyes can be ignored. In the present study, two-color microarray experiment by loop design was considered as an example to illustrate the proposed method. The proposed method can also be used to deal with the single-color microarray data by removing the effects relevant to arrays and dyes in the following models.

Identification of DEGs for one-factor microarray design

Let y_{ijkl} be the logarithm of measurement from the i-th gene, the j-th array, the k-th dye, and the l-th treatment. A full model can be expressed as:

$$y_{ijkl} = \mu + G_i + A_j + D_k + T_l + GA_{ij} + GD_{ik} + GT_{il} + \varepsilon_{ijkl} \tag{1}$$

where μ is the overall mean expression level; G_i is the effect of the i-th gene; A_j is the effect of the j-th array, $A_j \sim N\left(0,\ \sigma_A^2\right)$; D_k is the effect of the k-th dye, $D_k \sim N\left(0,\ \sigma_D^2\right)$; T_l is the effect of the l-th treatment level, $T_l \sim N\left(0,\ \sigma_T^2\right)$; GA_{ij} is the interaction effect between the i-th gene and the j-th array, $GA_{ij} \sim N\left(0,\ \sigma_{GA}^2\right)$; GD_{ik} is the interaction between gene i and dye k, $GD_{ik} \sim N\left(0,\ \sigma_{GD}^2\right)$; GT_{il} is the interaction between the i-th gene and the l-th treatment, $GT_{il} \sim N\left(0,\ \sigma_{GT}^2\right)$; and $\boldsymbol{\varepsilon}_{ijkl}$ is the residual effect accounting for the uncontrollable random errors, $\varepsilon_{ijkl} \sim N\left(0,\ \sigma_\varepsilon^2\right)$. All the terms in model 1 have been clearly interpreted by Kerr et al. (2000).

For the two-color cDNA microarray experiment, a normalization procedure is performed to minimize the global systematical variations involved in the experiment from the raw fluorescence measurements. The raw fluorescence measurements are normalized by $\gamma_{ijkl} = y_{ijkl} - \left(\widehat{\mu} + \widehat{A}_i + \widehat{D}_k + \widehat{T}_l\right)$ using least-square estimation (LSE) method. The normalized data (γ_{ijkl}) can be regarded as a variable for each gene after removing the global biological and technical variations. After that, it can be fitted by the following gene-specific model (Lu et al. 2005; Wolfinger et al. 2001):

$$\gamma_{ijkl} = \mu_i + A_{ij} + D_{ik} + T_{il} + \varepsilon_{ijkl} \tag{2}$$

where μ_i is the mean expression level of gene i; A_{ij}, D_{ik}, and T_{il} are equivalent to GA_{ij}, GD_{ik}, and GT_{il} in model 1 and are assumed to be normally distributed random variables with zero means and variance components σ_{GA}^2, σ_{GT}^2, and σ_{GD}^2, respectively; $\boldsymbol{\varepsilon}_{ijkl}$ is the residual error of gene l, $\varepsilon_{ijkl} \sim N\left(0,\ \sigma_\varepsilon^2\right)$.

Model 2 can be rewritten in matrix notations as:

$$\gamma = \mathbf{W}_{BV}\mathbf{b}_{BV} + \mathbf{W}_{GT}\mathbf{b}_{GT} + \varepsilon \tag{3}$$

where $\mathbf{b}_{GT} = [T_{i1}\ T_{i2}\ \cdots]'$; $\mathbf{b}_{BV} = [\mu_i\ A_{i1}\ A_{i2}\ \cdots\ D_{i1}\ D_{i2}\ \cdots]'$; $\mathbf{W}_{BV}$ and $\mathbf{W}_{GT}$ are the design matrices corresponding to $\mathbf{b}_{\mathrm{BV}}$ and $\mathbf{b}_{\mathrm{GT}}$, respectively; γ is the vector of the normalized data; $\boldsymbol{\varepsilon}$ is the vector of residual effects.

Under the null hypothesis H_0: $\mathbf{b}_{GT}=\mathbf{0}$, we could have an F statistic based on Henderson method III (Searle et al.

1992) to test the significance of differential expression for each gene:

$$F = \frac{SSR\,(\mathbf{b}_{GT}|\mathbf{b}_{BV})/(r_{\mathbf{W}} - r_{\mathbf{W}_{BV}})}{SSE/(n - r_{\mathbf{W}})} \tag{4}$$

where $\mathbf{W} = \left[\mathbf{W}_{BV} \vdots \mathbf{W}_{GT}\right]$; r_{w} and $r_{\mathbf{W}_{BV}}$ are the ranks of $\mathbf{W}$ and $\mathbf{W}_{\mathrm{BV}}$, respectively. The *F* statistic is used to test each of all the genes step-by-step to screen out the DEGs under the treatment of interest. The cutoff *P* value is adjusted at a certain level (say α) to control the experimental-wise false discovery rate (Storey and Tibshirani 2003).

