BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2726–2737
doi:10.1093/bioinformatics/bth319

The Global Error Assessment (GEA) model for the selection of differentially expressed genes in microarray data

Robert Mansourian[1,†], David M. Mutch[1,4,†], Nicolas Antille[1], Jerome Aubert[2], Paul Fogel[3], Jean-Marc Le Goff[2], Julie Moulin[1], Anton Petrov[5], Andreas Rytz[1], Johannes J. Voegel[2] and Matthew-Alan Roberts[1,*]

[1]Nestlé Research Center, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland, [2]Galderma Research & Development, 635, route des Lucioles—B.P.087, F-06902 Sophia Antipolis Cedex, France, [3]Paul Fogel Consultant, 4 rue Le Goff, F-75005 Paris, France, [4]Center for Integrative Genomics, Université de Lausanne, CH-1015 Lausanne, Switzerland and [5]BioDiscovery, Inc. 4640 Admiralty Way, Suite 710 Marina Del Rey, CA 90292, USA

Received on March 15, 2004; accepted on April 8, 2004
Advance Access publication May 14, 2004

ABSTRACT

Motivation: Microarray technology has become a powerful research tool in many fields of study; however, the cost of microarrays often results in the use of a low number of replicates (k). Under circumstances where k is low, it becomes difficult to perform standard statistical tests to extract the most biologically significant experimental results. Other more advanced statistical tests have been developed; however, their use and interpretation often remain difficult to implement in routine biological research. The present work outlines a method that achieves sufficient statistical power for selecting differentially expressed genes under conditions of low k, while remaining as an intuitive and computationally efficient procedure.

Results: The present study describes a Global Error Assessment (GEA) methodology to select differentially expressed genes in microarray datasets, and was developed using an *in vitro* experiment that compared control and interferon-γ treated skin cells. In this experiment, up to nine replicates were used to confidently estimate error, thereby enabling methods of different statistical power to be compared. Gene expression results of a similar absolute expression are binned, so as to enable a highly accurate local estimate of the mean squared error within conditions. The model then relates variability of gene expression in each bin to absolute expression levels and uses this in a test derived from the classical ANOVA. The GEA selection method is compared with both the classical and permutational ANOVA tests, and demonstrates an increased stability, robustness and confidence in gene selection. A subset of the selected genes were validated by real-time reverse transcription–polymerase chain reaction (RT–PCR). All these results suggest that GEA methodology is (i) suitable for selection of differentially expressed genes in microarray data, (ii) intuitive and computationally efficient and (iii) especially advantageous under conditions of low k.

Availability: The GEA code for R software is freely available upon request to authors.

Contact: mroberts@purina.com

INTRODUCTION

One of the first and most important steps in microarray data analysis is to determine those genes that were significantly and differentially regulated according to the condition or experimental parameter being studied. As all subsequent biological interpretation will depend on the accuracy of determining differential gene expression, an efficient and robust statistical analysis is a fundamental prerequisite for experimental interpretation.

Many of the first experiments to benefit from the global view of microarrays utilized a simple fold-change (FC) cut-off for the selection of differentially expressed genes. However, as microarray analysis matures, it is becoming clear that such a selection method makes several assumptions that are out of context with the rest of the experimental and biological data at hand. First, a FC cut-off (typically between 1.8 and 3.0) will treat all results as equal; i.e. a lowly expressed gene

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

with a FC of 3 is just as differentially regulated as a highly expressed gene with a FC of 3. Although exceptions can be found, it is intuitive that there is less confidence in a 3 FC observation at 3 transcripts/sample than for a 3 FC at 1000 transcripts/sample. Therefore, confidence in biological interpretation is difficult to ascertain when selecting microarray results via FC. Second, this commonly used approach does not accommodate for background noise, measurement variability, match/mismatch probe affinity, non-specific binding or low copy numbers—characteristics typical of microarray data that may not be homogeneously distributed (Kothapalli *et al*., 2002). Indeed, these aspects of microarrays are now being analyzed to fully understand how they affect the biological and mathematical interpretation of the data (Ramakrishnan *et al*., 2002).

Using standard statistical measures for each individual gene, such as a student's t-test or a classical ANOVA, can lead to inaccurate estimates of variances [statistical tests for individual genes will be referred to as 'by gene tests' (BGTs)] (Baldi and Long, 2001; Cui and Churchill, 2003). As the use of microarrays is relatively resource intensive for most laboratories, a low k is the experimental norm. A low k decreases the power of BGTs to differentiate between regulated and non-regulated genes. Secondarily, even in the case where reasonable numbers of replicates are achieved, there is always the desire to derive greater statistical power from the inherent multi-dimensional, yet simultaneous, measurements characteristic of microarrays. Therefore, the research community is keen to develop methodologies with these properties and new methods are frequently proposed with the ultimate goal of extracting the most biologically and mathematically significant results from genomic experiments (Baldi and Long, 2001; Brazma *et al*., 2001; Durbin *et al*., 2002; Ghosh, 2002; Huang and Pan, 2002; Kepler *et al*., 2002; Sasik *et al*., 2002; Thomas *et al*., 2001; Troyanskaya *et al*., 2002; Woolf and Wang, 2000). Some of the present authors have previously published a 'Limit fold change' (LFC) model, which attempts to utilize the inherent characteristics of microarray data to overcome the low statistical power of BGTs (Mutch *et al*., 2002). Increasing support for this type of methodology is now appearing in the literature (Baggerly *et al*., 2001; Claverie, 1999; Draghici *et al*., 2003; Hess *et al*., 2001; Jain *et al*., 2003; Kamb and Ramaswami, 2001; Lin *et al*., 2003; Nadon *et al*., 2001). These publications call for a necessity to 'borrow statistical power' through pooling replicates from different genes together during significance testing. In our present research, we extended the concept of 'borrowing statistical power' for estimating noise variance and applied this to ANOVA-based regulation significance tests. We also attempt to make a quantitative estimate of how efficient such a technique is when applied to actual microarray data. We use several approaches, including a comparison with real-time RT–PCR results, to demonstrate the advantages of our method when compared with BGTs. Despite the fact that the methods described in the aforementioned articles are based on approaches different from the ANOVA F-test utilized in this work (Jain *et al*., 2003; Lin *et al*., 2003), the qualitative results stemming from the comparisons with BGTs can be extrapolated to the class of 'pooled replicate noise' methods in general.

The new methodology is termed as Global Error Assessment (GEA) model, indicating its use of calculated inter-array error. More specifically, this model directly generates a robust estimate of the mean squared error (MSE), or equivalently of the SD, by estimating a localized error from the measurement information of several hundred genes with similar expression levels (neighboring genes). The robust MSE of this group of neighboring genes is a highly powerful estimate of the denominator of the F-statistic used in BGTs. It is this principal difference between GEA and other ANOVA-based tests that enable GEA to more powerfully determine differentially expressed genes.

