BMC Genomics

BioMed Central

Methodology article

Open Access

Tests for differential gene expression using weights in oligonucleotide microarray experiments

Pingzhao Hu[1], Joseph Beyene[2,3] and Celia MT Greenwood*[1,2]

Address: [1]Program in Genetics and Genomic Biology, The Hospital for Sick Children Research Institute, 15-706 TMDT, 101 College Street, Toronto, ON, M5G 1L7, Canada, [2]Department of Public Health Sciences, University of Toronto, Health Sciences Building, 155 College St, Toronto, ON, M5T 3M7, Canada and [3]Program in Population Health Sciences, The Hospital for Sick Children Research Institute, 555 University Ave, Toronto, ON, M5G 1X8, Canada

Email: Pingzhao Hu - phu@sickkids.ca; Joseph Beyene - joseph@utstat.toronto.edu; Celia MT Greenwood* - celia.greenwood@utoronto.ca

* Corresponding author

Published: 22 February 2006

*BMC Genomics*2006, **7**:33 doi:10.1186/1471-2164-7-33

Received: 01 September 2005
Accepted: 22 February 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/33

Abstract

Background: Microarray data analysts commonly filter out genes based on a number of ad hoc criteria prior to any high-level statistical analysis. Such ad hoc approaches could lead to conflicting conclusions with no clear guidance as to which method is most likely to be reproducible. Furthermore, the number of tests performed with concomitant inflation in type I error also plagues the statistical analysis of microarray data, since the number of tested quantities in a study significantly affects the family-wise error rate. It would, therefore, be very useful to develop and adopt strategies that allow quantification of the quality of each probeset, to filter out or give little credence to low-quality or unexpressed probesets, and to incorporate these strategies into gene selection within a multiple testing framework.

Results: We have proposed a unified scheme for filtering and gene selection. For Affymetrix gene expression microarrays, we developed new methods for measuring the reliability of a particular probeset in a single array, and we used these to develop measures for a set of arrays. These measures are then used as weights in standard t-statistic calculations, and are incorporated into the multiple testing procedures. We demonstrated the advantages of our methods using simulated data, publicly available spiked-in data as well as data comparing normal muscle to muscle from patients with Duchenne muscular dystrophy (DMD), in which a set of truly differentially expressed genes is known.

Conclusion: Our quality measures provide convenient ways to search for individual genes of high quality. The quality weighting strategies we proposed for testing differential gene expression have demonstrable improvement on the traditional filtering methods, the standard t-statistic and a regularized t-statistic in Affymetrix data analysis.

Background

Affymetrix GeneChip™ microarrays are used to measure gene expression for thousands of transcripts simultaneously. Each transcript is measured by 11–20 probesets, where a probeset consists of two almost identical sequences of length 25 bp. One member of the pair is the perfect match (PM) probe, where the sequence is the exact complement of a section of the mRNA. The other member

of the pair, the mismatch probe (MM), is identical to the PM probe except for the nucleotide in 13th position and is intended to measure and control for non-specific binding signals. The accuracy and sensitivity of the measurement for any one gene or EST depends on the uniqueness and the binding properties of the probes.

It is well known that most probesets perform consistently and reliably, in that similar estimates of expression are obtained from two replicates of an experiment. However, in many cases, the signals from a probeset can be hard to interpret. There may be substantial variability across the probesets in the estimated level of expression, or in the PM-MM differences. PM signals can be smaller than MM signals suggesting high levels of non-specific binding. It is also well known that a large proportion of the genes are expressed in only a few tissues or at a particular developmental stage, and hence many of the genes are not expected to have a measurable transcribed product. Su et al. [1] have estimated that in most cases, only 30–40% of the genome will be expressed in any one tissue. In such situations, the probesets give measurements for PM and MM that fluctuate near the lower detectable limit.

The latest release for the Affymetrix GeneChip Expression Analysis Platform, GeneChip Human Genome U133 Plus 2.0 Array [2], provides comprehensive coverage of the transcribed human genome on a single array. The array includes more than 54,000 probesets with 38,500 well-characterized human genes. When analyzing such a large number of genes, the adjustment of significance levels through multiple testing procedures such as the Bonferroni method [3], or Benjamini and Hochberg [4], may be dramatic enough to make it very difficult to identify differentially expressed genes. However, if we knew which transcripts were either not expressed in the tissue under study, or were measured unreliably due to poor probe specificity, we could exclude these transcripts from the analysis and pay a smaller penalty for multiple testing.

In 2001, Affymetrix produced a new algorithm for summarizing the results of a probeset, and the algorithm includes a detection p-value, which represents the probability that the probeset (gene) expression is above zero (i.e., turned on), and measured reliably and consistently across the probe. In addition, "Present/Marginal/Absent" (P/M/A) calls for each transcript are based on the detection p-value together with thresholds that can be altered by the user [5]. The calls are often used as filters to keep genes whose transcripts are detectable in a particular experiment [6-8], and this filtering concept has been integrated into some software such as dChip [9,10]. For example, Blalock, et al. [7] removed a probeset from further statistical analysis if there were > = 60% absent calls, for that probeset, in at least one treatment group. However, the stringency of the filtering procedure can strongly affect downstream analyses and the final results. Too much filtering may exclude some important genes, whereas too little filtering will reduce power by increasing the number of tests performed. Moreover, it is possible that the statistical error introduced by imperfect filtering criteria makes the overall error worse. Pounds and Cheng [11] recently argued that it is better to define gene filters based on the detection p-value than on the calls, so that more of the available information about probeset reliability can be used. They developed two pooled filters for each probeset based on detection p-values.

An alternative strategy to filtering is to treat probe-specific variability as a quality index to measure the reliability of each probeset. Although many summaries of Affymetrix GeneChip expression have been proposed (e.g. MAS5 [5], dChip [9,10], RMA [12], PLIER [13] and gcRMA [14]), there have been few studies on quality measures for Affymetrix GeneChip expression data. In previous studies, typically, once a filtering decision was made, the estimated intensities of each probeset were considered to be equally well measured, so that probesets with highly variable signals were considered as reliable as probesets with inconsistent signals. However, probesets that detect transcripts expressed at a very high level would be expected to show a more robust signal with greater quality than probesets that are performing poorly or detecting very low level transcripts.

Seo et al. [15] used the detection p-values to develop weighted Pearson correlations between expression measurements from two different arrays. The weight was defined to be a function of the two detection p-values for each probeset. They incorporated these weighted correlations into unsupervised clustering analysis, with the goal of choosing the best algorithm for summarizing across the probesets.

In this study, we focus on the problem of testing for differential expression between groups of arrays, while adjusting for multiple testing in a weighted framework. We redefine some of the widely used filtering methods [6-9] and also propose some new methods for measuring the quality of a particular probeset in a single array, in an experimental group of arrays, and in an entire study. These measures are then used as weights in t-statistic calculations, and are incorporated into the multiple testing procedures. We applied these methods to a dataset of expression in muscle tissue [16], Choe's spiked-in data [17] and simulated data. One of our quality measures for an experimental group is based on the measure developed very recently by Pounds and Cheng [11]; however, we go further in combining measures across groups, and in dis-

Table 1: Test statistics and quality measures. For the quality-based tests, the summarized quality measure, across treatment groups, is defined by $Q_g = \max_{w\in\{1\ldots W\}} R_g^w$.

