Published in final edited form as:

Stat Anal Data Min. 2013 April 1; 6(2): 144–155. doi:10.1002/sam.11180.

AAPL: Assessing Association between P-value Lists

Tianwei Yu^{1,*}, Yize Zhao¹, and Shihao Shen²

¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

²Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

Abstract

Joint analyses of high-throughput datasets generate the need to assess the association between two long lists of p-values. In such p-value lists, the vast majority of the features are insignificant. Ideally contributions of features that are null in both tests should be minimized. However, by random chance their p-values are uniformly distributed between zero and one, and weak correlations of the p-values may exist due to inherent biases in the high-throughput technology used to generate the multiple datasets. Rank-based agreement test may capture such unwanted effects. Testing contingency tables generated using hard cutoffs may be sensitive to arbitrary threshold choice. We develop a novel method based on feature-level concordance using local false discovery rate. The association score enjoys straight-forward interpretation. The method shows higher statistical power to detect association between p-value lists in simulation. We demonstrate its utility using real data analysis. The R implementation of the method is available at <http://userwww.service.emory.edu/~tyu8/AAPL/>.

1 INTRODUCTION

Omics studies generate test results on thousands of biological features, e.g. genes, SNPs and metabolites. In order to boost statistical power or gain extra insight, there is an interest to jointly analyze datasets from different labs, or from different types of measurements on the same biological features, e.g. expression and splicing of genes. Such joint analyses often involve semi-quantitative assessments on the agreement between experiments. In some situations, comparison can be done by utilizing the raw data and modeling experiment-specific variation and bias [1]. One example is the meta-analysis of microarrays, where datasets from multiple labs are combined to increase the statistical power. Before merging datasets, it is important to examine and compare the characteristics of the datasets, for example by dimension reduction and visualization [2, 3].

In some other situations, joint analysis of the raw data may be difficult due to computational load or nature of the data. Connectivity map (cmap), which compares a query gene expression signature to a large reference database, uses ranked list comparison based on the Kolmogorov-Smirnov statistic [4]. Other examples include comparing whether the list of differentially expressed genes in two datasets generated from different platforms agree with each other, testing the association between gene expression and alternative splicing, and testing the association of differential splicing between tissues. In such situations, comparisons need to be made on certain feature-level summary values, e.g. test statistics, p-values, and ranks.

*To whom correspondence should be addressed. tianwei.yu@emory.edu.

Currently there are a few options for such comparisons. One option is to compare ranked lists [5–7]. However it is known that rankings are unstable – small perturbations in the data can cause considerable changes in the rankings [8, 9]. In addition, weak correlations of the ranks may exist due to inherent biases in the high-throughput technology used to generate the multiple datasets included in joint analysis. For example, high-throughput RNA sequencing (RNA-Seq) has become a powerful technology for transcriptome analysis [10]. As RNA-Seq represents a sampling of the entire transcriptome, the detection of differential gene expression and alternative splicing by RNA-Seq is inherently biased towards highly expressed genes. Such a bias may result in spurious correlation of the ranks when comparing multiple RNA-Seq datasets. Another option is to use a hard cutoff of false discovery rate (FDR) within each dataset, and tabulate the significant/insignificant features between datasets [11]. Then a test on contingency table can be used to assess the significance of the association. The result may be sensitive to the arbitrary choice of cutoff values. There is a collection of agreement methods in the statistical literature, which are reviewed in [12, 13]. However, such measures are not directly applicable given the lack of interpretability of the scores and the skewness of the data.

Here we report a novel method to assess the association between the two p-value lists. We introduce an association score that has a straight-forward interpretation, and procedures to assess its significance and confidence interval. Utilizing the local false discovery rate (lfdr) concept [14], the method minimizes the impact of the features that are not significantly associated with the outcome variable in both datasets. It also avoids the arbitrary selection of cutoff values. We demonstrate its superiority over the aforementioned methods in simulations, and its utility using a real dataset.

2 METHODS

2.1. The association score

We utilize local false discovery rate (lfdr) [14] derived from the p-values to make probabilistic assessment. The local false discovery rate is a Bayes posterior probability that a feature is null given its test statistic or p-value [14]. The calculation of local false discovery rate (lfdr) is based on density estimation. Comparatively the false discovery rate (FDR) relies on tail area instead of density [15, 16]. A number of methods were proposed for the calculation of lfdr [14, 17–19].

Let the two sets of p-values be $p_1^{(1)}, p_2^{(1)}, \dots, p_N^{(1)}$ and $p_1^{(2)}, p_2^{(2)}, \dots, p_N^{(2)}$. The local false discovery rate (lfdr) is derived from each set of p-values,

$$lfdr_i^{(k)} = \text{Prob}(\text{gene } i \text{ is null in } k^{\text{th}} \text{ experiment} | p_i^{(k)}), i=1, \dots, N.$$

In this study we utilize the kernel-based local fdr (kerfdr) method [17]. The kerfdr approach transforms the p-values using a probit transformation, and models the mixture of null and non-null statistics using a two-component mixture model with semi-parametric density estimation [17]. Any other lfdr procedure can be used. The lfdr value for each gene is a probabilistic statement of how likely the gene is non-null in each test.

