Genomics & Informatics Vol. 9(1) 28-36, March 2011

Finding Interesting Genes Using Reliability in Various Gene Expression Models

Eun-Kyung Lee[1]*, Dianne Cook[2] and Heike Hoffman[2]

[1]Department of Statistics, Ewha Womans University, Seoul 120-750, Korea, [2]Department of Statistics, Iowa State University, Ames, IA, USA

Abstract

Most statistical methods for finding interesting genes are focusing on the summary values with large fold-changes or large variations. Very few methods consider the probe level data. We developed a new measure to detect reliability that incorporates the probe level data. This reliability measure is useful for exploring the microarray data without ignoring the probe level data. It is easy to calculate, and it can be used for all the other statistical methods as a good guideline to find real differentially expressed genes. Instead of filtering out genes before the analysis, we use whole genes in the analysis and make decisions with new reliability measures.

Keywords: microarray, quality control, reliability, probe level analysis, Affymetrix

Introduction

In Affymetrix microarrays, the expression of each gene is measured by comparing the hybridization of the sample mRNA to a set of probes. A probe is composed of 11-20 pairs of oligonucleotides, each 25 base pairs in length. The first type of probe in each pair is known as a perfect match (PM), which is taken from the target gene sequence. The second type is known as a mismatch (MM), created by changing the middle (13th) base of the PM sequence. The purpose of measuring MM values is to control for nonspecific binding of mRNA, but the actual use of MM values has become controversial (Lazaridis *et al*., 2001).

A gene expression experiment is usually conducted with multiple arrays, also called chips, to compare gene expression across different treatments (genotype, cofactor, replicate). Each chip is prepared by labeling it with a fluorescent dye and hybridizing it to an array. The arrays are scanned into images, which are numerically processed to obtain fluorescence intensity values for each PM and MM sequence. Gene expression microarrays are powerful bioinformatic tools, but the variability arising throughout the measurement process can obscure the biological signals of interest. Quantification of the measurement error helps to extract significant biological signals.

The chip-to-chip variability is controlled by normalization, which yields the same distribution of PM and MM values for each chip. (See Bolstad *et al*., (2003) for a comparison of methods.) After normalization, the values from a probe set are summarized into a single gene expression measure, quantifying the gene activity. Common approaches are the average difference, the model-based expression index (Li and Wong, 2001), the MAS 5.0 algorithm from Affymetrix, and the robust multi-chip average (rma) (Irizarry *et al*., 2003). Fig. 1 shows profile plots for two genes, 265670_s_a and 249093_at, each having 11 PM values. The rma summary expression value is a solid line, and the PM values are dashed lines. The horizontal axis shows the numbers for the eight chips, organized into three genotypes (A, B, C) and their replicates, and the vertical axis shows the log-scale expression values. The gene 265670_s_a has much less variability in PM value than the gene 249093_at, and thus, it is considered to have a more reliable measure of gene expression.

Commonly used models for finding differentially expressed genes are ANOVA (Churchill, 2004) and the HLM for one-way ANOVA (Chu *et al*., 2002). These models compare the expression values across experimental treatments for each gene. Genes that have differential expression on different treatments, relative to its variance over all chips, will have a low p-value. Both of these methods use the summary expression value. In practice, with our data we have found that they also yield somewhat unsatisfactory results. Many genes on the resulting lists have relatively flat, uninteresting profiles, and yet, many genes filtered from the list have relatively structured, interesting profiles. This has motivated us to consider using the PM value in addition to the summary expression value to evaluate the significance of genes.

Table 1 and Table 2 show ANOVA results for two sample genes, 265670_s_at and 249093_at. The p-val-

*Corresponding author: E-mail lee.eunk@ewha.ac.kr
Tel +82-2-3277-6857, Fax +82-2-3277-3607
Accepted 2 March 2011

