The Plant Journal (2007) **52**, 561–569 doi: 10.1111/j.1365-313X.2007.03227.x

TECHNICAL ADVANCE

EVE (external variance estimation) increases statistical power for detecting differentially expressed genes

Anja Wille[1,2], Wilhelm Gruissem[3], Peter Bühlmann[1] and Lars Hennig[3,*]
[1]*Seminar for Statistics, ETH Zurich, CH-8092, Zurich, Switzerland,*
[2]*Colab, ETH Zurich, CH-8092 Zurich, Switzerland, and*
[3]*Institute of Plant Sciences & Zurich–Basel Plant Science Center, ETH Zurich, CH-8092 Zurich, Switzerland*

Received 22 January 2007; revised 21 May 2007; accepted 11 June 2007.
[*]For correspondence (fax +41 1 632 1044; e-mail lars.hennig@ipw.biol.ethz.ch).

Summary

Accurately identifying differentially expressed genes from microarray data is not a trivial task, partly because of poor variance estimates of gene expression signals. Here, after analyzing 380 replicated microarray experiments, we found that probesets have typical, distinct variances that can be estimated based on a large number of microarray experiments. These probeset-specific variances depend at least in part on the function of the probed gene: genes for ribosomal or structural proteins often have a small variance, while genes implicated in stress responses often have large variances. We used these variance estimates to develop a statistical test for differentially expressed genes called EVE (external variance estimation). The EVE algorithm performs better than the *t*-test and LIMMA on some real-world data, where external information from appropriate databases is available. Thus, EVE helps to maximize the information gained from a typical microarray experiment. Nonetheless, only a large number of replicates will guarantee to identify nearly all truly differentially expressed genes. However, our simulation studies suggest that even limited numbers of replicates will usually result in good coverage of strongly differentially expressed genes.

Keywords: Affymetrix, Arabidopsis, biological noise, gene expression, microarray, statistical test.

Introduction

In all living organisms, the number of active genes in any given cell at any given time is much lower than the total number of genes. Cellular identity and physiology depends largely on the particular subsets of expressed genes. Therefore, determination of gene activity is a central question in biology. Many techniques exist to measure gene activity, but microarrays are currently the method of choice to profile transcript abundance at a genome-wide scale (Stoughton, 2005). In recent years, microarray technology has not only become more robust but also much more affordable for many laboratories. The growing body of published microarray studies called for common experimental annotations, which were established with MIAME (Minimum Information About a Microarray Experiment) and domain-specific extensions such as MIAME-Plant (Brazma *et al.*, 2001; Zimmermann *et al.*, 2006). Several public databases exist that permit online queries on thousands of annotated microarray experiments, including Array-Express, Genevestigator and NASCarrays (Brazma *et al.*, 2003; Craigon *et al.*, 2004; Zimmermann *et al.*, 2005).

For the model plant Arabidopsis, the Affymetrix ATH1 GeneChip® microarray is probably the most commonly used (Hennig *et al.*, 2003; Redman *et al.*, 2004; Zimmermann *et al.*, 2005). For example, the ATH1 array was used for the AtGenExpress project to establish an expression map of Arabidopsis and forms the basis for the Genevestigator software (Altmann *et al.*, 2004; Schmid *et al.*, 2005). The ATH1 array has 22 746 probesets, which probe the transcript abundance of 23311 genes (TAIR annotation of 5 April 2006).

Studies using microarrays can have many different designs, but most often the biologist looks for genes differentially expressed between a control and a treatment, a mutant or a transgenic plant. Such studies possibly represent the simplest of any microarray experiments. However, the identification of differentially expressed genes is not a trivial problem (Cui and Churchill, 2003; Smyth *et al.*, 2003), because microarray data are often noisy and in addition suffer from the 'curse of dimensionality': thousands of genes (variables) are measured for only few cases (replicates). The low number of replicates in combination with the use of a multiple-testing-adjusted *P*-value strongly reduces the power to identify differentially expressed genes. In particular, a two-sample *t*-test often does not perform satisfactorily with real-world data because the limited number of replications does not permit accurate estimation of the variance.

Studies to detect differentially expressed genes aim to identify sets of genes that possibly play are role in the various regulatory mechanisms of control and treatment, and differential expression of genes is typically measured on an individual level, i.e. by quantifying the univariate association of individual genes to a treatment variable or class label. However, genes frequently interact, and gene expression signals are usually not independent. Genetic interaction and specific regulatory pathways, which are major causes for lack of gene expression signal independence, cannot be detected with simple tests for differential expression, and are modeled at a later stage of the analysis.

In general, the identification of differentially expressed genes includes three steps: (i) normalization to minimize systematic errors (bias), (ii) transformation to minimize the variance–mean dependence of the probe intensities, and (iii) statistical testing for significant differences between signal means. For microarrays manufactured by Affymetrix, which have multiple probes for every tested transcript, calculation of probeset summaries constitutes an additional step. The combination of multiple measurements for each transcript into a single summary signal per probeset is not trivial, and multiple algorithms have been proposed to this end. In plant biology, GCRMA (Wu *et al.*, 2004) is possibly the most widely accepted, but MAS5 (Liu *et al.*, 2002) and RMA (Irizarry *et al.*, 2003) are commonly used as well. While MAS5 does not involve any transformation of signals and thus has a very strong variance–mean dependence, RMA and GCRMA involve log transformation that efficiently stabilizes the variance. Even better variance stabilization can be obtained using the vsn algorithm (Huber *et al.*, 2002).

