Evaluating preprocessing and differential expression combinations for Affymetrix GeneChip microarrays via spike-in, RT-PCR and cross-laboratory datasets

Ya-Li Wang[1], Guan-Hua Huang[1]*

[1]Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

*Corresponding author

Email addresses:

YLW: yali0629@hotmail.com

GHH: ghuang@stat.nctu.edu.tw

Abstract

Microarray technology for gene expression has been widely used for several years and a large number of computational analysis tools have been developed. We focus on the most popular platform, Affymetrix GeneChip arrays. Despite the rich research on selecting the optimal method of preprocessing and/or detecting differential expression, this paper is unique in the following aspects. First, we have explored suitable combination of preprocessing and differential expression methods. Second, we have evaluated both accuracy and inter-laboratory consistency on a variety of benchmark datasets with distinct characteristics. Third, we have compared stochastic-model-based and physical-mode-based preprocessing algorithms and gene-specific and empirical-Bayes' differential expression detection. We consider popular preprocessing methods: MAS 5.0, PLIER, RMA, dChip and PDNN, and differential expression methods: fold-change, two sample t-test, SAM, limma and EBarrays. Two spike-in datasets and a "real-world-sample" microarray dataset accompanying RT-PCR measurements are used to assess accuracy, and ROC curves are used for the evaluation. To evaluate inter-laboratory consistency, we use a dataset from the MAQC project, which contains arrays generated at two different laboratories using replicated samples. Inter-laboratory overlap rates of differentially expressed gene lists are compared. Our results show that accuracy is more sensitive to preprocessing methods, whereas inter-laboratory consistency is more sensitive to differential expression methods. We conclude that the signal intensity levels are the main factor that explains different performances between methods. We also recommend performing loess normalization at the probe set level.

Keywords: accuracy; inter-laboratory consistency; overlap rate; ROC curve

Introduction

Microarray for gene expression is a device designed to simultaneously measure the expression levels of many thousands of genes in a particular tissue or cell type. It is widely used in many areas of biomedical research, especially Affymetrix GeneChip platform. Millions of probes with length of 25 nucleotides are designed on an Affymetrix array. Two categories of probes are designed: “perfect match (PM)” probe perfectly matches its target sequence and “mismatch (MM)” probe is created by changing the middle (13th) base of its paired perfect match probe sequence. The purpose of designing MM probes is to detect the nonspecific binding because their perfect match partners may be hybridized to nonspecific sequences. A paired PM and MM is called a “probe pair” and each gene will be represented by 11-20 probe pairs typically. Owing to this distinctive design, preprocessing Affymetrix expression arrays usually involves three main steps, which are background adjustment, normalization and summarization. Another fundamental goal of a microarray experiment is to identify those genes that are differentially expressed within different samples. For example, a disease may be caused by large expression of particular genes resulting in variation between diseased and normal tissues. The method used to detect the genes that express differentially between different samples is called the differential expression method.

Various preprocessing and differential expression methods have been proposed, and their developers using different datasets and criteria claim there are some features superior to other methods. Irizarry et al. [1] developed a benchmark and a webtool to permit users to decide the most appropriate preprocessing methods for their application. Their benchmark was based on Affymetrix’s HGU95 and HGU133 Latin square experiments and GeneLogic’s dilution experiment, where gene groups

were spiked-in at various known concentrations and can be used as a reference to evaluate the accuracy of the expression values. They focused on the comparison of various preprocessing methods, not differential expression methods. In the Golden Spike Experiment, Choe et al. [2] presented a new spike-in dataset that had much larger number of spiked-in cRNAs and lower concentration fold changes in spiked-in cRNAs than the Affymetrix Latin square dataset and the GeneLogic dataset. Unlike being done by Irizarry et al. [1] that used each preprocessing algorithm as a whole, Choe et al. [2] considered every compatible combination of the options in various algorithms. Combinations of various preprocessing methods with t-statistic-variant's differential expression methods were implemented to the Golden Spike dataset. They concluded that subtracting MM signals from the PM probe intensities and performing normalization at probe set levels can greatly improve the ability of identifying truly differentially expressed genes. Comparing with the results of Irizarry et al. [1], the first conclusion was in apparent conflict and the probe-set-level normalization was not done there. Pearson [3] outlined six stages in the analysis pipeline to the Golden Spike dataset and conducted a very extensive comparison of various combinations of preprocessing and differential expression methods. He recommended the use of only equal spiked-in probe sets as true negatives and the use of probe-set-level normalization. Spike-in datasets had been criticized for too far removed from real world applications to be very useful. Qin et al. [4] had evaluated the performance of preprocessing methods on microarray datasets from real-world samples, using RT-PCR as the "gold-standard" assay. They concluded that using mismatch data for background adjustment achieved the best agreement between array and RT-PCR. In contrast to previous studies that focused on the identification of genes that expressed differentially under various experimental conditions, the MAQC project [5] aimed at

assessing the reproducibility of gene expression profiling data across platforms and laboratories. They found that the overlap across the two sites in genes that were identified as differentially expressed was high when the genes were selected by rank ordering the genes based on fold change, and, furthermore, gene lists were more consistent than those obtained by t-statistic-variant p-value.

