Gene expression

Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes

Min Zhang1,†, Lin Zhang2,†, Jinfeng Zou1, Chen Yao2, Hui Xiao1, Qing Liu1, Jing Wang1, Dong Wang1, Chuangang Wang1 and Zheng Guo1,2,∗

1College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China and
2Bioinformatics Centre and School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

Received on October 13, 2008; revised and accepted on April 28, 2009
Advance Access publication May 5, 2009
Associate Editor: Limsoon Wong

ABSTRACT

Motivation: According to current consistency metrics such as percentage of overlapping genes (POG), lists of differentially expressed genes (DEGs) detected from different microarray studies for a complex disease are often highly inconsistent. This irreproducibility problem also exists in other high-throughput post-genomic areas such as proteomics and metabolomics. A complex disease is often characterized with many coordinated molecular changes, which should be considered when evaluating the reproducibility of discovery lists from different studies.

Results: We proposed metrics percentage of overlapping genes-related (POGR) and normalized POGR (nPOGR) to evaluate the consistency between to DEG lists for a complex disease, considering correlated molecular changes rather than only counting gene overlaps between the lists. Based on microarray datasets of three diseases, we showed that though the POG scores for DEG lists from different studies for each disease are extremely low, the POGR and nPOGR scores can be rather high, suggesting that the apparently inconsistent DEG lists may be highly reproducible in the sense that they are actually significantly correlated. Observing different discovery results for a disease by the POG and nPOGR scores will evidently reduce the uncertainty of the microarray studies. The proposed metrics could also be applicable in many other high-throughput post-genomic areas.

Contact: guozhe@ems.hrbmu.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The reproducibility problem is one of the fundamental issues in validation of high-throughput biological discoveries (Marshall, 2004; Ramshoff, 2004; Ramshoff, 2005) because it is often considered that ‘a study that cannot be reliably reproduced has little value’ (Klebanov et al., 2007). However, lists of differentially expressed genes (DEGs) detected from different microarray studies for a disease are often highly inconsistent (Ein-Dor et al., 2005; Michiels et al., 2005; Tan et al., 2003), which frequently raises doubts about the reliability of microarrays (Ein-Dor et al., 2005; Frantz, 2005; Miklos and Maleksa, 2004). It has been suggested that many factors such as large biological variations, small sample sizes (Ein-Dor et al., 2005; Zhang et al., 2008) and improper statistical methods (Klebanov et al., 2007; Shi et al., 2006; Tong et al., 2006) may produce inconsistent DEG lists for a disease. Recently, we demonstrated that in the presence of small technical noise and wide differential expressions in a cancer, it is still highly likely that the observed percentage of overlapping genes (POG) of DEG lists from two studies is low, although each DEG list may comprise mostly true DEGs (Zhang et al., 2008). Up to now, the concept of reproducibility of DEG lists for a disease are often intuitively defined, using terms such as concordance (Shi et al., 2006), agreement (Shi et al., 2006), stability (Ein-Dor et al., 2006; Qiu et al., 2006) or rediscovery rate (Xu and Li, 2003). Most metrics for evaluating the consistency between two gene lists, including the frequently used POG (Ein-Dor et al., 2006; Irizarry et al., 2005; Shi et al., 2006), rely on counting gene overlaps and treat genes as if they would be completely independent with each other. However, a complex disease is often characterized with many coordinated molecular changes (Klebanov et al., 2006; Qiu et al., 2005), which should be taken into account when evaluating whether different studies have actually detected the same or similar results for a disease.

In this article, considering the correlated molecular changes in a complex disease, we propose a novel metric termed percentage of overlapping genes-related (POGR) and its normalized formation nPOGR for evaluating consistency between two DEG lists for the disease. Obviously, to be utilized as a disease marker, a DEG should have a steady expression pattern in the disease. Thus, we also consider genes’ upregulated or downregulated status when evaluating their reproducibility. Based on microarray datasets for three diseases, we show that though there are few genes shared between the DEG lists from different studies for a disease, most of the POGR and nPOGR scores are rather high, suggesting high reproducibility of the DEG lists in the sense that the apparently inconsistent DEG lists are significantly correlated. After normalizing the effects of list lengths and random data correlations in datasets, the nPOGR metric may be more appropriate for comparing the...
consistency levels of DEG lists with different lengths and from different studies.

