A novel significance score for gene selection and ranking
Yufei Xiao1,2,† Tzu-Hung Hsiao2, Ulitha Suresh2, Hung-I Harry Chen2, Xiaowu Wu3,4, Steven E. Wolf3,4, and Yidong Chen1,2,5,∗

1Department of Epidemiology and Biostatistics, 2Computational Biology and Bioinformatics Division, Greehey Children’s Cancer Research Institute, 3Department of Surgery, University of Texas Health Science Center at San Antonio, TX 78229, 4United States Army Institute of Surgical Research, Fort Sam Houston, TX 78234 and 5Cancer 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, π-value, by combining expression fold change and statistical significance (P-value), and explored its statistical properties. When compared to various existing methods, π-value based approach is more robust in selecting DE genes, with the largest area under curve in its receiver operating characteristic curve. We applied π-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 π-value helps elucidating otherwise overlooked important biological functions.

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

Contact: py@cherokeehs@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, π-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 DE genes, 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 or Smyth, 2004; Tusher 2001, empirical Bayes method (moderated 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.

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:

1. 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.

2. The ‘large fold change, large variance’ (LFVL) 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, 2000; empirical Bayes method (moderated t-test), 2002; regularized t-test, 2004. 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 < 0.05. Such approach is ad hoc Cui and Churchill, 2003 or Match et al. 2002, or a better statistical model Fu et al. 2004, or...
variable threshold according to intensity measurement [Marion 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 2005; Montazeri et al. 2013]. In McCarthy and Smyth [2005], 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 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. 2014].

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 $M$ and $N$ 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 set probe and gene interchangeably unless specified otherwise. Let $G_{x_k}$ be log-transformed expression level of the $i$-th gene in the $j$-th sample of class $C_2$ ($k = 1, 2$), then the entire processed microarray data can be represented in two data matrices, $G_1$ and $G_2$, where

$$G_k = \begin{bmatrix} G_{x_{11}}(k) & G_{x_{12}}(k) & \cdots & G_{x_{1M}}(k) \\ G_{x_{21}}(k) & G_{x_{22}}(k) & \cdots & G_{x_{2M}}(k) \\ \vdots & \vdots & \ddots & \vdots \\ G_{x_{M1}}(k) & G_{x_{M2}}(k) & \cdots & G_{x_{MN}}(k) \end{bmatrix}, \quad k = 1, 2.$$ 

Let the mean expression of the $i$-th gene in sample class $C_k$ be $\bar{G}_{x_k}(i)$, then the log-ratio (LR) and log-fold change (LFC) of this gene’s expression are denoted as $s_i = \bar{G}_{x_1}(i) - \bar{G}_{x_2}(i)$ and $\phi_i = \log_2(\bar{G}_{x_1}(i) / \bar{G}_{x_2}(i))$, respectively. In the literature, FC refers to the quantity $2^\phi$, and an alternative definition is $2^\sqrt{\pi x e}$. 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 an $t$-test or its variants is applied, and when multiple genes are involved, $P$-values are often adjusted to account for the effect of multiple testing, 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 2005].

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

$$\pi_i = \phi_i \cdot \log_2(p_i).$$

$\pi$-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

$$\pi_i = \log_2(1 - \exp(-\lambda \phi_i)),$$

thus $\pi = 10^{-\tau}$ can be viewed as transformed $P$-value, and $0 \leq \pi \leq 1$. When $\phi_i$ is less than 2-fold change, we will have $\pi > p_i$, and $\pi$ retains the regular $P$-value. For $\phi_i > 1$, namely more than 2-fold change, we have $\pi < p_i$, and $\pi$ has been 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, we have implemented a significance assessment tool for dependence condition given any input microarray dataset (Supplementary Material).