Quality Measures	Array specific measure	Group-specific measure R_g^w	Notes
Q_g^0	1.0	1.0	No filtering
Q_g^{call}	$q^*_{giw} = \begin{cases} 1 & if\ call\ is\ P \\ 0 & otherwise \end{cases}$	$\min_{i\in\{1\ldots I\}} q^*_{giw}$	All arrays must have present call to be included
Q_g^{meanp}	q_{giw} = detection p-value	$1-\bar{q}_{gw} = 1-\sum_{i\in w}(q_{giw})/n_w$	n_w is the number of arrays in group w
$Q_g^{\exp}$	q_{giw} = detection p-value	$\exp\lfloor \lambda_{gw}\log\nu \rfloor$ for sensitivity parameter ν	$\hat{\lambda}_{gw} = -n_w / \sum_{i\in w}\log(q_{giw})$
Q_g^{beta}	q_{giw} = detection p-value	$\nu^{\hat{a}_{gw}}$	$\hat{a}_{gw} = \bar{q}_{gw}/(1-\bar{q}_{gw})$
Q_g^{we}	q_{giw} = detection p-value	N/A	Weighted means $\bar{x}_{gw}$ and weighted standard deviations S_{gw}
Q_g^{LPE}	N/A	N/A	Local pooled error test [18]

cussing the usefulness of these measures when testing for differential gene expression.

Results

Test statistics and quality measures

We propose a weighting paradigm for including quality measures into analysis when testing for differential expression. Suppose that an unweighted test statistic for gene g is represented by t_g, and that the quality measure is called Q_g. We propose to evaluate the significance of gene g using the weighted test statistic $t_g^* = t_g Q_g$. The impact of the weighting on the pattern of results, across all genes, is then taken into account when calculating significance adjusted for multiple testing, or when estimating the false discovery rate. Conceptually, giving a low weight, $Q_g \approx 0$, to a particular gene can be thought of as excluding that gene from consideration. Therefore, our modifications to multiple-testing methods are based on adjusting the number of genes tested, based on the quality measures.

Measure Q_g must represent a summary across treatment groups $w \in \{1 \ldots W\}$. We propose

$$Q_g = \max_{w\in\{1\ldots W\}}[R_g^w], \qquad (1)$$

where R_g^w is a treatment-group specific measure of quality. By using the maximum value across groups in equation (1), a gene that is clearly expressed in one or more groups will be considered as a gene measured with high quality.

We examine the performance of several choices for the group-specific quality measure R_g^w. In particular, R_g^w = 1.0 leads, after using equation (1), to a measure Q_g^0 that is, in fact, no quality filtering: all genes are included in the analysis at full weight. The common practice of analyzing only genes with present calls can be based on { R_g^w = 1.0 when all arrays in group w have present calls, and R_g^w = 0 otherwise}. We will denote this as Q_g^{call}.

Instead of using the Affymetrix present/absent calls, more sensitivity can be gained by basing R_g^w on the actual detection p-values, q_{giw} for gene g, array i, and group w. Since a highly-expressed gene corresponds to a small detection p-value, we propose R_g^w = 1 - $\bar{q}_{gw}$, where $\bar{q}_{gw}$

is the mean detection p-value for group w, leading to a quality measure that we call Q_g^{meanp}. The mean may not give enough weight to small p-values, therefore we also investigated two measures that are based on parametric model assumptions for the distribution of detection p-values. Firstly, we assumed that -log(q_{giw}) follow an exponential distribution with group-specific mean λ_{qw}. Under this distribution, the probability that the detection p-values will be smaller than a threshold v can be written as $R_g^w = \exp(\lambda_{gw} \log v)$, with corresponding $Q_g^{\exp}$ following from equation (1).

Although, under the null hypothesis of no signal, the detection p-values can be expected to follow a uniform distribution (so that -log(q) is exponential with mean 1.0), when there is a signal, the p-value distribution is better described by the two-parameter Beta distribution. However, detection p-values are based on rank tests and there is often little variability in the p-values across arrays for highly expressed genes. This leads to difficulties in estimating two parameters. Pounds and Cheng [11] proposed a one-parameter Beta distribution for detection p-values, so that $R_g^w = p(q_{giw} \le v) = \int_0^v Beta(q; a_{gw})dq = v^{agw}$, and we incorporate this measure into our corresponding across-group quality measure Q_g^{beta}.

For comparison, two very different test statistics were also used: (1) Weights based on the detection p-values were incorporated directly in the calculation of group-specific means and standard deviations – we refer to this test statistic as Q_g^{we} to follow our naming convention, although no quality measure is defined; (2) In addition, we used the local pooled error (LPE) test [18], which is based on pooling errors within genes and between replicate arrays for genes in which expression values are similar. This test is referred to as Q_g^{LPE}. Unweighted multiple testing corrections were implemented for these two tests. Table 1 summarizes the different test statistics and quality measures used in the paper.

Analysis of simulated data

The performance of our proposed methods was evaluated using simulated probe level data generated from a model incorporating probe level effects, optical noise, and non-specific binding, as well as true signals [14]. We have run five simulation models following the simulation procedures described in Materials and Methods. Treatment effects on the signal were varied between 0.5 (very small) and 2.5 (very large) in the five models. To compare performance, we used Summarized Receiver Operating Characteristic (SROC) curves, where the test sensitivities and specificities (true positive and true negative proportions) for a range of p-value cutoffs (or FDR cutoffs for results with multiple testing adjustments) were averaged over 100 simulated datasets. Figures 1 and 2 show SROC curves of models 2 and 4, where large and small treatment effect sizes, respectively, were chosen in the generated models.

The SROC curve's overall behavior can be measured by the Area Under the Curve (AUC) [19]. Table 2 shows AUCs for five simulation models under different weighting and multiple testing strategies. As seen from the table and figures, weighted t-statistics based on quality measure Q_g^{call}, in which a gene was excluded if one or more arrays have an absent call, had the lowest AUC in all models. Whether p-values were unadjusted or adjusted by the weighted Benjamini-Hochberg (WBH) method, $Q_g^{\exp}$ outperformed the other quality scores in models 1, 2, and 3, where a moderate to large treatment effect was chosen. Although Q_g^{meanp} had the overall best performance for models 4 and 5, where a small treatment effect was used, $Q_g^{\exp}$ has the best sensitivity when specificity is larger than 95% (shown in Figure 2). We also observed that $Q_g^{\exp}$ had similar performance to Q_g^{LPE} for models 1, 2 and 3, but $Q_g^{\exp}$ outperformed Q_g^{LPE} for models 4 and 5. Q_g^{beta} has slightly worse performance than $Q_g^{\exp}$ for models 1,2 and 3, but their performance is almost the same for models 4 and 5; conversely, the statistic Q_g^{we} performed better, relative to the other methods, for small effect sizes. The performance of no filtering at all, Q_g^0, was also a good choice for small changes in gene expression. Therefore, the best choice of weighting statistic in the presence of adjustments for multiple testing appears to depend on the size of the treatment effects. We also found that it is better to define quality measures directly from Affymetrix detection p-values rather than from the Affymetrix Present/Marginal/Absent calls.