Despite the rich research on selecting the optimal method of preprocessing and/or differential expression, results have been limited or inconclusive for the following three reasons. First, preprocessing and differential gene expression discovery should be regarded as necessarily linked in the sense that preprocessing strongly determines which gene will be found to be differential [3, 6]. However, fewer studies have explored the best combination of two methods. Second, existing studies perform evaluation solely based on one benchmark dataset, which can result in a phenomenon called “over-training” [7]. There is clearly a need to assess methods on a wide range of suitable data sets, so that we can resolve conflicting results in the literature and comfortably extrapolate the conclusions from benchmark datasets to general use. Third, methods based on different models and ideas should be compared. Stochastic-model-based algorithms have been shown to improve the preprocessing of array data, whereas physical-mode-based algorithms also demonstrate the superior performance over stochastic-model-based algorithms [8]. In defining differentially expressed genes, current practice is to perform a gene-specific analysis, however, the empirical Bayes method, which can share information among genes, has gained popularity.

In the current study, we use various benchmark datasets to evaluate combinations of the most popular preprocessing and differential expression detection methods in terms of accuracy and inter-laboratory consistency. This study does not

intent to identify the “best” combination of preprocessing and differential expression detection methods from existing literature. In fact, it is unlikely to identify the best combination because of the huge amount of existing methods and the availability of the software. We aim to explore, under different analytic purposes (accuracy and inter-laboratory consistency) and various microarray datasets with distinct characteristics, the conditions that best fit to preprocessing and/or differential expression combinations. Here we consider four commonly used preprocessing algorithms with each taking a distinct adjustment strategy. They include stochastic-model-based algorithms: Microarray Suite software Version 5.0 (MAS5: [9]), Probe Logarithmic Intensity Error (PLIER: [10]), DNA-Chip Analyzer (dChip: [11, 12]) and Robust Multi-array Analysis (RMA: [13]), and physical-mode-based algorithm: Position-Dependent Nearest-Neighbor (PDNN: [8]). There are five popular differential expression methods considered: Fold-change (FC), two sample t-test, Significance Analysis of Microarrays (SAM: [14]), Linear Models and Empirical Bayes methods (limma: [15, 16]) and Parametric Empirical Bayes methods (EBarrays: [17, 18, 19]). Four benchmark datasets in total are used for evaluation. Two are spike-in datasets used to assess the accuracy: one from Affymetrix Latin square datasets and one from the Golden Spike Project. One “real-world-sample” microarray dataset accompanying RT-PCR measurements from the MAQC project is also used for accuracy. ROC curves are used for the evaluation. To evaluate the inter-laboratory consistency, we use another dataset from the MAQC project, which was generated using samples hybridized to Affymetrix platform at two different laboratories. Overlap rates of differentially expressed gene lists from two laboratories are compared.

Material and methods

Benchmark datasets

HGU133 Spike-in (Affymetrix human genome U133 spike-in dataset)

This dataset consists of 42 arrays, where 14 gene groups have been spiked-in at various known concentrations ranging from 0.125 to 512pM [20]. In addition to 14 spike-in gene groups, a common background cRNA have been added at all arrays. The 14 spike-in gene groups are arranged in the format of a 14×14 cyclic Latin square design with each concentration appearing once in each row and column. Each gene group contains 3 spike-in genes, and each experimental group contains 3 replicates. We consider 42 spike-in genes to be truly differentially expressed genes (DEGs) and other non-spike-in genes to be truly non-differentially expressed genes (NDEGs).

Golden Spike (a wholly defined control spike-in dataset)

Choe *et al*. [2, 21] generated a new control dataset that contains two sets of triplicates hybridizing to Affymetrix GeneChips. The two sets are spike-in samples and control samples, resulting in a total of 6 arrays. This dataset has three main distinct features as comparing to HGU133 Spike-in: (1) 1331 spike-in genes (probe sets) spiked-in at known differential concentrations between the spike-in and control samples, a much larger fraction of gene expression differences; (2) a lower fold changes ranging between 1.2-fold and 4-fold concentration difference; and (3) a defined background sample of 2535 genes presented at identical concentrations in both spike-in and control samples, rather than a biological RNA sample of unknown composition. In this paper, the 1331 genes spiked-in at different concentration between the spike-in and control samples are treated as DEGs. All the others including the 2535 genes spiked-in at 1:1 ratios and the non-spike-in (empty) genes are treated as NDEGs.

MAQC RT-PCR (MicroArray Quality Control project TaqMan Assays versus Affymetrix HG-U133 Plus 2.0 GeneChip)

Data here are taken from the MAQC project [22]. The two RNA sample types: Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR) were distributed among independent test sites with each test site using different microarray platforms to generate gene expression measurements. There were one test site using Applied Biosystems TaqMan Assays and three test sites using Affymetrix HG-U133 Plus 2.0 GeneChip. For TaqMan data, four replicate assays for each of the sample types were processed. For Affymetrix GeneChip data, there were five replicates for each of the sample types. We evaluate the performance of preprocessing and differential expression combinations on Affymetrix GeneChip data using TaqMan results as the gold-standard. Here, we only use the Affymetrix GeneChip data from test site 3 because this test site showed the best agreement with the TaqMan results (Figure 6b in [5]). Only 773 genes that were detected both on the TaqMan assays and on the Affymetrix GeneChip are included in this analysis. The two-sample t-test is performed to compare normalized TaqMan data between the UHRR sample type and the HBRR sample type (Detailed normalized approaches used can be found in Supplementary data of [5]). As a result, 618 genes with false discovery rate <0.05 are considered as DEGs, and other 155 genes are considered as NDEGs. Of note, we have chosen the t-test to determine DEGs and NDEGs in TaqMan data because the MAQC and many other studies also used the t-test for selecting differentially expressed genes in RT-PCR data [23, 24, 25]. We have also used SAM to redo the analysis of detecting differentially expressed genes in TaqMan data. There are 586 genes that are identified as DEGs and 20 genes that are identified as NDEGs by both t-test and SAM; as a result, the concordance rate between two tests in identifying differentially expressed genes is (586+20)/773=0.78. Two test results

are consistent and we are thus confident of differentially expressed genes selected by t-test.