To increase statistical power in data sets with few replicates, regularization techniques that 'borrow' statistical information across genes are often applied to improve variance estimation. To this end, several approaches have been proposed (Efron *et al.*, 2001; Kendziorski *et al.*, 2003; Tusher *et al.*, 2001). One very popular implementation of this idea is the LIMMA algorithm (Smyth, 2004). LIMMA uses expression data for other genes to obtain a modified (shrinkage) estimate of the variance of the gene of interest. Other approaches include pooling of variances across genes with similar variances (Jain *et al.*, 2003; Newton *et al.*, 2001; Quackenbush, 2002; Rocke and Durbin, 2001). Here, differentially expressed genes are identified based on a *Z* test with pooled variance estimates.

Although this approach often works quite well, it is not necessarily biologically reasonable that variance estimates can be combined for different genes. From a biologist's point of view, it is much more reasonable to assume that the measurement for each transcript has a specific variance that depends on (i) the probe properties of the microarray, and (ii) the transcriptional and post-transcriptional control of the gene (biological noise; Chubb *et al.*, 2006; Newman *et al.*, 2006). One approach along this line suggests the use of available microarray data from Gene Expression Omnibus (GEO) to obtain an estimate of the gene-specific variance (Kim and Park, 2004). For each gene, estimates are pooled across different experiments (not genes) to obtain a more reliable estimate of the variance, which will improve statistical power when used in tests, particularly for studies with low replicate numbers.

One disadvantage of the GEO-adjusted algorithm relies on invariant gene-specific variances, which are rarely encountered in real life. Here, we propose EVE (external variance estimates) as an extension of the original GEO-adjusted algorithm. We derived sets of gene-specific variance estimators from many experiments with few replicates each, and include possible dependencies on the signal intensity. Using a large database of microarray experiments, we have implemented EVE for Arabidopsis ATH1 microarray data, and provide gene-specific variance estimators. We find that, in situations with few available replicates, EVE outperforms both conventional *t*-tests and LIMMA.

Results and discussion

Estimation of probe-set-specific variances from databases

In order to improve power when identifying differentially expressed genes, Kim and Park (2004) suggested obtaining gene-specific variances from a large database of microarray experiments. Because variances of gene expression measurements depend not only on the gene but also on the detection method, we prefer to use the term probeset-specific variance rather than gene-specific variance. Large databases often contain experiments that involve diverse tissues and organs with widely varying gene expression activities. Therefore, extracted probeset-specific variances may reflect tissue effects rather than the variation of replicated measurements. The GEO-adjusted approach used either a global variance estimate based on all data or a pooled variance estimate based on experimental sets. These experimental sets usually combined data from only one type of tissue but different treatments, genotypes or clinical states of normal and tumour samples. The relative performance of both estimators was heavily influenced by the database composition (Kim and Park, 2004). Because the detection of

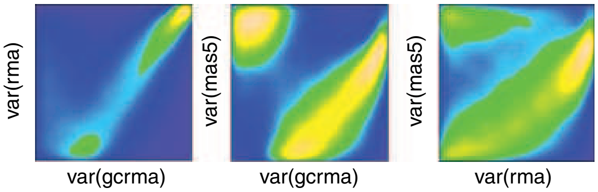

Figure 1. Probeset-specific variances for replicated microarray experiments depend on the algorithm generating summary signals.
Variances were calculated for every probeset based on four replicates per sample after processing the raw data with GCRMA, RMA or MAS5. Rank-transformed variances are displayed as heat plots. (a) RMA versus GCRMA; (b) MAS5 versus GCRMA; (c) MAS5 versus RMA. Data points cluster along the diagonal if probeset-specific variances are independent of the processing algorithm. Data were obtained from Vandepoele *et al.* (2005).

differentially expressed genes requires an estimator of the variation in replicated measurements, we use neither global nor pooled variances but instead variances derived exclusively from biological replicates combining only data from one type of tissue, treatment and genotype. When combining these variance estimates from many experiments with few replicates each, a reliable estimate of the gene-specific variance can be obtained (see Experimental procedures). This approach is much more realistic and biologically justified than the originally proposed GEO-adjusted method.

Various probe-level analysis methods exist for Affymetrix GeneChip microarrays, and different methods give different summary signals. Similarly, the variance in replicated experiments depends on the probe-level analysis method. This is visualized using a published data set that contains four replicates for each of two Arabidopsis genotypes (Vandepoele *et al.*, 2005). Variances were calculated for every probeset based on four replicates per sample after processing the raw data with GCRMA, RMA or MAS5. Variances were compared using heat plots (Figure 1), and the spreading of data points off the diagonal shows the effect of the processing algorithm. Probeset-specific variances were often similar for RMA- and GCRMA-processed data but differed considerably for MAS5-processed data. Together, these results demonstrate that probeset-specific variances often depend on the probe-level analysis method used. In the remainder of this study, we mainly focus on GCRMA-processed data, and present some results for data processed by MAS5, RMA or vsn.

We calculated probeset-specific variances for each of 258 duplicated and 131 triplicated experiments in the AtGenExpress data set (Altmann *et al.*, 2004; Schmid *et al.*, 2005), and extracted conservative estimates for the overall probeset-specific variance (see Experimental procedures).