Identification of DEGs by two-factor microarray design

For the microarray design with two treatment factors, model 1 can be extended to:

$$y_{ijklm} = \mu + G_i + A_j + D_k + T_l + S_m + TS_{lm} + GA_{ij} + GD_{ik} + GT_{il} + GS_{im} + GTS_{ilm} + \varepsilon_{ijklm} \tag{5}$$

where T_l is the effect of the *l*-th level of the first treatment, $T_l \sim N\left(0,\ \sigma_T^2\right)$; S_m is the effect of the *m*-th level of the second treatment, $S_m \sim N\left(0,\ \sigma_S^2\right)$; TS_{lm} is the interaction effect between the *l*-th level of the first treatment and the *m*-th level of the second treatment, $TS_{lm} \sim N\left(0,\ \sigma_{TS}^2\right)$; GT_{il} is the interaction between the *i*-th gene and the *l*-th level of the first treatment, $GT_{il} \sim N\left(0,\ \sigma_{GT}^2\right)$; GS_{im} is the interaction between the *i*-th gene and the *m*-th level of the second treatment, $GS_{im} \sim N\left(0,\ \sigma_{GS}^2\right)$; GTS_{ilm} is the interaction between the *i*-th gene, the *l*-th level of the first treatment, and the *m*-th level of the second treatment, $GTS_{ilm} \sim N\left(0,\ \sigma_{GTS}^2\right)$; the remaining variables and parameters have the same definition as those in model 1.

Analogue to that of single-treatment design, a normalization process is conducted to eliminate the effects of global biological and technical variations from the raw intensity measurements by $\gamma_{ijklm} = y_{ijklm} - \left(\widehat{\mu} + \widehat{A}_j + \widehat{D}_k + \widehat{T}_l + \widehat{S}_m + TS_{lm}\right)$ using LSE method. Then, the normalized data (γ_{ijklm}) is fitted by the following gene-specific model for discerning the significant gene,

$$\gamma_{ijklm} = \mu_i + A_{ij} + D_{ik} + T_{il} + S_{im} + TS_{ilm} + \varepsilon_{ijklm}. \tag{6}$$

Using model 6, we can test whether a gene is differentially expressed under the first treatment (*T*), the second treatment (*S*), or the combination of the two treatments (*T*×*S*). Therefore, model 3 can be expressed in three different matrix forms for each of the aforementioned treatments and treatment interaction as

$$\gamma = \mathbf{W}_{BV}\mathbf{b}_{BV} + \mathbf{W}_{GT}\mathbf{b}_{GT} + \varepsilon \tag{7}$$

$$\gamma = \mathbf{W}_{BV}^1\mathbf{b}_{BV}^1 + \mathbf{W}_{GS}\mathbf{b}_{GS} + \varepsilon \tag{8}$$

$$\gamma = \mathbf{W}_{BV}^2\mathbf{b}_{BV}^2 + \mathbf{W}_{GTS}\mathbf{b}_{GTS} + \varepsilon \tag{9}$$

where $\mathbf{b}_{GT} = [T_{i1}\ T_{i2}\ \cdots]'$, $\mathbf{b}_{GS} = [S_{i1}\ S_{i2}\ \cdots]'$, and $\mathbf{b}_{GTS} = [TS_{il1}\ TS_{il2}\ \cdots TS_{ilm}\ \cdots]'$; $\mathbf{b}_{BV} = [\mu_i\ A_{i1}\ A_{i2}\ \cdots\ D_{i1}\ D_{i2}\ \cdots\ S_{i1}\ S_{i2}\ \cdots]'$, $\mathbf{b}_{BV}^1 = [\mu_i\ A_{i1}\ A_{i2}\ \cdots\ D_{i1}\ D_{i2}\ \cdots T_{i1}\ T_{i2}\ \cdots]'$, and $\mathbf{b}_{BV}^2 = [\mu_i\ A_{i1}\ A_{i2}\ \cdots\ D_{i1}\ D_{i2}\ \cdots T_{i1}\ T_{i2}\ \cdots\ S_{i1}\ S_{i2}\ \cdots]'$; γ is the vector of the normalized data; $\boldsymbol{\varepsilon}$ is the vector of residual effects.