2 METHODS

2.1 Datasets

For prostate cancer, the cDNA microarray data consist of 62 tumors and 41 normal prostate samples (Singh et al., 2002). A total of 6797 genes were measured by both platforms. For lung cancer, the cDNA microarray data consist of 13 squamous cell lung cancer and five normal lung specimens (Garber et al., 2001), while the data by Affymetrix human U95A oligonucleotide arrays consist of 21 squamous cell lung carcinomas and 17 normal lung specimens (Bhatia et al., 2001). A total of 5530 genes were measured by both platforms. The two datasets for Duchenne muscular dystrophy (DMD) are based on Affymetrix Santa Clara, CA HG-U95AV2 and HG-U133A GeneChip, respectively. One dataset contains 24 samples from 12 DMD patients and 12 unaffected controls (Haslett et al., 2002) and the other consists of 36 samples from 22 DMD patients and 14 controls (Pesatori et al., 2007). A total of 8572 genes were measured by both platforms. For each disease, we only analyze the genes that are present on both platforms.

The cDNA data is log2 transformed and then normalized as median 0 and standard deviation 1 per array, as adopted in Oncomine database (Rohwedder et al., 2007). The ClueLists with missing rates above 20% are deleted, and the other missing values are replaced by using the KNN imputation algorithm (k = 15) (Troyanskaya et al., 2001). The Affymetrix GeneChip data are preprocessed by RMA (Robust Multi-array Analysis). The most recent (July, 2009) SOURCE database (Dethlefsen et al., 2003) is used for annotating ClueLists to GeneIDs.

2.2 DEGs and co-expression gene pairs

DEGs are most commonly defined as genes with non-zero expression difference between two groups. Usually, each gene is tested against the null hypothesis of equal means of expression levels across groups, while the multiple tests are controlled by false discovery rate (FDR) (Benjamini and Hochberg, 1995). However, as defined above, the test of false positives among the claimed DEGs. When using the conventional t-test to select DEGs, small per-gene variances could make genes with small fold-changes statistically significant. To deal with this problem, Tusher et al. (2001) proposed the significance analysis of microarrays (SAM) method to select DEGs by adding a small "fudge factor" to the denominator of the t-test statistic as follows:

\[ t(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s_0 + a} \]  

\[ a(i) = \sqrt{\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2}} \times \left[ \sum (x_1(i) - \bar{x}_1(i))^2 + \sum (x_2(i) - \bar{x}_2(i))^2 \right]^{1/2} \]  

where \( x_1(i) \) and \( x_2(i) \) are the expression vectors of gene \( i \) in \( n_1 \) disease samples and \( n_2 \) control samples, respectively. The "fudge factor" \( a(i) \) is chosen to minimize the coefficient of variance of \( t(i) \).

Because the FDR estimation procedure of SAM might be overly conservative (Xie et al., 2005, Zhang, 2007), we apply the FDR estimation method suggested by (Zhang, 2007) to the SAM program (samr_1.25 R package).

Similarly as Carter et al. (2005) did, in each dataset, we estimate the random correlation distribution by independently permuting the sample labels of each gene and then compute the Pearson correlation coefficients for all the gene pairs. For an observed correlation \( r \) between two genes in the original data, the \( P \)-value is calculated as the percentage of the permuted correlations exceeding \( r \). The FDR control procedure proposed by (Benjamini and Hochberg, 1995) is used to select gene pairs with significantly positive correlation.

