Research

Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis

Reinhard Hoffmann*, Thomas Seidl† and Martin Dugas‡

Addresses: *Department of Bacteriology, Max von Pettenkofer Institut, Pettenkoferstrasse 9a, 80336 Munich, Germany. †Section of Gene Function and Regulation, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK. ‡Department of Medical Informatics, Biometrics and Epidemiology, University of Munich, Marchioninistrasse 15, 81377 Munich, Germany.

Correspondence: Reinhard Hoffmann. E-mail: r_hoffmann@m3401.mpk.med.uni-muenchen.de

Published: 14 June 2002

Genome ***Biology*** 2002, **3(7)**:research0033.1–0033.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/7/research/0033

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 7 February 2002
Revised: 4 April 2002
Accepted: 24 April 2002

Abstract

Background: Oligonucleotide microarrays measure the relative transcript abundance of thousands of mRNAs in parallel. A large number of procedures for normalization and detection of differentially expressed genes have been proposed. However, the relative impact of these methods on the detection of differentially expressed genes remains to be determined.

Results: We have employed four different normalization methods and all possible combinations with three different statistical algorithms for detection of differentially expressed genes on a prototype dataset. The number of genes detected as differentially expressed differs by a factor of about three. Analysis of lists of genes detected as differentially expressed, and rank correlation coefficients for probability of differential expression shows that a high concordance between different methods can only be achieved by using the same normalization procedure.

Conclusions: Normalization has a profound influence of detection of differentially expressed genes. This influence is higher than that of three subsequent statistical analysis procedures examined. Algorithms incorporating more array-derived information than gene-expression values alone are urgently needed.

Background

cDNA or oligonucleotide microarrays have made possible the measurement of thousands of mRNA levels in parallel, enabling researchers for the first time to generate comprehensive cellular gene-expression profiles. Among several competing techniques, photolithographically synthesized high-density oligonucleotides are widely used. Current chip layouts allow for the parallel measurement of >12,000 gene-expression levels on a single array. In this approach, every gene is represented by a set of oligonucleotides perfectly matching the target sequence (PM oligo) and by a corresponding set with a 1 base-pair (bp) mismatch in a central position (MM oligo). The latter serves as an internal control for hybridization specificity. Relative transcript abundance is reported as the so-called 'average difference' value, that is the average of all PM-MM differences across the gene-specific set of probes [1,2]. An alternative approach fits a linear model onto the differences between PM and MM hybridization intensities and takes a model-based expression value as a measure of transcript abundance [3,4].

The technique is standardized in such a way that generation of gene-expression data is straightforward and quite easy to do. Analysis of processed fluorescence-intensity data, in contrast, is not. Analysis of a typical microarray experiment involves the following steps: pre-scaling of the fluorescence

intensity across the different arrays belonging to one experiment to correct for differences in probe labeling, probe concentration, hybridization efficiency, and potentially other factors (in the context of microarray analysis, this process is generally termed normalization); detection of differentially expressed genes; in the case of experimental setups comparing more than two conditions, a clustering step to group together genes with similar expression patterns; and higher-level analysis, for example by combining functional annotations of genes having predefined interesting expression patterns with previous knowledge about the experimental system under investigation.

Most frequently, high-density oligonucleotide data are normalized by a simple 'global scaling' procedure. This involves multiplication of every gene-expression value with a constant factor so that the mean intensities of the arrays to be compared are identical. A conceptually related approach involves fitting a linear regression model on the data and scaling the fluorescence intensities so that the resulting regression model has a slope of 1 and a y-intercept of 0 [5]. This approach suffers from two significant drawbacks: first, it relies on the implicit assumption that the total mRNA content of different cell types compared is the same. This is not always the case, especially if cell types of different size and/or cell-cycle status are compared. Control of this effect is attempted by loading identical amounts of cRNA onto the chips. However, it has been shown that the mean expression level on any array can be subject to significant variation across arrays [6]. Second, the normalization is linear and cannot account for nonlinearity in the underlying data. Previous studies [7,8] show that simple linear regression models incompletely fit the data and that two or more linear models with different slope for different ranges of fluorescence-intensity values result in a better fit.

Two conceptually related solutions to these problems have been proposed. They assume that an 'invariant set' exists, containing genes that do not change significantly between two experimental conditions. First, all fluorescence values on the arrays are ranked according to intensity. Then, items with similar ranks between two arrays are identified and considered unchanged. These items are used for nonlinear normalization. This procedure can be performed either on the feature level (taking raw fluorescence values as input) [3,4] or on the probe set level, taking average-difference values as input [8].

A similar multitude of strategies exist to detect differentially expressed genes. The easiest approach is to define genes as differentially expressed that change more than an arbitrarily chosen threshold. More sophisticated analyses additionally apply statistical tests such as Student's t-test for comparisons between two experimental conditions. However, this and other parametric tests rely on certain assumptions, namely that the underlying data are normally distributed with equal variances across experimental conditions [9]. These assumptions must not necessarily be met, and analysis of our own (T.S. and R.H., unpublished observations) and other [10] datasets show that they are often not fulfilled. Non-parametric tests such as the Mann-Whitney test, in contrast, do not rely on such strong assumptions, but a larger number of replicate experiments is desirable.

A particular problem is the analysis of 'multiclass experiments' containing more than two experimental conditions, such as cellular developmental stages. Many researchers carry out pairwise comparisons of all possible pairs of combinations, resulting in a list of genes that are detected as differentially expressed at least once. This leads to increased type-I error rates, with the final data set having a type-I error rate up to $1-(1-\alpha)^n$, where α is the type-I error rate of individual pairwise comparisons and n is the number of pairwise comparisons [9]. Five pairwise comparisons at the 95% confidence level thus result in a confidence level for the resulting dataset of 77%. Classical statistics offer ANOVA algorithms for such problems. Here, differential gene expression is detected by comparing variances within experimental conditions to variances across experimental conditions [9]. Both parametric (F) and nonparametric (H or Kruskal-Wallis) tests exist, with the associated problems described above.

