BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 4 2010, pages 486–492
doi:10.1093/bioinformatics/btp685

Gene expression

Mixture-model based estimation of gene expression variance from public database improves identification of differentially expressed genes in small sized microarray data

Mingoo Kim[1,2,†], Sung Bum Cho[1,2,†] and Ju Han Kim[1,2,*]

[1]Seoul National University Biomedical Informatics (SNUBI) and [2]Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Korea

Received on June 10, 2009; revised on December 3, 2009; accepted on December 9, 2009

Advance Access publication December 16, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The small number of samples in many microarray experiments is a challenge for the correct identification of differentially expressed gens (DEGs) by conventional statistical means. Information from public microarray databases can help more efficient identification of DEGs. To model various experimental conditions of a public microarray database, we applied Gaussian mixture model and extracted bi- or tri-modal distributions of gene expression. Prior variance of Baldi's Bayesian framework was estimate for the analysis of the small sample-sized datasets.

Results: First, we estimated the prior variance of a gene expression by pooling variances obtained from mixture modeling of large samples in the public microarray database. Then, using the prior variance, we identified DEGs in small sample-sized test datasets using the Baldi's framework. For benchmark study, we generated test datasets having several samples from relatively large datasets. Our proposed method outperformed other benchmark methods in terms of detecting gold-standard DEGs from the test datasets. The results may be a challenging evidence for usage of public microarray databases in microarray data analysis.

Availability: Supplementary data are available at http://www.snubi.org/publication/MixBayes

Contact: juhan@snu.ac.kr

1 INTRODUCTION

Differential expression analysis of large-scale microarray studies requires more than five replicates in each comparison group for stable results (Hwang *et al.*, 2002; Pavlidis *et al.*, 2003). Many microarray studies, however, are performed with fewer than five samples in each group due to high-cost limitation or scarcity of biological source materials. In the analysis of the small sample-sized microarray data, it is difficult to correctly identify differentially expressed genes using standard group-comparison statistics because estimation of gene-specific variances, with which to determine the statistical significance of observed changes in gene expression, becomes unstable with a small number of replicates.

Many methods have been introduced to address this variance estimation problem. A popular approach has been certain type of regularization of t-test. In the significance analysis of microarrays (SAM) (Tusher *et al.*, 2001), a non-specific small constant is added to all variance estimates so that they are not to be too small. In Cyber-T (Baldi and Long, 2001), a posterior variance in Bayesian framework is used for the variance estimation of a gene combining a prior variance from neighboring genes and a data variance of the gene. Empirical Bayes methods compensate for the small number of replicates by combining information across arrays (Efron and Raftery, 2001; Kendziorski *et al.*, 2003; Maureen *et al.*, 2006). The Bayesian approaches have tried to improve the identification of differentially expressed genes by using information across other genes having similar expression.

On the other hand, a very different approach was suggested by Kim and Park (2004) to estimate the 'natural' variance of individual genes using a large number of experiments performed previously. This became possible with large public databases of microarray experiments such as the Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002) and ArrayExpress (Brazma *et al.*, 2003). This approach has a natural strength over the Bayesian methods in that gene-specific variance is estimated not from the expression of other genes but from the prior values of expression of the same gene.

However, the GEO-adjusted method used the information in GEO database without considering any information in experimental data for estimating gene-specific variances. Moreover, the variance estimate is non-specific to the experimental dataset. Expression variance is not only gene-specific but also condition-specific. While one may want to obtain an estimation of gene-specific variance under certain condition that is comparable to that of the experimental dataset, direct computation over the whole GEO database returns the global variance rather than the variance within the desired condition.

Because GEO database is an aggregate of many experiments across many different conditions, we cannot assume that a gene has a single distribution across the whole GEO database. As demonstrated in Figure 1, the distribution of expression of a gene in GEO database may be composed of multiple distributions. Therefore, it makes more sense to assume that a gene expression has a multi-distributional structure in GEO database, instead of single compositional structure.

