Bioinformatics, 32(5), 2016, 682–689
doi: 10.1093/bioinformatics/btv631
Advance Access Publication Date: 30 October 2015
Original Paper

Gene expression

Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes

Ben Li[1], Zhaonan Sun[2], Qing He[1], Yu Zhu[2],* and Zhaohui S. Qin[1,3],*

[1]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA, [2]Department of Statistics, Purdue University, West Lafayette, IN 47906, USA and [3]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 15, 2015; revised on September 30, 2015; accepted on October 26, 2015

Abstract

Motivation: Modern high-throughput biotechnologies such as microarray are capable of producing a massive amount of information for each sample. However, in a typical high-throughput experiment, only limited number of samples were assayed, thus the classical 'large *p*, small *n*' problem. On the other hand, rapid propagation of these high-throughput technologies has resulted in a substantial collection of data, often carried out on the same platform and using the same protocol. It is highly desirable to utilize the existing data when performing analysis and inference on a new dataset.
Results: Utilizing existing data can be carried out in a straightforward fashion under the Bayesian framework in which the repository of historical data can be exploited to build informative priors and used in new data analysis. In this work, using microarray data, we investigate the feasibility and effectiveness of deriving informative priors from historical data and using them in the problem of detecting differentially expressed genes. Through simulation and real data analysis, we show that the proposed strategy significantly outperforms existing methods including the popular and state-of-the-art Bayesian hierarchical model-based approaches. Our work illustrates the feasibility and benefits of exploiting the increasingly available genomics big data in statistical inference and presents a promising practical strategy for dealing with the 'large *p*, small *n*' problem.
Availability and implementation: Our method is implemented in R package IPBT, which is freely available from https://github.com/benliemory/IPBT.
Contact: yuzhu@purdue.edu; zhaohui.qin@emory.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput technologies such as microarray and next generation sequencing have become indispensable tools in biomedical research. These technologies can generate a rich set of information for each biological sample, which can be summarized into a comprehensive picture of the underlying biological processes or systems. However, due to the relatively high cost and complexity of sample preparation, the number of samples surveyed in each experiment is much smaller than the number of features surveyed in each sample. The key characteristic of such datasets can be summarized as 'large *p*, small *n*' (Fan and Lv, 2010). This presents a tremendous challenge when conducting statistical inference on these data such as detecting differentially expressed (DE) genes – a fundamental problem in gene expression data analysis.

Hierarchical models (Good, 1965), which are designed to 'borrow strength' across features, provide a solution under the high-dimensional data setting and have been shown to be effective in dealing with high-throughput genomics data (Kerr and Churchill, 2001; Newton *et al.*, 2001; Parmigiani *et al.*, 2003; Smyth, 2004). The hierarchical model framework has been increasingly utilized in genomics research (Ji and Liu, 2010). It is widely accepted that hierarchical models possess key advantages over naïve, separate inference on individual features. The key idea is that hierarchical models assume that distribution parameters of all features (e.g. probes) are random draws from an upper-level prior distribution (hyperprior). As a result, the posterior distributions of these parameters 'regress' toward the middle. Despite its success, such a strategy could be a double-edged sword. On the one hand, this approach alleviates the problem of poor inference results due to small sample size; on the other hand, it inadvertently introduces biases when estimating the variances of genes that have intrinsic high or intrinsic low variance. It is a reasonable strategy if no additional information except the current experimental data is available. However, in reality, given the explosion of genomics datasets that are publicly available, there is abundant information that can be utilized and should be considered. A unique and fundamental advantage of the Bayesian inference framework lies in its capability to incorporate existing prior information. Bayesian inference achieves seamless integration of prior knowledge and observed data hence is desirable in solving real practical problems (Gelman, 2004). Because technologies like microarray have been widely adopted, there are plenty of publicly available data (referred to as historical data hereafter). We believe such information should be taken advantage of, and the Bayesian framework provides an attractive avenue for implementing such a strategy. Although historical data have been exploited in other contexts (for example, Sui *et al.* (2009) applied a historical database of microarray experiments to adjust background for DNA microarrays), we found none of the existing methods for detecting DE genes explicitly utilizes historical data in a Bayesian setting.