Recently, an alternative procedure for detection of differentially expressed genes, called significance analysis of microarrays (SAM), has been described [11]. Here, a relative difference in gene expression is computed, incorporating means and standard deviations across experimental conditions. Next, the dataset is permuted several times, and the relative difference is computed again, on the basis of the permuted datasets. For the majority of genes, these two values are approximately equal. For some genes, however, the difference between the two scores exceeds a certain threshold parameter. These genes are called differentially expressed. A false-discovery rate [12] can be computed on the basis of how many genes are called in the permuted datasets with the given threshold.

Obviously, there are a large number of analysis options for gene-expression data. The influence of normalization and statistical analysis on the detection of differentially expressed genes has not been investigated to date. In this study, we carry out a thorough comparison of different normalization and statistical procedures to define the key components for detection of differentially expressed genes in a multiclass experiment.

Results

The aim of the present study was to evaluate different normalization and statistical analysis methods for their influence on detection of differentially expressed genes. We focused on a typical multiclass experiment. The dataset used

comprises high-density oligonucleotide array-derived gene-expression data of five consecutive cellular populations of an ordered cellular differentiation pathway. The biology-oriented analysis and interpretation of the dataset has been described elsewhere [13].

Normalization

Figure 1 shows signal intensity scatterplots of one randomly chosen array set, consisting of two arrays with different gene content. Not normalized values are on the *x*-axis, and the normalized counterparts on the *y*-axis. In all scatterplots, the subA and subB arrays can easily be distinguished by the different slopes of the respective data points. This underlines the necessity for separate normalization of different subarrays of one set.

The two normalization methods based on invariant features produce a significant amount of scatter compared to the not normalized data (Figure 1a,b), especially the model-based expression values as compared to the traditional average difference values. However, the model-based expression values calculated after invariant feature normalization (*y*-axis in Figure 1b) differ by two factors from the not normalized average difference values (*x*-axis in Figure 1b). The scatter in Figures 1a and b reflects data processing on the fluorescence level with recalculation of gene-expression metrics after normalization, in contrast to the other methods. Comparing average difference and model-based expression values derived after invariant feature normalization (Figure 1c), it becomes evident that model-based expression values tend to be higher than average difference values in the low-signal area of the plot. Thus, low-abundance genes give higher signals when model-based expression values are used. This might reflect either a greater sensitivity of the model-based approach or an overestimation of the expression level.

The invariant set normalization method results in very similar slopes for the subA and subB arrays, respectively (Figure 1d). Since pre-computed average difference values are used, the normalization does not result in recalculation of the expression-level values from fluorescence data, resulting in less scatter than in Figure 1a,b. The global scaling method, as expected, produces two 'straight lines' of data points representing different normalization factors for the two array types (Figure 1e).

To explore further the differential impact of the normalization procedures, we compared the normalization curves generated by the two nonlinear (invariant feature and invariant set) normalization schemes (Figure 2). In Figure 2a, all the approximately 400,000 feature intensities from two different arrays are plotted against each other; the invariant features are displayed as red circles. The position of the invariant features far off and below the diagonal (blue line) indicates substantial need for normalization. Figure 2b shows a normalization curve for the same two non-normalized arrays obtained by the invariant set method. The invariant data points are indicated by red circles. In contrast to Figure 2a, the invariant sets are located very closely to or directly on the diagonal (blue dots), indicating that normalization by the invariant set method will affect expression level values mildly. This illustrates that different normalization methods will have a profound effect on the expression-level values, even if normalization has been carried out to the same baseline array.

Statistical evaluation and detection of differentially expressed genes

The four normalized datasets were subjected to three popular methods for identifying differentially expressed genes, namely parametric and nonparametric ANOVA models and the permutation-based SAM procedure. This resulted in 12 datasets containing data about probability of differential gene expression.

At the 99% confidence level - a median false discovery rate (FDR) of 1% in SAM, respectively - large sets of genes are detected by each of the combinations of normalization and statistical analysis methods. However, the number of genes detected differs dramatically (Figure 3a). The combination of linear global scaling with parametric ANOVA yields 1,526 differentially expressed genes. In contrast, the combination of invariant feature normalization with calculation of average difference values and SAM results in only 608 genes with a median FDR of 1% or lower. Looking at the average numbers of differentially expressed genes, it is highest with the parametric ANOVA model and smallest with the SAM procedure across different normalization procedures. Similarly, linear global scaling yields the largest, and invariant feature normalization with calculation of average difference values smallest, set of differentially expressed genes across the different statistical evaluations performed (Figure 3a).

However, the confidence level alone is seldom used for detection of differentially expressed genes. Usually, a fold-change criterion as well as an absolute difference criterion is added. Figure 3b shows what percentage of the genes shown in Figure 3a also pass additional criteria (twofold change and absolute difference of at least 100 units). These additional criteria lead to a reduction of the number of detected genes in all datasets, but to a markedly different extent. On average, 78, 82 and 88% of the genes detected using only the 99% confidence criterion are still detected when applying the additional criteria in the parametric ANOVA, nonparametric ANOVA, and SAM datasets, respectively (Figure 3b). In contrast, this holds true for only 63% of the genes normalized with the invariant probe set method. Genes detected by the combination with parametric ANOVA are most significantly affected, with only 54% fulfilling the additional criteria.

