Published online December 14, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 22 ***e193***
doi:10.1093/nar/gni193

Integrating probe-level expression changes across generations of Affymetrix arrays

Laura L. Elo[1,2,*], Leo Lahti[2,3], Heli Skottman[2,4], Minna Kyläniemi[2], Riitta Lahesmaa[2] and Tero Aittokallio[1,2]

[1]Department of Mathematics, FIN-20014 University of Turku, Finland, [2]Turku Centre for Biotechnology, PO Box 123, FIN-20521 Turku, Finland, [3]Laboratory of Computer and Information Science, Helsinki University of Technology, PO Box 5400, FIN-02015 HUT, Finland and [4]Institute for Regenerative Medicine Regea, University of Tampere and Tampere University Hospital, FIN-33520, Tampere, Finland

Received August 16, 2005; Revised October 31, 2005; Accepted November 28, 2005

ABSTRACT

There is an urgent need for bioinformatic methods that allow integrative analysis of multiple microarray data sets. While previous studies have mainly concentrated on reproducibility of gene expression levels within or between different platforms, we propose a novel meta-analytic method that takes into account the vast amount of available probe-level information to combine the expression changes across different studies. We first show that the comparability of relative expression changes and the consistency of differentially expressed genes between different Affymetrix array generations can be considerably improved by determining the expression changes at the probe-level and by considering the latest information on probe-level sequence matching instead of the probe annotations provided by the manufacturer. With the improved probe-level expression change estimates, data from different generations of Affymetrix arrays can be combined more effectively. This will allow for the full exploitation of existing results when designing and analyzing new experiments.

INTRODUCTION

The enormous popularity of gene expression profiling with microarrays in recent years has resulted in a rapid accumulation of data in many laboratories and public databases. As microarray experiments are expensive and often involve biological samples that are difficult to obtain, sample sizes in typical microarray studies are relatively small, leading to several false-positive and false-negative findings. Therefore, methods that can effectively extract information from previous studies are of practical interest for minimizing the number of additional experiments needed without compromising the reliability of the results. However, combining data across studies performed at different times and perhaps in different laboratories is a challenging task where both biological and technical sources of variability must be considered carefully.

A major problem in integrative analysis is that gene expression data generated with different microarray platforms are not directly comparable, and even within the same technique different protocols for sample preparation, array hybridization and data analysis can result in severe variations among data sets. Accordingly, the early cross-platform comparisons often showed poor correlation between their intensity measurements (1,2). More recent studies have showed that implementation of standardized protocols for all steps of the microarray study can markedly increase reproducibility between platforms and even across laboratories (3,4). However, some of the variation can be beyond the capacity of standard normalization techniques if the remaining discrepancies between data sets originate from measuring different splice variants of the same gene (5).

As the compositions of microarrays are regularly updated to incorporate new genes with improved target sequences, it is difficult to combine data even from different generations of the same microarray platform. In particular, Affymetrix high-density oligonucleotide arrays utilize multiple (typically 8–16) 25mer probes, the so-called probe set, to measure the expression level of a transcript target. Although the use of several probes for each target leads to more robust estimates of transcript activity, it is clear that probe qualities may significantly affect the results of a study. It has been noticed that a considerable number of probes on various high-density oligonucleotide arrays do not uniquely match their intended targets (6–9). By matching the probe sequences to the most up-to-date genomic sequence data, it is possible to assess the quality of the probes. Redefinition of probe sets according to the latest probe sequence information can increase their accuracy and cross-platform consistency with other array types (6,8,9).

*To whom correspondence should be addressed. Tel: +358 2 333 8002; Fax: +358 2 333 8000; Email: laliel@utu.fi

Previous works on different generations of Affymetrix arrays have concentrated mainly on the reproducibility of their expression results. In a comparison of two Affymetrix arrays, HuGeneFL and HG-U95A, Nimgaonkar *et al*. (10) concluded that the reproducibility is high only when the corresponding probe sets share many exact probes. Hwang *et al*. (7) advanced the comparison analysis by selecting subsets of probes with overlapping sequence segments and recalculating expression values using the selected probes only. While such probe filtering could significantly improve the reproducibility between Affymetrix HG-U95Av2 and HG-U133A arrays, some useful information from the non-overlapping probes measuring identical targets may be lost. In fact, from the investigator's point of view, the enhanced comparability is of practical importance only when the probes match identical targets.

In the present work, we continue the integrative analysis across generations of Affymetrix arrays by considering explicitly the actual targets of probe sequences rather than their similarities. As most current arrays with an enhanced probe design protocol contain high quality probes that do not share sequence similarity with the older probes, we do not filter probes based on overlap but utilize all available probe-level information across generations. We carry out a thorough examination of two in-house data sets, containing expression data from human HG-U133A and HG-U133Plus2.0 arrays and murine MG-U74Av2 and mouse MOE430 2.0 arrays. Additionally, we consider two publicly available data sets, containing expression data from human HG-U95Av2 and HG-U133A arrays. Each data set contains technical replicates hybridized to two array types, allowing us to isolate the array-effects from the underlying biological variation. Since the technical replicates are assumed to produce the same results on both arrays, the comparability of the arrays can be directly evaluated. We also investigate several different probe set pairing approaches in the comparison studies.

Toward combining results from multiple studies, we propose a novel meta-analytic framework, based on the selected probe set pairing method and our probe-level estimate of expression changes (referred to as PECA). The performance of this procedure is demonstrated on a public data set, which also contains several biological samples hybridized to both HG-U95Av2 and HG-U133A arrays. The meta-analysis method is evaluated in terms of its stability when the sample size is reduced. As agreement between the pure expression measurements do not consider the platform-specific probe-effects, which arise from inherent differences in the hybridization efficiency of different probes, we also use relative expression changes when evaluating the methods. Besides removing the probe-effects, expression changes are often more meaningful for the investigator, as the main interest in most studies is in identifying a set of candidate genes that are differentially expressed between groups of samples instead of their plain expression levels.

MATERIALS AND METHODS

Human embryonic stem cell data (hESC)

Two human embryonic stem cell (hESC) lines, HS306 and HS293, from Karolinska University Hospital (Huddinge, Sweden) were derived and cultured in serum replacement medium on human foreskin fibroblast feeder cells as described previously (11). The total RNA was isolated from 5 to 10 hESC colonies using the RNAeasy mini kit (Qiagen, Valencia, CA). The sample preparation was performed according to the Affymetrix two-cycle GeneChip® Eukaryotic small sample target labeling assay version II (Affymetrix, Santa Clara, CA). The samples were hybridized to human HG-U133A and HG-U133Plus2.0 arrays (Table 1).

Mouse Chlamydia pneumonia infection data (mCPI)

Female inbred Balb/c mice obtained from Harlan Netherlands (Horst, The Netherlands) were infected with *Chlamydia pneumoniae* as described previously (12). The axillary lymph nodes from 12 control mice and the mediastinal lymph nodes from 12 infected and 12 re-infected mice were pooled. The total RNA from CD4+ cells were isolated using the Trizol method (Invitrogen Co., Carlsbad, CA) and further purified with RNAeasy mini kit. The sample preparation was performed according to the Affymetrix two-cycle GeneChip® Eukaryotic small sample target labeling assay version II. The samples were hybridized to murine MG-U74Av2 and mouse MOE430 2.0 arrays (Table 1).

Human acute lymphoblastic leukemia data (ALL)

The public data sets from the microarray studies of Yeoh *et al*. (13) and Ross *et al*. (14) contained expression data from ALL patients with different leukemia subtypes. A total of 360 patient samples were hybridized to HG-U95Av2 arrays and 132 of the same samples were also hybridized to

Table 1. Hybridization scheme

Data set	Condition	Samples	HG-U133Plus2.0	HG-U133A	HG-U95Av2	MOE430 2.0	MG-U74Av2
hESC	HS293	2	1	1	—	—	—
hESC	HS306	2	1	1	—	—	—
mCPI	Control	1	—	—	—	1	2
mCPI	Infected	1	—	—	—	1	1
mCPI	Re-infected	1	—	—	—	1	2
ALL	T-ALL	14	—	1	1	—	—
ALL	E2A-PBX1	18	—	1	1	—	—
IM	Dermatomyositis	5	—	1	1	—	—
IM	Other myopathy	9	—	1	1	—	—

The third column indicates the number of samples in each condition. The rest of the columns are the number of hybridizations per sample in each sample set on different array types.

