BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 3 2008, pages 374–382
doi:10.1093/bioinformatics/btm620

Gene expression

A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments

Fangxin Hong[1,*,†] and Rainer Breitling[2]

[1]Department of Biostatistics, Division of Information Sciences, City of Hope National Medical Center, Beckman Research Institute, 1500 Duarte Rd, Duarte, CA 91010, USA and [2]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands

Received on June 8, 2007; revised on December 4, 2007; accepted on December 8, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: The proliferation of public data repositories creates a need for meta-analysis methods to efficiently evaluate, integrate and validate related datasets produced by independent groups. A t-based approach has been proposed to integrate effect size from multiple studies by modeling both intra- and between-study variation. Recently, a non-parametric 'rank product' method, which is derived based on biological reasoning of fold-change criteria, has been applied to directly combine multiple datasets into one meta study. Fisher's Inverse χ^2 method, which only depends on *P*-values from individual analyses of each dataset, has been used in a couple of medical studies. While these methods address the question from different angles, it is not clear how they compare with each other.

Results: We comparatively evaluate the three methods; *t*-based hierarchical modeling, rank products and Fisher's Inverse χ^2 test with *P*-values from either the *t*-based or the rank product method. A simulation study shows that the rank product method, in general, has higher sensitivity and selectivity than the *t*-based method in both individual and meta-analysis, especially in the setting of small sample size and/or large between-study variation. Not surprisingly, Fisher's χ^2 method highly depends on the method used in the individual analysis. Application to real datasets demonstrates that meta-analysis achieves more reliable identification than an individual analysis, and rank products are more robust in gene ranking, which leads to a much higher reproducibility among independent studies. Though *t*-based meta-analysis greatly improves over the individual analysis, it suffers from a potentially large amount of false positives when *P*-values serve as threshold. We conclude that careful meta-analysis is a powerful tool for integrating multiple array studies.

Contact: fxhong@jimmy.harvard.edu

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

1 INTRODUCTION

High-throughput microarray technology has become a popular tool for large-scale comparative analysis of gene expression profiles. As a result, there are rapidly growing collections of publicly available datasets that can be used for subsequent analysis (Moreau *et al*., 2003). However, direct comparison among heterogeneous datasets is not possible due to the complicated experimental variables embedded in array experiments (Irizarry *et al*., 2005; Kuo *et al*., 2002). Meta-analysis, which consists of a set of statistical techniques to combine results from several studies, appears to be a good and practical solution. Recently, its applicability to microarray data was demonstrated by different groups. Rhodes *et al*. (2002) applied meta-analysis to combine four datasets on prostate cancer to determine genes that are differentially expressed between clinically localized prostate and benign tissue. Parmigiani *et al*. (2004) performed a cross-study comparison of gene expression for the molecular classification of lung cancer. Park and Stegall (2007) combined publicly available and their own microarray datasets to investigate the detection of cytokine gene expression in human kidney. Meta-analysis has been shown to have increased statistical power to detect small but consistent effects that might be false negatives in the individual analyses (Choi *et al*., 2003). It also has significantly improved reproducibility when compared with independent studies, which may lead to improved reliability (Hong *et al*., 2006). Therefore, meta-analysis provides researchers with an indispensable tool to interrogate existing databases for candidate biomarkers and biological pathways.

Meta-analysis may be broadly defined as the quantitative review and synthesis of the results of related but independent studies (Normand, 1999). The objectives include increasing power to detect an overall treatment effect and assessment of the amount of variability between studies. The former is the common task of most microarray experiments, which aim at detecting differentially expressed genes among multiple conditions (control versus treatment). Since the early days, many simple and sophisticated statistical methods have been proposed for this purpose in the individual experiments (Breitling *et al*., 2004; Efron *et al*., 2001; Newton *et al*., 2004; Tusher *et al*., 2001), and their performance has been thoroughly

[*]To whom correspondence should be addressed.
[†]Present address: Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA.

compared (Pan, 2002). However, complicated experimental variables and different platforms used in microarray experiments lead to more statistical issues than encountered in standard meta-analysis, thus most standard methods can not be applied directly for combining array datasets.

In recent years, several meta-analysis methods have been proposed using different approaches. The most straightforward one is Fisher's Inverse χ^2 test (Fisher, 1925), which computes a combined statistic from the P-values obtained from the analysis of the individual datasets, $S = -2\log(\Pi_i P_i)$, where S follows a χ^2 distribution with $2I$ degrees of freedom under the joint null hypothesis. This method is easy to use and does not require additional analysis. However, by working with the P-values, it is impossible to estimate the average magnitude of differential expression.