Under each of the null hypothesis $H_0 : \mathbf{b}_{\mathrm{GT}}=\mathbf{0}$, H_0: $\mathbf{b}_{\mathrm{GS}}=\mathbf{0}$ or $H_0 : \mathbf{b}_{\mathrm{GTS}}=\mathbf{0}$, the significance of *T*, *S*, or *T*×*S*-specific differentially expressed genes can be tested by the *F* statistic analog to that in Eq. 4. The adjusted cutoff *P* value is used to control the experimental-wise false discovery rate.

Software development

Based on the proposed method, a computer program written by C++ language has been developed for gene expression data analysis and incorporated in the software package QTModel (http://ibi.zju.edu.cn/software/qtmodel). This software provides the access for the selection of DEGs ranked by their statistic scores, with data either from simple designs or from complex designs involving more than one experimental factor. In addition, various statistical algorithms are available in this software to estimate the variance components, fixed effects, and random effects of the factors involved in the experiment.

Results

Analysis of human acute leukemia data (one-treatment factor)

A dataset of human acute leukemia collected by Golub et al. (1999) was reanalyzed in the present study to illustrate the proposed method. In this dataset, the messenger RNA samples obtained from 38 leukemia patients with clinically defined 27 acute lymphoblastic leukemia (ALL) (19 B cell ALL and eight T cell ALL) samples and 11 acute myeloid leukemia (AML) samples were assayed by Affymetrix Hu6800 Genechips (Golub et al. 1999). The essential objective of this experiment was to find out the differentially expressed genes which are highly correlated with class distinction of B cell ALL, T cell ALL, and AML for establishing an efficient molecular diagnosis system of human acute leukemia. The raw fluorescence measurements were log-transformed and normalized by Affymetrix software, and the normalized data were reanalyzed by the proposed method.

In addition, the dataset is also reanalyzed by the SAM package (Tusher et al. 2001) and the MAANOVA package (Kerr et al. 2000). The SAM method uses a modified *t* test by adding a small positive constant to the denominator of the conventional gene-specific *t* test. Permutation test is used to generate an empirical distribution of the *t* statistics for hypothesis test. On the other hand, the MAANOVA uses ANOVA model to fit the data and constructs *F* statistic based on the methods of variance component estimation, such as ML, REML, and MINQUE in which REML is used as the default configuration in the MAANOVA package. Since the methods of SAM and MAANOVA are not able to handle missing data, the K-nearest neighbors imputation (KNNimpute) method (Troyanskaya et al. 2001) was adopted to estimate the missing values. In each of the three analysis procedures, the cutoff *P* value was adjusted to control the false discovery rate at 0.05.

The results showed that MAANOVA identified 188 DEGs, SAM identified 163 DEGs, and the proposed method detected 194 DEGs. All the DEGs detected by SAM can be identified by both MAANOVA method and our method, while other 34 genes which could not be declared as significant DEGs were identified by MAANOVA or the proposed method (Table 1). Of these 34 genes, 26 of which were detected by MAANOVA, and 31 of them can be identified by our method.

After that, we used the DEGs identified by the three methods to classify the samples by hierarchical cluster with Euclidian distance and unweighted pair group method with arithmetic mean linkage criterion using the software ClusterProject (http://ibi.zju.edu.cn/software/clusterproject). These samples have been clinically characterized. Therefore, we can compare the classes obtained by cluster analysis with those by clinical diagnosis to examine the reliability of those DEGs detected by three different methods. The results showed that in all the three cases, ALL and AML could be classified by cluster analysis of DEGs (Fig. 1). Moreover, both MAANOVA method and our method could exactly distinguish the B cell ALL and T cell ALL, but SAM method confounded one B cell ALL sample into the class of T cell ALL.

Analysis of mouse brain data (two-treatment factors)

A mouse brain dataset collected by Sandberg and his colleagues was reanalyzed in the present study. In this dataset, the expression of 13,069 probe sets corresponding to approximately 10,000 genes and expressed sequence tags is assayed using two different oligonucleotide arrays (Mu11KsubA and Mu11KsubB; Sandberg et al. 2000). The experimental samples are collected from six brain regions, i.e., cortex (Cx), cerebellum (Cb), midbrain (Mb), hippocampus (Hp), amygdale (Ag), and entorhinal cortex (Ec) which were prepared in duplicates from two commonly used inbred strains, C57BL/6 and 129SvEv. This dataset provides abundant source of targets for surveying behavioral and neurophysiological differences between the mouse strains, as well as structural and functional differences among the brain regions. The raw fluorescence measurements were log-transformed and normalized by Affymetrix software, and the normalized data were reanalyzed by the proposed method. The cutoff *P* value was adjusted to control the false discovery rate at 0.01.