HG-U133A arrays. We selected for our analyses 32 samples that were hybridized to both array types and represented two genetically distinct leukemia subtypes: 14 T-ALL samples and 18 E2A-PBX1 samples (Table 1). To obtain equal sample sizes in both groups, we randomly excluded 4 E2A-PBX1 samples from the analysis.

Human inflammatory myopathies data (IM)

The publicly available data set from the study of Hwang *et al*. (7) contained muscle tissue samples from 14 patients with inflammatory myopathies. The patients were divided into two groups: five patients had dermatomyositis and nine patients had other inflammatory myopathies. Each sample was hybridized to HG-U95Av2 and HG-U133A arrays. To make the present results directly comparable with the results obtained by Hwang *et al*. we included all the samples into our study.

Probe sequence data

Probe sequences and their 'bestmatch' tables were downloaded from the Affymetrix web pages (www.affymetrix.com). Other array-wise information on probes and probe sets, including GeneID annotations, were provided with annotation data packages of the Bioconductor project (15). Genomic mRNA sequences for alignments were downloaded from Entrez nucleotide (16) for human (March 3, 2005) and mouse (April 29, 2005), excluding EST, STS, GSS, 'working draft' and 'patents' sequences, and sequences with a 'XM_' tag, as in (7). The Entrez mRNA sequences were assigned to GeneID identifiers by using the gene2accession conversion file obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA) for human (March 23, 2005) and mouse (April 28, 2005). This resulted in a total of 209 650 and 183 461 mRNA sequences for human and mouse, respectively. The probes in the AFFX-control sets were omitted from the analysis.

Probe verification

To guarantee the quality and comparability of the 25mer probes, we verified them using the Entrez mRNA sequence database (16). Perfect matches of the probes to mRNA sequences were searched with BLAT v. 26 (17). A given probe often matches several mRNA targets. In such cases, it is common that the mRNA sequences are merely separate sequence submissions of the same gene. To distinguish between probes with unique and multiple gene targets, we assigned the Entrez mRNA sequences to GeneID identifiers (18).

The probes were classified according to their manufacturer annotations and our Entrez verifications. *Verified* probes are detected to match Entrez mRNA sequences with a unique GeneID. Probes with no matching GeneID targets are *mistargeted*, and those assigned to several GeneIDs are *non-specific*. A probe is *conflicting* if its verified target is different from the one in the array-wise annotations. *A verified probe set* is a subset of the corresponding original probe set, obtained by masking the mistargeted, non-specific and conflicting probes from the original set. An *alternative probe set* is a collection of probes on a given array that are verified to uniquely measure a given GeneID. An alternative probe set contains verified probes only, but these may include probes from various original probe sets.

Probe set pairing

A common approach to compare different generations of Affymetrix arrays is to use the so-called 'bestmatch' tables provided by the array manufacturer. The best match pairs are based on the similarity between the target sequences of the probe sets. Since the HG-U133Plus2.0 array contains all the probe sets from the HG-U133A and HG-U133B arrays, plus 9921 additional probe sets, the HG-U133A and HG-U133Plus2.0 arrays can be compared by selecting the same probe sets from the two arrays. We consider these pairs as best match pairs as well, although this is a much stricter pairing criterion than the one usually characterizing the best match pairs.

An alternative approach for probe set pairing is to use GeneID identifiers. Original and verified probe sets on both arrays can be assigned to GeneIDs by using the array-wise annotations. As these are not available for alternative probe sets, we used the verified GeneIDs from our Entrez studies. We only considered those GeneIDs for which corresponding probes existed on original, verified and alternative probe sets.

Probe-level expression change averaging (PECA)

We based the selection of genes differentially expressed between two particular groups of samples on probe-level microarray data instead of probe set-level summary intensities obtained with, for instance, robust multi-array average (RMA) (19) or Affymetrix microarray suite (MAS) (www.affymetrix.com). More specifically, we first calculated the selected test statistic separately for each probe in the data and then averaged over the probes within each probe set. In the calculations, we used perfect match (PM) intensities, which were quantile-normalized (20) and log-transformed before the analysis. We refer to this procedure as PECA.

We considered two types of PECA-measures within a microarray study: the signal log-ratio and the Hedges' g, which is a commonly used effect size estimate in meta-analysis (21). Let the normalized logarithmic PM intensities of the probe j in the probe set i under the two conditions within a study be $x_{ij} = (x_{ij1}, \ldots, x_{ijn_1})$ and $y_{ij} = (y_{ij1}, \ldots, y_{ijn_2})$ where the total number of samples within the study is $n = n_1 + n_2$. The signal log-ratio is then defined as $d_{ij} = \bar{x}_{ij} - \bar{y}_{ij}$, and the Hedges' g as $g_{ij} = a(\bar{x}_{ij} - \bar{y}_{ij})/s_{ij}$, where $\bar{x}_{ij}$ and $\bar{y}_{ij}$ are the means of the two groups, s_{ij} is the pooled standard deviation, and $a = 1 - 3/(4n - 9)$ is a correction term that makes the Hedges' g-estimate unbiased. After calculating the probe-level estimates, the probe set-level estimates were formed by averaging over the probes within each probe set. In the present study, the probe sets were defined using the various probe verification criteria and the PECA-estimates were calculated separately within each study on each array generation.

Meta-analysis of effect sizes

Suppose that m studies produce effect size estimates e_k and measures of variability s_k^2, $k = 1, \ldots, m$. Assume that all studies estimate the same parameter μ and any differences between the estimates are due to sampling error $\epsilon_k \sim N(0, s_k^2)$. Then the meta-analysis estimate for μ is the weighted average

over the effect size estimates

$$\hat{\mu} = \frac{\sum_{k=1}^{m} w_k e_k}{\sum_{k=1}^{m} w_k},$$

where the weight w_k is defined as $w_k = s_k^{-2}$. The variance of $\hat{\mu}$ is $s_{\hat{\mu}} = 1/\sum w_k$, and hence the hypothesis H_0:$\mu = 0$ can be considered by using the test statistic $Z = \hat{\mu}/s_{\hat{\mu}}$, which is distributed as $N(0,1)$ under the null hypothesis H_0. For a detailed description of this technique, see (21). Such meta-analytic method was applied in the present study to combine the expression changes across the array types.

TESTING PROCEDURE

We first evaluated the effect of the different probe set pairing and probe verification criteria on the reproducibility of RMA- and MAS-normalized signal intensities between each array pair on which the same sample was hybridized (between-study comparison). We then investigated the comparability of relative expression changes and the agreement of differentially expressed genes between these array pairs using the GeneID matched alternative probe sets (between-study comparison). At this stage, the expression changes were calculated within each array generation (within-study analysis) using the PECA-procedure (probe-level estimation) and the summary intensities from RMA and MAS 5.0 (probe set-level estimation). Finally, we used the meta-analysis approach to combine the expression changes from the different array generations (between-study analysis). The meta-analysis was carried out using PECA-estimated Hedges' *g*-values (probe-level between-study analysis) as well as Hedges'*g*-values calculated from the RMA-derived intensities (probe set-level between-study analysis).

Reproducibility of signal intensities

To assess the level of reproducibility of signal intensities between technical replicates across array generations, we calculated the Pearson correlation coefficient between each array pair from the same sample. The intensity values were obtained with RMA and MAS 5.0. We compared the intensities between the best match pairs of original probe sets and verified probe sets as well as GeneID pairs with three different collections of probes: (i) original Affymetrix probe sets, (ii) verified probe sets and (iii) alternative probe sets. If multiple probe sets corresponded to the same GeneID, their values were averaged (22). On each array, the variability in the intensity values among the probe sets corresponding to the same GeneID was investigated for the 10 GeneIDs with the largest number of probe sets.

Comparability of relative expression changes

The comparability of relative expression changes between alternative probe sets on two array generations was investigated by considering signal log-ratios and Hedges'*g*-values between two particular groups of samples in the hESC, mCPI, ALL and IM data. In addition, we randomly generated 100 subsamples of sizes 2–5 from the ALL data set to study more carefully the performance of the Hedges'*g* with small sample sizes. In each array comparison, two replicate estimates corresponding to the same samples on the different arrays were obtained. We used the Pearson correlation coefficient between these estimates as a measure of comparability between the arrays. The expression changes were calculated using the PECA-approach and the RMA- and MAS-normalized intensities.