For an association score, we consider the following ratio: the expected proportion of concordant genes, i.e. genes called null/non-null in both lists, divided by its expected value if the two lists were independent (derived in section 2.2),

$$S = \frac{\frac{1}{2} \sum_i (1 - \text{lfdr}_i^{(1)}) (1 - \text{lfdr}_i^{(2)}) + \frac{1}{N} \sum_i \text{lfdr}_i^{(1)} \text{lfdr}_i^{(2)}}{1 - m_1 - m_2 + 2m_1 m_2} = 1 + \frac{\frac{2}{N} \sum_i \text{lfdr}_i^{(1)} \text{lfdr}_i^{(2)} - 2m_1 m_2}{1 - m_1 - m_2 + 2m_1 m_2},$$

where N is the total number of genes, m_1 is the sample mean of the lfdr values in experiment one, and m_2 is the sample mean of the lfdr values in experiment two. When the two lists have some level of agreement, i.e. genes called significant in one list is more likely to be called significant in the other list, the value of S is greater than one. If the two lists have no association, the value of S is close to one. Another scenario is the two lists are inversely related, in which case the value of S is less than one.

2.2. Permutation test and its normal approximation

Given the empirical distribution of lfdr values derived from the two lists, we can resort to permutation test to find the significance level. The null hypothesis is that the assignment of significance level is independent between the two lists. We can randomly permute one of the lfdr lists M times and compute the test statistic $\{S^{(m)}\}_{m=1, \dots, M}$. The proportion of the sampled permutations with test statistics more extreme than the observed S is taken as the p-value of the two-sided test,

$$p = \frac{2}{M} \min\left(\sum_{m=1}^M I(S^{(m)} \geq S), \sum_{m=1}^M I(S^{(m)} \leq S)\right).$$

In the permutation test, the empirical distributions of $\text{lfdr}^{(1)}$ and $\text{lfdr}^{(2)}$ are fixed, i.e. the values of N , m_1 , and m_2 do not change. Only the part $\sum_i \text{lfdr}_i^{(1)} \text{lfdr}_i^{(2)}$ determines the significance. Meanwhile, since the range of value for all the $\text{lfdr}_i^{(k)}$ ($i=1, 2, \dots, N; k=1, 2$) are zero to one, it is easy to prove that both $\{\text{lfdr}_i^{(1)}\}_{i=1, 2, \dots, n}$ and $\{\text{lfdr}_i^{(2)}\}_{i=1, 2, \dots, n}$ satisfy the condition required by the permutation central limit theorem [20]. Thus we can use the permutation central limit theorem to illustrate that under the null hypothesis of independence, $\sum_i \text{lfdr}_i^{(1)} \text{lfdr}_i^{(2)}$ approximately follows the normal distribution with

$$\begin{aligned} \mu &= \sum_i E(\text{lfdr}_i^{(1)}) E(\text{lfdr}_i^{(2)}) = N m_1 m_2, \\ \sigma^2 &= \sum_i \left\{ E\left(\text{lfdr}_i^{(1)}\right)^2 V\left(\text{lfdr}_i^{(2)}\right) + E\left(\text{lfdr}_i^{(2)}\right)^2 V\left(\text{lfdr}_i^{(1)}\right) + \right. \\ &\quad \left. V\text{lfdr}_i^{(1)} V\text{lfdr}_i^{(2)} + i \neq j E\text{lfdr}_i^{(1)} \text{lfdr}_j^{(1)} E\text{lfdr}_i^{(2)} \text{lfdr}_j^{(2)} - E\text{lfdr}_i^{(1)} E\text{lfdr}_j^{(1)} E\text{lfdr}_i^{(2)} E\text{lfdr}_j^{(2)} \right\} \\ &= \frac{N^2}{N-1} V_1 V_2. \end{aligned}$$

where N is the length of the p-value vector, m_1 is the sample mean of the lfdr values in experiment one, V_1 is the sample variance of the lfdr values in experiment one, m_2 is the sample mean of the lfdr values in experiment two, and V_2 is the sample variance of the lfdr values in experiment two. Because N is large in the current setting, we can see that under the null hypothesis,

$$S \sim \text{Normal}\left(1, \frac{1}{N} \left(\frac{2}{1 - m_1 - m_2 + 2m_1 m_2}\right)^2 V_1 V_2\right).$$

2.3. The possible range of the score

It may be useful to assess the score's deviation from one in relation to its maximum possible and minimum possible values, because the statistic S is not confined to a fixed interval. In practical terms, it is unlikely that the similarity score can get close to the maximum/minimum possible values.

In the permutation setting, the possible range of S can be obtained numerically. It can be easily shown that its maximum possible value is achieved by ordering both lists of lfr values in an increasing order, and taking the value of the statistic from the re-ordered lists. Here we consider only the dot product because other parts are constants.