2.3 POG and POGR scores

The POG metric is usually used to measure the consistency of gene lists by simply counting overlapped genes between the lists without considering the regulation directions of genes (Chen et al., 2007; Shi et al., 2005). We can evaluate reproducibility of DEG lists in such a broad sense by just counting the overlapped and related genes, as briefly shown in 'Results' section (Section 3.3). However, a DEG for a disease should be regulated in the same direction (up-regulated or down-regulated) in different datasets. Thus, strictly, we define a gene is shared by two gene lists only if it is both overlapped and regulated in the same direction in both datasets. In the following text, we mainly analyze the reproducibility of gene lists with such a strict definition.

Suppose \( k \) genes are shared between list 1 with length \( l_1 \) and list 2 with length \( l_2 \), then the POG score from list 1 to list 2 is \( \text{POG}_{12} = k/l_1 \), and the score from list 2 to list 1 is \( \text{POG}_{21} = k/l_2 \).

Then, considering the correlated molecular changes in a disease, we define the POGR metric to evaluate the consistence between two DEG lists. Let \( O_{12} \) (or \( O_{21} \)) represent the number of genes in list 1 (or list 2) which are not shared but significantly positively correlated with at least one gene in list 2 (or list 1), then the POGR scores are calculated as:

\[ \text{POGR}_{12} = (k + O_{12})/l_1 \quad \text{or} \quad \text{POGR}_{21} = (k + O_{21})/l_2 \]  

Both the POG and POGR score takes values ranging from 0 to 1. Usually, a high POG or POGR score is desired for two discovery lists for a disease. However, both metrics are dependent on the lists lengths (see 'Results' section) and thus lack statistical significance. Also, when expressions of genes are widely correlated in a disease, the POGR score between two gene lists randomly extracted from the original datasets may be high. To partially solve this problem, for an obtained POGR score, we can estimate the probability \( (P \)-value) of observing the score in the original datasets by random chance. The null hypothesis is that the POGR score of two DEG lists (with lengths \( l_1 \) and \( l_2 \)) is the same as the expected score for pairs of gene lists (with lengths \( l_1 \) and \( l_2 \)) randomly extracted from the original datasets. Here, the \( P \)-value is calculated as, among the scores of the 10 000 random pairs of gene lists, the proportion of the scores exceeding the observed one. Similarly to the basic idea of gene set enrichment analysis (GSEA) (Khribanov et al., 2007; Subramanian et al., 2005), which tests whether the members of a gene set are randomly distributed throughout all ranked genes according to correlations of their expressions with the sample labels, we test whether the expression correlations are randomly distributed throughout all genes or primarily among DEGs.

Here, two genes are defined as a correlated pair if and only if they are significantly positively correlated in both datasets. We note that other choices exist. For example, for a DEG list reported in a previous study, it is reasonable to evaluate its reproducibility in a new study by only requiring significant correlation in the new dataset, especially when the original data of the previous study are unavailable. The negatively correlated gene pairs are excluded because we specifically require that the related genes should be regulated in the same direction. However, negatively correlated genes may be also biologically meaningful, which warrants further investigation.

2.4 Normalization of the POG and POGR scores

Because the POG and POGR scores depend on list length, both metrics cannot be applied to compare the consistency levels of gene lists with different lengths. Thus, we propose an approach to normalize the effect of list lengths on POG and POGR scores. Additionally, we also normalize the POGR scores to reduce the effects of general data correlations in datasets.

For two lists (with lengths \( l_1 \) and \( l_2 \)) randomly extracted from two datasets, suppose \( E(k) \) is the expected number of their shared genes, then the POG expectation is \( \text{EP}_{12} = E(k)_1 \) or \( \text{EP}_{21} = E(k)_2 \). Similarly to the kappa metric (Fleiss, 1971), we normalize the POGR score between two DEG lists (with lengths \( l_1 \) and \( l_2 \)) as the proportion of the observed scores beyond the expected ones.
chance to the corresponding maximum potential scores beyond chance as follows:

\[
\begin{align*}
\text{nPOG}_{12} &= \frac{POG_{12} - E(POG_{12})}{1 - E(POG_{12})} \\
\text{nPOG}_{21} &= \frac{POG_{21} - E(POG_{21})}{1 - E(POG_{21})}
\end{align*}
\]

(4)

where \(k\) is the observed number of the shared genes between the two DEG lists. The \(E(POG_{12})\) or \(E(POG_{21})\) can be estimated by the average of the scores for 10,000 pairs of random lists (with lengths \(l_1\) and \(l_2\)).