The results showed that 124 genes were identified to be differentially expressed between the two inbred strains (adjusted cutoff P=0.000172). Of these strain-specific genes, 63 of them having *P* value less than 10^{-5} were analyzed by hierarchical cluster by the software Cluster (Eisen et al. 1998; Fig. 2). It was observed that all samples collected from C57BL/6 strain were grouped together, and those from 129SvEv were classified as one group using the 63 strain-specific genes.

Table 1 The DEGs that can not be identified by SAM but are detected by MANOAVA and the proposed method

Gene ID	Our method	MAANOVA	Gene ID	Our method	MAANOVA
D26156_s_at	√	√	M83221_at	√	×
D30758_at	√	√	M95678_at	√	×
D43682_s_at	√	√	M98399_s_at	√	√
D49950_at	√	√	M98833_at	√	√
D50310_at	√	√	S82470_at	√	×
D63479_s_at	√	×	U12471_cds1_at	√	√
D83032_at	√	√	U15085_at	√	√
HG2788-HT2896_at	√	×	U61734_s_at	√	×
HG627-HT5097_s_at	√	√	U62136_at	√	√
HG688-HT688_f_at	√	√	U73960_at	√	√
L38696_at	√	√	X58431_rna2_s_at	√	×
M12959_s_at	√	√	X69398_at	√	×
M16336_s_at	×	√	X70394_at	√	√
M21119_s_at	×	√	X76061_at	√	√
M26311_s_at	×	√	X77584_at	√	√
M37815_cds1_at	√	√	X89985_at	√	×
M63379_at	√	√	X99584_at	√	√

"√" and "×" represent "detection" and "not detection"