An interesting alternative to the classic ANOVA test is a permutational analog of ANOVA (Dudoit *et al*., 2002). The benefit of utilizing this method lies in the attempt to estimate the actual distribution of the test statistic (F) through the use of thousands of computer permutations. Although more robust than the classical ANOVA, it still suffers from a lack of power under conditions of low k. Furthermore, it is computationally intensive and difficult to implement for the average analyst. Therefore, we also compared GEA with the permutational ANOVA test along these criteria.

To test the methodology, GEA, classical ANOVA and permutational ANOVA were applied to microarray measurements from a biological experiment that compared control and interferon-gamma (IFN-γ) treated skin cells *in vitro*. In this respect, low biological variability and strong induction of genes known to be affected by treatment were important so as to lend credibility to any results obtained from downstream statistical or bioinformatic processing. The DK-7 cell line was selected for the present study due to its previously established high reproducibility (data not shown). Furthermore, IFNs have been chosen as stimulators due to their known effects, such as induction of proinflammatory cytokines, cell adhesion molecules and keratinocyte markers like keratin 17, on skin epithelial cells (Freedberg *et al*., 2001; Sebok *et al*., 1998; Teunissen *et al*., 1998; Wei *et al*., 1999). IFNs are a family of related cytokines that act through their cognate receptor to initiate a signaling cascade, involving the JAK kinase family of tyrosine kinases and the STAT family of transcription factors as well as alternative pathways, that lead to the transcriptional modulation of known IFN-stimulated genes (ISGs) (Ramana *et al*., 2002; Schindler and Darnell, 1995). ISGs have been well documented in previous studies, including a previous microarray analysis (Der *et al*., 1998), and thereby provide a means to confirm or refute microarray results via alternate techniques.

Important to the statistical methodology presented here, the experiment utilized a significant number of replicates to confidently model the gene selection function and enable GEA's comparison with both the classical and permutational ANOVA tests. In addition, a small subset of those genes identified as differentially regulated were confirmed by real time RT–PCR and compared to findings previously reported in the literature. Of equal importance it is simple, intuitive and computationally efficient allowing it to be easily implemented under standard computing environments. Based on all these results, the GEA model provides microarray users with a novel and statistically powerful method for the identification of differentially expressed genes.

MATERIALS AND METHODS

Quality control for RNA integrity

At confluence, keratinocytes from the DK-7 cell line were exposed to 100 IU/ml of IFN-γ in serum-free medium. After 24 h, RNA was extracted using the Qiagen Rneasy Mini Kit (Qiagen SA, Cedex, France). All the samples were monitored by agarose gel and with the Agilent 2100 Bioanalyser (Agilent Biotechnologies, Germany) and consistently demonstrated high-quality RNA (28S/18S ratio $\sim$2, but always $<$3).

cRNA preparation, array hybridization and scanning

According to Affymetrix protocol, 5 μg total RNA was the starting material for all individual samples. In general, total RNA was converted into biotinylated cRNA, hybridized in the Affymetrix probe array cartridge, stained and then quantified. First and second strand cDNA synthesis was performed using the SuperScript Choice System (Invitrogen AG, Basel, Switzerland), according to manufacturer's instructions, but using an oligo-dT primer containing a T7 RNA polymerase binding site. Labeled cRNA was prepared with the RNA Transcript Labeling kit (Enzo Biochem Inc., NY). Biotinylated CTP and UTP were used together with unlabeled NTPs in the reaction, and unincorporated nucleotides were removed with GeneChip® Cleanup Module (Affymetrix, Inc., Santa Clara, CA).

cRNA (20 μg) was fragmented at 94°C for 35 min in buffer containing 200 mM Tris-acetate, pH 8.1, 500 mM KOAc and 150 mM MgOAc. Prior to hybridization, fragmented cRNA in hybridization mix (buffer containing 100 mM MES, 1 M NaCl, 20 mM EDTA, 0.01% Tween-20, 0.5 ng/μl BSA, 0.1 ng/μl herring sperm and Affymetrix controls), was heated to 95°C for 5 min, cooled to 45°C and loaded onto an Affymetrix probe array cartridge. The probe array was incubated for 16 h at 45°C at constant rotation (60 rpm), then exposed to Affymetrix washing and staining protocol. This protocol includes:

- One wash with non-stringent buffer (6$\times$ SSPE, 0.01% Tween-20 and 0.005% antifoam).
- One wash with stringent buffer (100 mM MES, 0.1 M NaCl and 0.01% Tween-20).
- First stain with 0.01 mg/ml streptavidin–phycoerythrin conjugate (Molecular Probes) in buffer containing 100 mM MES, 1 M NaCl, 0.05% Tween-20 and 4 mg/ml of BSA.
- One wash with non-stringent buffer (6$\times$ SSPE, 0.01% Tween-20 and 0.005% antifoam).
- Second stain with 3 μg/ml of biotinylated anti-streptavidin +0.2 mg/ml of IgG in buffer containing 100 mM MES, 1 M NaCl, 0.05% Tween-20 and 4 mg/ml of BSA.
- Third stain with 0.01 mg/ml streptavidin–phycoerythrin conjugate (Molecular Probes) in buffer containing 100 mM MES, 1 M NaCl, 0.05% Tween-20 and 4 mg/ml of BSA.
- One wash with non-stringent buffer (6$\times$ SSPE, 0.01% Tween-20, 0.005% antifoam).

Probe arrays were scanned at 488 nm using an Argon-ion Laser (made for Affymetrix by Agilent). Readings from the quantitative scanning were analyzed with Affymetrix Gene Expression Analysis Software (MAS 5.0).

Experimental design

The GEA model is explained using a simple experiment comparing control versus IFN-γ treated skin cells (one design factor on two levels). To evaluate the biological and the experimental variability, the following experimental design was planned as shown in Figure 1: 'Control' and 'IFN-γ stimulated' keratinocytes were cultured in triplicate using three individual petri Æ10 culture dishes and RNA was independently extracted from each cultured replicate. For each RNA sample, cRNA synthesis was performed in triplicate and each cRNA pool was hybridized to an Affymetrix U133 Gene Chip.