We normalize the POGR in a similar way. For two lists (with lengths \(l_1\) and \(l_2\)) randomly extracted from two datasets, suppose \(E(O_{12})\) or \(E(O_{21})\) is the expected number of genes in list 1 (or list 2), which are not shared but significantly correlated with at least one gene in list 2 (or list 1), then the POGR score expected by random chance is \(E(POGR_{12}) = [E(k) + E(O_{12})]/[E(k) + E(O_{12})] \times E(O_{12})\). The POGR score can be normalized as:

\[
\begin{align*}
\text{nPOGR}_{12} &= \frac{POGR_{12} - E(POGR_{12})}{1 - E(POGR_{12})} \\
\text{nPOGR}_{21} &= \frac{POGR_{21} - E(POGR_{21})}{1 - E(POGR_{21})}
\end{align*}
\]

(5)

where \(E(O_{12})\) or \(E(O_{21})\) can be estimated as the average number of genes in list 1 (or list 2), which are not shared but significantly positively correlated with genes in list 2 (or list 1) in 10,000 pairs of gene lists (with lengths \(l_1\) and \(l_2\)) randomly extracted from two datasets.

When the POGR score of two DEG lists (with lengths \(l_1\) and \(l_2\)) is less than the expected score for genes of gene lists (with lengths \(l_1\) and \(l_2\)) randomly extracted from the original datasets, the nPOGR score will be negative. In this situation, we set the score as 0. Thus, the nAPoGR score achieves 0 if and only if POGR \(\leq E(POGR)\) \(< 1\). It achieves 1 if and only if POGR = 1 and E(POGR) \(< 1\). Note that the score is invalid when E(POGR) = 1, such as in the extreme situation that all genes are significantly correlated in both datasets for a disease. However, this does not appear in the datasets for the three diseases under this study and would rarely appear in other real applications. Because an nPOGR score is a ratio, it is better to interpret the score together with the maximum potential agreement beyond chance \([1 - E(POGR)]\) if it can achieve for pairs of lists with the same lengths from the original datasets.

We can estimate the significance level (P-value) for an observed nPOGR score of two DEG lists (with lengths \(l_1\) and \(l_2\)) to test the null hypothesis that this score is the same as the expected score for pairs of gene lists (with lengths \(l_1\) and \(l_2\)) randomly extracted from the original datasets. This P-value can be calculated from the same set of 10,000 pairs of random gene lists for calculating the P-value of the corresponding POGR score. According to formula (5), for a pair of genes randomly extracted from the original datasets, its nPOGR score will be larger than the observed nPOGR score if and only if its POGR score is larger than the observed POGR score, since the E(POGR) is a constant estimated from the 10,000 random gene list pairs. Thus, estimated as the proportion of the random nPOGR scores exceeding the observed nPOGR score, the P-value of the observed nPOGR score is the same as the P-value of the corresponding POGR score.

After excluding the effects of list lengths and general data correlations on chance agreement, the nPOGR score may be more appropriate for comparing the consistency levels of DEG lists with different lengths and from different studies. On the other hand, because the nPOGR score measures the relative agreement beyond chance, it should be necessary to use the POGR scores to reflect the apparent agreement of DEG lists based on the original data correlation.