Dominating functional categories differ between the high-variance and low-variance ranges

The concept of using gene- or probeset-specific variances to detect differentially expressed genes relies on the biological notion that some genes tend to show stronger variation between replicate samples than others. To test this assumption we identified the 500 probesets with the largest mean variance and the 500 probesets with the smallest mean variance, and analyzed the distribution of the probed genes in gene ontology (GO) categories. We found that several GO categories were highly enriched among the low- or high-variance genes (Figure 2). Genes with high variance vary strongly between replicates. Many of these genes function in responses to stress or biotic and abiotic stimuli (Figure 2a) and often encode membrane proteins (Figure 2c). In contrast, genes with small variance vary only weakly between replicates. Many of these genes function in ribosomal protein synthesis (Figure 2a) and often encode proteins with structural functions or components of the ribosome (Figure 2b,c). Together, these results meet the intuitive expectations of many biologists, and confirm our hypothesis that variability of expression measures between replicates has a strong gene-specific component.

Probeset-specific variances often vary depending on the expression level

In microarray experiments, it has often been observed that variance depends on signal intensity, as evident in funnel-shaped MA plots. This is biologically plausible because signals can be derived with greater precision for strong than for weak signals. In contrast to MA plots, which visualize the variance for many probesets under a single experimental condition, we consider the variance for a single probeset under many experimental conditions. In this case, the variance–mean relationship is not obvious, and the benefit of traditional variance stabilization methods is not clear. To explore the variance–mean relationship, we constructed a heat plot of the probeset-specific variance versus the mean from the 380 reference experiments in the AtGenExpress database (Figure 3a). On a genome-wide level, there is no clear correlation between signal means and variances. Next, we tested whether the signal variance is independent of the signal mean for each probeset. We fitted a linear model for variance versus mean and corrected for multiple testing according to the method described by Benjamini and Hochberg (1995): 16 086 probesets had a slope significantly different from zero ($P < 0.05$). Although the slope was predominantly positive, there were also probesets with negative slopes for variance versus mean (Figure 3b). For comparison, there were 22 190, 14 116 and 10 996 probesets with a slope significantly different from zero ($P < 0.05$) when MAS5, RMA or vsn, respectively, were used for normalization, transformation and calculation of probeset summaries. Together these results established that (independent of the data-processing algorithm), the probeset-specific variance is often not constant but a function of the signal mean.

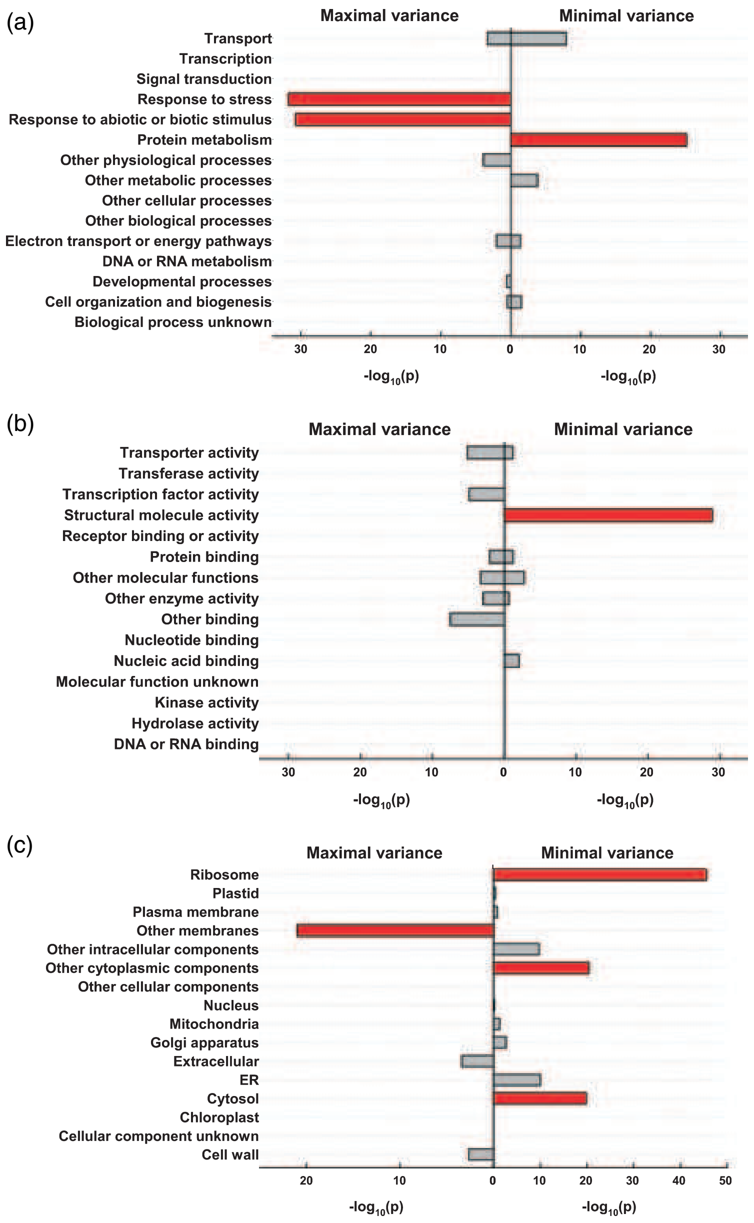

Figure 2. Genes with large and small probeset-specific variances belong to different functional categories.
For each probeset, the minimal and maximal replicate variance observed in the 380 reference experiments of the AtGenExpress data set (Altmann *et al.*, 2004; Schmid *et al.*, 2005) was determined. Five hundred probesets with the largest variance and 500 with the smallest average variance were selected and analyzed for representation of Gene Ontology (GO) categories. Probabilities were determined by a hypergeometric test with multiple testing corrections (Bonferroni). Displayed are log-likelihoods for enrichment of the biological process (a), molecular function (b) and cellular localization (c) categories. Red bars denote the most significant enrichments for each gene set.

