BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 2 2011, pages 225–231
doi:10.1093/bioinformatics/btq650

Gene expression

Advance Access publication November 18, 2010

SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures

Hong-Qiang Wang[1], Lindsey K. Tuominen[1] and Chung-Jui Tsai[1,2,*]

[1]Warnell School of Forestry and Natural Resources and [2]Department of Genetics, University of Georgia, Athens, GA 30602, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The pre-estimate of the proportion of null hypotheses (π_0) plays a critical role in controlling false discovery rate (FDR) in multiple hypothesis testing. However, hidden complex dependence structures of many genomics datasets distort the distribution of p-values, rendering existing π_0 estimators less effective.

Results: From the basic non-linear model of the q-value method, we developed a simple linear algorithm to probe local dependence blocks. We uncovered a non-static relationship between tests' p-values and their corresponding q-values that is influenced by data structure and π_0. Using an optimization framework, these findings were exploited to devise a Sliding Linear Model (SLIM) to more reliably estimate π_0 under dependence. When tested on a number of simulation datasets with varying data dependence structures and on microarray data, SLIM was found to be robust in estimating π_0 against dependence. The accuracy of its π_0 estimation suggests that SLIM can be used as a stand-alone tool for prediction of significant tests.

Availability: The R code of the proposed method is available at http://aspendb.uga.edu/downloads for academic use.

Contact: cjtsai@warnell.edu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on September 28, 2010; accepted on November 5, 2010

1 INTRODUCTION

Experiments in the 'omics' fields often involve hundreds to thousands of dependent variables, such as those encountered in genetic linkage (Lander and Kruglyak, 1995), gene expression (Stafford and Yidong, 2007) and metabolic profiling (Clarke and Haselden, 2008). Multiple testing correction is necessary in order to identify statistically significant variables for subsequent analyses, without a flood of false positives called by chance. The false discovery rate (FDR) approach (Benjamini and Hochberg, 1995) and its many variants, including the positive FDR (pFDR)-based q-value statistic (Storey, 2002), have been widely used for false discovery control in multiple hypothesis testing. The q-value represents the minimum pFDR that can occur for any possible θ greater than or equal to a p-value point. Given a set of p-values ranked in an increasing order, $p_i, i=1,2,\ldots,m$ (m is the total number of tests), the q-value is calculated as:

$$q(p_i)=\frac{\pi_0 m}{i}p_i \tag{1}$$

This formula indicates that π_0 is the only unknown parameter to be pre-estimated. The accuracy of the π_0 estimate directly affects the q-value calculation and the optimal control of FDR. A reliable π_0 estimate also provides a simple prediction of the number of genes that are differentially expressed under the experimental conditions in gene expression analysis.

A widely used method is the λ-estimator (Storey, 2002), i.e.

$$\hat{\pi}_0(\lambda)=\frac{\#\{p_i>\lambda\}}{(1-\lambda)m} \tag{2}$$

where $\lambda\in(0,1)$ is a pre-chosen cutoff and $\#\{p_i>\lambda\}$ is the number of p-values greater than λ. Considering the non-linear relationship between $\pi_0(\lambda)$ and λ, we refer to this estimator as the non-linear model. The underlying assumption is that the largest p-values are most likely to come from a uniform distribution of null features in the range (0,1). In practice, there is a bias versus variance tradeoff for choosing an optimal λ for the estimation of π_0 (Storey, 2002). To balance this trade-off, Storey and Tibshirani (2003) used a natural cubic spline smoothing (CSS) method to fit the non-linear relationship across a range of λ, as implemented in the QVALUE software. Jiang and Doerge (2008) proposed an average estimate (AE) method, which takes advantage of multiple non-linear estimators to reduce the π_0 estimation variance. Markitsis and Lai (2010) recently proposed a censored beta mixture model (CBMM), based upon the beta-uniform mixture (BUM) method of Pounds and Morris (2003), to approximate the p-value distribution.

Both the original and the q-value-based FDR procedures were developed for independent test statistics. Although these methods can be applied to weakly dependent data (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003), they are unreliable for datasets with inherent multiplicity and complex dependence (Clarke and Hall, 2009). The BUM-based methods also cannot effectively handle irregular p-value distribution (Markitsis and Lai, 2010). To specifically handle multiple testing under dependence, Efron (2007a) developed an empirical Bayesian framework, called Locfdr, to remedy the effects of data correlation. However, Locfdr is only applicable to data with a large (≥0.9) π_0 (Efron, 2007a).

We have developed a linear π_0 estimator from the non-linear method of Storey (2002) to explore local properties of p-value distributions as a means to better capture data dependence. When applied to data with a uniform null p-value distribution, the slope and intercept of the linear model reflect the proportions of null

[*]To whom correspondence should be addressed.

(π_0) and alternative hypotheses (π_1), respectively. Based on this framework, we devised a sliding linear model (SLIM) for estimating π_0 that can be applied over a broader range of p-value distributions. An interrelationship between p-values and q-values was observed and exploited to optimize SLIM. We compared the performance of SLIM with that of four other methods on three types of simulated datasets mimicking various p-value distribution scenarios and data dependence structures, as well as on one *Populus* microarray experiment. The results show that SLIM performs better than the previous methods under the conditions examined.

2 GENERATION OF SIMULATION DATASETS

Considering that dependence often leads to a distorted p-value distribution, we generated two simulated datasets with uniform and non-uniform null p-values, referred to as uniform and non-uniform datasets, respectively, to investigate the effect of various p-value distributions on π_0 estimation. The third simulated dataset mimics microarray gene expression data by explicitly adding dependence structures.