In the present study, we performed comparative study about estimating the gene-specific and condition-adjusted variances of

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

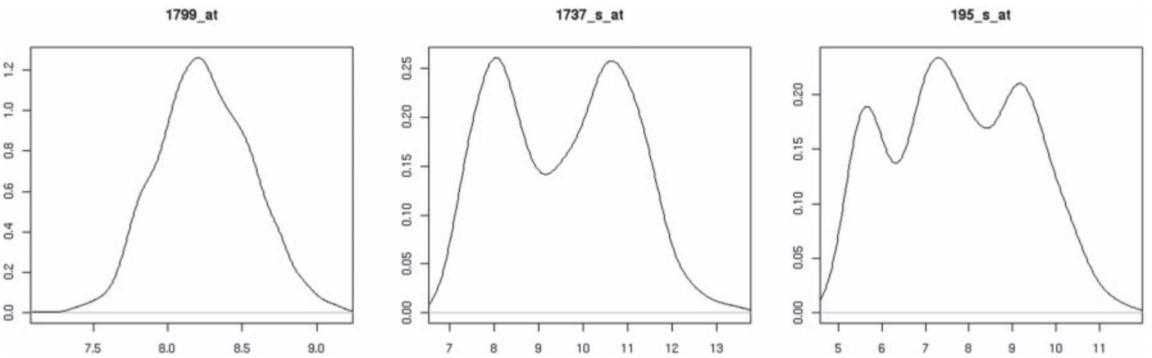

Fig. 1. Examples about distributions of GEO-wide gene expression. Expression density plots were obtained from ∼1400 microarrays present in GEO database. While the probe 1799_at seems to show uni-modal distribution, the probes 1737_s_at and 195_s_at seem to show bi- and tri-modal distributions, respectively. Without application of the Gaussian mixture model, the density *R* function generated these bi- and tri-modal distribution plots. It may not be sensible to estimate gene-specific variance assuming that a gene has a single expression distribution across GEO database. Using Gaussian mixture model, we decomposed the distributions of 1737_s_at and 195_s_at into two and three Gaussian distributions, respectively. In the Affymetrix U95A platform, using Gaussian mixture model, 6173 (48.9%) and 4384 (34.7%) among the 12 625 probes are modeled to have bi- and tri-modal distributions, respectively.

gene expression for two group comparisons in microarray data having less than five samples in each group.

The Bayesian framework improves the previous GEO-adjusted method (Kim and Park, 2004) using both reference and experiment information. We found that G_{MixBayes} outperforms the regularized t-test and the GEO-adjusted methods in the identification of estimating prior variances from GEO database using Gaussian mixture model, and then integrates the priors with the data variances from differentially expression genes (DEGs) in gene expression microarray studies. We propose GEO-MixtureBayesian method (G_{MixBayes} in short), the experimental data into posterior variances. The Gaussian mixture model improves the prior variance estimation from GEO database in terms of performance exploiting the multi-distributional structure in GEO database.

2 METHODS

2.1 Datasets

We used two kinds of datasets, i.e. test and reference datasets. Test datasets are used to compare various benchmark methods. The test datasets should have two groups for comparison and enough number of replicates for each group such that we can reliably make a 'gold-standard' DEG list and have enough number of repeats for repeated microarray sampling.

Reference datasets are used to estimate gene-specific prior variances. They are expected to have large quantity across various conditions such that we can stably estimate the condition-adjuested variability of each gene.

In the present study, we used four test (Table 1) and 33 reference datasets (see Supplementary Table 1). The test datasets includes the prostate and Duchenne muscular dystrophy datasets (Haslett *et al*., 2002; Singh *et al*., 2002) used in the study of Kim and Park (2004), and two more datasets (Strunnikova *et al*., 2005; Stearman *et al*., 2005) having large number of replicates for both comparison groups and with CEL files available. We chose to analyze Affymetrix HG-U95A chip datasets such that we can make more stable comparison and avoid the complexity of between-platform comparisons. We chose reference datasets that have more than 10 replicates and CEL files available throughout GEO database. In total, we obtained 33 datasets containing 1327 microarrays, which is three times as large as that of 471 microarrays in the reference datasets of the study of Kim and Park (2004). We normalized all datasets into a single matrix using the RMA package (Bolstad *et al*., 2003).

Table 1. Summary of test datasets

Dataset	GEO ID	Replicates[a]
Duchenne muscular dystrophy (Haslett *et al*., 2002)	GSE1004 GDS563	21 (10/11)
Macular degeneration and dermal fibroblast response to sublethal oxidative stress (Strunnikova *et al*., 2005)	GSE1719 GDS963	36 (18/18)
Pulmonary adenocarcinoma (Stearman *et al*., 2005)	GSE2514 GDS1650	39 (19/20)
Prostate cancer (Singh *et al*., 2002)	NA/NA[b]	102 (50/52)

[a] Numbers of total (normal/disease) samples.