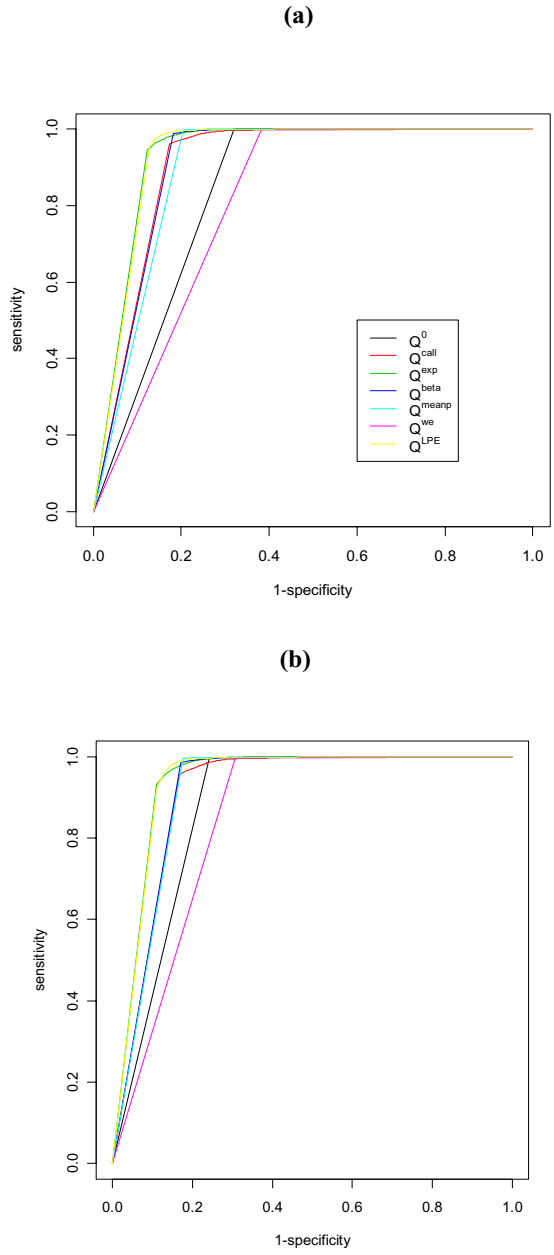

Figure 1
Receiver Operator Characteristic (ROC) plots for tests of differential expression in the simulated data with treatment effect δg = 2.0. Six weighted tests and the local pooled error (LPE) test are compared. (a) p-values unadjusted for multiple testing, (b) p-values adjusted by the weighted Benjamini and Hochberg (WBH) multiple testing method.

Duchenne muscular dystrophy vs. normal muscle

Haslett et al. [16] compared gene expression between 12 quadriceps biopsies from Duchenne muscular dystrophy (DMD) patients and 12 quadriceps biopsies of normal skeletal muscle. Using a method called geometric fold change (GFC), they identified 133 differentially expressed genes (139 probesets) with a permutation-based false discovery rate of 2.3×10^{-3}; Of these, 12 genes (13 probesets) were confirmed by RT-PCR. This set will be referred to as the "RT-PCR" probesets.

Tables 3 and 4 compare the agreement between probesets selected by our test statistics and Haslett et al. [16]. In Table 3, for each test statistic, we selected the top T = 30, 50, 100 or 139 significantly expressed probesets, ranked based on our adjusted false discovery rates, and we counted the number of these probesets that were also in the top T probesets identified by GFC and ranked based on their absolute fold changes. Table 4 shows a similar comparison for the RT-PCR set.

The measure Q_g^{call} had the worst performance in concordance for the lists of top T probesets in Table 3, and for identifying RT-PCR probesets in Table 4. The weighted t-statistics based on $Q_g^{\exp}$ and Q_g^{beta} performed quite similarly, and gave results that were close also to three other statistics (Q_g^0 Q_g^{meanp} Q_g^{we}), but $Q_g^{\exp}$ showed slightly better results in Table 3 than these other methods. Interestingly, the FDR values in Table 3 associated with $Q_g^{\exp}$ are visibly larger than for the other methods. Although there is more agreement between GFC and Q_g^{LPE} in Table 3 than for other methods when the number of selected probesets is small (< = 100), the agreement decreases when more probesets are selected, even though the FDR estimates are smaller. This statistic is obviously very different from the others. When examining Table 3 for T = 30, agreement for all methods with the GFC paper was slightly better using MAS5. This may not be surprising since Haslett et al. [16] used MAS5 also, although, in Table 3, this relationship not entirely consistent across different values of T. In Table 4, however, the agreement is always better using MAS5.

Analysis of Choe's spiked-in data

The availability of data from spiked-in experiments (Choe et al. [17]) provides an excellent opportunity to examine the performance of our weighted statistics on real data where the answers are known. Selected transcripts were added at a range of known concentrations; some were chosen to have differential expression between two groups of samples (true positive changes in expression); others were spiked-in at the same concentration in the two groups (true negative changes in expression).

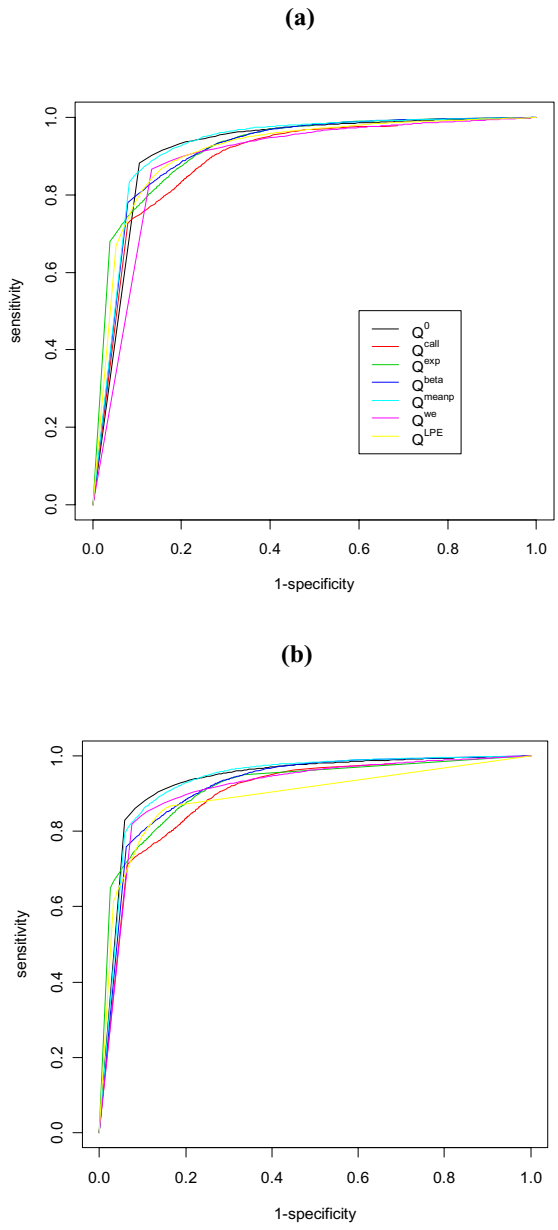

Figure 2
Receiver Operator Characteristic (ROC) plots for tests of differential expression in the simulated data with treatment effect δg = 1.0. (a) p-values unadjusted for multiple testing, (b) p-values adjusted by the weighted Benjamini and Hochberg (WBH) multiple testing method.

Figures 3 and 4 show the ROC curves for the comparison of the two groups for MAS5 and RMA, respectively. Table 5 shows the AUC under different weighting and multiple testing strategies. For both the RMA and MAS5 strategies, it is still clear that whether p-values were unadjusted or adjusted by the WBH method, $Q_g^{\exp}$ outperforms the other quality scores. However, unlike the results in the DMD as well as the simulated data, here the quality measure Q_g^{call} had better performance than Q_g^0, Q_g^{meanp} and Q_g^{we}. In general, the results based on the MAS5 strategy are slightly better than those based on RMA for any quality-based test statistic, but the relative rankings of the different statistics remain almost the same across the two summarization methods. The LPE test appears to be particularly sensitive to the summarization method. It performs very well for the MAS5 method without multiple testing corrections and very poorly when the BH multiple testing is used. Test statistics based on $Q_g^{\exp}$ and Q_g^{beta} had better performance than the LPE test when p-values were adjusted by the WBH in both MAS5 and RMA.

Discussion

In traditional gene selection methods, pre-filtering and selection are two separate procedures, and commonly used pre-filtering methods are based on Affymetrix calls [6-9]. We unified gene filtering and gene selection procedures together, where the importance of a gene is measured by its quality score, defined across all arrays and experimental groups, rather than by a given cutoff call value. The methods, therefore, overcome the shortcomings of the traditional methods to filter out genes before any high-level data analysis, such as gene selection, is carried out.