Recursive partitioning versus global variance estimation

Because replicate variances are usually not independent of signal intensities, we calculated the probeset-specific variance in two ways: (i) we averaged all available variances, and (ii) we used classification and regression trees as implemented in the recursive partitioning algorithm *rpart* (Breiman *et al.*, 1984; Therneau and Atkinson, 1997) to partition the set of variances according to signal intensities (see Experimental procedures for details). Both procedures are equivalent for genes whose replicate variance does not depend on the signal intensity and where no partitioning takes place.

In the recursive partitioning, the signal intensity region is recursively broken into smaller intervals in which replicate variances are assumed to be constant. The original region is split in such a way that replicate variances show little variation within each of the selected intervals. In contrast, average replicate variances differ largely between intervals. With this partitioning algorithm, a non-linear dependency between replicate variances and signal intensities can be modeled.

Table 1 shows the frequency of nodes generated by *rpart*. About 650 probesets have replicate variances that are independent of signal intensity (no partitioning), but the majority of probesets have replicate variances that can be

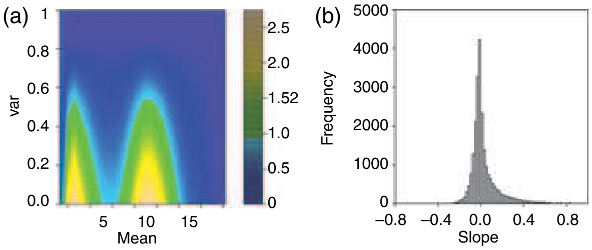

Figure 3. Relationship of probeset-specific variance and signal mean.
(a) Probeset-specific variances and means were calculated for the 380 reference experiments in the AtGenExpress data set and displayed as a heat plot.
(b) Histogram of the slopes obtained when a linear model for variance versus mean was fitted for every probeset.

Table 1 Partitioning of replicate variances according to signal intensities

Bin size	Signal range[a]	n_{GCRMA}	n_{MAS5}	n_{RMA}	n_{vsn}
1	4.4 (1.8)/2.6 (1.4)	649	147	1005	1909
2	5.2 (2.6)/3.1 (1.8)	11 076	13 656	12 318	12 397
3	5.6 (2.8)/3.5 (1.8)	8665	7875	7986	6860
4	5.8 (2.8)/3.6 (1.7)	2265	1065	1397	1535
5	6.4 (2.9)/3.4 (1.8)	152	66	104	108
6	7.0 (2.6)/5.3	3	1	0	1

[a] Median (MAD) for GCRMA/vsn.

partitioned in two or three signal intensity regions. Next, we analyzed the effect of the data pre-processing algorithm. Although vsn helps to increase the number of probesets that do not require partitioning to 1900, even with vsn the large majority of probesets require partitioning in up to five bins. When averaged over all probesets, the degree of partitioning (i.e. the number of bins created) depends on the spread of signal intensities observed for any given probeset as measured by $\max_{(\text{signal})} - \min_{(\text{signal})}$ for both GCRMA- and vsn-processed data (Table 1). Together, these results demonstrate that, for most probesets, it is preferable not to pool all variances but only those in certain signal intensity regions obtained by partitioning the data.

EVE outperforms the t-test and LIMMA

Due to small sample size, sample variances are often poorly estimated in microarray data, which leads to inflated *t*-statistics and reduced power of *t*-tests. Like Kim and Park (2004), we therefore reasoned that detecting differentially expressed genes using the estimated probeset-specific variances should increase power. Because we use an external variance estimate, we call our algorithm EVE. In contrast to a conventional *t*-test, which estimates both population mean and variance from the sample data, EVE estimates only the population mean from the sample data and uses the tabulated probeset-specific variance from external data to detect differentially expressed genes by *Z*-tests.

Receiver–operator curves (ROC) based on simulated data (see Experimental procedures) illustrate the advantage of a priori knowledge of variances (Figure 4). For all settings, EVE outperformed the standard *t*-test. Because estimation of the variance by the *t*-test is particularly poor for low numbers of replicates, it is plausible that the difference between the *t*-test and EVE was greatest for two or three replicates (see Zien *et al*., 2003). In contrast, if 10 replicates were simulated, the performance of EVE and the *t*-test was more similar. Nonetheless, in real-world experiments, there are rarely more than four replicates for microarray data.

To establish whether EVE has advantages for real-world research, we tested the performance of EVE on real experimental data. However, a major problem with real-world data is that the truth is usually not known a priori, i.e. it is not known which genes are differentially expressed and which are not. We choose two strategies to allow at least a partial characterization of EVE's performance on real data despite the lack of a 'gold standard'.