[b] Data is not available in GEO and downloaded from author's website.

2.2 Benchmark outline

For the purpose of comparison, we used '*repeated sampling procedure*' which was used both by Pavlidis *et al*. (2003) and Kim and Park (2004). First, a gold-standard DEG list was determined for each dataset by standard t-test comparing all samples in the dataset and sorted the resultant DEGs in a descending order by the absolute value of T-statistic. Second, a small number of microarrays ($n = 2$–5) were sampled from each comparison group to obtain DEGs by applying benchmark methods. For brevity, we use the same notation 'NvN' to denote a two-group comparison with N arrays versus N arrays following Kim and Park (2004). The list of DEGs was compared to the gold-standard list. We measured the performance of various methods as the number of top K genes from the gold-standard list that were correctly returned by the methods. This testing on sampled microarrays and comparison to gold-standard lists was repeated. For reliable conclusion, the performances are averaged over 500 repeats. In the present study, the sampling number ranged from two to five.

2.3 Test statistics for differential expression

Test statistics appear in this study can be categorized into four groups. The first group consists of standard methods which have been conventionally used in microarray study and do not consider reference datasets. This group contains mean-fold, standard t-test and regularized t-test. The second group consists of GEO-adjusted methods which replace the data variance with the gene-specific variance estimated from GEO database. This group contains previously proposed GEO-global test and -pooled test, and the newly proposed GEO-mixture test. The third are Bayesian methods which use the posterior variance combining the data variance and the prior variance from reference datasets. The fourth are hybrid methods which re-rank genes based on the merged rank of two different methods. This scheme was proposed by Kim and Park (2004) and claimed to have superior performance when GEO-based methods are merged with regularized t-test. Note that the hybrid methods are different from others in that they just vote and do not calculate any actual statistics. In the followings, we list the formulas of the six basic statistics. These cover all the basic forms and other statistics are simple variants from these.

- Mean-fold (F_{Mean})

$$\mu_{i1}-\mu_{i2} \tag{1}$$

where μ_{i1} and μ_{i2} are means for groups 1 and 2, respectively, for the i-th gene. In log scale, this is equal to fold ratio which is often preferred by biologists.

- Standard t-test (T_{Stan})

$$\frac{\mu_{i1}-\mu_{i2}}{\sqrt{\frac{\sigma_{i1}^2}{n_1}+\frac{\sigma_{i2}^2}{n_2}}}, \tag{2}$$

where n_1 and n_2 are the sample sizes in the groups 1 and 2, and σ_{i1}^2 and σ_{i2}^2 are the variance estimates in the groups 1 and 2, respectively, for the i-th gene.

- Regularized t-test (T_{Reg})

$$\frac{\mu_{i1}-\mu_{i2}}{\sigma_r^+\sqrt{\frac{\sigma_{i1}^2}{n_1}+\frac{\sigma_{i2}^2}{n_2}}} \tag{3}$$

σ_r^2 is the fifth percentile of all variances of the other genes (Kim and Park, 2004).

- GEO-global test (G_{Global})

$$\frac{\mu_{i1}-\mu_{i2}}{\sigma_{Global,i}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{4}$$

- GEO-pooled test (G_{Pooled})

$$\frac{\mu_{i1}-\mu_{i2}}{\sigma_{Pooled,i}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \tag{5}$$

where $\sigma_{\text{Global,i}}^2$ and $\sigma_{\text{Pooled,i}}^2$ are the gene-specific variances for the i-th gene, estimated by the global and pooled methods, respectively, from GEO database (see Section 2.4 for details).

- GEO-mixtureBayesian test (G_{MixBayes})

$$\frac{\mu_{i1}-\mu_{i2}}{\sqrt{\frac{\sigma_{\text{MixPos},i1}^2}{n_1}+\frac{\sigma_{\text{MixPos},i2}^2}{n_2}}} \tag{6}$$

$\sigma_{\text{MixPos},i}^2$ is the posteior variance for the i-th gene, calculated from the prior variance $\sigma_{\text{Mixture},i}^2$ and the data variance σ_i^2 in the Bayesian framework. Please notice that while σ_{MixPos}^2 can adjusts to different data, σ_{Global}^2 and σ_{Pooled}^2 are non-specific to experimental data (see Section 2.4 for the prior estimation and Section 2.5 for the Bayesian integration).