Choi *et al.* (2003) adopted the classic biostatistical meta-analysis framework (Normand, 1999) in array analysis and used a t-like statistic (defined as effect size) as the summary statistic for each gene from each individual dataset. They then proposed a hierarchical modeling approach to assess both intra- and inter-study variation in the summary statistic across multiple datasets. This model-based method estimates an overall effect size as the measurement of the magnitude of differential expression for each gene through parameter estimation and model fitting. The approach has been implemented into a Bioconductor (Gentleman *et al.*, 2004) package *GeneMeta* that facilitates its application.

Recently, the non-parametric rank product (RP) method has been introduced in another Bioconductor package (*RankProd*) (Hong *et al.*, 2006) to identify differentially expressed genes, which has direct applicability in meta-analysis. It is based on the rank product method of detecting differentially expressed genes (Breitling *et al.*, 2004) and offers several advantages over linear models or t-tests, including a biologically intuitive fold-change (FC) criterion, fewer assumptions and better robustness, which leads to increased power in low sample number and/or large noise settings (Breitling and Herzyk, 2005). Both the t-based and the RP method utilize permutation tests to assess the statistical significance, reporting the false discovery rate (FDR) of the identification based on combined statistics. And both of them generate P-values which can also serve as input for Fisher's Inverse Chi- square test.

In this article, we comparatively evaluate the three methods; t-based hierarchical modeling using *GeneMeta*, the rank product method with *RankProd* and Fisher's Inverse χ^2 test with P-values from the individual analysis of a single dataset with each of the first two methods. ROC curves and pAUC (Pepe, 2000) are utilized in a simulation study to quantify the sensitivity and specificity of each method. We also apply the methods to two sets of experimental microarray studies, one with relatively small and one with relatively large between-study variation. And we address the performance based on two main criteria: (1) reproducibility measured by a CAT plot (Irizarry *et al.*, 2005) and (2) identification power measured by integration-driven discovery rate (Choi *et al.*, 2003). We also briefly discuss other available meta-analysis methods, such as Bayesian approaches (Wang *et al.*, 2004), linear models (Ghosh *et al.*, 2003) and rank aggregation (DeConde *et al.*, 2006).

2 METHODS

Let T and C stand for two experimental conditions (treatment versus control), and let there be $i = 1, \ldots, I$ independent studies (datasets) and (n_{iT}, n_{iC}) replicates for the ith study. Thus, for a given gene, the data is recorded as T_{ij}/C_{ij} which is the (logged) gene expression level in the jth replicate in treatment/control of the ith study. In the following sections, we briefly describe the three methods.

2.1 The *t*-based modeling approach

The t-test and its variations are the most widely used approaches in array analysis to identify differentially expressed genes. Meta-analysis based on the t-statistic was reviewed by Normand (1999) in the context of biostatistical applications, and it was adopted for microarray analysis recently (Choi *et al.*, 2003). Briefly, a standardized mean difference was obtained as an effect size index for the measurement of differential expression of a gene in any given study.

$$d_i = \frac{\bar{T}_i - \bar{C}_i}{S_p} \tag{1}$$

$\bar{T}_i$ and $\bar{C}_i$ represent the means of treatment and control group in the ith study, and S_p indicates the estimated variation. Then we model the effect size index d_i across studies by a hierarchical model:

$$\begin{aligned} d_i &= \theta_i + \varepsilon_i, \varepsilon_i \sim N(0, s_i^2) \\ \theta_i &= \mu + \delta_i, \delta_i \sim N(0, \tau^2) \end{aligned} \tag{2}$$

where μ denotes the parameter of interest (treatment effect), and τ^2 and s_i^2 represent the between-study and within-study variation. The model has two different forms: a fixed-effect model (FEM) and a random effect model (REM), depending on whether between-study variation is ignorable. Choi *et al.* (2003) suggest to use Cochran's Q statistic (Cochran, 1954) to test homogeneity of study effect, which is assessing the hypothesis that τ^2 is zero. Failure to reject the null hypothesis should indicate the appropriateness of the FEM. Otherwise REM will be used instead, where the estimator by DerSimonian and Laird (1986) is used to estimate τ^2. Then this estimate is submitted to estimate μ and its variance $Var[\mu]$ by a point estimator

$$\hat{\mu}(\tau^2) = \frac{\sum(s_i^2 + \tau^2)^{-1} d_i}{\sum(s_i^2 + \tau^2)^{-1}}, Var[\hat{\mu}(\tau^2)] = \frac{1}{\sum(s_i^2 + \tau^2)^{-1}}$$

A Z-score will be derived from $\hat{\mu}(\tau^2)$ and $Var[\hat{\mu}(\tau^2)]$ to assess the standardized average treatment effect for each gene across studies, z_g, $g = 1, \ldots, G$.