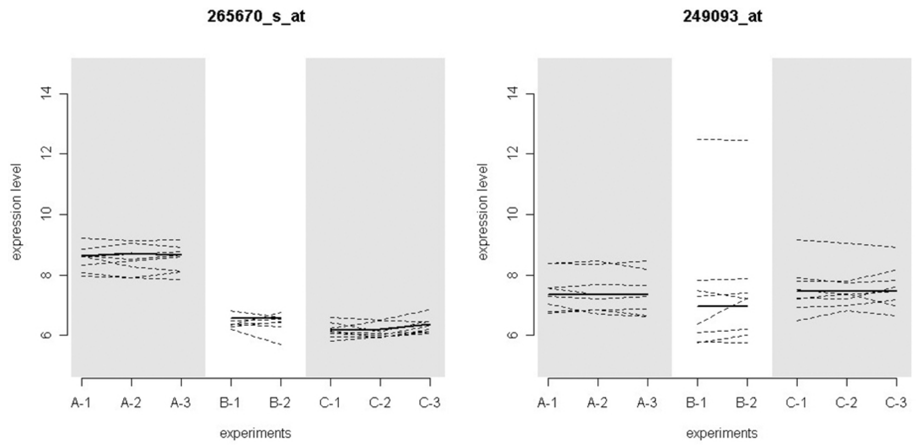

Fig. 1. The expression values of gene 265670_s_at and gene 249093_at. The x-axis represents the chips (arrays), and the y-axis represents the expression values. The first three chips are the replicates of A, and the area is shaded. The next two chips are the replicates of B, and the others are the replicates of C; this area is also shaded. The solid lines in each treatment represent the RMA summary values, and all values in the same treatment are connected. The dotted lines represent the normalized PM values.

Table 1. ANOVA Table for 265670_s_at

Source	df	MS	F	p-value
Treatment	2	4.9763	922.9	$3.79 * 10^{-7}$
Error	5	0.0054		

Table 2. ANOVA Table for 249093_at

Source	df	MS	F	p-value
Treatment	2	0.1653	2053.2	$5.16 * 10^{-8}$
Error	5	0.0001		

ues for both genes are very small, less than 10^{-7}, suggesting that both genes have significantly different expression between treatments. A closer look, however, is revealing. The gene 265670_s_at has a very large mean square treatment effect, 4.9763, and a small mean square error, resulting in a high F value and low p-value. In contrast, the gene 249093_at has a small mean square treatment effect, 0.1653, and a very small mean square error effect, 0.0001, which also yields a high F value and low p-value. The profiles of these two genes are also plotted in Fig. 1. Although both genes are significant according to the ANOVA table, the genes are not alike. One gene has very flat profiles, which makes it less interesting than the other gene.

The two genes have another difference. The profiles of the PM values are also shown in Fig. 1. The gene 249093_at has a lot more variability in PM values than 265670_s_at. The variability in the PM value should be incorporated into a model for detecting significant expression differences.

Chu *et al.* (2002) proposed using the HLM for one-way ANOVA, which is a classical linear mixed model, incorporating probe-level data, for modeling expression. To fit the model, they use standard maximum likelihood (ML) methods using the Proc Mixed procedure in SAS and apply it for each gene. They also estimate the summary values from this model and use them for further analysis, like clustering. However, ML fitting for the model might be too sensitive to outliers. Fig. 2 shows the difference between RMA summary values and summary values from the HLM model for two genes, 265670_s_at (■) and 249093_at (○). For 265670_s_at, the estimated values from the HLM model are similar to the RMA summary values. However, the gene 249093_at, which has more varied PM values, shows quite different estimated values from the HLM model compared to the RMA summary values. This is due to the variation in the PM values.

In this paper, we define the reliability of a summary gene expression value using the variation of probe-level data and show how to use this reliability measure with models for gene expression analysis and associated variance modifications, such as SAM (Tusher *et al.*, 2001) and eBayes (Smyth, 2004). Reliability quantifies the uncertainty in gene expression measurements that can assist in filtering out genes with too much variability in favor of genes that have consistent measurements and significant expression.

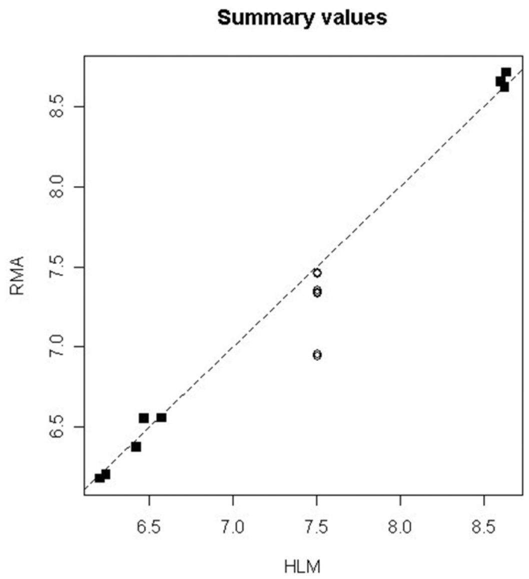

Fig. 2. The black squares correspond to the gene 265670_s_at, and the empty circles correspond to the gene 249093_at. The x-axis represents the expression values estimated from the HLM, and the y-axis represents the RMA expression values. The dashed line is y=x. For 265670_s_at, the estimated values from the HLM are similar to the RMA summary values. However, the gene 249093_at shows quite different estimated values from the mixed model compared to the RMA summary values.