Agreement of differentially expressed genes

The agreement of the most differentially expressed genes between the array generations was investigated by ranking the genes according to signal log-ratios and Hedges' *g*-estimates and calculating the proportion of common genes among the top *N* genes in both array types. If two array generations are comparable, the corresponding lists of differentially expressed genes should contain many overlapping genes (3). Again, we used PECA-estimates and the corresponding estimates obtained using the RMA- and MAS-intensity values in the context of alternative probe sets.

Performance of the meta-analysis

The meta-analysis of PECA-based Hedges'*g*-values was compared with the meta-analysis calculated from the RMA-based summary intensities (23). We also compared the performance of both meta-analyses with the corresponding analyses on the individual data sets. The performance of the methods was evaluated by considering the stability of their results when the number of biological samples was reduced (24). We randomly generated 100 subsamples of sizes 2–5 from the ALL data set and applied each method to them. The results of each subset were then compared with the results obtained from the whole data set by determining the proportion of common genes among the top 100 genes.

RESULTS

While most of the probes on the arrays studied could be confirmed to uniquely match a GeneID, a considerable number of probes were rejected since they were either mistargeted, non-specific or conflicting (Table 2). The number of mistargeted probes was especially high on the HG-U133Plus2.0 and MOE430 2.0 arrays, whereas non-specific and conflicting probes were less common. The high number of mistargeted probes on the two arrays is mainly due to the large number of EST-targeted probe sets on these arrays. Our probe verification did not check probes for matches against EST sequences that often lack GeneID assignment but have been used for

Table 2. Probe verification summary

Array type	Probes	Verified (%)	Mistargeted (%)	Non-specific (%)	Conflicting (%)
HG-U133Plus2.0	604 258	58.2	40.2	1.6	2.6
HG-U133A	247 965	82.5	14.4	3.0	3.1
HG-U95Av2	199 084	82.6	14.4	3.0	2.8
MOE430 2.0	496 468	68.2	30.8	1.1	4.9
MG-U74Av2	197 993	73.1	24.2	2.7	1.3

Probes matched to mRNAs with a unique GeneID in Entrez database are considered verified. Mistargeted probes could not be assigned to a GeneID, whereas non-specific probes have several GeneID targets. If the verified target of a probe is different from the annotations provided by the Bioconductor array packages, the probe is considered conflicting.

Table 3. Numbers of probe sets included into the comparisons

Data set	Array comparison	Best match pairs	GeneID pairs	Multiple original sets (%)	Multiple verified sets (%)
hESC	HG-U133A vs. HG-U133Plus2.0	—	12661	36.7	32.2
mCPI	MG-U74Av2 vs. MOE430 2.0	8595	7735	26.4	14.9
ALL, IM	HG-U95Av2 vs. HG-U133A	8429	8240	25.2	18.9

The best match pairs provided by Affymetrix are based on the similarity of the target sequences of the probe sets. The GeneID pairs were obtained by assigning the probe sets to GeneID identifiers. Only GeneIDs for which probes existed on original, verified and alternative probe sets were considered. If multiple probe sets corresponded to the same GeneID, their values were averaged. The last two columns show the proportion of GeneIDs with multiple probe sets when GeneID pairs of original and verified probe sets were formed.

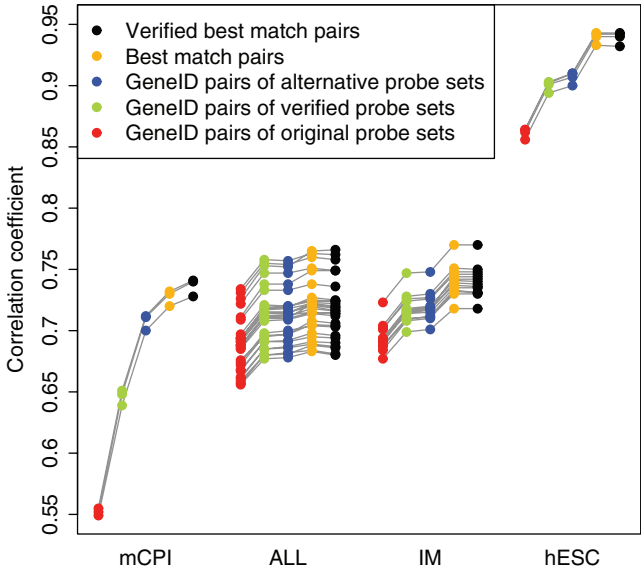


Figure 1. The RMA intensity correlations between technical replicates on two array generations. The Pearson correlation was calculated between each array pair from the same sample. The gray lines show which correlations were obtained from the same array pair with the different probe matching criteria. In the hESC array comparison, the best match probe sets contained exactly the same probes on both array generations, which resulted in very high correlations. The advantages of probe verification and alternative mappings were largest when arrays with different probe collections were compared, as in the mCPI, ALL and IM array comparisons.

the design of several probe sets on these arrays. By simply ignoring the mistargeted and non-specific probes, we were still left with a large number of good-quality probes with a unique GeneID assignment. The typical sizes of the alternative probe sets were approximately the size of the original Affymetrix probe set or its multiplier (see Supplementary Figure 1). The proportion of alternative sets with <5 probes was relatively small, varying between 0.4% (MOE430 2.0) and 2.1% (MG-U74Av2). The numbers of probe sets included into each comparison are listed in Table 3, along with the proportions of GeneIDs with multiple original Affymetrix probe sets.

Effect of probe matching methods on the array reproducibility

Figure 1 illustrates for each array comparison the RMA-based intensity correlations between the pairs of arrays to which the same sample was hybridized. Similar results were obtained with MAS intensities (data not shown). In each array comparison, GeneID pairs of the manufacturer-defined original probe sets performed worst. Probe verification of these sets improved the correlations. Moreover, it was observed that probe verification improved the consistency of the measurements within an array (see Supplementary Figure 2). In the mCPI array comparison, the alternative probe sets produced higher correlations than the verified probe sets, whereas in the ALL, IM and hESC array comparisons the verified sets and the alternative sets performed equally well. In the ALL comparison, also the best match pairs performed similarly, whereas in the mCPI, IM and hESC comparisons, the best match pairs could still improve the correlations. As expected, the improvement was largest in the hESC data, where the best match pairs contained only probes that were the same on both arrays. Interestingly, the verification of the original Affymetrix probe sets used in the best match pairs did not considerably affect the reproducibility of the signal intensities in any of the array comparisons.

Effect of probe-level effect size estimates on the array comparability

The correlations of signal log-ratios and Hedges' *g*-estimates between each replicate study with different array types are shown in Figure 2. In all comparisons, the PECA-estimates showed consistently the best comparability between the array types. The estimates calculated using MAS summary intensities performed generally poorest. With signal log-ratios, the RMA-based estimates usually reduced only slightly the comparability as compared with the PECA-estimates (Figure 2A). With Hedges' *g*-values, however, the benefit from using PECA was considerably higher, especially with small sample sizes (Figure 2B and C). In the hESC data, the correlation increased from below 0.1 with RMA to ~0.7 with PECA (Figure 2B). Similar results were obtained with the ALL data when only two samples from both patient groups were included into the analysis (Figure 2C). As the number of samples increased, the differences between the methods became smaller.