2.1 Uniform and non-uniform datasets

The uniform and non-uniform datasets were produced using a procedure modified from Storey (2002). For the former, we set $m = 10\,000$, and generated $m\pi_0$ null hypotheses from a normal distribution with mean $\mu_0 = 0$ and SD $\sigma = 1$, as well as $m(1-\pi_0)$ alternative hypotheses from a normal distribution with $\mu_1 = 5$ and $\sigma = 1$. We varied π_0 among 0.1–0.9 to track estimation performance. The non-uniform datasets were similarly generated except that the null hypotheses comprised a mixture of two normal distributions: one with $\mu_0 = -1$ and $\sigma = 1$ and the other, $\mu_0 = 1$ and $\sigma = 1$. The mixture coefficient was set to $\upsilon = 0.5$. The p-values for each dataset were calculated assuming a normal null distribution $N(0,1)$.

2.2 Gene expression simulation datasets

We followed the procedure of Qin *et al.* (2008) to produce the microarray gene expression simulation data, consisting of two experimental conditions with sample sizes of n_1 and n_2, and $G = 10\,000$ gene probes. To add hidden dependence structures, a correlation background $X[G \times n(n = n_1 + n_2)]$ was generated by (i) randomly selecting clump size m from $\{1,2,\dots,100\}$ and clump-wise correlation ρ from $U(0.5,1)$. For a given (m,ρ) pair, we (ii) generated noise vectors e_j of dimension $m \times 1$ from $N(0_m,(1-\rho)I_m + \rho 1_m 1_m')$, and (iii) set $x_{.j} = \mu + diag(\omega)e_{.j}$ as the background expression values of the m genes in the clump at sample $j = 1,2,\dots,n = n_1 + n_2$, where μ is an $m \times 1$ vector of elements $\mu_g \sim 1000\chi_5^2$ and ω is an $m \times 1$ vector of elements $\omega_g = e^{\beta_0/2}\mu_g^{\beta_1/2}$, and β_0 and β_1 are two constants for all G genes. In this exercise, we set $\beta_0 = -5$, $\beta_1 = 2$ and $n_1 = n_2 = 6$.

Differential expression data were generated by setting $\pi_1 G$ as the number of genes differentially expressed between the two conditions, with one half upregulated and the other half downregulated. Steps (i) through (iii) were repeated to form a correlation background for the $\pi_1 G$ regulated genes in the n samples. We then added (or subtracted) a term $2^{-1/2}\delta_g\omega_g$ to (from) the samples of one of the two conditions, where the coefficient δ_g was sampled from a uniform distribution $U(5,10)$ such that the true expression ratio is $1 + 2^{-1/2}e^{\beta_0/2}\delta_g \sim U(1.2901, 1.5804)$. Finally, we randomly replaced $\pi_1 G$ rows of the background X with the $\pi_1 G$ differentially expressed genes, and varied π_0 between 0.1 and 0.9 to simulate various data configurations, with 1000 iterations each. For each 'gene', we calculated the p-value for differential expression between the two conditions using the moderated t-statistic (Smyth, 2004).

3 THE LINEAR MODEL-BASED FRAMEWORK FOR π_0 ESTIMATION

Let $\#\{p_i \le \lambda\}$ be the number of hypotheses with p-values less than a cutoff λ, we transform the Storey's non-linear model to:

$$
\begin{aligned}
&\pi_0 = \frac{\#\{p_i > \lambda\}}{(1-\lambda)\,m}\\
&\Rightarrow 1-\pi_0(1-\lambda) = 1 - \frac{\#\{p_i > \lambda\}}{m} \Rightarrow 1-\pi_0(1-\lambda) = \frac{\#\{p_i \le \lambda\}}{m}\\
&\Rightarrow \#\{p_i \le \lambda\} = m\,(1-\pi_0(1-\lambda)) = m\,(1-\pi_0+\pi_0\lambda) \qquad (3)\\
&\Rightarrow \#\{p_i \le \lambda\} = (m\lambda)\pi_0 + m\,(1-\pi_0)
\end{aligned}
$$

Next, let y and x substitute $\#\{p_i \le \lambda\}$ and $m\lambda$, respectively, and Equation (3) can simply be rewritten as:

$$y = \pi_0 x + m(1-\pi_0) = \pi_0 x + m\pi_1 \qquad (4)$$

where $\pi_1 = 1-\pi_0$ is just the proportion of the alternative hypotheses. Equation (4) clearly reveals that the total number of hypotheses called significant with a p-value cutoff λ comprises two portions: one associated with null hypotheses (false positive), $\pi_0 x$ [or $(m\lambda)\pi_0$], and the other with alternative hypotheses (true positive), $m\pi_1$. Let $\gamma = y/m$ be the proportion of hypotheses called significant by λ, Equation (4) then becomes:

$$\gamma = \pi_0\lambda + \pi_1 \qquad (5)$$

The associated (λ,γ) plot actually represents the cumulative probability distribution (CPD) of p-values (Fig. 1A). Intuitively, from Equation (5) one may linearly fit the CPD of p-values over a proper range, and calculate the slope of this fitting line as the estimated π_0. Without loss of generality, given a range of λ, $\Delta = [\lambda_s, \lambda_e], 0.05 \le \lambda_s < \lambda_e \le 1$, the π_0 estimation can be written as:

$$\hat{\pi}_0(\lambda_s,\lambda_e) = \frac{\gamma_e - \gamma_s}{\lambda_e - \lambda_s} \qquad (6)$$

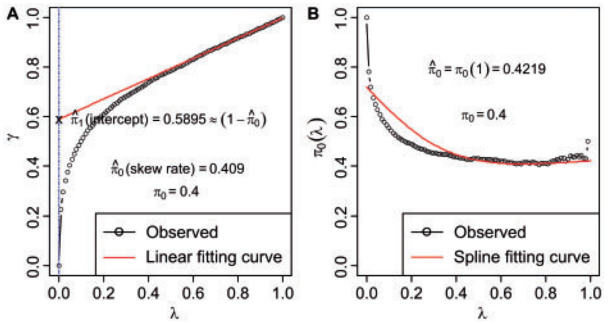

Fig. 1. Estimation of π_0 using (**A**) linear and (**B**) non-linear models. For illustration, the proportion of true null hypotheses is $\pi_0 = 0.4$.

where γ_s and γ_e represent the cumulative probabilities at λ_s and λ_e, respectively, and 0.05 is set according to the conventional p-value cutoff for statistical significance.