Our hypothesis is that the expression value of each gene (or its surrogate – probe on the microarray) has its unique distribution which reflects its intrinsic biological properties. For example, when historical data collected under diverse conditions were aggregated together, compounded with limited signal range of the microarray technology, measurements of house-keeping genes tend to show high means but relatively small variances across conditions; whereas genes responding to stimuli tend to have large variances since their expression values can go either way. Therefore, assuming proper normalization has been performed across samples, to perform statistical inference, we believe it is perhaps a better strategy to use data that are collected from different experiments but the same gene, than data collected from the same experiment but different genes. To illustrate the point, using 566 normal solid tissue microarray datasets obtained by Affymetrix GeneChip U133A from the global gene expression map of microarray data built by Lukk *et al.* (2010) (details about the datasets can be found in the Results section), we plot standard deviation versus mean on 22 283 probes of their normalized and log-transformed expression values (Fig. 1). We observe a crescent shape in the plot, probes with low or high means tend to have small variance (measured by standard deviation in figures and tables), while probes with mid-level means tend to have large variance. We choose 100 probes from each of the three spots that correspond to low mean/small variance, mid-level mean/large variance and high mean/small variance, respectively, and perform a Gene Ontology (GO) (Ashburner *et al.*, 2000) enrichment analysis on each set of the corresponding genes using DAVID (Huang da *et al.*, 2009a,b). More Details can be found in the Supplementary Tables S1–S3 and Figure S1. The result appears to support our hypothesis. We find that the genes in Group 3 are mostly involved in housekeeping activities evidenced by enriched functional categories such as translation elongation or ribosome-related. Genes in Group 2 are mostly known for being responsive to stimuli. Genes in Group 1 show no functional enrichment, perhaps because they are barely expressed.

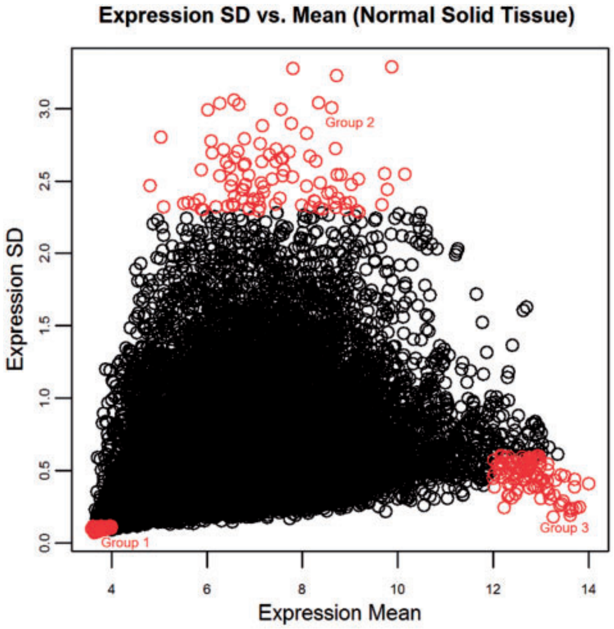

Fig. 1. Standard deviation (SD) versus mean for each probe across 566 normal solid tissue samples. The red zones for Group 1, 2, and 3 represent probes with low means and small SDs, probes with mid-level means and large SDs, and probes with high means and small SDs, respectively. The respective GO term enrichment results are presented in the Supplementary Materials

2 Methods

2.1 Problem setup

In an experiment to identify DE genes, the goal is to detect, among all the genes tested (often in the tens of thousands), the ones that show statistical significant difference in their expression measures between two conditions. For each individual gene, this is a hypothesis testing problem. In a typical study using microarray or other high throughput technologies, the sample size (i.e. number of replicates) is often extremely small, which presents a significant challenge for testing due to unreliable variance estimates. To overcome this, various methods have been proposed in the literature, aiming to obtain more robust variance estimates. For example, SAM (Tusher *et al.*, 2001) proposed to add a small constant to stabilize variance estimation when performing Student's *t*-test. A Bayesian hierarchical model provides an alternative approach by borrowing information from other genes to stabilize estimated variances. In this approach, the variance estimate for each gene can be regarded as the weighted average of the sample variance of this gene and the overall sample variance across all genes. The underlying assumption is that all genes share some commonalities, so much so that the parameters of the prior distributions (of the model parameters) can be regarded as random samples drawn from another distribution (hyper-prior). Such an approach has been widely adopted to analyze microarray gene expression data and other high-throughput genomics data. As an example, Limma is an empirical Bayesian method, which utilizes a hierarchical model to borrow information from other genes so that the resulting estimate of variance can be improved (Smyth, 2004).

2.2 Basic statistical framework

Let X_{ijk} denotes the log-transformed gene expression value after appropriate preprocessing and normalization, where i denotes the gene (or probe), j denotes the condition (control group or treatment group) and k denotes the replicate with $i=1,\ 2,\ldots,I,\ j=1,\ 2$ and $k=1,\ 2,\ldots,\ n$. The basic assumption for the gene expression value is:

$$X_{ijk}|\mu_{i,j},\sigma_i^2 \sim N(\mu_{i,j},\sigma_i^2) \tag{1}$$

where $\mu_{i,j}$ denotes the mean for the ith gene in the jth group and σ_i^2 is the variance for the ith gene. We test whether the mean expression for a certain gene is significantly different between the two groups. For the ith gene, the hypotheses are: $\mathrm{H}_0:\ \mu_{i,1} = \mu_{i,2}$ versus $\mathrm{H}_\mathrm{A}:\ \mu_{i,1} \neq \ \mu_{i,2}$.