Effect of probe-level effect size estimates on the array agreement

Figure 3 shows the agreement of the most differentially expressed genes between the array types when two groups of samples in the ALL, IM and hESC data were compared. The best agreement was consistently achieved with the PECA-estimates, whereas with MAS-based estimates the correspondence of the top genes between the arrays was poorest. Especially in the hESC array comparison, the superiority of the PECA-method was drastic as compared to the probe set-level methods. For example, with signal log-ratios, the percentage of common genes among top 30 genes was ~25%

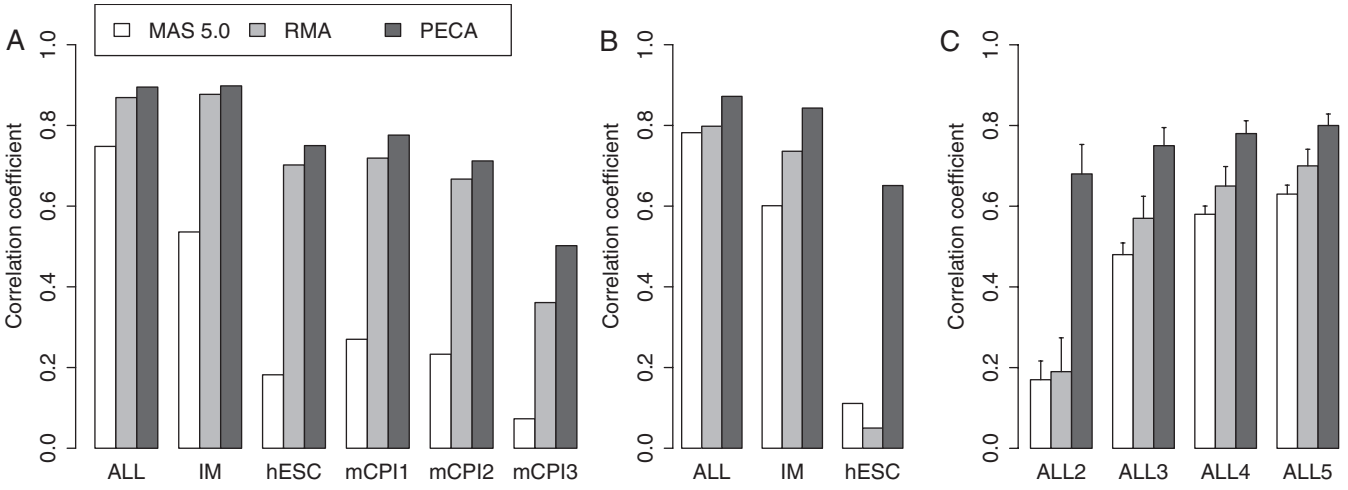


Figure 2. Observed correlations between the expression changes across different arrays as assessed with (**A**) signal log-ratios and (**B** and **C**) Hedges' *g*-estimates. In the ALL array comparison between HG-U95Av2 and HG-U133A arrays, expression changes between two distinct leukemia subtypes (14 samples per group) were calculated. In addition to the whole ALL data set, Hedges' *g*-estimates were calculated for 100 randomly sampled subsets of sizes 2–5 (ALL2–ALL5). Graph C shows the average correlations calculated over these subsets along with their standard deviations. In the IM array comparison between HG-U95Av2 and HG-U133A arrays, expression changes were calculated between patients with dermatomyositis (five samples) and patients with other inflammatory myopathies (nine samples). In the hESC array comparison between HG-U133A and HG-U133Plus2.0 arrays, expression changes between two hESC cell lines (two samples per group) were estimated. In the mCPI array comparison between MG-U74Av2 and MOE430 2.0 arrays, signal log-ratio between an infected and a control sample (mCPI1), between a re-infected and a control sample (mCPI2), and between an infected and a re-infected sample (mCPI3) were calculated. In each two-group comparison, the PECA-estimates of expression changes were compared with the corresponding expression change estimates obtained with RMA- and MAS-based intensity values, which are widely used in microarray data analysis.

with MAS, 60% with RMA and 70% with PECA (Figure 3C). With Hedges' *g*-estimates, there were no common genes among the top 30 genes with either MAS or RMA, whereas the PECA-estimates resulted in ~50% overlap of the genes (Figure 3F). Within the array generations, the proportion of common genes among the top 100 genes between RMA and PECA was typically ~80%, while it was 50% or less between MAS and PECA and between MAS and RMA.

Effect of sample sizes on the meta-analysis performance

Figure 4 illustrates the consistency of the 100 most differentially expressed genes identified using 2–5 biological samples from the ALL data as compared with the genes identified from the whole ALL data. As expected, the agreement among the top genes increased when the number of samples increased. The overall agreement of the results obtained with RMA-intensity values was again substantially lower than the agreement of the PECA-based results. The meta-analysis based on the PECA-estimated Hedges' *g*-values was most stable. With two samples, there were on average over 55% of common genes when the PECA-based meta-analysis was applied but only 35% with the RMA-based meta-analysis. When an individual data set of size 2 was considered, the stability of both approaches was reduced as compared with the meta-analysis. In particular, with the RMA-based analysis, the agreement decreased from 35% with the meta-analysis to ~15% with an individual data set. However, even the meta-analysis could not raise the stability of the RMA-based estimates to the same level as the PECA-estimates. To obtain an agreement of over 50% of genes, the RMA-based meta-analysis typically required four samples, whereas only two samples were needed with the PECA-estimates, even when an individual data set was analyzed.

DISCUSSION

We have introduced a meta-analytic approach, which considers the latest probe-level information when combining the results of multiple Affymetrix microarray studies. We first showed that alternative probe sets provide a good option as compared with the manufacturer-defined probe sets when arrays with different probe collections are compared. Using these alternative sets, we then demonstrated that the comparability of expression changes across different array generations can be considerably improved with PECA-estimation as compared with the estimation based on RMA- or MAS-based summary intensities, especially when the sample sizes are small. The key finding was that by using the PECA-estimates one can more effectively combine the results of individual Affymetrix studies in the context of meta-analysis. In particular, we showed that the consistency of the differentially expressed genes can be improved by integrating PECA-based expression changes across studies. Taken together, these results suggest that available Affymetrix microarray studies of the particular condition can be effectively exploited when analyzing new experiments.

Conventionally, the probe-level expression data are summarized into simple numerical estimates of probe set-level gene expression. A major drawback of this approach is that a substantial amount of probe-level information is discarded. This issue has only lately become a focus of interest. It has been shown that by using probe-level expression data when identifying differentially expressed genes the quality of the resulting gene lists can be improved: Lemon *et al*. (25) and Master *et al*. (26) based their methods on probe-level *t*-tests; Barrera *et al*. (27) applied two-way ANOVA methods to probe-level data; and Chen *et al*. (28) measured probe-level differences in percentiles of ranks. The MAS software also