MAQC Rats (MicroArray Quality Control project rat toxicogenomic study)

This dataset is a part of the complete dataset from a rat toxicogenomic study [26], which is one of the reference datasets of the MAQC project [22]. The dataset was generated using 36 RNA samples from rats treated with three chemicals (aristolochic acid, riddelliine and comfrey). In total there were six treatment/tissue groups: kidney from aristolochic acid-treated rats (K_AA), kidney from vehicle control (K_CTR), liver from aristolochic acid-treated rats (L_AA), liver from riddelliine-treated rats (L_RDL), liver from comfrey-treated rats (L_CFY) and liver from vehicle control (L_CTR). Within each treatment/tissue group, there were six biological replicates. Aliquots of these samples were prepared and distributed to each of the five test sites for gene expression profiling using microarrays from one of the four different platforms (Affymetrix, Agilent, Applied Biosystems and GE Healthcare). There are two test sites using Affymetrix platform, and we adopt only the data from these two test sites. Each test site generated 36 arrays respectively.

Preprocessing methods

Preprocessing Affymetrix expression arrays usually involves three main steps: background adjustment, normalization, and summarization. Table 1 gives a summary of the preprocessing methods compared. Software for carrying out the analysis and relevant references are also provided in Table 1.

Differential expression methods

Table 2 gives a summary of the differential expression methods used, including the test model, significance score, software for implementing the method and relevant references. The significance score is used to represent the level of significance of

each gene, and the genes are then ranked by this score when drawing the ROC curve. The significance scores are based on two-sided tests. In the Golden Spike dataset, all differentially expressed spike-ins are up-regulated (the spike-in samples contain increased concentration of spiked-in cRNAs compared to the control samples). Therefore, it is more reasonable to use one-sided up-regulated tests than two-sided tests. However, in typical microarray experiments, there is a balance between up- and down-regulation, and we do not know which gene is up-regulated and which is down-regulated. We are more interested in the performance of differential expression tests regardless the direction of change. As such, two-sided tests are used.

ROC curves for accuracy

The accuracy of a combination of preprocessing and differential expression methods is shown by how well the combination actually measured what it is supposed to measure. To properly compare the combinations in terms of accuracy, the true differentially expressed genes of the dataset must be known. Three benchmark datasets are used for accuracy: two datasets HGU133 Spike-in and Golden Spike that provide the results of spike-in experiments where gene fragments have been added at known concentrations, and the MAQC RT-PCR dataset where the results from the TaqMan are used as the gold standard. The Receiver Operating Characteristic (ROC) curve is used. The ROC curve, which is widely used to evaluate the differential expression methods in microarray analysis, is a graphical plot of the sensitivity (x-axis) versus 1-specificity (y-axis) for a binary classifier system as its discrimination threshold is varied. Sensitivity and specificity are measurements of how well a binary classification test correctly identifies the truth. Sensitivity is defined as the probability that the test lead to make positive decision given that the truth is actually a positive case. This is also known as the true positive (TP) rate. Specificity is defined

as the probability that a negative decision is made when the truth is negative. In other words, 1-specificity represents the probability that the positive decision is made when the truth is negative, and the meaning is equivalent to the false positive (FP) rate. It provides tools to select possibly optimal methods by comparing the area under the ROC curve (AUC). The AUC measures discrimination, that is, the ability of the test to correctly classify those positive case and negative case in fact. The bigger the AUC, the better the overall performance of the test. In this paper, we use absolute TP and FP instead of rates for the ROC curve because it is easier to interpret [1]. Its AUC is then standardized by dividing the total area.

Here we make a brief description of how to accomplish an “average” ROC curve for the HGU133 Spike-in dataset. Different experimental groups imply that the spike-in genes are spiked-in at different concentrations. Here, we restrict attention to the comparison between concentrations with a 4-fold or smaller change. For each pair of experimental groups with a 4-fold or smaller difference in concentration, all 42 spike-in genes have the same concentration fold-change between two groups and thus are considered as DEGs. We compute the number of TPs and FPs for a large range of thresholds. To form an average ROC curve, we compute the average TP number over all these pairs of experimental groups for each FP value. An average ROC curve is created by plotting the FP versus its average TP [1].

Overlap rates for inter-laboratory consistency

The inter-laboratory consistency is based on the degree to which the differentially expressed genes obtained by the combination can be in agreement across laboratories. It is essential to assess this lab-to-lab variability before one can meaningfully merge and/or compare conclusions from different studies. As a result, the high level of inter-laboratory consistency can be a crucial criterion for picking up the preprocessing and

differential expression combinations. To properly compare the method combinations in terms of inter-laboratory consistency, we use the MAQC Rats dataset, where replicated RNA samples from rats treated with three chemicals were distributed to each of the two laboratories for gene expression profiling. For each method combination, we plot the graph where the x-axis represents the number of genes selected as differentially expressed genes and the y-axis represents the overlap rate of two differentially expressed gene lists (from two laboratories) for a given number of differentially expressed genes. The overlap rate is measured using the Jaccard similarity coefficient. For example, when we employ two sample t-test as differential expression method and use p-value 0.05 as the threshold, two gene lists according to the two test sites are produced respectively by collecting the genes which have p-value smaller than 0.05. The numerator of the overlap rate is defined as the number of overlapping genes of two gene lists, and the denominator of overlap rate is defined as the total number of genes in the union of two gene lists. Thus, if there are genes $\{a,b,c,d,e\}$ have p-value smaller than 0.05 for the first test site and genes $\{c,d,e,f\}$ have p-value smaller than 0.05 for the second test site. The overlap rate is $\frac{3}{6}=0.5$.