First, we used a permutation test to estimate the proportion of false-positive hits returned by EVE using the same data set with four replicates described before (Vandepoele *et al*., 2005). From the complete data set, we generated all possible subsets representing experiments with two, three or four replicates from treatment and control each by sampling without replacement. Then we determined the number of probesets classified as differentially expressed by EVE using multiple-testing correction according to the method described by Benjamini and Hochberg (1995) and a *P*-value threshold of 0.05 (Table 2). To estimate the number of false positives, we generated all possible subsets from the data that contain a balanced mixture of wild-type and transgenic samples (i.e. both subsets contain equal numbers of treatment and control samples), and again determined the number of probesets classified as differentially expressed by EVE (Table 2). This was performed separately for GCRMA- and vsn-processed data, and we observed satisfying false-discovery rates of between 2% and 5% in all cases, demonstrating the validity of our approach for real-world biological data. Because EVE consistently identified more differentially expressed genes when using vsn-processed data at similar false-discovery rates, it is likely that data processing by vsn increases the power of detection of differentially expressed genes.

Second, we tested how many of the differentially expressed genes could be identified when only two or three replicates were used, in comparison to using all four replicates. Again all possible subsets representing experiments with two, three or four replicates were generated by sampling without replacement. Identification rates were

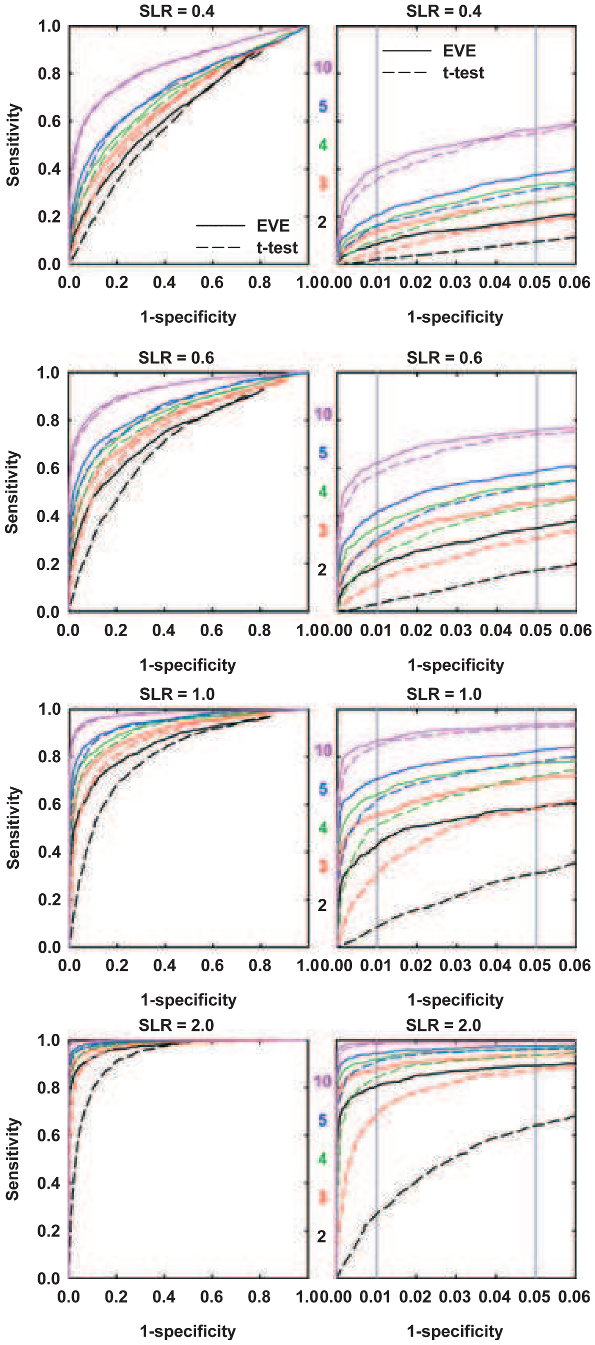

Figure 4. Performance of EVE on simulated data.
The simulation of data was based on the probeset-specific variance estimates (assuming a Normal distribution), and included 1000 randomly selected differentially expressed genes with signal log ratios (SLR) of 0.4, 0.6, 1 or 2, respectively. Shown are receiver–operator curves (ROC) for the *t*-test (broken lines) and EVE (solid lines) for various numbers of replicates randomly drawn from the simulated data sets. Colors represent two (black), three (red), four (green), five (blue) and ten (pink) replicates.

calculated as the mean ratio of identified probesets based on *n* replicates (*n* = 2, 3 or 4) relative to identified probesets based on four replicates (Figure 5a). For three replicates, the identification rate for EVE dropped to about 60%, and for two replicates the identification rate dropped further to about 40%. In contrast, with three replicates, LIMMA identified just 40% of the originally identified probesets (found with four replicates) and the *t*-test identified 10%. With only two replicates available, LIMMA found 10%, while the *t*-test usually identified none of the probesets identified with four replicates. Similarly, when only probesets that were identified by all three algorithms were used as the 'gold standard', EVE performed considerably better for two and three replicates than the *t*-test or LIMMA (Figure 5b). Thus, for this data set, EVE clearly out-performed the *t*-test in detecting affected probesets with reduced sample size. In addition, EVE with only two replicates performed similarly, on average, to LIMMA with three replicates.