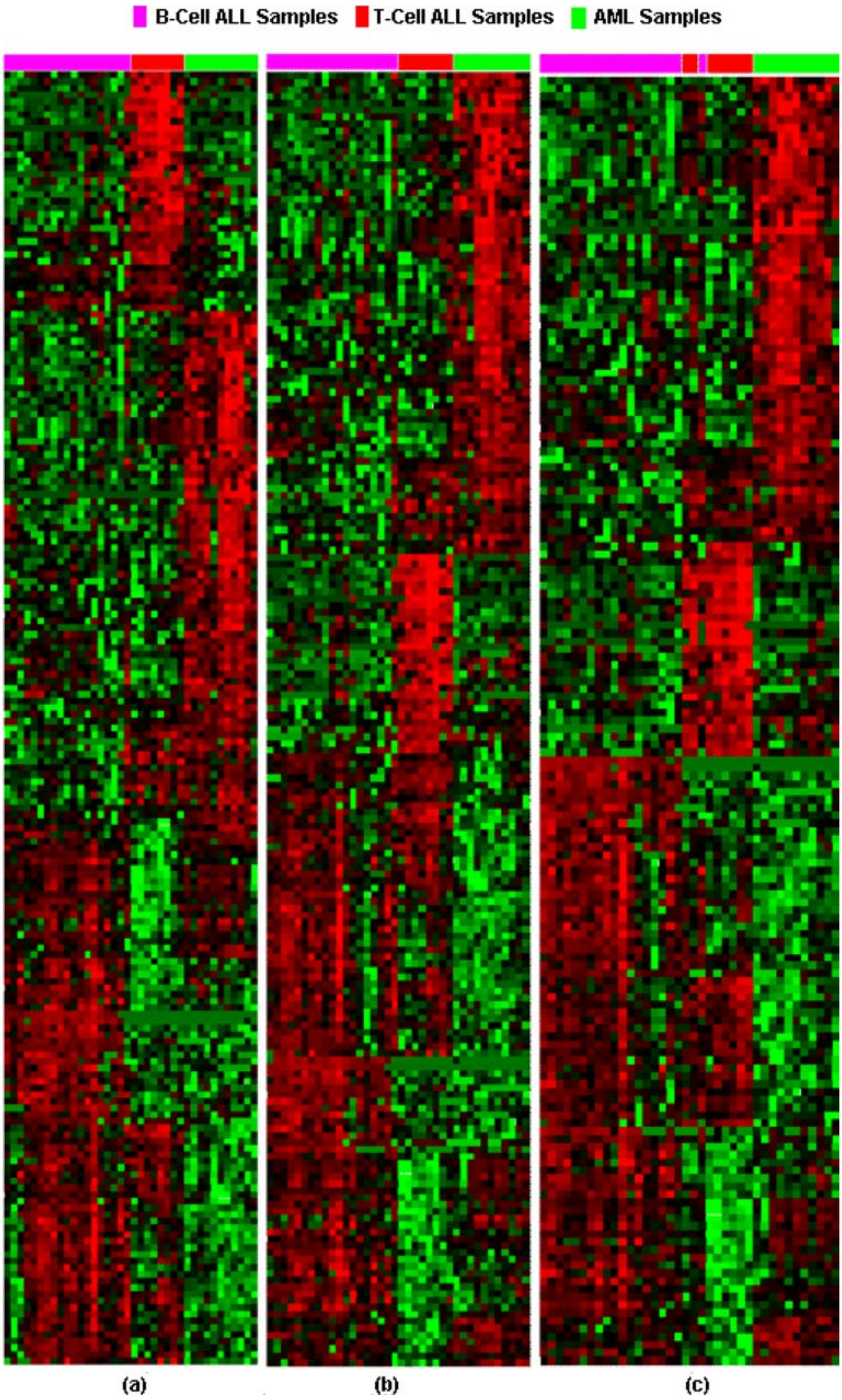


Fig. 1 Classification of human acute leukemia subtypes by cluster analysis of differentially expressed genes detected by the proposed method (**a**), the MAANOVA method (**b**), and the SAM method (**c**), respectively