Four groups of tissues suffering different treatments versus their controls are considered for the differential expression comparison: L_AA vs. L_CTR, L_CFY vs. L_CTR, L_RDL vs. L_CTR and K_AA vs. K_CTR. Patterns for each of the four tissue/treatment groups versus their controls were similar, thus, the “average” graph showing the overlap rates averaging over four tissue/treatment groups is used to facilitate the comparison.

Results

Only part of the area under the ROC curve is used for evaluating accuracy

For the ROC curve of the HGU133 Spike-in dataset, the growth in TP of most combinations became flat gradually when FP>100 (Supplementary figure 1). Furthermore, in this dataset, TP rates of the best performance had reached about 95% when FP=100. As for combinations that performed badly in FP<100, their TP rates can increase as FP>100, but their performances still cannot catch up to those that performed well in FP<100 (Supplementary figure 1). Thus, we focused on the part of FP<100 and reported the AUC up to 100 FPs in the HGU133 Spike-in dataset.

Unlike most microarray experiments that contain only a small percentage of differentially expressed genes, the Golden spike dataset had nearly 10% of the genes to be differentially expressed. Only considering the ROC curve with FP<100 as we did for HGU133 Spike-in was not suitable for this dataset. From Supplementary figure 2, we found that it was reasonable to focus on the part of the ROC curve with FP rate<0.1 (FP $\approx$ 1266).

In the MAQC RT-PCR dataset, there are more differentially expressed genes (618 genes) than non-differentially expressed genes (155 genes). Supplementary figure 3 shows that the curves from all combinations merge together after FP>50. We thus only considered the ROC curve with FP<50.

Two combinations cannot be executed in R, and we had no information about their performance. They were HGU133 Spike-in + dChip(PM-MM) + EBarrays(GG) and HGU133 Spike-in + PDNN + EBarrays(GG). These failures were caused by the convergence of the estimating procedure *emfit* in EBarrays to boundary solutions. Thus, there were 42 combinations for the Golden Spike and MAQC RT-PCR datasets, but only 40 combinations for the HGU133 dataset.

We used the AUC up to 100 FPs in the HGU133 Spike-in dataset, AUC up to the 0.1 FP rate in the Golden spike dataset and AUC up to 50 FPs in the MAQC RT-PCR dataset as criteria for ranking the performance in accuracy. If there was a distinct difference in the AUC between two continuously ranked combinations, combinations were separated from there and divided into two groups. By this way, total combinations were clustered into four groups for the HGU133 Spike-in dataset, three groups for the Golden spike dataset and five groups for the MAQC RT-PCR dataset as shown in Supplementary tables 1-3. Their corresponding ROC curves are shown in Figures 1-3.

Pre-processing methods RMA, PLIER16 and PDNN produce superior accuracy for the HGU133 Spike-in dataset

There were four important findings from analyzing the HGU133 Spike-in dataset. (1) RMA, PLIER16 or PDNN cooperated with most differential expression methods had excellent performances, except when the Welch t-test or t.test was employed as the differential expression method (Supplementary table 1 and Figure 1). (2) Conversely, the combinations with preprocessing methods of MAS5 or dChip(PM-MM) were inferior to other compared combinations. Notice that this inferiority became worse when a differential expression method EBarrays or FC was used. (3) For a fixed differential expression method, performances varied greatly when employing different preprocessing methods, except for t-test and Welch t-test that performed badly persistently. (4) All combinations using the t-test outperformed those using the Welch t-test. The simple differential expression method FC can perform surprisingly well when combining with RMA or PLIER16.

Pre-processing method dChip has the best accuracy for the Golden Spike dataset

Results for the Golden Spike dataset were different from results for the HGU133 Spike-in dataset. (1) dChip had an outstanding performance on Golden Spike (Supplementary table 2 and Figure 2). All combinations were clearly divided into three groups. Especially, combinations with dChip(PM-only) cooperated with every differential expression method were classified into the best group (Supplementary table 2). However, dChip(PM-MM) had an extreme performance. When it was cooperated with differential expression methods: t-test, Welch t-test, limma and SAM, the performance was outstanding; otherwise, it performed poorly. (2) MAS5 combined with FC/EBarrays(GG)/EBarrays(LNN) performed poorly (Supplementary table 2). (3) For a fixed differential expressed method, performances varied greatly when employing different preprocessing methods (Figure 2).