In a typical hierarchical model designed for analyzing microarray data, the probability models for gene expression values under the two conditions can be written as follows (Note that the following models are adapted from the ones originally proposed by Ji and Wong (2005) for modeling tiling array data):

$$X_{ijk}|\mu_{i,j},\sigma_i^2 \sim N(\mu_{i,j},\sigma_i^2) \tag{2}$$

$$\mu_{ij}|\mu_0,\tau_0^2 \propto 1 \tag{3}$$

$$\sigma_i^2|\nu_0,\omega_0^2 \sim \mathrm{Inv}-\ \chi^2(\nu_0,\omega_0^2) \tag{4}$$

where the mean parameter μ_{ij} is assumed to be uniform and variance parameter σ_i^2 is assumed to follow an inverse-χ^2 distribution with hyper-parameters ν_0 and ω_0^2. An empirical Bayes shrinkage estimator for σ_i^2 is then used as the variance estimator $\widehat{\sigma_i^2}$, which is subsequently used to perform an adjusted t-test.

2.3 Informative prior Bayesian test (IPBT)

In this study, we propose an alternative approach, which is in some sense 'perpendicular' to the Bayesian hierarchal model for detecting DE genes. Instead of borrowing information from different genes measured in the same experiment, our proposed approach borrows information from the measurements of the same gene in different experiments conducted in the past, using the same technology, same type of chip, on the same type of cells (or similar). The idea of utilizing past experience can be readily achieved under a Bayesian inference framework in the form of prior distributions.

The key idea of our approach is to specify an informative, gene-specific prior distribution for each gene based on abundant historical data and then conduct Bayesian hypothesis testing. Hence, we name our approach informative prior Bayesian test (IPBT). Because different genes have different biological functions, it is often the case that their expression quantities display rather diverse distributions. Therefore, in contrast to the Bayesian hierarchical model, IPBT assumes that each gene has its own unique prior distributions.

2.3.1 IPBT model

The full model is:

$$X_{ijk}|\mu_{i,j},\sigma_i^2 \sim N(\mu_{i,j},\sigma_i^2) \tag{5}$$

$$\mu_{i,j}|\mu_{i0},\frac{\sigma_i^2}{k_i} \sim N(\mu_{i0},\frac{\sigma_i^2}{k_i}) \tag{6}$$

$$\sigma_i^2|\nu_i,\omega_i^2 \sim \mathrm{Inv}-\ \chi^2(\nu_i,\omega_i^2) \tag{7}$$

where (μ_{i0},k_i) and (ν_i,ω_i^2) are the hyper-parameters. The main difference between IPBT and the hierarchical model in (2)–(4) is that here hyper parameters (ν_i,ω_i^2) for the variance $\sigma_i^2 s$ are gene-specific. This gives each gene its specific prior distribution and allows more flexibility. Supplementary Figure S2 in the Supplementary Materials summarizes the difference between IPBT and the Bayesian hierarchical model. In IPBT, for each gene, the parameter of interest is σ_i^2 for which we infer using a Bayesian procedure. Details about the model and its inference procedure can be found in the Supplementary Materials (Section 2.1).

2.3.2. Inference and computation

Substantial amount of microarray data have been accumulated in public repositories such as gene expression omnibus (GEO) and ArrayExpress. Using such historical data, we can obtain gene-specific and informative prior distributions for most of the cell types studied.

Applying IPBT, we perform statistical hypothesis testing to detect DE genes in the form of Student's t-test (with adjusted variance estimates) to allow a direct and fair performance comparison with other existing methods. The test statistics is:

$$t_i^* = \frac{\overline{X}_{i1} - \overline{X}_{i2}}{\sqrt{\frac{2\widehat{\nu_i^*}/(\widehat{\nu_i^*}-2)}{n}\widehat{\omega_i^*}^2}} \tag{8}$$

where $\overline{X}_{i1}$ and $\overline{X}_{i2}$ are sample means for control and treatment group, respectively. $\widehat{\nu_i^*}$ and $\widehat{\omega_i^*}$ are estimates of the posterior distribution parameters.

The adjusted variance estimate is essentially the weighted average of the estimated variances obtained from historical data and current data, respectively. This indicates that IPBT indeed enables natural integration of historical data into the current experiment to assist in DE gene detection. Because we choose conjugate priors (i.e. normal inverse-χ^2 distributions) in IPBT, the posterior distribution can be directly estimated. More details can be found in Section 2.1 of the Supplementary Materials. We show that using adjusted t-test is equivalent to using Bayes factor in terms of ranking DE genes in our model (details can be found in Section 2.2 of the Supplementary Materials).

3 Results

3.1 Simulation

We conduct a simulation study to compare IPBT with four alternative methods for detecting DE genes: (i) Student's t-test, (ii) SAM, achieved by R package 'siggenes'; (iii) Limma, achieved by R package 'Limma'; and (iv) Z test using the true variance.