2.4 Estimation of prior variances from reference datasets

$$\sigma_{\text{Global}}^2=\frac{1}{n-1}\sum_{j\in D}\left\{\sum_{k\in D_j}\left(x_{jk}-\bar{x}\right)\right\}, \tag{7}$$

$$\sigma_{\text{Pooled}}^2=\frac{1}{|D|}\sum_{j\in D}\left\{\frac{\sum_{k\in D_j}\left(x_{jk}-\bar{x}_j\right)}{n_j-1}\right\}, \tag{8}$$

$$\sigma_{\text{Mixture}}^2=\frac{1}{|M|}\sum_{y\in M}\left\{\frac{\sum_{k\in M_y}\left(x_{yk}-\bar{x}_y\right)}{n_y-1}\right\}, \tag{9}$$

where σ_{Global}^2 and σ_{Pooled}^2 are the estimates proposed by Kim and Park for gene-specific variances derived from GEO database. In Equations (7) and (8), D and j indicates the reference datasets and the number of the reference datasets. X_{jk} is a gene expression value of the k-th gene. $\bar{x}$ is the mean of the k-th gene of the whole microarray samples of the reference datasets. $\bar{x}_j$ is the mean of k-th gene of the j-th reference dataset. $\sigma_{\text{Mixture}}^2$ is the estimate newly proposed in the present study, estimating variances from the microarray samples re-grouped by Gaussian mixture model. In Equation (9), M and y is the mixture distribution and the number of distributional components in the mixture distribution, respectively. $\sigma_{\text{Mixture}}^2$ is a pooled variance of the each compositional distribution of the mixture model. In our revised Bayesian framework, these variance estimates are used as prior variance. The information we need is not only the gene-specific global variance but also how much the expression of a gene varies within the replicates under certain condition, i.e. *within-condition variance*. Since the GEO database is an aggregate of heterogeneous experiments under different conditions, what σ_{Global}^2 measures is not *condition-adjusted* but the total variance, that is, *the sum of within and between condition variances*. σ_{Pooled}^2, which averages the variances of datasets, also has the risk of measuring the variability between conditions, but in a lesser degree because the heterogeneity of a dataset is expected to be smaller than that of whole GEO database.

On the other hand, we decomposed the distribution of a gene in GEO database into a number of Gaussian distributions, representing *conditions* of the gene. Then we calculated $\sigma_{\text{Mixture}}^2$ by averaging the variances of the Gaussian distributions. We found that $\sigma_{\text{Mixture}}^2$ was much smaller than σ_{Global}^2, indirectly verifying that $\sigma_{\text{Mixture}}^2$ effectively excluded *between-condition variance*. For the computation of Gaussian mixture model, we used Mclust, an R package which uses the EM algorithm for mixture modeling and the BIC criteria for model count determination (Fraley and Raftery, 1999).

2.5 Bayesian framework for variance integration

Bayesian framework has been used for differential expression analysis of microarray study (Baldi and Long, 2001; Gottardo *et al.*, 2003). Previously, the prior variance was estimated from the neighboring genes or fixed to a non-specific value. The GEO database, however, can provide natural estimates of prior. We can estimate the prior of a gene using the prior GEO expression values of the same gene instead of the expression values of other genes. For Bayesian integration of the prior variance and the data variance, we use the fomula by Baldi and Long (2001) that models log-expression values by normal distributions, parameterized by corresponding means and variances with hierarchical prior distributions.

$$\sigma_{\text{Posterior}}^2=\frac{\nu_{\text{Prior}}\sigma_{\text{Prior}}^2+(n-1)\sigma_{\text{Data}}^2}{\nu_{\text{Prior}}+n-2}. \tag{10}$$

In essence, the posterior variance is represented as a weighted average of prior variance and data variance. The parameter, $\kappa=\nu_{\text{Prior}}+n$, determines the degree of confidence in the prior variance σ_{Prior}^2 versus the data variance σ_{Data}^2. Different posterior variances are derived from the same data variance depending on the prior variances. In case of G_{MixBayes}, σ_{MixPos}^2 is derived when $\sigma_{\text{Mixture}}^2$ is used for prior.

3 RESULTS

3.1 Performance comparison

For the four test datasets, our proposed method, G_{MixBayes} is compared to the standard methods, F_{Mean}, T_{Stan} and T_{Reg}, and the previous GEO-adjusted methods, G_{Global} and G_{Pooled}. Sample sizes were chosen from two to five, under which standard methods were known to be ineffective (Pavlidis *et al.*, 2003). This range was also used by Kim and Park (2004).