First, suppose we have $a_1 \leq a_2$ and $b_1 \leq b_2$. Then because $(a_1 - a_2)(b_1 - b_2) \geq 0$, we have $a_1 b_1 + a_2 b_2 \geq a_1 b_2 + a_2 b_1$. Now considering two longer lists, $a_1 \leq a_2 \leq \dots \leq a_N$ and $b_1 \leq b_2 \leq \dots \leq b_N$. We start with the dot product with any pairing, $a_1 b_m + a_2 b_k + \dots + a_q b_1 + \dots + a_N b_r$. Because $a_1 b_1 + a_q b_m \geq a_1 b_m + a_q b_1$, we can obtain an equal or higher value by changing the pairing to $a_1 b_1 + a_2 b_k + \dots + a_q b_m + \dots + a_N b_r$. Following the same procedure, we then iteratively increase the value of the dot product by considering the pair (a_2, b_2) and so on. In the end, we reach $a_1 b_1 + a_2 b_2 + \dots + a_N b_N$, and its value is greater or equal to the dot product we started with. Because the starting point represents any pairing, we have shown that the dot product with matched order achieves the maximum possible value. Similarly, by replacing b_i with $(1 - b_i)$ to reverse the order and using the above argument, it can be shown that the minimum possible value of S is achieved by ordering one list in increasing order and the other in decreasing order, and taking the value of the statistic from the reordered lists.

2.4. Bootstrap confidence interval

Although the above test provides inference regarding whether the two lists are associated, it doesn't provide the confidence interval for S . Theories of confidence interval on dot product or correlation coefficient cannot be applied in the current setting because the distributions of the lfr values are highly skewed – the distributions are often “U” shaped with density peaking around zero and one. We resort to bootstrapping to obtain the confidence interval of S . Unlike in the permutation test, the pairing between the two lists is not disrupted. In each bootstrap iteration, we resample N pairs of lfr values with replacement, and calculate the test statistic from the resampled data. After B iterations, we then take the 2.5th and 97.5th percentiles of the test statistic values as the 95% confidence interval of S .

2.5. Simulation study

We tested the performance of the proposed method in comparison with three other methods. The first was the ordered list comparison of the p-values [5]. Notice this method is not specifically designed for p-value lists. The second was Fisher's exact test on the contingency table we first dichotomized the p-values into “significant” and “non-significant” based on FDR cutoffs, and then conducted the Fisher's exact test on the 2×2 contingency table. Two FDR cutoffs were tested: 0.1 and 0.2. The third method was testing the significance of the Pearson's correlation coefficient between the two p-value lists.

Each simulated dataset contained two lists of p-values $\{p_i^{(1)}\}_{i=1,\dots,N}$ and $\{p_i^{(2)}\}_{i=1,\dots,N}$. In order to generate correlated null p-values between the two lists, we first generated two lists of test statistics $\{s_i^{(1)}\}_{i=1,\dots,N}$ and $\{s_i^{(2)}\}_{i=1,\dots,N}$. For genes that were null in both lists, their test statistics $\begin{pmatrix} s_i^{(1)} \\ s_i^{(2)} \end{pmatrix}$ were drawn from the bivariate normal distribution with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance-

covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Different values of ρ were used (0, 0.2, 0.4 and -0.2) to reflect weak to moderate correlation. For genes that were null in one of the lists, their null test statistics were drawn from the standard normal distribution. Non-null test statistics were drawn from the normal distribution $N(-3, 0.5)$ or $N(-4, 1)$. Supporting Figure 1 shows the level of overlap between the null and non-null distributions.

After generating the test statistics, we took the standard normal CDF of the test statistics to generate the p-values, $p_i^{(k)} = \Phi(s_i^{(k)})$, $i = 1, \dots, N$, $k = 1, 2$. This way the null gene p-values in each list were uniformly distributed, and the non-null gene p-values were close to zero. For genes that were null in both lists, their correlation coefficient between p-values were very close to the correlation coefficient between the test statistics. Different levels of overlap between non-null genes in the two lists were tested.

3 RESULTS

Simulation results

Setting the alpha level at 0.05, we compared the statistical power of each method at different simulation settings. Figure 1 shows the results when the total number of genes is 4000, and the non-null test statistics are drawn from $N(-3, 0.5)$. More complete results can be found in the Supplementary Materials. Four methods were compared – our lfr-based method, ordered list comparison, testing of contingency table, and testing the significance of correlation coefficient between p-value lists. For our method, normal approximation yielded identical results as permutation test. Thus only the normal approximation results are plotted.

In every subplot of Figure 1, the vertical dotted line represents the number of overlaps when the two sets of non-null genes were drawn independently. A method should yield a power of 0.05 or lower to maintain correct size of the test. The further away from the vertical line, the higher level of association exists between the two lists, where a better method should yield higher power. Every row of Figure 1 represents a different level of correlation between the p-values of null genes in both lists. A better method should be less affected by the correlation of the null genes. Every column of Figure 1 represents a different scenario in terms of the number of non-null genes in each list.

For the ordered list comparison, we used the default method in the R package `OrderedList` [5] from Bioconductor [21]. The method returns the best result after internally comparing several parameter settings, which causes the size of the test to be incorrect. When the two non-null lists were independently generated (vertical dotted line), `OrderedList` showed a much higher frequency than 0.05 to call the two lists significantly associated. With incorrect size, the nominal power at other simulation settings become questionable.

Other tests maintained the correct size when the null gene p-values are not correlated (Fig. 1, top row). The test based on p-value correlation shows markedly lower power than the rest. Our method shows higher power than the contingency table-based tests. When weak positive correlation between null gene p-values is present (Fig. 1, second row), the test based on p-value correlation becomes incorrect in size. The performance of our method and the contingency table-based test is only slightly impacted. When the correlation between null gene p-values becomes moderate (Fig. 1, third row), We see the curve of all methods shift further to the left. The contingency table based test using FDR cutoff of 0.1 (cyan curve) was least impacted by the correlation, while the contingency table based test using FDR cutoff of 0.2 (orange curve) became incorrect in size when the number of significant genes is larger (Fig. 1, third row, right panel). We also tested the situation of weak negative

correlation between null gene p-values (Fig. 1, fourth row). In this scenario both ordered list test and test of p-value correlation were both clearly wrong in size. Our lfr-based test and contingency table test were impacted very little by the null-gene p-value correlation.