Data analysis

The GeneChip U133 Set, comprised of A and B chips, contains 45 000 different probe sets that corresponds to $\sim$39 000 transcripts derived from $\sim$33 000 well-substantiated human genes. The Affymetrix 'GeneChip software' integrates multiple types of information in order to determine the relative mRNA abundance for a given gene, which is termed the average difference intensity or ADI (http://www.affymetrix.com/); however, the statistical models discussed herein are not Affymetrix software-specific and can be applied to datasets produced by alternate methods. This ADI is then normalized using quantile normalization and natural logarithm transformation. The complete dataset (accession no. GSE1132) is available on the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). All data processing steps described below rely on this normalized

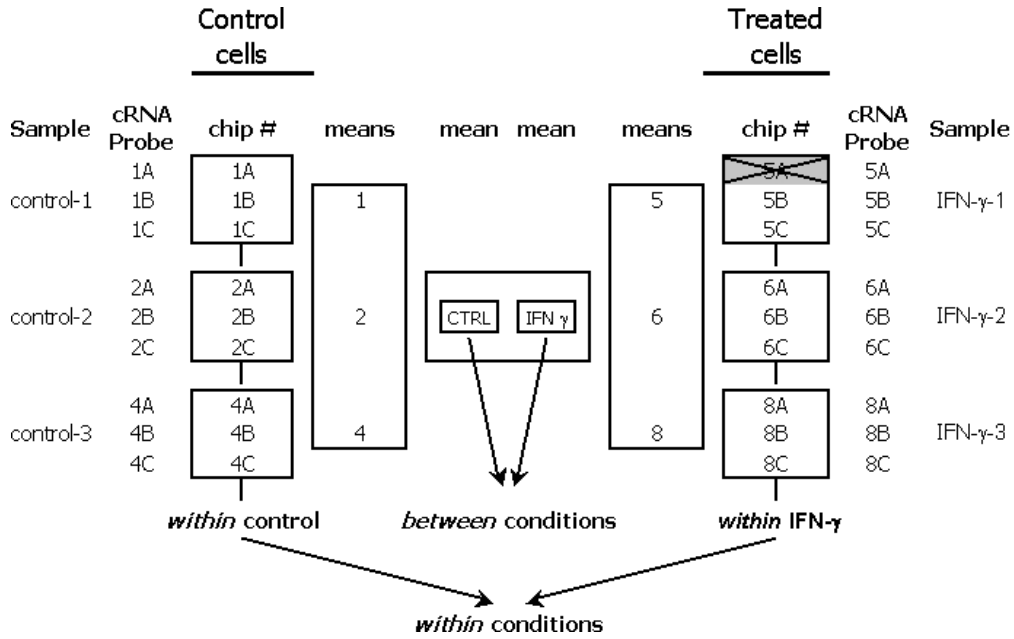

Fig. 1. Experimental design. 'Control' and 'IFN-γ stimulated' keratinocytes were cultured in triplicate using three individual Petri ∅10 culture dishes and RNA was independently extracted from each Petri dish. For each RNA sample, cRNA synthesis was performed in triplicate (A, B, C) and each cRNA pool was hybridized to an Affymetrix U133 Gene Chip. Quality controls were performed according to Affymetrix protocols and all chips, except chip 5A, were found correctly hybridized.

ADI; however, the models presented herein are compatible with other normalization procedures (data not shown).

BGT–ANOVA

By gene testing proceeds by applying an ANOVA to the normalized ADI of each gene. The procedure is explained in the case of n treatments with k replicates each. In our case, $n = 2$ and $k = 3$ (corresponding to the three Petri dishes) or $k = 9$ (corresponding to the nine measurements).

- Estimate the total variability using the sum of squares (SST) and split it into two sources: between treatment and within treatment variability (SST = SSA + SSE).
- Compute the variances of these two sources using mean squares (MS): MSA = SSA/$\mathrm{df_A}$ and MSE = SSE/$\mathrm{df_E}$ with $\mathrm{df_A} = n - 1$ and $\mathrm{df_E} = nk - n$.
- Compute the test statistic F = MSA/MSE, which follows Snedecor's F-distribution with degrees of freedom $\mathrm{df_A}$ and $\mathrm{df_E}$.
- Select genes for which MSA $>$ Limit_α = MSE $* F^{-1}(1 - \alpha \mathrm{df_A}, \mathrm{df_E})$, where α is the significance level.

BGT–permutational ANOVA

The classic ANOVA test described above is based on the assumption of Gaussian distribution imposed on the data points. Even though this assumption is widely accepted for log-transformed microarray data, it is frequently seen that it cannot be assumed and applied to all datasets. The most interesting alternative to classic ANOVA test is permutational analog of ANOVA (Dudoit *et al*., 2002). The benefit of utilizing this method lies with its attempt to estimate actual distribution of the test statistic F described in the previous paragraph. The procedure consists of two steps repeated at least 10 000 times:

- Randomly permute experiment columns of the data table not permuting the experiment labels. Now for every gene you have your measurements randomly distributed between experiments.
- Perform classic ANOVA test on these data and record the value of F-statistic.

These two steps repeated a number of times will produce an estimate of distribution for F-statistics for every particular gene. Actual F-values for original expression measurements compared to this distribution will yield a P-value.

Global Error Assessment

GEA applies ANOVA, but uses a robust estimation of the within treatment variability. Robustness is achieved by two means:

- Averaging within treatment variability of genes that are expressed at a similar level.

Table 1. Sequence information for those genes selected for confirmation by real-time RT–PCR

Gene name	Accession number/AoD	Forward primer	Probe	Reverse primer
gIP10	Hs00171042_m1	N/A	N/A	N/A
STAT1	M97936	GTGGAAAGACAGCCCTGCAT	CGCACCCTCAGAGGCCGCTG	ACTGGACCCCTGTCTTCAAGAC
MXB	M30818	CGAATGAGTGCTGTGTAAGTGATG	TGCTCAAGCCCAGGCCTTGGAC	AAAGGGACCGGCTAACAGTCA
MXA	Hs00182073_m1	N/A	N/A	N/A
isg15	NM_005101	GGGACCTGACGGTGAAGATG	TGGCGGGCAACGAATTCCAGG	GCCAATCTTCTGGGTGATCTG
IRF7	U73036	GCCTGGTCCTGGTGAAGCT	CCTGGCTGTGCCGAGTGCACCT	AGGAAGCACTCGATGTCGTCAT
ISG56K	M24594	GCCTCCTTGGGTTCGTCTATAA	CCCTGGAGTACTATGAGCGGGCCC	TTCTCAAAGTCAGCAGCCAGTCT
IFI-6-16	NM_002038	GGCTACGCCACCCACAAGT	CTGGCTACTCCTCATCCTCCTCACT-ATCGA	GGCCAAGAAGGAAGAAGAGGTT

N/A, not available.

- Using robust estimates of the average variability, instead of classical ones.

The following procedure is therefore implemented:

- Calculate the mean normalized ADI, as well as the MSA and the MSE of each gene.
- Sort genes by ascending mean normalized ADI and group them into bins of 200 consecutive genes (corresponding to ∼100 bins for an Affymetrix GeneChip).
- The MSE of the 200 genes in each bin are summarized using a robust estimation: $\mathrm{MSE_{Robust}} = \mathrm{Median}_{i=1,\ldots,200}\,(\mathrm{MSE}) * \mathrm{df_E}/\chi^{-1}\,(0.5, \mathrm{df_E})$, where χ^{-1} is the inverse of the one-tailed probability of the χ^2 distribution.
- For each gene, compute the test statistic $F = \mathrm{MSA}/\mathrm{MSE_{Robust}}$, which follows Snedecor's F-distribution with degrees of freedom $\mathrm{df_A} = n - 1$ and $\mathrm{df_{E,Robust}} = 200 * (nk - n)$.
- Select genes for which $\mathrm{MSA} > \mathrm{Limit_{Robust},\alpha} = \mathrm{MSE_{Robust}} * F^{-1}(1-\alpha, \mathrm{df_A}, \mathrm{df_{E,Robust}})$, where α is the significance level.