This effect reflects systematic differences in the datasets. Table 1 shows fold-change and difference statistics for different

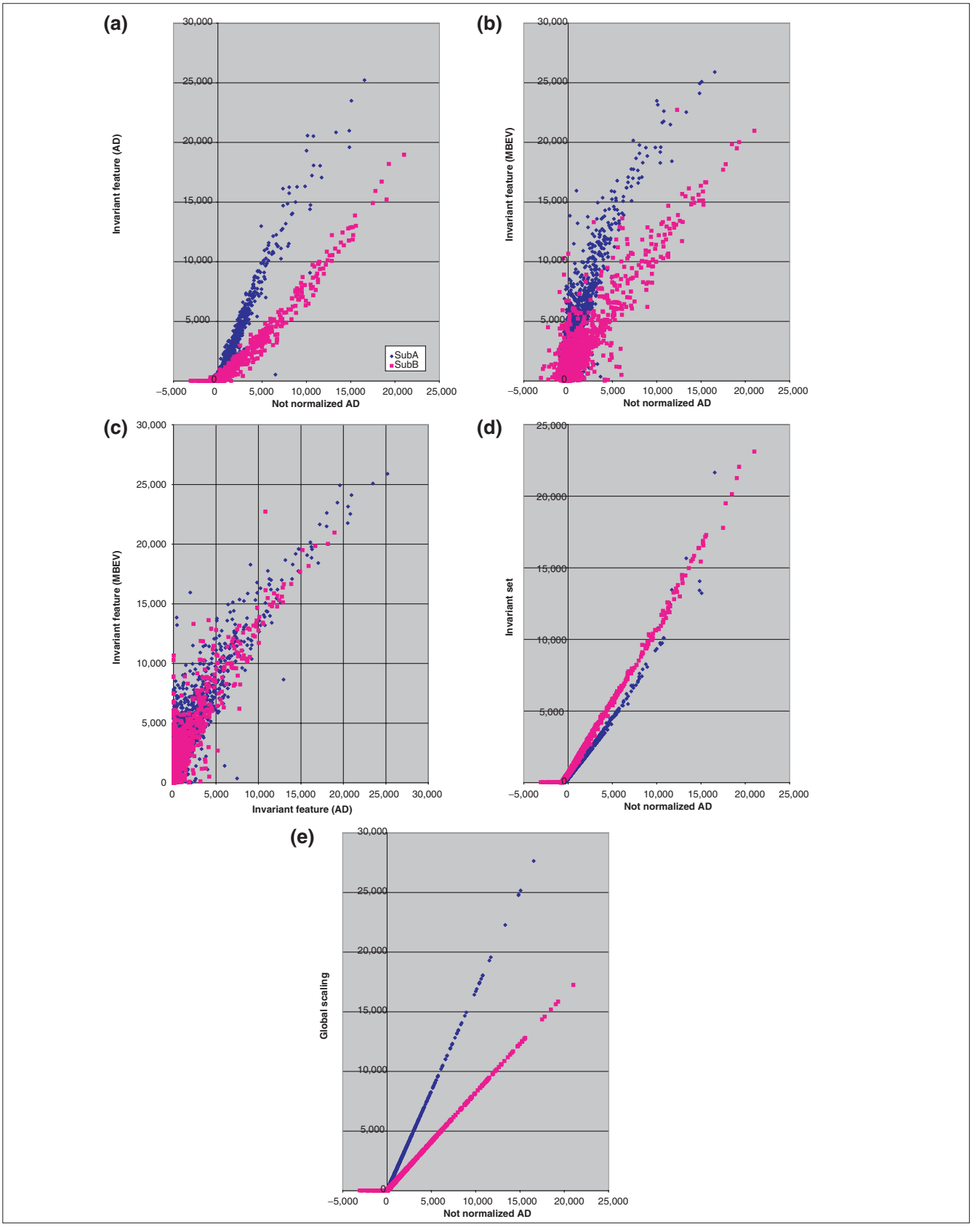

Figure 1 *(see the legend on the following page)*

subsets of the gene-expression datasets. Considering all genes on the arrays (Table 1a), the dataset normalized by the invariant probe set method contains the smallest fold changes with the smallest differences between mean and median fold change. Notably, the highest fold change achieved by any gene is 72-fold in this dataset, as opposed to several hundred fold in the other datasets. Similarly, the mean fold change of genes detected with 99% confidence by any of the three methods is three- to fourfold in the invariant probe set normalized dataset, as opposed to around tenfold or higher in datasets employing the other normalization procedures. Also, this subset of genes shows the smallest difference between mean and median (Table 1b). This indicates that the fold-change values are less skewed toward small values in the dataset normalized by the invariant probe set method. Analyzing the absolute differences between the maximum and minimum value across the five cellular stages examined, these effects are still present, but markedly reduced (Table 1c,d).

We next asked how well different combinations of normalization and statistical-analysis strategies agree in detecting differentially expressed genes. Figure 4 shows how many genes are detected as differentially expressed by how many data-analysis combinations. All three criteria for filtering the data (confidence, fold change, absolute difference) were applied as above. Interestingly, a large proportion (382 genes) is detected by all 12 combinations. An almost equally large number is detected by one particular combination only. Among 201 genes detected by two combinations, these two combinations involve the same normalization procedure in 155 cases, in contrast to 40 cases that involve the same statistical algorithm. Similarly, among 208 genes detected by three combinations, the same normalization is involved in all three detecting combinations in 145 cases (whereas the same statistical algorithm is involved in 11 cases only). This might indicate that the normalization procedure has a profound influence on which genes are detected as differentially expressed in a subsequent statistical analysis step. The sharp drop between three- and four-method combinations is most likely due to the fact that three statistical algorithms have been employed. The sharp rise between 11 and 12 method combinations is most likely due to the fact that genes detected by 11 combinations form a very stable subset already, and thus are more likely to be detected by the twelfth method as well.

We thus investigated how many genes are detected simultaneously when comparing any two different combinations of methods. Table 2 shows the percentage of genes identified as differentially expressed by the method combination defined by the column header out of the combination defined by the row designation. Again, all three filter criteria were applied. Values range from 41% (of the genes identified by global scaling and parametric ANOVA which are also found by application of invariant feature normalization with calculation of average difference values and SAM) to 100%. Strikingly, the genes identified by SAM are always a subset of the genes identified by parametric ANOVA, if the same normalization procedure has been employed (figures of 100% in Table 2). A high proportion (>93%) of the genes identified by nonparametric ANOVA are also identified by parametric ANOVA, and a similar high proportion of the genes identified by SAM are also identified by nonparametric ANOVA using the same normalization (Table 2). Thus, a type of hierarchy evolves: genes detected by SAM generally are a subset of genes detected by nonparametric ANOVA, and genes detected by nonparametric ANOVA generally are a subset of genes detected by parametric ANOVA. Interestingly, the genes detected after invariant probe set normalization generally are a subset of the genes detected after global scaling, with the hierarchy as described above (values marked by asterisks in Table 2).