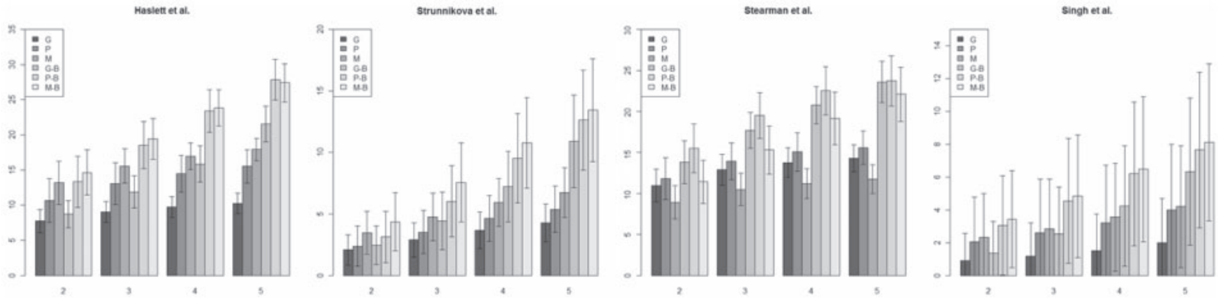

Fig. 2. Performance comparison in four test datasets. Each graph shows the performances of the six methods for each test dataset. The x-axis represents '*the number of samples in comparison*'. The y-axis represents '*the number of gold-standard top 50 genes present in the testing top 50 genes*'. Test methods are denoted with single letters: F_{Mean}, F; T_{Stan}, S; T_{Reg}, R; G_{Global}, G; G_{Pooled}, P; G_{MixBayes}, B. Values are averaged over 500 replications.

3.1.1 Comparison of each dataset Figure 2 demonstrates the performance of the six test statistics for the four datasets. The methods show relatively high performance for the *Haslette's dataset and* relatively low performance for the *Singh's dataset*. This variation is an expected one since the performance is dependent not only on the test statistics but also on the characteristics of the dataset such as the number of samples and the degrees of within-group homogeneity and between-group separation. Despite the variations, common tendencies in the performance curves are well demonstrated with the different datasets. First, G_{MixBayes} shows the best performance in most of the comparisons (13 out of 16 comparisons). Second, regularized t-test shows improved performance than standard t-test. This confirms the previous studies of variance regularization. Third, the performance of the statistics estimating variances from test datasets, T_{Stan}, T_{Reg} and G_{MixBayes} increases as the number of samples is increased. But the performance of the statistics ignoring test dataset variances, F_{Mean}, G_{Global} and G_{Pooled} dose not increase as much as the increment of sample size. Note that G_{MixBayes} is the only method that uses both test and reference datasets for gene-specific variance estimation, while G_{Global} and G_{Pooled} utilizes reference datasets only.

3.1.2 Comparison summary Figure 3 exhibits the summary of the performances in Figure 2. G_{MixBayes} was the best performer across all sample range, returning 17 and 37% more of top genes than the second best F_{Mean} in 2v2 and 3v3 tests, respectively, and 13 and 3% more of top genes than the second best T_{Reg} in 4v4 and 5v5 tests, respectively. G_{Global} and G_{Pooled} were better than T_{Stan} and T_{Reg} in 2v2 test but their improvement in 2v2 was not as large as to be comparable to T_{Stan} in 5v5 and T_{Reg} in 3v3, as claimed in the Kim and Park's study. In each dataset, the Kim and Park's GEO methods outperformed only in Stearman's dataset where the performances of G_{Global} and G_{Pooled} in 2v2 were comparable to T_{Stan} in 5v5 and T_{Reg} in 4v4. The detailed information about the magnitude of improvement of G_{MixBayes} is available in the supplementary web site.

3.1.3 Reproducibility of the results There can be many possible reasons why some of the performance improvements of *Kim and Park* are not observed in our study. First of all, the results of Kim and Park are only based on single test dataset, *Singh's dataset*. As can be seen in Figure 2, there are variations in performance among test datasets. Thus it is probable that the result of Kim and Park are dependent on the specific dataset. Second, the configuration of reference datasets, GEO database specifically, has changed. Kim and Park used ~500 microarrays from GEO for reference, while we used ~1400 microarrays. This difference in reference datasets may affect the performance of the test statistics, especially more for G_{Pooled} and G_{Global} which depend on reference datasets only. This may explain the reason why we failed to replicate the reported performance improvement of G_{Pooled} and G_{Global} in *Singh's test dataset*. The list of the specific 500 datasets used as reference datasets in the study of Kim and Park were not able to be reconstructed from the current 1400 arrays simply because the information was not available (personal communication with the authors). Although T_{Reg} outperforms the Kim and Park's GEO methods in general, the opposite results were observed in 2v2 comparison. In 3v3 comparison, T_{Reg} was not superior to the Kim and Park's GEO methods with the Haslett dataset. We observed similar results with increasing number of DEGs in the test datasets (see Supplementary Material). These results indicated that the Kim and Park's method may outperform with small number of samples ($n < 4$).

