BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 801–807
doi:10.1093/bioinformatics/btr671

Gene expression

Advance Access publication February 9, 2012

# A novel significance score for gene selection and ranking

Yufei Xiao [1,2,*], Tzu-Hung Hsiao[2], Uthra Suresh [2], Hung-I Harry Chen [2], Xiaowu Wu [3,4], Steven E. Wolf [3,4] and Yidong Chen [1,2,5,*]

[1]Department of Epidemiology and Biostatistics, [2]Computational Biology and Bioinformatics Division, Greehey Children's Cancer Research Institute, [3]Department of Surgery, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, [4]United States Army Institute of Surgical Research, Fort Sam Houston, TX 78234 and [5]Cancer Therapy and Research Center, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

Associate Editor: Janet Kelso

**ABSTRACT**

**Motivation:** When identifying differentially expressed (DE) genes from high-throughput gene expression measurements, we would like to take both statistical significance (such as *P*-value) and biological relevance (such as fold change) into consideration. In gene set enrichment analysis (GSEA), a score that can combine fold change and *P*-value together is needed for better gene ranking.

**Results:** We defined a gene significance score $\pi$-value by combining expression fold change and statistical significance (*P*-value), and explored its statistical properties. When compared to various existing methods, $\pi$-value based approach is more robust in selecting DE genes, with the largest area under curve in its receiver operating characteristic curve. We applied $\pi$-value to GSEA and found it comparable to *P*-value and *t*-statistic based methods, with added protection against false discovery in certain situations. Finally, in a gene functional study of breast cancer profiles, we showed that using $\pi$-value helps elucidating otherwise overlooked important biological functions.

**Availability:** http://gccri.uthscsa.edu/Pi_Value_Supplementary.asp

**Contact:** xy@ieee.org; cheny8@uthscsa.edu

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

Received on May 11, 2011; revised on October 12, 2011; accepted on November 14, 2011

## 1 INTRODUCTION

In this article, we introduce a gene significance score, $\pi$-value, for robust selection of differentially expressed (DE) genes, and demonstrate its application in gene set enrichment analysis (GSEA). In developing the concept, we are inspired by two facts. First, there is discrepancy between the statistical and biological meanings of differential expression. Second, a proper score is required to evaluate and rank genes in GSEA.

It was pointed out in McCarthy and Smyth 2009 that statistically speaking, genes showing systematic difference between two conditions are considered DE, whereas in biological term, DE means the difference in gene expression is sufficiently large. In some biological applications, DE genes were selected by gene expression level change (such as using 2-fold change as cutoff), and the variability of gene expression was not taken into account. Meanwhile, some statistical methods consisted of finding 'informative genes' whose expression levels correlate with or are predictive of class labels, such as in Ambroise and McLachlan 2002; Golub *et al.* 1999. These approaches usually found not all DE genes, but a subset of DE genes that distinguished sample classes. To detect all DE genes, we usually conduct hypothesis testing of $\mu_0 = \mu_1$ against $\mu_0 \neq \mu_1$ or the like, such as methods based on expression ratio statistic Chen *et al.*, 1997, distinctness by Kolmogorov–Smirnov distance and similarity of expression profiles by Pearson's correlation coefficient Huang *et al.*, 2010, *t*-statistic and its variants Smyth, 2004; Tusher *et al.*, 2001.

In hypothesis test-based statistical methods, *P*-value of the test statistic is the basis of gene selection, as a result of which, we face the following two issues:

- The 'small fold change, small variance' (SFSV) issue. When a gene's variance is small, a slight expression level change may result in significant *P*-value and conclusion of statistical significance. However, a small expression change has questionable biological justification, which frequently leads to false discovery.

- The 'large fold change, large variance' (LFLV) issue. A gene with considerable fold change (FC) may possess a large variance, such as dysregulated genes in a disease condition. In such a situation, a large expression change may be accompanied by a non-significant *P*-value, making it possible for us to miss biologically meaningful but highly volatile changes.

Since variance estimation is part of the issues, which is often worsened by small sample size, improvements were proposed to adjust estimated variance, such as SAM test Tusher *et al.*, 2001, empirical Bayes method (moderated *t*-test) Smyth, 2004 and regularized *t*-test Baldi and Long, 2001. Nevertheless, even with accurate variance estimation, problem may arise when we select genes solely by *P*-value resulting from a hypothesis test.

To address the SFSV issue, it is common to combine *P*-value and FC criteria, such as 2-fold change and $P \leq 0.05$. Such approach is *ad hoc* Cui and Churchill, 2003; Yanofsky and Bickel, 2010. Meanwhile, methods that incorporate or improve the original FC criterion were proposed, such as a better defined FC limit Mutch *et al.*, 2002, a better statistical model Fu *et al.*, 2006, or

*To whom correspondence should be addressed.

variable threshold according to intensity measurement Mariani *et al.*, 2003; Yang *et al.*, 2002. Methods that directly incorporate FC criterion into test statistics provide significant improvement to meet the practical needs McCarthy and Smyth, 2009; Montazeri *et al.*, 2010. In McCarthy and Smyth 2009, TREAT, a hypothesis test relative to a FC threshold, was proposed with $H_0 : |\mu_0 - \mu_1| \leq \tau$ and $H_1 : |\mu_0 - \mu_1| > \tau$. TREAT requires that a threshold $\tau$ be chosen prior to the test, and if $\tau$ changes, all genes need to be tested again and *P*-values re-calculated. While TREAT has adequately addressed the SFSV issue, the LFLV issue remains a problem, which is fairly common in human clinical studies where patient-to-patient variation far exceeds that of well-controlled cell culture or animal studies, and better variance estimation does not help Wu *et al.*, 2010.