3 RESULTS

3.1 Reproducibility of studies for three diseases

For prostate cancer, by SAM with 1% FDR control, we detect 1054 and 1343 DEGs in the datasets of oligo and cDNA microarrays, respectively. The reproducibility of the two studies is measured in two directions. In one direction, for the list of DEGs detected in the oligo microarray data, we measure its reproducibility in the cDNA microarray data as the percentage of these DEGs, which also appear in the cDNA result. The POGR12 score is as low as 38% and the nPOG12 score is only 30%, intuitively suggesting that most of the results from the oligo microarray data are irreproducible in the cDNA microarray result. Then, considering significantly co-expressed gene pairs with 0.1% FDR control, the POGR21 score achieves 90%, intuitively suggesting that the DEG list from the oligo microarray data may be highly reproducible in the cDNA microarray data when considering correlated molecular changes. The corresponding nPOGR is 74%, indicating that after subtracting the POGR score by random chance, the observed score achieves a high proportion of the maximum potential agreement beyond chance (0.38). In another direction, for the DEG list detected in the cDNA data, we measure its reproducibility in the oligo microarray data as the percentage of the detected DEGs, which also appear in the oligo microarray result.

For the full DEG lists, although the POGR21 score is only 30% and the nPOG21 is as low as 25%, the POGR21 score reaches 90% and the nPOG is as high as 76%, with the maximum potential agreement value as 0.40.

Similar results can be observed for the lists of the top-ranked DEGs from the two studies. For example, although the lists of the top-10 DEGs from the two datasets share only three genes, the POGR12 score is as high as 90%. The corresponding nPOGR is 89%, reaching a high proportion of the maximum potential value of 0.94. In another direction, both the POGR21 and nPOGR21 scores are 100%, indicating that the list of the top 10 DEGs from the oligo microarray data is fully reproducible in the cDNA microarray data when considering correlated molecular changes.

Then, we study the reproducibility of two relatively small-scaled experiments for lung cancer. By SAM with 1% FDR control, we detect 2157 DEGs in the oligo microarray data with 38 samples and only 336 DEGs in the cDNA microarray data with 18 samples. Finding significant correlations with 0.1% FDR control, most of the POGR and nPOGR scores are low and only slightly higher than the corresponding POG and nPOG scores (Table 1). Loosening the FDR level to 5% for significant correlations, the DEG lists from the cDNA microarray data are highly reproducible in the larger oligo microarray data, with all the POGR scores and most of the corresponding nPOGR scores reaching above 90% (all with \(P < 2 \times 10^{-3}\) (see Supplementary Table S1). In another direction, all the POGR and nPOGR scores for the lists of the top-ranked 10, 50 and 100 DEGs increase to above 70%. However, for the list of all the 2157 DEGs from the larger oligo microarray data, the POGR12 and nPOG12 scores measuring its reproducibility in the smaller cDNA microarray data are only 40% and 34%, respectively. This could be explained by the fact that the smaller samples of the cDNA microarray data may have insufficient statistical power in identifying enough DEGs to characterize the disease.

Finally, for a non-cancer disease (DMD), by SAM with 1% FDR control, we detect 805 and 800 DEGs in two datasets, respectively. For the full lists, the POG and nPOG scores are at a rather high level around 50%. Considering significant correlations with 0.1% FDR control level, the POGR and nPOGR scores in both directions increase to above 60%. Notably, for the top-ranked DEG lists, most POGR and nPOGR scores are above 80%, much higher than the corresponding POG and nPOG scores (Table 1). For example, for the
Table 1. Consistence scores between DEG lists for the same disease when using 0.1% FDR control for significant correlations