Irizarry et al. [27] showed that genes spiked-in at equal levels had lower fold changes than non-spike-in (empty) genes in the Golden Spike dataset. As pointed out by many researchers, this artifect can invalidate the comparison if the set of all unchanging (equal and empty) probes is used as the true negatives [3, 27]. We thus removed the empty probe sets and only included 2535 genes spiked-in at 1:1 ratios as the true negatives and see if above conclusion still hold. Resulting AUCs and corresponding ROC curves are shown in Supplementary table 4 and Supplementary figure 4. Conclusions are consistent with mild differences. (1) dChip still had an outstanding performance on Golden Spike (Supplementary table 4 and Supplementary figure 4). Combinations with dChip cooperated with every differential expression method were on the top of the ranking (Supplementary table 4). (2) Preprocessing methods that implemented PM-MM (dChip(PM-MM) and MAS5) did pretty well.

Their performances in the HGU133 Spike-in dataset and in the Golden Spike dataset were quite different. (3) PLIER16 performed poorly (Supplementary table 4).

dChip is good for experiments with high signal intensities, whereas RMA and PDNN are good for low signal intensities

To resolve the inconsistence between the conclusions reached by HGU133 Spike-in and Golden Spike, histograms of the log2-transformed signal intensities for DEGs and NDEGs are given in Figure 4. Clearly, DEGs of the Golden Spike dataset had much higher intensity levels than those of the HGU133 Spike-in dataset. For NDEGs, two datasets had pretty close signal intensity levels. We thus hypothesized that the signal intensity levels of probe sets can be the cause of this inconsistence. To verify this hypothesis, we re-plotted ROC curves for the Golden Spike dataset, using DEGs and NDEGs whose intensity levels lay between the 1st quartile and the 3rd quartile of spike-in and non-spike-in intensity levels of the HGU133 Spike-in dataset, respectively (Figure 5). There were 209 DEGs and 5059 NDEGs being selected from Golden Spike for re-drawing. Interestingly, PDNN and RMA outperformed dChip in these “low intensity” ROC curves. We also re-plotted “high intensity” ROC curves for the HGU133 Spike-in dataset, following analogous process above. Only experiment pairs of 9 versus 10, 11 versus 12 and 13 versus 14 had 4 or more spike-in probe sets that had intensity levels falling between the 1st quartile and the 3rd quartile of DEG intensity levels of the Golden Spike dataset. Figure 5 displays “high intensity” ROC curves for the HGU133 Spike-in dataset when comparing experiment11 with experiment 12. We found that dChip(PM-only) outperformed PLIER16, and dChip(PM-only) can outperform RMA and PDNN when cooperated with differential expression methods SAM and limma. From these results, we thus conclude that dChip is good for experiments with high signal intensities, whereas RMA and PDNN are good for experiments with low signal intensities.

Because the absolute intensity level does not have equivalent meaning across platforms/experiments, we thus suggest the readers to compare their preprocessed log2-transformed intensities with Figure 4 or Supplementary table 5 to judge the experiment to have high or low signal intensities.

Probe-set-level loess normalization is recommended, especially for experiments with high signal intensities

Choe *et al*. [2] found that a second normalization at the probe set level generally yielded superior results. Figure 6 shows the ROC curves for expression levels with loess or quantile normalization at the probe set level. For HGU133 Spike-in, the probe-set-level normalization did not change the accuracy. However, for Golden Spike, the loess normalization can improve the accuracy, and PDNN and RMA performed better than dChip when performing the probe-set-level normalization. Therefore, we recommend that preprocessing method RMA or PDNN plus the loess normalization at the probe set level are used for experiments with high signal intensities.

Preprocessing methods PDNN and PLIER16 result in the best agreement with the TaqMan, while differential expression method EBarrays has the worst agreement

Figure 3 displays the ROC curves for the real Affymetrix HG-U133 Plus 2.0 GeneChip data from the MAQC project, using genes that were significantly differentially expressed in TaqMan arrays as DEGs. PDNN and PLIER16 cooperated with SAM/limma/t.test/FC had the highest area under the curve (Supplementary table 3). dChip(PM-MM) had relatively poorer performance than other preprocessing methods. These results were similar to the results from the HGU133 Spike-in dataset. In fact, the average signal intensities for the MAQC RT-PCR dataset were closer to the average signal intensities for the HGU133 Spike-in dataset than to the average signal intensities for the Golden Spike dataset (Supplementary table 5). This is

consistent with our hypothesis that the signal intensity levels are the main factor that explains different performances between methods. In general, combinations with the same preprocessing methods tended to perform similarly except that combinations of most preprocessing methods and EBarrays had the worst performance. The role for the preprocessing methods in accuracy was not as prominent as observed in the HGU133 Spike-in and Golden Spike datasets.

Inter-laboratory consistency depends more on differential expression methods than on preprocessing methods with FC having the best performance

In comparing the method combinations in terms of inter-laboratory consistency in the MAQC Rats dataset, we found that the trends of overlap rate change for most combinations were similar when the number of genes selected as differentially expressed was greater than 10,000. Thus, our comparison in inter-laboratory consistency focused on the value of the x-axis of the overlap plot less than 10,000. From Figure 7, we found that inter-laboratory consistency depended more on differential expression methods than on preprocessing methods because combinations were clustered by differential expression methods. When FC differential expression method cooperated with non-MM-corrected preprocessing methods (RMA, PDNN, PLIER16 and dChip(PM-only)), combinations had highest overlap rates. However, FC with MM-corrected preprocessing methods (MAS5 and dChip(PM-MM)) had a rapidly drop-off consistency. Differential expression methods SAM and limma were also doing well in inter-laboratory consistency, with the overlap rate greater than 0.5. However, differential expression method EBarray did poorly in inter-laboratory consistency.