Motivated by aforementioned difficulties, we propose a gene significance score called $\pi$-value, which combines FC and *P*-value into one score. It provides a decision trade-off between FC and *P*-value, and offers a new means to rank and select genes. Our unique contributions are as follows: (i) $\pi$-value does not introduce any new statistical test, and its computation is simple; (ii) it is convenient to adjust the number of selected DE genes by changing the threshold, and no re-computation is needed; (iii) it addresses both SFSV and LFLV issues. Unlike most existing methods, this approach may retain some genes with large FCs but non-significant *P*-values, because they may be biologically important and worthy of further investigation; and (iv) $\pi$ assigns one single score to a gene by its FC and *P*-value, and thus it is useful in gene ranking and can be naturally applied to GSEA.

## 2 MATERIALS AND METHODS

### 2.1 Definition

Given pre-processed and normalized microarray data of two sample classes ($C_1$ and $C_2$) corresponding to two biological conditions (such as test and control groups), we assume there are $M_1$ and $M_2$ samples under $C_1$ and $C_2$, respectively, and $N$ probe sets on each array. Each probe set usually represents a gene, and we will use probe set and gene interchangeably unless specified otherwise. Let $g_{(k)ij}$ be $\log_2$-transformed expression level of the $i$-th gene in the $j$-th sample of class $C_k$ ($k$ = 1, 2), then the entire processed microarray data can be represented in two data matrices, $G_1$ and $G_2$, where

$$G_k = \begin{bmatrix} g_{(k)11} & g_{(k)12} & \cdots & g_{(k)1M_k} \\ g_{(k)21} & g_{(k)22} & \cdots & g_{(k)2M_k} \\ \vdots & \vdots & \cdots & \vdots \\ g_{(k)N1} & g_{(k)N2} & \cdots & g_{(k)NM_k} \end{bmatrix}, \quad k = 1, 2.$$

Let the mean expression of the $i$-th gene in sample class $C_k$ be $g_{(k)i\cdot}$, then the log-ratio (LR) and log-fold change (LFC) of this gene's expression are denoted as $x_i = g_{(1)i\cdot} - g_{(2)i\cdot}$ and $\phi_i = |g_{(1)i\cdot} - g_{(2)i\cdot}|$, respectively. In the literature, FC refers to the quantity $2^{\phi_i}$, and an alternative definition is $2^{\phi_i} \cdot \text{sign}(g_{(1)i\cdot} - g_{(2)i\cdot})$.

To select DE genes between two sample classes, we can employ one of the following decisions:

1. Correlation decision: selection of genes based on their correlation between expression levels and the corresponding phenotypes.
2. FC decision.
3. *P*-value decision associated with equal-mean hypothesis test. Usually a *t*-test or its variants is applied, and when multiple genes are involved, *P*-values are often adjusted to account for the effect of multiple testings, such as the Benjamini–Hochberg method Benjamini and Hochberg, 1995.
4. Decision cascade: a common *ad hoc* combination of two or more decisions.
5. *A priori* information fusion, such as incorporating FC threshold into *t*-test in TREAT McCarthy and Smyth, 2009.

Here we propose *a posteriori* information fusion scheme, to combine FC and *P*-value into one score after their individual evaluation, rather than fuse them earlier. Given $\phi_i$ and $p_i$, LFC and *P*-value of the $i$-th gene resulting from a hypothesis test, we define $\pi$-value as

$$\pi_i = \phi_i \cdot (-\log_{10} p_i). \tag{1}$$

$\pi$-value is non-negative, and the larger it is, the more significant gene expression change is. The unique combination of FC and *P*-value is motivated by our intention to transform *P*-value by FC, while still maintaining the characteristics of *P*-value. To see what $\pi$ suggests, we observe that

$$\pi_i = \log_{10} \frac{1}{p_i^{\phi_i}} \Rightarrow \Pi = 10^{-\pi_i} = p_i^{\phi_i}, \tag{2}$$

thus $\Pi = 10^{-\pi_i}$ can be viewed as a transformed *P*-value, and $0 \leq \Pi \leq 1$. When $\phi_i = 1$, namely 2-fold change, we will have $\Pi = p_i$, and $\Pi$ equals the regular *P*-value. For $\phi_i < 1$, namely less than 2-fold change, we have $\Pi > p_i$ and *P*-value is penalized; for $\phi_i > 1$, namely more than 2-fold change, we have $\Pi < p_i$, *P*-value is enhanced. Therefore, by defining $\pi$ to be the product of LFC and log-transformed *P*-value, we adjust gene's statistical significance by the amount of FC. A discussion of decision boundaries of FC, *P*-value and $\pi$-value on volcano plot is provided in Section 2.3. IN Supplementary Materials

In Equation (1), $p_i$ can be *P*-value from regular *t*-test, SAM Tusher *et al.*, 2001 or moderated *t*-test Smyth, 2004. Moreover, $\pi$ is not tied with specific hypothesis testing; for instance, it can be based on test of variance (such as *F*-test) or TREAT. Finally, $p_i$ can be raw *P*-value or adjusted *P*-value. These facts make $\pi$ a versatile gene significance score and suited for other applications such as GSEA. In this article, we will simply use *P*-value of the regular *t*-test in $\pi$.