Quantitative Real-Time RT–PCR

Of the total RNA preparation used for microarray analysis 500 ng was used for the first-strand cDNA synthesis (TaqMan reverse transcription reagent, N8080234 and a random hexamer primer) according to the manufacturer's instructions (Applied Biosystems, Foster City, USA). Semi-quantitative PCR was performed using the ABI PRISM 7900 Sequence detection system (Applied Biosystems). Primers and TaqMan probes were either designed using Primer Express software (Applied Biosystems) or ordered from Applied Biosystems through their Assays on Demand (AoD) service (Table 1). The PCR reactions were carried out according to the manufacturer's instructions. All results were normalized to GAPDH, which was not differentially regulated.

RESULTS AND DISCUSSION

The GEA model

Binning The GEA method more accurately characterizes MSE by calculating the robust mean SD of genes within a bin of 200 nearest neighbors. This is accomplished by sorting genes by mean ADI in ascending order and then placing them into bins. Various bin sizes were examined to determine how bin size would affect GEA model. Bin sizes of 25, 50, 100, 200 and 300 genes were examined as shown in Figure 2. In the present experiment, greater variability was observed with small bins, decreasing sharply and then leveling off at higher expression. It should be noted that there are a variety of normalization methods and alternative probe set expression level calculations (Bolstad *et al*., 2003; Irizarry *et al*., 2003) that would have various effects on the variability distribution seen in Figure 2. A full analysis of data pretreatment procedures is beyond the scope of the present work; however, under any of these procedures variability would still be expected to remain heterogeneous across the expression range and therefore take advantage of the binning and local calculation of error.

Importantly, the trend for the relationship between variability and expression level remained stable across the range of bin sizes, indicating that small changes in bin size do not have major effects. A bin size of 200 appeared to be optimal because it provides an accurate local estimate of MSE while simultaneously approaching a smoothed trend line. No further investigation for smoothing this non-continuous trend to continuity was deemed necessary. All subsequent GEA calculations are based on the MSE per bin of 200 genes, as described in Materials and methods section.

Classical ANOVA and GEA for two treatments and three replicates

Gene selections resulting from classical ANOVA and GEA are compared for the experiment with two treatments (control

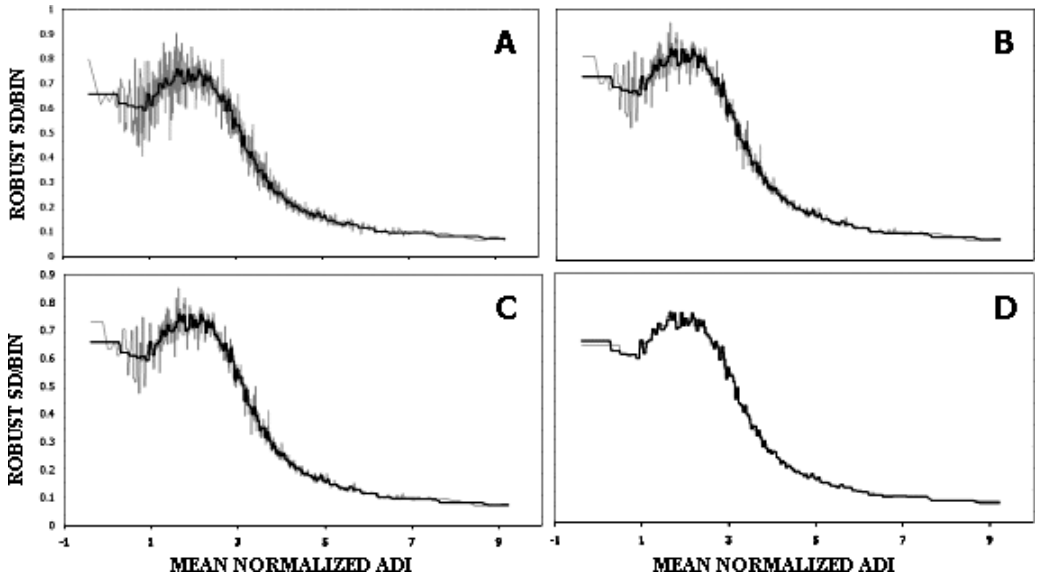

Fig. 2. Effects of varying bin size on the GEA model. Bin sizes of 200 are compared with bins of 25 (**A**), 50 (**B**), 100 (**C**) and 400 (**D**) genes. The x-axis represents the mean normalized ADI per bin. The y-axis represents the robust SD per bin, which is the square root of MSE_{Robust} (defined in the Materials and methods section).

versus IFN-γ treated skin cells) with three replicates (three Petri dishes). For an initial comparison the number of genes selected is unimportant, only in so much as it is equal between the two methodologies. In this manner, the alpha values generated can be compared instead of the actual gene content. Only for convenience an initial selection of 10% (or 4.5 k probe sets) among the 45k present on the combined U133A and B chips was used. The calculated significance levels were $\alpha = 0.0144$ for classical ANOVA and $\alpha = 0.0010$ for GEA. These significance levels show the higher confidence estimated via the GEA selection. This confidence is directly dependent of the higher degrees of freedom of the binned MSE in GEA. One obtains the following distribution among selected genes: 2790 genes are selected by both techniques; 1709 genes are selected by GEA only; and 1710 genes are selected by classical ANOVA only.

These differences can be explained by the fact that classical ANOVA computes the MSE individually, and is therefore influenced by random noise. Uncertainty in the estimate of MSE leads to both false negatives (high MSA and accidentally large MSE) as well as false positives (low MSA and accidentally very low MSE). Furthermore, GEA is based on the observation that the experimental variability (MSE) of gene expression is locally related to expression level, which implies that MSE is similar for neighboring genes. Therefore, the GEA method computes a robust estimate of the MSE of each gene by averaging the MSE of gene neighbors in each bin. This is an effort to establish a localized baseline of variability to compare with potential treatment effects, i.e. differential gene expression.

Classical ANOVA and GEA for two treatments and nine replicates

The initial analysis indicated that the between-petri variability was of the same magnitude as the within-petri variability (data not shown). It was therefore concluded that, for the purpose of demonstration, the data could be analyzed as though there was one sample consisting of nine replicates, instead of three samples consisting of three replicates. It is important to note that this analysis was only performed to show the effects of a larger sample size on gene selection with both GEA and classical ANOVA. The selection differences between classical ANOVA $\alpha = 0.0013$ and GEA $\alpha = 0.0004$ are striking. Under these conditions, 3239 genes are selected with both classical ANOVA and GEA, 1260 genes by GEA only and a different 1260 genes by classical ANOVA only.