Methods

We start from the 2-level hierarchical linear model (HLM) for one gene. Let be P_{lk} the log$_2$-transformed probe level intensity value of the l-th chip and the k-th oligonucleotide probe, l =1, 2, ⋯, L, and k =1, 2,⋯ , K (usually between 11-20). The chip-to-chip variability is usually controlled by normalizing the log2-transformed probe intensity values using the quantile method. Then, the probe level model is

$$P_{lk} = M_l + \gamma_{lk} \tag{1}$$

where M_l represents the summary expression value of the *l*-th chip and γ_{lk} is the normally distributed random error with a mean of 0 and a probe level variance σ^2. We assume the common probe level variance for each chip. Because we do not know the true value for M_l, we set the level-2 (summary value level) model as follows:

$$M_l = \mu + \varepsilon_l \tag{2}$$

where μ is the overall mean, and ε_l is the random effect associated with the *l*-th chip, assumed to be normally distributed with a mean of 0 and a variance τ^2. We assume that ε_l and γ_{lk} are independent.

Combining (1) and (2) yields the model

$$PM_{lk} = \mu + \varepsilon_l + \gamma_{lk} \tag{3}$$

which is a one-way ANOVA model with a random effect, where μ is the overall mean, ε_l is a summary value level random effect, and γ_{lk} is a probe level random effect. Then, the variance of log$_2$ probe intensities is

$$Var(P_{lm}) = Var(\varepsilon_l + \gamma_{lk}) = \tau^2 + \sigma^2 \tag{4}$$

The σ^2 parameter represents the probe level variability, and τ^2 captures the variability of the summary values. Most models for finding significant genes start from the summary value level model, using the summary values and ignoring the probe level variability. Our interest focuses not only on τ^2 but also on σ^2. If σ^2 is large relative to τ^2, the summary values of the chips might be less reliable.

In the 2-level HLM, the reliability ρ of the estimated value $\hat{M}_l$ is defined as follows (Raudenbush and Bryk, 2002):

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2 / K} \tag{5}$$

where τ^2 represents the variation of the summary value M_l, and σ^2 represents the variation of log$_2$ probe intensity values PM_{lm}. Therefore, this reliability can be written as:

$$\rho = \frac{VAR(M_l)}{VAR(M_l) + VAR(PM_{lm}) / K} \tag{6}$$

This measure is defined from the structure of HLM and represents the reliability of the summary expression values of the specific gene. This population version of the reliability is usually estimated by fitting the hierarchical model, where a simple mean of the log$_2$-transformed probe intensity values is used as the summary value. However, a simple mean of the log$_2$ probe intensities is rarely used for the summary value in microarray data analysis. There are a lot of methods to calculate summary values in robust way (Bolstad *et al*., 2003), including RMA. Recently, (Millenaar *et al*., 2003) compared several calculation methods for summary values. They concluded that the user needs to try several different methods. However, it is usually not easy.

We extend the approach by estimating reliability for any type of summary value method. Let y_l be a summary value of the l-th chip, and $P_{l1}, P_{l2}, \cdots, P_{lK}$ are probe-level data. y_l is calculated from $P_{l1}, P_{l2}, \cdots, P_{lK}$ using one of the summarization methods. Then, our new reliability measure is defined as follows:

$$\rho^* = \frac{\sum_{l=1}^{L} (y_l - \bar{y})^2 / (L-1)}{\sum_{l=1}^{L} (y_l - \bar{y})^2 / (L-1) + \frac{1}{LK(K-1)} \sum_{l=1}^{L} \sum_{k=1}^{K} (P_{lk} - \bar{P}_{l.})^2} \tag{7}$$

This new reliability measure can be calculated easily with several different summary expression values. If RMA is the summary estimation method, y_l is a summary value from RMA (Irizarry *et al.*, 2003) and $P_{l1}, P_{l2}, \cdots, P_{lK}$ are log$_2$ PM values. If MBEI (Li and Wong, 2001) values are used, y_l is a summary value from the dchip method and $P_{l1}, P_{l2}, \cdots, P_{lK}$ are log$_2$ PM - log$_2$ MM values.

This new reliability statistic, ρ^*, is easy to calculate, and it can be applied for a variety of summary value methods. The population version of reliability ρ in the equation (6) is only for the 2-level HLM model and can only be used for $\hat{M}_l$. On the other hand, the reliability statistic ρ^* can be used for diagnostics in models with various summary values. Therefore, it is possible to compare various methods to calculate summary values using this measure.