Our measure Q_g^{call} is essentially the same as the traditional method of keeping only probesets where all arrays record a present call. Our results clearly demonstrated that the weighted t-statistic based on Q_g^{call} has the worst performance in the real data and all simulation models; Pounds and Cheng [11] also found that simple filtering based on Present/Marginal/Absent calls was a poor choice. However, this statistic performed well in the spiked-in data. This artificially-created dataset contained only transcripts that were added in known concentrations, therefore the expression signals are likely to be much more consistent across arrays than signals from different individuals.

The best choice of weighting statistic in the presence of adjustments for multiple testing appears to depend on the size of the treatment effects. The exponential model meas-

Table 2: Area under the curves (AUCs) for five simulation models *

Model	Multiple Testing	Quality Measures						
		Q_g^0	Q_g^{call}	$Q_g^{\exp}$	Q_g^{beta}	Q_g^{meanp}	Q_g^{we}	Q_g^{LPE}
Model 1 δ_g = 2.5	Unadjusted	0.777	0.380	0.925	0.892	0.871	0.743	0.924
	WBH	0.812	0.382	0.929	0.895	0.885	0.772	0.932
Model 2 δ_g = 2.0	Unadjusted	0.840	0.387	0.934	0.907	0.898	0.809	0.932
	WBH	0.878	0.387	0.937	0.912	0.911	0.846	0.937
Model 3 δ_g = 1.5	Unadjusted	0.895	0.394	0.933	0.918	0.922	0.868	0.932
	WBH	0.925	0.395	0.937	0.923	0.932	0.902	0.936
Model 4 δ_g = 1.0	Unadjusted	0.918	0.399	0.921	0.913	0.926	0.889	0.915
	WBH	0.935	0.401	0.917	0.918	0.934	0.912	0.891
Model 5 δ_g = 0.5	Unadjusted	0.876	0.408	0.875	0.876	0.884	0.850	0.850
	WBH	0.882	0.410	0.858	0.861	0.882	0.860	0.770

For $Q_g^{\exp}$ and Q_g^{beta}, the sensitivity parameter ν was set to 0.05.

* Simulated probe level data was summarized using RMA

ure, $Q_g^{\exp}$, appears to have a slight advantage over the one-parameter beta distribution Q_g^{beta}, although for small effect sizes they perform very similarly. The distribution of quality scores for the exponential model gives, in general, lower weight to probesets with small detection p-values than the beta distribution. This may lead to a small reduction in sensitivity – the effect of these lower weights can be seen when examining the FDR cutoffs in the DMD data, and in the figures. However, these lower weights also improve specificity which tends to lead to better overall prediction.

For quality measures $Q_g^{\exp}$ and Q_g^{beta}, the performance of the weight depends on the sensitivity parameter ν. In order to analyze the effect of different values of parameter ν on our results, we did another simulation study. Table 6 summarizes the sensitivity and specificity of the proposed test procedures to detect differentially expressed genes (using the weighted Benjamini-Hochberg procedure to control the false discovery rate at 5%), as a function of ν when the treatment effect was set to 2.0. The sensitivity decreases and specificity increases as ν gets smaller. The sensitivity associated with Q_g^{beta} is slightly better than $Q_g^{\exp}$ across this table, but the specificity is worse. It can be argued that the optimal choice of ν is the one that comes closest to perfect prediction. For ν = 0.05 and 0.01, the Euclidean distances from the point where sensitivity = 1 and specificity = 1 are 0.153 and 0.157, for the exponential distribution, and 0.219 and 0.221, for the beta distribution, respectively. Therefore, we fixed ν = 0.05 in our analysis; this choice can be thought of as a p-value of 0.05, leading to the interpretation that a representative transcript from a particular experimental group will be called present when the expected (under the assumed distribution) p-value <0.05. In fact, the significance level of 0.05 is widely used in statistical hypothesis testing and is comparable to the thresholds for Present/Marginal/Absent calls used in the Affymetrix software.

It is now a common practice to use procedures for controlling multiple testing when identifying differentially expressed genes. Various procedures, such as the Bonferroni correction, the Benjamini and Hochberg false discovery rate [4], or the Benjamini and Yekutieli false discovery rate [20], have been widely used. Due to the quality adjustment in our proposed t-statistics, we can no longer assume that the p-values p_{tg} have a uniform distribution under the null hypothesis, and we can not assume that the test statistics have identical distributions. Therefore, we developed weighted multiple testing procedures. Our findings showed that the proposed weighted Benjamini and Hochberg (WBH) adjustment procedure is better than the weighted Bonferroni (WB) and weighted Benjamini and Yekutieli (WBY) adjustment procedures (results not shown). However, we also observed that the proposed t-statistics using WB and WBY had poor sensitivity, regardless of the type of quality measure score (data not shown). We are planning to further extend weighted multiple testing methods in order to redefine Storey and Tibshirani's positive false discovery rate (pFDR) [21],

Table 3: Agreement in probeset selections between our methods and Haslett et al. [16]: Given a chosen number of selected probesets, how many of the probesets selected by GFC were also selected by the methods in this paper (corresponding false discovery rate WBH-FDR)

Summarization Method	Number of probesets selected	Q_g^0	Q_g^{call}	$Q_g^{\exp_e}$	$Q_g^{beta_e}$	Q_g^{meanp}	Q_g^{we}	Q_g^{LPE}
MAS5	30[a]	10[b] (3.6e-06[c])	6 (7.6e-06)	10 (6.2e-04)	10 (4.9e-06)	10 (3.6e-06)	11 (3.6e-06)	17 (2.3e-21)
	50	21 (1.4e-05)	13 (1.6e-05)	25 (2.3e-03)	22 (1.0e-05)	21 (1.2e-05)	22 (7.8e-06)	32 (6.0e-13)
	100	56 (3.5e-05)	40 (1.5e-04)	59 (4.1e-03)	57 (2.2e-05)	56 (2.6e-05)	56 (1.9e-05)	61 (3.3e-09)
	139 (2.3e-03)	94 (1.1e-04)	65 (3.3e-04)	96 (9.2e-03)	95 (8.4e-05)	94 (9.3e-05)	94 (6.1e-05)	76 (4.1e-07)
RMA	30	8[d] (1.3e-05)	7 (2.5e-05)	9 (1.9e-03)	8 (7.5e-06)	8 (9.7e-06)	8 (5.6e-06)	20 (1.8e-18)
	50	24 (4.3e-05)	13 (5.0e-05)	26 (3.4e-03)	24 (2.8e-05)	24 (3.2e-05)	24 (2.0e-05)	30 (1.8e-11)
	100	60 (1.4e-04)	42 (3.3e-04)	63 (1.1e-02)	60 (9.8e-05)	60 (1.3e-04)	59 (8.0e-05)	62 (3.5e-07)
	139	85 (2.5e-04)	69 (7.4e-04)	92 (1.9e-02)	92 (2.0e-04)	89 (2.2e-04)	85 (1.5e-04)	77 (2.2e-05)

[a] The top 30 probesets were selected by each method.

[b] 10 probesets overlapped between the top 30 probesets selected by GFC and the top 30 probesets selected by Q_g^0, when data were summarized using MAS5.

[c] The WBH-FDR for the top 30 probesets selected by Q_g^0 was 3.6e-06.

[d] 8 probesets overlapped between the top 30 probesets selected by GFC (data normalized by MAS5) and the top 30 probesets selected by Q_g^0 (data normalized by RMA).

[e] Sensitivity parameter $\nu = 0.05$.

where the test statistics are assumed to be identically distributed.