However, a comprehensive and fair comparison of all three algorithms should include three different 'gold standards'. These three 'gold standards' should consist of the probesets identified by either LIMMA, the *t*-test or EVE based on all available data. In general, results of all three algorithms corresponded quite well (Figure 5c–e). Similarly to the previous results, EVE identified 40% of the standard probesets using just two replicates regardless whether the standard was LIMMA, the *t*-test or EVE. When three replicates were used, this number increased to about 60%. Again, both the *t*-test and LIMMA performed poorly with two replicates and generally worse than EVE with three replicates. Very similar results were obtained when vsn instead of GCRMA was used to process the data (not shown). Notably, even when all available data were used, EVE did not identify all probesets identified by LIMMA or the *t*-test and vice versa. Because we do not know which of the probesets identified by any algorithm are true-positive and which are false-positive results, we cannot decide whether 'missed' probesets are caused by false negatives in one algorithm or by false positives in the other. However, the permutation experiments (Table 2) resulted in acceptably small false-discovery rates, suggesting that most of the probesets exclusively identified by EVE represent true signals.

Conclusions

Efficiently identifying differentially expressed genes from microarray data is a non-trivial task that is usually complicated by low replicate numbers and large gene numbers. Therefore, the statistical power of standard procedures is often limited, and non-conservative multiple testing procedures can lead to many false positives. The major obstacle is correct estimation of the variance based on low replicate numbers. Several algorithms have been proposed to overcome these problems, and LIMMA is currently among the most popular. Nonetheless, *t*-tests and derived variants such as ANOVA are also commonly used. Analyzing several hundreds of replicated experiments, we found that

Table 2 Estimation of false-positive rates for EVE

	vsn			GCRMA		
Replicates	Number of different probesets[a]	FP[b]	FDR[c]	Number of different probesets[a]	FP[b]	FDR[c]
2	1682 (492)	70 (87)	4.1% (5.2%)	1343 (328)	51 (56)	3.8% (4.2%)
3	2724 (838)	56 (79)	2.0% (2.3%)	2140 (586)	58 (73)	2.7% (3.4%)
4	3864	196 (288)	5.1% (7.5%)	3065	142 (193)	4.6% (6.3%)

[a] Number of positive hits ($P < 0.05$); median (MAD) of multiple sample permutations.
[b] Number of false-positive hits; median (MAD) of multiple sample permutations.
[c] False-discovery rate; median (MAD) of multiple sample permutations.

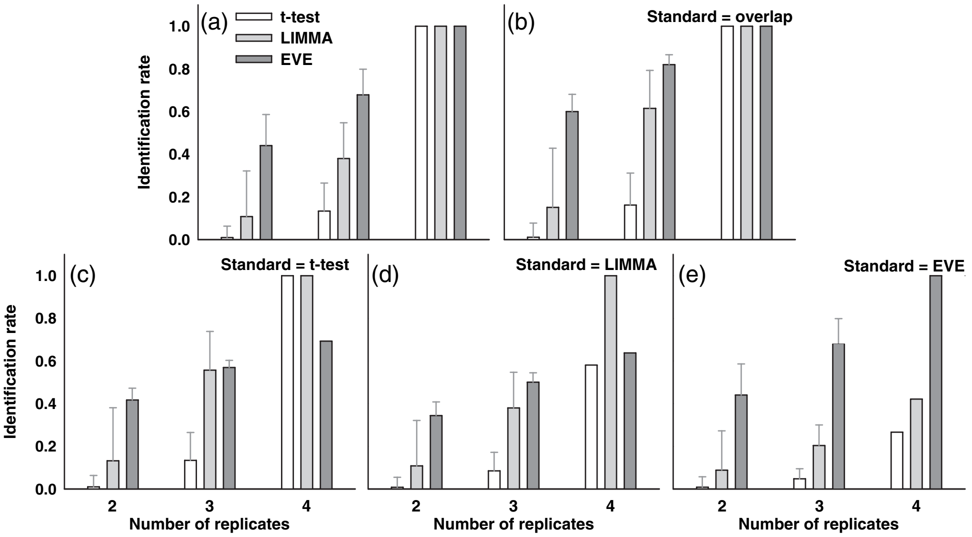

Figure 5. Performance of EVE on real-world data.

GCRMA-based microarray data from Vandepoele *et al.* (2005) with four replicates were analyzed with a standard *t*-test (white bars), LIMMA (light-gray bars) and EVE (partitioned variances) (dark-gray bars). Alternatively, all possible subsets representing experiments with two or three replicates were generated by sampling without replacement. In all cases, probesets were counted if $P < 0.05$ after multiple-testing correction according to the method described by Benjamini and Hochberg (1995).

(a) Fraction of probesets identified by the *t*-test, LIMMA or EVE with two or three replicates, compared to the number of probesets identified by the same algorithm with four replicates.