These results point to a pronounced role for the normalization process in the detection of differentially expressed genes. We sought to determine which one of the two steps (normalization versus statistical analysis) has a greater influence on the detection of differentially expressed genes. We thus calculated rank correlations for probabilities of differential expression for all genes represented on the arrays between pairs of all 12 combinations of normalization and statistical analysis procedures (Table 3). Strikingly, a correlation of >0.9 can only be achieved between two statistical analysis methods if the same normalization has been carried out. Rank correlations between probabilities for differential expression are markedly lower (as low as 0.4 to 0.5 in some instances) if two different normalization procedures are compared, irrespective whether the same statistical analysis method has been used. This indicates that the normalization procedure employed has a tremendous effect on the subsequent detection of differentially expressed genes. The importance of the normalization step has not been properly regarded in the past.

Discussion

The present study examined the influence of normalization and statistical analysis on detection of differentially

Figure 1 *(see the figure on the previous page)*
Pre- and post-normalization signal intensity scatterplots. The *x*-axis in all panels except (c) represents the not normalized average difference values (AD) derived from the Affymetrix GeneChip software after scanning. **(a)** *y*-axis is invariant-feature normalization with calculation of AD values. **(b)** *y*-axis is invariant-feature normalization with model-based expression values (MBEV). **(c)** *x*-axis is invariant-feature normalization with calculation of AD values; *y*-axis is invariant-feature normalization with MBEV. **(d)** *y*-axis is invariant probe set normalization. **(e)** *y*-axis is global scaling. Blue dots, subA-array; pink dots, subB-array.

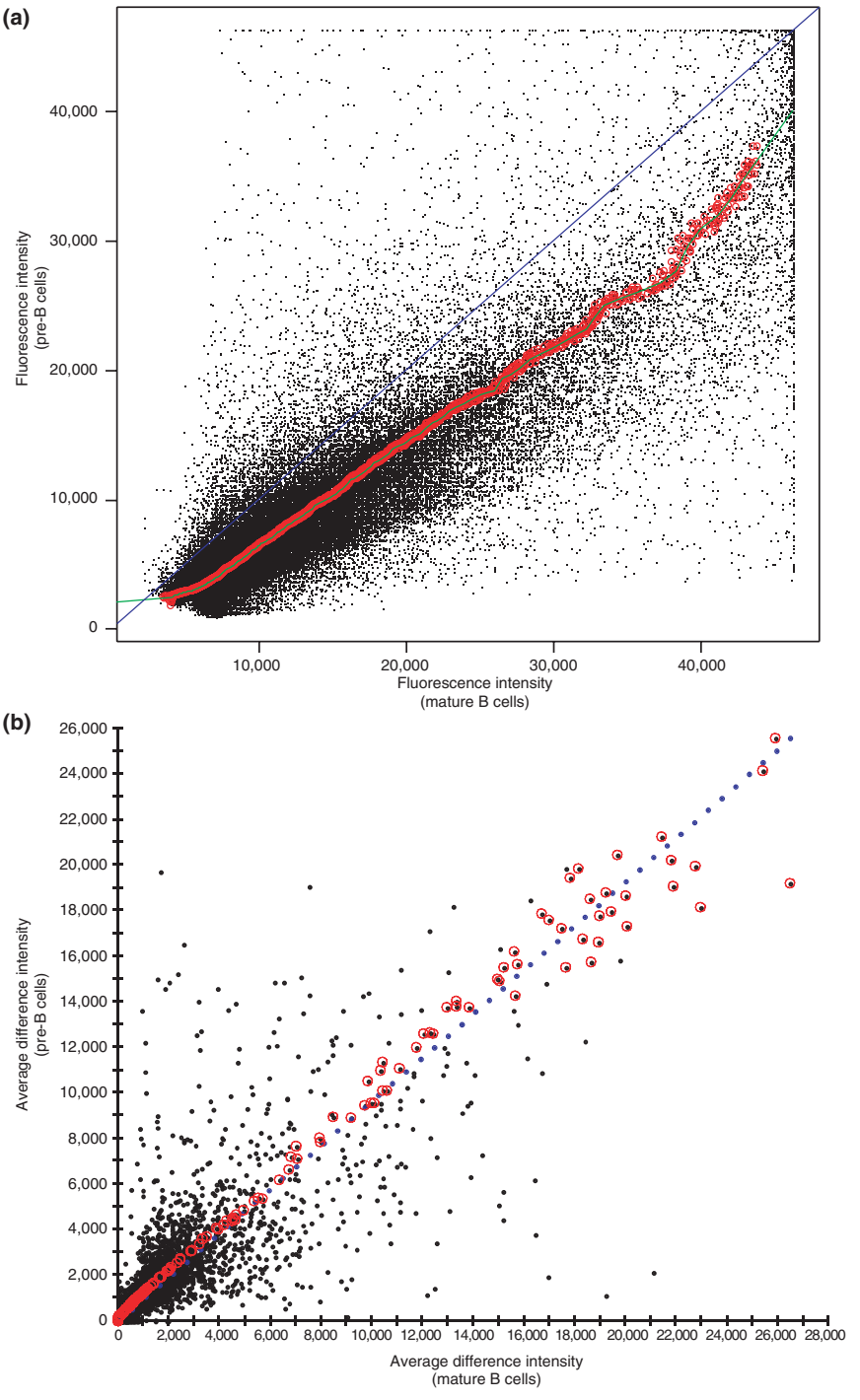

Figure 2
Invariant sets and normalization curves generated by nonlinear normalization methods using identical randomly chosen arrays. **(a)** Invariant-feature method. *x*-axis, Mature B cells, replicate 3, not normalized; *y*-axis, baseline: pre-BI cells, replicate 3. Black dots, feature intensities; red circles, invariant set of features (connected by the green line). The blue line forms the diagonal with slope of 1. **(b)** Invariant set method. Cells as in (a). Black dots, probe set average difference values; red circles, invariant set. Blue dots form the diagonal with slope of 1. Axes are labeled with average difference intensities.