To assess the statistical significance of the combined results, one would obtain P-values from a standard normal distribution $N(0,1)$ using these Z-scores. However, it is preferred in array analysis that permutation is used instead, to account for small sample size and to avoid the violation of the normality assumption. In this case, column-wise permutation is performed within each study to create randomized data and z-scores under the null distribution, z_g^{*b} for permutation $b = 1, 2, \ldots, B$. The ordered statistics $z_{(g)}$ ($z_{(1)} \leq \ldots \leq z_{(p)}$) and $z_{(g)}^{*b}$ ($z_{(1)}^{*b} \leq \ldots \leq z_{(p)}^{*b}$) were obtained, and the FDR was estimated for a given gene g by

$$FDR_g = \frac{(1/B) \sum_b \sum_{(g)} I(|z_{(g)}^{*b}| \geq z_g)}{\sum_{(g)} I(|z_{(g)}| \geq z_g)}$$

where $I(\bullet)$ is the indicator function. In the package *GeneMeta*, the FDR estimation is carried out for up-regulation, down-regulation and two-side comparison, respectively. Based on similar reasoning, we can also extend the permutation to compute P-values as follows;

$$P_g = (1/GB)\sum_b \sum_{(g)} I(|z^{*b}_{(g)}| \geq z_g)$$

where G is the total number of genes under study.

2.2 The rank product approach

The rank product is a non-parametric statistic that was first proposed to detect differentially expressed genes in a single dataset (Breitling *et al*., 2004). It is derived from biological reasoning about the fold-change (FC) criterion and detects genes that are consistently found among the most strongly up-regulated (or down-regulated) genes in a number of replicate experiments. Moreover, the method offers a natural way to overcome the heterogeneity among multiple datasets and therefore can be extended to meta-analysis, which generates a single significance measurement for each gene in the combined study (Hong *et al*., 2006).

Here we describe the rank product meta-analysis algorithm using two datasets as an example with (n_{1T}, n_{1C}) and (n_{2T}, n_{2C}) replicates, respectively.

(1) For a one-channel array, compute pair-wise ratios or fold-changes within each dataset; T_{1j}/C_{1l}, $j = 1, \ldots, n_{1T}$, $l = 1, \ldots, n_{1C}$ $\Rightarrow K_1 = n_{1T} \times n_{1C}$ comparisons and T_{2j}/C_{2l}, $j = 1, \ldots, n_{2T}$, $l=1, \ldots, n_{2C} \Rightarrow K_2 = n_{2T} \times n_{2C}$ comparisons. (For two-channel arrays, T_{1j}/C_{1j}, $j = 1, \ldots, n_1$ and T_{2j}/C_{2j}, $j = 1, \ldots, n_1$ with $n_{iT} = n_{iC} = n_i$, so $K_1 = n_1$ and $K_2 = n_2$.)

(2) Rank fold-change (FC) within each comparison (largest FC $\Rightarrow$ rank 1)$\Rightarrow$ r_{gik}: rank of gene g in ith study under kth comparison, $k = 1, \ldots, K_i$.

(3) Combine K_1 ranks from dataset 1 and K_2 ranks from dataset 2, determine rank product for each gene as $RP_g = (\prod_i \prod_k r_{gik})^{(1/K)}$ where $K = K_1 + K_2$.

(4) Independently permute expression values within each single array relative to gene ID, repeat step (1)–(3) and obtain the null rank product statistic $RP^{*(b)}_g$.

(5) Repeat step (4) B times and form a reference distribution with $RP^{*(b)}_g$, determine P-value and FDR associated with any given gene g similarly as the one used in the t-based modeling approach.

$$P_g = (1/GB)\sum_b \sum_{(g)} I(|RP^{*b}_{(g)}| \leq RP_{m=g})$$

$$FDR_g = \frac{(1/B)\sum_b \sum_{(g)} I(|RP^{*b}_{(g)}| \leq RP_g)}{\sum_{(g)} I(|RP_{(g)}| \leq RP_g)}$$

One-channel experiments, for the purpose of this discussion, include Affymetrix gene-chips and two-color cDNA arrays with reference design; direct two-color cDNA arrays are usually two-channel experiments. The permutations are done by permuting the expression value (ratio for two-channel experiments) within each array, rather than by permuting the samples across arrays as in the t-based approach. Basically, the algorithm computes pairwise FC with replicates for each gene between treatment and control in both directions, respectively, and transforms FC into rank among all genes under study, then searches for genes that are consistently top ranked across replicates. Converting FC into ranks increases robustness against noise and heterogeneity across studies. Indeed, a recent study (Yuen *et al*., 2002) found that, although the fold-changes of differentially expressed genes had poor consistency across array platforms, the rank orders were comparable. This method is also implemented in a Bioconductor package (*RankProd*).