For data with a uniform p-value distribution, Equation (6) may be applied directly for π_0 estimation. This is shown in Figure 1, using the uniform simulation dataset. We obtained a linear fitting curve via Equation (6) over a range $\lambda \in [0.7, 1]$ (Fig. 1A), and a non-linear fitting curve using the CSS algorithm (Storey and Tibshirani, 2003) with default parameters (Fig. 1B). Given a true π_0 of 0.4, Equation (6) takes advantage of the data linearity on the right side of the (λ, γ) plot to yield a $\hat{\pi}_0 = 0.4089$, which is more accurate than $\hat{\pi}_0 = 0.4219$ generated by CSS. Large variation of $\pi_0(\lambda)$ at $\lambda \in [0.8, 1]$ makes fitting the non-linear model more error prone.

Equation (6) only uses the partial information in $\Delta = [\lambda_s, \lambda_e]$, e.g. [0.7, 1] in Figure 1A, for the π_0 estimation, and will be inadequate to handle datasets with non-uniform p-value distributions. In order to retain as much information as possible about the null hypotheses' p-value distribution, we devised the following strategy for the estimation of π_0. We first divide the (λ, γ) plot into a series of λ-segments (S). Thus, n segments can be formed as:

$$S = \{s_i, s_i = [\lambda_i, \lambda_{i+1}]\}_1^n, \quad 0.05 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{n+1} = 1 \tag{7}$$

We then linearly regress γ by λ using Equation (6) for each segment and obtain n local estimates of $\hat{\pi}_0^i = \hat{\pi}_0(\lambda_i, \lambda_{i+1}), i = 1, 2, \ldots, n$, in accordance with Equation (7). In view of the overall null p-value distribution being a mixture of local distributions, we estimate π_0 as:

$$\hat{\pi}_0 = D^{-1}(\alpha) \tag{8}$$

where D represents the cumulative distribution function of $\hat{\pi}_0^i$, D^{-1} represents its inverse (i.e. quantile) function, and $0 \le \alpha \le 1$ represents a given quantile point. The selection of an appropriate value of α will be addressed in Section 4.2. This π_0 estimator, which we termed the sliding linear model (SLIM), considers the collective effect of null p-values over the majority of their distribution range. Therefore, SLIM should be able to deal effectively with null p-values having a complex distribution pattern.

4 PARAMETERS OF THE SLIM ESTIMATOR

In this section, we first consider some properties of the q-value and their implications for the development of SLIM. We then discuss the optimization of SLIM parameters for π_0 estimation.

4.1 Properties of *q*-values

Based on Equation (1), the CPD curves of p-values and q-values in the (λ, γ) plot intersect at $\lambda = p_z$, where $z = \pi_0 m$. This is illustrated in Figure 2A and B (gray solid arrows) using the uniform simulation datasets with various π_0 scenarios. We observe that the q-value of a given test is larger than p-value (i.e. $q_i > p_i$) when $p_i < p_z$ and smaller than p-value ($q_i < p_i$) when $p_i > p_z$. This non-static relationship suggests that the common use of an arbitrary q-value cutoff, often at the same level as the p-value or FDR cutoff, is not universally appropriate.

Given π_0, let $p_{\max}$ be the maximum p-value among the truly significant (as opposed to called significant) tests; we infer that there exists a corresponding, hidden maximum FDR, denoted by FDR$^{\max}$. That is, no false negatives are encountered at $p_{\max}$, and all errors are due to false positives, estimated $p_{\max}\pi_0$. The FDR$^{\max}$ thus can be written as:

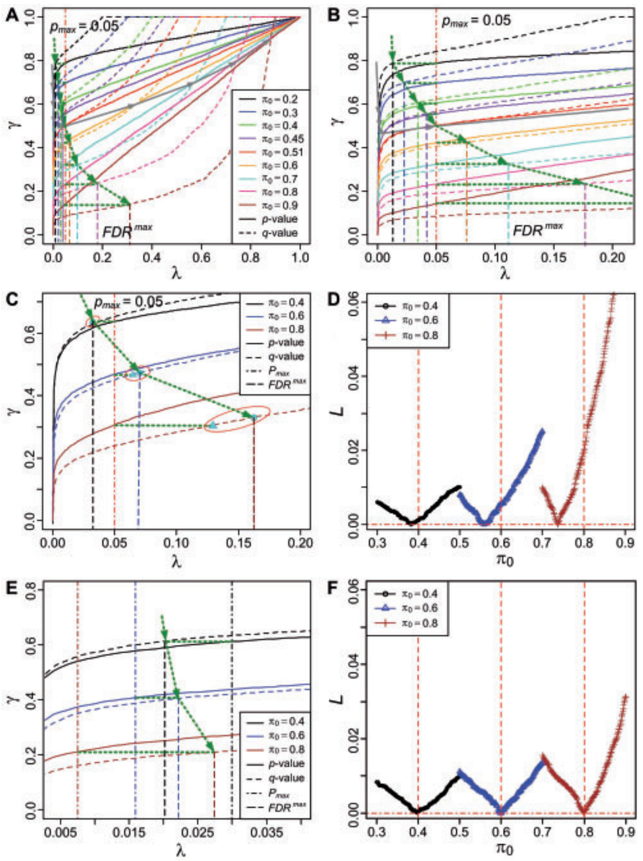

Fig. 2. Non-static relationships between the p- and q-values. (**A** and **B**) Uniform simulation datasets (B gives a close-up view of A); (**C–F**) non-uniform data. Gray solid arrows (A and B) indicate the intersection points of the two corresponding CPD curves, and green dashed arrows (A–C and E) depict the change of FDR$^{\max}$ across different π_0 scenarios. (D) and (F) illustrate the change of L (difference between the fractions of tests called significant by q-values and p-values) under a range of hypothetical π_0 values (true $\pi_0 \pm 0.1$) for three π_0 scenarios. A fixed $p_{\max} = 0.05$ was used in A through D, while a $p_{\max} = 0.03$, 0.02 and 0.01 was used for the three π_0 scenarios in E and F. Data presented in D are also shown in Table 2.