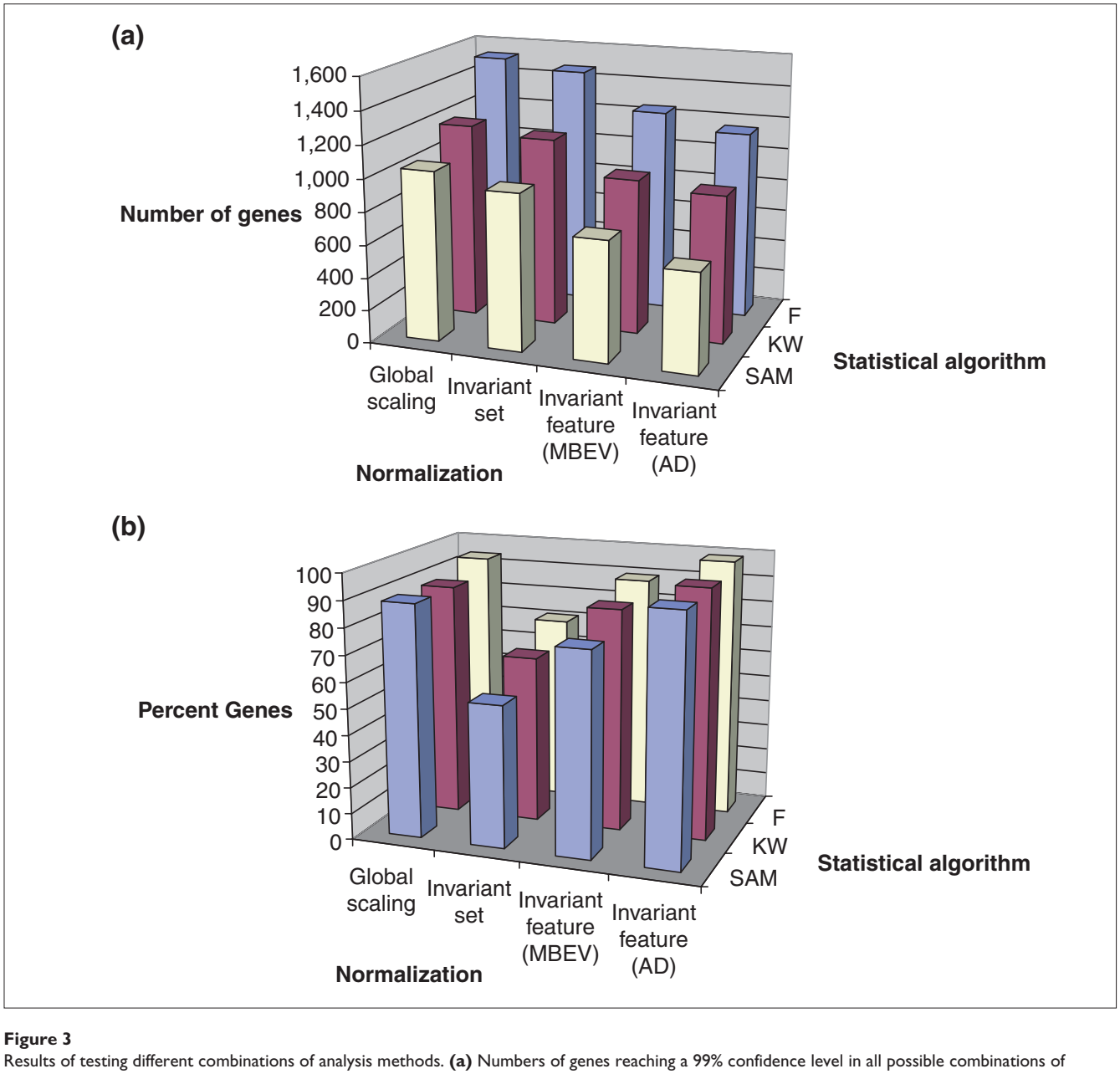

Figure 3
Results of testing different combinations of analysis methods. **(a)** Numbers of genes reaching a 99% confidence level in all possible combinations of normalization and statistical analysis algorithms. *x*-axis, normalization methods; *y*-axis, statistical analysis algorithms. Column height, number of genes. **(b)** Percentage of genes from Figure 2a that additionally reach a ratio of at least 2 and an absolute difference of at least 100 units. Layout is as in Figure 2a. AD, average difference; F, F-statistics; KW, Kruskal-Wallis; MBEV, model-based expression values; significance analysis of microarrays (SAM).

expressed genes in a oligonucleotide microarray experiment. The dataset used describes five different cellular stages of an ordered differentiation pathway [13,14]. We focused on statistical algorithms designed for proper analysis of such multiclass experimental designs.

A first striking observation is that the number of genes detected as differentially expressed varies by a factor of almost three, depending on which combination of normalization method and statistical analysis has been carried out. The genes detected by a confidence criterion alone show large differences in mean and median fold changes, and thus show different susceptibility to the use of additional criteria for filtering. This affects primarily the dataset normalized by the invariant probe set method. Also, genes in this dataset show smaller fold changes, and the maximum fold change achieved is 6- to 11-fold lower than in the other datasets. Most probably, this is due to the shifting of data as a first step in the normalization so that only 2% of the raw values are below 20. This assigns a higher value to each data point

Table 1

Results of different methods of statistical analysis

		Invariant feature (AD)	Invariant feature (MBEV)	Invariant set	Global scaling
(a)	Mean	3.55	2.45	1.59	3.37
	Median	2.09	1.46	1.30	2.12
	Max	429.01	667.16	71.63	848.28
(b)	Mean FC				
	F	15.05	9.69	3.06	10.28
	KW	18.48	12.24	3.35	12.21
	SAM	21.91	13.75	3.63	13.08
	Median FC				
	F	5.77	2.97	2.08	3.89
	KW	7.89	3.63	2.25	4.77
	SAM	9.20	4.11	2.41	4.92
(c)	Mean	428.71	750.20	324.17	384.47
	Median	85.52	259.83	122.04	102.37
	Max	25141.89	24616.80	28650.70	24794.22
(d)	Mean difference				
	F	2248.04	3346.51	1427.59	1784.93
	KW	2545.81	3881.37	1655.16	2044.61
	SAM	3157.21	4451.51	1827.97	2216.36
	Median difference				
	F	1116.52	1954.57	664.49	954.38
	KW	1328.38	2444.28	815.85	1138.39
	SAM	1865.47	2927.57	952.92	1248.04