2.3 Fisher's inverse Chi-square approach

Moreau *et al*. (2003) reviewed several simple methods (called omnibus procedures; Hedges and Olkin, 1985) that are available to test the statistical significance of combined results based on P-values from each single study. One method (Tippet, 1931) is to take the minimum P-value ($p_{\min}$) for each gene observed over different datasets but test this minimum P-value at a higher stringency than the single study rejection threshold α: reject 'no differential expression' if $p_{\min} < 1\text{-}(1 - \alpha)^{(1/I)}$. This method is sensitive to outliers, so a variant uses the nth smallest P-value as an alternative (Wilkinson, 1951).

Another method is Fisher's Inverse χ^2 test. It computes a combined statistic from the P-values obtained from the individual datasets,

$$S = -2\log(\Pi_i p_i)$$

which follows a χ^2 distribution with $2I$ degrees of freedom under the joint null hypothesis and thus P-values of the combined statistic can be calculated. Some researchers also extended Fisher's method by giving different weights to P-values from each dataset (Good, 1955). Weight assignment can depend on the reliability of each P-value as a result of data quality or on other factors considered important. To address the issues of losing power due to a single very poor entry, Zaykin *et al*. (2002) proposed a truncated product method (TPM) to calculate Fisher's product using a thresholding criterion where only P-values less than or equal to some pre-specified cutoff value τ contribute to the combined product,

$$S^z = -2\log(\Pi_i P_i^{I(P_i \leq \tau)})$$

In addition to reducing false negatives, TPM also guards against false positives by requiring the presence of at least one significantly small P-value, which is rooted in the concern that a combination of marginally significant P-values might suggest unreasonably high significance of the combined statistic (Rosenthal, 1991). Pyne *et al*. (2006) proposed a conservative way of controlling the combined FDR at a specified level α by thresholding each experiment at FDR level α' ($\alpha' \leq \alpha^2/4I^2$) with experiment-specific P-value cutoffs $\tau_{i,\,\alpha'}$ following the procedures by Benjamini and Hochberg (1995) or Storey (2002). One aspect we need to point out is that Fisher's product should be applied to P-values for up-regulation and down-regulation separately, as random opposite expression differences in a small sample setting would result in marginally small P-values that lead to a high significance of the combined statistic.

It is easy to notice that both the t-based and the rank product approach can be applied to an individual dataset and P-values from such individual analyses from either of them can be used in Fisher's product method. In this study, we will apply Fisher's product using P-values obtained from regular individual analyses with either t-based or rank product approach, respectively. Since it is hard to threshold with only two P-values available in our examples (Section 3), we will use the original form of Fisher's product as the test statistic and derive P-values for the combined results from the null distribution. The outcome will then be compared to that from meta-analysis with a t-based or rank product approach.

2.4 Evaluation

We evaluate the performance of the above three methods by comparing their power of detecting differential expression (sensitivity) and reliability (specificity). We adopted receiver operating characteristic (ROC) curves and associated partial area under the curve (pAUC) (Pepe, 2000) in the simulated dataset, and Correspondence At the Top (CAT) plots (Irizarry *et al*., 2005) in the real-data applications, for which the true differentially expressed genes are unknown. The ROC curve is primarily a descriptive device displaying the range of trade-offs between false positive rates and false negative rates for a given test. The closer to the upper left-hand corner of the ROC space, the better the performance of the given test. pAUC is the area under the ROC curve in a restricted range of false-positive rates, often used as summary index of test accuracy within a practical region of false positive rates.

CAT plots are a reliability assessment tool introduced by Irizarry *et al.* (2005), which assesses the agreement of the identification among studies. Genes identified in multiple independent studies are likely to be the truly significant ones, thus high reproducibility among independent studies suggests a high specificity. It has been shown that correlation or scatter plots of $\log_2$-fold changes are poor measurements of the agreement among studies, as they are heavily influenced by large amounts of non-differentially expressed genes. In practice, we are only interested in a small subset of genes that appear to be differentially expressed. Therefore, it is more important to assess agreement for genes that are likely to be called significant (Irizarry *et al.*, 2005). The procedure for creating a CAT plot is to make a list of n top candidate genes for each of the two studies, which can be individual or meta studies, and plot the proportion in common against the list size n. In other words, one plots the proportion of the top n genes identified in one study that are 're-discovered' in the top n genes of another study, hence the alternative designation as 'plot of rediscovery rate'. We have three independent datasets in our application below, thus we will perform meta-analysis with two of them and compare the results with that from the third dataset (reference study) to draw CAT plots. In addition, the proportion in common will be calculated among any two of three individual analyses. The average proportion will be included in the plot as the performance of the regular individual analysis.