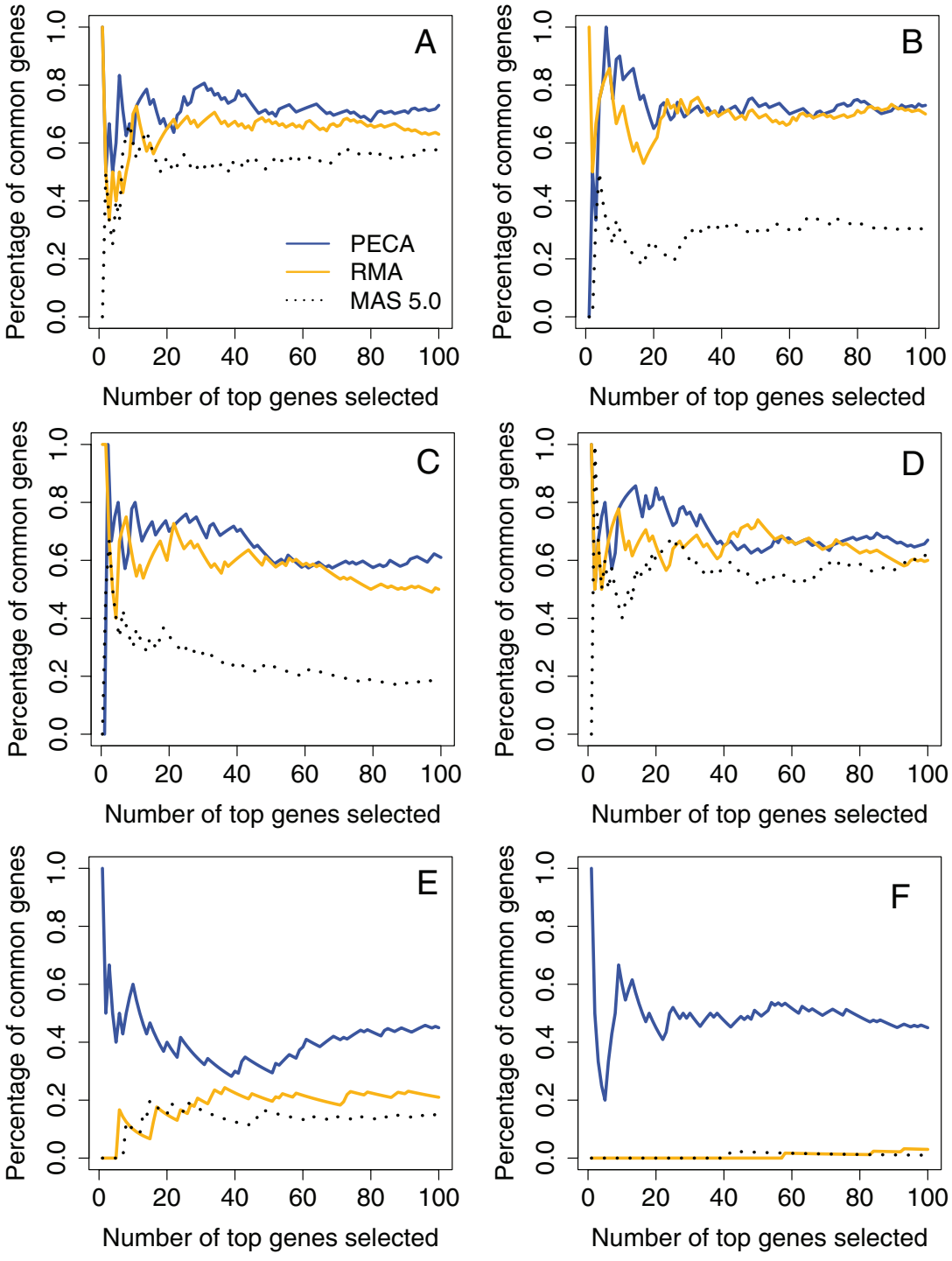


Figure 3. Agreement of differentially expressed genes between technical replicates. The proportion of overlapping genes in the two top *N* lists is plotted as a function of the list size. The genes were ranked with signal log-ratios in (**A**) ALL data (14 samples per group), (**B**) IM data (5 and 9 samples in the groups) and (**C**) hESC data (2 samples per group), and with Hedges' *g*-estimates in (**D**) ALL data, (**E**) IM data and (**F**) hESC data. The observed peaks at the beginning of the curves arise from a single shared top one gene in the two lists.

uses probe-level information in determining differential expression but the algorithm is restricted to comparisons between two arrays only. Our proposed PECA-method can be considered as a generalization of the MAS algorithm and the approaches of (25,26). The method can be used with any number of arrays, and in addition to *t*-statistic, it can improve other measures as well, especially when there are only few samples in the data (see Figure 2). Moreover, the computational burden of PECA is approximately the same as that of RMA- or MAS-normalizations.

We have also carried out an additional study in the Affymetrix spike-in data, where we showed that the

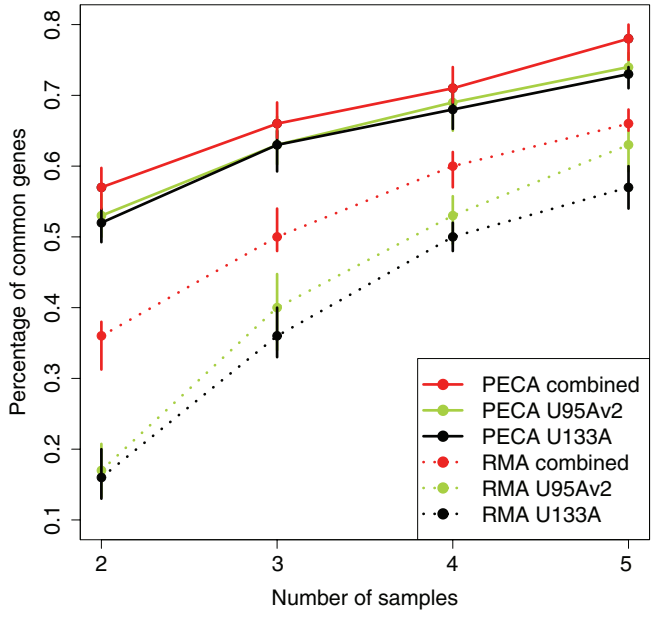


Figure 4. Agreement of differentially expressed genes using Hedges' *g* in the ALL data as the number of samples was varied. The performance was measured by calculating the proportion of common genes among the top 100 genes obtained with the whole data and with randomly selected smaller subsets of sizes 2–5. The results are presented as median percentage over 100 subsets (points) along with the interquartile ranges (error bars). RMA-based (dotted lines) and PECA-based (solid lines) estimation was used with the individual HG-U95Av2 (green) and HG-U133A (black) data sets and with the meta-analysis approaches (red). The RMA-based meta-analysis (RMA combined) represents the meta-analytic approach that has previously been proposed for microarray data (23).

PECA-estimated signal log-ratios and Hedges' *g*-values outperformed the corresponding values calculated from the RMA-normalized intensity values, especially when the sample size was small (see Supplementary Figure 3). In the context of microarray analysis, a common approach to overcome the problem of small sample sizes is to use a modified version of the ordinary *t*-statistic (29). Therefore, we evaluated its performance as well. In general, the PECA-estimated Hedges' *g* performed at least as well as the RMA-based modified *t*-statistic. In particular, with sample sizes 2 and 3, it yielded clearly better AUC-values and the PECA-estimated modified *t*-statistic could not improve its performance further. Although in this study we concentrated on the simple two-group comparisons only, it is possible to generalize the PECA-type analysis to situations, where there are more than two groups to be compared.

In a previous study, Hwang *et al*. (7) suggested that probe filtering could markedly improve the reproducibility of the top ranked genes as assessed with the two-sample *t*-test with unequal variances. After filtering the probe sets according to sequence similarity, they identified 30–40% common genes among the top 20 genes and ~25% common genes among the top 100 genes in the IM data. In our analysis with the PECA-estimated Hedges' *g*, the percentage of commonly identified genes was 40–60% among the top 20 genes and ~45% among the top 100 genes in the same data (see Figure 3E). In general, the percentage of common genes with PECA-estimates was over 40% even when there were only two samples in both groups. With the largest ALL data set, the percentage of common genes increased to 60–80%. These enhanced results clearly demonstrate the importance of the probe-level information in increasing the comparability between array generations. Similar approach could also be used to improve the agreement across different platforms (30).

Similar to (7), we aligned the probes to mRNA sequences with BLAT, which uses heuristics to speed up the search. To evaluate the accuracy of the BLAT search, we aligned the probes of the HG-U133A array also with the Bioconductor matchprobes package, which is based on exact string matching methods. The results obtained with BLAT and matchprobes were virtually the same (BLAT missed 52 of the 241 898 unique probes). The most essential difference between the two methods was in computation time. With an ordinary desktop PC, it took several days to align the HG-U133A probes against human mRNA sequences in Entrez using matchprobes, whereas BLAT made it in hours.

According to our results, the benefits gained from probe verification and alternative mappings are largest when arrays with different probe collections are compared, as in the mCPI, ALL and IM array comparisons (see Figure 1). Although the best match pairs of the original and verified probe sets performed similarly, they rely extensively on manufacturer annotations, including potentially erroneous probes. The alternative probe sets, on the contrary, are expected to refine as the public transcript databases grow in size and improve in accuracy. In the hESC array comparison, correlations between alternative probe sets were somewhat lower than correlations between best match probe sets. This was due to the fact that the original probe sets contained exactly the same probe sequences on both arrays, whereas the alternative probe sets on the HG-U133Plus2.0 array contained also probes that were not included in the HG-U133A array. Also in this case, however, the biological relevance of the alternative probe sets may be higher, since the original probe sets with identical probes would correlate highly even if they were erroneous in biological sense.

Meta-analysis has traditionally been used in medical and social sciences to combine results of different studies (21). Only recently, meta-analysis has also been applied to microarray experiments. Rhodes *et al*. (31) computed gene-specific *P*-values separately for each study and combined them using the Fisher statistic. Choi *et al*. (23) and Stevens and Doerge (32), on the other hand, combined the actual expression data by employing fixed effects and random effects models. In general, a random effects model is more reasonable than a fixed effects model because microarray studies are typically heterogeneous due to, for example, biological variation and differences between experimental methods. However, with only two studies to be combined, which is a typical case with microarrays, we based our integration method on a fixed effects model (33). An analogous approach can be used in the context of a random effects model when there are more studies to be combined.

We showed that the meta-analysis based on the PECA-estimated Hedges' *g*-values was more stable than the Choi *et al*. (23) meta-analysis based on the RMA-estimated summary intensities (see Figure 4). The stability of the methods was evaluated in terms of overlapping top genes obtained when using the whole ALL data set or random subsamples from it. It was assumed that the whole data set provides a

plausible approximation for the true ranking of the genes (24). The spike-in results supported this assumption (see Supplementary Figure 3). Because the reference ranking in the ALL data set was constructed from the same set as the subsamples, the overlap of the top genes might be overestimated with large subsample sizes. As we were interested in the performance of the methods with the smallest sample sizes 2–5, however, such procedure gives valuable information on the stability of the methods. While in this study it was beneficial to have the same samples hybridized to both arrays, the real benefits of the proposed meta-analytic procedure come from combining studies with diverse biological samples.