We show other simulation results in Supporting Figures 2~4. Generally, when the total number of genes is 4000 and the non-null test statistics were generated from $N(-4, 1)$, the null gene p-value correlation has even less impact on our method and the contingency table-based test using FDR cutoff of 0.1 (Supporting Figure 2). However the contingency table-based test using FDR cutoff of 0.2 showed worse performance when the null gene p-value correlation is moderate, indicating the sensitivity of the contingency table-based test on the cutoff value (Supporting Figure 2, third row). We also tested the performance of the methods using simulations where total number of genes is 2000. We kept the same number of non-null genes (200 and 500). In this case the proportion of non-null genes are higher, and the performance of all methods are better. As to the comparison between the methods, the overall trend remains the same (Supporting Figures 3 & 4).

Prostate cancer data

To demonstrate the utility of the method, we used the prostate cancer datasets in the R package MetaPCA [3], which were compiled from four separate gene expression studies [22–25] and subjected to gene filtering based on study-wise mean, variance and missing proportions [3]. All four datasets contain samples in three groups – normal, primary tumor and metastasized tumor. Within each dataset, we first tested differential expression using limma [26] between the normal and primary tumor groups, and between primary and metastasized tumor groups. We then tested the association between the p-value lists of pairs of experiments (Table 1). All the significant test results were in the positive direction, i.e. the two lists showed higher level of agreement compared to random.

Between normal and primary tumor, the expected number of differentially expressed genes were 649 (Yu), 747 (Lapointe), 780 (Tomlins), and 618 (Varambally) respectively. Using the lfr-based test, three of the four experiments showed certain levels of agreement (Table 1), while the Tomlins dataset appeared to generate a significant gene list that was not associated with any of the other three. Testing of contingency table and testing correlation coefficient between p-value lists generated similar result, while the ranked list-based test yielded significant p-values between Tomlins and two other experiments. We then examined the scatter plots of the lfr values using color representation of smoothed density (Figure 2). Between the Tomlins and other datasets, we noticed substantial density at the lower-right and/or upper-left corner, indicating different calls made by different experiments. By checking the original publication, we found the Tomlins dataset was the only one using laser-capture microdissection (LCM) and whole-transcriptome amplification [23], while the other three datasets were generated from tissue samples. This could explain the difference of gene selection at the global level between the Tomlins and the other three datasets.

Between primary and metastasized tumor, the expected number of differentially expressed genes are 915 (Yu), 758 (Lapointe), 824 (Tomlins) and 910 (Varambally) respectively. According to the lfr-based test, the level of agreement between the Yu and Lapointe experiments dropped to a marginal level, while both experiments still showed some agreement with the Varambally experiment (Table 1). Again the Tomlins dataset showed no agreement with the other three. The other three test generated similar results, except the ranked list-based test called significant association between the Lapointe and Yu datasets, and the contingency table-based test showed marginal significance between the Varambally and Yu datasets. Another way of looking at the data is to examine if the test of the two stage transitions, i.e. normal *v.s.* primary tumor and primary tumor *v.s.* metastasized tumor, point to similar genes. We tested between the two p-value lists within each experiment. One of the

datasets (Varambally) showed no agreement between the two stage transitions, while two other datasets (Lapointe and Tomlins) showed strong agreement. Examining the scatter plots of the *lfdr* values using color representation of smoothed density (Figure 3), we noticed that the lack of agreement of Yu and Varambally datasets were caused by substantial density at the lower-right corner, i.e. large number of genes called insignificant in normal v.s. tumor test, but significant in tumor v.s. metastasis test, which could well be explained by the progression of tumor leading more genes to be differentially expressed. Thus the lack of agreement cannot be taken at its face value here.

In all three sets of comparisons, we notice that in all cases where a majority of the tests claimed significant association, the *lfdr*-based test always shows the smallest p-value. On the other hand, in the cases where all tests yielded insignificant results, the *lfdr*-based test doesn't show any tendency of generating smaller p-values than the others (Table 1).

Comparing gene expression in multiple types of cancer

We analyzed gene expression data derived from tissue samples of multiple cancer types using the Affymetrix U133A or U133 Plus 2.0 arrays. Most of the data were downloaded from the Gene Expression Omnibus (GEO) [27]. The data we studied include five breast cancer datasets GSE5364 [28], GSE5847 [29], GSE10780 [30], GSE15852 [31], GSE29431 (no reference available); four lung cancer datasets GSE7670 [32], GSE10072 [33], GSE18842 [34], GSE19804 [35]; five colorectal cancer datasets GSE18105 [36], GSE21510 [37], GSE22598 [38], GSE24514 [39], GSE23878 (no reference available); three renal cancer datasets GSE14762 [40], GSE6344 [41], GSE781 [42]; and four oral cancer datasets GSE30784 [43], GSE31056 [44], GSE9844 [45], and a dataset published by Ziober *et al* that is not in GEO [46]. For each dataset, we first obtained the gene level p-values by testing between normal and cancer phenotype using LIMMA [26]. We then selected the overlapping probesets of the arrays. As all the datasets involved were generated using Affymetrix U133A or U133 Plus 2.0 arrays, 22277 probesets were present in all the datasets. Thirdly, we matched the probesets to ENTREZ gene IDs. A total of 12713 ENTREZ genes were found in all datasets. In the case where an ENTREZ gene was represented by more than one probesets, the geometric mean of the corresponding p-values was taken as the p-value of the gene.