Source codes for creating ROC curves, overlap plots and histograms of signal intensities are available

We have made the source codes used to create this paper available as Supplementary file 2. R codes for creating ROC curves, overlap plots, and histograms of the log2-transformed signal intensities are contained in Supplementary file 2. This would be useful for readers to use the proposed methods to compare their combinations.

Conclusion and discussion

Through the analysis of spike-in, RT-PCR and cross-laboratory benchmark datasets, our results provide general guidelines for selecting preprocessing and differential expression methods in analyzing Affymetrix GeneChip array data. It seems that accuracy is more sensitive to preprocessing methods, whereas inter-laboratory consistency is more sensitive to differential expression methods. We conclude that preprocessing method dChip is good for experiments with high signal intensities, whereas preprocessing methods RMA and PDNN are good for experiments with low signal intensities. Loess normalization at the probe set level is found to be specifically useful for experiments with high signal intensities. In general, differential expression methods SAM and limma have better performance than other differential expression methods. Differential expression method EBarray usually performs poorly when cooperating with MM-corrected preprocessing methods.

We find that accuracy is more sensitive to preprocessing methods, whereas inter-laboratory consistency is more sensitive to differential expression testing. Hoffmann et al. [28] has also shown that the normalization procedure has a much stronger effect on the subsequent detection of differentially expressed genes than the differential expression test. Possible explanations are: first, the spike-in and MAQC RT-PCR datasets use technical replicates instead of biological ones, which create biases that can only be corrected by appropriate preprocessing. The differential

expression testing that aims to account for random variation thus does not make improvement in accuracy. In the MAQC Rats dataset, groups of six 6-week-old Big Blue rats are used for replications; as a result, these biological replicates contain variations that can differentiate the performance of various differential expression methods. Second, in the spike-in experiments, only 3 replications are made for each condition, which is apparently not enough for statistical testing to reach reasonable power. The MAQC RT-PCR and MAQC Rats datasets contain 5 or 6 replicates and thus provide the differential expression tests to demonstrate their effects.

Our analysis emphasizes that the numerical difference in absolute intensity levels can affect the performances of various preprocessing and differential expression methods. Irizarry et al. [27] has also pointed out the difference in expression intensity levels between HGU133 Spike-in and Golden Spike; thus they do not expect an algorithm that performs well on Golden Spike to necessarily perform well on HGU133 Spike-in, and vice versa. Our results verify their conjecture. After careful examination of two stochastic algorithms RMA and dChip that perform well in HGU133 Spike-in and Golden Spike respectively, we have gained some insights on why their performances vary. First, researchers demonstrate that background noise makes it harder to detect differential expression for genes that are present at low intensities [1]. RMA has made special efforts to first adjust intensities to remove the background effect and then obtain an expression measure using a linear additive model on the background-adjusted intensities, whereas dChip performs a joint modeling of the background noise and the expression measure. For large values of intensities, the joint model is preferable. However, expression measures close to 0 can have particularly large variance and thus the two-stage approach is more appropriate. Second, RMA and dChip both recognize that linear scale measures are not optimal.

RMA applies a linear additive model for log-transformed intensities, and dChip adopts a multiplicative model for original intensities. Both algorithms assume a linear model for intensities on a log scale, but the constant variance assumption of the linear model is assumed on log-transformed intensities in RMA and on original intensities in dChip. It is shown that for low intensity measures the log transformation can help the satisfaction of the constant variance assumption [29]. Third, RMA uses a robust estimating procedure (median polish) to protect against outlier probes, and dChip develops a selection procedure of outlier probes and discard them in estimating intensity measures. It seems two approaches can have different effects on different values of intensities.

It is essential to assess the lab-to-lab variability (the lab effect) before one can meaningfully merge and/or compare conclusions from different studies. Many researches have shown a minimal lab effect [5, 30]. Our results provide extra information: some statistical testing approaches used for detecting differentially expressed genes can inflate this usually small lab effect, whereas different preprocessing procedures perform consistently (to truth or falseness) across laboratories.

There are some interesting findings about differential expression methods. When EBarrays cooperated with a MM-corrected preprocessing method, its accuracy and inter-laboratory consistency are lower than other combinations'. However, the other empirical Bayes approach limma performs excellent no matter what preprocessing algorithms it goes with. Although EBarrays and limma both adopt empirical Bayes approach, their basic modeling frames are different. EBarrays considers a finite mixture and hierarchical model; whereas limma uses a linear model to perform the modified t-test in which the empirical Bayes approach is used to shrink

the estimated sample variances. Our finding can reflect that the noise created by subtracting MM is inflated when implementing the "full" hierarchical Bayes approach, but this noise does not bias the selection of differentially expressed genes when using the standard t-test with modified variance estimates. This conclusion is reconfirmed by the observation that SAM also does well in both accuracy and inter-laboratory consistency even when cooperating with MM-corrected preprocessing methods. Lo and Gottardo [31] also conclude that limma has better performance than EBarrays. They have found that the assumption of a common coefficient of variation across genes in EBarrays can adversely affect the resulting inferences, and have presented an extended model that is shown to have comparable performance to limma.