Previous meta-analysis studies on microarray data have not paid much attention to the quality of the effect size estimates (23,32). With small sample sizes, especially the Hedges' *g*-estimates are prone to unpredictable changes, since gene-specific variability can easily be underestimated resulting in large statistics' values due to chance alone. As only few replications are performed in most microarray experiments, it is critical to improve the effect size estimation with small sample sizes. The general idea of improving the reliability of the microarray results by pooling together results from existing studies is feasible only if the data are properly pre-processed. As probe verification is increasingly used in pre-processing of microarray data or for confirming the final results of a microarray study, it is natural to combine it with other probe-level analysis methods. We demonstrated that summarizing the expression changes over the verified probes only consistently helps in integrating data across studies made with different Affymetrix generations in the same laboratory. The biological findings from the hESC and mCPI data sets are published elsewhere [(34), (Kyläniemi, M., Haveri, A., Vuola, J., Puelakkainen, M. and Lahesman, R., unpublished data)].

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

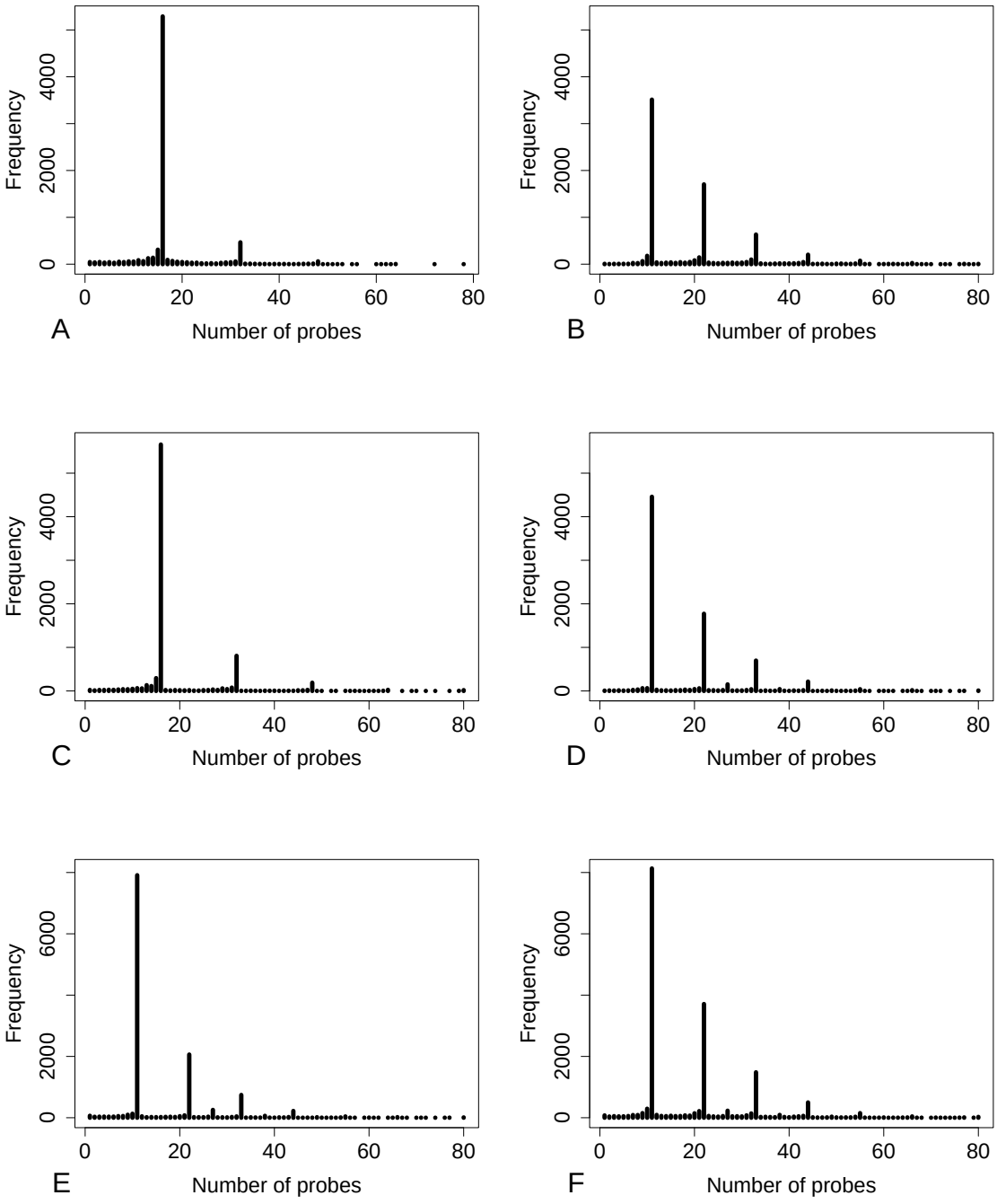
The authors thank Miina Miller, the Finnish DNA Microarray Centre, Turku Centre for Biotechnology, for technical assistance. The work was supported by the Academy of Finland (grant 203632), the National Technology Agency, Turku University Hospital Research Fund and the Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi). Funding to pay the Open Access publication charges for this article was provided by the Academy of Finland.

Conflict of interest statement. None declared.