### 2.2 Statistical characterization

Assuming LR follows a normal distribution, *P*-value follows a uniform [0,1] distribution, and LR and *P*-value are independent, we can show that the probability density function (PDF) of $\pi$ is (Section 1.1 in Supplementary Materials),

$$f_\Pi(z) = \sqrt{\frac{2}{\pi\sigma^2}} \lambda \int_0^\infty \frac{1}{x} e^{-(x^2/2\sigma^2 + \lambda z/x)} dx, \quad z \geq 0, \tag{3}$$

where $\lambda = \ln 10$, and $\sigma^2$ is the variance of LR. In reality, LR and *P*-value are dependent. The distribution of $\pi$ under independence and dependence assumptions are shown in Figure 1 (simulation details are provided in Section 1.1 in Supplementary Materials). In either case, the PDF of $\pi$ peaks at 0, decreases monotonically with $\pi$ and decreases approximately exponentially when $\pi$ is sufficiently large. The dependent PDF drops less steeply than the independent case, although they have similar trends.

To use $\pi$ for detecting DE genes at a given significance level $\alpha$, we need to determine its critical value with respect to $\alpha$. *P*-value's distribution does not change by dataset and platform. However, LR follows a normal distribution whose variance $\sigma^2$ can vary by dataset and platform, which impacts the distribution of $\pi$ as well as critical values. The method of computing critical values under independence or dependence is provided in the Supplementary Materials. Table 1 shows an example of the critical values of $\pi$ with respect to different $\alpha$'s under the independence and dependence assumptions ($\sigma = 0.48$ for the LR), respectively, and the simulation details are in the Supplementary Materials. Table 1 is not a generalized result, but rather a demonstration of the difference between the dependence and independence conditions. We have implemented a significance assessment tool for dependence condition given any input microarray dataset (Supplementary Material).

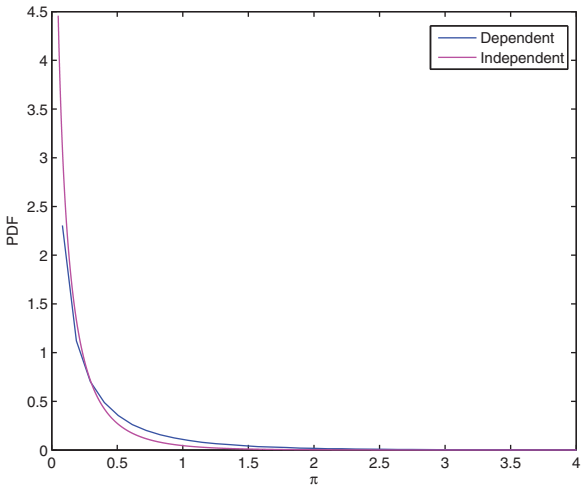


**Fig. 1.** PDF of $\pi$ when LFC and $P$-value are dependent (blue) or independent (magenta).

**Table 1.** CV of $\pi$ under dependence or independence

| $\alpha$ | $CV^{dep}$ | $CV^{ind}$ | $\alpha$ | $CV^{dep}$ | $CV^{ind}$ |
|---|---|---|---|---|---|
| 0.2 | 0.4079 | 0.2572 | 0.001 | 3.8434 | 2.1281 |
| 0.1 | 0.7319 | 0.4292 | 0.0005 | 4.4122 | 2.4503 |
| 0.05 | 1.1082 | 0.6274 | 0.0002 | 5.1559 | 2.8988 |
| 0.02 | 1.6657 | 0.9246 | 0.0001 | 5.7964 | 3.2543 |
| 0.01 | 2.1270 | 1.1733 | 0.0001 | 6.4407 | 3.6233 |
| 0.005 | 2.6138 | 1.4408 | 0.00002 | 7.1713 | 4.1308 |
| 0.002 | 3.2939 | 1.8213 | 0.00001 | 7.7298 | 4.5291 |

CV, critical value.

### 2.3 GSEA

In the original GSEA Subramaniana *et al.*, 2005, a set of genes are first ranked by $P$-value or FC, then a Kolmogorov–Smirnov like test is performed to evaluate the significance of differential expression as a whole set. Later, generalized GSEA methods were developed Ackermann and Strimmer, 2009; Jiang and Gentleman, 2007, such as in Tian *et al.* 2005, where a $t$-statistic based enrichment score (ES) is suggested for a set of $n$ genes,

$$ES_t = \frac{1}{n}\sum_{i=1}^{N} I_i t_i, \tag{4}$$

where $N$ is the total number of genes, and $I_i$ is an indicator function, with $n = \sum_{i=1}^{N} I_i$. Similarly, we can use $P$-value for enrichment score. To follow the convention that a larger score indicates higher enrichment, we can define

$$ES_p = \frac{1}{n}\sum_{i=1}^{N} I_i(-\log_{10} p_i). \tag{5}$$

A discussion of $ES_t$ and $ES_p$ is provided in Section 1.3 in Supplementary Materials.

