BMC Genomics

BioMed Central

Methodology article

Open Access

Operon information improves gene expression estimation for cDNA microarrays

Guanghua Xiao*[1], Betsy Martinez-Vaz[2], Wei Pan[1] and Arkady B Khodursky[2]

Address: [1]Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, Minneapolis, MN 55455-0378, USA and [2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Saint Paul, MN, 55108, USA

Email: Guanghua Xiao* - guanghx@biostat.umn.edu; Betsy Martinez-Vaz - bzayas@biosci.cbs.umn.edu; Wei Pan - weip@biostat.umn.edu; Arkady B Khodursky - khodu001@umn.edu

* Corresponding author

Published: 21 April 2006

BMC Genomics 2006, **7**:87 doi:10.1186/1471-2164-7-87

Received: 30 September 2005
Accepted: 21 April 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/87

Abstract

Background: In prokaryotic genomes, genes are organized in operons, and the genes within an operon tend to have similar levels of expression. Because of co-transcription of genes within an operon, borrowing information from other genes within the same operon can improve the estimation of relative transcript levels; the estimation of relative levels of transcript abundances is one of the most challenging tasks in experimental genomics due to the high noise level in microarray data. Therefore, techniques that can improve such estimations, and moreover are based on sound biological premises, are expected to benefit the field of microarray data analysis

Results: In this paper, we propose a hierarchical Bayesian model, which relies on borrowing information from other genes within the same operon, to improve the estimation of gene expression levels and, hence, the detection of differentially expressed genes. The simulation studies and the analysis of experiential data demonstrated that the proposed method outperformed other techniques that are routinely used to estimate transcript levels and detect differentially expressed genes, including the sample mean and SAM t statistics. The improvement became more significant as the noise level in microarray data increases.

Conclusion: By borrowing information about transcriptional activity of genes within classified operons, we improved the estimation of gene expression levels and the detection of differentially expressed genes.

Background

Genome-wide monitoring of transcription by means of DNA microarrays is used to infer transcriptional and regulatory networks in living organisms. In most of microarray experiments, transcript levels of thousands of genes are measured with a relatively small number of replications, so the estimates of true expression levels from microarray data may be poor, mostly due to a small sample size. To address this problem, several statistical methods have been proposed to borrow information from other genes to improve detection of the differentially expressed ones [1-9]. The main idea is to borrow information from other genes to estimate either the distributions of genes's expression levels or the distribution of error terms. The underlying assumption is that it should be possible to improve the estimates of expression levels of genes by borrowing information about transcriptional activity across the sets of genes that are biologically, or

Table 1: Mean squared errors for different methods

	Proposed model	Sample mean	Difference
Setting 1	0.062	0.068	8.6%
Setting 2	0.129	0.146	12.8%
Setting 3	0.222	0.255	15.1%

physically, related. In some cases, the expression levels may significantly vary across the genes, then borrowing information from unrelated genes may not improve, or even worsen, the estimates of gene expression levels [10]. However, if, based on biological knowledge, we can expect that some genes are more likely to express at similar levels (i.e. co-express), then we can improve the inference by using information about the activity of those genes.

An operon [11] is a set of linearly juxtaposed genes transcribed as a single mRNA; operons are commonly found in prokaryotic genomes such as *Escherichia coli*. Transcription of operons of *E. coli* has been examined, and operons have been predicted in many studies [12-19], which provides background information about the *E. coli* regulatory network. The genes within the same operon usually have similar expression levels, hence show some local structure in expression profiles [20], and this fact has been successfully used in operon prediction [21,22].

Based on the existing information about the structure of operons, we propose a hierarchical Bayesian model which improves gene expression estimation by borrowing information from genes within the same operon. Most existing methods for detecting differently expressed genes [1-9] borrow information from other genes in the whole genome, while our proposed method only borrows information from other genes within the same operon, which has sound biological basis. Wren *et al* [23] have proposed a simulated annealing approach to adjust gene expression data by using existing microarray measurements obtained on the same organism, which effectively reduced the noise and made it possible to compare different microarray experiments. But their method relies on reference microarray experiments that cover the dynamic range of transcript abundances for most of the genes, which may be difficult to select or unavailable. Instead of using existing microarray measurements, we use existing information about the operons' structure to reduce the noise in microarray data.

A more accurate estimation of transcript abundances of individual genes will improve our ability to evaluate transcriptional activity on a genome-wide scale, and hence facilitate the exploration of gene regulatory networks. Our proposed method provides a better way to estimate relative transcript levels, which is critical for distinguishing differentially expressed (DE) genes from equally expressed (EE) genes. Herein, we refer to the logarithm of a ratio of the fluorescent intensities of the test and control samples as the observed gene expression level. The genes with estimated expression levels significantly different from zero are identified as DE genes, otherwise as EE genes.

Using more than 200 microarray experiments, we obtained the evidence of co-transcription of genes within *E. coli* operons on a genome-wide scale. We applied the proposed method to three simulated and one experimentally obtained data sets. The simulation studies and the real data application demonstrated that the proposed method performed better than the sample mean and the SAM t statistics in estimating the gene expression levels as well as in detecting differentially expressed genes. The improvement became more significant as the noise level in microarray data increased.

Results

Simulation study

We carried out three simulations, with similar settings and the noise level gradually increasing from simulation 1 to 3. In simulations, we assumed that the genes within an operon are co-transcribed. Since the true expression levels in a simulated data set were known, we could calculate the mean squared errors of the estimated expression levels (summarized in Table 1). Without incorporating the information from operons, the sample mean of the observed expression levels would be a natural estimator of a gene's expression level. The mean squared errors of the two estimates, the posterior mean from the proposed model and the sample mean, are shown in Table 1 for comparison. It can be easily seen that the incorporation of operon information leads to a better estimate, with a smaller mean squared error, of expression levels. When the noise level increased, the improvement from incorporating operon information also increased.