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEGs</th>
<th>POG</th>
<th>nPOG</th>
<th>POGR</th>
<th>nPOGR</th>
<th>Maxa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td>0.30</td>
<td>0.90</td>
<td>0.89</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Top 30</td>
<td>0.14</td>
<td>0.14</td>
<td>0.78</td>
<td>0.69</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.15</td>
<td>0.14</td>
<td>0.80</td>
<td>0.66</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>1054 vs. 1343</td>
<td>0.38</td>
<td>0.30</td>
<td>0.90</td>
<td>0.74</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>From 102 samples to 103 samplesb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td>0.30</td>
<td>1.00</td>
<td>1.00</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Top 30</td>
<td>0.14</td>
<td>0.14</td>
<td>0.92</td>
<td>0.89</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.15</td>
<td>0.14</td>
<td>0.94</td>
<td>0.90</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>From 38 samples to 18 samplesb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Top 30</td>
<td>0.20</td>
<td>0.19</td>
<td>0.20</td>
<td>0.19</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.31</td>
<td>0.30</td>
<td>0.32</td>
<td>0.31</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>2157 vs. 336</td>
<td>0.13</td>
<td>0.09</td>
<td>0.13</td>
<td>0.09</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>From 24 samples to 36 samplesb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.70</td>
<td>0.70</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Top 30</td>
<td>0.42</td>
<td>0.42</td>
<td>0.92</td>
<td>0.92</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.54</td>
<td>0.54</td>
<td>0.94</td>
<td>0.94</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>805 vs. 800</td>
<td>0.53</td>
<td>0.50</td>
<td>0.72</td>
<td>0.68</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>From 36 samples to 24 samplesb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td>0.20</td>
<td>0.80</td>
<td>0.80</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Top 30</td>
<td>0.42</td>
<td>0.42</td>
<td>0.80</td>
<td>0.80</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Top 100</td>
<td>0.54</td>
<td>0.54</td>
<td>0.85</td>
<td>0.85</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>800 vs. 805</td>
<td>0.53</td>
<td>0.49</td>
<td>0.66</td>
<td>0.61</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>

aMax means the maximum potential agreement beyond chance (see ‘Methods’ section).
bTwo datasets for each disease are marked by their sample sizes. ‘From dataset1 to dataset2’ means that the reproducibility of a DEG list detected in dataset1 is evaluated in dataset2.

As shown in Table 1, for the prostate cancer dataset with 24 samples, the POG and nPOG scores for evaluating its reproducibility in the dataset with 36 samples are both 20%, while the corresponding POGR and nPOGR scores are both 70%. In another direction, the corresponding POGR and nPOGR scores are both 80%. Thus, the DEG lists from one of the two studies are highly reproducible in another from the view of correlated molecular changes.

Here, all the POG and POGR scores in Table 1 are significantly higher than the average POG and POGR scores of the random list pairs extracted from the corresponding original datasets ($P < 1 \times 10^{-4}$).

3.2 Factors affecting the scores

As shown in Figure 1, for 10 000 pairs of lists with the same length randomly selected from each dataset, both of the average POG and POGR scores increase as the list length increases.

On the other hand, both the corresponding nPOG and nPOGR scores fluctuate around zero as the list length increases (Fig. 1). Thus, compared with their non-normalized counterparts, the normalized metrics nPOG and nPOGR are more stable and independent of the list lengths and general data correlations in the datasets. For randomly extracted lists, when using less stringent FDR control levels for selecting significant correlations, the nPOGR scores are still around zero though the POGR scores may increase (Fig. 2).

As shown in Figure 3, in each of the datasets for prostate cancer, the correlation distribution for all gene pairs is similar to the correlation distribution for DEG pairs selected by SAM with 1% FDR control, indicating that the expressions of genes in the prostate cancer data tend to be widely correlated. In such a situation, the nPOGR scores decrease obviously from the POGR scores (Table 1). In the two datasets for lung cancer, the distributions of the correlations of DEG pairs are distinct from the distributions

![Fig. 1. The effect of list lengths on POG, nPOG, POGR, and nPOGR. The x-axis represents list lengths ranging from 100 to 2000, and the y-axis represents the average scores of 10 000 pairs of gene lists randomly selected from the original datasets. Here, 0.1% FDR level is used for detecting significant correlations. The legend for all the six subgraphs is the same as shown within the last one.](https://example.com/fig1.png)
Fig. 2. The effect of FDR control levels on POG, nPOG, POGR and nPOGR. The x-axis represents FDR control levels ranging from 0.01% to 5%. The y-axis represents the average scores of 10,000 pairs of gene lists with a given length (1000) randomly selected from the original datasets. The legend for all the six subgraphs is the same as shown within the last one.

of all gene pairs, indicating that the expression correlations in these datasets are largely limited to DEGs. Similar results are observed in the two datasets for DMD. In this situation that most non-DEGs are not significantly correlated, the nPOGR scores decrease slightly from the POGR scores (Table 1).