First, we examined the global agreement between all pairs of datasets using all the 12713 genes. The p-values between most dataset pairs are small, indicating a certain level of association across all the cancer types under study. The large number of genes made the statistical power very high to detect even weak associations. For better visual comparison, we use a heatmap of the test statistic (Fig. 4). The lung cancer and oral cancer datasets clearly showed higher within-type association, while the breast cancer datasets appeared to be heterogeneous.

It is expected that different biological pathways may behave differently in various cancer types. With our new tool to assess association of the p-values of groups of genes between datasets, we examined which biological functions tend to be differentially regulated between cancer types. We note that here we assess the behavior of biological processes only based on the genes' first-order relations with the clinical phenotype, i.e. control v.s. a certain type of cancer. In order to map genes to biological functions/pathways, we resorted to the functional annotations by Gene Ontology (GO) [47]. We downloaded the GO slim terms that are a set of relatively broad ontologies pertaining to various aspects of the biological system [47]. We assessed the association between datasets using each GO slim term that is assigned more than 200 human ENTREZ genes. A total of 74 GO slim terms were included in this study. The number of genes assigned to these terms range from 202 to 4064.

For each dataset, we first found the p-values of all the 12713 genes. We then transformed the p-values into lfr values using kerfdr. For each GO slim term, we selected the genes assigned to the term. Using the lfr values of these genes, we conducted the test of association between all pairs of datasets. With a p-value cutoff of 0.001 to offset the large number of tests conducted, we found which dataset pairs were associated based on each GO slim term. Our purpose was to find which GO slim terms behave more consistently within cancer types, compared to between cancer types. For this purpose, the ‘ideal’ GO slim term should generate significant test result on pairs of datasets of the same cancer type, while generating insignificant results on pairs of datasets of different cancer types. Thus we considered dataset pairs from the same cancer type “true positives”, and pairs from different cancer types “true negatives”. We then used the balanced accuracy, i.e. the average value of sensitivity and specificity to summarize the result from each GO slim term,

$$\text{balanced accuracy} = \left(\frac{\# \text{ within-type pairs called significant}}{\text{total } \# \text{ within-type pairs}} + \frac{\# \text{ between-type pairs called insignificant}}{\text{total } \# \text{ between-type pairs}} \right) / 2$$

Because the breast cancer datasets were highly inconsistent, we put more emphasis on other cancer types when summarizing the results. We ordered the GO slim terms based on their balanced accuracy on the datasets excluding the breast cancer datasets. The GO slim terms showing highest and lowest balanced accuracy are shown in Table 2. The complete results of all 74 GO slim terms is in Supporting Table 1. We noticed that the rankings agree very well with and without the breast cancer datasets.

The top 10 GO slim terms showed two clear themes (Table 2). The first theme is cellular and intercellular structures, including GO:0048646 (anatomical structure formation involved in morphogenesis), GO:0007155 (cell adhesion), GO:0008092 (cytoskeletal protein binding), and GO:0007010 (cytoskeleton organization). This is easy to interpret as different cancer types are derived from tissues of distinct structural characteristics. The second theme is small molecule metabolism and energy generation, including GO:0006520 (cellular amino acid metabolic process), GO:0006629 (lipid metabolic process), GO:0008289 (lipid binding), GO:0005739 (mitochondrion), GO:0008233 (peptidase activity), and GO:0016491 (oxidoreductase activity). It is known that the metabolism of small molecules including key amino acids is associated with the proliferation of cancer cells [48], and cancer cells may reprogram metabolism to attain nutrient self-sufficiency [49]. The results on GO slim terms indicate different types of cancer may behave differently in the small molecule metabolism reprogramming. We further examined the results between pairs of datasets using heatmaps. By plotting the results from the GO slim term GO:0006520 (cellular amino acid metabolic process), we found that most within-type dataset pairs (except breast cancer) were called significant, while most between-type dataset pairs were called insignificant (Fig. 5). The same pattern was observed using the two terms with the next highest balanced accuracy scores GO:0005739 (mitochondrion) (Supporting Figure 5) and GO:0008289 (lipid binding) (Supporting Figure 6).

The bottom 10 GO slim terms were less cohesive in terms of their functional relations. An interesting observation was three terms related to cell cycle, GO:0051301 (cell division), GO:0007067 (mitosis) and GO:0000228 (nuclear chromosome), were included. We examined the results between pairs of datasets using heatmaps. By plotting the results from the GO slim term that generated the lowest balanced accuracy, GO:0000228 (nuclear chromosome), we found that datasets from breast cancer and renal cancer do not show within-cancer type associations, while datasets from oral, lung, and colorectal cancers showed heavy cross-cancer type associations (Fig. 6). Another GO slim term in the bottom 3,

GO:0007067 (mitosis), showed a similar pattern (Supporting Figure 8). These results indicate there might be some relations between the cell cycle behaviors of oral, lung and colorectal cancer cells, while the cell cycle behaviors of renal and breast cancers are divergent, even within the same cancer type. Another GO slim term in the bottom three, GO:0005886 (plasma membrane), generated results showing broad dependencies both within and between cancer types (Supporting Figure 7), indicating a broad association across cancer types of the behavior of the genes in this GO term.