To evaluate the performance of the proposed method in detecting DE genes, the estimated expression level of each gene was used to rank the genes, and the highly ranked genes were identified as DE genes. Since the identities of DE genes in the simulation studies were known, we compared the performances of the proposed method, sample mean, and SAM t statistics in detecting DE genes using receiver operating-characteristic (ROC) curves (Figure 1). In a ROC curve, the *sensitivity* is plotted against 1 – *specificity*. The sensitivity is denned as a fraction of true DE genes being correctly detected and the specificity is a fraction of the true EE genes being correctly identified. The ROC curves in Figure 1 demonstrate that the performance of the sample mean and of the SAM t statistic were very close,

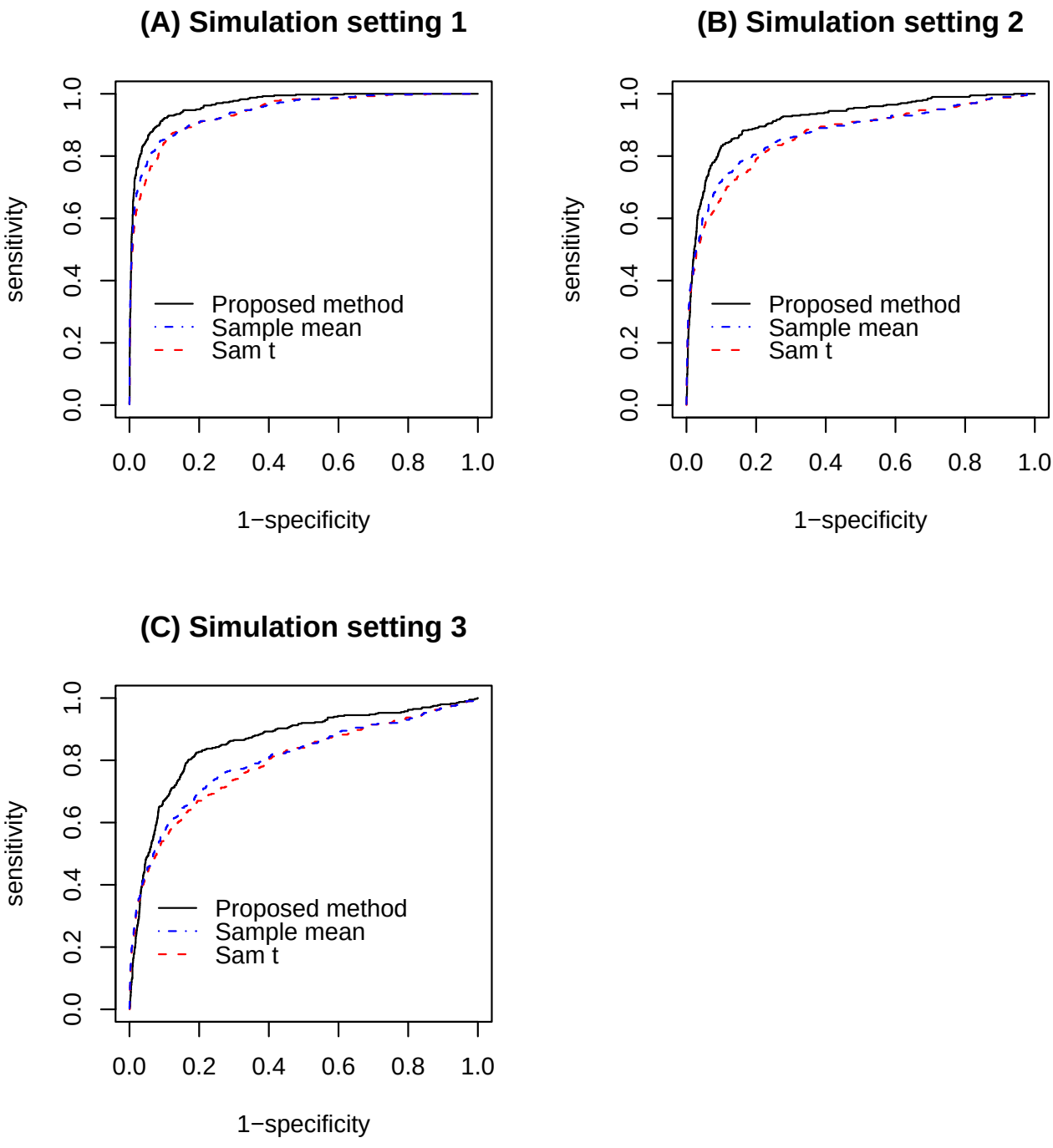

Figure 1
ROC curves of simulation settings. The Figure (A), (B), and (C) are the ROC for simulations 1,2 and 3, respectively. It shows that the sample mean and the SAM t statistic have similar performance in detecting DE genes, and our hierarchical model outperformed both of them.