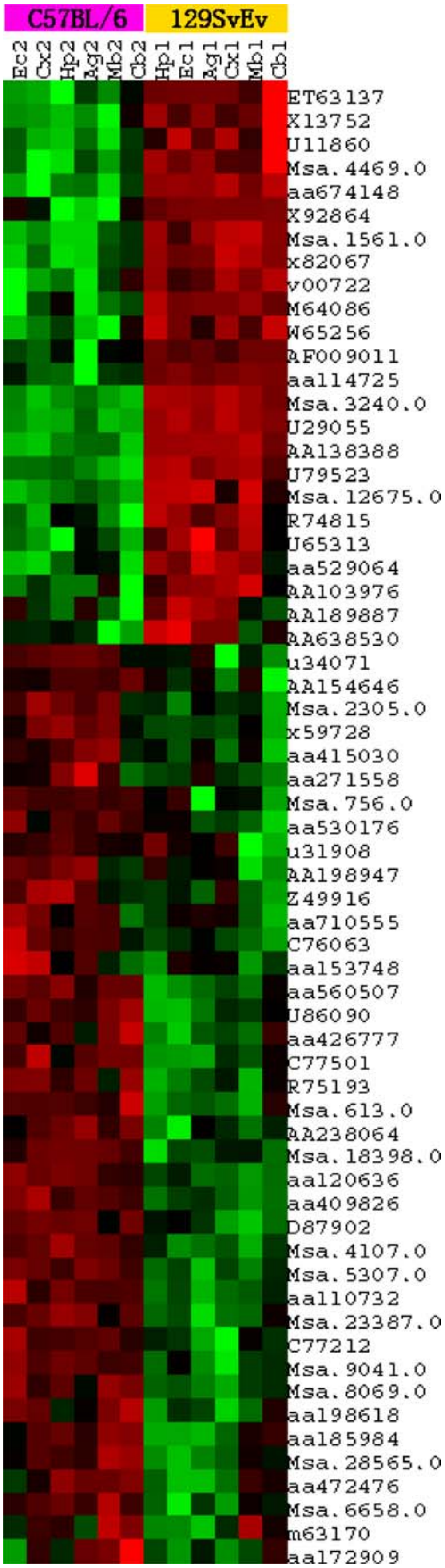


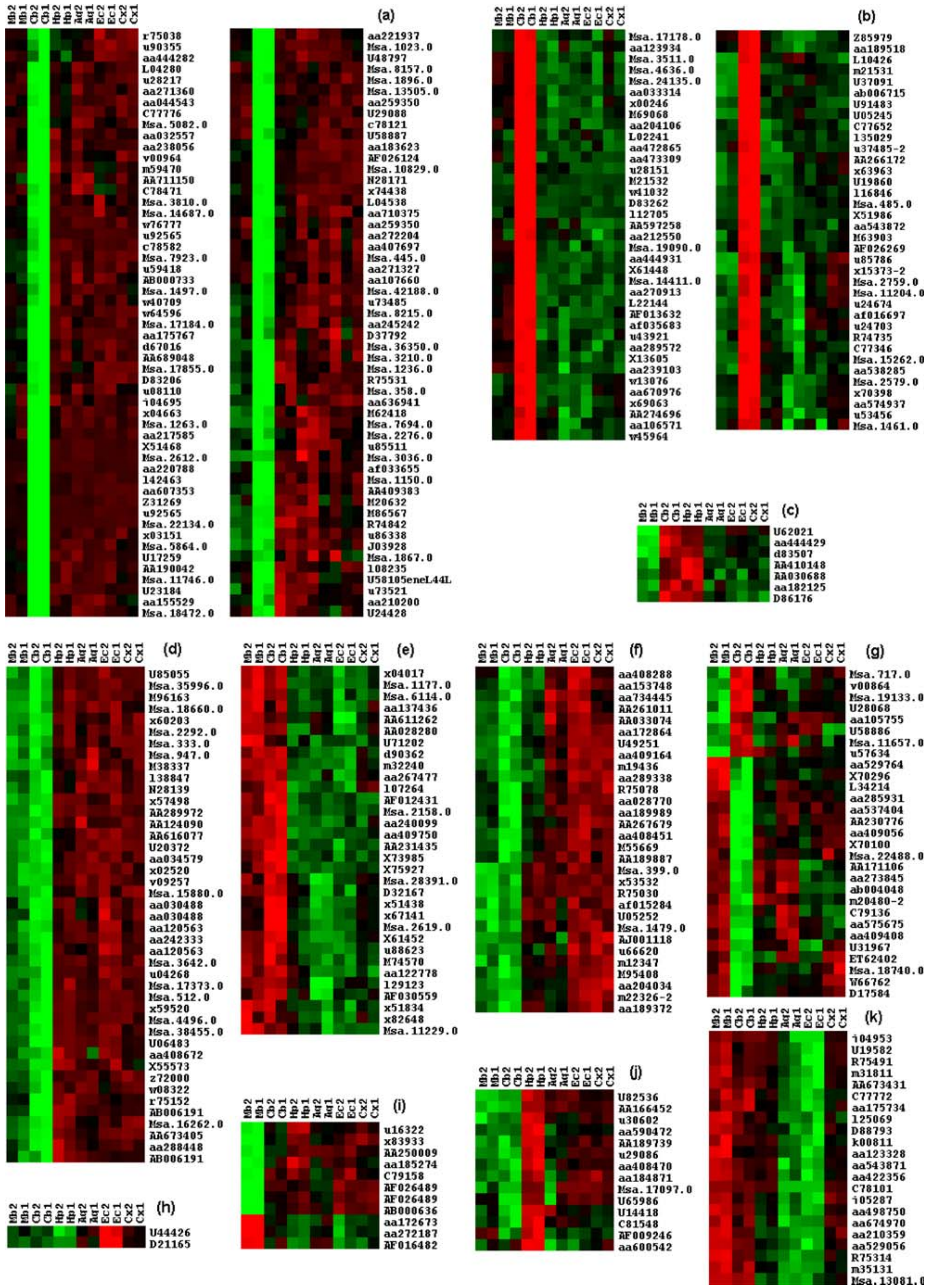
Fig. 2 Classification of two mouse inbred strains, C57BL/6 and 129SvEv, using strain-specific genes identified by the proposed method. Of all the 164 strain-specific genes, 63 of which with P values less than 10^{-5} are visualized in this figure (for space consideration)

Moreover, 1,333 genes were identified having differential expression among different brain regions (adjusted cutoff P=0.003765). Of these regional-specific genes, 411 having P values less than 10^{-5} were filtered for hierarchical cluster analysis and classified those regional-specific genes (RSGs) into 11 groups (Fig. 3). It was observed that a large number of these regional-specific genes were independently up- (Fig. 3b) or down- (Fig. 3a) regulated in cerebellum, which is in agreement with the results in previous analysis utilizing the same dataset (Pavlidis and Noble 2001). Most of RSGs which had differential expression in midbrain were positively or negatively co-expressed in cerebellum (Fig. 3d,e, and g), and only ten of them were independently up- or down-regulated in midbrain (Fig. 3i). For those genes presented in Fig. 3d, they were down-regulated in both cerebellum and midbrain, but were up-regulated in forebrain (represented by cortex, amygdala, hippocampus, and entorhinal cortex), while for those genes presented in Fig. 3e, they were up-regulated in cerebellum and midbrain and down-regulated in forebrain. A few genes were identified with differential expression in hippocampus, and they were mostly positively or negatively co-expressed in cerebellum (Fig. 3c,j). For those genes presented in Fig. 3f, they were enriched in entorhinal cortex and cortex, but were deficient in cerebellum and/or midbrain. Furthermore, 22 were down-regulated in both amygdale and entorhinal cortex, but were enriched in midbrain. Only two genes were identified having distinct up-regulating pattern in this study (Fig. 3h).