We will use the integration-driven discovery (IDD) to measure the extra power offered by integrating multiple datasets (Choi *et al.*, 2003). IDD was originally defined using cutoffs in Z-score and we modified it to P-values in order to accommodate the outcomes from all three methods. As we will generate P-values for both up-regulation and down-regulation for each gene, we use $(p_i^{\text{up}}, p_i^{\text{dn}})$ and $(p^{\text{up}}, p^{\text{dn}})$ to denote the significance in each comparison for the individual study i and combined meta-analysis. The IDD is defined as

$$(p^{\text{up}} \le p_{\text{th}}) \text{ and } (p_i^{\text{up}} > p_{\text{th}} \text{ for } i = 1, \ldots, I)$$

or

$$(p^{\text{dn}} \le p_{\text{th}}) \text{ and } (p_i^{\text{dn}} > p_{\text{th}} \text{ for } i = 1, \ldots, I)$$

IDD is the number of extra genes identified by meta-analysis compared with the union set of all individual studies at the same p-threshold level (p_{th}). Integration-driven discovery rate (IDR), defined as the ratio of IDD to the total number of discoveries, will be listed for a series of small P-value thresholds for each method. However, a low IDR, which suggests a subtly increased power, might be due to a relatively high power of the given method in the individual studies, which indeed indicates *better* performance. Moreover, a high IDD might lead to a potentially decreased specificity, as more false positives might appear with more genes identified, particularly as the p-threshold increases. Therefore, we consider the measurement of 'reproducibility' or reliability by CAT plot a better evaluation criterion.

3 RESULTS

3.1 Simulation

To evaluate the performance of the three meta-analysis methods, we first simulated expression levels of $G = 5000$ genes under two conditions T(treatment) and C(control) from 3 ($I = 3$) independent studies based on a t-based model as

Table 1. pAUC from the false positive range 0–0.05 for all six senarios presented in simulation studies. 'F-meta' stands for meta-analysis using Fisher's product with P-values from either of the two methods

$\tau_g^2 \sim IG(a, b)$		a = 3, b = 0.4 (LV)			a = 3, b = 0.04 (SV)		
Sample Size		K = 3	K = 5	K = 10	K = 3	K = 5	K = 10
Single	RP	0.028	0.027	0.027	0.032	0.040	0.039
	t-based	0.019	0.021	0.027	0.032	0.032	0.038
Meta	RP	0.034	0.033	0.033	0.039	0.039	0.038
	t-based	0.027	0.026	0.034	0.039	0.040	0.037
F-meta	RP	0.034	0.033	0.033	0.039	0.039	0.038
	t-based	0.033	0.022	0.005	0.039	0.038	0.034

$$\begin{aligned} C_{gik} &= \alpha_g + \gamma_{gi} + \varepsilon_{gik}, \\ T_{gik'} &= \alpha_g + \gamma_{gi} + \beta_{gi} + \varepsilon_{gik'} \end{aligned} \tag{3}$$

and

$$\beta_{gi} = \mu_g + \delta_{ig}$$

where C_{gik} and $T_{gik'}$ are logged expression levels of gene g in the ith study under control ($k = 1, \ldots, K_1$ replicate) and treatment ($k' = 1, \ldots, K_2$ replicate). For simplicity, we let $K_1 = K_2 = K$. α_g is the mean expression under control, and γ_{gi} represents its variation among studies. β_{gi} stands for the expression difference in the ith study, which contains a true difference μ_g and its variation δ_{ig} among studies. We notice that γ_{gi} can be ignored as it will be canceled in both t-based and rank product analysis.

In order to mimic experimental microarray studies, we generated data (C_{gik}, T_{gik}) based on rough parameter estimates using a point estimator from the experimental data used in the following sections. For example, since the mean logged expression for all genes ranged from 3–13 in the hormone data (Section 3), we simulated α_g from *Unif*(3,14); we randomly selected 10% true differentially expressed genes with $\mu_g \sim Unif(-3, 3)$. We allow δ_{ig} to have gene-specific between-study variation, $\delta_{ig} \sim N(0, \tau_g^2)$, and use an inverse gamma distribution to generate $\tau_g^2 \sim IG(a, b)$. We also simulated within-study error ε_{gik}, $\varepsilon_{gik'}$ from normal distributions with different error variation in three independent studies as $N(0, s_i^2)$, $s_i^2 = 0.03$, 0.05, 0.08. To explore the effect of between-study variation (τ_g^2) and sample size (K), we simulated 6 scenarios listed in the Table 1, with (1) relatively large variation (LV) $a = 3$, $b = 0.2$, $E(\tau_g^2) = 0.2$, $\text{Var}(\tau_g^2) = 0.04$ and (2) relatively small variation (SV) $a = 3$, $b = 0.02$, $E(\tau_g^2) = 0.02$, $\text{Var}(\tau_g^2) = 0.0004$ and three settings of sample size, K=3, 5,10, which covers the range where meta-analysis is likely to be most beneficial.