3.1.1 Comparison of variance estimation between hierarchical model and IPBT

To illustrate the impact of different methods on genes' variance estimation, we conduct the following simulation study. Using the mean and standard deviation obtained from normal solid tissue samples in the global gene expression map of microarray data developed by Lukk *et al.* (2010), we simulate two samples of expression data and treat them as current control data. We randomly select ten samples from normal solid tissue samples and use them as historical data when estimating standard deviation with IPBT. Figure 2 shows the plots of standard deviations obtained using various methods versus

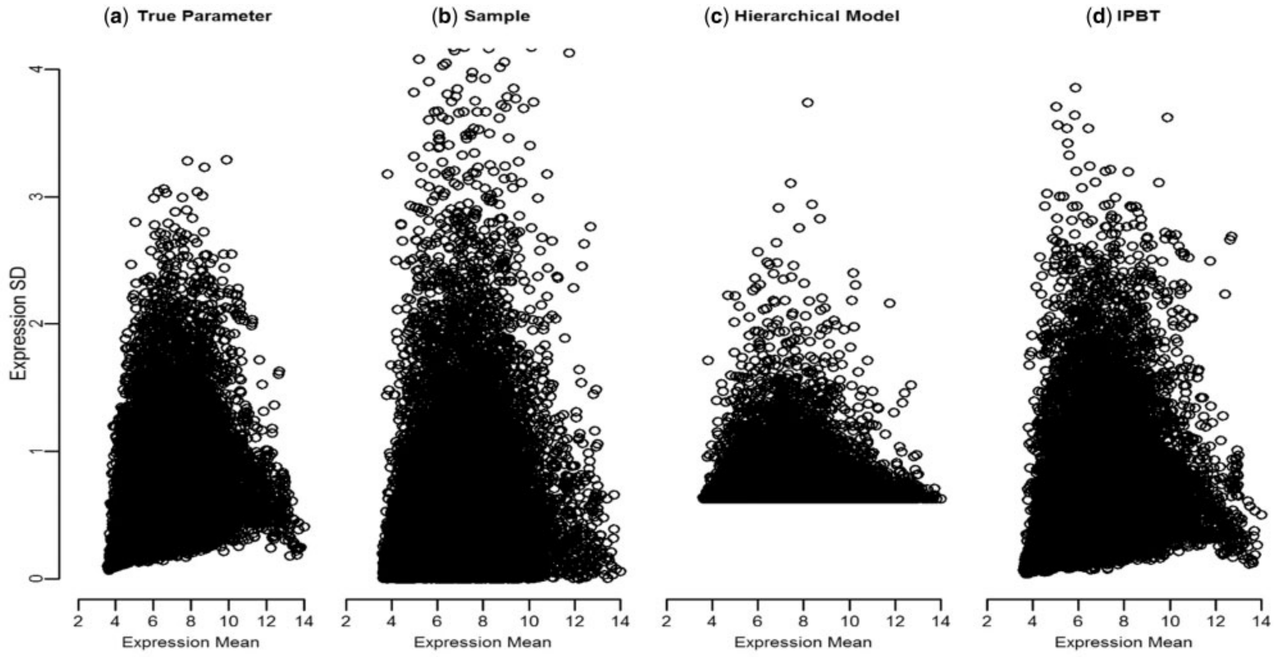

Fig. 2. Standard Deviation (SD) Estimates generated from different methods with probes sorted by their true expression mean values. (a) True SDs (b) Sample SD of the current samples (c) SD estimates from Bayesian hierarchical model (d) SD estimates from IPBT

their true means. Figure 2(a) shows the pre-specified true standard deviation of each gene versus its true mean expression value. Figure 2(b) shows the sample standard deviations calculated from the two 'current' samples, which include extreme small standard deviations caused by limited sample size. Figure 2(c) gives the standard deviations estimated from the Bayesian hierarchical model, which show shrinkage towards the middle effect compared to Figure 2(b) and clearly suffer from the over-shrinkage problem. Figure 2(d) shows the variance estimates from IPBT, which show little over-shrinkage problem.

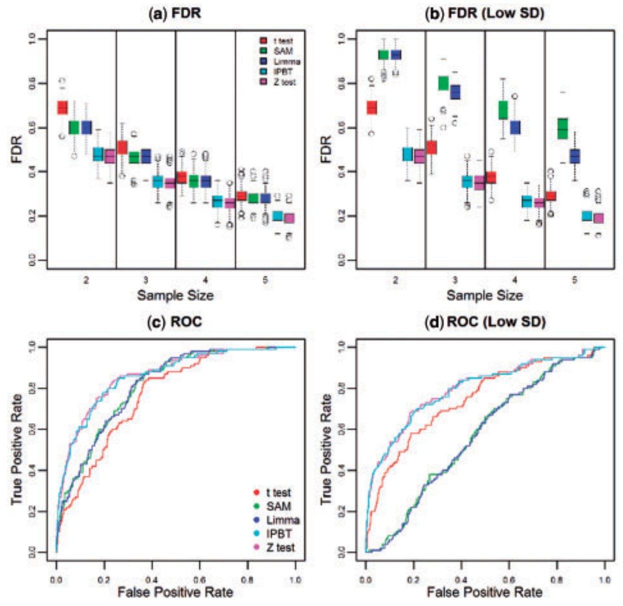

Fig. 3. FDR for detecting DE genes comparing various methods with different sample size for **(a)** random chosen DE genes and **(b)** low standard deviation DE genes. ROC curves for detecting DE genes comparing different methods in one simulation for **(c)** random chosen DE genes and **(d)** low standard deviation DE genes