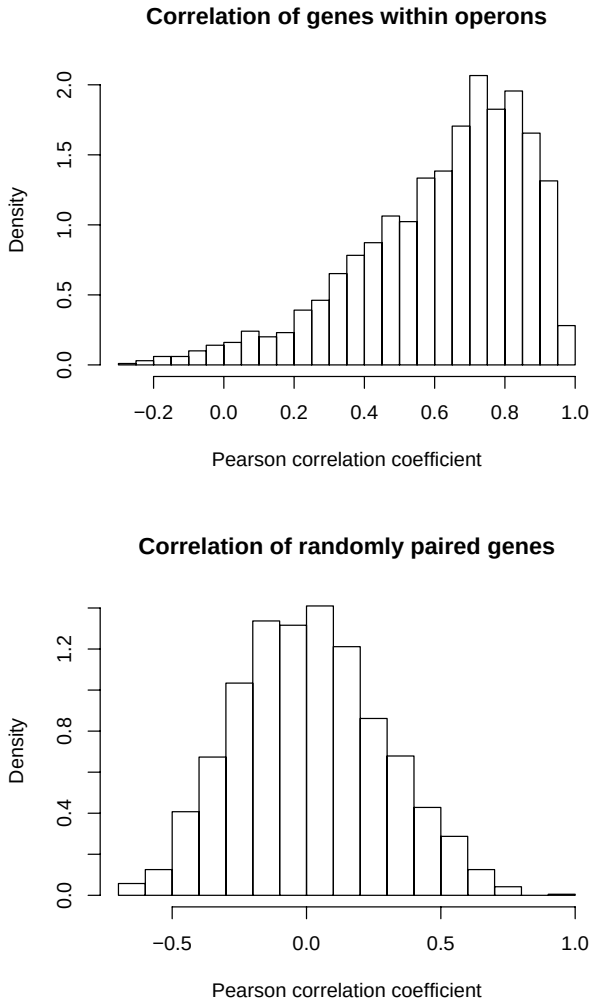

Figure 2

Histogram of the correlation coefficients. Histogram of pairwise correlation coefficients for (A)the genes within operons and (B) random gene pairs. The correlations are calculated across experimental conditions. The correlation of genes organized in operons is much higher than that of random genes, strongly indicating the co-expression of genes within operons.

and our hierarchical model, incorporating operon information, outperformed both of them. The difference in the performance became greater as the noise level increased. For example, as the specificity equals to 0.8, the sensitivities of the methods using the sample mean or SAM t were about 0.91, 0.80, and 0.68 for simulations 1,2, and 3, respectively, while the sensitivities of the proposed method were 0.95, 0.89 and 0.83, respectively.

Application to E. coli data

To verify the assumption that the genes organized in operons are co-expressed, we pooled together data from 217 microarray experiments, obtained in 53 conditions [24]. The distribution of pairwise correlations between expression profiles of genes in operons was greatly skewed towards positive values, with the mean correlation of 0.62 (Figure 2A). Unlike the profiles of genes organized in operons, expression profiles of randomly picked pairs of genes were not correlated; the corresponding distribution of correlation coefficients was almost symmetric around 0, with the mean correlation of 0.012 (Figure 2B). This result demonstrated the similarity of transcriptional activity of genes within operons and served as a motivation for borrowing information from other genes within the same operon.

The proposed method [see Additional file 1] was used to analyze differential transcriptional activity in an *E. coli* mutant lacking the *flhDC* gene, a master regulator of transcription of genes whose products mediate bacterial motility and chemotaxis [see Additional file 2]. The genes were ranked by their estimated expression levels, i.e. their posterior means of μ_i obtained from the proposed model. For the sake of comparison, the sample mean and SAM t statistics were also used to rank the genes [see Additional file 3]. Using the functional annotation from Macnab [25] as a standard, we obtained the number of false positives at different cut-off levels (total positives). The comparison revealed that ranking genes by the proposed method produced fewer false positives than the ranking based on the SAM t or sample mean statistics (Figure 3).

To find a reasonable cutoff value for differently expressed genes, we calculated the false discovery rate (FDR) [26,27]based on the posterior probability [7]. In this experiment, we also estimated the FDR by using the functional annotation from Macnab [25] as a reference. Comparison of the estimated False Discovery Rates revealed that the estimated FDR from the posterior probability was a little lower than that derived from the annotation, which could be due to the partial incompleteness of the reference (Fig. 4). Overall, the estimated FDR from the posterior probability is close to the FDR using the reference list of genes, indicating that our method for estimating the FDR is adequate. We set the cutoff for the FDR to be 0.01, which identified the top 44 genes as DE genes. At such a cutoff, the estimated number of false negatives is 14 and the estimated false negative rate is about 0.003 (see the "Methods" section for details). The top 44 genes are listed in Table 2. Note that, in Table 2, the gene expression level is on the log scale and the FDR corresponds to a specific number of DE genes and not to each individual gene itself. According to Macnab's classification [25], 41 genes out of the 44 were expected to be differentially expressed

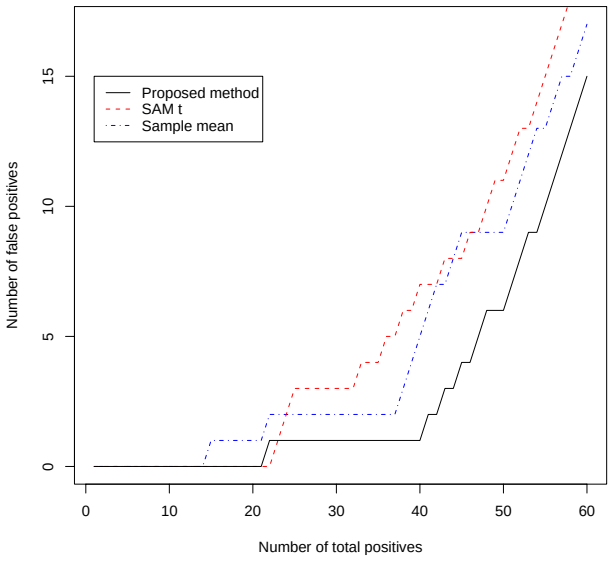

Figure 3
Number of false positives vs. Number of total positives. For the *E. coli* motility data, the genes are ranked by using the proposed method, SAM t statistic and sample mean. The number of false positives is plotted against the number of total positives for each ranking criterion. It shows that ranking genes by proposed method has less false positives than ranking genes by SAM t or sample mean.