These results demonstrate that by increasing the number of replicates, the differences between classical ANOVA and GEA are reduced because the individual MSE is estimated more accurately. Furthermore, increasing the number of replicates from 3 to 9 indicates that the classical ANOVA selection approaches that of the GEA selection; however, even with nine replicates it is still far from achieving an identical selection. It is estimated that the classical ANOVA would eventually completely converge with GEA at some large number of replicates. Furthermore, comparing those genes selected by either analysis, GEA appears to be selecting almost the same genes, whereas classical ANOVA does not, as shown in Table 2.

The comparison can also be made from a different perspective. By choosing a highly confident GEA P-value the question

Table 2. Effects of replication on ANOVA- and GEA-based gene selection

	Classical ANOVA	Perm. ANOVA	GEA
Genes selected for both three and nine replicates	3375	—	4128
Genes selected for three replicates only	1125	—	371
Genes selected for nine replicates only	1124	4500	371

Number of genes selected by classical ANOVA, permutational ANOVA and GEA as a function of the number of replicates used for computation. A target of 4500 genes was set for each method generating a P-value threshold of 0.0013, 0.0197 and 0.0004, respectively. Only values for nine replicates could be obtained for permutational ANOVA, as the minimum confidence (0.001) could not be achieved with less than 7 replicates.

Table 3. Selection confidence of ANOVA- and GEA-based selection methods for most significant expression outliers

Selection method	P-value level	No of. selected genes	No. of genes in common with GEA selected genes
GEA	1.00×10^{-30}	531	531
Perm. ANOVA	0.003	2418	527
Classical ANOVA	0.02	3358	523

can then be asked at what confidence level would the classical technique require to achieve full concordance. As shown in Table 3, a GEA P-value of 1×10^{-30} selects 531 genes. As demonstrated graphically in Figure 3, these genes clearly differentiate themselves from the underlying variability derived from comparison of replicate controls. Point coordinates x, y on the graph corresponds to the expression (in log scale) of a particular gene, as observed under the two conditions (IFN-γ and control treatments). Blue points correspond to paired expressions in control treatments and make up the 'variability cloud', green open squares correspond to paired expressions in control and treated conditions—most of them corresponding to unmodulated genes and thus overlapping with background blue points. The pear shape of the variability cloud shown in Figure 3 corroborates the inverse relationship between variability and expression demonstrated in Figure 2. Red crosses correspond to selected genes according to GEA, with very high significance ($P < 1 \times 10^{-30}$). One can visually see that the higher the expression level, the closer these points are to the contour of the variability cloud. However, a minimal distance separates these points from the contour. In other words, paired expressions of selected genes in control and treated conditions, according to GEA, are clearly different from paired expressions of genes in control conditions.

To achieve concordance (523 out of 531 genes), the classical ANOVA must relax to a P-value of 0.02 or greater. Orange squares correspond to those genes selected by classical ANOVA, which selects a total of 3358 genes, almost all of which overlap with the underlying variability cloud as seen in Figure 3a. It can be concluded from this analysis that GEA does in fact derive increased statistical power from the binned MSE.

Permutational ANOVA test and GEA

The permutational ANOVA test was picked for additional comparison because it is considered to be more robust and powerful than the classic ANOVA (Dudoit *et al*., 2002). Having a large number of replicates for each experiment allows us to build a reliable estimate of statistic distribution through permutations (in the present case 10 k permutations were used. The same perspective as was used with classical ANOVA was applied to the comparison between GEA and permutational ANOVA (Table 2 and Figure 3b.) In Table 2, it can be seen that a completely parallel comparison cannot be performed using permutational ANOVA, as the number of available replicates determines a minimum level of available P-values. A highly confident P-value of 0.05 or better is far out of reach for this method when only three replicates are available. A minimum of seven replicates would be required to meaningfully achieve a P-value of 0.05 or better. This 'non-comparison' is itself important as it indicates that GEA can be confidently used under conditions of low experimental replicates where permutational ANOVA could not.

A graphical analysis similar to that conducted between GEA and classical ANOVA is shown in Figure 3b. Using the same highly confident GEA P-value ($P < 1 \times 10^{-30}$) as before, a confidence level for permutational ANOVA was selected to achieve full concordance, as demonstrated in Table 3 ($P < 0.003$; 10k permutations). The orange squares in Figure 3b correspond to those genes selected by the permutation ANOVA test; however, it can also be seen that most GEA selections are co-selected by the permutation ANOVA. The permutation ANOVA returns 527 out of 531 genes selected by GEA; however, these 527 are among a total of 2418 genes, most of them borderline or within the underlying variability cloud. These results demonstrate additional confidence in GEA (co-selected by a second method with high confidence), but also a superior selectivity that is achieved compared to permutational ANOVA.

Another significant advantage of GEA over the permutational test is lower computational complexity. Performing a large number of permutations on a standard laboratory PC can take hours or even days, whereas the GEA methodology can be applied to a dataset and return results in a matter of minutes. Using the present study as an example, the GEA calculation took <5 min whereas the 10k permutational ANOVA took ~6 h to run on a 1.6 MHz Pentium 4 computing platform.

P-value adjustment for multiple tests

Performing multiple, yet simultaneous, significance tests raises the pointed question of P-value adjustment. A P-value