3.1.2. Simulation strategy

The simulation study considers 1000 genes and k (ranging from 2 to 5) samples for both the treatment and control groups. We randomly select 10% of the 1000 genes (i.e. 100 genes) as designated DE genes. Gene expression values in both the treatment and control groups are assumed to follow normal distributions. The distribution parameters are obtained from real data in the global gene expression map. First, 1000 genes are randomly selected (without replacement) genome-wide. Then for each gene, we derive its sample mean and sample variance from the 566 normal samples in the collection. For the treatment group, the mean and variance of a gene's expression value are assumed to be equal to their counterparts in the control group except for the 100 DE genes for which the mean expression values are set to be two standard deviations higher. For historical data used by IPBT, we first randomly select 188 normal samples out of 566 (without replacement) from the global gene expression map, then obtain their gene expression values corresponding to the 1000 genes selected earlier.

3.1.3. DE gene detection result

To evaluate the performance, we calculate the empirical false discovery rate (FDR) (Benjamini and Hochberg, 1995; Tusher *et al.*, 2001) (also known as false discovery proportion – the proportion of incorrect DE calls among all the ones called) from the top 100 genes ranked by the test statistics. Detailed definition of empirical FDR (simply referred to as FDR from now on) is in the Supplementary Materials. The simulation procedure is repeated 500 times for each method. The distributions of the 500 FDRs of the methods are summarized using box plots and shown in Figure 3(a). Our method clearly outperforms all other methods except for the Z test using true variances (considered the gold standard). The performances of our method and Z test are fairly close. Remarkably, the FDR of DE genes

Table 1. Consistency for detecting DE genes

Overlap times	3	4	5	High (4 + 5)	Moderate (3 + 4 + 5)
Student's *t*-test	31.20	14.64	2.92	17.56	48.76
SAM	28.43	25.24	10.55	35.79	64.22
Limma	28.63	25.67	10.56	36.23	64.86
IPBT	**33.12**	**30.39**	**11.54**	**41.93**	**75.05**
Z test	33.31	31.47	11.23	42.70	76.01

Overlap times (Low variance)	3	4	5	High (4 + 5)	Moderate (3 + 4 + 5)
Student's *t*-test	21.55	6.87	0.83	7.70	29.25
SAM	17.28	5.73	0.83	6.56	23.84
Limma	18.02	6.13	0.93	7.06	25.08
IPBT	**33.57**	**28.67**	**10.69**	**39.36**	**72.93**
Z test	33.28	32.49	12.71	45.20	78.48

* The best results (after excluding *Z* test) are in **bold**.

detected by IPBT is even smaller than the FDR of DE genes detected by the Student's *t*-test with larger sample size (i.e. increased by one).

We use Receiver Operator Characteristic (ROC) curves to further compare IPBT with the other methods. Figure 3(c) shows a typical ROC curve for one single simulation with two replicates. Detailed area under the curve (AUC) corresponding to Figure 3(c) is listed in Supplementary Table S7 ('Random Choice' column). The ROC curves again show that IPBT performs better than all the other methods in detecting DE genes except for the *Z* test, and the performances of our method and the *Z* test are similar. More results can be found in the simulation section of the Supplementary Materials (Supplementary Fig. S3; Table S6).

We further compare the consistency and stability of these methods in detecting DE genes. In each simulation, historical data remain unchanged, but five different sets of the control and treatment data were generated from the same underlying distributions. For each set of the control and treatment data, we apply all four methods to detect DE genes. We summarize the number of overlaps among the five lists of DE genes. The simulation procedure is repeated 500 times, and the average number of overlaps is used as a measure of consistency in detecting DE genes. We consider an average number of overlaps greater than or equal to four as an indication of high consistency and greater than or equal to three as moderate consistency. The average numbers of overlaps are reported in Table 1 which again shows that IPBT outperforms other methods except for the *Z* test in consistency, and the performances of IPBT and the *Z* test are close.

3.1.4 Impact of inaccurate variance estimation

In the previous simulation study, for each gene, we use the same distribution to generate current data and historical data. This represents an idealistic scenario and may not hold true in reality. To examine the robustness of our method, we conduct an additional simulation study in which both parameters in the normal distribution that produces the historical data are shifted such that the distributions that generate historical data and current data are no longer identical. The amount of shift is randomly drawn from a uniform distribution in the interval of (−20%, 20%). We investigate three types of noise-added historical data: unbiased, over-dispersion and under-dispersion. Supplementary Figure S5 in the Supplementary Materials shows that IPBT with noisy historical data still outperforms other methods. Although the performance of IPBT deteriorated when noisy historical data are used, it is still better than Student's *t*-test, SAM and Limma in terms of FDR and is close to the gold standard *Z* test result in all scenarios. This result demonstrates the robustness of IPBT and implies its broad applicability even with potentially noisy historical data.