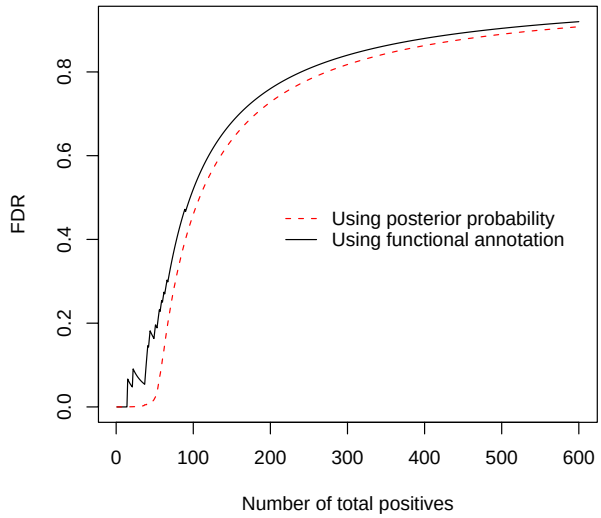

Figure 4
FDR estimation. Estimate FDR by using posterior probability and the functional annotation from Macnab *et al*. The solid line is for the FDR estimate using existing functional annotation while the dashed line for using posterior probability. It indicates that estimate the FDR by using posterior probability yields reasonable result.

in the flhDC- dependent manner, whereas the lists of 44 genes identified by using SAM t and sample mean contained only 36 and 38 expected genes, respectively.

We examined some genes and operons in more detail, to demonstrate the advantages of borrowing information from within an operon. For example, an operon *argT-hisJQMP* contains 5 genes (argT, hisJ, hisM, hisP, hisQ) and is not expected to be differentially expressed under the examined experimental condition, according to our biological knowledge [25]. But from the microarray data, the mean expression level (an average log ratio) of the gene *argT was* -2.73, which ranked 15th among all the expression levels in the *E. coli* genome. This "high expression" of the *argT* could possibly be caused by random noise in microarray measurements and/or in biological samples, since the mean expression levels of other genes within the same operon were close to 0 (those for genes hisP, hisM, hisQ, hisJ were 0.00, -0.05, 0.03, and 0.05, respectively). However, accounting for expression levels of other genes within the same operon lowered the estimate of the expression level (posterior mean of μ_i of the *argT to* -0.02, rank of 1456, indicating that this gene was not differentially expressed. Analysis of another operon, *fliDST*, illustrates a complimentary case. Transcription of the *fliDST operon* (containing 3 genes, *fliD*, *fliS*, *fliT*) is known to be controlled by the FlhDC [25], and thus, under the experimental condition, differential expression of genes in that operon would be expected. While the expression level of the *fliT*, estimated by the sample mean at -0.90, ranked only 65th, the estimated expression level of the gene after borrowing information from two other genes in the operon was -3.25, which ranked 19th.

In general, through borrowing information, our Bayesian method worked in a way giving more consistent estimates of the expression levels for the genes of the same operon. For example, compared with using the sample mean to estimate expression levels, the Bayesian method tended to yield smaller standard deviations of the expression estimates for within-operon genes (see Figure 5).

Discussion

In this paper, we proposed and applied a hierarchical Bayesian model, to estimate relative gene expression levels and detect differentially expressed genes by borrowing expression information within operons. The performance of the proposed method was compared with that of the sample mean and SAM t statistics. Through the simulation

Table 2: Top 44 genes and their estimated relative transcription levels

Name	B number	Operon	Verified	Estimated expression	FDR
fliC	B1923		+	-4.71	0.000
flgB	B1073	flgBCDEFGHIJK	+	-3.90	0.000
flgE	B1076	flgBCDEFGHIJK	+	-3.82	0.000
flgL	B1083		+	-3.81	0.000
flgF	B1077	flgBCDEFGHIJK	+	-3.81	0.000
flgD	B1075	flgBCDEFGHIJK	+	-3.81	0.000
flgK	B1082	flgBCDEFGHIJK	+	-3.76	0.000
flgI	B1080	flgBCDEFGHIJK	+	-3.76	0.000
flgJ	B1081	flgBCDEFGHIJK	+	-3.75	0.000
flgH	B1079	flgBCDEFGHIJK	+	-3.74	0.000
flgC	B1074	flgBCDEFGHIJK	+	-3.73	0.000
flgG	B1078	flgBCDEFGHIJK	+	-3.71	0.000
fliA	B1922	fliAZY	+	-3.44	0.000
fliY	B1920	fliAZY	+	-3.35	0.000
fliZ	B1921	fliAZY	+	-3.32	0.000
fliD	B1924	fliDST	+	-3.30	0.000
fliS	B1925	fliDST	+	-3.25	0.000
fliT	B1926	fliDST	+	-3.25	0.000
tap	B1885		+	-3.22	0.000
tar	B1886		+	-2.97	0.000
tsr	B4355		+	-2.64	0.000
fadL	B2344		-	-2.37	0.001
fliH	B1940	fliFGHIJK	+	-2.03	0.001
fliG	B1939	fliFGHIJK	+	-2.01	0.001
fliF	B1938	fliFGHIJK	+	-1.99	0.001
fliK	B1943	fliFGHIJK	+	-1.99	0.001
fliJ	B1942	fliFGHIJK	+	-1.92	0.000
flgA	B1072	flgAMN	+	-1.91	0.001
fliI	B1941	fliFGHIJK	+	-1.89	0.001
flgM	B1071	flgAMN	+	-1.87	0.001
flgN	B1070	flgAMN	+	-1.81	0.001
flxA	B1566		+	-1.70	0.002
aer	B3072		+	-1.66	0.002
cheZ	B1881		+	-1.63	0.002
cheR	B1884		+	-1.60	0.003
fliM	B1945	fliMNOPQR	+	-1.55	0.003
cheY	B1882		+	-1.51	0.005
fliL	B1944	fliMNOPQR	+	-1.50	0.005
fliN	B1946	fliMNOPQR	+	-1.49	0.005
ycgR	B1194		-	-1.44	0.007
cheW	B1887	motAB-cheAW	+	-1.42	0.007
cheA	B1888	motAB-cheAW	+	-1.39	0.007
b1904	B1904		-	-1.38	0.009
mot A	B1890	motAB-cheAW	+	-1.32	0.009

Note: the estimated gene expression levels is log scale.

studies, we showed that the proposed method outperformed the sample mean and the SAM t statistics in estimating gene expression levels and detecting DE genes. The proposed method was used to analyze differential expression in an *E. coli* mutant with a defect in transcription of motility/chemotaxis genes, giving results more consistent with the existing biological knowledge than those obtained by using the other statistics.