$\pi$-value can be employed in GSEA in two ways: the first is to use the original GSEA method, but rank genes by $\pi$-value instead of $P$-value or FC; and the second is to use generalized GSEA by defining

$$ES_\pi = \frac{1}{n}\sum_{i=1}^{N} I_i \pi_i = \frac{1}{n}\sum_{i=1}^{N} I_i \phi_i(-\log_{10} p_i). \tag{6}$$

Whichever GSEA method we adopt, $\pi$-value-based algorithm has the advantage of combining FC and $P$-value information.

In generalized GSEA, a gene set is considered enriched if the statistical significance ($P$-value) of its enrichment score is below a threshold, and a method to compute $P$-value of ES is provided in Algorithm S4 of the Supplementary Materials.

### 2.4 Data resampling methods

Given a gene expression dataset, which usually contains both non-DE and DE genes, it is useful to extract or generate a dataset that satisfies null hypothesis, called *background data* or *null data*. Null data can be used for obtaining the empirical distribution and critical values of $\pi$-value, for instance. We can also impose differential expression on pre-chosen genes on the null data, and regard the pre-chosen DE genes as the *ground truth*. This technique is useful in evaluating the performance of gene selection criterion where true DE genes must be known in order to compute sensitivity and FDR. Two algorithms are developed for generating background data from a real dataset. The first algorithm, maResampling, is based on resampling and suitable for large unbalanced samples. The second one, maBootstrapping, is a parameterized bootstrapping scheme for small or large balanced samples. They are listed as Algorithms S2 and S3 in the Supplementary Materials.

### 2.5 Datasets

Dataset 1 (Breast Cancer Data): a total of 286 lymph node-negative primary breast cancer samples (GSE2034) described in Wang *et al.* 2005 were downloaded from GEO (http://www.ncbi.nih.gov/geo), and more details are provided in the Supplementary Materials. To derive a null dataset from it, we used the resampling method in Algorithm S2 in Supplementary Materials, $(G_1, G_2)$ = maResampling $(G, 50, 50)$. To add DE genes, we randomly chose 1114 (5%) genes and make them DE by adding $\Delta\mu_i$ to randomly selected 557 genes in data matrix $G_1$, and adding $\Delta\mu_i$ to the remaining 557 genes in $G_2$. Each $\Delta\mu_i$ is uniformly distributed in [0.5, 4].

Dataset 2 (Burn Injury Data): in a recent study on burn injured rats, we have obtained 68 gene expression profiles, from which a data matrix $G$ consisting of five replicate arrays is taken for further manipulation (Section 1.5 in Supplementary Materials). To generate a null dataset from $G$, we adopted Algorithm S3 in Supplementary Materials and used function $(G_1, G_2)$ = maBootstrapping $(G, 5, 5)$. Adding DE genes to the null data is similar to Dataset 1.

## 3 RESULTS

### 3.1 Identification of DE genes

To test $\pi$-value in the identification of DE genes, we use the Breast Cancer Data derived null dataset ($M_1, M_2$ = 50) described before, with 1114 (5%) artificially designed DE genes as the known truth, whose mean differential expression levels are uniformly distributed in [0.2, 2]. We use receiver operating characteristic (ROC) curve as performance measure, and compare the following gene selection criteria. The moderated $t$-test, shrinkage $t$-test and TREAT are available in R package. The ROC curves are obtained from the average result of 100 times of simulations.

(1) $\pi$-value: $\pi$-value is computed from $P$-value of regular $t$-test.
(2) LFC: Log-fold change criterion;
(3) RegPval: $P$-value criterion using regular $t$-test;
(4) ModPval: $P$-value criterion using moderated $t$-test;
(5) TREAT: $P$-value criterion using TREAT with LFC threshold $\tau = 0.5$; and
(6) ShrPval: $P$-value criterion using shrinkage $t$-test.

Figure 2a shows that $\pi$-value criterion generally outperforms other criteria with the largest area under curve (AUC). The $\pi$-value