(a) Mean, median, and maximum fold changes (FC) of all genes on the two array types normalized with different methods. AD, average difference; MBEV, model-based expression values. The averages are given across all genes of the ratio between the maximum and the minimum value across the five different conditions, calculated after averaging replicate arrays. **(b)** Mean and median fold changes (FC) of genes detected as differentially expressed with 99% confidence by all possible combinations of normalization and statistical analysis algorithms, calculated as in (a). F, F-statistics; KW, Kruskal-Wallis; significance analysis of microarrays (SAM), **(c)** Mean, median, and maximum signal differences of all genes on the two array types normalized with different methods, calculated by analogy with the ratios in (a). **(d)** Mean and median signal differences of genes detected as differentially expressed with 99% confidence by all possible combinations of normalization and statistical analysis algorithms, calculated as in (c).

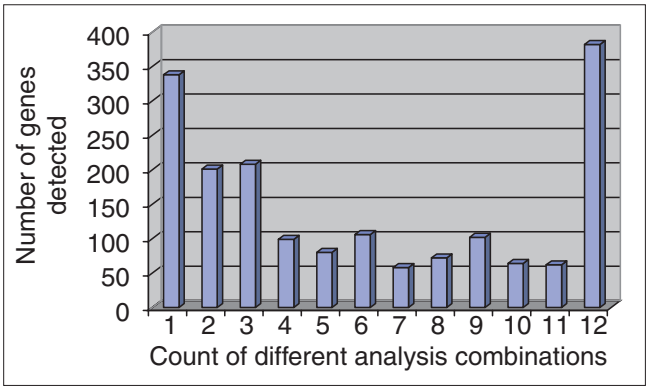

Figure 4
Numbers of genes detected by one or more of the 12 possible combinations between normalization and statistical analysis. *x*-axis, count of different combinations between normalization and statistical algorithm, *y*-axis, number of genes detected as differentially expressed in the respective number of different analysis combinations. A total of 12 different analysis combinations (four normalization procedures times three methods to detect differentially expressed genes) have been investigated.

while preserving the difference between any two data points, effectively reducing the ratio.

Analysis of the identity of probe sets indicates that the normalization method has a very high influence of detection of differentially expressed genes. In those cases where one gene is detected by two or more different combinations of normalization and statistical analysis algorithms, these combinations usually employ the same normalization. Moreover, genes are detected as differentially expressed to a high percentage in different combinations of analysis strategies only when the same normalization has been applied. Applying the same statistical algorithms, in contrast, does not have such a profound effect. Finally, a high rank correlation for probability of differential expression can only be achieved if the same normalization procedure has been applied. Thus, normalization appears to have a higher influence on the set of differentially expressed genes than the choice of statistical algorithm.

A few points should be kept in mind when interpreting the results presented here. First, the dataset chosen consists of five independent replicate experiments, a situation rarely encountered in microarray experiments.

Second, the dataset has been derived from highly purified cell populations. This situation is also different from most other microarray experiments. It might therefore be that the effects described here are even more pronounced in different experimental settings investigating less well-defined materials.

Third, the samples have been subjected to two rounds of RNA amplification. Samples prepared according to the standard Affymetrix protocols might behave differently. However, we do not expect this to be a significant issue, as the hybridization behavior of amplified and standard probes has been shown by us and other groups [15,16] to be similar. Moreover, we have high confidence in our dataset, as many genes with known expression patterns are detected in concordance with earlier results [13,17,18].

Fourth, the present analysis follows the general practice of detecting differentially expressed genes solely on the basis of

Table 2

Percentage of genes identical among all genes detected by different combinations of normalization and statistical analysis

	Invariant feature; AD, F	Invariant feature; AD, KW	Invariant feature; AD, SAM	Invariant feature; MBEV, F	Invariant feature; MBEV, KW	Invariant feature; MBEV, SAM	Invariant set; F	Invariant set; KW	Invariant set; SAM	Global scaling; F	Global scaling; KW	Global scaling; SAM
Invariant feature; AD, F		**93**	**100**	78	83	92	81	83	88	64	74	78
Invariant feature; AD, KW	74		**95**	67	79	82	74	79	81	55	67	69
Invariant feature; AD, SAM	56	66		51	59	74	58	61	66	41	50	54
Invariant feature; MBEV, F	69	75	82		**96**	**100**	77	80	82	58	67	70
Invariant feature; MBEV, KW	62	74	79	80		**95**	71	75	77	51	62	64
Invariant feature; MBEV, SAM	56	63	81	68	77		63	66	70	44	53	57
Invariant set; F	59	69	77	63	70	76		98	**100**	56	67	70
Invariant set; KW	56	67	74	60	68	74	**90**		**96**	52	66	66
Invariant set; SAM	55	64	74	58	65	73	85	89		49	61	66
Global scaling; F	80	87	92	81	86	90	95*	96*	98*		**100**	**100**
Global scaling; KW	72	83	88	73	82	85	89	96*	95*	78		**93**
Global scaling; SAM	70	78	88	71	77	84	85	89	94*	72	86	

Percentages are relative to the method defined in the column headings. Values of 100% and greater than 90% are highlighted in bold**.** *Percentage of genes identified after invariant probe set normalization that are also identified after global scaling (see text for details).

differences in the average difference value. This represents only a small proportion of the information generated by the analysis of the fluorescence images. Additional information about cross-hybridization, fractions of probe pairs contributing to the signal, and 'presence' or 'absence' of a transcript, among others, is readily available [1,2]. Unfortunately, no consensus exists on how to incorporate this additional information in a setting that cannot be handled by the manufacturer's software. Anecdotally, individual groups apply their own, often arbitrarily chosen, criteria to increase confidence in the results [19,20]. Data-analysis algorithms employing as much information as possible with incorporation of replicate experiments and the ability to analyze more than two conditions simultaneously are urgently needed.