FC is found to perform good in inter-laboratory consistency but not in accuracy. Guo et al. [26] also find that using fold change, rather than t-test, gets more reproducible. Another interesting observation from Guo et al. [26] is that t-test p-value ranking is dramatically affected by the choice of normalization methods; however, when FC ranking is used for gene selection, the impact of normalization methods on the overlap of gene lists becomes minimal (Supplementary Figure 3 for [26]). Our Figure 7 shows a different picture, where, especially when the number of differentially expressed genes is greater than 100, FC with MM-corrected preprocessing methods has a rapidly drop-off consistency, but limma, SAM and t-test cooperated with all preprocessing methods have a steadily increased consistency. This difference is caused by several reasons. First, Supplementary Figure 3 of [26] compares the overlap of gene lists derived from two normalization methods on data generated *at the same test site*; whereas our Figure 7 shows the overlap rates *between two test sites* that implement a specific preprocessing method. Second, [26] is based on the Applied Biosystems platform and we are based on the Affymetrix. Third, it

seems that the background adjustment (PM-only versus PM-MM) is the major factor that explained difference between methods, not normalization [1].

Authors' contributions

YLW and GHH formulated the original concept, participated in the design of the study, interpreted the results and drafted the manuscript. YLW performed the statistical analysis. All authors read and approved the final manuscript.

Acknowledgements

GHH was supported by grants from the National Science Council of Taiwan (NSC98-2118-M-009-002-MY2 and NSC98-3112-B-001-027), and the Program for Promoting Academic Excellence of Universities in the Ministry of Education of Taiwan (MOE-ATU). We are also grateful to the National Center for High-performance Computing for computer time and facilities.

References

[1] Irizarry RA, Wu Z, Jaffee HA: Comparison of Affymetrix GeneChip expression measures. Bioinformatics, 22, 789-794, 2006.

[2] Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol, 6, R16, 2005.

[3] Pearson RD: A comprehensive re-analysis of the Golden Spike data: towards a benchmark for differential expression methods. BMC Bioinformatics 9: 164, 2008.

[4] Qin LX, Beyer RP, Hudson FN, Linford NJ, Morris DE, Kerr KF: Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. BMC Bioinformatics, 7, 23, 2006.

[5] MAQC Consortium: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol, 24, 1151-1161, 2006.

[6] Steinhoff C, Vingron M: Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform, 7, 166-177, 2006.

[7] Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP: A benchmark for Affymetrix GeneChip expression measures. Bioinformatics, 20, 323-331, 2004.

[8] Zhang L, Miles MF, Aldape KD: A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol, 21, 818–821, 2003.

[9] Affymetrix: Statistical algorithms description document [http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf]. White Papers. Santa Clara, Affymetrix 2002.

[10] Affymetrix: Guide to probe logarithmic intensity error (PLIER) estimation [http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf]. Technical Notes. Santa Clara, Affymetrix 2005.

[11] Li C, Wong W: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA, 98, 31-36, 2001.

[12] Li C, Wong W: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol, 2, research0032.1-0032.10, 2001.

[13] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4, 249-264, 2003.

[14] Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 98, 5116–5121, 2001.

[15] Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol, 3, Article3, 2004.

[16] Smyth GK: Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York, Springer, 397-420, 2005.

[17] Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol, 8, 37-52, 2001.

[18] Kendziorski CM, Newton MA, Lan H, Gould MN: On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Stat Med, 22, 3899-3914, 2003.

[19] Newton MA, Kendziorski CM: Parametric Empirical Bayes Methods for Microarrays. In: The Analysis of Gene Expression Data: Methods and Software. Edited by Parmigiani G, Garrett ES, Irizarry RA, Zeger SL. New York, Springer, 254-271, 2003.

[20] Affymetrix Latin Square Data for Expression Algorithm [http://www.affymetrix.com/support/technical/sample_data/datasets.affx].

[21] The Golden Spike Project [http://www.ccr.buffalo.edu/halfon/spike/index.html].

[22] The MicroArray Quality Control (MAQC) project [http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/].

[23] Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY, Ma Y, Maqsodi B, Papallo A, Peters EH, Poulter K, Ruppel PL, Samaha RR, Shi L, Yang W, Zhang L, Goodsaid FM: Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol, 24, 1115-1122, 2006.

[24] Bosotti R, Locatelli G, Healy S, Scacheri E, Sartori L, Mercurio C, Calogero R, Isacchi A: Cross platform microarray analysis for robust identification of differentially expressed genes. BMC Bioinformatics, 8 Suppl 1, S5, 2007

[25] Wang Y, Barbacioru C, Hyland F, Xiao W, Hunkapiller KL, Blake J, Chan F, Gonzalez C, Zhang L, Samaha RR: Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. BMC Genomics, 7, 59, 2006.

[26] Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM, Hurban P, Phillips KL, Xu J, Deng X, Sun YA, Tong W, Dragan YP, Shi L: Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nat Biotechnol, 24, 1162-1169, 2006.

[27] Irizarry RA, Cope LM, Zhijin Wu Z: Feature-level exploration of a published Affymetrix GeneChip control dataset. Genome Biol, 7, 404, 2006.

[28] Hoffmann R, Seidl T, Dugas M: Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol, 3, research0033.1-11, 2002.

[29] Durbin BP, Hardin JS, Hawkins DM, Rocke DM: A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics, 18 Suppl 1, S105-110, 2002.

[30] Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martínez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W: Multiple-laboratory comparison of microarray platforms. Nat Methods, 2, 345-350, 2005.

[31] Lo K, Gottardo R: Flexible empirical Bayes models for differential gene expression. Bioinformatics, 23, 328-335, 2007.

Figures

Figure 1 - ROC curves for the HGU133 Spike-in dataset.
In the top panel, combinations using the same preprocessing method are assigned to the same color. In the bottom panel, combinations using the same differential expression method are assigned to the same color. Only FPs<100 are shown. There are 40 combinations in total.