There is no general rule to decide whether the reliability is high or not. We suggest a guideline that helps in decision-making. The definition of ρ^* can be re-stated as follows:

$$\frac{\mathrm{var}(probe\ level\ data)}{\mathrm{var}(summary\ values)} = \frac{K(1-\rho^*)}{\rho^*} \qquad (8)$$

where K is a constant. Fig. 3 shows the relationship between ρ^* and this ratio of variances. If the ratio is less than or equal to 1, the reliability ρ^* will be larger than 0.92. If we allow this ratio up to 2, ρ^* should be larger than 0.85. Therefore, if the ratio of variances between probe-level data and summary values is less than or equal to 1, we expect to have a reliability greater than 0.92. If the reliability is less than 0.92, in this case, we

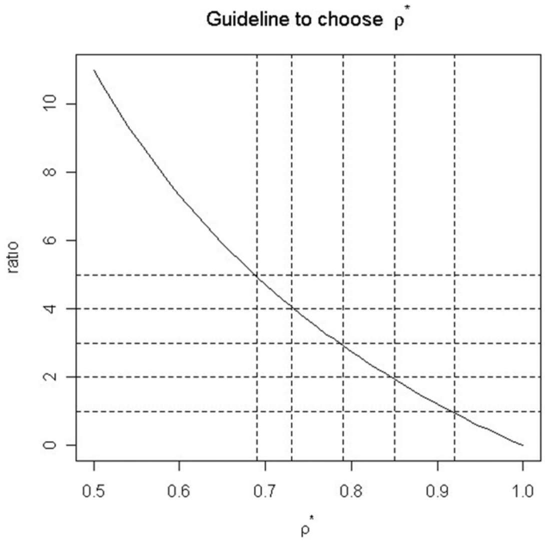

Fig. 3. Ratio of probe variance to summary expression variance plotted against reliability, ρ^*. Reliability larger than 0.92 corresponds to a probe variance less than or equal to the summary expression variance.

Table 3. This table is the top 15 genes that have the smallest p-value of the ANOVA test. The numbers under each method name is the rank. A/C and B/C represent the fold-changes between two treatments

Affy ID	Locus ID	GO.Function	Rel.*	eBayes	SAM	ANOVA	HLM	A/C	B/C
246785_at	AT5G27380	Molecular function unknown	0.846	129	26	1	985	-0.203	-1.579
262128_at	AT1G52690		0.993	1	1	2	2	4.415	0.827
259768_at	AT1G29390		0.930	33	12	3	277	-1.694	-2.119
265575_at	AT2G14260	Aminopeptidase activity; catalytic activity; hydrolase activity;	0.624	1162	307	4	5567	0.150	-0.750
262571_at	AT1G15430		0.852	611	152	5	1503	0.810	-0.208
262313_at	AT1G70900		0.915	133	41	6	333	-1.453	-1.176
256648_at	AT3G13580	Structural constituent of ribosome; transcription regulator activity;	0.897	111	44	7	409	1.497	0.216
251428_at	AT3G60140	Hydrolase activity, hydrolyzing O-glycosyl compounds;	0.987	2	2	8	3	4.198	2.620
258347_at	AT3G17520		0.846	757	224	9	1032	0.883	0.070
266897_at	AT2G45820	DNA binding;	0.921	97	38	10	465	0.398	-1.418
{bf 262605_at}	AT1G15170	Antiporter activity; drug transporter activity; transporter activity;	0.166	6977	2814	11	14802	-0.364	-0.274
255046_at	AT4G09650	Hydrogen-transporting ATP synthase activity, rotational mechanism;	0.927	12	7	12	202	-1.605	-2.917
251714_at	AT3G56370	ATP binding; kinase activity; protein serine/threonine kinase activity;	0.810	1158	339	13	1353	-0.678	-0.841
261279_at	AT1G05850	Chitinase activity	0.958	30	15	14	64	-0.519	-2.455
253606_at	AT4G30530	Catalytic activity	0.928	24	14	15	204	-1.237	-2.583

can not believe the result from the ANOVA.

Results and Discussion

The experiment is a completely randomized design, containing one treatment having 3 levels - A, B, C - and each is replicated:

Treatment	Replicates
A	Chip1, Chip2, Chip3
B	Chip4, Chip5
C	Chip6, Chip7, Chip8

The data were recorded on the new Affymetrix GeneChip Arabidopsis Genome Array. The PM values were normalized using the quantile method, and RMA was used to determine the summary expression value. We compare the gene lists resulting from the eBayes (Smyth, 2004) and SAM (Tusher *et al.*, 2001) methods, classical ANOVA, and HLM for one-way ANOVA, along with reliability. Table 3 shows the results ordered by the top 15 genes from classical ANOVA. Table 4 shows the results ordered by eBayes. The tables display the ranks of the genes on the list according to the different methods. Table 5 lists genes that have good reliability and small p-values that are unfortunately prone to being filtered off a list of interesting genes. Table 6 lists genes that have a small p-value but also poor reliability that perhaps should be filtered off a list of interesting genes.