Finally, testing more than 13,000 hypotheses on only five different conditions constitutes a significant multiple-testing problem. It is commonly accepted in multivariate statistics that the number of hypothesis should not exceed the number of parameters. Thus, when testing such a high number of hypotheses, the probability of at least one falsely rejected null hypothesis (the so-called family-wise error rate) is high. Although multiple solutions to this problem have been proposed (like SAM, controlling the false-discovery rate rather than the type-I error rate) [11,12], to date no consensus exists on how to deal with that problem in the context of gene-expression analysis.

The question naturally arises of which combination of algorithms is 'best' for analyzing gene-expression data. There is probably no general answer. One has to balance sensitivity, which attempts to detect all differentially expressed genes, against specificity, which attempts to reduce the number of false positives as much as possible. This is nicely illustrated by the set of genes mentioned above that are known to change. All of the 21 genes examined so far are detected in at least one method combination. However, two genes are detected by only one combination, and only seven of the genes known to change during B-cell differentiation are detected by all 12 combinations of methods. Thus, the more specific an algorithm is, the more likely is a loss of sensitivity. However, with the high number of differentially expressed genes typically detected in a microarray experiment, specificity might be a major issue.

As most of the genes detected by the permutation-based SAM method are enclosed in the ANOVA models, this algorithm appears to be inherently more specific than the classical ones. Regarding normalization, the invariant-feature method with calculation of average difference values yields

Table 3

Rank correlations for probability of differential expression between all possible combinations of normalization and statistical analysis procedures

	Invariant feature; AD, F	Invariant feature; AD, KW	Invariant feature; AD, SAM	Invariant feature; MBEV, F	Invariant feature; MBEV, KW	Invariant feature; MBEV, SAM	Invariant set; F	Invariant set; KW	Invariant set; SAM	Global scaling; F	Global scaling; KW	Global scaling; SAM
Invariant feature; AD, F	**1.000**	**0.922**	**0.995**	0.512	0.495	0.512	0.545	0.549	0.546	0.626	0.618	0.623
Invariant feature; AD, KW	**0.922**	**1.000**	**0.915**	0.487	0.487	0.487	0.540	0.555	0.541	0.604	0.621	0.600
Invariant feature; AD, SAM	**0.995**	**0.915**	**1.000**	0.517	0.500	0.517	0.553	0.555	0.553	0.625	0.617	0.623
Invariant feature; MBEV, F	0.512	0.487	0.517	**1.000**	**0.941**	**1.000**	0.440	0.435	0.440	0.443	0.444	0.446
Invariant feature; MBEV, KW	0.495	0.487	0.500	**0.941**	**1.000**	**0.941**	0.431	0.433	0.431	0.428	0.435	0.432
Invariant feature; MBEV, SAM	0.512	0.487	0.517	**1.000**	**0.941**	**1.000**	0.440	0.436	0.440	0.443	0.444	0.447
Invariant set; F	0.545	0.540	0.553	0.440	0.431	0.440	**1.000**	**0.933**	**1.000**	0.738	0.730	0.749
Invariant set; KW	0.549	0.555	0.555	0.435	0.433	0.436	**0.933**	**1.000**	**0.933**	0.719	0.743	0.729
Invariant set; SAM	0.546	0.541	0.553	0.440	0.431	0.440	**1.000**	**0.933**	**1.000**	0.739	0.731	0.750
Global scaling; F	0.626	0.604	0.625	0.443	0.428	0.443	0.738	0.719	0.739	**1.000**	**0.925**	**0.994**
Global scaling; KW	0.618	0.621	0.617	0.444	0.435	0.444	0.730	0.743	0.731	**0.925**	**1.000**	**0.918**
Global scaling; SAM	0.623	0.600	0.623	0.446	0.432	0.447	0.749	0.729	0.750	**0.994**	**0.918**	**1.000**

Values of >0.9 are highlighted in bold.

the smallest set of genes. However, this set is not a subclass of the genes detected after normalization with other methods (Table 2), as is the case for SAM. Ideally, a researcher would have a set of genes with known differential expression and a set known not to be differentially expressed. This could be used to define the conditions for analysis. In the absence of such a training set, it is probably a wise decision to use the algorithms likely to result in the most specific analysis.

Materials and methods

Gene-expression dataset

The B-cell precursor gene-expression dataset described here has been published in detail previously [13]. Total femoral bone marrow cells of 5-6-week-old C57/BL6 mice (n = 4 per experiment) were divided into into three equal samples. Cells were stained and five populations representing consecutive cellular differentiation stages were sorted. These stages were: pre-BI cells (c-Kit$^+$ B220$^+$), large pre-BII cells (surface immunoglobulin (sIg)$^-$ CD25$^+$ B220$^+$ large), small pre-BII cells (sIg$^-$ CD25$^+$ B220$^+$ small), immature B (sIgM$^+$ B220^{lo}) and mature B cells (sIg$^+$ B220^{hi}) [14]. A total of 50,000 (pre-BI, large pre-BII) or 150,000 (small pre-BII, immature and mature B cells) cells were sorted directly into TRIzol RNA isolation reagent (Life Technologies) at 50,000 cells/500 μl TRIzol. A cell purity of ≥98% was routinely achieved. RNA was then subjected to two rounds of *in vitro* transcription-based RNA amplification as described earlier [13,16,21]. Affymetrix Mu11k subA and subB GeneChip® arrays were hybridized, washed, stained and scanned according to the manufacturer's specifications. Five independent replicate experiments were performed; thus, a total of 50 chips is included in the current analysis (5 conditions x 5 replicates x 2 chip layouts). Scanned raw data images were processed with Affymetrix GeneChip v3.2 software, resulting in processed image (.cel) and numerical (.chp) files. The entire dataset can be obtained from the NCBI at [22] under accession GSE13.