A major advantage of the proposed approach is in borrowing expression information from other genes within the same operon. The approach is developed within a statistically sound Bayesian model and it offers necessary flexibility with respect to the amount of information that needs to be borrowed from other genes. By borrowing information we can obtain stabilized estimates of expression levels from rather noisy microarray data. As a result, the estimates of transcript levels within the same operon become more similar to each other, more so than without

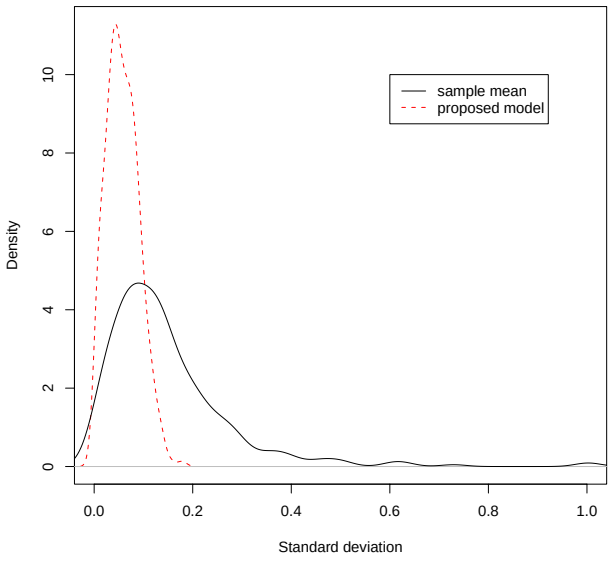

Figure 5
Within-operon standard deviation of the estimated gene expressions. Distribution of the standard deviation of the expression estimates of the genes from the same operon. The solid line is for the sample mean method while the dashed line for our proposed Bayesian model. Comparing to the sample mean method, our proposed method yields smaller within-operon standard deviations of the gene expressions.

borrowing information; this is consistent with a biological fact that genes within the same operon are transcribed as a single mRNA molecule. With the proposed method, the estimated expression levels of genes in "differentially expressed" operons are consistently high, and more importantly, transcript abundances of genes in "equally expressed" operons are stabilized towards zero. In the experimentally obtained microarray data, the within operon variation was smaller than the variation among the replicate data points for the same genes, indicating that the expression levels of the genes within an operon were very similar.

In our model, the ratio of parameters τ^2 and σ_i^2 determines how much information comes from the observed expression of a gene and how much comes from the average expression of an operon, when estimating the expression level of an individual gene within an operon. A smaller τ, as compared to σ_i^2, puts more weight on the average expression of an operon. Here we assumed the same τ value for all operons, implying that all operons had similar within-the-operon variability. However, this assumption might not be realistic. In some operons and physiological conditions, the genes might express very similarly, but in others, especially under the control of internal promoters, transcription of individual genes may be more heterogeneous. In the future, we will investigate the effect of an operon-specific τ. Operonal organization of genes is common in prokaryotes and also present in some eukaryotic organisms [28-30], and the proposed method can be extended to biological systems where the operonal structure is unknown. Many biological studies have demonstrated that co-expressed genes tend to cluster on the chromosome. Although the nature of this phenomenon is not quite understood, a positional clustering of co-expressed genes can be found in many eukaryotes including yeast [31,32], worm [33], fly [34,35], mouse [36], and human [37-39]. These findings indicate that the genes are likely to co-transcribe with their chromosomal neighbors. In those cases, instead of borrowing information from genes in the same operon, we can borrow information from gene neighbors on the chromosome. Another extension of our method would involve incorporation of gene annotation information into the analysis of expression data. The approach would be very similar to the one described in this paper: based on biological knowledge, the genes belonging to the same functional group are more likely to be co-expressed, so we can use a hierarchical model to borrow information from and for the genes within the same functional group to improve the estimates of gene expression levels.

Conclusion

The information about operon structure leads to a better estimation of gene expression levels. Using simulated and experimental data sets, we have demonstrated that the proposed method performs better than the sample mean and the SAM t statistics in estimating the relative levels of transcript abundances and detecting differentially expressed genes.

Methods

RegulonDB database and E. coli. microarray data

RegulonDB [40] is a database containing information about known operons in *E. coli*. According to the RegulonDB annotations, 1486 genes (about one third of all genes predicted in the *E. coli* genome) are organized in 600 operons.

The *E. coli* data set contains results of 217 microarrays collected in 53 different experimental conditions. The fluorescent intensities of the test and control samples were measured, and the average log ratio of the intensities for

each gene under the same condition was used here to represent an observed gene expression level under that condition [24].