(b–e) Fraction of probesets identified by the *t*-test, LIMMA or EVE with two, three or four replicates compared to the number of probesets identified by the *t*-test (c), LIMMA (d), EVE (e) or by all three tests (b) with four replicates. Values are means + SD.

probesets have typical, distinct variances that depend, at least in part, on the biological function of the probed gene: genes for ribosomal or structural proteins often have small variances, while genes implicated in stress responses often have large variances. This is entirely reasonable from a biological point of view, and strongly justifies the use of the probeset-specific variances for detecting differentially expressed genes, as similarly suggested in another recent study (Kim and Park, 2004). However, in contrast to the approach by Kim and Park, who used pseudo-replicates throughout, our approach is based only on real replicates and is thus much more realistic. We found that external variance estimation performs better than the *t*-test and LIMMA on real-world data. In contrast to LIMMA and some other algorithms for detecting differentially expressed genes, EVE is much better biologically justified. Here, we implemented EVE as proof of principle for the Arabidopsis ATH1 microarray. In the near future, similar large data sets are expected to become available for other species. This will make EVE an attractive option for data analyses in a growing number of experimental settings. Nonetheless, it must be kept in mind that, although the variances used in EVE help to 'borrow' statistical power from hundreds of external microarrays, they are not necessarily an appropriate estimate for the actual variance in every experiment. In other words, EVE helps to maximize the information gained from a typical microarray experiment, but does not change the rule that only a large number of replicates will guarantee identification of nearly all truly differentially expressed genes. However, our simulation studies suggest that if one is willing to ignore genes that have only weakly altered expression, even limited numbers of replicates will usually result in a good coverage of truly differentially expressed genes.

Experimental procedures

Microarray data

Data for determining probeset-specific variances were obtained from 380 reference experiments of the AtGenExpress data set (Altmann *et al.*, 2004; Schmid *et al.*, 2005). The AtGenExpress project compiled a large set of experiments that used the Arabidopsis ATH1 array, including 249 experiments with duplicated measurements and 131 experiments with triplicated experiments. The names of the cel-files from the 909 arrays that were used here are listed in Supplementary Table S1 and are available for download at the TAIR FTP server (http://www.arabidopsis.org).

Data for testing the algorithm were taken from Vandepoele *et al.* (2005). All Affymetrix *.CEL files were processed using GCRMA (Wu *et al.*, 2004), MAS5 (Liu *et al.*, 2002), vsn (Huber *et al.*, 2002) and RMA (Irizarry *et al.*, 2003).

Software

All data processing was performed using the statistic package R (version 2.5.0) that is freely available at http://www.r-project.org/ (Ihaka and Gentleman, 1996). Data display was performed using R and Sigmaplot 8.0 (SPSS; http://www.sigmaplot.com). An R script for using EVE, tabulated variances and partitioning data are available at http://www.pb.ethz.ch/downloads. For better user convenience, the R script can not only be used for processed data but accepts cel files as input that can be processed with either GCRMA or vsn. Additional material, including raw data used for this analysis, is available from the authors upon request.

Partitioning and pooling of variances

We used a data set of 380 experiments consisting of duplicated and triplicated measurements. For every probeset, j, m_{ij} and $s_{ij}{}^2$ denote the sample mean and sample variance of the signal intensity in the ith experiment consisting of n_i replicates, respectively. We used a recursive partitioning algorithm from the *rpart* R package (Breiman *et al.*, 1984) to partition the variances $s_{ij}{}^2$ according to m_{ij}, with the minbucket option set to 50, i.e. constructing bins wherein $s_{ij}{}^2$ is modeled to be constant with respect to m_{ij} when varying i. To accommodate varying numbers of replicates n_i, weights w_i were defined as:

$$w_i = (n_i - 1)/\sum(n_k - 1)$$

The pooled sample variance $s_{jpool}{}^2$ was calculated for every probeset j according to:

$$s_{jpool}{}^2 = \sum((n_i - 1) \times s_{ij}^2)/\sum(n_i - 1) = \sum(w_i \times s_{ij}^2)$$

Because observed variances are frequently considerably larger than the pooled variance $s_{jpool}{}^2$, which is based on the mean, and because $s_{ij}{}^2$ were often not normally distributed (not shown), we calculated for every probeset a robust conservative estimate $s_{jpool,q}{}^2$ of the variance as the 80th percentile of all $s_{ji}{}^2$.

Similarly, pooled variances were calculated for every probeset for every bin created by *rpart* as $s_{jpool,b}{}^2$, where $b = 1 \ldots B$, and B is the total number of bins.

The external estimation of variance algorithm

In the commonly used *t*-test, both sample mean and variance are estimated from the data. Because the variance is used in the denominator when calculating the *t*-statistic, incorrect estimates of the sample variance can easily inflate the *t*-statistic. In contrast, incorrect estimates of the sample mean affect the *t*-statistic much less (Appendix S1). In contrast to the *t*-test, the *Z*-test relies on known sample variances obtained from other sources. Because the estimation of the variance in the *t*-test depends heavily on sample size, the *t*-test and the *Z*-test perform similarly with large sample sizes, but the *Z*-test is much more powerful than the *t*-test for small sample sizes. Summarizing, in EVE, for every probeset these steps are performed:

(i) calculate the sample means T and C from the replicated measurements for treatment and control;

(ii) use these means to extract two conservative probeset-specific variances $s_{jpool,bT}{}^2$, $s_{jpool,bC}{}^2$ as above, one for treatment with T in b_T and one for control with C in b_C;

(iii) calculate the Z statistic

$$Z = (T - C)/\sigma_{T-C}$$

where $\sigma_{T-C}{}^2 = \sigma^2\,(1/N + 1/M)$, with N and M the numbers of replicates for treatment and control, respectively [the value σ^2 is conservatively estimated as $\max(s_{jpool,bT}{}^2, s_{jpool,bC}{}^2)$ from (ii)];

(iv) determine a P-value (Z is (0,1) normal).