Discussion

With the advent of microarray technology, a number of tools have been developed to identify DEGs for various kinds of microarray designs. However, some of the problems such as dealing with data from design with two-treatment factors and missing observations still remain unsolved. In the present study, a new statistical method and its corresponding software are developed to identify DEGs, which has the merits of being able to analyze the microarray data from experiments with one- or two-treatment

Fig. 3 Regional-specific genes (*RSGs*) identified by the proposed method. The RGSs with cutoff P value less than 10^{-5} are selected for cluster analysis. These genes are classified into 11 groups with each group being presented by one sub-figure. The other RGSs without distinct pattern are not shown


Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(a)
r75038
u90355
aa444282
L04280
u28217
aa271360
aa044543
C77776
Msa.5082.0
aa032557
aa238056
v00964
m59470
AA711150
C78471
Msa.3810.0
Msa.14687.0
w76777
u92565
c78582
Msa.7923.0
u59418
AB000733
Msa.1497.0
w40709
w64596
Msa.17184.0
aa175767
d67016
AA689048
Msa.17855.0
D83206
u08110
i04695
x04663
Msa.1263.0
aa217585
X51468
Msa.2612.0
aa220788
l42463
aa607353
Z31269
u92565
Msa.22134.0
x03151
Msa.5864.0
U17259
AA190042
Msa.11746.0
U23184
aa155529
Msa.18472.0
aa221937
Msa.1023.0
U48797
Msa.8157.0
Msa.1896.0
Msa.13505.0
aa259350
U29088
c78121
U58887
aa183623
AF026124
Msa.10829.0
N28171
x74438
L04538
aa710375
aa259350
aa272204
aa407697
Msa.445.0
aa271327
aa107660
Msa.42188.0
u73485
Msa.8215.0
aa245242
D37792
Msa.36350.0
Msa.3210.0
Msa.1236.0
R75531
Msa.358.0
aa636941
M62418
Msa.7694.0
Msa.2276.0
u85511
Msa.3036.0
af033655
Msa.1150.0
AA409383
M20632
M86567
R74842
u86338
J03928
Msa.1867.0
l08235
U58105eneL44L
u73521
aa210200
U24428
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(b)
Msa.17178.0
aa123934
Msa.3511.0
Msa.4636.0
Msa.24135.0
aa033314
x00246
M69068
aa204106
L02241
aa472865
aa473309
u28151
M21532
w41032
D83262
l12705
AA597258
aa212550
Msa.19090.0
aa444931
X61448
Msa.14411.0
aa270913
L22144
AF013632
af035683
u43921
aa289572
X13605
aa239103
w13076
aa670976
x69063
AA274696
aa106571
w45964
Z85979
aa189518
L10426
m21531
U37091
ab006715
U91483
U05245
C77652
l35029
u37485-2
AA266172
x63963
U19860
l16846
Msa.485.0
X51986
aa543872
M63903
AF026269
u85786
x15373-2
Msa.2759.0
Msa.11204.0
u24674
af016697
u24703
R74735
C77346
Msa.15262.0
aa538285
Msa.2579.0
x70398
aa574937
u53456
Msa.1461.0
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(c)
U62021
aa444429
d83507
AA410148
AA030688
aa182125
D86176
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(d)
U85055
Msa.35996.0
M96163
Msa.18660.0
x60203
Msa.2292.0
Msa.333.0
Msa.947.0
M38337
l38847
N28139
x57498
AA289972
AA124090
AA616077
U20372
aa034579
x02520
v09257
Msa.15880.0
aa030488
aa030488
aa120563
aa242333
aa120563
Msa.3642.0
u04268
Msa.17373.0
Msa.512.0
x59520
Msa.4496.0
Msa.38455.0
U06483
aa408672
X55573
z72000
w08322
r75152
AB006191
Msa.16262.0
AA673405
aa288448
AB006191
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(e)
x04017
Msa.1177.0
Msa.6114.0
aa137436
AA611262
AA028280
U71202
d90362
m32240
aa267477
l07264
AF012431
Msa.2158.0
aa240099
aa409750
AA231435
X73985
X75927
Msa.28391.0
D32167
x51438
x67141
Msa.2619.0
X61452
u88623
M74570
aa122778
l29123
AF030559
x51834
x82648
Msa.11229.0
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(f)
aa408288
aa153748
aa734445
AA261011
AA033074
aa172864
U49251
aa409164
m19436
aa289338
R75078
aa028770
aa189989
AA267679
aa408451
M55669
AA189887
Msa.399.0
x53532
R75030
af015284
U05252
Msa.1479.0
AJ001118
u66620
m12347
M95408
aa204034
m22326-2
aa189372
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(g)
Msa.717.0
v00864
Msa.19133.0
U28068
aa105755
U58886
Msa.11657.0
u57634
aa529764
X70296
L34214
aa285931
aa537404
AA230776
aa409056
X70100
Msa.22488.0
AA171106
aa273845
ab004048
m20480-2
C79136
aa575675
aa409408
U31967
ET62402
Msa.18740.0
W66762
D17584
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(k)
i04953
U19582
R75491
m31811
AA673431
C77772
aa175734
l25069
D88793
k00811
aa123328
aa543871
aa422356
C78101
i05287
aa498750
aa674970
aa210359
aa529056
R75314
m35131
Msa.13081.0
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(j)
U82536
AA166452
u30602
aa590472
AA189739
u29086
aa408470
aa184871
Msa.17097.0
U65986
U14418
C81548
AF009246
aa600542
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(i)
u16322
x83933
AA250009
aa185274
C79158
AF026489
AF026489
AB000636
aa172673
aa272187
AF016482
Mb2 Mb1 Cb2 Cb1 Hp2 Hp1 Aq2 Aq1 Ec2 Ec1 Cx2 Cx1
(h)
U44426
D21165