An *E. coli* motility expression data set ([41] series accession number: GPL2101) was obtained in a direct pairwise comparison between a knock-out mutant of the *flhDC*, a master regulator of the motility/chemotaxis regulon [25], and its isogenic wild type strain. Total RNA samples of a mutant *E. coli* (test samples) and an isogenic wild type *E. coli* (control samples) were labeled with red (Cy5) and green (Cy3) fluorophors. The intensities from the red and the green channels were normalized by the lowess method [42]. There were 4281 genes (G = 4281) with four replicates for each gene (n = 4). Let $Y_{i,j}$ be defined as the log ratio of the intensities between the test and control samples for gene *i* on array *j;* that is,

$$Y_{i,j} = \log_2 \frac{R_{i,j}}{G_{i,j}} = \log_2 \frac{(\text{Intensity of test sample})_{i,j}}{(\text{Intensity of control sample})_{i,j}}$$

Hierarchical models

We propose a hierarchical Bayesian model,

$$Y_{ij} \mid \mu_i, \sigma_i^2 \sim N\left(\mu_i, \sigma_i^2\right) \qquad (1)$$

where, $Y_{i,j}$ is the log ratio of gene *i* in replicate *j*, μ_i and σ_i are the true expression level and the standard deviation, respectively. In our method, the posterior mean of μ_i is used as the estimated expression level of gene *i*, while the sample mean, $\overline{y}_i$., is referred to as the observed expression level.

As prior knowledge, we assume that if several genes belong to the same operon, in accordance with the RegulonDB annotation, then their expression levels are from a normal distribution, with the mean λ_p and the variance τ^2. Specifically,

$$\begin{cases} \mu_i \mid \lambda_p, \tau^2 \sim N\left(\lambda_p, \tau^2\right) & \text{if } i \in O_p \text{ for some } p; \\ Pr(\mu_i) \propto 1 & \text{if } i \notin O_p \text{ for any } p. \end{cases}$$

where O_p denotes operon *p*. λ_p represents the expression level of the operon *p*, which is the average of the mean expression levels of all genes within the operon *p*. τ^2 is the within operon variation, and is assumed to be the same across all operons. A non-informative prior is assigned to λ_p, that is $Pr(\lambda_p) \propto 1$, to reflect the lack of prior information, σ_i^2 and τ^2 have vague priors, which are inverse Gamma distributions with the shape and rate parameters equal to 0.01 and 0.01 respectively [43]. If gene *i* is not in any operon, then

$$\mu_i \mid y_{ij}, \lambda_p, \sigma_i^2, \tau \sim N\left(\overline{y}_i, \sigma_i^2 / n\right) \qquad (2)$$

so the posterior mean of μ_i is just the the sample mean $\overline{y}_i$; if gene *i* is in operon *p*, then the conditional distribution of μ_i can be derived as:

$$\mu_i \mid y_{i,j}, \lambda_p, \sigma_i^2, \tau \sim N(\eta_i^2(\frac{n}{\sigma_i^2}\overline{y}_{i\cdot} + \frac{1}{\tau^2}\lambda_p), \eta_i^2) \qquad (3)$$

where

$$\eta_i^2 = (\frac{n}{\sigma_i^2} + \frac{1}{\tau^2})^{-1}$$

Equation (3) shows that, when borrowing information from the other genes within the same operon, the estimated expression level of the gene *i* becomes the weighted average of the observed expression level of gene *i* and the expression level of operon *p*, given that gene *i* belongs to operon *p*. The weights are inversely proportional to the variances. In this model, a key concept is to shrink the observed expression level $\overline{y}_i$, towards λ_p, the expression level of an operon, based on the knowledge of the operon structure. The degree of shrinkage is determined by the variability of $\overline{y}_i$. and λ_p. Without incorporating operon information, the estimated expression level would be close to the observed expression level, $\overline{y}_i$ In the hierarchical model, λ_p represents the expression level of operon *p*, and

$$\lambda_p \mid \mu_i, \sigma_i^2, \tau \sim N(\frac{1}{m_p}\sum_{i \in O_p} \mu_i, \frac{\tau^2}{m_p}) \qquad (4)$$

where, m_p is the number of genes in operon *p*, and $i \in O_p$ denotes that gene *t* is in operon *p*. Although the posterior distribution was not available in a closed form, we could derive a closed form of the full conditional distribution, and used Markov chain Monte Carlo (MCMC) to simulate the parameters from the posterior distribution. With this closed form expression, the model could be easily coded in R [see Additional file 1] for MCMC simulation using Gibbs sampling [44]. The expression level of gene *i* is estimated by the posterior mean of μ_i, and the genes are ranked by the absolute values of the posterior means of the μ_i's. Genes with high rankings were designated as differentially expressed (DE) genes.

SAM t statistic

To evaluate the performance of the proposed method in estimating gene expression levels and identifying DE genes, we compared the proposed method to the sample mean and SAM t statistics [1,3]. Because of its good performance, the SAM t statistic [1,3] is widely used to rank genes and detect DE genes [45]. We denote the SAM t statistic for the gene i as Z_i, then

$$Z_i = \frac{\bar{Y}_i}{S_i + S_0} \qquad (5)$$

where $\bar{Y}_i$ and S_i are the sample mean and sample standard deviation for the gene i, and So is the 90 th percentile of S_i's.

Simulation settings

We conducted three simulation studies to assess the usefulness of our method. The operon structure of the *E. coli* genome from the RegulonDB database [40] was used in simulation studies. We randomly chose 100 operons (involving about 340 genes) and assumed that genes from those operons were differentially expressed DE genes. Then we randomly picked a subset of non-operon genes to be DE genes and adjusted the total number of DE genes to 400. Let μ_i be the expression level of gene i, for i = 1, 2,..., 4821. For DE genes, μ_i's were simulated from an equal mixture of $N(1, 0.25^2)$ and $N(-1, 0.25^2)$ distributions, and genes within the same operon were from the same component of the mixture distribution. For EE genes, $\mu_i \sim N(0, 0.25^2)$. We simulated 4 replicates from a normal distribution for each gene, $Y_{ij} \sim N(\mu_i, \sigma_i^2)$, where $Y_{i,j}$ was the log ratio of transcript abundances for the gene i on the array j. To provide increasing noise levels for simulations 1, 2 and 3, the σ_i's were simulated from the *uniform*(0.25, 0.75), *uniform*(0.5,1.0), and *uniform*(0.75,1.25), respectively.