Subsequently, multiple-testing correction (FDR) is performed using *p.adjust* from the R package *stats*. The Benjamini–Hochberg procedure for controlling the FDR is not guaranteed to be valid under arbitrary dependence among tests, but some theoretical justifications for certain dependencies have been worked out (Benjamini and Yekutieli, 2001).

Simulation study

First, control signal intensities were simulated for each probeset i by drawing n times from $N(\mu_i,\sigma_i{}^2)$, where the signal mean μ_i is drawn from $U\,(1 \ldots 16)$ (the range of signal intensities returned by GCRMA), $\sigma_i{}^2$ is the variance tabulated for probeset i and signal mean μ_i, and with $n = 2 \ldots 10$ in nine independent analyses. Second, treatment signal intensities were simulated for each probeset i by drawing n times from $N(\mu_i + \mathrm{F},\sigma_i{}^2)$, where $F > 0$ for 1000 randomly selected probesets and $F = 0$ for the remaining probesets. In four independent simulations, F was 0.4, 0.6, 1 or 2, respectively, corresponding to fold changes of 1.3, 1.5, 2 and 4. Next, the simulated data were analyzed with the *t*-test and EVE. Receiver–operator curves (ROC) were constructed to display sensitivity versus specificity.

Assignment of genes to functional categories

The means of the variances observed for each probeset were determined, and the 500 probesets with the largest mean variance and the 500 probesets with the smallest mean variance were selected. Probed genes from both sets were grouped into collapsed functional gene ontology categories (obtained from http://www.arabidopsis.org). The significance of enrichment was estimated based on the hypergeometric test and multiple testing corrections according to the method described by Benjamini and Hochberg (1995).

Acknowledgements

We are grateful to the AtGenExpress consortium for the AtGenExpress data set. We thank Drs D. Goldstein and

M. Delorenzi (EFPL Lausanne) for GCRMA processing of the AtGenExpress data set, and Dr M. Mächler (ETH Zurich) for computational help. A.W. has been supported by the SEP program at ETH Zurich under Project TH-8/02-2. This research was supported by ETH Project grant TH-16/05-2 to L.H.

Supplementary material

The following supplementary material is available for this article online:

Table S1. A list of the CEL-files that were used to determine tabulated variances.

Appendix S1. Code and example data to use the EVE algorithm.

This material is available as part of the online article from http://www.blackwell-synergy.com

References

Altmann, T., Weigel, D. and Nover, L. (2004) AtGenExpress – Ein multinational koordiniertes Programm zur Erforschung des Arabidopsis Transkriptoms. *GenomXpress*, **3**, 13–14.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. [B]*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188.

Brazma, A., Hingamp, P., Quackenbush, J. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* **29**, 365–371.

Brazma, A., Parkinson, H., Sarkans, U. *et al.* (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Chubb, J.R., Trcek, T., Shenoy, S.M. and Singer, R.H. (2006) Transcriptional pulsing of a developmental gene. *Curr. Biol.* **16**, 1018–1025.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCarrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.* **32**, D575–D577.

Cui, X. and Churchill, G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210.1–210.10.

Efron, N., Tibshirani, R., Storey, J. and Tusher, V.G. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160.

Hennig, L., Menges, M., Murray, J.A.H. and Gruissem, W. (2003) Arabidopsis transcript profiling on Affymetrix genechip arrays. *Plant Mol. Biol.* **53**, 457–465.

Huber, W., Heydebreck, A.V., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

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.

Kendziorski, C.M., Newton, M.A., Lan, H. and Gould, M.N. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.* **22**, 3899–3914.

Kim, R.D. and Park, P.J. (2004) Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol.* **5**, R70.1–R70.10.

Liu, W.M., Mei, R., Di, X. *et al.* (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L. and Weissman, J.S. (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*, **441**, 840–846.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* **8**, 37–52.

Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* **32**(Suppl.), 496–501.

Redman, J.C., Haas, B.J., Tanimoto, G. and Town, C.D. (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J.* **38**, 545–561.

Rocke, D.M. and Durbin, B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.* **8**, 557–569.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.

Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 1–26.

Smyth, G.K., Yang, Y.H. and Speed, T. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.* **224**, 111–136.

Stoughton, R.B. (2005) Applications of DNA microarrays in biology. *Annu. Rev. Biochem.* **74**, 53–82.

Therneau, T.M. and Atkinson, E.J. (1997) *An Introduction to Recursive Partioning Using the rpart Routines. Technical Report 61.* Rochester, NY: Department of Health Science Research, Mayo Clinic.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T., Gruissem, W., Van de Peer, Y., Inze, D. and De Veylder, L. (2005) Genome-wide identification of potential plant E2F target genes. *Plant Physiol.* **139**, 316–328.

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. and Spencer, F. (2004) *A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Technical Report*. Baltimore, MD: John Hopkins University, Department of Biostatistics.

Zien, A., Fluck, J., Zimmer, R. and Lengauer, T. (2003) Microarrays: how many do you need? *J. Comput. Biol.* **10**, 653–667.

Zimmermann, P., Hennig, L. and Gruissem, W. (2005) Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci.* **10**, 407–409.

Zimmermann, P., Schildknecht, B., Craigon, D. *et al.* (2006) MIAME/Plant – adding value to plant microarrray experiments. *Plant Methods*, **2**, 1.1–1.3.