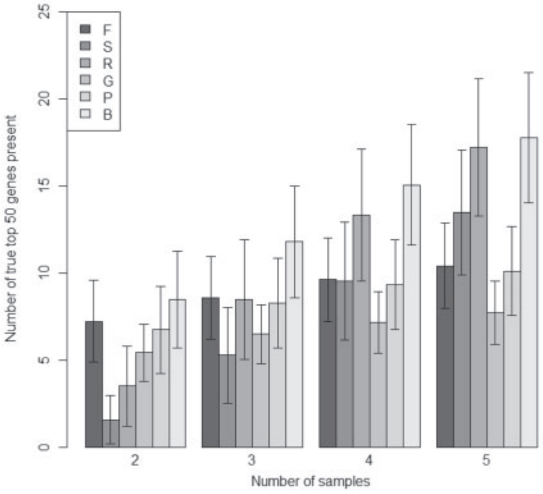

Fig. 3. Performance summary. The performances in Figure 2 are averaged. Test methods are denoted with single letters: F_{Mean}, F; T_{Stan}, S; T_{Reg}, R; G_{Global}, G; G_{Pooled}, P; G_{MixBayes}, B.

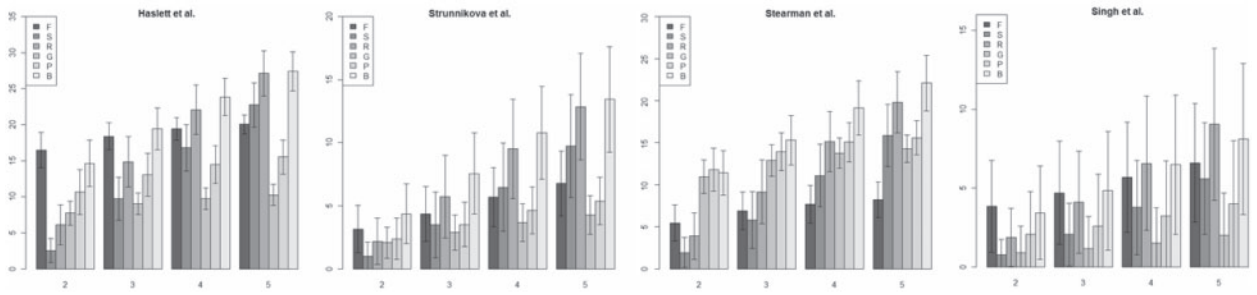

Fig. 4. Sources of improvement. Performance curves of test methods in four test datasets are shown. Dotted lines represent methods after Bayesian integration and closed lines un-integrated versions. The *x*-axis represents the number of samples in each comparison group and *y*-axis represents the number of matched top 50 genes to the gold-standard DEG list. Values are averaged over 500 replications. Test methods are denoted as: G_{Global}, G; G_{Pooled}, P; G_{MixBayes} in dotted lines and G_{Mixture} in closed lines, M.

3.1.4 Comparison in hybrid method with regularized t-test Another advantage of previous *GEO-adjusted* methods, G_{Pooled} and G_{Global}, is that they perform well when combined with T_{Reg} at gene rank level. The hybrid method averaged 75% of the value of the lower rank and 25% of the value of the higher rank to merge ranks in T_{Reg} and the corresponding *GEO-adjusted* methods. This 75%/25% ratio was highly tuned empirically for better results in the previous study (Kim and Park, 2004). In the present study using the proposed 75%/25% ratio, G_{Pooled} and G_{Global} could find 52 and 75% more of the top genes, respectively, in hybrid use with T_{Reg} than in single use (the percentages are averages over the four datasets). The performance of G_{MixBayes}, however, is little improved with hybrid use with T_{Reg}. Specifically, its performance was improved by 13% in *Haslette's*, 1% in *Strunnikova's*, 2% in *Stearmin's* and 1% in *Singh's test datasets*. This limited improvement in the hybrid scheme with G_{MixBayes} is possibly because G_{MixBayes} already incorporates the benefit of T_{Reg} in the level of statistical calculation through the Bayesian integration of test- and reference-dataset variances. There might have been less room for improvement by adding information from T_{Reg} at the level of gene ranks. Although their substantial improvements by hybrid method, G_{Global} and G_{Pooled} outperform G_{MixBayes} only in *Stearmin's test dataset* (*t*-test, $P < 0.05$). In the rest, G_{MixBayes} outperformed significantly (*t*-test, $P < 0.05$).