By classical ANOVA (Table 3), the most significantly differentially expressed gene, 246785_at, is ranked as low as 129 by eBayes, 26 in the SAM method, and 985

Table 4. This table is the top 15 genes that have smallest p-value of the eBayes method. The numbers under each method name is the rank. A/C and B/C represent the fold-changes between two treatments

Affy ID	Locus ID	GO.Function	Rel.*	eBayes	SAM	ANOVA	HLM	A/C	B/C
262128_at	AT1G52690	Molecular function unknown	0.993	1	1	2	2	4.415	0.827
251428_at	AT3G60140	Hydrolase activity, hydrolyzing O-glycosyl compounds	0.987	2	2	8	3	4.198	2.620
254098_at	AT4G25100	Iron superoxide dismutase activity	0.995	3	3	27	1	-0.774	-4.819
264514_at	AT1G09500	Alcohol dehydrogenase activity; cinnamyl-alcohol dehydrogenase activity	0.970	4	4	19	17	3.515	1.609
261309_at	AT1G48600	S-adenosylmethionine-dependent methyltransferase activity	0.977	5	6	28	19	-2.781	-3.808
262047_at	AT1G80160	lactoylglutathione lyase activity	0.984	6	5	24	5	3.396	2.313
257315_at	AT3G30775	Proline dehydrogenase activity	0.971	7	9	53	14	4.270	2.250
251438_s_at	AT5G33355	Molecular function unknown	0.975	8	11	32	7	3.406	2.118
252984_at	AT4G37990	Aryl-alcohol dehydrogenase activity; mannitol dehydrogenase activity	0.947	9	8	25	85	3.170	1.562
246114_at	AT5G20250	Hydrolase activity, hydrolyzing O-glycosyl compounds	0.965	10	10	30	26	3.259	1.739
264524_at	AT1G10070	Branched-chain-amino-acid transaminase activity; catalytic activity	0.949	11	16	52	71	3.282	1.811
255046_at	AT4G09650	Hydrogen-transporting ATP synthase activity, rotational mechanism	0.927	12	7	12	202	-1.605	-2.917
245148_at	AT2G45220	Enzyme inhibitor activity; pectinesterase activity	0.952	13	13	33	47	3.013	1.141
264580_at	AT1G05340	Molecular function unknown	0.989	14	20	107	4	4.420	1.011
267002_s_at	AT2G34430	Chlorophyll binding; O-acetyltransferase activity; chlorophyll binding	0.917	15	21	76	161	-2.300	-3.862

Table 5. The list of 5 genes that are filtered out using small fold-changes but are highly reliable and highly significant. The numbers under each method name is the rank. A/C and B/C represent the fold-changes between two treatments

Affy ID	Locus ID	GO.Function	Rel.*	eBayes	SAM	ANOVA	HLM	A/C	B/C
262781_s_at	AT3G26340	Endopeptidase activity; threonine endopeptidase activity	0.926	820	802	825	478	0.416	-0.911
254578_at	AT4G19410	Carboxylic ester hydrolase activity	0.918	784	816	964	380	0.548	-0.875
252591_at	AT3G45600	Molecular function unknown	0.917	811	784	784	540	0.361	-0.953
261872_s_at	AT4G21660	Molecular function unknown	0.917	554	639	946	518	0.867	-0.838
260705_at	AT1G32400	Protein binding	0.905	814	822	910	570	0.830	-0.489

Table 6. The list of 5 genes that are highly significant but have very low reliability. The numbers under each method name is the rank. A/C and B/C represent the fold-changes between two treatments

Affy ID	Locus ID	GO.Function	Rel.*	eBayes	SAM	ANOVA	HLM	A/C	B/C
253743_at	AT4G28940	Undetermined	0.504	725	612	442	5958	-0.111	1.065
263048_s_at	AT2G05310	Chloroplast	0.598	789	810	927	9899	-0.427	-1.391
252024_at	AT3G52880	Cytosol; undetermined	0.622	797	754	730	5085	0.284	-1.005
260917_at	AT1G02700	Mitochondrion	0.507	939	698	403	6830	-0.197	0.893
252325_at	AT3G48560	Cytosol; chloroplast	0.589	957	947	973	6810	0.165	-1.058