For each scenario, we applied the rank product and t-based method to each of the three datasets and calculated average results as the performance of the individual analysis. We then selected two datasets at a time [$C(3, 2) = 3$ times] for meta-analysis and also got average outcomes from three meta-analyses. By comparing the identifications with the true differentially expressed genes, we plotted the ROC

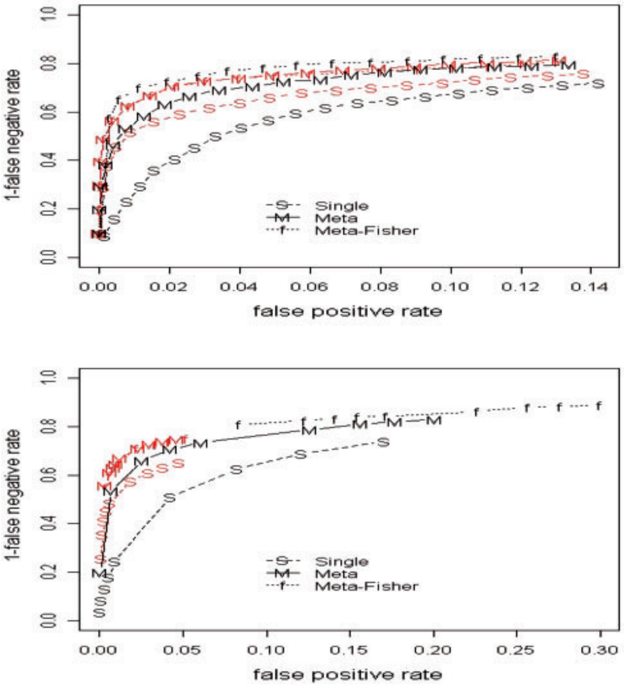

Fig. 1. Partial ROC curves for simulated datasets (K=3, LV) with rank product (red) and t-based method (black) for the top 1000 identified genes (top) and for a series of selected P-value thresholds 0.0001,. . .,0.01 (bottom). '-s-': individual analysis; '-M-': meta-analysis with rank product/t-based approach; '-f-': meta-analysis using Fisher's product with P-values from either of the two methods. Note the '-f-' and '-M-' in red are overlapped.

curves for the top n = 50, 100,. . ., 1000 identified genes (Supplementary Fig. 1) and for a series of selected P-value thresholds 0.0001. . .0.01 (Supplementary Fig. 2) in all six scenarios. Figure 1 shows an example (K = 3, LV). pAUC values within the standard false positive rate region (0–0.05) are summarized for all six scenarios in Table 1.

The series of ROC plots and pAUC summaries in Table 1 highlights several findings. First, as expected, meta-analysis with either method achieves better outcomes compared to single analysis, suggesting increased power and specificity. However, the improvement becomes less predominant as τ_g (between-study variation) decreases and K (sample size) increases. Second, rank products outperform the t-based method with a better tradeoff between sensitivity and specificity, especially in single analysis and in the low false positive rate region. Again, the advantage gets smaller when K increases and/or τ decreases. Third, the t-based method generates inflated P-values (Supplementary Fig. 3) leading to high false-positive rates even at small P-value thresholds, as indicated by the shift of the ROC curve to the right. P-values from rank products appear to be a more reliable measurement of significance. Finally, the performance of Fisher's method highly depends on the quality of P-values from the simple individual analysis. The inflation of P-values escalates when applying Fisher's method with P-values from the t-based method, which leads to an unacceptably large number of false discoveries in Fisher's method.

3.2 Controlled human array data

This dataset was originally presented by Irizarry *et al.* (2005) as the controlled experiment for a multiple-laboratory comparison of three different microarray platforms: Affymetrix Genechips, two-color spotted cDNA arrays, and two-color long oligonucleotide arrays. Two RNA samples were created in which only a few genes were expected to be differentially expressed. Two technical copies were made for each of them, given to researchers in 10 labs from the Washington, DC–Baltimore area, and processed according to what was considered best practice in each lab. We selected data from lab 1, 2 and 3 out of 5 labs which utilized Affymetrix gene-chips platform as the testing datasets. We expect both intra-study and between-study variations to be small due to the technically replicated RNA sample and the same platform being utilized in all three labs.