3.2 Sources of improvement

Here, higher performance of G_{MixBayes} over G_{Pooled} and G_{Global} was well demonstrated in the four test datasets. G_{MixBayes} is modified from earlier *GEO-adjusted* methods in two aspects: estimation of prior variance using Gaussian mixture model and Bayesian integration of both testing and reference dataset variances. To assess the individual contributions of the two modifications for the performance improvement, we performed two additional comparisons:

(1) Comparing G_{Pooled} and G_{Global} to G_{Mixture}, an un-integrated version of G_{MixBayes}

(2) Comparing G_{MixBayes} to the Bayesian version of G_{Pooled} and G_{Global}

In the above comparison, G_{Mixture} used the pooled variance of mixture distributions of the reference datasets without using the experimental dataset. The Bayesian version G_{Pooled} and G_{Global} was computed by substituting prior variance of equation (10) with the $\sigma^2_{\mathrm{Pooled}}$ and $\sigma^2_{\mathrm{Global}}$.

As demonstrated in Figure 4, both of our modifications improved the performance. Bayesian integration improved the performance of all methods in all test datasets. It improved more as increasing number of samples were in test datasets, resulting steep performance curves. This is because the more stable variance estimation of test datasets was used for Bayesian integration. We also found performance improvements when the mixture prior estimation is used alone. In the three methods using the prior variance only, G_{Mixture} outperforms G_{Pooled} and G_{Global} in three out of four test datasets. The improvements were constant across all sample sizes because the sample size only affected the variance estimates of test datasets which were not used for these prior-only methods. The improving effects were additive when Bayesian integration and mixture prior estimation were used together. This may be because they incorporated different information in the steps of variance calculation. The maximum performance was achieved when both Bayesian integration and mixture prior were used.

3.3 The effect of other parameters

There are a number of parameters that may affect the results of the present analysis. We evaluated the performance of the benchmark statistics with the following different parameter set.

3.3.1 GEO change The change of reference datasets in GEO database may affect the estimation. This may be in part the reason why G_{Global} and G_{Pooled} did not perform as well in the present study as claimed in the previous one. To test whether our findings are only specific to current state of GEO, we compared the performance of G_{MixBayes}, G_{Global} and G_{Pooled} across various states of GEO database (Fig. 5). By means of random-sampling from GEO, we compared the methods using reference datasets with 500–1200 arrays. For reliability, we repeated the computation 100 times for each sample size. Five hundred is the same size to the study of Kim and Park, though the composition is not identical. The maximum size that we could produce enough replications from the total of 1400 microarrays was 1200. Across all conditions, we found that G_{MixBayes} consistently outperformed both G_{Global} and G_{Pooled}. Therefore, we believe that the findings are not limited to the specific state of reference datasets that G_{MixBayes} may outperform in the future composition of GEO.