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),

$$f(\zeta(\pi)) = \frac{\pi}{\sigma^2 \sqrt{2\pi}} \exp\left(-\frac{\pi^2}{2\sigma^2}\right), \quad \zeta(\pi) = 0,$$

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 [null distribution is 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 $a$, we need to determine its critical value with respect to $a$. $P$-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 [null] shows an example of the critical values of $\pi$ with respect to different $a$’s under the independence and dependence assumptions ($\sigma = 0.48$ for the LR), respectively, and the simulation details are in the Supplementary Materials. Table [null] 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).
Table 1. CV of π under dependence or independence

<table>
<thead>
<tr>
<th>α</th>
<th>CV_{dep}</th>
<th>CV_{ind}</th>
<th>α</th>
<th>CV_{dep}</th>
<th>CV_{ind}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>3.8434</td>
<td>2.1281</td>
<td>0.002</td>
<td>3.4978</td>
<td>2.0580</td>
</tr>
<tr>
<td>0.01</td>
<td>3.6233</td>
<td>2.1408</td>
<td>0.04</td>
<td>2.8988</td>
<td>2.2088</td>
</tr>
<tr>
<td>0.05</td>
<td>3.2543</td>
<td>2.5234</td>
<td>0.01</td>
<td>3.6407</td>
<td>2.2128</td>
</tr>
<tr>
<td>0.005</td>
<td>3.6407</td>
<td>3.6233</td>
<td>0.0002</td>
<td>3.7298</td>
<td>4.5291</td>
</tr>
</tbody>
</table>

2.3 GSEA

In the original GSEA, a set of genes is 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 to evaluate the significance of differential expression as a whole set. Later, generalized GSEA methods were developed to evaluate the significance of differential expression as a whole set.

In general, we adopt GSEA method, as in Tian et al. 2005, such as in Tian et al. 2005, Ackermann and Strimmer 2009, where a t-statistic based enrichment score (ES) is suggested for a set of n genes,

\[ E_{S} = \sum_{i=1}^{N} I_{i}, \]  \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

\[ E_{S} = \sum_{i=1}^{N} I_{i} (log_{10} P_{i}). \]  \tag{5}

A discussion of \( E_{S} \) and \( E_{S_{p}} \) 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

\[ E_{S_{p}} = \sum_{i=1}^{N} I_{i} (log_{10} P_{i}). \]  \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 criteria 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, maltesampling, is based on resampling and suitable for large unbalanced samples. The second one, maltoolbootstrapping, 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, \((G_1, G_2) = \text{maltesampling} (G, 50, 50)\). To add DE genes, we randomly chose 1114 (5%) genes and make them DE by adding \( \Delta \mu \) to randomly selected 557 genes in data matrix \( G_1 \), and adding \( \Delta \mu \) to the remaining 557 genes in \( G_2 \). Each \( \Delta \mu \) 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 \((G_1, G_2) = \text{maooltoolbootstrapping} (G, 5, 5)\). 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 (M1, M2 = 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 \( t = 0.5 \);

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

\[ \text{Fig. 1. PDF of } \pi \text{ when LFC and } P \text{-value are dependent (blue) or independent (magenta).} \]  

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer Data</td>
<td>A total of 286 lymph node-negative primary breast cancer samples (GSE2034) described in Wang et al. 2005.</td>
</tr>
<tr>
<td>Burn Injury Data</td>
<td>68 gene expression profiles from a recent study on burn injured rats.</td>
</tr>
</tbody>
</table>

\[ \text{Table 1. CV of } \pi \text{ under dependence or independence} \]  