3.1.5 Detect DE genes with low intrinsic variance

As Figure 2 demonstrates, Hierarchical model-based methods inflate the variance of the genes which have intrinsic low variance hence lower power to detect DE genes of this kind. IPBT, on the other hand, does not suffer from this shortcoming. To further investigate how over-correction affects the detection of DE genes, we conduct another simulation study under the scenario that the DE genes have low intrinsic variance, and the results are reported in Figure 3(b), (d), Supplementary Table S7 ('Low Variance' column) and Table 1 (Low variance). All the results show that Bayesian hierarchical model performs even worse than Student's *t*-test, whereas IPBT maintains superior performance that is similar to the performance of the *Z* test. These results confirm the robustness of IPBT because it avoids the 'over-correction' issues for those genes with low intrinsic variance. More results can be found in the simulation section of the Supplementary Materials (Supplementary Fig. S4; Table S6).

3.2 Real data analysis

3.2.1 The global gene expression map

Lukk *et al.* (2010) built a global gene expression map that includes microarray data from 5372 human samples and contains 369 different tissues, cell lines and disease states. Among them, we calculate informative priors for 96 groups with at least ten samples including normal solid brain tissue and normal solid heart tissue. The dataset (processed and normalized by robust multiarray analysis (RMA) (Irizarry *et al.*, 2003)) was downloaded from arrayExpress (ID: E-MTAB-62). The 96 groups have a median sample size of 25.5 and a mean sample size of 48. More details about the gene expression map can be found in the Supplementary Materials.

3.2.2 Comparison of current and historical data

Our model assumes that the historical data is informative for estimating gene expression variance. We validate this using two sets of data from the global gene expression map. One set contains all the data from heart, and the other one contains all the data from brain. For each set, we download the raw data (CEL files) and subsequently process and normalize the data using RMA (Irizarry *et al.*, 2003), using R package 'oligo'. In the first study, we randomly choose five normal heart samples (out of 36) and five disease heart samples (out of 51) and use them as the current data. Data from the 31 remaining normal heart samples are used as historical data. Figure 4(a) and (b) shows the standard deviations of the genes in the control group (normal samples) and treatment group (disease samples) against historical data, respectively. The strong positive correlation patterns demonstrated in the plot confirm that using historical information as informative priors in the inference procedure is feasible. We also conduct similar analyses on the brain samples. We randomly choose five normal brain samples (out of 39) as controls and five disease brain samples (out of 31) and use them as the current data. Data from the 34 remaining normal brain samples are used as historical data. Corresponding results

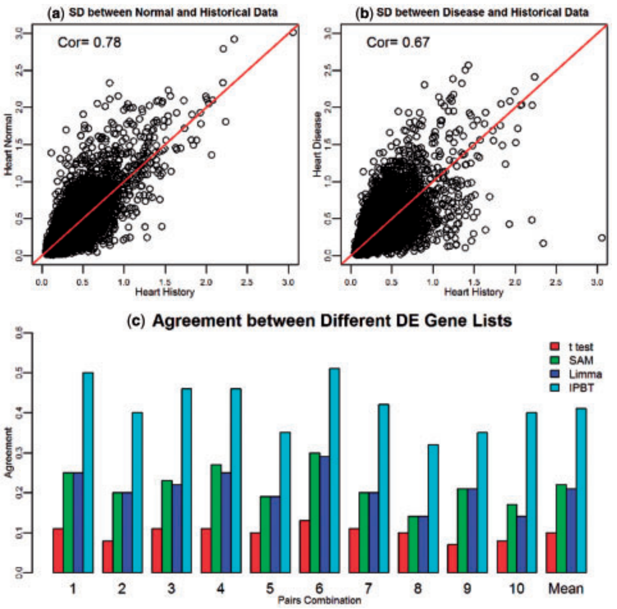

Fig. 4. (a) Comparison of standard deviations (SD) obtained from the five heart normal samples and that obtained from the heart historical data. **(b)** Comparison of SDs obtained from the five heart disease samples and that obtained from the heart historical data. **(c)** Agreements between all pair combinations of top 1000 genes from all 5 DE gene lists

which display similar pattern are shown in Supplementary Figure S6(a), (b).

3.2.3 DE gene detection

For real data, since it is not possible to know what the real DE genes are, we use agreement as the measurement of performance. This strategy has been commonly used in microarray data analysis studies (Lim *et al.*, 2015; Lim and Wong, 2014). In this study, again using the global gene map data, we randomly select two normal heart samples and two disease heart samples. Data from the remaining 34 normal heart samples are used as historical data. We then apply IPBT and competing methods on these data to obtain a list of top 1000 DE genes for each method. We repeat the above sampling and testing procedure five times. Then for each method, we calculate the agreement between every pair of the 1000 DE gene lists. Figure 4(c) summarizes the results, which shows significant higher agreement for our IPBT method compared to others. We also compared the DE gene calling consistency as we did in the simulation study, and the results are summarized in Supplementary Table S8. Again, IPBT performs the best among all methods tested. The procedure is repeated for brain data, comparing two normal brain samples and two disease brain samples. The results are shown in Supplementary Figure S6 and Table S9. IPBT again achieves the best agreement and consistency. To get a comprehensive picture of performance, we also conduct performance comparison on each of the five testing sets individually. Detailed descriptions of the study and results are given in the Supplementary Materials (Section 5.2, Supplementary Figure S7–S11).