Similar to the simulation study, we performed an individual analysis for each of three datasets, and seleted two datasets at a time [$C(3, 2)$=3 times] for meta-analysis. The average outcomes from three individual and three meta-analyses are used in the comparison. A Q-test indicated that an FEM is appropriate in the t-based method, which confirms that the inter-study variation is ignorable. As we only expect very few differentially expressed genes, we use several low P-values as the cutoff point in Table 2 and list the number of genes identified in the individual and meta-analysis as well as the integration-driven discovery rate (in parenthesis). As shown in Table 2, meta-analysis is able to identify more genes at the same p-level, suggesting an increased power and a potentially low false negative rate. However, unexpectedly large amounts of genes are identified by t-based methods, particularly with Fisher's product, making it suspicious of potentially high false positive rates as already indicated in the simulation study. We notice that rank products identified a significant amount of genes at low FDR level (<0.05), while the model-based method failed to identify any genes at the same FDR level. Due to the small scale of between-study and within-study variation, unsurprisingly, most of the genes identified in the individual studies were confirmed or re-identified in meta-analysis. Furthermore, we see higher significance (smaller P-values) in meta-analysis for genes with consistent but small changes (data not shown), suggesting an increased confidence regarding the identification.

Treating the top n genes identified in the third dataset as the reference, we compared the top n genes identified in dataset 1 and 2, as well as the top n genes from meta-analysis combining dataset 1 and 2. The percentage in common among the top genes is presented in CAT plots. Since we would only expect a small number of differentially expressed genes, the CAT plots (Fig. 2) are drawn for the top 400 genes (200 up-regulated and 200 down-regulated) identified from each analysis. Meta-analysis, in general, gains higher reliability compared to single analysis, suggesting that the result is more likely to be reproduced by an independent study. However, the rediscovery rate is much higher for rank products compared to that of the t-based method for both individual and meta-analysis. For example, the rediscovery rate is above 60% in all analyses with rank products, while it is below 50% with the t-based method. This is consistent with the simulation study,

Table 2. Controlled data set: average number of genes identified at different *P*-levels and integration-driven discovery rate (IDD) in meta-analysis (shown in parenthesis) using rank product (Meta-RP) and *t*-based approach (Meta-*t*). The results with Fisher's product are listed as 'Meta-F'

P	Rank Product			*t*-based		
	Single	Meta-RP	Meta-F	Single	Meta-*t*	Meta-F
1e-4	47	166 (0.72)	150 (0.69)	7	87 (0.98)	496 (0.99)
5e-4	100	233 (0.58)	221 (0.55)	39	229 (0.92)	721 (0.94)
1e-3	137	281 (0.52)	259 (0.48)	75	332 (0.84)	839 (0.89)
5e-3	280	449 (0.40)	424 (0.36)	280	672 (0.62)	1243 (0.73)

Table 3. Plant hormone data set

P	Rank Product			*t*-based		
	Single	Meta-RP	Meta-F	Single	Meta-*t*	Meta-F
0.001	136	260 (0.5)	247 (0.47)	37	94 (0.98)	74 (0.64)
0.005	292	447 (0.40)	417 (0.35)	159	295 (0.92)	220 (0.51)
0.01	424	566 (0.32)	550 (0.29)	317	491 (0.83)	368 (0.39)
0.05	1132	1039 (0.13)	999 (0.10)	1425	1352 (0.38)	1105 (0.11)

'single': Union set of genes identified from the individual analyses.

where rank products yield more robust gene ranking for the top genes, leading to higher reproducibility and increased specificity. This might be due to the robustness of rank products against noise. Surprisingly, Fisher's product appears to have higher reproducibility than the *t*-based method, suggesting that meta-analysis using a *t*-based model might add another level of instability, which is prevented by Fisher's approach.

3.3 Plant hormone data

Here we meta-analyzed three array experiments that have been carried out in two laboratories (Shimada and Chory) to study the effect of a particular hormone on plant growth. Each of the three studies compares gene expression profiles at 3 h after hormone treatment to non-treatment control plants (Vert *et al.*, 2005). Shimada's group in Japan first conducted two very similar experiments, each with two replicates; one with the Affymetrix 8K GeneChip, representing 1/3 of the Arabidopsis genome, and one with Affymetrix ATH1 arrays, representing (approximately) the whole genome. Chory's group in the USA reported a third similar experiment using the Affymetrix ATH1 array with three replicates for each condition (Nemhauser *et al.*, 2004). All three datasets were preprocessed using gcRMA (Wu *et al.*, 2004) in Bioconductor and ~7000 common genes were extracted and used in the evaluation. Quality checks (not shown) indicate severe heterogeneity among the three datasets and a strong 'lab effect' as the two datasets from Shimada's group, although using different types of GeneChip, are much more similar to each other than to the data from Chory's group. To reduce the disruption by lab effect in computing the rediscovery rate, we used the two datasets from Shimada's group to practice meta-analysis and treat the Chory data as the reference. A *Q*-test for heterogeneity found that the REM was appropriate when the *t*-based method is used for this dataset.