$$\mathrm{FDR}^{\max} = \frac{p_{\max}\pi_0}{\pi_1 + p_{\max}\pi_0} = \frac{p_{\max}\pi_0}{1-(1-p_{\max})\pi_0} \tag{9}$$

Taking $p_{\max} = 0.05$ as an example, we can calculate the FDR$^{\max}$ by Equation (9) for different π_0 scenarios (Supplementary Table S1). Since the rank of $p_{\max}$ is $\pi_1 m + p_{\max}\pi_0 m$, the q-value at $p_{\max}$ becomes $q = \mathrm{FDR}^{\max}$ according to Equation (1). Therefore, the FDR$^{\max}$ can be taken as the maximum q-value among all alternative tests. This means that the two hypothesis testing criteria, p-value and q-value, will call the same set of tests significant at their respective cutoffs, $p_{\max}$ and FDR$^{\max}$, because the q-value procedure does not change the order of hypotheses ranked by p-values.

Let L be the absolute difference between the fractions of tests called significant by the two (p-value and q-value) methods; we have $L = 0$ at the true value of π_0. Supplementary Figure S1 gives a geometric interpretation of the relationship in the (γ, λ)

plot. Figure 2A and B illustrates the relationship for uniform data: at $\lambda = \mathrm{FDR}^{\max}$, the CPD of q-values has a γ intercept same as that of the CPD of p-values at $\lambda = p_{\max}$ (green horizontal dashed lines). For non-uniform data, the irregular distribution of null p-values can skew this relationship, as shown for three π_0 scenarios at $p_{\max} = 0.05$, where $\hat{\pi}_0$ at the minimum L deviated from the true values (Fig. 2C and D). However, given a proper $p_{\max}$ (e.g. 0.03, 0.02 and 0.01 for the three π_0 scenarios, respectively), the relationship held again, and the minimum $L = 0$ gave accurate π_0 estimates (Fig. 2E and F). Extensive testing showed that L always reached a minimum around the true π_0 over a range of $p_{\max}$ (Supplementary Fig. S2). Together, the data suggest that the observed relationship between q- and p-values holds irrespective of data structure, and can be exploited to guide the parameter tuning of SLIM.

4.2 Parameter tuning of the SLIM estimator

The SLIM depends on three parameters: λ_1, n and α. λ_1 specifies the lower boundary of the first sliding segment used for π_0 estimation across the λ axis of the (λ, γ) plot. Considering that the smaller p-values are most likely to come from alternative hypotheses, a small λ_1 is preferred to maximize the range of data coverage while minimizing the influence of true positives on π_0 estimation. Analysis with various simulation datasets and π_0 scenarios showed that the mean relative errors (MREs) of π_0 estimates reached a minimum at $\lambda_1 < 0.2$ (Supplementary Fig. S3). We therefore set $\lambda_1 = 0.1$ as default. The parameter n specifies the number of λ-segments, and influences how the distribution of null p-values is estimated. For data encompassing a complex p-value distribution, a sufficiently large n is necessary to capture clusters of similar p-values. However, an overly large n may lead to smaller segments with abrupt changes in slope and increase π_0 estimation errors. We have found $n = 10$ to be generally robust against a range of data scenarios.

The quantile parameter α captures the collective effect of the n local π_0 estimates from the sliding segments, thus playing a crucial role in the global π_0 estimate by SLIM. Following the relationship between p- and q- values, the selection of a proper α can be solved as an optimization problem, i.e.

$$\text{Minimize } L = \left| \tilde{\pi}_q^{\alpha} - \tilde{\pi}_p^{\alpha} \right|, \quad s.t. \quad 0 \leq \alpha \leq 1 \tag{10}$$

where $\tilde{\pi}_q^{\alpha} = \#\{q_i < \mathrm{FDR}^{\max}\}/m$ and $\tilde{\pi}_p^{\alpha} = \#\{p_i < p_{\max}\}/m$ represent the fraction of tests called significant by q-values and p-values, respectively, under $\hat{\pi}_0^{\alpha} = D^{-1}(\alpha)$. Equation (10) aims to choose the optimal α by minimizing the difference L between $\tilde{\pi}_q^{\alpha}$ and $\tilde{\pi}_p^{\alpha}$ at a given $p_{\max}$ to approximate the global π_0. We devised an α-searching procedure to solve the optimization problem. We first form a candidate set of α, $A = \{\alpha_i = \frac{1}{B} i, i = 1, 2, \ldots, B\}$, from which a proper α will be selected. To gain sufficient granularity, we set B to be no less than 100. For each α_i, we then calculate $\hat{\pi}_0^{\alpha} = D^{-1}(\alpha)$ and the difference L between $\tilde{\pi}_q^{\alpha}$ and $\tilde{\pi}_p^{\alpha}$ using $\pi_0 = \hat{\pi}_0^{\alpha}$ and a given $p_{\max}$ (see below). Finally, the $\tilde{\alpha}$ with the minimum L is chosen as the final value of α.