<table>
<thead>
<tr>
<th>α</th>
<th>CV_{dep}</th>
<th>CV_{ind}</th>
<th>α</th>
<th>CV_{dep}</th>
<th>CV_{ind}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>3.8434</td>
<td>2.1281</td>
<td>0.002</td>
<td>3.4978</td>
<td>2.0580</td>
</tr>
<tr>
<td>0.01</td>
<td>3.6233</td>
<td>2.1408</td>
<td>0.04</td>
<td>2.8988</td>
<td>2.2088</td>
</tr>
<tr>
<td>0.05</td>
<td>3.2543</td>
<td>2.5234</td>
<td>0.01</td>
<td>3.6407</td>
<td>2.2128</td>
</tr>
<tr>
<td>0.005</td>
<td>3.6407</td>
<td>3.6233</td>
<td>0.0002</td>
<td>3.7298</td>
<td>4.5291</td>
</tr>
</tbody>
</table>

\[ \text{CV, critical value.} \]
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-\(t\) 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’ \(\Delta\mu_j\) uniformly distributed in \([0, 2]\) (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 again demonstrate that \(\pi\)-value and \(\pi\)-value based enrichment scores are denoted by \(E_{S\pi}, E_{S\pi}^p\) and \(ES_T\), 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 \(C_j\) and \(C_{j'}\), with \(\mu_{ij} \sim N(\mu_j, \sigma_j), \mu_{ij'} \sim N(\mu_j + \Delta\mu, \sigma_j),\) and \(\Delta\mu\) varies from 0.0 to 1.0 across the gene sets. The detection rates of \(t\)-, \(\pi\)- 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 an enrichment score is to detect enrichment. \(ES_T\) is less sensitive than \(E_{S\pi}\) and \(ES_T^p\) 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 \(\Delta\mu = 0\), the gene set is probably not truly enriched, a low detection rate is thus desirable. Figure 3 shows \(ES_T\) 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 \(\mu_{ij} \sim N(\mu_j, \min\sigma_j, \sigma_{\min})\) and \(\mu_{ij'} \sim N(\mu_j + \Delta\mu, \min\sigma_j, \sigma_{\min})\). \(\sigma_{\min}\) varied from 0.0 to 1.5 across the gene sets. As expected and shown in Figure 3b, when \(\sigma_{\min} \sim 0\) (no minimum gene variance is imposed), all three enrichment scores have near perfect detection rate. However, when \(\sigma_{\min}\) gradually increases to 1.5, \(ES_T\) can still maintain about 90% detection rate, while the detection rate of \(ES_T^p\) drops to <60%, and \(ES_T^p\) has the worst detection rate, <20%.

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 \(\Delta\mu_j\) uniformly distributed in \([0, 2]\). (b) Number of false discoveries among top \(n\) selected genes.

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 5), and the procedure is summarized in the Supplementary Materials. \(t\)-statistic, \(P\)-value and \(\pi\)-value based enrichment scores are denoted by \(E_T, E_{\pi}^p\) and \(ES_T\), 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 \(C_j\) and \(C_{j'}\), with \(\mu_{ij} \sim N(\mu_j, \sigma_j), \mu_{ij'} \sim N(\mu_j + \Delta\mu, \sigma_j),\) and \(\Delta\mu\) varies from 0.0 to 1.0 across the gene sets. The detection rates of \(t\)-, \(\pi\)- 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 an enrichment score is to detect enrichment. \(ES_T\) is less sensitive than \(E_{\pi}^p\) and \(ES_T^p\) 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 \(\Delta\mu = 0\), the gene set is probably not truly enriched, a low detection rate is thus desirable. Figure 3 shows \(ES_T\) 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 \(\mu_{ij} \sim N(\mu_j, \min\sigma_j, \sigma_{\min})\) and \(\mu_{ij'} \sim N(\mu_j + \Delta\mu, \min\sigma_j, \sigma_{\min})\). \(\sigma_{\min}\) varied from 0.0 to 1.5 across the gene sets. As expected and shown in Figure 3b, when \(\sigma_{\min} \sim 0\) (no minimum gene variance is imposed), all three enrichment scores have near perfect detection rate. However, when \(\sigma_{\min}\) gradually increases to 1.5, \(ES_T\) can still maintain about 90% detection rate, while the detection rate of \(ES_T^p\) drops to <60%, and \(ES_T^p\) has the worst detection rate, <20%.