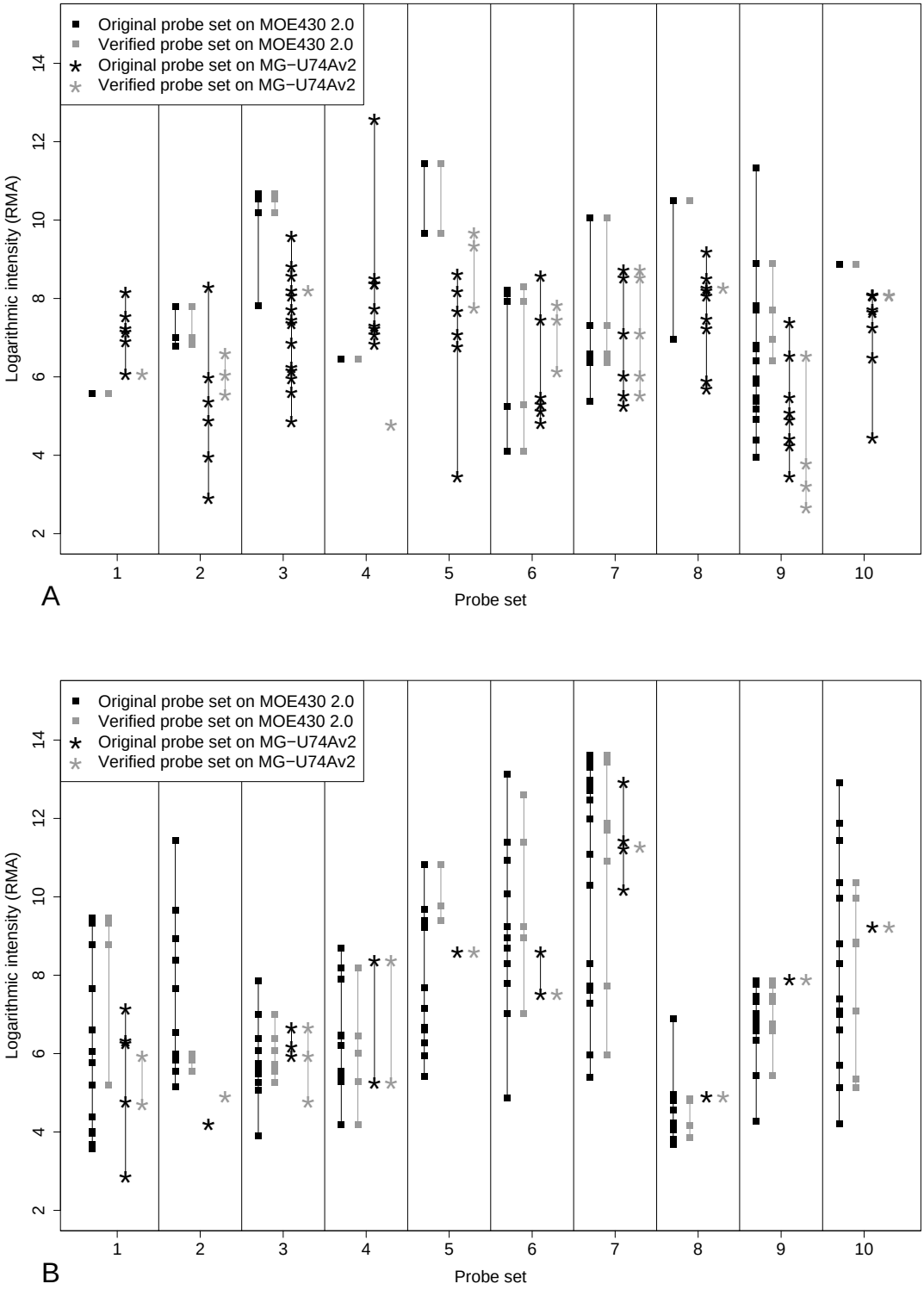
REFERENCES

1. Kothapalli,R., Yoder,S.J., Mane,S. and Loughran,T.P.Jr (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.
2. Kuo,W.P., Jenssen,T., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
3. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G.N., Geoghegan,J., Germino,G. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods*, **2**, 345–349.
4. Bammler,T., Beyer,R.P., Bhattacharya,S., Boorman,G.A., Boyles,A., Bradford,B.U., Bumgarner,R.E., Bushel,P.R., Chaturvedi,K., Choi,D. *et al*. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, **2**, 351–356.
5. Larkin,J.E., Frank,B.C., Gavras,H., Sultana,R. and Quackenbush,J. (2005) Independence and reproducibility across microarray platforms. *Nature Methods*, **2**, 337–343.
6. Gautier,L., Moller,M., Friis-Hansen,L. and Knudsen,S. (2004) Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, **5**, 111.
7. Hwang,K.B., Kong,S.W., Greenberg,S.A. and Park,P.J. (2004) Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, **5**, 159.
8. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res*., **32**, e74.
9. Zhang,J., Finney,R.P., Clifford,R.J., Derr,L.K. and Buetow,K.H. (2005) Detecting false expression signals in high-density oligonucleotide arrays by an *in silico* approach. *Genomics*, **85**, 297–308.
10. Nimgaonkar,A., Sanoudou,D., Butte,A.J., Haslett,J.N., Kunkel,L.M., Beggs,A.H. and Kohane,I.S. (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics*, **4**, 27.
11. Inzunza,J., Gertow,K., Stromberg,M.A., Matilainen,E., Blennow,E., Skottman,H., Wolbank,S., Ahrlund-Richter,L. and Hovatta,O. (2005) Derivation of human embryonic stem cell lines in serum replacement medium using postnatal human fibroblasts as feeder cells. *Stem Cells*, **23**, 544–549.
12. Penttilä,J.M., Anttila,M., Puolakkainen,M., Laurila,A., Varkila,K., Sarvas,M., Mäkelä,P.H. and Rautonen,N. (1998) Logical immune responses to *Chlamydia Pneumoniae* in the lungs of BALB/c mice during primary infection and reinfection. *Infect. Immun*., **6**, 5113–5118.
13. Yeoh,E.J., Ross,M.E., Shurtleff,S.A.Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A., Cheng,C. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.
14. Ross,M.E., Zhou,X., Song,G., Shurtleff,S.A., Girtman,K., Williams,W.K., Liu,H.C., Mahfouz,R., Raimondi,S.C., Lenny,N., Patel,A. and Downing,J.R. (2003) Classification of pediatric lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.
15. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
17. Kent,W.J. (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Res*., **12**, 656–664.
18. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
19. 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.
20. Bolstad,B.M., Irizarry,R.A. and Åstrand, M, Speed,T. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
21. Hedges,L.V. and Olkin,I. (1985) *Statistical Methods for Meta-analysis*. Academic Press, Orlando, FL.
22. Park,P., Cao,Y.A., Lee,S.Y., Kim,J., Chang,M.S., Hart,R. and Choi,S. (2004) Current issues for DNA microarrays: platform comparison, double

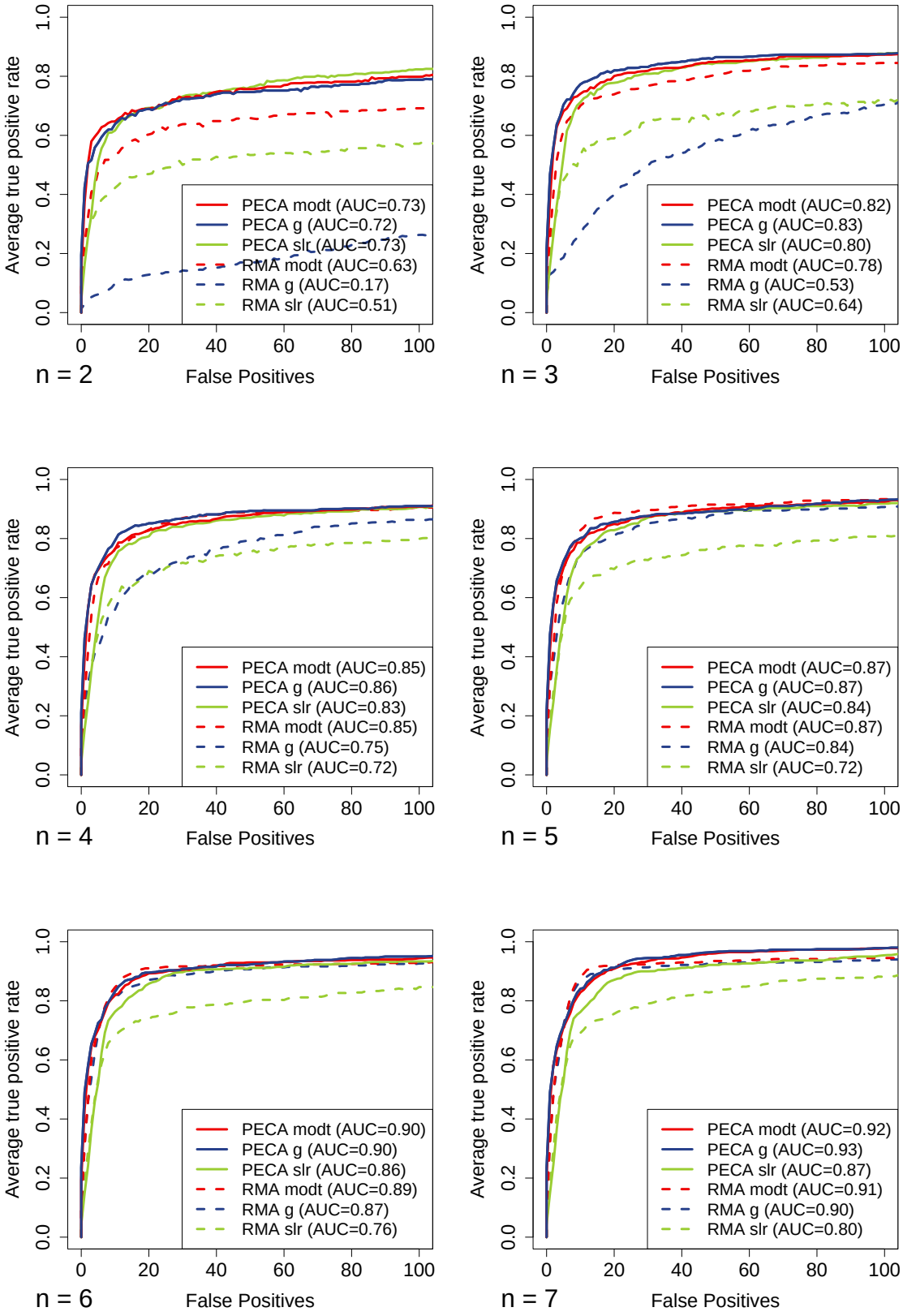
linear amplification, and universal RNA reference. *J. Biotechnol.*, **112**, 225–245.
23. Choi,J.K., Yu,U., Kim,S. and Yoo,O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.
24. Kim,R.D. and Park,P.J. (2004) Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol.*, **5**, R70.
25. Lemon,W.J., Liyanarachchi,S. and You,M. (2003) A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biol.*, **4**, R67.
26. Master,S.R., Stoddard,A.J., Bailey,L.C., Pan,T.C., Dugan,K.D. and Chodosh,L.A. (2005) Genomic analysis of early murine mammary gland development using novel probe-level algorithms. *Genome Biol.*, **6**, R20.
27. Barrera,L., Benner,C., Tao,Y.C., Winzeler,E. and Zhou,Y. (2004) Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays. *BMC Bioinformatics*, **5**, 42.
28. Chen,D.T., Chen,J.J. and Soong,S.J. (2005) Probe rank approaches for gene selection in oligonucleotide arrays with a small number of replicates. *Bioinformatics*, **21**, 2861–2866.
29. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.
30. Tan,P.K., Downey,T.J., Spitznagel,E.L.Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
31. Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.
32. Stevens,J.R. and Doerge,R.W. (2005) Combining Affymetrix microarray results. *BMC Bioinformatics*, **6**, 57.
33. Garrett-Mayer,E., Parmigiani,G., Zhong,X., Cope,L. and Gabrielson,E. Cross-study validation and combined analysis of gene expression microarray data. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, **65**.
34. Skottman,H., Mikkola,M., Lundin,K., Olsson,C., Stromberg,A.M., Tuuri,T., Otonkoski,T., Hovatta,O. and Lahesmaa,R. (2005) Gene expression signatures of seven individual human embryonic stem cell lines. *Stem Cells*, **9**, 1343–1356.