We examined the effect of varying $p_{\max}$ on π_0 estimation using simulated uniform, non-uniform and gene expression data. Although the results varied depending on data structure and π_0 scenario, too large (>0.6) and too small (<0.01) a $p_{\max}$ tended to give rise to large MREs of $\hat{\pi}_0$, based on 1000 random iterations (Supplementary Fig. S4). In most cases, the MRE was less than 0.1 for a $p_{\max}$ between 0.01 and 0.1, suggesting that SLIM is relatively insensitive to $p_{\max}$ (i.e. a true $p_{\max}$ is not necessary in practice, since it is an unknown property). We recommend setting $p_{\max}$ at 0.05 as default. For further optimization, users may wish to test multiple $p_{\max}$ and examine the CPD curves of p-values and q-values in the resultant (λ, γ) plots, as illustrated in Figure 2, in order to select an optimal $p_{\max}$.

Overall, we transform the α optimization problem into two user-definable and insensitive parameters B and $p_{\max}$, thereby simplifying the implementation of SLIM. A procedural framework for SLIM is depicted in Supplementary Figure S5.

5 SIMULATION EXPERIMENTS

We applied SLIM to the three types of simulation datasets from Section 2 in comparison with four other methods: CSS (Storey and Tibshirani, 2003), AE (Jiang and Doerge, 2008), CBMM (Markitsis and Lai, 2010) and Locfdr (Efron, 2007a).

5.1 Uniform and non-uniform simulation data

Table 1 summarizes the mean errors and SD of π_0 estimates for each of the five methods across nine π_0 scenarios on the uniform and non-uniform data. The results show that SLIM achieved overall lower mean errors for π_0 estimation than the other four methods, especially for the non-uniform data. CBMM also worked well, but exhibited a decreasing accuracy as π_0 increases. As reported by Efron (2007a), Locfdr is applicable only to datasets with a large π_0, and it was indeed most accurate for the non-uniform data with $\pi_0 = 0.9$ (Table 1). CSS and AE worked reasonably well with uniform data,

Table 1. Comparison of SLIM with four previous methods on uniform and non-uniform simulation datasets

π_0	$\tilde{\alpha}$	CSS	AE	CBMM	Locfdr	SLIM
Uniform data						
0.1	0.35–0.89	7.5/9.4	20.4/33.4	**1.1/1.4**	NA	3.8/1.6
0.2	0.26–0.78	11.0/13.7	19.4/30.6	**1.8/1.8**	NA	3.9/1.8
0.3	0.21–0.69	13.8/18.2	12.4/16.5	**1.9/2.0**	NA	3.4/4.3
0.4	0.25–0.65	15.8/18.4	10.7/12.4	1.8/2.2	NA	**1.7/1.2**
0.5	0.33–0.78	16.8/19.8	11.1/12.3	**2.3/3.1**	NA	2.4/2.5
0.6	0.31–0.71	17.6/20.3	11.6/9.8	2.4/3.1	NA	**1.9/1.9**
0.7	0.29–0.82	18.6/21.1	8.4/8.3	2.5/3.2	NA	**1.4/1.4**
0.8	0.29–0.80	19.5/23.0	9.0/9.6	2.6/3.1	43.5/36.1	**1.9/1.6**
0.9	0.32–0.83	19.2/24.0	7.4/8.7	3.6/4.0	8.8/11.9	**1.6/2.4**
Non-uniform data						
0.1	0.74–0.90	78.1/7.2	29.7/3.8	12.5/1.3	NA	**11.7/1.6**
0.2	0.76–0.90	147.5/9.8	61.5/6.4	23.4/2.2	NA	**17.7/1.7**
0.3	0.76–0.89	231.6/14.5	91.7/6.8	36.8/2.5	NA	**28.2/4.5**
0.4	0.75–0.89	299.3/14.5	117.2/5.1	48.6/3.2	NA	**35.5/2.7**
0.5	0.74–0.90	379.0/12.3	130.5/13.6	60.1/4.0	NA	**42.8/3.9**
0.6	0.70–0.89	NA	138.2/7.4	71.2/3.1	NA	**39.9/2.4**
0.7	0.68–0.88	NA	163.2/6.3	83.3/2.8	NA*	**60.6/3.7**
0.8	0.71–0.89	NA	183.9/7.6	93.4/2.6	76.4/20.4	**68.2/4.7**
0.9	0.77–0.90	NA	NA	102.7/3.7	**30.7/30.6**	77.1/4.6

The smallest value for each π_0 scenario is highlighted in boldface. NA, the method did not work (in the cases of CSS and AE, $\hat{\pi}_0 = 1$ in all random iterations); *, mean error and SD could not be determined because Locfdr did not work for some of the 1000 random datasets. Data are presented as mean error/SD ($\times 10^{-3}$) of π_0 estimates from 1000 random iterations.

but lost their accuracy for non-uniform data, suggesting their general deficiency in dealing with complex p-value distributions. Table 1 also reports the range of α determined by the optimization scheme of Equation (10): $\tilde{\alpha}$ was around the mid-range (∼0.3 to 0.8) for the uniform data, and trended toward larger values (∼0.7 to 0.9) for the non-uniform data.

Based on one of the non-uniform datasets ($\pi_0 = 0.8$), we compared the computation cost of SLIM and three other previous methods (CSS did not work for this dataset). The CPU time (in second) was 1.1 (SLIM), 13.00 (AE), 0.17 (Locfdr) and 0.50 (CBMM) using an intel®Core2 Duo 3 GHz processor with 3 GB of RAM and the Microsoft Windows XP operating system. The result suggests that SLIM is not computationally demanding.