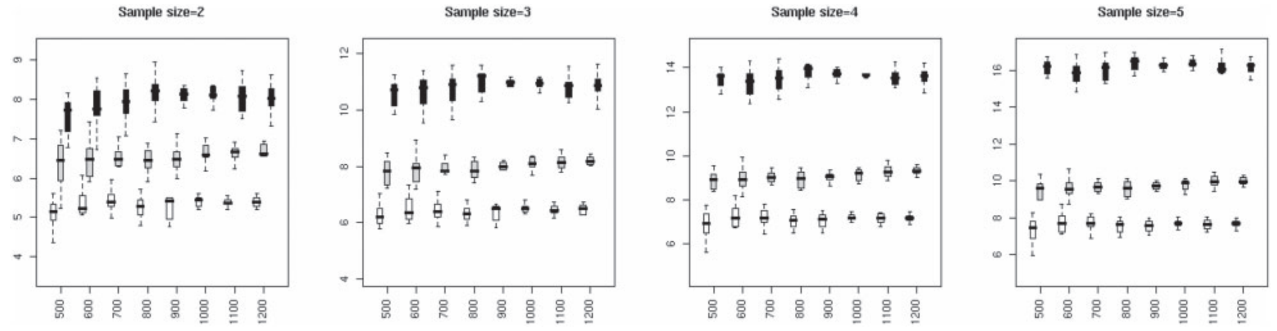

Fig. 5. GEO change. Graph shows the performances of the three GEO-adjusted methods: G_{Global}, G_{Pooled}, G_{MixBayes}, across different numbers of microarrays (the x-axis) in the reference datasets and different number of samples for estimating gene-specific variances. G_{Mixture} consistently outperformed the other two methods. Black box denotes G_{MixBayes}, grey box G_{Pooled} and white box G_{Global}. The x-axis represents the number of microarrays in the reference dataset and y-axis represents the number of matched top 50 genes to the gold-standard DEG list.

3.3.2 Top K genes In this study, the performance of test method is measured as the number of genes common between top 50 in gold-standard DEG list and the test DEG list. Depending on the number of top K genes, the performance can change. We reproduced the results under other top K genes and found that our results are not significantly affected by K (see Supplementary Material).

3.3.3 Prior confidence In Bayesian integration, the prior confidence parameter should be determined by user. We used $\kappa = 6$ in this study. Similar results were obtained in other κ values ranging from 6 to 15 (see Supplementary Material).

4 DISCUSSION AND CONCLUSIONS

In this analysis, we improved identification of differentially expressed genes in datasets having a small number of samples by obtaining prior variance from mixture modeling of microrarray data in the public database.

The success of our methods seemed to come from estimation of prior variances using mixture modeling. The Baldi's Bayesian framework performs well in the microarray data having small samples. This is re-validated in our analysis (Fig. 4). Instead of estimating prior variance based on uni-modal distribution, we selected prior variance from the multi-modal mixture model. As shown in Figure 1, a large number of genes had bi- or tri-modal distributions in their expression values. Simple pooled variances of such genes are likely to be larger than that of a single density of the mixture model because the pooled variance is equivalent to the sum of the variances from each component of the mixture model. This is especially the case with the Kim and Park's method. Therefore, the prior variances were tended to be small and it influenced the posterior variances to be small in the Bayesian estimation. This might contribute to the better performance of our approach because test statistics will increase as the variance decreases with a same mean difference obtained from the experimental data. This may be the same reason for outperforming the regularized t-test.

The number of data is still increasing in public microarray data repositories. With the success of the microarray application in genomic research, more investigators are willing to use microarray experiment. However, the number of samples and cost are the main obstacles to the application of microarray. This research provided a candidate solution for the analysis of small sample-sized data using public repositories.

ACKNOWLEDGEMENTS

This study was supported by a grant from the Ministry of Education, Science and Technology, Republic of Korea (M20706000020-07M0600-02010) and in part by a grant of the Korea Health 21 R&D Project, Ministry of Health, Welfare and Family Affairs, Republic of Korea (0405-BC02-0604-0004). S.B.C's educational training was supported by a grant from the Ministry of Education, Science and Technology, Republic of Korea (M10729070001-07N2907-00110).

Conflict of Interest: none declared.

REFERENCES

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

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

Edgar,R. *et al*. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.

Efron,B. and Tibshirani,R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol*., **23**, 70–86.

Fraley,C. and Raftery.A E. (1999) MCLUST: software for model-based cluster analysis. *J. Classif.,* **16**, 297–306

Gottardo,R. *et al*. (2003) Statistical analysis of microarray data a Bayesian approach. *Biostatistics*, **4**, 577–620.

Haslett,J.N. *et al*. (2002) Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle. *Proc. Natl Acad. Sci. USA*, **99**, 15000–15005.

Hwang,D. *et al*. (2002) Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**, 1184–1193.

Kendziorski,C.M. *et al*. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med*., **22**, 3899–3914.

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

Maureen,A.S. *et al*. (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC bioinformatics*, **19**, 538.

Pavlidis,P. *et al*. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.

Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Stearman,R.S. *et al*. (2005) Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model. *Am. J. Pathol*., **167**, 1763–1775.

Strunnikova,N. *et al*. (2005) Differences in gene expression profiles in dermal fibroblasts from control and patients with age-related macular degeneration elicited by oxidative injury. *Free Radic. Biol. Med.*, **39**, 781–796.

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