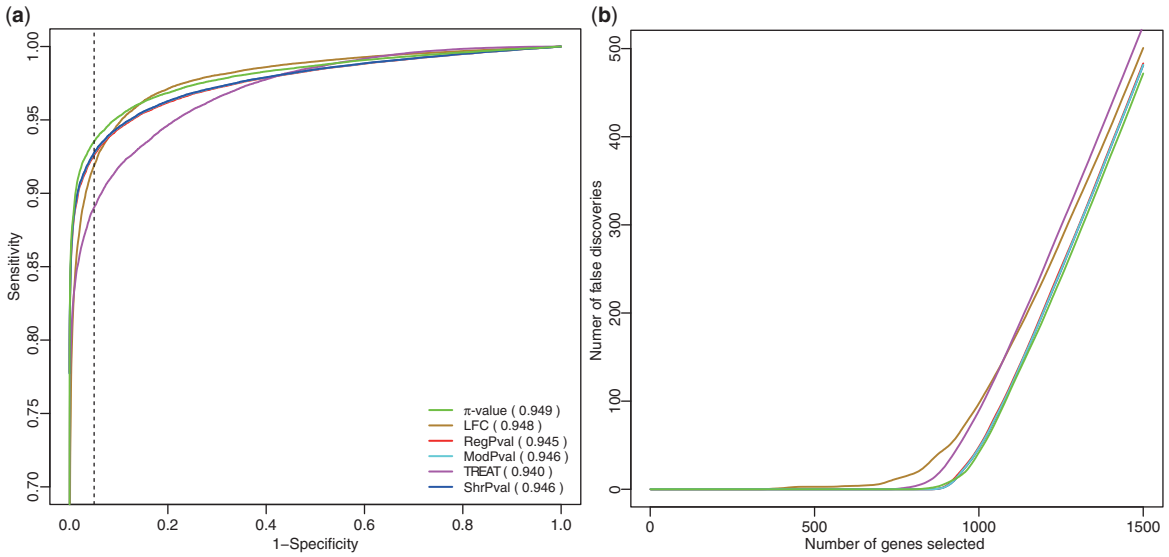


**Fig. 2.** ROC curves and FDR for various gene selection criteria. (**a**) ROC curves. AUC for each ROC curve is provided in the parentheses. DE genes have $\Delta\mu_i$ uniformly distributed in [0.2, 2]. (**b**) Number of false discoveries among top $n$ selected genes.

criterion performs best in the high specificity region. For example, when specificity equals 0.95, indicated by a dashed line in Figure 2a, $\pi$-value has the highest sensitivity, 0.94. Regular $t$-test, moderated $t$-test and shrinkage-$t$ are less sensitive, with FC criterion and TREAT being the least sensitive. The number of false discoveries among the top $n$ selected genes for each criterion is shown in Figure 2b. $\pi$-value, regular $t$-test, moderated-$t$ and shrinkage-$t$ have similar performance and have fewer false discoveries than FC and TREAT. When $n<920$ and $n\geq 1000$, $\pi$-value has the lowest FDR among all criteria. The robustness of $\pi$-value criterion is further demonstrated in Supplementary Figure S2 when we reduce the synthetic DE genes' differential expression $\Delta\mu_i$ to the range of [0.1, 1]. $\pi$-value still has higher sensitivity in the high specificity region [0.8, 1] (left of the dashed line, Supplementary Fig. S2a and the lowest FDR Supplementary Fig. S2b). We have also evaluated the capability of $\pi$-value through an additional simulated dataset and an empirical dataset (Section 2.2 in Supplementary Materials). In simulations adopted from a hierarchical model Smyth, 2004, $\pi$-value has comparable performance as moderated-$t$, TREAT and shrinkage-$t$ in similar and balanced variances conditions. Simulation results on an empirical dataset again demonstrate that $\pi$-value has comparable or better performances. It is especially noteworthy that (in the hierarchical model simulations), although FC and regular $t$-test do not perform robustly in the condition of similar and balanced variances, $\pi$-value, which integrated the two methods, has good FDR and AUC values competitive to other more complex methods (moderated-$t$, TREAT and shrinkage-$t$) which need extra prior assumption and complicated computations. All the above results indicate that $\pi$-value is a desirable criterion for DE gene selection.

## 3.2 Application to GSEA

$\pi$-value provides a new ranking method for genes based on combination of expression change and statistical significance, and it can be readily applied to GSEA.

We use Burn Injury Data derived dataset to demonstrate the application of $\pi$-value to generalized GSEA [Equation (6)], and the procedure is summarized in the Supplementary Materials. $t$-statistic, $P$-value and $\pi$-value based enrichment scores are denoted by $ES_t, ES_p$ and $ES_\pi$, respectively.

*$\pi$-based enrichment score is less sensitive to low FC but equally sensitive when LFC $\geq$0.4.* To study the sensitivity of enrichment scores with respect to LFCs of genes in GSEA, we designed the following experiment. We created several gene sets, each containing 20 genes. In each gene set, 10 genes have different mean expressions in class $C_1$ and $C_2$, with $g_{(1)ij}\sim\mathcal{N}(\mu_i,\sigma_i)$, $g_{(2)ij}\sim\mathcal{N}(\mu_i+\Delta\mu,\sigma_i)$, and $\Delta\mu$ varies from 0.0 to 1.0 across the gene sets. The detection rates of $t$-, $P$- and $\pi$-based enrichment scores with respect to $\Delta\mu$ are shown in Figure 3a. As expected, the higher the LFC $\Delta\mu$ is, the more sensitive a enrichment score is to detect enrichment. $ES_\pi$ is less sensitive than $ES_t$ and $ES_p$ when $\Delta\mu<0.4$; however, it is equally sensitive when $\Delta\mu\geq 0.4$, and all three enrichment scores achieve near 100% detection rate. Finally, it is noted that when $\Delta\mu$ is near 0, the gene set is probably not truly enriched, a low detection rate is thus desirable. Figure 3 shows $ES_\pi$ has the lowest detection rate when $\Delta\mu\sim 0$, indicating the lowest FDR.