Several other modified t-statistics, such as SAM [22], penalized t-statistics [23], or the local pooled error method used here [18], have been developed for microarray data analysis. All of these methods focus on overcoming the shortcoming of the ordinary t-statistics for ranking genes, due to unstable variance estimates that may arise when sample size is small. Each method used a slightly different strategy to estimate a penalty parameter for smoothing unstable variance estimates, using information from all genes rather than relying solely on variance estimates from an individual gene. However, as we discussed in the "Background" Section, the quality of this information is not the same across genes. Therefore, the estimate of the penalty parameter may not be reliable if we assume that all gene-specific information has the same quality. Here, we take another strategy to improve the performance of ordinary t statistics by putting a high weight for the genes with high quality and a low weight for those with low quality in the t-statistic calculations. The initial comparison of our strategy with the LPE test demonstrated that our weighted tests based on $Q_g^{\exp}$ and Q_g^{beta} are promising – AUC results were fairly similar, and better in several cases. Although our weighting strategy could be combined with a revised variance-smoothing algorithm to potentially improve the performance of the test statistics in small samples, this was not the focus of this work.

It should be noted that the probe-level data analysis (such as background correction, normalization and summarization methods) may influence the results of the test procedures we discussed. Our initial analysis of real data and spiked-in data show that the weighted test statistics based on the MAS5 strategy may have slightly better performance than those based on the RMA strategy. In future investigations, we plan to explore how different probe-level data analysis methods, such as MAS5, dCHIP, RMA, PLIER and gcRMA, and different detection p-value calculation methods [17], may influence our weighting strategy in detail.

Our results showed that some true differences are missed with any filtering method, as was noted by others [11]. It is known that the mismatch probes measure some true signal [12], and therefore is possible that, especially for genes expressed at a low level, quality scores for real data are lower than in our simulated data. Defining an alternative to the detection p-value that does not depend on the mismatch data may lead to better sensitivity. However, if mismatch probes do contain signal, and a study wishes to identify small changes in gene expression, the decision to use any filter at all must be carefully considered, given that some true effects are likely to be excluded. In addition, when effects are small, it must be realized that using multiple testing corrections will also lead to the exclusion of real effects.

Table 4: Comparison of our methods and Haslett et al. [16]. Identification of differentially-expressed probesets validated by RT-PCR: given a chosen number of probesets selected by GFC, and the number of probesets validated by RT-PCR within this set, how many of the RT-PCR probesets were selected by the methods in this paper.

Method	# of probesets selected: # of GFC selections validated by RT-PCR	Q_g^0	Q_g^{call}	$Q_g^{\exp_e}$	$Q_g^{beta_e}$	Q_g^{meanp}	Q_g^{we}	Q_g^{LPE}
MAS5	30[a] : 8[b]	5[c]	1	5	5	5	5	4
	50 : 11	8	2	7	7	8	7	6
	100 : 12	11	3	10	10	11	10	11
	139 : 13	12	4	11	11	12	12	12
RMA	30 : N/A	2[d]	1	4	2	2	2	6
	50 : N/A	6	2	7	6	6	6	7
	100 : N/A	9	4	9	9	9	8	8
	139 : N/A	10	4	11	10	10	10	11

[a] The top 30 probesets were selected by each method.
[b] 8 probesets in the top 30 were validated by RT-PCR in Haslett et al. [16].
[c] 5 RT-PCR probesets were among the top 30 probesets selected by Q_g^0. The 5 RT-PCR probesets are among the 8 RT-PCR probesets selected by GFC.
[d] 2 RT-PCR probesets were among the top 30 probesets selected by Q_g^0 (data normalized by RMA). The 2 RT-PCR probesets are among the 8 RT-PCR probesets selected by GFC, using the MAS5 normalization.
[e] Sensitivity parameter $\nu = 0.05$.

The proposed quality measures may also be useful in other applications of microarray experiments. For example, in microarray meta-analysis we could weight based on not just the quality of a study [24], but the quality of each measurement [25,26].

Methods

Quality measures

The quality of a probeset may depend on many quantities such as spatial arrangement on arrays, upper and lower threshold effects, etc. Here we focus on measuring reliability and consistency of a probeset's expression using the probeset's detection p-value. The distribution of expression within a probeset, leading to a detection p-value, is influenced by all stages of the microarray process including scanner brightness, background, RNA quality, chip design, etc. [5]. Therefore, it can be thought as a synthesis index to represent the probeset's quality.

In this paper, we defined quality measures hierarchically at 3 levels (Table 1): (a) gene and array level, (b) gene and group level, summarizing across arrays within each group, (c) gene level, summarizing across arrays as well as groups. Suppose there are $i = 1, 2, ..., I$ arrays in total, each containing $g = 1, 2, ..., G$ genes. Further, assume that there are W treatment groups, each consisting of n_w arrays, for $w = 1, 2, ..., W$. Therefore, for the first aspect (a), we measured the quality of the measure of expression for one transcript based on the detection p-value or the Affymetrix Present/Absent/Marginal call [5]. Two measures were defined, which we denote by q_{gi}, the detection p-value, and q^*_{gi}, based on the present/absent call (see Table 1).

Approaches for summarizing across a set of arrays take two forms. Firstly, the quality of a gene's measurement across a particular group of arrays can be defined; we call this R_g^w. Several different quality scores are described in Table 1, with more detail below for the exponential and beta distributions. Secondly, group-level quality scores can be used to create a single summary measure that applies to all arrays being analyzed. For the latter, we use $Q_g = \max_{w \in \{1 ... W\}}[R_g^w]$. Use of the maximum leads to the desirable property that genes present under one set of experimental conditions but absent under another will be retained for analysis with a high quality score.

To develop a model-based quality measure, we argued as follows: if a gene is not expressed or cannot be measured, then the detection p-values (q_{gi}) are expected to follow a uniform distribution. Equivalently, if a gene's expression cannot be detected, then we can assume a common distribution for this gene for all arrays in group w, such that $-\log(q_{gi}) \sim Exp(1)$. That is, the negative $\log(q_{gi})$ follows an exponential distribution with a rate parameter value of 1 [25]. We therefore made the assumption that the q_{gi} for gene g and array i follows the one-parameter exponential distribution with a group specific mean λ_{gw}, $w = 1, ..., W$,

Table 5: Area under the curves (AUCs) of Choe's spiked-in data (ν = 0.05)

Method	Multiple Testing	Quality Measures						
		Q_g^0	Q_g^{call}	$Q_g^{\exp}$	Q_g^{beta}	Q_g^{meanp}	Q_g^{we}	Q_g^{LPE}
RMA	Unadjust	0.800	0.881	0.885	0.882	0.850	0804	0.854
	WBH	0.779	0.871	0.874	0.868	0.832	0.776	0.847
MAS5	Unadjust	0.815	0.889	0.901	0.896	0.869	0.814	0.922
	WBH	0.800	0.877	0.895	0.884	0.856	0.789	0.653

so that $-\log(q_{gi}) \sim Exp(\lambda_{gw})$. The maximum likelihood estimate of λ_{gw} is the inverse of the group-specific sample mean. Define ν, a sensitivity parameter, to be a desired threshold for the detection p-values, representing 1 minus the probability that any probeset in a particular treatment group shows a detectable signal. Then $P(q_{giw} \le \nu)$ leads to $R_g^w = P(-\log(q_{giw}) \ge -\log\nu) = e^{\lambda_{gw}\log\nu}$. Although the distribution of the q_{gi} may not follow an exponential distribution exactly, this simple assumption may give adequate results for developing quality measures.