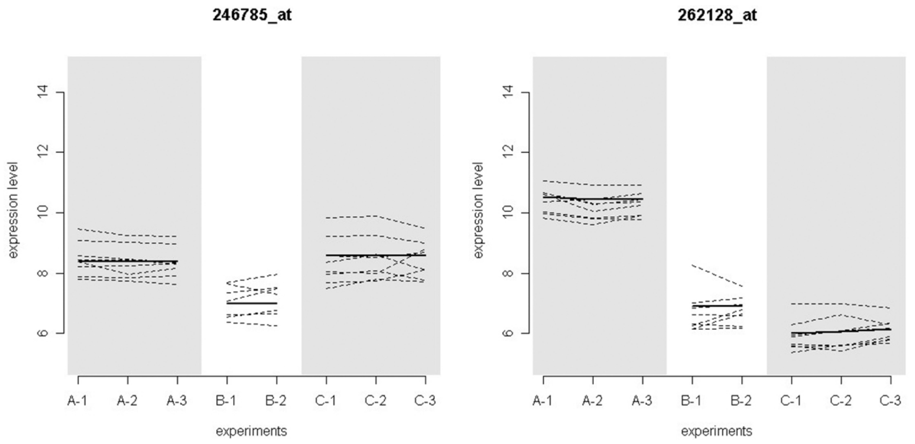

Fig. 4. Profiles of summary expression and PM values for genes ranked highly by ANOVA in Table 3. Two genes, 246785_at (left) and 262128_at (right), are ranked 1 and 2, respectively, on the ANOVA list. 262128_at has quite consistent PM values and big differences between the treatments, which clearly makes it an interesting gene. 246785_at has slightly less consistency and less difference between the treatments, which makes it really less interesting than the number 2 ranked gene.

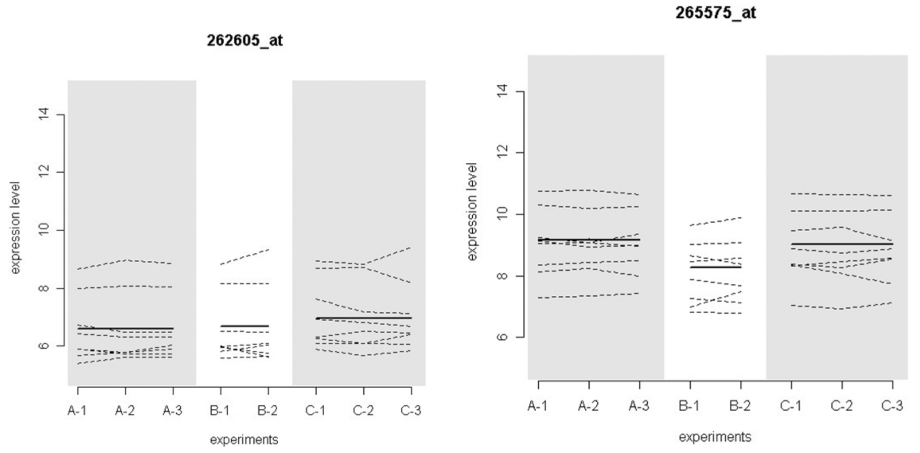

Fig. 5. Profiles of summary expression and PM values for genes ranked highly but having low reliability in Table 3: 262605_at (left) and 265575_at (right) have large variability in PM values. They have very low reliability, and yet they are ranked in the top 15 by ANOVA with p-values less than 1.0-8; the profiles suggest that these two genes are not at all interesting.

by HLM, and the reliability of this gene is less than perfect at 0.846. The methods disagree on the significance of this gene. The second-ranked gene, 262128_at, has a good reliability (0.993) and is ranked highly by eBayes, SAM, and HLM; so, the methods all agree that this is an interesting gene. Fig. 4 shows the profiles of PM values for these genes. The difference in variability of the PM values supports the position that 262128_at is undoubtedly an interesting gene but that 246785_at is much less so.

The worst genes on this list (Table 3) are 262605_at and 265575_at, with reliabilities of 0.166 and 0.624, respectively. Fig. 5 shows the profiles of PM values for these genes. There is a lot of variability in the PM values. Although they are at the top of the ANOVA model's list of significant genes, they should not be on a final list of interesting genes.

Table 4 shows that there is considerable agreement about genes between eBayes and SAM. Both methods adjust their test statistics by shrinking an individual gene's variance estimate towards the variance of all genes. All of the differentially expressed genes on this list also have high reliability. The HLM disagrees with eBayes and SAM about the interestingness of several of the genes.

A comparison of all of the genes is possible using brushing in a scatterplot matrix (Becker and Cleveland, 1988). Fig. 6 is the screenshot of the brushing being done with GGobi (Swayne *et al.*, 2003). The scatterplot matrix has 5 diagnostic statistics: reliability, eBayes F values, ANOVA F values, SAM F values, and HLM χ^2 values. (Large test statistics, F, χ^2 correspond to small p-values.) Each point in this plot represents one gene, with its corresponding values on the diagnostics plotted against the appropriate axis.

The diagonal elements are density plots of each diagnostic. The eBayes, ANOVA, SAM, and HLM test statistics are all strongly left-skewed, which says that there are a lot more small values of these statistics and few high (important) values. Reliability has an unusual distribution - fewer genes have high or low reliability and many more genes are in the mid-range of the scale.