4. DISCUSSION

Overall, our lfdr-based test achieves high statistical power to detect associations between p-value lists, and it is resistant to weak to moderate correlations between null gene p-values. Our test statistics uses the expected proportion of concordant genes. The permutation test assumes that the distribution of the p-values in each list is fixed. Given that the user doesn't re-examine the raw data, this is a reasonable assumption that is also used by the ranked list comparison and the contingency table approach. The inference using permutation test and normal approximation yielded identical results, suggesting that the normal approximation, which has very little computing cost, is appropriate when the number of genes are in the thousands.

Because our test is centered around the dot product between the two lfdr vectors, the permutation test is equivalent to the test of significance of Pearson's correlation coefficient between the lfdr vectors. Although we do prefer the more interpretable test statistic presented in this manuscript, our results does provide a justification of directly testing the significance of Pearson's correlation coefficient between the lfdr vectors, even though the distribution of lfdr values is far from normal. The inference of confidence interval by bootstrapping doesn't converge to the theoretical confidence interval derived from dot product or correlation coefficient that rely on normality assumption, which is caused by the extreme skewness of the lfdr value distribution.

Our method is based on the local fdr concept, and the result depends on the correct estimation of the local fdr values from the p-values, which requires the correct tuning of the lfdr procedure. Using the simulation settings presented in Figure 1, we tested three other local fdr procedures available in R: locfdr [50], twilight [51], and fdrtool [19], all at default settings. The procedures locfdr and fdrtool generated almost identical results as kerfdr (Supporting Figure 9). Twilight results are almost identical to the others when there is no correlation between null gene p-values. In the presence of correlation, the twilight results are more impacted by the correlation (Supporting Figure 9).

Overall, we presented a method to assess the association between p-value lists. The method can be used to find agreement between experiments by directly comparing the p-value lists. Similar to rank-based agreement methods, our method can help assess the level of agreement between datasets before meta-analysis, or find associations between different types of measurements on the same set of biological features.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by NIH grants P20HL113451, U19AI057266 and U19AI090023. The authors thank Dr. Yi Xing and Dr. George Tseng for providing motivating examples and helpful discussions. The comments from two anonymous reviewers helped greatly improve the manuscript.

References

1. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 2008; 5(9):e184. [PubMed: 18767902]
2. Yu T, Ye H, Chen Z, Ziober BL, Zhou X. Dimension reduction and mixed-effects model for microarray meta-analysis of cancer. *Front Biosci.* 2008; 13:2714–2720. [PubMed: 17981746]
3. Kang, D.; Tseng, G. MetaPCA: Meta-analysis in the Dimension Reduction of Genomic data. R package version 0.1.4. 2011. <http://CRAN.R-project.org/package=MetaPCA>
4. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science.* 2006; 313(5795):1929–1935. [PubMed: 17008526]
5. Yang X, Bentink S, Scheid S, Spang R. Similarities of ordered gene lists. *J Bioinform Comput Biol.* 2006; 4(3):693–708. [PubMed: 16960970]
6. Lottaz C, Yang X, Scheid S, Spang R. OrderedList--a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics.* 2006; 22(18):2315–2316. [PubMed: 16844712]
7. Yang X, Sun X. Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers. *BMC Bioinformatics.* 2007; 8:118. [PubMed: 17411443]
8. Boulesteix AL, Slawski M. Stability and aggregation of ranked gene lists. *Brief Bioinform.* 2009; 10(5):556–568. [PubMed: 19679825]
9. Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics.* 2008; 24(2):258–264. [PubMed: 18024475]
10. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009; 10(1):57–63. [PubMed: 19015660]
11. Morrissey ER, Diaz-Uriarte R. Pomelo II: finding differentially expressed genes. *Nucleic Acids Res.* 2009; 37(Web Server issue):W581–586. [PubMed: 19435879]
12. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. *J Am Stat Assoc.* 2002; 97(457):257–270.
13. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat.* 2007; 17 (4):529–569. [PubMed: 17613641]
14. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol.* 2002; 23(1):70–86. [PubMed: 12112249]
15. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met.* 1995; 57(1):289–300.
16. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat.* 2003; 31(6):2013–2035.
17. Guedj M, Robin S, Celisse A, Nuel G. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics.* 2009; 10:84. [PubMed: 19291295]
18. Liao JG, Lin Y, Selvanayagam ZE, Shih WJ. A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics.* 2004; 20(16):2694–2701. [PubMed: 15145810]
19. Strimmer K. A unified approach to false discovery rate estimation. *BMC Bioinformatics.* 2008; 9:303. [PubMed: 18613966]
20. Wald A, Wolfowitz J. Statistical tests based on permutations of the observations. *Annals of Mathematical Statistics.* 1944; 15:358–372.
21. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004; 5(10):R80. [PubMed: 15461798]
22. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A.* 2004; 101(3):811–816. [PubMed: 14711987]
23. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, et al. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet.* 2007; 39(1):41–51. [PubMed: 17173048]

24. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*. 2005; 8(5):393–406. [PubMed: 16286247]
25. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J Clin Oncol*. 2004; 22(14):2790–2799. [PubMed: 15254046]
26. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004; 3:Article3. [PubMed: 16646809]
27. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*. 2006; 411:352–369. [PubMed: 16939800]
28. Yu K, Ganesan K, Tan LK, Laban M, Wu J, Zhao XD, Li H, Leung CH, Zhu Y, Wei CL, et al. A precisely regulated gene expression cassette potentially modulates metastasis and survival in multiple solid cancers. *PLoS Genet*. 2008; 4(7):e1000129. [PubMed: 18636107]
29. Boersma BJ, Reimers M, Yi M, Ludwig JA, Luke BT, Stephens RM, Yfantis HG, Lee DH, Weinstein JN, Ambs S. A stromal gene signature associated with inflammatory breast cancer. *Int J Cancer*. 2008; 122(6):1324–1332. [PubMed: 17999412]
30. Chen DT, Nasir A, Culhane A, Venkataramu C, Fulp W, Rubio R, Wang T, Agrawal D, McCarthy SM, Gruidl M, et al. Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res Treat*. 2010; 119(2):335–346. [PubMed: 19266279]
31. Pau Ni IB, Zakaria Z, Muhammad R, Abdullah N, Ibrahim N, Aina Emran N, Hisham Abdullah N, Syed Hussain SN. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. *Pathol Res Pract*. 2010; 206(4):223–228. [PubMed: 20097481]
32. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC, Lin CH, Whang-Peng J, Hsu SL, Chen CH, et al. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*. 2007; 8:140. [PubMed: 17540040]
33. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one*. 2008; 3 (2):e1651. [PubMed: 18297132]
34. Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Farez-Vidal ME. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer*. 2011; 129(2):355–364. [PubMed: 20878980]
35. Lu TP, Tsai MH, Lee JM, Hsu CP, Chen PC, Lin CW, Shih JY, Yang PC, Hsiao CK, Lai LC, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*. 2010; 19 (10):2590–2597. [PubMed: 20802022]
36. Matsuyama T, Ishikawa T, Mogushi K, Yoshida T, Iida S, Uetake H, Mizushima H, Tanaka H, Sugihara K. MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. *Int J Cancer*. 2010; 127(10):2292–2299. [PubMed: 20162577]
37. Tsukamoto S, Ishikawa T, Iida S, Ishiguro M, Mogushi K, Mizushima H, Uetake H, Tanaka H, Sugihara K. Clinical significance of osteoprotegerin expression in human colorectal cancer. *Clin Cancer Res*. 2011; 17(8):2444–2450. [PubMed: 21270110]
38. Okazaki S, Ishikawa T, Iida S, Ishiguro M, Kobayashi H, Higuchi T, Enomoto M, Mogushi K, Mizushima H, Tanaka H, et al. Clinical significance of UNC5B expression in colorectal cancer. *Int J Oncol*. 2012; 40(1):209–216. [PubMed: 21922135]
39. Alhopuro P, Sammalkorpi H, Niittymäki I, Bistrom M, Raitila A, Saharinen J, Nousiainen K, Lehtonen HJ, Heliovaara E, Puhakka J, et al. Candidate driver genes in microsatellite-unstable colorectal cancer. *Int J Cancer*. 2012; 130(7):1558–1566. [PubMed: 21544814]
40. Wang Y, Roche O, Yan MS, Finak G, Evans AJ, Metcalf JL, Hast BE, Hanna SC, Wondergem B, Furge KA, et al. Regulation of endocytosis via the oxygen-sensing pathway. *Nat Med*. 2009; 15(3):319–324. [PubMed: 19252501]
41. Gumz ML, Zou H, Kreinest PA, Childs AC, Belmonte LS, LeGrand SN, Wu KJ, Luxon BA, Sinha M, Parker AS, et al. Secreted frizzled-related protein 1 loss contributes to tumor phenotype of clear cell renal cell carcinoma. *Clin Cancer Res*. 2007; 13 (16):4740–4749. [PubMed: 17699851]

42. Lenburg ME, Liou LS, Gerry NP, Frampton GM, Cohen HT, Christman MF. Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*. 2003; 3:31. [PubMed: 14641932]
43. Chen C, Mendez E, Houck J, Fan W, Lohavanichbutr P, Doody D, Yueh B, Futran ND, Upton M, Farwell DG, et al. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev*. 2008; 17 (8):2152–2162. [PubMed: 18669583]
44. Reis PP, Waldron L, Perez-Ordóñez B, Pintilie M, Galloni NN, Xuan Y, Cervigne NK, Warner GC, Makitie AA, Simpson C, et al. A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC Cancer*. 2011; 11:437. [PubMed: 21989116]
45. Ye H, Yu T, Temam S, Ziober BL, Wang J, Schwartz JL, Mao L, Wong DT, Zhou X. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics*. 2008; 9:69. [PubMed: 18254958]
46. Ziober AF, Patel KR, Alawi F, Gimotty P, Weber RS, Feldman MM, Chalian AA, Weinstein GS, Hunt J, Ziober BL. Identification of a gene signature for rapid screening of oral squamous cell carcinoma. *Clin Cancer Res*. 2006; 12(20 Pt 1):5960–5971. [PubMed: 17062667]
47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; 25(1):25–29. [PubMed: 10802651]
48. Jain M, Nilsson R, Sharma S, Madhusudhan N, Kitami T, Souza AL, Kafri R, Kirschner MW, Clish CB, Mootha VK. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science*. 2012; 336(6084):1040–1044. [PubMed: 22628656]
49. Govardhan KS, Ramyasri K, Kethora D, Ravishakar Y, Prasenjit M. Harnessing impaired energy metabolism in cancer cell: small molecule- mediated ways to regulate tumorigenesis. *Anticancer Agents Med Chem*. 2011; 11(3):272–279. [PubMed: 21434854]
50. Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc*. 2004; 99(465):96–104.
51. Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics*. 2005; 21(12):2921–2922. [PubMed: 15817688]