Estimation of false positives and false negatives

Using the posterior distributions, we can evaluate the FDR for specific number of DE genes. Using $Pr(|\mu_i| > \delta | Y_{i,j})$ to estimate the probability of gene i to be a DE gene, we can estimate the number of false positives for a cut off value k [7]:

$$\text{Expected Number of False Positives (k)} = \sum_{i=1}^{k}\left(1 - Pr\left(|\mu_i| > \delta \mid Y_{i,j}\right)\right) = \sum_{i=1}^{k} Pr\left(|\mu_i| < \delta \mid Y_{i,j}\right)$$

Here, the genes are ranked based on the estimated mean expression level μ_i. In this study, we set δ = 1,

which corresponding to the commonly used 2-fold cutoff.

The false discovery rate (FDR) for the cut off k can be derived as:

$$FDR(k) = Expected\left(\frac{\text{Number of False Positives}}{\text{Number of Total Positives}}\right) = \frac{1}{k}\sum_{i=1}^{k} Pr\left(|\mu_i| < \delta \mid Y_{i,j}\right)$$

Similarly, the number of false negatives for the cut off k can be calculated as:

$$\text{Expected Number of False Negatives (k)} = \sum_{i=k+1}^{G} Pr(|\mu_i| < \delta \mid Y_{i,j}))$$

Algorithm for Gibbs sampler

The algorithm is implemented below:

Set initial values:

$$\mu_i^{(0)} = \bar{\gamma}_i$$
$$\sigma_i^{2^{(0)}} = \frac{1}{2}V_i + \frac{1}{2}V_0$$
$$\tau^{2^{(0)}} = V_0$$

FOR t FROM 1 TO T, draws random samples:

$$\lambda_p^{(t)} \sim N\left(\frac{1}{n_p}\sum_{i\in p}\mu_i^{(t-1)}, \frac{1}{n_p}\tau^{2^{(t-1)}}\right)$$
$$\tau^{2^{(t)}} \sim IG\left(\frac{1}{2}\sum_p n_p + 0.01, \frac{1}{2}\sum_p\sum_{i\in p}(\lambda_p^{(t)} - \mu_i^{(t-1)})^2 + 0.01\right)$$
$$\sigma_i^{2^{(t)}} \sim IG\left(\frac{n}{2} + 0.01, \frac{1}{2}\sum_{j=1}^{n}(\mu_i^{(t-1)} - \gamma_{ij})^2 + 0.01\right)$$
$$\mu_i^{(t)} \sim N\left(\frac{n}{\sigma_i^{2^{(t)}}}\bar{\gamma}_i + \frac{1}{\tau^{2^{(t)}}}\lambda_p^{(t)}, \left(\frac{n}{\sigma_i^{2^{(t)}}} + \frac{1}{\tau^{2^{(t)}}}\right)^{-1}\right) \text{ if } i \in p \text{ for some } p$$
$$\mu_i^{(t)} \sim N\left(\bar{\gamma}_i, \frac{1}{n}\sigma_i^{2^{(t)}}\right) \text{ otherwise}$$

END FOR

where V_i is the sample variance of gene i, and V_0 is the median of V_i's. n_p is the number of genes in operon p. T is the total number of iteration. To diminish the effect of the initial values, we discard the results from the early iterations ($t \le T_B$, where T_B is the burn in time). The posterior mean of μ_i, of gene i is calculated by:

$$\hat{\mu}_i = \frac{1}{T - T_B} \sum_{t=T_B+1}^{T} u_i^{(t)}$$

In our proposed method, the expression level of gene i is estimated by $\hat{\mu}_i$. In the real data example, T_B and T are 500 and 2000, respectively.

Authors' contributions

GX initiated the study, implemented the methods and conducted data analysis. WP participated in development of the methods and co-wrote the paper. BMV generated the *E. coli* motility data. ABK generated the *E. coli* data set containing 217 microarrays, supervised the project and co-wrote the paper. All authors contributed to the writing, read and approved the final manuscript.

Additional material

Additional File 1

Operon.r – The R code used in the study.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-87-S1.R]

Additional File 2

Motility.txt – E. coli motility data set.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-87-S2.txt]

Additional File 3

Result.txt – The result for E. coli motility data, including posterior mean of proposed method, sample mean and SAM t statistics
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-87-S3.txt]

Acknowledgements

The authors are grateful to the reviewers for helpful comments. GX is supported by a Merck fellowship, and BMMV was supported in part by a Ford postdoctoral fellowship. This work was supported in part by NIH grant GM066098 (ABK) and HL65462 (WP), and a University of Minnesota AHC faculty research development grant (WP and ABK).