3.3 Comparison between DEG lists from different diseases

Finally, we analyze the consistence of DEG lists from datasets for different diseases, using the larger dataset for each disease. The DMD dataset shares a total of 9200 and 8495 genes with the prostate and lung cancer datasets, respectively, and the two cancer datasets have 6797 genes in common. Similarly to the above analysis, DEGs are selected by SAM with 1% FDR control and the correlated gene pairs are determined with 0.1% FDR control.

As shown in Table 2, the POGR and nPOGR scores of DEG list pairs from different diseases are obviously much smaller than the scores from the same disease. For example, the nPOGR score is 0 for the full DEG lists from prostate and DMD, while the nPOGR score is 74% for the full DEG lists from two prostate cancer datasets and 68% for the full DEG lists from two DMD datasets. For DEG lists from a cancer dataset and a DMD dataset, most of the POGR and POGR scores are not much different from the scores for gene lists randomly selected from the original datasets (P > 1%). When using looser FDR control levels for selecting correlated genes, the nPOGR scores of DEG lists from DMD and cancer still fluctuate around zero (Supplementary Table S2).

Notably, the nPOGR scores between the DEG lists from the two cancers are higher than the scores between DMD and cancer (Table 2 and Supplementary Table S2), suggesting that the expression patterns of the altered genes in different cancers may be more consistent. For example, between the full DEG lists from prostate and lung cancer, the nPOGR12 score is 15% and the corresponding POGR12 score is 35%, which is significantly higher than the average score of randomly selected lists.

As mentioned in ‘Methods’ section (Section 2.3), we can also evaluate the reproducibility of DEG lists in a broad sense by just counting the overlapped and correlated genes, without regulation direction requirement. Calculated similarly as the approaches described in ‘Methods’ section, for DEG lists selected from the datasets for the same disease, all the scores are almost unchanged (Supplementary Table S3 and Table S4), reflecting that DEGs from the same disease are regulated in a similar way. However, interestingly, for DEG lists from the cancer and DMD datasets, some scores are unexpectedly high (Supplementary Table S5 and Table S6), especially when using less stringent FDR control levels for selecting correlated gene pairs. To understand these surprising results, we find that the full DEG list from DMD are non-randomly overlapped with the full DEG list from the prostate cancer data (POG = 30%, P < 1 × 10^{-4}) as well as the full DEG list from the lung cancer data (POG = 53%, P < 1 × 10^{-4}), indicating that many genes (DEGs) disturbed in DMD are also differentially expressed in cancer. Without the requirement for regulation direction, the overlapped DEGs between DMD and cancer tend to be correlated with other DEGs in each disease and thus will introduce higher POGR as well.
As nPOGR scores. However, when considering regulation directions of genes, all the scores become rather low, suggesting that the genes differentially expressed in different diseases may be regulated in different ways. Thus, it is important to consider the regulation directions of DEGs when evaluating their reproducibility.

4 DISCUSSION

Intuitively, a high POG score is expected for two studies for a disease. Similarly, a high POGR score suggests apparently high reproducibility of DEG lists in the sense that different studies have actually detected similar results for a disease. Because both POG and POGR scores are dependent on list lengths, they cannot be simply used to compare the reproducibility of gene lists with different lengths. Statistically, for each obtained POG or POGR score, we can calculate the P-value of observing the score in random lists from the original data, which may help interpret the intuitive score.