*$\pi$-based enrichment score is robust with respect to intrinsic variance.* To study how gene variance affect the detection of enriched gene sets, we designed genes sets similar to the previous experiment, but with fixed differential expression and controlled minimum variance. In particular, we fixed $\Delta\mu$ to be 0.5 to guarantee the high confidence of enrichment, and let the DE genes have $g_{(1)ij}\sim\mathcal{N}(\mu_i,\min(\sigma_i,\sigma_{\min})$ and $g_{(2)ij}\sim\mathcal{N}(\mu_i+\Delta\mu,\min(\sigma_i,\sigma_{\min})$. $\sigma_{\min}$ varied from 0.0 to 1.5 across the gene sets. As expected and shown in Figure 3b, when $\sigma_{\min}=0$ (no minimum gene variance is imposed), all three enrichment scores have near perfect detection rate. However, when $\sigma_{\min}$ gradually increases to 1.5, $ES_\pi$ can still maintain about 90% detection rate, while the detection rate of $ES_t$ drops to <60%, and $ES_p$ has the worst detection rate, <20%.

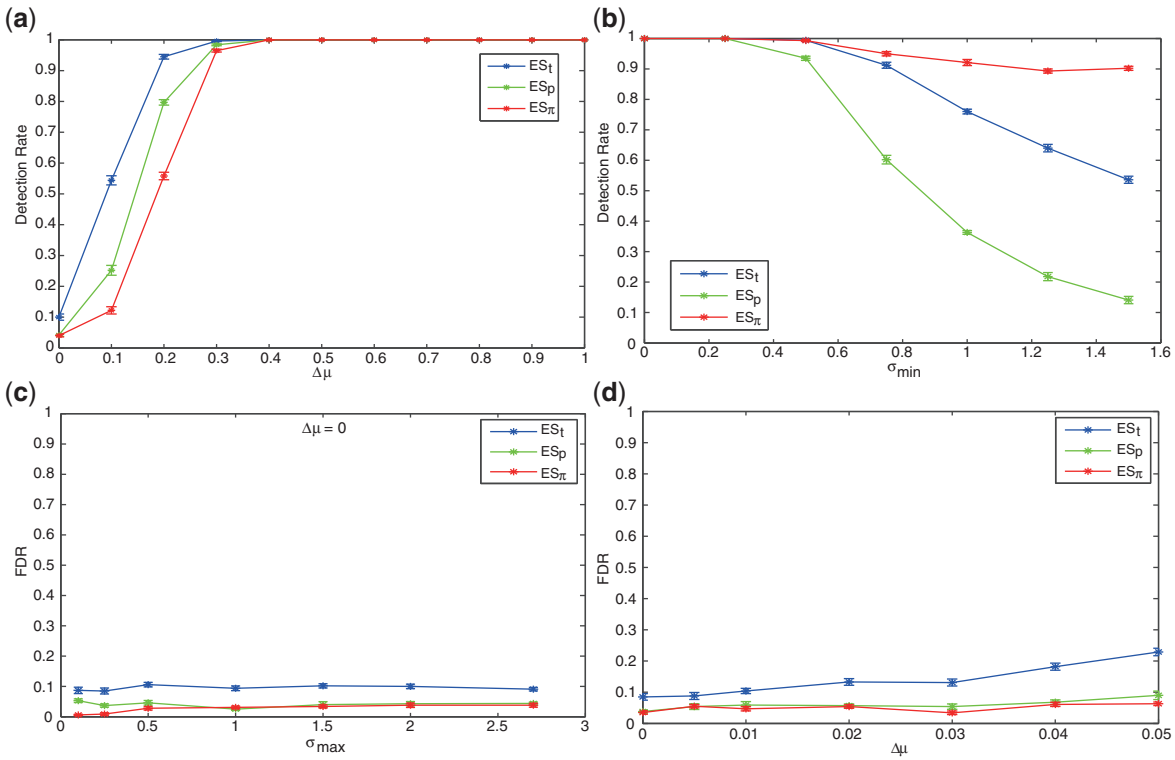


**Fig. 3.** Top figures: detection rates with respect to (**a**) varied LFC and (**b**) varied minimum variance. In (a), $x$-axis denotes LFC $\Delta\mu$; in (b), $x$-axis denotes the minimum SD $\sigma_{\min}$, with $\Delta\mu = 0.5$ fixed. Bottom figures: false discovery rates (FDRs) with respect to (**c**) zero-fold change and varied maximum variance, and (**d**) slight FC. In (c), $\Delta\mu = 0$, and $x$-axis denotes the upper limit of SD $\sigma_{\max}$; in (d), $\Delta\mu$ varies from 0.0 to 0.05, and $\sigma_{\max}$ is not restricted.