The off-diagonal plots in the scatterplot matrix are the (5* 4/2=10) pairwise plots of the 5 diagnostics, which shows how the statistics are related to each other. The values for eBayes and SAM are very strongly related, with a slight non-linear relationship. eBayes and SAM have a reasonably close association with ANOVA, with the exception of a relatively small number of the genes. eBayes and SAM have a moderate association with HLM - for most genes, the HLM value seems to be the same as the eBayes or SAM value plus a constant. ANOVA has almost no relationship with HLM. Reliability has a much different relationship with each of the other

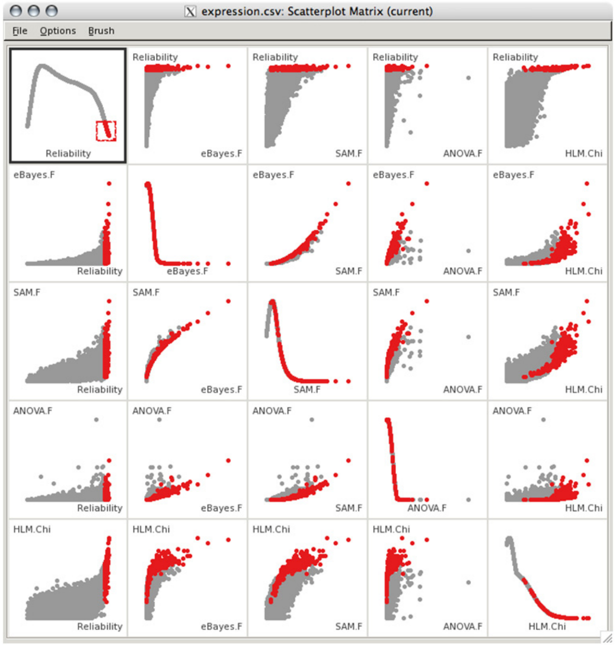

Fig. 6. The scatter plot matrix from GGobi. Red highlighted genes have high reliability ones. If they also have high values in the analysis, they must be truly differentially expressed ones.

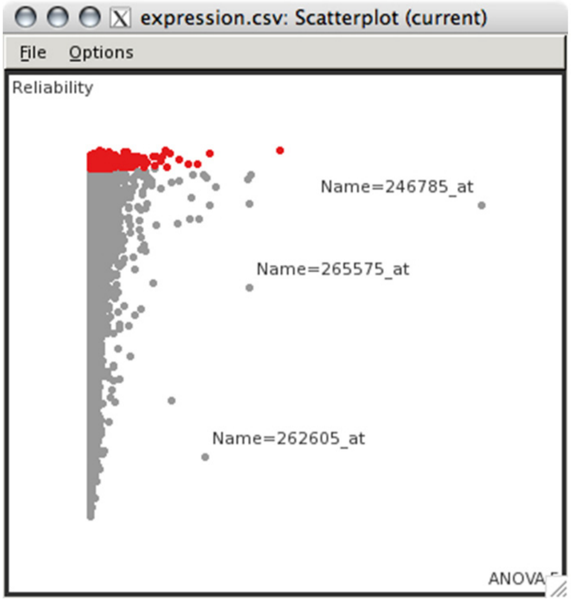

Fig. 7. The scatter plot matrix of reliability and ANOVA F statistics in GGobi. Highlighted genes (red) have high reliability ones. Genes with high a ANOVA rating and low reliability are identified as 246785_at, 265575_at, and 262605_at.

diagnostics. Reliability and eBayes have the strongest association, where genes with high test statistic values also generally have high reliability. Many genes with high reliability have low test statistic values. This association is similar for SAM and HLM but not so for ANOVA; there are several genes that have high ANOVA test statistics but relatively low reliability.

Brushing on the scatterplot matrix is done in the density plot of reliability. Genes with the highest reliability are brushed (red). These same genes are colored (red) in the other plots. The values that these genes have on the other diagnostics is interesting; they cover the full range of F statistic values for eBayes, SAM, and ANOVA but moderate to high values of χ^2 for HLM. HLM uses the PM values, on which reliability is calculated, in the model, while the methods do not; so, there should be some similarity between the diagnostic values. That there are some differences must be due to the use of non-robust estimates of the summary values. Comparing eBayes and ANOVA, the high-reliability genes have a very strong relationship, and there are many genes rated highly by ANOVA that have low reliability. The most interesting feature of the diagnostics revealed by the scatterplot matrix is that there are three outliers, which are the three genes that eBayes, SAM, and HLM all rate very highly: 262128_at, 251428_at, and 254098_at. There is also an outlier in the values of ANOVA, which have low reliability, and this corresponds to the gene that ANOVA rates very highly, counter to all of the other diagnostics: 246785_at. We also see several other outliers in the ANOVA and Reliability plot that have low reliability, two of which correspond to 265575_at and 262605_at (Fig. 7).