Table 2. Average number of correctly identified DE probes across all 91 group pairs on Spike-in Experiments data among the top *k* probes

Top *k*	5	10	15	20	25	30	35	40
Student's *t*-test	3.1	5.9	8.6	11.2	13.8	16.3	18.7	21.0
SAM	3.3	6.2	8.9	11.6	14.3	16.7	19.5	22.3
Limma	3.3	6.2	8.8	11.6	14.1	16.8	19.5	22.3
IPBT	**3.9**	**7.4**	**10.4**	**13.3**	**16.4**	**19.1**	**22.0**	**25.0**

* The best results are in **bold**.

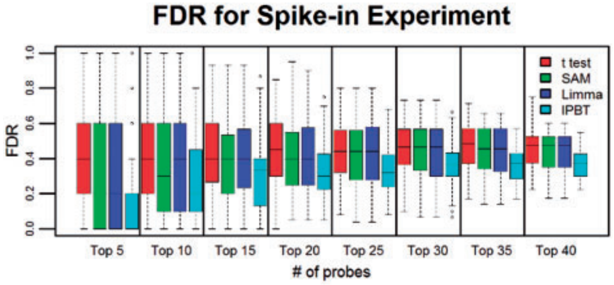

Fig. 5. All the detection methods are applied to all 91 pairs of hybridizations. Box plots of FDRs are shown for all 91 group pairs when calling top k probes significant

3.2.4 Latin Square hgu133a Spike-in experiment data

This data set consists of three replicates of 14 separate hybridizations of 42 spiked transcripts in a complex human background (HeLa cells) at concentrations ranging from 0.125 to 512 pM (Affymetrix, Santa Clara, CA). Since the spike-in genes are known, this dataset has been widely used in evaluating the performance of Microarray preprocessing algorithms (McCall *et al.*, 2010; Wu and Irizarry, 2004) and DE gene analysis methods (Lo and Gottardo, 2007). In our study, each time we select two out of the 14 separate hybridizations as the control and treatment groups (each group has three replicates) respectively. All 91 pairs are tested for DE gene detection. After excluding the probes that do not exist in Affymetrix GeneChip U133A, 34 probes are *bona fide* differentially expressed each time. We use 42 datasets from HeLa cells (cervical adenocarcinoma cell line) from the global gene expression map as the historical data.

In this study, each method generates a DE probe list (ranked by the test statistics) in every pair of the control and treatment groups and we obtain the proportion of correct DE calls (match the 34 *bona fide* DE probes). Table 2 summarizes the average number of correctly identified DE probes among the top k ($k = 5, 10, \ldots, 40$)probes across all 91 control and treatment combinations. Figure 5 shows the box plots of FDRs for the top *k* probes called significant. IPBT consistently detects more *bona fide* DE probes hence has a lower FDR in terms of the median across all the experiments. In addition, IPBT is more robust since it consistently shows the smallest interquartile ranges in the boxplot. All these results show that IPBT performs better than other methods.

4 Discussion

In this study, we present a novel strategy of reutilizing relevant information contained in historical data to improve DE gene detection. Simulation studies and real data applications show that our method IPBT significantly outperforms other existing methods in terms of both accuracy and consistency in detecting DE genes. In particular, when the DE genes have relatively low intrinsic variances, methods based on the Bayesian hierarchical models perform poorly whereas IPBT maintains its superior performance.

Bayesian hierarchical model provides an attractive statistical framework for handling 'large *p*, small *n*' inference problems.

Because it can 'borrow' information from all genes in the genome to aid the inference on a single gene so that the poor performance due to limited sample size can be improved. However, as we showed in this study, the Bayesian hierarchical model approach can suffer from the 'over-correction' problem and produce false negatives. In addition, the empirical Bayesian approach assumes a common prior for every gene, which will limit the effectiveness of the approach for genes with dramatically different behaviors. In contrast, IPBT assumes gene-specific, informative priors. With the rapid proliferation of high-throughput genomics big data, deriving these informative priors is no longer an issue.

Meta-analysis is a powerful tool for combining multiple studies of a related hypothesis and has been applied to microarray data (Conlon *et al.*, 2007; Tseng *et al.*, 2012). Our approach is different from meta-analysis because historical data used in IPBT may come from experiments with a different hypothesis, and the historical data is used indirectly in the form of informative priors in Bayesian inference.