Figure 1.

Comparing the size and power of the methods. Rows: different levels of correlation between null genes in both tests. Columns: different numbers of non-null genes. N1: number of non-null genes in list 1; N2: number of non-null genes in list 2. Vertical line: expected number of overlaps when the non-null genes are independently drawn; horizontal line: the alpha level 0.05 as a reference for examining the size of the tests. The total number of genes is 4000.

For genes that are null in both tests, their test statistic was generated from a bivariate normal

distribution with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance-covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Otherwise the null test statistic was simulated from the standard normal distribution; the non-null statistic was simulated from a normal distribution with mean of -3 and standard deviation of 0.5 . The p-values were generated from the test statistics using standard normal CDF.

Figure 2. Scatter plots of the lfdr values using color representation of smoothed density between experiment comparison.

Figure 3. Scatter plots of the lfdr values using color representation of smoothed density within experiment comparison.

Figure 4.
Heatmap of the test statistic between pairs of datasets using all genes.

Figure 5.

Testing the association between pairs of experiments using the gene list of GO term GO:0006520 – cellular amino acid metabolic process. Significance levels of between-dataset associations are presented as a heatmap. P-value cutoff is 0.001.

Figure 6.

Testing the association between pairs of experiments using the gene list of GO term GO:0000228 – nuclear chromosome. Significance levels of between-dataset associations are presented as a heatmap. P-value cutoff is 0.001.

Table 1

Test results of the prostate cancer data*.

Between experiments, normal v.s. primary:			
	Lapointe	Tomlins	Varambally
Yu	5.62×10^{-11}	0.951	0.000611
	0.00011	0.524	0.00461
	1.07×10^{-6}	0.777	0.00582
	<0.001	0.094	<0.001
Lapointe		0.978	3.06×10^{-8}
		0.205	4.22×10^{-4}
		0.951	0.00101
		<0.001	<0.001
Tomlins			0.478
			0.908
			0.863
			<0.001
Between experiments, primary v.s. metastasis:			
	Lapointe	Tomlins	Varambally
Yu	0.0678	0.862	1.89×10^{-7}
	0.156	0.384	0.0335
	0.0366	0.592	0.00975
	<0.001	0.167	<0.001
Lapointe		0.0788	2.11×10^{-4}
		0.565	0.157
		0.653	0.0483
		0.112	<0.001
Tomlins			0.724
			1
			0.907
			0.032
Within experiment, between normal v.s. primary and primary v.s. metastasis:			
Yu	Lapointe	Tomlins	Varambally
0.0118	5.37×10^{-17}	1.54×10^{-6}	0.561
0.32	4.69×10^{-10}	0.00235	0.723
0.194	9.9×10^{-11}	0.00353	0.724
0.012	<0.001	<0.001	0.03

* In every box, there are four p-values. From top to bottom: the l_q-based test, test of contingency table with FDR cutoff of 0.2, test of significance of Pearson's correlation coefficient, and test of ranked list. P-values ≤ 0.01 are colored red.

Table 2

The balanced accuracy (average of sensitivity and specificity) of predicting whether two datasets belong to the same cancer type using their p-value agreement of GO slim terms (cutoff = 0.001).

GO term	name	All datasets	Excluding breast cancer [*]
Top terms:			
GO:0006520	cellular amino acid metabolic process	0.763	0.884
GO:0005739	mitochondrion	0.783	0.867
GO:0008289	lipid binding	0.786	0.859
GO:0048646	anatomical structure formation involved in morphogenesis	0.766	0.858
GO:0006629	lipid metabolic process	0.78	0.852
GO:0007155	cell adhesion	0.734	0.852
GO:0008233	peptidase activity	0.754	0.848
GO:0016491	oxidoreductase activity	0.76	0.848
GO:0008092	cytoskeletal protein binding	0.749	0.844
GO:0007010	cytoskeleton organization	0.757	0.837
Bottom terms:			
GO:0007165	signal transduction	0.634	0.682
GO:0043167	ion binding	0.643	0.682
GO:0006464	protein modification process	0.64	0.678
GO:0006950	response to stress	0.631	0.677
GO:0051301	cell division	0.62	0.677
GO:0016746	transferase activity, transferring acyl groups	0.623	0.675
GO:0004871	signal transducer activity	0.637	0.668
GO:0007067	mitosis	0.611	0.663
GO:0005886	plasma membrane	0.611	0.651
GO:0000228	nuclear chromosome	0.609	0.649

* Rows are ordered based on the balanced accuracy values excluding the breast cancer datasets.