Figure 2 - ROC curves for the Golden Spike dataset.
In the top panel, combinations using the same preprocessing method are assigned to the same color. In the bottom panel, combinations using the same differential expression method are assigned to the same color. Only FP rates<0.1 are shown. There are 42 combinations in total.

Figure 3 - ROC curves for the MAQC RT-PCR dataset.
In the top panel, combinations using the same preprocessing method are assigned to the same color. In the bottom panel, combinations using the same differential expression method are assigned to the same color. Genes that are significantly differentially expressed in TaqMan arrays are treated as the true positives. Only FPs<50 are shown. There are 42 combinations in total.

Figure 4 - Histograms of the log2-transformed signal intensities for truly differentially expressed and non-differentially expressed probe sets.
Left panels are plots created from the GHU133 Spike-in dataset with various preprocessing methods, and right panels are plots created from the Golden Spike dataset. In the HGU133 Spike-in dataset, curves of “spikein” are created using 42 spike-in probe sets, and curves of “non-spikein” are using other non-spike-in probe sets. In the Golden Spike dataset, curves of “spikein” are created using 1331 genes spiked-in at different relative concentration between the spike-in and control samples, and curves of “non-spikein” are using 2535 genes spiked-in at 1:1 ratios and the non-

spike-in (empty) genes. In the plot, each “box” represents the boxplot of the corresponding signal intensities.

Figure 5 - ROC curves when using a subset of probe sets.
For the HGU133 Spike-in dataset, we re-plotted the ROC curve of experiment 11 versus 12, using spike-in and non-spike-in probe sets whose intensity levels lay between the 1st quartile and the 3rd quartile of DEG and NDEG intensity levels of the Golden Spike dataset. Only FPs<100 are shown. For the Golden Spike dataset, we re-plotted the ROC curve, using DEGs and NDEGs whose intensity levels lay between the 1st quartile and the 3rd quartile of spike-in and non-spike-in intensity levels of the HGU133 Spike-in dataset. Only FP rates<0.1 are shown.

Figure 6 - Comparison of ROC curves with or without probe-set-level normalization.
Left panels are ROC curves without a second normalization at the probe set level. Middle panels are ROC curves with a loess normalization at the probe set level. Right panels are ROC curves with a quantile normalization at the probe set level.

Figure 7 - Overlap rates between two test sites
In the top panel, combinations using the same preprocessing method are assigned to the same color. In the bottom panel, combinations using the same differential expression method are assigned to the same color. Only the numbers of differentially expressed genes<10,000 are shown. The x-axis (the number of differentially expressed genes) is in the $\log_{10}$ scale.

Tables

Table 1 - Summary of the preprocessing methods compared

Model	Method	Background adjustment	Normalization	Summarization	Software	Reference
Stochastic model	MAS5	Locational adjustment & MM intensities subtraction	Scale normalization	Tukey biweight average	*mas5* function in the Bioconductor *affy* package	[9]
	PLIER16	Probe-specific background adjustment	Quantile normalization	Fit a model with feature response parameters and multiplicative errors	*justPlier* function in the Bioconductor *plier* package. The constant 16 was back to the estimated intensity.	[10]
	dChip (PM-MM)	MM intensities subtraction	Invariant set	Fit a model-based expression index model	*dChip 2006* Software downloaded from the author's webpage	[11]
	dChip (PM only)	PM only	Invariant set	Fit a model-based expression index model	*dChip 2006* Software downloaded from the author's webpage	[12]
	RMA	Convolution (global) background correction	Quantile normalization	A robust linear model is fitted (median polish)	*rma* function in the Bioconductor *affy* package	[13]
Physical model	PDNN	PM only	Quantile normalization	Use a free energy model that accounts for background and signal	*PerfectMatch* Software (Version 2.3) downloaded from the author's webpage	[8]

Table 2 - Summary of the differential expression detection methods compared

Method	Test model	Significance score	Software	Reference
FC	Average probeset intensity ratio	|log fold-change ratio|	R package	
t.test	Two-sample t-test with equal variance	$-\log_{10}$(p-value)	*rowttests* function in the Bioconductor *genefilter* package	
Welch.t	Two-sample t-test allowing unequal variance	$-\log_{10}$(p-value)	*rowFtests* function in the Bioconductor *genefilter* package	
SAM	Modified version of the standard t-statistic to adjust the high variance caused by a low expression level	$\lvert d_{(g)} - \bar{d}_{(g)}^{*} \rvert$, where d_g is the relative difference and $\bar{d}_{(g)}^{*}$ is the expected relative difference from permutation	R *samr* package	[14]
limma	Linear model with empirical-Bayes-adjusted standard deviation	B-statistic used in limma	Bioconductor *limma* package	[15, 16]
EBarrays(GG)	Empirical Bayes gamma-gamma mixture model	Posterior probability of differential expression	Bioconductor *EBarrays* package	[17, 18]
EBarrays(LNN)	Empirical Bayes lognormal-normal mixture model	Posterior probability of differential expression	Bioconductor *EBarrays* package	[17, 18]

Supplementary files

Supplementary file 1 - All supplementary figures and tables of the paper.

Supplementary file 2 - Source codes used to create this paper.
R codes for creating ROC curves and overlap plots, and histograms of the log2-transformed signal intensities are contained in this supplementary file. Please read through the file “readme.txt” before using these codes.