factors and deal with the problem of missing observations. In the present method, we use the F test based on Henderson method III to test the significance of treatment effects of each gene. As compared with the mixed model methods of Wolfinger et al. and Lu et al. using t test based on variance component estimates, the present method has much lower computational load because both the methods of Wolfinger et al. and Lu et al require to iteratively calculate the inverse of an $n \times n$ matrix in each test (where n is the number of observations). In addition, the MAANOVA method can also handle the data from design with two-treatment factors; however, it still has the difficulty in dealing with missing data. In ANOVA approach, the number of replicates for each condition/treatment should be equal; therefore, it is not feasible to handle missing values. Take the human acute leukemia dataset for an example; the samples were assayed by Affymetrix genechips and all the absent calls were treated as missing observation in identification of DEGs. When applying the MAANOVA method to analyze this dataset, it is required to fill up the missing values using the methods such as the KNNimpute method, which may bring extra artificial errors. In the present investigation, an F statistic is constructed in the framework of mixed linear model based on Henderson method III, providing a general form to deal with both the balanced and unbalanced data, thus could overcome the drawback of ANOVA method.

Pavlidis and Noble (2001) also analyzed the same dataset of mouse brain by two-way ANOVA method. In general, our result is consistent with that of Pavlidis and Noble. For example, both studies revealed that cerebellum is the most distinct of brain regions (Fig. 3a,b), and there are many genes that could distinguish forebrain from cerebellum and midbrain (Fig. 3d–g). However, we do not agree that midbrain was the second most distinct of brain regions. We found that the majority of genes differentially expressed in midbrain were also co-expressed in cerebellum. There is no evidence in the present analysis that cortex has the least distinct expression pattern. It is indicated that hippocampus has the least distinct expression pattern, and most of DEGs in cortex are co-expressed in entorhinal cortex. In addition, more regional-specific gene expression patterns are identified in the present analysis as compared to those of Pavlidis and Noble.

References

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21(Suppl.):33–37

Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6:59–75

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14:457–460

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863–14868

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A, Fan JB, Barnard S, Chee MS (2004) Decoding randomly ordered DNA arrays. Genome Res 14:870–877

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet 29:389–395

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7:819–837

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313:1929–1935

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680

Lu Y, Zhu J, Liu P (2005) A two-step strategy for detecting differential gene expression in cDNA microarray data. Curr Genet 47:121–131

Pavlidis P, Noble WS (2001) Analysis of strain and regional variation in gene expression in mouse brain. Genome Biol 2:RESEARCH0042

Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A 97:11038–11043

Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci U S A 100:9440–9445

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116–5121

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8:625–637

Zou YY, Yang J, Zhu J (2006) A robust statistical procedure to discover expression biomarkers using microarray genomic expression data. J Zhejiang Univ Sci B 7:603–607