5.2 Gene expression simulation data with hidden dependence structures

The simulated gene expression analysis showed that SLIM outperformed the other four methods based on the mean errors and SDs of $\hat{\pi}_0$ (Fig. 3). We calculated the number of alternative hypotheses predicted by each method, i.e. $(1-\hat{\pi}_0)\times 10000$, as a proxy of the number of differentially expressed (DE) genes for each π_0 scenario. Because the true null and alternative hypotheses are known, the false positive rate (FPR), false negative rate (FNR), FDR and accuracy of each method can be evaluated. SLIM achieved an overall low FPR, FNR and FDR (Table 2). The ability of SLIM to balance between false positive and false negative errors led to an overall high accuracy (∼0.99). The previous methods suffered from various trade-offs. For instance, CBMM generated the lowest FNR in most scenarios at the expense of higher FPR and FDR, leading to reduced accuracy compared with SLIM. Locfdr was effective for large π_0 scenarios, producing a low level of FPR and FDR in these cases, but at the expense of higher FNR and an overall low accuracy. These results demonstrate the power and robustness of SLIM in coping with data dependence structures. We also observed that the π_1 $(=1-\pi_0)$ estimate of SLIM may be used as a cutoff for DE gene selection with sufficient FDR control and a high level of accuracy (Table 2). In practice, the FDR$^{\max}$ of the maximum p-value among the $(1-\hat{\pi}_0)\times m$ DE genes may be taken as the FDR for these DE genes. For the simulation data, the FDR$^{\max}$ values of SLIM-called DE genes were found to be very close to the actual FDR (most of the differences were around 0.001).

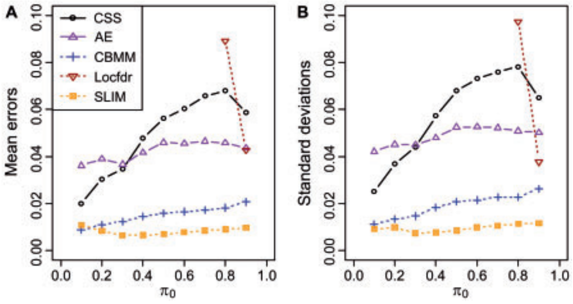

Fig. 3. Comparison of mean errors (**A**) and SDs (**B**) of π_0 estimated by the five methods in simulated gene expression analysis.

To illustrate the influence of hidden dependence structures on the π_0 estimation, representative p-value distributions of the three types of simulation data were shown in histograms and (λ,γ) plots. Compared with the uniform data (Fig. 4A), the non-uniform data have a skewed distribution of null hypotheses (Fig. 4B), which complicates π_0 estimation. The addition of dependence structures (Fig. 4C–G) led to various irregular p-value distribution patterns. These differences in histograms are reflected in the CPD curves of p-values in the (λ,γ) plot (Fig. 4H). The non-uniform data show sinusoidal deviations from the straight line generated by the uniform data, while the gene expression simulation datasets are intermediate in shape between those for the uniform and non-uniform data. These comparisons help explain why π_0 estimation methods developed on the assumptions of data independence and uniformity do not work as effectively for datasets containing dependence structures.

6 APPLICATION TO REAL-WORLD GENE EXPRESSION DATASETS

To test SLIM in practice, we examined two Affymetrix microarray datasets from a *Populus* stress response experiment (GEO no. GSE14515, Yuan *et al.*, 2009). The experiment monitored gene

Table 2. FPR, FNR, FDR and accuracy of the five methods for the gene expression simulation data

π_0	DE gene no.	FPR					FNR					FDR					Accuracy				
		CSS	AE	CBMM	Locfdr	SLIM	CSS	AE	CBMM	Locfdr	SLIM	CSS	AE	CBMM	Locfdr	SLIM	CSS	AE	CBMM	Locfdr	SLIM
0.1	9000	0.101	**0.066**	0.082	NA	0.119	0.013	0.034	0.004	NA	**0.002**	0.011	**0.007**	0.009	NA	0.013	0.978	0.962	**0.989**	NA	0.986
0.2	8000	0.071	0.048	0.053	NA	**0.047**	0.022	0.039	0.005	NA	**0.004**	0.017	0.012	0.013	NA	**0.011**	0.968	0.959	0.986	NA	**0.988**
0.3	7000	0.055	0.035	0.040	NA	**0.024**	0.029	0.040	**0.006**	NA	**0.006**	0.022	0.014	0.017	NA	**0.010**	0.964	0.961	0.984	NA	**0.989**
0.4	6000	0.051	0.032	0.034	NA	**0.018**	0.048	0.051	**0.007**	NA	**0.007**	0.031	0.020	0.022	NA	**0.012**	0.951	0.957	0.982	NA	**0.989**
0.5	5000	0.040	0.027	0.029	NA	**0.013**	0.075	0.068	**0.009**	NA	0.010	0.036	0.024	0.027	NA	**0.013**	0.942	0.953	0.981	NA	**0.989**
0.6	4000	0.029	0.021	0.024	NA	**0.010**	0.110	0.086	**0.012**	NA	0.013	0.037	0.028	0.034	NA	**0.015**	0.938	0.953	0.981	NA	**0.989**
0.7	3000	0.025	0.017	0.021	NA∗	**0.008**	0.164	0.117	**0.017**	NA∗	0.018	0.047	0.034	0.043	NA∗	**0.019**	0.933	0.953	0.980	NA∗	**0.989**
0.8	2000	NA	0.014	0.019	0.011	**0.007**	NA	0.174	**0.024**	0.404	0.026	NA	0.047	0.064	**0.011**	0.028	NA	0.954	0.980	0.911	**0.989**
0.9	1000	NA	NA	0.019	0.008	**0.007**	NA	NA	**0.045**	0.354	0.046	NA	NA	0.121	0.054	**0.053**	NA	NA	0.979	0.957	**0.989**

The best performer (lowest error rate or highest accuracy) from each π_0 scenario is shown in boldface. FPR is calculated as the number of false positives over the number of false positives plus true negatives; FNR is the number of false negatives over the number of false negatives plus true positives; FDR is the number of false positives over the number of false positives plus true positives; and the accuracy is the number of true positives plus true negatives over the total number of tests. NA: the method did not work; *: Locfdr did not work for some of the 1000 random datasets.