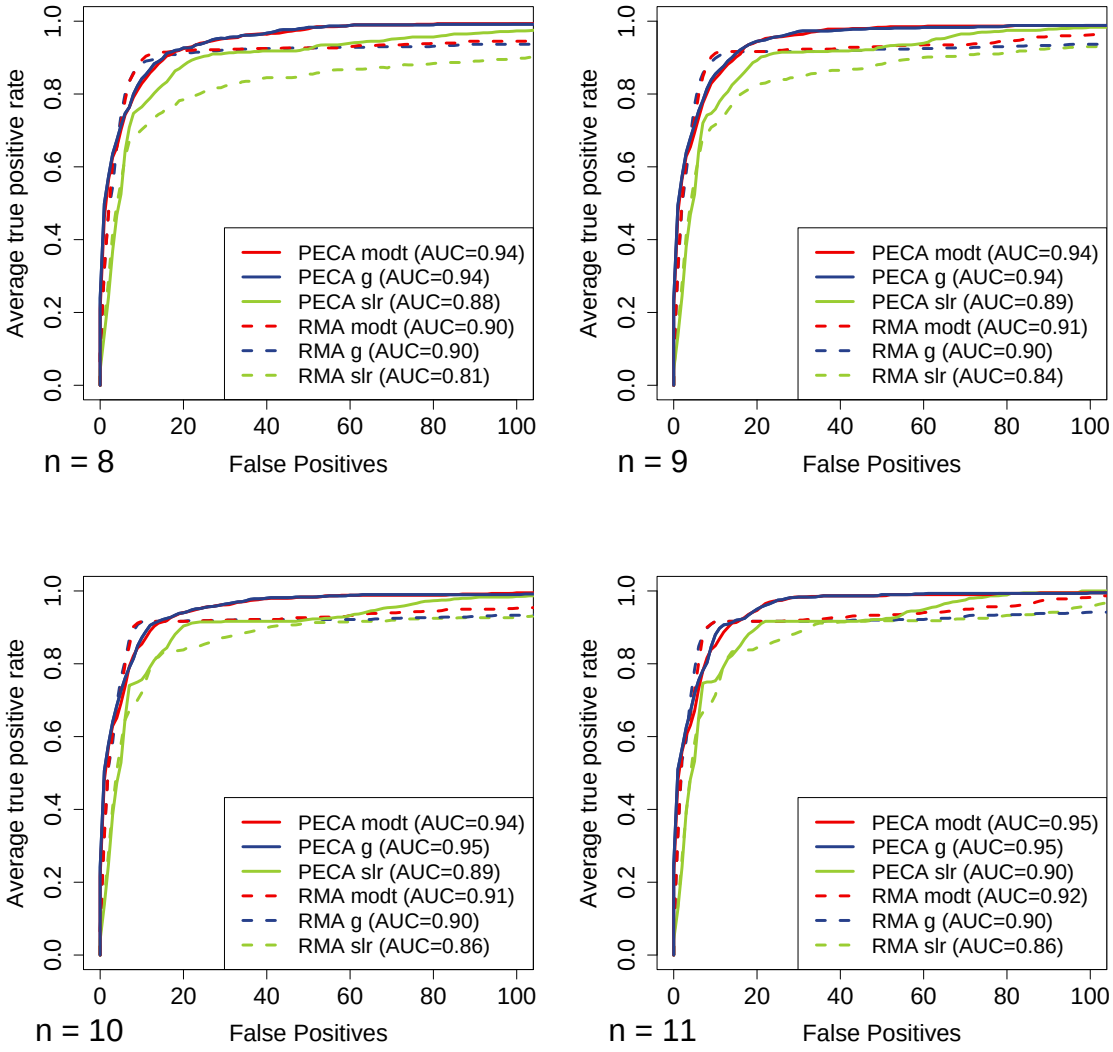
Supplementary figure 1. The probe number distributions of the alternative probe sets on the arrays in the different array comparisons: (A) MG-U74Av2 (mCPI), (B) MOE430 2.0 (mCPI), (C) HG-U95Av2 (ALL, IM), (D) HG-U133A (ALL, IM), (E) HG-U133A (hESC), and (F) HG-U133Plus2.0 (hESC). The manufacturer-defined probe sets on the MOE430 2.0, HG-U133A and HG-U133Plus2.0 arrays contained 11 probes, whereas the MG-U74Av2 and HG-U95Av2 arrays were originally designed to have probe sets of size 16. An alternative probe set contains all the verified probes on an array that match to a unique GeneID. A typical size of an alternative probe set was the size of the original Affymetrix probe set or its multiplier. The proportion of alternative sets with less than 5 probes varied between 0.4 % (MOE430 2.0) and 2.1 % (MG-U74Av2). Thus, small probe sets were not overrepresented on any array. The total numbers of alternative sets included in the array comparisons were 7735 in the mCPI array comparison, 8240 in the ALL and IM array comparisons, and 12661 in the hESC array comparison.



Supplementary figure 2. Variability of the RMA-based intensity values of the probe sets corresponding to the same GeneID for the 10 GeneIDs with the largest numbers of probe sets on the (A) MG-U74Av2 and (B) MOE430 2.0 arrays. The intensity values are shown for the original (black) and verified (grey) probe sets corresponding to the same GeneID on both arrays. The closer the points are along a vertical line, the better is the agreement between the intensity values for the same GeneID on an array. It can be noticed that the probe verification can improve the consistency of the measurements within an array.



Supplementary figure 3.



Supplementary figure 3. (cont.) Averaged Receiver Operator Characteristic (ROC) curves for the PECA signal log-ratio, PECA Hedges' g -statistic, PECA modified t-statistic, RMA signal log-ratio, RMA Hedges' g -statistic, and RMA modified t-statistic (29). The different analysis methods were applied to the Affymetrix HG-U95Av2 spike-in data containing two groups of 12 samples (www.affymetrix.com). In this carefully controlled experiment, it is known that 12 spiked genes are differentially expressed between the two groups, whereas all the other genes are not. Since the truth is known, it is easy to determine true positives (TP) and false positives (FP). We randomly sampled 100 subsets of each possible size from 2 to 11 and determined the average ROC curve over them. The average TP was calculated over the sampled subsets for each FP value, and then plotted against FP. Similarly as Cope et al. (*Bioinformatics* **19**, 185-193, 2003), we restricted the analysis up to 100 FPs, since the lists of genes with more than 100 errors are typically not useful. As a summary statistic, we report the average area under the curve (AUC) up to 100 FPs . The AUCs were standardized so that the largest possible value is 1. The PECA signal log-ratios and Hedges' g -values outperformed the signal log-ratios and Hedges' g -values calculated from the RMA-normalized intensity values, especially with the smallest sample sizes. Moreover, the PECA-estimated Hedges' g performed at least equally well as the modified t-statistic calculated from the RMA-normalized intensity values, being clearly better with sample sizes 2 and 3. The PECA-estimated modified t-statistic could not improve the performance further.