$\pi$-based enrichment score protects against false discovery. The FDR can be assessed when $\Delta\mu = 0$ or is very small (Fig. 3c and d). Figure 3c shows the result of an experiment where we designed several gene sets with no FC and varied maximum variance (similar to previous experiments, we controlled $\sigma_{\max}$ of 10 selected genes in each gene set). The FDR for $ES_\pi$ is consistently <5%, almost always smaller than $ES_t$ and $ES_p$, and $ES_t$ has the largest FDR. Since $\Delta\mu \to 0$ constitutes a heavy penalty on $\pi$-value, a highest degree of protection against FDR is observed when $\sigma_{\max}$ is smaller ($\leq 0.5$), such that smaller variance is correlated with larger $t$ and smaller $P$-value, leading to higher $ES_t$ and $ES_p$. The protection of $ES_\pi$ against FDR can be further extended when FC is non-zero but too small for the gene set to be considered enriched. In an extended experiment, we let $\Delta\mu$ vary from 0 to 0.05 without variance control, and observed again that $ES_\pi$ has a consistent lower FDR than $ES_t$ and $ES_p$ (Fig. 3).

### 3.3 Application to gene expression profiling of estrogen receptor sensitive breast cancer

In the previous sections, we based our simulations on controlled datasets where differential expressions were artificially added to null dataset (satisfying null hypothesis), for convenient evaluation of performance. Now we will apply $\pi$-value to the original breast cancer dataset.

The breast cancer dataset (GSE2034) contains 77 estrogen receptor negative (ER−) and 209 estrogen receptor positive (ER+) samples of breast tumor, and we use it as the reference to compare three enrichment scores, $ES_t, ES_p$ and $ES_\pi$. It is well documented that ER+ and ER− breast cancer patients respond to drugs differently and have distinct prognosis, and it is shown that estrogen regulates pathway in breast cancers Lewis-Wambi and Jordan, 2009. Therefore, we choose an estrogen-related Gene Ontology term 'response to estrogen stimulus' (GO:0043627) to form our interested gene set. Using 1091 biological process terms (each term corresponding to a gene set) recorded in the Gene Ontology database, we have conducted a generalized GSEA using the method outlined in Section 2.4 in Supplementary Materials. Results in Table 2 (top 3 rows) show that our interested gene set (estrogen stimulus) is highly enriched with high rankings among 1091 gene sets for all three enrichment scores $ES_t, ES_p$ and $ES_\pi$. The rankings ($\text{rank}_{\text{ref}}$) of the gene set among all gene sets using $ES_t, ES_p$ and $ES_\pi$ are 67, 47 and 27, respectively. The $P$-values ($p_{\text{ref}}$) of $ES_t, ES_p$ and $ES_\pi$ are 0.003, 0.002 and 0.001, respectively.

In order to test the performance of different enrichment scores in small sample setting, we used the original 283 samples as a pool, and randomly drew a subsample consisting of 27 ER+ and 10 ER− profiles 1000 times, to which we applied GSEA based on the three scores. The results are listed in Table 2 (bottom 3 rows). Among three enrichment scores, $ES_\pi$ provides the highest average ranking ($\text{rank}_{\text{avg}} = 58$) of the interested gene set and also has the most significant average $P$-value ($p_{\text{avg}}$). Using $P < 0.05$ as an enrichment criterion, $ES_\pi$ has the highest enrichment rate in 1000

**Table 2.** Results of GSEA

| Estrogen stimulus | $ES_t$ | $ES_p$ | $ES_\pi$ |
|---|---|---|---|
| $\text{rank}_{\text{ref}}$ | 67 | 47 | 27 |
| $p_{\text{ref}}$ | 0.003 | 0.002 | 0.001 |
| $\text{rank}_{\text{avg}}$ | 138 | 130 | 58 |
| $p_{\text{avg}}$ | 0.03 | 0.03 | 0.007 |
| Enrichment rate (%) | 60.0 | 60.8 | 94.3 |

Rank refers to the ranking of a gene set's ES score in descending order. Average $P$-value of ES is obtained from 1000 times of simulations. Enrichment rate is determined by the percentage of the gene set satisfying the criterion $P$-value of ES $\leq 0.05$, in 1000 times of simulations.

times of simulations. These results show that $\pi$-value has desirable performance in GSEA, especially in small sample setting.

## 4 DISCUSSION

We defined $\pi$-value as a gene significance score combining FC and $P$-value, and derived its distribution. To identify DE genes by $\pi$-value, we can specify a significance level $\alpha$ and use the corresponding critical value as threshold. Under certain assumptions, critical values can be obtained theoretically; otherwise, they can be estimated from simulation, as shown in the Supplementary Materials.

One may choose *ad hoc* $\pi$ threshold of 1.3 or 2.0, noting that at 2-fold change, $P=0.05, 0.01$ translate to $\pi=1.3, 2.0$, and the thresholds are also close to the dependent CVs at $\alpha=0.05, 0.01$ (Table 1). Moreover, the definition of $\pi$-value in Equation (1) implies that 2-fold change is a neutral position: a smaller FC will penalize $P$-value, and a larger FC will boost $P$-value. As an extension to this concept, if we consider $n$-FC as a reasonable decision boundary for DE genes, we can adopt an alternative definition of $\pi = (\phi/\log_2 n)\cdot(-\log_{10} p)$, and choose decision threshold accordingly. Further comparison of $\pi$-value, $P$-value and FC to serve as DE gene selection criteria is discussed in the Supplementary Materials.