However, even a small P-value for a POG or POGR score cannot guarantee a high percentage of shared genes between two lists. By normalizing the effects of list length and correlation distribution in datasets, the proposed nPOGR metric suggests a more appropriate way for comparing the reproducibility of gene lists with different lengths and from different studies.

As previously demonstrated by us (Zhang et al., 2008), though DEG lists from current small-scaled microarray studies may only capture a small fraction of DEGs in a disease and thus show low POG scores, each separately determined list may comprise mostly true DEGs. Usually, the global expression changes of genes in a disease may introduce great uncertainty to the findings at the individual genes level. Under such a situation, even if there are many ‘true’ DEGs in a disease, we may be only interested in analyzing the most significant ones. The proposed POGR metric is especially helpful for evaluating such short lists of interesting genes. It should be noted that distinct methods for selecting DEGs, such as SAM (Tusher et al., 2001), analysis of variance (ANOVA) (Hardeo and Mohammed, 2000) and empirical Bayes with t-statistic (Efron et al., 2001; Oshlack et al., 2007), may capture different statistical aspects of expression changes and contribute to the observed inconsistency between the derived DEGs (Jeffery et al., 2006; Qi et al., 2006).

Thus, it warrants further investigation to study the reproducibility of DEG lists found by different DEG selection approaches. At the functional module level, the strictly inconsistent gene lists generated in different experiments for a disease could be rather functionally consistent (Gao et al., 2006; Hosack et al., 2003; Zhu et al., 2007). Specifically, we have demonstrated functional modules related to a disease could be captured by only a small fraction of DEGs in the disease (Yang et al., 2008). Therefore, similarly to POGR metric, a metric considering gene overlaps from the view of functional similarity will probably provide another way for evaluating the reproducibility of the apparently inconsistent discoveries in microarray studies, which deserves our future work. Other metrics or views could be considered for observing different discovery results for a disease. Notably, Mao et al (2009) showed that the discriminating genes identified in a study may preserve their disease discriminating ability in another study though these genes might not be selected independently in the later dataset because they might not be the optimum ones in both studies. Their result implies high concordance of DNA microarrays by the signals and classifier transferability of the discriminating genes, rather than simply relying on counting gene overlaps. This concept is somewhat similar to ours. However, our proposed metrics should be more statistically interpretable and applicable in classical studies on evaluating differential gene expressions in diseases.

Finally, we note that the irreproducibility problem of finding disease markers also exists in other post-genomic areas such as proteomics (Ransohoff, 2005), metabolomics (David I. and

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEGs</th>
<th>POG</th>
<th>nPOG</th>
<th>POGR</th>
<th>nPOGR</th>
<th>Max</th>
<th>DEGs</th>
<th>POG</th>
<th>nPOG</th>
<th>POGR</th>
<th>nPOGR</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Prostate to DMDb</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.98</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td>vs. Top 30</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.00</td>
<td>0.93</td>
<td>vs. Top 30</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.00</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td>Top 100</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.89</td>
<td>Top 100</td>
<td>0.01</td>
<td>0.00</td>
<td>0.06</td>
<td>0.00</td>
<td>0.89</td>
</tr>
<tr>
<td>From Lung to DMDb</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>vs. Top 30</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>vs. Top 30</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>From Lung to Prostateb</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>From Lung to DMDb</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>From Prostate to DMDb</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>From Lung to Prostateb</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>Top 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>vs. Top 50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
<td>Top 100</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Indicated as in Table 1.
Douglas B., 2006) and large-scale screens for cancer mutations (Chen et al., 2007; Frantz, 2005). The apparently inconsistent discoveries generated in these post-genomic areas could also be evaluated by considering correlated disease markers. Because this problem is of fundamental importance in the field of systems biology, it deserves our future researches.

ACKNOWLEDGEMENTS

National Natural Science Foundation of China (Grant Nos. 30170515, 30370388, 30571034).

Conflict of Interest: none declared.

REFERENCES