From this analysis, it looks like eBayes, SAM, and HLM all do a reasonable job in finding significantly differentially expressed genes. However, they can miss a few genes that perhaps should be detected, and they can also accept a few genes that perhaps should not be detected. Reliability helps to uncover these.

Table 5 lists 5 genes that have high reliability (>0.9) and relatively high ranks (<1,000). But, they also have small fold-changes that might exclude them if filtering is conducted before the analysis. Profiles for two of these genes are shown in Fig. 8. These genes are characterized by small differences in the summary expression values for each treatment but also by very consistent PM values. Based on the PM values, to which eBayes and SAM are blind, it could very well be argued that these are significantly differentially expressed genes.

Table 6 lists 5 genes that have very low reliability (<0.65) and relatively high ranks (<1,000). They would be included in the significantly differently expressed genes, but the reliability of these genes is very low, casting doubt on the interestingness of these genes. Based on the PM values, to which eBayes and SAM are blind, it could very well be argued that they are not significantly differentially expressed genes.

In this paper, we defined a new reliability statistic that can be used with any other statistical method to find differentially expressed genes. This measure is very easy to calculate, relative to applying the eBayes and SAM methods. Therefore, it can be used easily for the exploratory data analysis step as well as the intensive data analysis step. It also provides a guide to the variance of each gene's summary expression value. This statistic can also be used with the other statistical

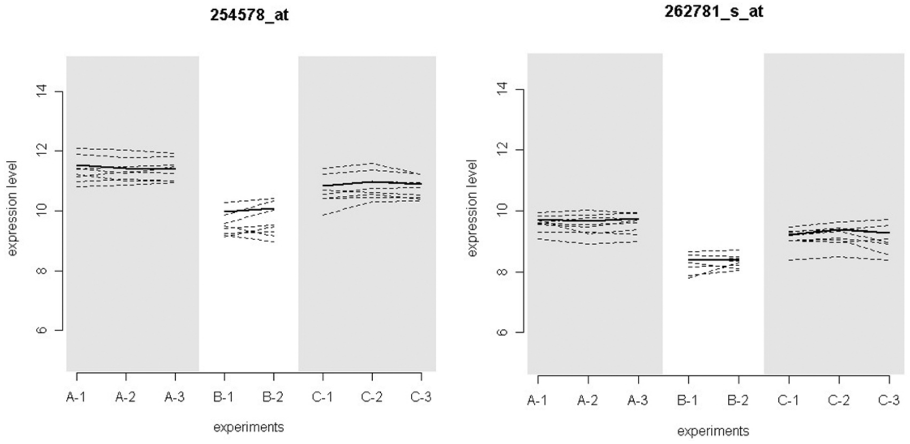

Fig. 8. Profiles of a special group of genes that have high reliability but would be excluded from a list of significantly differentially expressed genes by all methods. The profiles show that these genes may be important. They have small differences in the summary expression values for each treatment but also very consistent PM values.

methods, like clustering and classification. The R package ProbeR for calculating and exploring reliability measure with GGobi is available at CRAN (http://www.r-project.org).

Acknowledgments

We thank Steve Rodermel for the data. This work was supported by the Priority Research Centers Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (2009-0093827) and National Science Foundation Arabidopsis 2010 grants DBI-0209809 and DBI-0520267.

References

Becker, R.A. and Cleveland, W.S. (1988). Brushing Scatter-plots, in Cleveland, W. S. and McGill, M. E. (eds) Dynamic Graphics for Statistics, Wadsworth, Monterey, CA. 201-224.

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

Chu, T.M., Weir, B., and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math. Biosci.* 176, 35-51.

Churchill, G.A. (2004). Using ANOVA to analyze microarray data. *Biotechniques* 37, 173-175.

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.

Lazaridis, E.N., Sinibaldi, D., Bloom, G., Mane, S., and Jove, R. (2001). A simple method to improve probe set estimates from oligonucleotide arrays. *Math. Biosci.* 176, 53-58.

Li, C., and Wong W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 176, 53-58.

Millenaar, F.F., Okyere, J., May, S.T., Zanten, M., Voesenek, L.A., and Peeters, A. (2006). How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics* 7, 137.

Raudenbush, S.W., and Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA:Sage Publications.

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

Swayne, D.F., Lang, D.T., Buja, A., and Cook, D. (2003). GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics and Data Analysis* 43, 423-444. http://www.ggobi.org.

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.