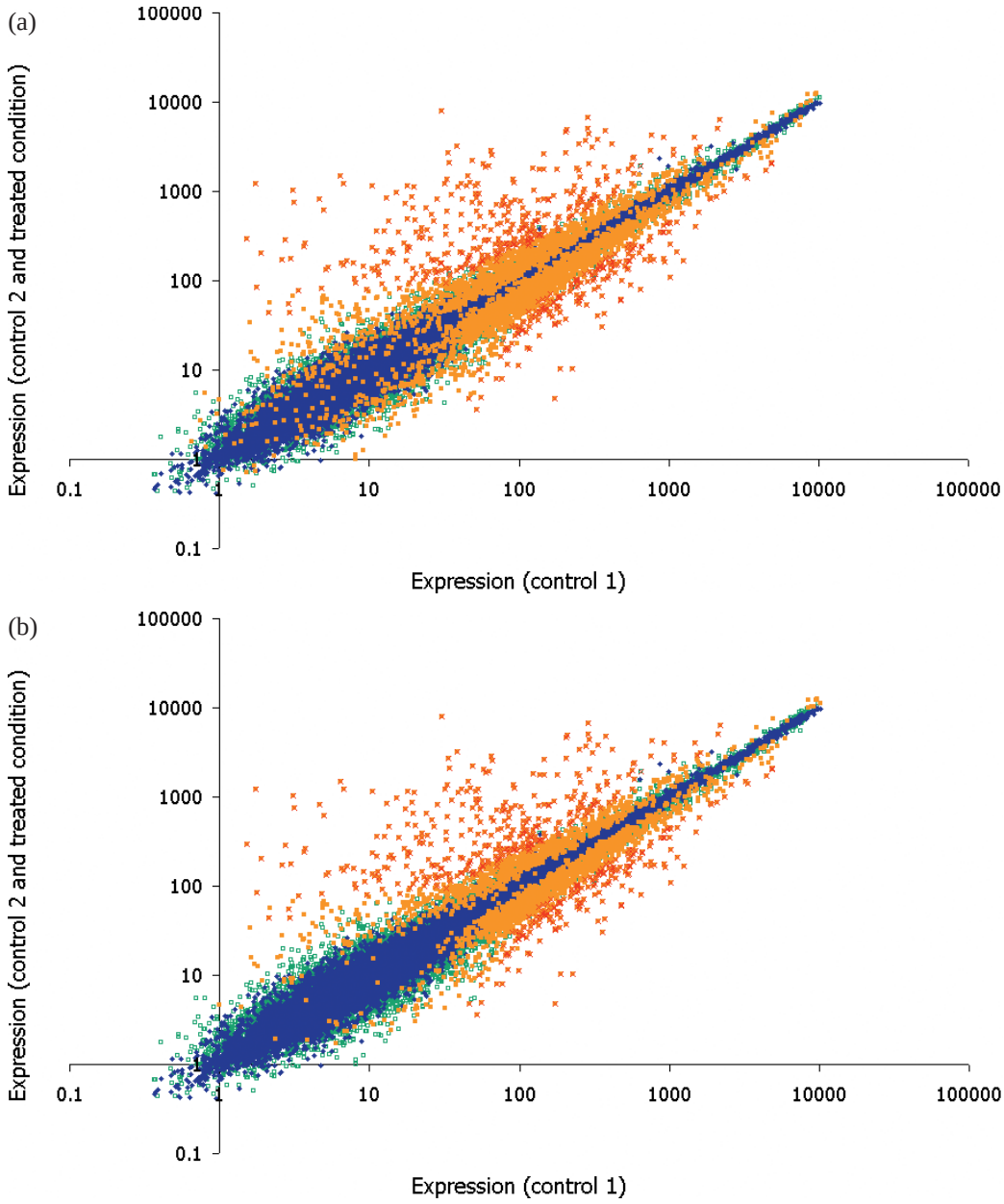

Fig. 3. (a) Graphical validation of detected genes (classical ANOVA versus GEA). Blue squares define the 'variability cloud' of the dataset. Green open squares represent those genes under treated condition. Red crosses indicate those genes detected with a GEA P-value $< 1 \times 10^{-30}$. Orange squares indicate those genes detected with a classical ANOVA P-value < 0.02. (b) Graphical validation of detected genes (permutational ANOVA versus GEA). Blue squares define the 'variability cloud' of the dataset. Green open squares represent those genes under treated condition. Red crosses indicate those genes detected with a GEA P-value of $<1 \times 10^{-30}$. Orange squares indicate those genes detected with a permutational ANOVA P-value <0.003.

can be significantly low simply because of the noise present in the system. By increasing the number of tests there is an increased chance that this will arise. There are multiple procedures that allow adjustment of P-values with respect to the multiple testing problem (Bonferroni method, Holm's procedure, etc.). However useful these may be for estimating the true P-value, they are not of great importance for the comparison of tests, e.g. the GEA versus ANOVA comparisons made here. In addition, multiple testing corrections do not solve for the lack of power in a situation with a low number of replicates.

An adjustment of P-values is still an approximation, and can sometimes aggravate the situation. As an example, the Westfall and Young adjustment procedure (Dudoit *et al.*, 2002;

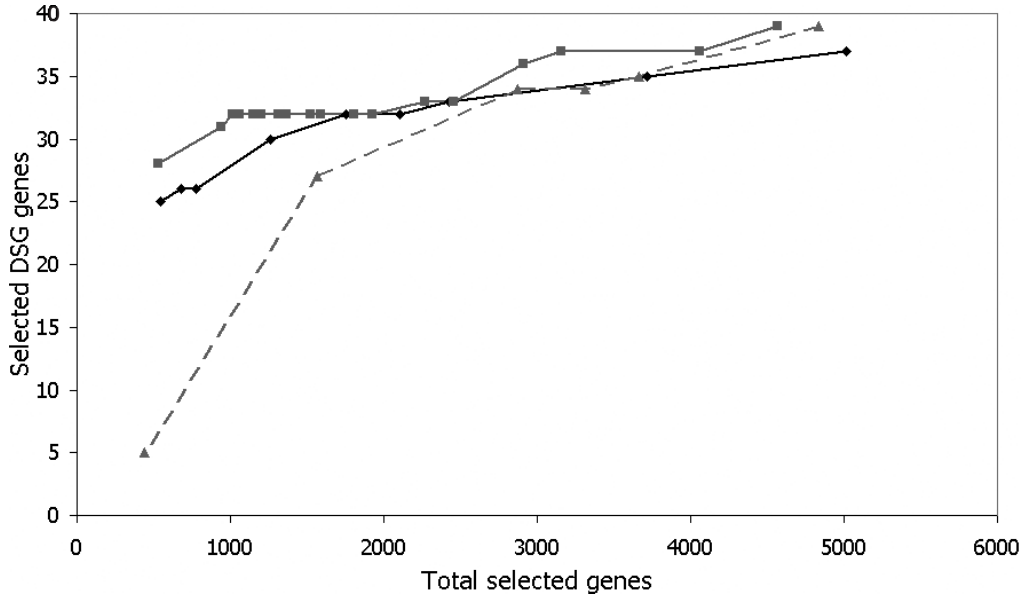

Fig. 4. Detection of previously identified IFN-γ stimulated genes. Number of IFN-γ stimulated genes selected by a particular method versus total number of genes selected by the method at particular thresholds. Squares correspond to GEA ($P < 1 \times 10^{-30}$), Diamond shapes represent results of classic ANOVA approach ($P < 0.0004$), and Triangles show performance of permutational ANOVA ($P < 0.0012$). 'DSG' refers to Der *et al*. IFN-γ stimulated genes.

Westfall and Young, 1993) was applied to the permutational ANOVA results here, producing all P-values equal to 1. This was not helpful and so was not included in the comparison of selection methods. Therefore, analysts of microarray results are encouraged to consider the multiple testing problem, in addition, but not as a substitute, to the differential gene expression methodologies such as those presented here.