There is much room for improvement in IPBT. First, the informative prior used in IPBT is gene-specific so DE gene analysis is done gene-by-gene. In reality we know some genes are correlated with each other such as genes located in the same pathway or sharing similar biological functions. A potential extension of IPBT is to introduce correlation among genes. Correlation information can be derived from biological knowledge or historical data. Recent studies have demonstrated the benefit of incorporating correlation information in the inference of DE genes (Lim and Wong, 2014; Soh *et al.*, 2011).

Second, the current IPBT method uses normal distribution to model log transformed expression measures. The distribution choice is made mainly for mathematical convenience. One can replace normal distribution with other non-normal ones to achieve robustness in inference in the same way as Ganjali *et al.* (2015) have done in their study of DE gene detection.

Third, we assume the expression values used by IPBT have already been background-corrected and normalized. This is possible with the powerful normalization techniques such as RMA. It is however, desirable if additional consideration is factored in the model to account for subtle experiment-to-experiment biases in the data as shown in studies such as Arima *et al.* (2011) and Lewin *et al.* (2006). This will potentially make IPBT more flexible and further improve its performance.

We are planning to pursue such extensions to IPBT in future follow-up studies.

5 Conclusions

In this study, we investigate the feasibility and effectiveness of deriving informative priors from historical microarray data and using them to help detect DE genes in studies with limited sample size. Through simulation and real data analysis, we show that our method significantly outperforms competing methods including the popular and state-of-the-art Bayesian hierarchical model-based approaches.

Taking advantage of the resource of global gene expression map developed by Lukk *et al.* (2010), we have calculated informative priors for 96 different groups of cell types using the Affymetrix U133A GeneChip as a community resource for DE gene study (all groups in the global gene expression map with at least 10 samples). We made the calculated informative priors freely available for the research community, which can be downloaded from https://github.com/benliemory/IPBT.

The strategy we propose in this paper is not limited to the microarray platform. RNA-Seq (Mortazavi *et al.*, 2008) is considered a better alternative for measuring gene expression because it can provide more information about the transcriptome (alternative splicing, gene fusion, etc.). We did not use RNA-Seq data since currently much less 'historical' data is available compared to microarray, due to the comparatively higher cost and shorter time of the adoption of RNA-Seq. As the total volume of RNA-Seq data increases, the IPBT framework can be applied to RNA-Seq as well. Cross-platform models may also be considered.

Our work illustrates the feasibility and benefits of exploiting the increasingly available genomics big data in statistical inference and presents a promising strategy for dealing with the 'large p, small n' problem.

Acknowledgement

The authors want to thank the Associate Editor and three anonymous reviewers for their insightful and constructive comments which greatly improve the manuscript.

Funding

B.L. and Z.S.Q. are partially supported by National Science Foundation grant DMS1000617 and National Institutes of Health grant P01GM085354. Z.S and Y.Z. are partially supported by National Science Foundation grant DMS1000443.

Conflict of Interest: none declared.

References

Arima,S. *et al.* (2011) Exploiting blank spots for model-based background correction in discovering genes with DNA array data. *Stat. Modell.*, **11**, 89–114.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300.

Conlon,E.M. *et al.* (2007) Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, **8**, 80.

Fan,J. and Lv,J. (2010) A selective overview of variable selection in high dimensional feature space. *Stat. Sin.*, **20**, 101–148.

Ganjali,M. *et al.* (2015) Robust modeling of differential gene expression data using normal/independent distributions: a Bayesian approach. *PLoS One*, **10**, e0123791.

Gelman,A. (2004) *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Good,I.J. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: M.I.T. Press.

Huang da,W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Huang da,W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Ji,H. and Liu,X.S. (2010) Analyzing 'omics data using hierarchical models. *Nat. Biotechnol.*, **28**, 337–340.

Ji,H. and Wong,W.H. (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics (Oxford, England)*, **21**, 3629–3636.

Kerr,M.K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Lewin,A. *et al.* (2006) Bayesian modeling of differential gene expression. *Biometrics*, **62**, 1–9.

Lim,K. *et al*. (2015) A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small. *J. Bioinform. Comput. Biol.*, **13**, 1550018.

Lim,K. and Wong,L. (2014) Finding consistent disease subnetworks using PFSNet. *Bioinformatics (Oxford, England)*, **30**, 189–196.

Lo,K. and Gottardo,R. (2007) Flexible empirical Bayes models for differential gene expression. *Bioinformatics (Oxford, England)*, **23**, 328–335.

Lukk,M., *et al*. (2010) A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.

McCall,M.N. *et al*. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Newton,M.A. *et al*. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **8**, 37–52.

Parmigiani,G. *et al*. (2003) *The Analysis of Gene Expression Data : Methods and Software*. New York: Springer.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.

Soh,D. *et al*. (2011) Finding consistent disease subnetworks across microarray datasets. *BMC Bioinformatics*, **12**, S15.

Sui,Y. *et al*. (2009) Background adjustment for DNA microarrays using a database of microarray experiments. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **16**, 1501–1515.

Tseng,G.C. *et al*. (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, **40**, 3785–3799.

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

Wu,Z. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.*, **22**, 656–658; author reply 658.