Among various selection criteria for identifying DE genes, $\pi$-value is the most robust. We have found that

- $\pi$-value performs well at the high specificity region, and its ROC curve stays above all other criteria in the region. $\pi$-value also has the lowest FDR when the number of selected genes increases.
- Overall, $\pi$ outperforms other criteria by having the largest AUC (area under curve) of ROC.

In the simulations, $\pi$-value was computed based on regular $t$-test, therefore its computational complexity is less than moderated $t$-test, TREAT, or some other improvements to the regular $t$-test.

In generalized GSEA, $\pi$-based enrichment score is also more robust than $t$-statistic and $P$-value based scores. In the experiments, we designed half of the genes in a gene set to be DE, to mimic the real world scenario. We have found that

- $ES_\pi$ has comparable detection rate of enriched gene sets as $ES_t$ and $ES_p$ when differential gene expression $\Delta\mu \geq 0.4$, although it is less sensitive when $\Delta\mu < 0.4$.
- $ES_\pi$ has a robust high detection rate when DE genes in an enriched gene set have increased variance.
- $ES_\pi$ protects against false discovery when $\Delta\mu$ is close to zero, especially when part of the genes have low variance.

In the application to breast cancer, we utilized various scoring methods of GSEA to evaluate the performance on small subsets of ER+ and ER− breast tumor samples. Based on 1000 times of simulations, we have found that $\pi$-value based enrichment score provides higher ranks of the estrogen-related GO term, higher enrichment rate and smaller $P$-values. The results demonstrate the advantages of $\pi$-value to serve as enrichment score in GSEA applications.

In conclusion, we find that $\pi$-value is a robust score for detecting DE genes, reflected in the ROC curve consisting of sensitivity versus specificity and the false discovery rate plot. In generalized GSEA, $\pi$-value based enrichment score has a comparable performance with $P$-value and $t$-statistic based enrichment scores in general. In special situations, $ES_\pi$ behaves more robustly than $ES_p$ and $ES_t$, in that it improves sensitivity in the LFLV situation, and protects against false discovery in the SFSV situation. When applied to breast cancer data, $\pi$-value also shows its potential in identifying enriched gene sets involved in key biological functions. Most existing microarray analysis infrastructure, including software and public domain databases, provide FC and $P$-value information. FC is an important factor in biological discovery, which represents the variation of mRNA abundance across different biological conditions. $P$-value of $t$-test signifies the confidence of difference between case and control, based on which null hypothesis will be rejected or accepted. By integrating the two factors, $\pi$-value can improve the performance of microarray analysis without using prior information or recalculating a new statistic. Therefore, it provides us an option to improve on the current infrastructure without major changes. Not limited to regular-$t$ test, $\pi$-value can be extended to integrate other kinds of $P$-values, such like moderated $t$-test, ANOVA or linear regression, to satisfy various demands in microarray analysis.

## ACKNOWLEDGMENTS

The authors thank Greehey Children's Cancer Research Institute for its technical support of the high-performance computing facility. We thank Dr Jon Gelfond for constructive criticism of the manuscript.

*Funding*: Combat Casualty Care Division, United States Army Medical Research and Materiel Command; Technologies for Metabolic Monitoring (TMM)/Julia Weaver Fund; National Institutes of Health (R01 GM063120-04) to X.W. and S.E.W. National Institutes of Health/NCI cancer center grant (P30 CA054174-17); Biostatistics and Informatics Shared Resources (BMISR); NIH/NCRR CTSA grant (1UL1RR025767) to Y.C.

*Conflict of Interest*: none declared.

## REFERENCES

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**.

Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci.*, **99**, 6562–6566.

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

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*, **57**, 289–300.

Chen,Y. *et al.* (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.

Cui,X. and Churchill,G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.

Fu,W.J. *et al.* (2006) Statistical models in assessing fold change of gene expression in real-time RT-PCR experiments. *Comput. Biol. Chem.*, **30**, 21–26.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Huang,H. *et al.* (2010) Discovering disease-specific biomarker genes for cancer diagnosis and prognosis. *Technol. Cancer Res. Treat.*, **9**, 219–229.

Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

Lewis-Wambi,J.S. and Jordan,V.C. (2009) Estrogen regulation of apoptosis: how can one hormone stimulate and inhibit? *Breast Cancer Res.*, **11**, 206.

Mariani,T.J. *et al.* (2003) A variable fold change threshold determines significance for expression microarrays. *FASEB J.*, **17**, 321–323.

McCarthy,D.J. and Smyth,G.K. (2009) Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, **25**, 765–771.

Montazeri,Z. *et al.* (2010) Shrinkage estimation of effect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: applications to differential gene expression. *Stat. Appl. Genet. Mol. Biol.*, **9**, Article 23.

Mutch,D. *et al.* (2002) The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, **3**, 17.

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.

Subramaniana,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tian,L. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

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

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Wu,X. *et al.* (2010) Effect of high dose insulin treatment on skeletal muscle gene expression after severe burn. *J. Burn Care Res.*, **31** (Suppl. 2), S48.

Yang,I.V. *et al.* (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, **3**, research0062.

Yanofsky,C.M. and Bickel,D.R. (2010) Validation of differential gene expression algorithms: application comparing fold-change estimation to hypothesis testing. *BMC Bioinformatics*, **11**, 63.