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.
A novel gene significance score

Fig. 3. Top figures: detection rates with respect to (a) varied LFC and (b) varied minimum variance. In (a), x-axis denotes \( \Delta \mu \), in (b), x-axis denotes the upper limit of SD \( \sigma_{\text{max}} \); in (d), \( \Delta \mu \) varies from 0 to 0.05, and \( \sigma_{\text{max}} \) 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. 3a 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 \( \sigma_{\text{max}} \) of 10 selected genes in each gene set). The FDR for \( E_{\text{St}} \) is consistently <5%, almost always smaller than \( E_{\text{Sp}} \) and \( E_{\text{St}} \), and \( E_{\text{Sp}} \) 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 \( \sigma_{\text{max}} \) is smaller \((\leq 0.5)\), such that smaller variance is correlated with larger \( \tau \) and smaller \( P \)-value, leading to higher \( E_{\text{St}} \) and \( E_{\text{Sp}} \). The protection of \( E_{\text{St}} \) 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 \( E_{\text{St}} \) has a consistent lower FDR than \( E_{\text{St}} \) and \( E_{\text{Sp}} \) (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, \( E_{\text{St}} \), \( E_{\text{Sp}} \), and \( E_{\text{St}} \). 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 \( E_{\text{St}} \), \( E_{\text{Sp}} \), and \( E_{\text{St}} \). The rankings \((\text{rank}_{\text{avg}})\) of the gene set among all gene sets using \( E_{\text{St}} \), \( E_{\text{Sp}} \) and \( E_{\text{St}} \) are 67, 47 and 27, respectively. The \( P \)-values \((\text{rank}_{\text{avg}})\) of \( E_{\text{St}} \), \( E_{\text{Sp}} \) and \( E_{\text{St}} \) 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, \( E_{\text{Sp}} \) provides the highest average ranking \((\text{rank}_{\text{avg}} = 58)\) of the interested gene set and also has the most significant average \( P \)-value \((\text{rank}_{\text{avg}})\). Using \( P < 0.05 \) as an enrichment criterion, \( E_{\text{Sp}} \) has the highest enrichment rate in 1000
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 \textit{ad hoc} \( \pi \) threshold of 1.3 or 2.0, noting that at 2-fold change, \( P = 0.05 \) translate to \( \pi = 1.3, 2.0 \), and the thresholds are also close to the dependent CVs at \( \alpha = 0.05 \), 0.01 (Table I). Moreover, the definition of \( \pi \)-value in Equation (I) implies that 2-fold change is a neutral position: a smaller FC will imply that \( 2 \)-fold change is a neutral position: a smaller FC will imply 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 \( \tau \)-statistic based enrichment scores in general. In special situations, \( E_{\text{st}} \) behaves more robustly than \( E_{\text{st}} \) and \( E_{\pi} \) 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.

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 \( \tau \)-statistic based enrichment scores in general. In special situations, \( E_{\text{st}} \) behaves more robustly than \( E_{\text{st}} \) and \( E_{\pi} \) 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.

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.

\[ \text{Conflict of Interest: none declared.} \]

References


---

**Table 2. Results of GSEA**

<table>
<thead>
<tr>
<th>Estrogen stimulus</th>
<th>( E_{\text{st}} )</th>
<th>( E_{\pi} )</th>
<th>( E_{\text{st}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank&lt;sub&gt;ref&lt;/sub&gt;</td>
<td>67</td>
<td>47</td>
<td>27</td>
</tr>
<tr>
<td>( P_{\text{ref}} )</td>
<td>0.003</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>rank&lt;sub&gt;avg&lt;/sub&gt;</td>
<td>138</td>
<td>130</td>
<td>58</td>
</tr>
<tr>
<td>( P_{\text{avg}} )</td>
<td>0.03</td>
<td>0.03</td>
<td>0.007</td>
</tr>
<tr>
<td>Enrichment rate (%)</td>
<td>60.0</td>
<td>60.8</td>
<td>94.3</td>
</tr>
</tbody>
</table>

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.
A novel gene significance score