Theoretically, p-values are expected to follow a two-parameter Beta distribution. However, estimating these two parameters is occasionally impossible due to the fact that the detection p-values are derived from rank tests, and all arrays may have exactly the same p-value when genes are highly expressed. Pounds and Cheng [11] recently proposed a one-parameter Beta distribution to model detection p-values within each experimental group, $R_g^w = p(q_{giw} \le \nu) = \int_0^\nu Beta(q; a_{gw})dq = \nu^{agw}$, with sensitivity parameter ν. The parameter $\hat{a}_{gw}$ can be estimated by $\hat{a}_{gw} = \bar{q}_{gw}/(1-\bar{q}_{gw})$ [27], where $\bar{q}_{gw} = \sum_{i\in w} q_{giw} \Big/ n_w$

Tests of differential expression with quality weights

In traditional meta-analysis, quality measures are often used when combining results from different studies [28]. Without loss of generality, we can assume that we are comparing two groups of microarrays (with N_A arrays in group A and N_B in group B), and testing for differentially-expressed genes with the two-sample Welch t-statistic, not assuming equal variances. For gene g, the test statistic is therefore $t_g = \frac{\bar{x}_{gA} - \bar{x}_{gB}}{\sqrt{s_{gA}^2/N_A + s_{gB}^2/N_B}}$, where $\bar{x}_{gA}$ and $\bar{x}_{gB}$ denote the sample average intensities in groups A and B, respectively, and s_{gA}^2 and s_{gB}^2 denote the corresponding sample variances. For more than two groups, an F statistic could be used instead. The quality measure (Q_g) for gene g is then incorporated by $t_g^* = t_g Q_g$. Therefore, $t_g^* = t_g$ when $Q_g = 1$, and t_g^* goes to zero with low quality scores.

We converted this modified t-test statistic to a p-value by reference to a standard t-distribution with degrees of freedom based on Satterthwaite's approximation [29], assuming unequal variances between groups A and B. Of course, it is clear that t_g^* will no longer follow the t-distribution, since the kurtosis will be substantially altered as Q_g gets closer to zero. However, the next section describes how these altered p-values were used in modified adjustments for multiple testing.

This simple approach of weighting the t-test statistics can be contrasted with the approach of weighting the expression intensities within the test statistic. Using standard formula for weighting based on quality, we constructed $t_{qg} = \frac{\bar{x}_{qgA} - \bar{x}_{qgB}}{\sqrt{s_{qgA}^2/N_A + s_{qgB}^2/N_B}}$, where $\bar{x}_{qgw}$ and s_{qgw}^2 are the group-specific weighted means and standard deviations [30], weighted by 1 - q_{qi}. Here, $w = A, B$. We obtained p_{qtg} from a t-distribution with degrees of freedom based on Satterthwaite's approximation [29]. For brevity, we refer to this approach as Q_g^{we} even though there was no specific quality measure attached to each gene.

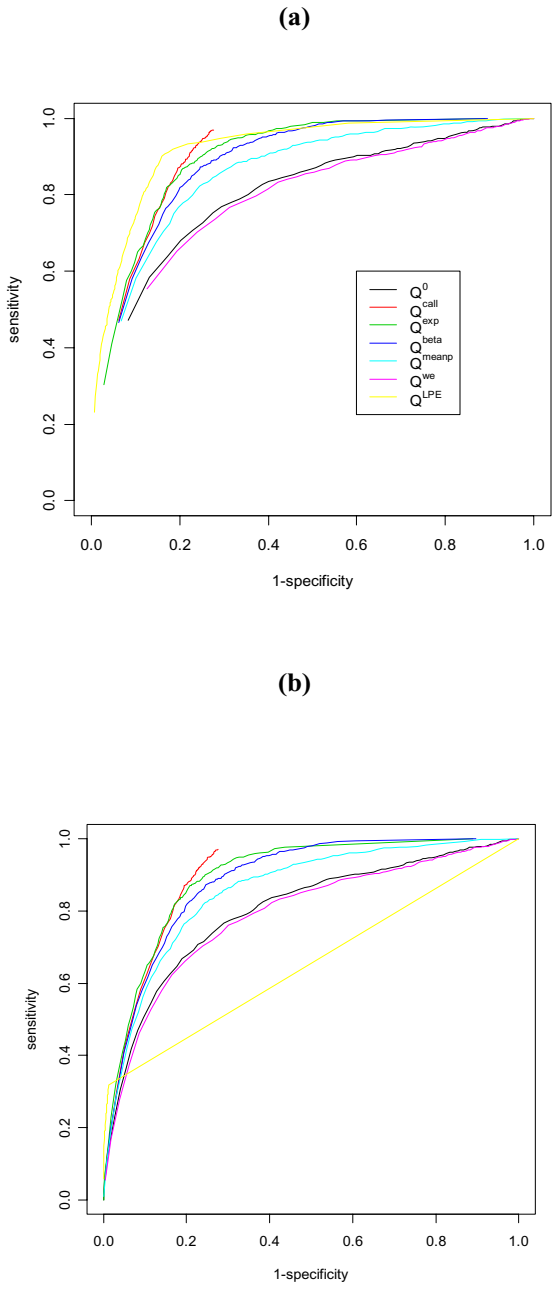

Figure 3
Receiver Operator Characteristic (ROC) plots for tests of differential expression in Choe's spiked-in data summarized by MAS5. (a) p-values unadjusted for multiple testing, (b) p-values adjusted by the weighted Benjamini and Hochberg (WBH) multiple testing method.

Multiple hypotheses testing with weights

Several authors have considered the problem of including weights in multiple hypothesis testing situations [31-33]. Often, weights have been introduced when some of the hypotheses H_i are deemed more important than others. For example, Holm [31] first introduced weights into his sequentially rejective multiple hypothesis testing procedure, where weight was used to indicate the importance of the hypotheses. In that spirit, testing for differences in genes that are not expressed or non-specific in such genes should be considered less important than testing for differences among genes that are specific and well-expressed.

To give another perspective, our quality-weighted statistics t^* can also be considered as an implementation of a filtering method. For example, Q_g^{call} is a straightforward implementation of filtering based on including only genes with present calls on all arrays, and our other measures effectively exclude genes with small quality measures when the weighted statistic t^* is close to zero. Therefore, we propose to adjust the effective number of genes tested to correspond to the number of tested genes of high quality.

Assume that quality measures $Q_1, Q_2, \ldots, Q_G$ and significance values $p_1, p_2, \ldots, p_G$ have been calculated for all G genes. Let $p_1^* \le p_2^* \le,\ldots,\le p_G^*$ denote the ordered significance values, and let $Q_1^*, Q_2^*, \ldots, Q_G^*$ denote the quality measures in the same order. We redefined the Benjamini and Hochberg (BH) multiple testing procedure as $\tilde{p}_g^{WBH} = \min_{u=g,\ldots,G}\{(\min(\frac{\sum_{z=1}^{G} Q_z^*}{\sum_{z=1}^{u} Q_z^*} p_u^*, 1)\}$. The sums of the quality measures in this formula estimate the number of high quality genes instead of the number of genes. We call this approach the Weighted Benjamini and Hochberg (WBH) method. Similar modifications lead to a weighted Benjamini and Yekutieli (WBY) procedure (not shown).

We applied WBH to the t-statistics modified by the quality measures Q_g^{call}, $Q_g^{\exp}$, Q_g^{beta} and Q_g^{meanp}. For the t-statistics modified by quality measures Q_g^0 and Q_g^{we} as well as local pooled error (LPE) test Q_g^{LPE}, the standard form of the Benjamini and Hochberg (BH) multiple testing procedure was used.