Confirmation of results

A thorough literature review revealed a high concordance between known IFN-γ effects and the panel of genes selected by GEA. The classical ANOVA, permutational ANOVA and GEA were then compared to determine their respective abilities to detect ISGs. A previous study lists ISGs that were known from the literature along side those discovered through microarray analysis (Der *et al*., 1998). Using various selection criteria, a total of 49 IFN-γ stimulated (increased expression) genes were previously detected in a study by Der *et al*. As there are significant differences in the protocols of the two experiments, 100% concordance was considered unlikely; however, a significant number of these genes were expected to appear in the current study. The ability of each technique to detect this set of 49 genes is illustrated in Figure 4, in which the number of the Der *et al*. IFN-γ stimulated genes (DSGs) selected in the present dataset by a particular method is plotted against the total number of genes selected. In addition, the multiple P-value thresholds for each method is provided to demonstrate the overall performance of the various methods. It can be seen that GEA is almost uniformly superior to both forms of ANOVA in detecting these genes. The GEA test detected 28 DSGs with a high degree of confidence ($P < 1 \times 10^{-30}$ selecting DSGs in a total pool of 531 differentially expressed genes), whereas a classical ANOVA was only able to identify 25 DSGs within total selection of approximately the same size ($P < 0.0004$ selecting DSGs in a total pool of 548 differentially expressed genes; 9 replicates). The permutational ANOVA performed less well than both the classical ANOVA and GEA, as only 8 DSGs in a pool of 540 selected genes were identified with this method of analysis ($P < 0.0012$).

Importantly, the study of Der *et al*. also lists those 30 genes, previously identified in the literature using alternative experimental methodolgies, to be stimulated by IFN-γ (Der *et al*., 1998). All 30 ISGs were selected by GEA in the present experiment (100% detection at $P < 1 \times 10^{-30}$ selecting 30 ISGs in a pool of 531 differentially expressed genes). In contrast, only 25 ISGs were selected by classical ANOVA using all nine replicates (83% detection at $P < 0.0004$ selecting ISGs in a total pool of 548 differentially expressed genes). From another perspective, over 2000 genes would need to be selected in order to achieve 100% detection of ISGs by classical ANOVA. In comparison, the permutational ANOVA test identified just seven of the known ISGs (23% detection at $P < 0.0012$ selecting ISGs in a total pool of 540 differentially expressed genes). Alternatively, 2400 genes would have to be selected by the permutational ANOVA to achieve a 100% detection rate.

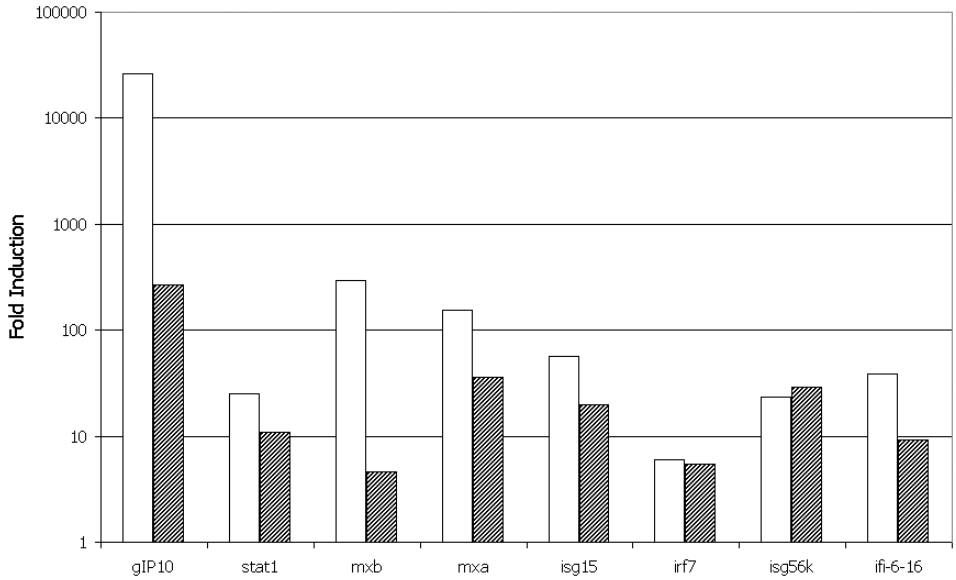

Fig. 5. Comparison between real-time RT–PCR and Affymetrix results. Fold induction values of eight ISGs are plotted (IFN-γ versus control), with RT–PCR data represented by empty bars and microarray data by stripped bars. RT–PCR results are normalized to the housekeeping gene GAPDH. All genes validated by RT–PCR were selected by GEA with P-values $<1 \times 10^{-30}$. The abbreviations listed above correspond to the following gene information. They are listed here according to abbreviation, Affymetrix U133A probe set ID, gene title and then accession number. Gene List: STAT1, 200887_s_at, signal transducer and activator of transcription 1, 91 kDa, 'NM_007315, NM_139266'; Mxa, 202086_at, myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse), NM_002462; isg56k, 203153_at, interferon-induced protein with tetratricopeptide repeats 1, NM_001548; ifi-6-16, 204415_at, interferon, alpha-inducible protein (clone IFI-6-16), 'NM_002038 NM_022872, NM_022873'; gIP10, 204533_at, chemokine (C–X–C motif) ligand 10, NM_001565; mxb, 204994_at, myxovirus (influenza virus) resistance 2 (mouse), NM_002463; isg15, 205483_s_at, interferon, alpha-inducible protein (clone IFI-15K), NM_005101; irf7, 208436_s_at, interferon regulatory factor 7, NM_001572.

Eight of the many known ISGs (gIP10, Stat1, MxA, MxB, IRF7, ISG56K) and the housekeeping gene GAPDH were analyzed further by real-time RT–PCR as an initial *in vitro* validation. Each gene was examined in duplicate in each of the three control conditions and the three IFN-γ treated conditions. Following normalization with GAPDH, the fold induction values were determined and these values were compared with the observed microarray results (Figure 5). As the sensitivity for these two techniques is quite different, it was not expected that identical fold-changes would be seen for real-time RT–PCR and Affymetrix Gene Chips (Holland, 2002). Therefore, concordance was determined based on direction, i.e. an increase in gene expression measured with Gene Chips was also seen by real-time RT–PCR. As shown in Figure 5, all genes showed an increased expression when measured by either technique. Therefore, this initial validation study confirmed that the experiment was functioning predictably (as previously determined in the literature) and that the GEA selection results had biological significance. At the same time, using the previously described set for classical ANOVA ($P < 0.0004$ selecting a total pool of 548 differentially expressed genes) only six of the eight would have been selected; whereas all genes measured by real-time RT–PCR were selected by GEA ($P < 1 \times 10^{-30}$ selecting a total pool of 531 differentially expressed genes). Lastly, within a pool of similar size permutational ANOVA selected only three of the eight real-time RT–PCR validated genes ($P < 0.0012$ selecting a total pool of 540 differentially expressed genes).