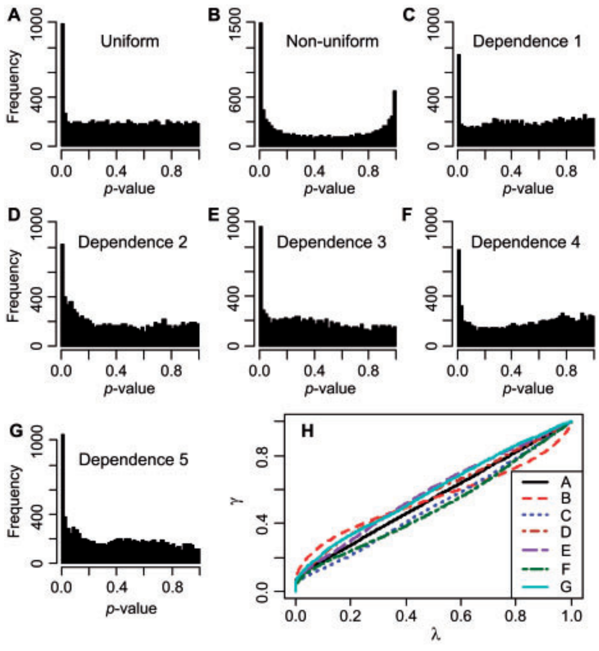

Fig. 4. Histograms and (λ,γ) plots of p-values in different simulation datasets. (**A**–**G**) Histograms for uniform (A) and non-uniform (B) data, as well as data with five different dependence structures (C–G). (**H**) The corresponding CPD curves of p-values in the (λ,γ) plot.

Table 3. Estimates of $\hat{\pi}_0$ and the corresponding DE gene numbers by different methods for the two *Populus* nitrogen stress datasets

		CSS	AE	CBMM	Locfdr	SLIM	$\tau_p=0.01$	$\tau_p=0.05$	$\tau_p=0.1$
YL	$\hat{\pi}_0$	0.45	0.47	0.41	0.96	0.73	–	–	–
	DE no.	7333	7099	7916	594	3649	1913	4134	5475
ML	$\hat{\pi}_0$	0.46	0.49	0.43	0.97	0.75	–	–	–
	DE no.	7075	6751	7542	398	3285	1648	3787	5109

expression changes in young (YL) and mature (ML) leaves in response to 4 week nitrogen depletion (normal versus low N), each with two biological replicates. Raw hybridization signals were processed using the R package affyPLM, and $m=13\,335$ probes that passed quality control filtering (raw intensities ≥ 100 in both replicates of at least one condition) were obtained for further analysis. We used the moderated t-statistics (Smyth, 2004) to summarize the expression differences between treatments and to calculate the corresponding p-values for each gene.

As shown in Table 3, SLIM obtained a $\hat{\pi}_0$ of 0.73 and 0.75 for the YL and ML datasets, respectively, with a corresponding $\tilde{\alpha}$ of 0.7 and 0.72. In comparison, the estimates by CSS, AE and CBMM were much lower (<0.5), while those by Locfdr were very high (>0.95). Because the true π_0 is unknown, the reproducibility of the five methods for π_0 estimation was evaluated. We followed the bootstrap procedure of Markitsis and Lai (2010) to re-sample the p-values 500 times with replacement and equal probabilities, and to determine the 95% confidence interval (CI) of the resultant π_0 estimates from the 500 resampling datasets. SLIM and CBMM both exhibited excellent stability as shown in the boxplot (Fig. 5A for the YL dataset), with a tight 95% CI (SLIM: 0.723–0.731 and CBMM: 0.404–0.413). In contrast, π_0 estimates by the Locfdr had the widest 95% CI. Similar results were obtained for the ML dataset (Supplementary Fig. S6A).

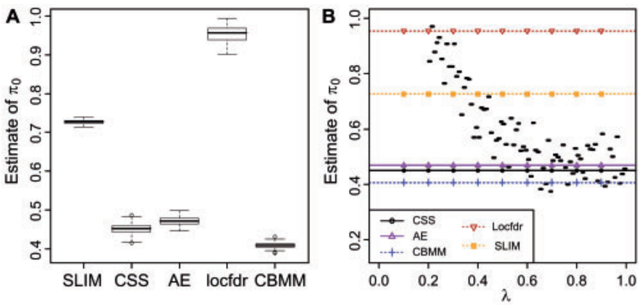

Fig. 5. Comparison of the five methods on the *Populus* YL dataset. (**A**) Boxplots of bootstrapped π_0 estimates. (**B**) The relationship between local π_0 estimates across various segments $\lambda\in[0.2,1]$ (short black bars) and the global π_0 estimates (horizontal lines).