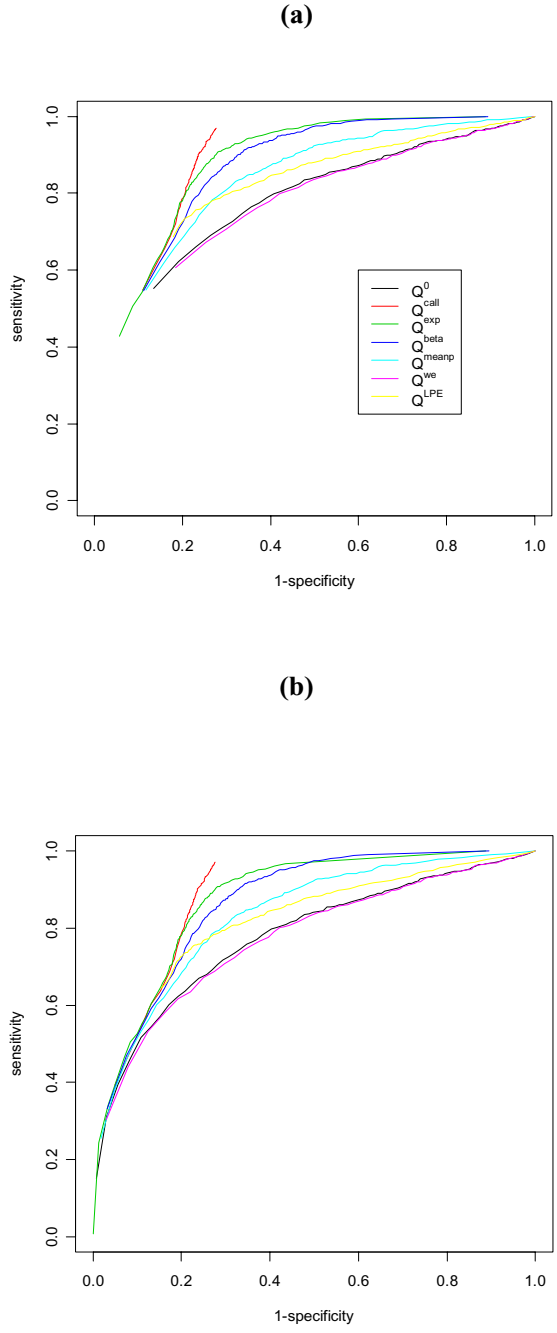

Figure 4
Receiver Operator Characteristic (ROC) plots for tests of differential expression in Choe's spiked-in data was summarized by RMA. (a) p-values unadjusted for multiple testing, (b) p-values adjusted by the weighted Benjamini and Hochberg (WBH) multiple testing method.

Simulation method

Affymetrix probe level data were generated based on a unified model proposed by Wu et al. [14]. When treatment effects are considered for different conditions, such as cancer and normal tissues, gene expression across these conditions can be modeled as:

$$\begin{aligned} Y_{gij} &= O_{gij}^{PM} + N_{gij}^{PM} + S_{gij}^{PM} \\ &= O_{gij}^{PM} + \exp(\mu_{gij}^{PM} + \varepsilon_{gij}^{PM}) + \exp(\gamma_g + \delta_g X_i + a_{gij} + b_i + \zeta_{gij}^{PM}) \\ W_{gij} &= O_{gij}^{MM} + N_{gij}^{MM} + \Phi S_{gij}^{MM} \\ &= O_{gij}^{MM} + \exp(\mu_{gij}^{MM} + \varepsilon_{gij}^{MM}) + \Phi_{gj} \exp(\gamma_g + \delta_g X_i + a_{gij} + b_i + \zeta_{gij}^{MM}), \end{aligned} \quad (2)$$

where Y_{gij} and W_{gij} are the *PM* and *MM* intensities for the probe j in probeset g on array i respectively; O denotes optical noise. N represents non-specific binding (NSB) noise. S is a quantity proportional to RNA expression and the coefficient $0 < \Phi < 1$ accounts for the fact that for some probe-pairs the *MM* detects signal. Following Wu et al. [14], we simulated O_{gij}^{PM} and O_{gij}^{MM} using independent draws from $\log_2(O_{gij}^{PM}) \sim N(5, 0.1)$ and $\log_2(O_{gij}^{MM}) \sim N(5, 0.1)$. We set $\mu_{gj}^{PM} = \mu_{gj}^{PM} = 4.6$, and assumed that ε_{gj}^{PM} and ε_{gj}^{MM} follow a bivariate normal distribution with mean 0, variance 1, and correlation 0.88. We then generated identically and independently distributed random variates $e \sim N(0,0.08)$, so that $\varepsilon_{gij}^{PM} = \varepsilon_{gj}^{PM} + e_{gij}^{PM}$ and similarly $\varepsilon_{gij}^{MM} = \varepsilon_{gj}^{MM} + e_{gij}^{MM}$. When mismatch probe j of gene g is attached by picking up stray signal, Φ_{gi} is generated as $\Phi_{gi} \sim Beta(0.5,5)$, otherwise, $\Phi_{gi} = 0$. The proportion of attached probes among total probes was set to 0.01. Since S follows a power law, we set its base to 2. Therefore, if we denote γ_g as the baseline log expression level for probeset g, we can select $\log_2(\gamma_g)$ expression levels from 0 to 12, which were generated from $\gamma_g \sim 12 * Beta(1,3) + 1$. δ_g is the expected differential expression of gene g across different conditions, which is varied in the simulations. b_i, which describes the need for normalization, was set to be zero. α_{gij} is the signal detecting ability of probe j in gene g on array i, which is assumed to follow a normal distribution with mean zero and signal detection variance σ_α^2. Multiplicative errors ζ_{gij}^{PM} and ζ_{gij}^{MM} were generated independently from $N(0, \sigma_\zeta^2)$. Values of σ_α^2 and σ_ζ^2 were varied in the simulations. Since the theoretical maximum value of an Affymetrix scanner is 2^{16},

Table 6: Sensitivity and specificity for detecting differentially expressed genes, as a function of the sensitivity parameter ν

ν	$Q_g^{\exp}$		Q_g^{beta}	
	Sensitivity	Specificity	Sensitivity	Specificity
0.6	0.9996	0.7562	0.9998	0.7207
0.5	0.9995	0.7763	0.9997	0.7372
0.4	0.9981	0.7930	0.9994	0.7490
0.3	0.9967	0.8070	0.9990	0.7590
0.2	0.9922	0.8211	0.9984	0.7668
0.1	0.9831	0.8373	0.9956	0.7752
0.05	0.9660	0.8507	0.9936	0.7812
0.01	0.9167	0.8665	0.9413	0.7873

we kept only generated Y_{gij} and W_{gij} less than 2^{16}, that is, $Y_{gij} = \min(O_{gij}^{PM} + N_{gij}^{PM} + S_{gij}^{PM}, 2^{16})$ and $W_{gij} = \min(O_{gij}^{MM} + N_{gij}^{MM} + \Phi S_{gij}^{MM}, 2^{16})$. The RMA algorithm [12] was used to summarize simulated probe-level data into a signal value, and the MAS5 algorithm [5] was used to obtain the detection p-value.

Design of simulation study

The simulation design is shown in Table 7. We assumed two groups, A and B, with N_A and N_B arrays, respectively. G genes were generated, of them the proportion of expressed genes is k, and the proportion of differentially expressed genes is d of the $G*k$ expressed genes. We set the number of up-and -down regulated genes to be the same in this study. Therefore, the G genes are divided into four groups: a non-expressed gene group where genes are not expressed across all arrays in group A and group B; a non-differentially expressed gene group where genes are expressed but not differentially between groups A and B; an up-regulated gene group where the mean gene expression of gene g in group B is larger by δ_g than that in group A; and a down-regulated gene group where the mean gene expression of gene g in group A is larger by δ_g than that in group B.