References

1. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *PNAS* 2001, **98(9**5116-5121 [http://www.pnas.Org/cgi/content/abstract/98/9/5116].
2. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17(6**509-519 [http://bioinformatics.oxfordjournals.Org/cgi/content/abstract/17/6/509].
3. Efron B, Tishirani R, Storey J, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J Amer Statist Assoc* 2001, **96:**1151-1160.
4. Pan W: **A comparative review of statistical methods for discovering dierentially expressed genes in replicated microarray experiments.** *Bioinformatics* 2002, **18(4**546-554 [http://bioinformatics.oxfordjournals.Org/cgi/content/abstract/18/4/546].
5. Broet P, Richardson S, Radvanyi F: **Bayesian Hierarchical Model for Identifying Changes in Gene Expression from Microarray Experiments.** *Journal of Computational Biology* 2002, **9(4**671-683 [http://www.liebertonline.com/doi/abs/10.1089/10665270276027738 1].
6. Kendziorski CM, Newton MA, Lan H, Gould MN: **On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.** *Statistics in Medicine* 2003, **22(24):**3899-3914.
7. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostat* 2004, **5(2**155-176 [http://biostatistics.oxfordjournals.Org/cgi/content/abstract/5/2/155].
8. Lonnstedt I, Speed T: **Replicated microarray data.** *Statist Sinica* 2002, **12:**31-46.
9. Lewin A, Richardson S, Marshall C, A G, Aitman T: **Bayesian Modelling of Differential Gene Expression.** *Biometrics* 2005 in press. http://www.bgx.org.uk/papers.html
10. Liu D, Parmigiani G, Caffo B: **Screening for Differentially Expressed Genes: Are Multilevel Models Helpful?** *Johns Hopkins University, Dept. of Biostatistics Working Papers* 2004 [http://www.bepress.com/jhubiostat/paper34].
11. Miller JH, Reznikoff WS: *The operon* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1978.
12. Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C: **DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli.** *PNAS* 2000, **97(22**12170-12175 [http://www.pnas.org/cgi/content/abstract/97/22/12170].
13. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC: **Comparative Gene Expression Profiles Following UV Exposure in Wild-Type and SOS-Deficient Escherichia coli.** *Genetics* 2001, **158:**41-64 [http://www.genetics.Org/cgi/content/full/158/l/41].
14. Moreno-Hagelsieb G, Trevino V, Perez-Rueda E, Smith TF, Collado-Vides J: **Transcription unit conservation in the three domains of life: a perspective from Escherichia coli.** *Trends Genet* 2001, **17(4):**175-7.
15. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in Escherichia coli: Genomic analyses and predictions.** *PNAS* 2000, **97(12**6652-6657 [http://www.pnas.org/cgi/content/abstract/97/12/6652].
16. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002:S329-36.
17. Ermolaeva MD, White O, Salzberg SL: **Prediction of operons in microbial genomes.** *Nucleic Acids Res* 2001, **29(5):**1216-21.
18. Jacob E, Sasikumar R, Nair KNR: **A fuzzy guided genetic algorithm for operon prediction.** *Bioinformatics* 2005, **21(8):**1403-7.
19. Westover BP, Buhler JD, Sonnenburg JL, Gordon JI: **Operon prediction without a training set.** *Bioinformatics* 2005, **21(7):**880-8.
20. Jeong KS, Ahn J, Khodursky AB: **Spatial patterns of transcriptional activity in the chromosome of Escherichia coli.** *Genome Biology* 2004, **5:**R86.
21. Sabatti C, Rohlin L, Oh MK, Liao JC: **Co-expression pattern from DNA microarray experiments as a tool for operon prediction.** *Nucleic Acids Res* 2002, **30(13):**2886-93.
22. Bockhorst J, Craven M, Page D, Shavlik J, Glasner J: **A Bayesian network approach to operon prediction.** *Bioinformatics* 2003, **19(10):**1227-35.
23. Wren JD, Yao M, Langer M, Conway T: **Simulated annealing of microarray data reduces noise and enables cross-experimental comparisons.** *DNA Cell Biol* 2004, **23(10):**695-700.
24. Sangurdekar DP, Srienc F, Khodursky AB: **A classification based framework for quantitative description of large-scale microarray data.** *Genome Biology* 2006, **7(4):**R32.
25. Macnab RM: **Genetics and biogenesis of bacterial flagella.** *Annu Rev Genet* 1992:131-58.
26. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J R Stat Soc B* 1995, **57:**289-300.

27. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *PNAS* 2003, **100(16**9440-9445 [http://www.pnas.org/cgi/content/abstract/100/16/9440].
28. Lercher MJ, Blumenthal T, Hurst LD: **Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes.** *Genome Res* 2003, **13(2):**238-43.
29. Blumenthal T, Gleason KS: **Caenorhabditis elegans operons: form and function.** *Nat Rev Genet* 2003, **4(2):**112-20.
30. Blumenthal T: **Operons in eukaryotes.** *Brief Funct Genomic Proteomic* 2004, **3(3):**199-211.
31. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26(2):**183-6.
32. Kruglyak S, Tang H: **Regulation of adjacent yeast genes.** *Trends Genet* 2000, **16(3):**109-11.
33. Roy PJ, Stuart JM, Lund J, Kim SK: **Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans.** *Nature* 2002, **418(6901):**975-9.
34. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI: **Large clusters of co-expressed genes in the Drosophila genome.** *Nature* 2002, **420(6916):**666-9.
35. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome.** *J Biol* 2002, **1:**5.
36. Li Q, Lee BTK, Zhang L: **Genome-scale analysis of positional clustering of mouse testis-specific genes.** *BMC Genomics* 2005, **6:**7.
37. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291(5507):**1289-92.
38. Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AHC: **The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.** *Genome Res* 2003, **13(9):**1998-2004.
39. Yager TD, Dempsey AA, Tang H, Stamatiou D, Chao S, Marshall KW, Liew CC: **First comprehensive mapping of cartilage transcripts to the human genome.** *Genomics* 2004, **84(3):**524-35.
40. Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12.** *Nucl Acids Res* 2004, **32(90001**D303-306 [http://nar.oxfordjournals.org/cgi/content/full/32/suppl/D303].
41. **NCBI Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
42. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucl Acids Res* 2002, **30(4**e15 [http://nar.oxfordjournals.Org/cgi/content/full/30/4/el5].
43. Carlin B, Louis T: *Bayes and Empirical Bayes Methods for Data Analysis* Boca Raton, FL: Chapman and Hall/CRC Press 2000; 2000.
44. Gelfand A, Smith A: **Sampling Based Approaches to Calculating Marginal Densities.** *Journal Amer Stat Assoc* 1990, **85:**398-409.
45. Xie Y, Jeong KS, Pan W, Khodursky A, Carlin BP: **A case study on choosing normalization methods and test statistics for two-channel microarray data.** *Comp Fund Genom* 2004, **5:**432-444.