Several lines of evidence suggest that the higher estimates of SLIM and Locfdr are more reasonable than those of CSS, AE or CBMM. First, Locfdr was designed for datasets having a large π_0 (Efron, 2007a), and our simulation analysis confirmed that it is indeed not applicable to π_0 scenarios <0.7, regardless of the dependence structure (Tables 1 and 2). The mere fact that Locfdr worked suggests that the two datasets are more likely to have a π_0 greater than 0.7. Second, we estimated the proxy number of DE genes (alternative hypotheses) by each method using $(1-\hat{\pi}_0)\times 13\,335$, and compared the results with DE genes selected by p-value cutoffs (Table 3). The underestimate of $\hat{\pi}_0$ by the CSS, AE and CBMM methods is evident, judging from the large numbers of corresponding DE gene proxies—greater than those obtained by $p\leq 0.1$—in both datasets. Third, based on the local estimates across various segments of $\lambda\in[0.2,1]$, the $\hat{\pi}_0$ of CSS, AE and CBMM appeared to be biased toward $\lambda\in[0.6,1]$, as shown for the YL data in Figure 5B. The local estimates were obtained using the linear estimator of Equation (6) and with segments $s_i=[0.2+0.01\times(i-1), 0.2+0.01\times i], i=1,2,\ldots,80$. Similar results were obtained for the ML dataset (Supplementary Fig. S6B). These findings suggest that the CSS, AE and CBMM methods largely disregard the contribution of null p-values from [0,0.6], leading to their underestimation of the π_0. The Locfdr has a tendency to overestimate π_0 (Fig. 5B) and incur a higher FNR based on our simulation analysis (Table 2). Accordingly, it predicted fewer DE genes for the two datasets, compared with those called by a stringent p cutoff $\tau_p=0.01$. The numbers of SLIM-predicted DE genes for the YL and ML datasets (3649 and 3285, respectively) were less than those (4134 and 3787) called by $\tau_p=0.05$, consistent with the default setting of $p_{\max}=0.05$ in SLIM. The calculated $\text{FDR}^{\max}$ was 0.12 and 0.13 for YL and ML, respectively, based on Equation (9). Examination of the CPD curves of p- and q-values showed that their respective γ intercepts at $p_{\max}$ and $\text{FDR}^{\max}$ are in close proximity, with a negligible L ($\sim 10^{-4}$) in both datasets (Supplementary Fig. S7). This suggests that the SLIM-estimated π_0 should be near their theoretical values. On these bases, we argue that SLIM provides the most reasonable π_0 estimate in practice. Users can further reduce the list of DE genes, if deemed necessary, by other criteria, such as q-value, FDR or fold-change cutoffs.

7 DISCUSSION

An important issue in multiple hypothesis testing is how to deal with the dependencies hidden among thousands of tests. Efron (2007b) has shown that correlation among variables considerably changes the theoretical null distribution patterns. We also observed that dependence structures lead to distorted distributions of null p-values, and this likely underlies the relatively large π_0 estimation errors by methods developed under the assumption of data independence. The Locfdr approach (Efron, 2007a) was specifically designed to handle data containing dependence structures, but it is only applicable when π_0 is large. CBMM uses a censored beta-uniform mixture model to fit the distorted p-value distribution, alleviating to some degree the difficulty caused by dependence. SLIM is based on a linear model transformed from the non-linear λ estimator (Storey, 2002). The superior performance of SLIM can be ascribed to its data partitioning and optimization schemes. SLIM uses a sliding linear model to partition data into local dependence blocks. This reduces data complexity, while enabling SLIM to utilize information from a broader range of p-value distribution for π_0 estimation. Using simulated data, we uncovered a non-static relationship between p-values and q-values of a given set of tests that is influenced by data structure and π_0 scenarios. SLIM employs an optimization scheme to explicitly exploit this relationship by minimizing the difference (L) between the fractions of tests called significant by the p-value and q-value methods. The optimization scheme is particularly important to balance between positive and negative errors, thereby achieving FDR control. Thus, SLIM effectively handles hidden dependence without the need to empirically adjust the null p-value distributions.

The selection of a proper q-value cutoff in multiple hypothesis testing is not trivial, especially given the dependence of the q-value calculation on the π_0 estimation. Recalling that the number of significant tests in a given experiment is simply $\pi_1 m$, we argue that an accurate estimation of π_0 can serve as an alternative to q-value-based significance testing. Using simulated data, SLIM was shown to outperform the other methods, achieving the lowest FDR overall, accompanied by the highest degree of accuracy in declaring significant tests. This suggests that SLIM can be used as a stand-alone tool in multiple testing for determination of significant tests.

In summary, SLIM is a robust estimator especially suited for datasets with non-uniform p-value distribution patterns due to data dependence. SLIM is computationally efficient and easy to implement. It requires four user-selected parameters, n, λ_1, $p_{\max}$ and B (the latter two are proxies of the quantile parameter α). We recommend $n=10$, $\lambda_1=0.1$, $B=100$ and $p_{\max}=0.05$ as the default settings. Users may wish to test a higher B to ensure sufficient granularity, and a range of $p_{\max}$ for optimal α selection and π_0 estimate. In addition to microarray analysis, SLIM has been applied to metabolite profiling analysis in our laboratory, and should be applicable to a wide range of experiments

Funding: US National Science Foundation Plant Genome (grant DBI-0836433 to C.-J.T.); National Science Funding of China (No. 30900321 to H.-Q.W.).

Conflict of Interest: none declared.

REFERENCES

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

Clarke,S. and Hall,P. (2009) Robustness of multiple testing procedures against dependence. *Ann. Stat.*, **37**, 332–358.

Clarke,C.J. and Haselden,J.N. (2008) Metabolic profiling as a tool for understanding mechanisms of toxicity. *Toxicol. Pathol.,* **36**, 140–147.

Efron,B. (2007a) Size, power and false discovery rates. *Ann. Stat.*, **35**, 1351–1377.

Efron,B. (2007b) Correlation and large-scale simultaneous significance testing, *J. Am. Stat. Assoc.,* **102**, 93–103.

Jiang,H. and Doerge,R.W. (2008) Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Inform.*, **6**, 25–32.

Lander,E. and Kruglyak,L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, **11**, 241–247.

Markitsis,A. and Lai,Y. (2010) A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, **26**, 640–646.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Qin,H. *et al.* (2008) An efficient method to identify differentially expressed genes in microarray experiments. *Bioinformatics*, **24**, 1583–1589.

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

Stafford,P. and Yidong,C. (2007) Expression technology - a review of the performance and interpretation of expression microarrays. *IEEE Signal Proc. Mag.,* **24**, 18–26.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B,* **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies, *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Yuan,Y. *et al.* (2009) Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in *Populus* and *Arabidopsis. Proc. Natl Acad. Sci. USA*, **106**, 22020–22025.