We ran five simulation models following the above design. The specific parameters used in the five models are: number of genes: 1000; sample size: 25 arrays in groups A and B, respectively (for a total of 50 arrays); number of probes within each probeset: 16; proportion of

Table 7: Simulation structure of Affymetrix microarray data

		Group A			Group B		
		Array 1	.....	Array N_A	Array 1	.....	Array N_B
Up-regulated Gene Group	g_1		γ_g			$\gamma_g = \gamma_g + \delta_g$	
	...						
Down-regulated Gene Group	$g_{d*k*G/2}$ $g_{d*k*G/2+1}$		$\gamma_g = \gamma_g + \delta_g$			γ_g	
	...						
Non-differentially Expressed Gene Group	g_{d*k*G} $g_{d*k*G+1}$			$\gamma_g \neq 0$ $\delta_g = 0$			
	...						
Non- Expressed Gene Group	g_{k*G} g_{k*G+1}			$\gamma_g = 0$ $\delta_g = 0$			
	...						
	G						

expressed genes: 0.5; proportion of differentially expressed genes: 0.1; signal detection variance: 0.25; multiplicative error variance: 0.05 and sensitivity parameter ν: 0.05.

Duchenne muscular dystrophy data

Haslett et al. [16] hybridized the total RNA to HG-U95Av2 GeneChips. They used MAS5 to obtain signal intensities and normalized with a linear regression. Tests of differential expression were based on geometric fold change (GFC) [16]. The differential expression of 12 genes (13 probesets) was confirmed by quantitative RT-PCR analysis of seven DMD biopsies and four unaffected biopsies. We used only 23 arrays (only 11 DMD arrays) for our re-analysis since one file was truncated. Raw data were converted to signal estimates using both MAS5 [5] and RMA [12], both were implemented using the affy package in Bioconductor [34].

Spiked-in data of Choe et al. [17]

The 'spiked-in' experiment (Choe et al. [17]) for Affymetrix Genechips provides a controlled dataset of 3,860 RNA species with known sequence and known concentration. Two different samples were prepared and hybridized in triplicate to Affymetric GeneChips; these are called the 'constant' (C) and 'spike' (S) samples. Out of 3,860 RNA species, 2,551 of them were created to have the same concentrations in both samples while the rest (1,309) were spiked in with different concentrations between the S and C samples. Ten fold-change levels, ranging from 1.2 to 4-fold, were assigned to the spiked-in RNAs. Basically, all the RNAs with positive log fold changes can be thought of as differentially expressed genes. In this study, we considered the top 1000 probesets as differentially expressed genes (as did Choe et al. [17]).

Authors' contributions

CG initiated, designed and managed the study. CG also proposed the methods in this manuscript. JB participated in designing and managing the study. PH conducted data analysis and drafted the manuscript. CG and JB revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge helpful suggestions from two anonymous reviewers that greatly improved the quality of the manuscript. This work was supported by the Ontario Genomics Institute and Genome Canada.

References

1. Su AI, Cooke MP, Keith AC, Yaron H, John RW, Tim W, Anthony PO, Raquel GV, Lisa MS, Aziz M, Ardem P, Garret MH, Peter GS, John BH: **Large-scale analysis of the human and mouse transcriptomes.** *Proceedings of the National Academy of Sciences USA* 2002, **99:**4465-4470.
2. **Affymetrix-Human Genome Arrays** [http://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf]
3. Ge Y, Dudoit S, Speed TP: **Resampling-based multiple testing for microarray data hypothesis.** *Test* 2003, **12:**1-44.
4. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society B* 1995, **85:**289-300.
5. **Affymetrix – Statistical Algorithms Description Document** [http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf]
6. Hittel DS, Kraus WE, Hoffman EP: **Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome.** *Journal of Physiology* 2003, **548:**401-410.
7. Blalock EM, Chen KC, Sharrow K, Herman JP, Porter NM, Foster TC, Landfield PW: **Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment.** *The Journal of Neuroscience* 2003, **23:**3807-3819.
8. Chen YW, Nader G, Baar KR, Hoffman EP, Esser KA: **Response of rat muscle to acute resistance exercise defined by transcriptional and translational profiling.** *Journal of Physiology* 2002, **545:**27-41.
9. Li C, Wong WH: **DNA-Chip Analyzer (dChip): user manual.** [http://biosun1.harvard.edu/complab/dchip/filter_gene.htm].
10. Li C, Wong W: **Model-based analysis of oligonucleodide array: Expression index computation and outlier detection.** *Proceedings of the National Academy of Sciences USA* 2001, **98:**31-36.
11. Pounds S, Cheng C: **Statistical development and evaluation of microarray gene expression data filters.** *Journal of Computational Biology* 2005, **12:**482-495.
12. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix genechip probe level data.** *Nucleic Acids Research* 2003, **31:**e15.
13. **Affymetrix – Technical Manual** [http://www.affymetrix.com/support/technical/manual/expression_manual.affx]
14. Wu Z, Irizarry RA, Gentleman R, Martinez MF, Spencer F: **A Model Based Background Adjustement for Oligonucleotide Expression Arrays.** *Journal of the American Statistical Association* 2004, **99:**909-915.
15. Seo J, Bakay M, Chen Y, Hilmer S, Sheneiderman B, Hoffman EP: **Optimizing signal/noise ratios in expression profiling: Project-specific algorithm selection and detection p value weighting in Affymetrix microarrays.** *Bioinformatics* 2004, **20:**2534-2544.
16. Haslett JN, Sanoudou D, Kho AT, Bennett RR, Greenberg SA, Kohane IS, Beggs AH, Kunkel LM: **Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle.** *Proceedings of the National Academy of Sciences USA* 2002, **99:**15000-15005.
17. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biology* 2002, **6:**R16.
18. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: **Local-pooled-error test for indentifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, **19:**1945-1951.
19. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143:**29-36.
20. Yekutieli D, Benjamini Y: **Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics.** *Journal of Statistical Planning and Inference* 1999, **82:**171-196.
21. Storey JD, Tibshirani R: **Statistical significance for genome-wide studies.** *Proceedings of the National Academy of Sciences USA* 2003, **100:**9440-9445.
22. Tusher V, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proceedings of the National Academy of Sciences USA* 2001, **98:**5116-5121.
23. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Statistical Applications in Genetics and Molecular Biology* 2004, **1:**Article 3.
24. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling inter-study variation.** *Bioinformatics* 2003, **19:**i84-i90.

25. Hu P, Greenwood MTC, Beyene J: **Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models.** *BMC Bioinformatics* 2005, **6:**128.
26. Tritchler D: **Modelling study quality in meta-analysis.** *Statistics in Medicine* 1999, **18:**2135-2145.
27. Lehman EL: **Theory of point estimation.** *Wiley* 1983.
28. Fleiss JL, Gross AJ: **Meta-analysis in epidemiology, with special reference to studies of the association between environmental tobacco smoke and lunc cancer: a critique.** *Journal of Clinical Epidemiology* 1991, **44:**127-139.
29. Satterthwaite FW: **An approximate distribution of estimates of variance components.** *Biometrics* 1946, **2:**110-114.
30. **SAS Institute Inc. – The MEANS Procedure** [http://www.caspur.it/risorse/softappl/doc/sas_docs/proc/z0608466.htm]
31. Holm S: **A simple sequentially rejective multiple test procedure.** *Scandinavian Journal of Statistics* 1979, **6:**65-70.
32. Benjamini Y, Hochberg Y: **Multiple hypotheses testing with weights.** *Scandinavian Journal of Statistics* 1997, **24:**407-418.
33. Hommel G, Kropf S: **Tests for differentiation in gene expression using a data-driven order or weights for hypotheses.** *Biometrical journal* 2005, **47:**554-562.
34. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy-analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20:**307-315.