In summary, these confirmatory results are highly significant. They indicate good sensitivity for true positives by GEA relative to the other techniques being compared here. The GEA method has been found to place high confidence in biologically significant genes known to be regulated by the experimental induction (IFN-γ) being studied. Lastly, the biologically significant genes have also been confirmed through a complementary technique like real-time RT–PCR.

CONCLUSION

The GEA model for differential gene expression selection was introduced to respond to the small sample size problem in microarray analysis by replacing the denominator of the F-ratio (the MSE normally estimated per gene) with the MSE calculated per bin, which is directly related to the mean absolute expression level within each bin. The GEA method was demonstrated to confidently determine differentially expressed genes under conditions of both low and high replication. These results were compared to a classical

ANOVA and an advanced permutational ANOVA test. The major differences in gene selection between ANOVA and GEA are that an ANOVA (classic or permutational) analyzes each gene separately using a single factor ANOVA, whereas GEA is far more powerful because the MSE is estimated locally based on information from nearest gene neighbors. This has recently been demonstrated by Mutch *et al*. (2003) for which the GEA model more accurately identified significant differences in gene expression than the classical ANOVA when compared with real-time RT–PCR data. As the bin size used in this analysis was 200, this implies that the number of degrees of freedom is increased by 200-fold. The IFN-γ experiment used here to develop GEA provided a number of literature sources to confirm or refute our findings. GEA was demonstrated to be uniformly superior for both high and low replicates in detecting known ISGs and had excellent concordance with previous microarray studies looking at IFN-γ stimulation. Lastly, but of equal importance GEA is a relatively simple, intuitive and computationally efficient method, which allows it to be easily implemented under standard computing environments.

REFERENCES

Baggerly,K.A., Coombes,K.R., Hess,K.R., Stivers,D.N., Abruzzo,L.V. and Zhang,W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8** 639–659.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.*, **8**, 1821–1832.

Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Der,S.D., Zhou,A., Williams,B.R. and Silverman,R.H. (1998) Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proc. Natl Acad. Sci., USA*, **95**, 15623–15628.

Draghici,S., Kulaeva,O., Hoff,B., Petrov,A., Shams,S. and Tainsky,M.A. (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. *Bioinformatics*, **19**, 1348–1359.

Dudoit,S., Yang,Y.H., Speed,T.P. and Callow,M.J. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105–S110.

Freedberg,I.M., Tomic-Canic,M., Komine,M. and Blumenberg,M. (2001) Keratins and the keratinocyte activation cycle. *J. Invest. Dermatol.*, **116**, 633–640.

Ghosh,D. (2002) Resampling methods for variance estimation of singular value decomposition analyses from microarray experiments. *Funct. Integr. Genomics*, **2**, 92–97.

Hess,K.R., Zhang,W., Baggerly,K.A., Stivers,D.N. and Coombes,K.R. (2001) Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol.*, **19**, 463–468.

Holland,M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.*, **277**, 14363–14366.

Huang,X. and Pan,W. (2002) Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Funct. Integr. Genomics*, **2**, 126–133.

Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Jain,N., Thatte,J., Braciale,T., Ley,K., O'Connell,M. and Lee,J.K. (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, **19**, 1945–1951.

Kamb,A. and Ramaswami,M. (2001) A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnol.*, **1**, 8.

Kepler,T.B., Crosby,L. and Morgan,K.T. (2002) Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.*, **3**, RESEARCH0037.

Kothapalli,R., Yoder,S.J., Mane,S. and Loughran,T.P.,Jr. (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.

Lin,H., Stoehr,J.P., Nadler,S.T., Schueler,K.M., Yandell,B.S., and Attie,A.D. (2003) Adaptive gene picking with microarray data: detecting important low abundance signals. In *The Analysis of Gene Expression Data: Methods and Software*. Springer-Verlag.

Mutch,D.M., Anderle,P., Fiaux,M., Mansourian,R., Vidal,K., Wahli,W., Williamson,G. and Roberts,M.A. (2003) Regional variations in Abc transporter expression along the mouse intestinal tract. *Physiol. Genomics*, **16**, 16.

Mutch,D.M., Berger,A., Mansourian,R., Rytz,A. and Roberts,M.A. (2002) The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, **3**, 17.

Nadon,R., Shi,P., Skandalis,A., Woody,E., Hubschle,H., Susko,E., Ramm,P. and Rghei,N. (2001) Statistical interference methods for gene expression arrays. *Proceedings of SPIE, Microarrays, Optical Technologies and Informatics BIOS 2001*, pp. 46–55.

Ramakrishnan,R., Dorris,D., Lublinsky,A., Nguyen,A., Domanus,M., Prokhorova,A., Gieser,L., Touma,E., Lockner,R., Tata,M. *et al*. (2002) An assessment of Motorola CodeLink

microarray performance for gene expression profiling applications. *Nucleic Acids Res.*, **30**, e30.

Ramana,C.V., Gil,M.P., Schreiber,R.D. and Stark,G.R. (2002) Stat1-dependent and – independent pathways in IFN-gamma-dependent signaling. *Trends Immunol.*, **23**, 96–101.

Sasik,R., Calvo,E. and Corbeil,J. (2002) Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics*, **18**, 1633–1640.

Schindler,C. and Darnell,J.E.,Jr. (1995) Transcriptional responses to polypeptide ligands: the JAK–STAT pathway. *Annu. Rev. Biochem.*, **64**, 621–651.

Sebok,B., Bonnekoh,B., Vetter,R., Schneider,I., Gollnick,H. and Mahrle,G. (1998) The antipsoriatic dimethyl-fumarate suppresses interferon-gamma-induced ICAM-1 and HLA-DR expression on hyperproliferative keratinocytes. Quantification by a culture plate-directed APAAP-ELISA technique. *Eur. J. Dermatol.*, **8**, 29–32.

Teunissen,M.B., Koomen,C.W., de Waal Malefyt,R., Wierenga,E.A., and Bos,J.D. (1998) Interleukin-17 and interferon-gamma synergize in the enhancement of proinflammatory cytokine production by human keratinocytes. *J. Invest. Dermatol.*, **111**, 645–649.

Thomas,J.G., Olson,J.M., Tapscott,S.J. and Zhao,L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Troyanskaya,O.G., Garber,M.E., Brown,P.O., Botstein,D. and Altman,R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461.

Wei,L., Debets,R., Hegmans,J.J., Benner,R. and Prens,E.P. (1999) IL-1 beta and IFN-gamma induce the regenerative epidermal phenotype of psoriasis in the transwell skin organ culture system. IFN-gamma up-regulates the expression of keratin 17 and keratinocyte transglutaminase via endogenous IL-1 production. *J. Pathol.*, **187**, 358–364.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing*. John Wiley and son, Inc.

Woolf,P.J. and Wang,Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol. Genomics*, **3**, 9–15.