Normalization

Four different normalization procedures were used. All normalizations were carried out separately for the subA and subB arrays, respectively. After normalization, all gene-expression values below 20 were set to 20 to eliminate low-level signals. SubA and subB chip types were combined into one gene-expression matrix.

Global scaling

For global scaling, average difference values were extracted from the .chp files. All average difference values of every

chip were summed up, and the mean of these sums across all chips of the same layout was calculated. The ratio of the actual average difference sum for any given chip and the mean of all average difference sums across all chips with the same layout served as a correction factor for this chip, with which all the average difference values were multiplied.

Invariant feature normalization and model-based expression values

For invariant-feature normalization, the program dchip [3,4] was used. Processed image (.cel) files served as input, and normalization was carried out according to the developer's specifications. Briefly, the program first identified a baseline array with median overall fluorescence intensity. Next, for every array, invariant features (defined as all the features with similar ranks of fluorescence intensity between two arrays) were identified. Finally, a piecewise linear running median line based on the invariant features was calculated and used as the normalization curve. After normalization, both traditional average difference values and model-based expression values (MBEV) were calculated and exported to Microsoft Excel.

Invariant set normalization

For invariant-set normalization on the probe-set level, all average difference levels were extracted from the .chp files and imported into 'The Equalizer' [8]. Normalization was performed according to the developer's specifications. Briefly, all values were first shifted (by adding a constant value) so that only 2% of the data points had an intensity below 20. Next, values with similar ranks (± 15) between two arrays were identified, taking the first array of the set as baseline. A curve was fitted on this similar-rank subset. Finally, all data points were shifted so that the original similar-rank subset has a slope of 1. Normalized values were exported to Microsoft Excel.

Statistical analysis

Three different methods were used for detection of differentially expressed genes. The F-test for parametric ANOVA and the H (Kruskal-Wallis) test for nonparametric ANOVA were implemented in Microsoft Excel using standard formulas [9]. The permutation-based method SAM [11] is freely available to academic researchers as an add-in for Microsoft Excel. To enable comparisons between SAM and the two ANOVA approaches, we considered a median false-detection rate of 1% in SAM as comparable to a 99% confidence level in ANOVA. Maximum fold changes and maximum differences in gene expression were calculated from the minimum and the maximum of the population-wise means across the five cellular populations examined.

References

1. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al.*: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14:**1675-1680.
2. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ: **Genome-wide expression monitoring in *Saccharomyces cerevisiae*.** *Nat Biotechnol* 1997, **15:**1359-1367.
3. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98:**31-36.
4. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.** *Genome Biol* 2001, **2:**research0032.1-0032.11.
5. Fambrough D, McClure K, Kazlauskas A, Lander ES: **Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes.** *Cell* 1999, **97:**727-741.
6. Hill AA, Brown EL, Whitley MZ, Tucker-Kellogg G, Hunter CP, Slonim DK: **Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls.** *Genome Biol* 2001, **2:**research0055.1-0055.13.
7. Schadt EE, Li C, Su C, Wong WH: **Analyzing high-density oligonucleotide gene expression array data.** *J Cell Biochem* 2000, **80:**192-202.
8. Stuart RO, Bush KT, Nigam SK: **Changes in global gene expression patterns during development and maturation of the rat kidney.** *Proc Natl Acad Sci USA* 2001, **98:**5649-5654.
9. Zar JH: *Biostatistical Analysis*, 4th edn. Upper Saddle River, NJ: Prentice Hall, 1999.
10. Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P: **Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. analysis of global gene expression in *Escherichia coli* K12.** *J Biol Chem* 2001, **276:**19937-19944.
11. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
12. Benjamini X, Hochberg X: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Stat Soc B* 1995, **57:**289-300.
13. Hoffmann R, Seidl T, Neeb M, Rolink A, Melchers F: **Changes in gene expression profiles in developing B cells of murine bone marrow.** *Genome Res* 2002, **12:**98-111.
14. Rolink A, Grawunder U, Winkler TH, Karasuyama H, Melchers F: **IL-2 receptor alpha chain (CD25, Tac) expression defines a crucial stage in pre-B cell development.** *Int Immunol* 1994, **6:**1257-1264.
15. Baugh LR, Hill AA, Brown EL, Hunter CP: **Quantitative analysis of mRNA amplification by *in vitro* transcription.** *Nucleic Acids Res* 2001, **29:**E29.
16. Luo L, Salunga RC, Guo H, Bittner A, Joy KC, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, *et al.*: **Gene expression profiles of laser-captured adjacent neuronal subtypes [Erratum *Nat Med* 1999 Mar;5(3):355].** *Nat Med* 1999, **5:**117-122.
17. Grawunder U, Leu TM, Schatz DG, Werner A, Rolink AG, Melchers F, Winkler TH: **Down-regulation of RAG1 and RAG2 gene expression in preB cells after functional immunoglobulin heavy chain rearrangement.** *Immunity* 1995, **3:**601-608.
18. Melchers F, Rolink A: **B-Lymphocyte Development and Biology.** In *Fundamental Immunology*. Edited by WE Paul. Philadelphia/New York: Lippincott-Raven; 1999:183-224.
19. Mills JC, Syder AJ, Hong CV, Guruge JL, Raaii F, Gordon JI: **A molecular profile of the mouse gastric parietal cell with and without exposure to *Helicobacter pylori*.** *Proc Natl Acad Sci USA* 2001, **98:**13687-13692.
20. Ehrt S, Schnappinger D, Bekiranov S, Drenkow J, Shi S, Gingeras TR, Gaasterland T, Schoolnik G, Nathan C: **Reprogramming of the macrophage transcriptome in response to interferon-gamma and *Mycobacterium tuberculosis*. Signaling roles of nitric oxide synthase-2 and phagocyte oxidase.** *J Exp Med* 2001, **194:**1123-1140.
21. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P: **Analysis of gene expression in single live neurons.** *Proc Natl Acad Sci USA* 1992, **89:**3010-3014.
22. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]