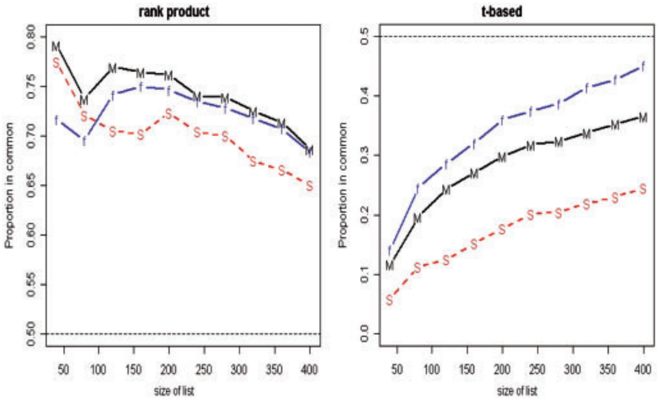

Fig. 2. CAT plot of the controlled data with rank product (left panel) and *t*-based approach (right panel). '-s-'(red): individual analysis; '-M-'(black): meta-analysis with rank product/*t*-based approach; '-f-' (blue): meta-analysis using Fisher's product with *P*-values from each of the two methods. The dotted horizontal line indicates 50% agreement.

While reporting integration driven discoveries in Table 3, we can also see that a larger percentage of discoveries in the simple analyses are not identified by meta-analysis using the *t*-based method compared to the rank product method. This indicates a potentially high false positive rate and/or low robustness in gene ranking of the *t*-based approach. For example, only 26% genes were re-discovered in meta-analysis at $p = 0.01$ level and even worse for lower *p*-thresholds. Similarly to the controlled human data, rank products identified more genes at the lower *p*-levels. Together, this again suggests that rank products have higher selectivity and sensitivity than the *t*-based method in case of small sample size and large noise. Since, biologically, plant hormones affect plant growth in a global way, we expect expression level changes for a large number of genes, so we chose to expand the re-discovery rate comparison to the top 1000 up-regulated and 1000 down-regulated genes. Similar to Figure 2, figure 3 confirmed the increased re-discovery rate with meta-analysis in both up- and down-regulated gene sets, and rank products have much higher reproducibility than the *t*-based method in this dataset, too. One should notice the extremely low reproducibility of the *t*-based method in single analysis (essentially just a *t*-test), indicating much higher false positive rates, which is consistent with what we observed in the IDR table (Table 3). Fisher's product appears to have similar performance (slightly worse) as rank products and the *t*-based method in this dataset, when using *P*-values from each method, respectively.

4 DISCUSSION

4.1 A comparative summary of the three methods

It is clear that Fisher's product uses only *P*-values from each single dataset; its performance heavily depends on the

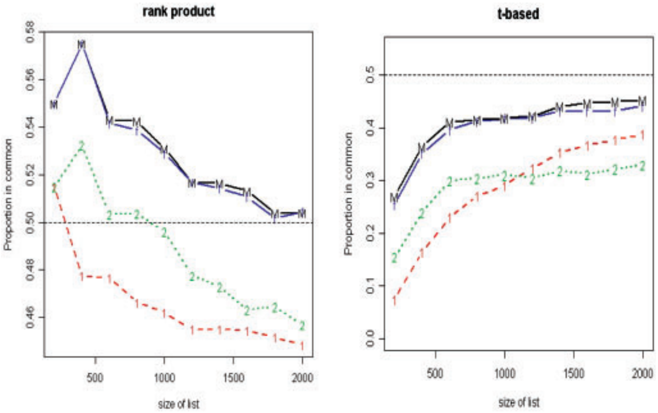

Fig. 3. CAT plot of the plant hormone data with rank product (left) and model-based approach (right). '-1-'(red) and '-2-'(green) shows overlap between dataset 1, 2, and the reference. '-M-'(black): meta-analysis with rank product/t-based approach; '-f-'(blue): meta-analysis using Fisher's product with P-values from each of the two methods. The dotted horizontal line indicates the position of 50% agreement.

underlying method used to calculate p. Therefore, we will focus our discussion on the two methods used to generate the P-values.

Both t-based and rank product approaches are derived or extended from the ones used in simple analysis, therefore they largely inherit their features in simple analysis. The t-based approach originates from Student's t-test and provides a flexible selection of a fixed-effect or random-effect model based on homogeneity tests and an overall measure of differential expression for each gene. The latter feature offers a direct comparison of the magnitude of a treatment on different genes. Rank products, however, do not have this feature but only provide the relative position of a gene compared with all other genes under study for judging its expression difference.

On the other hand, rank products have advantages over t-based approaches in terms of robustness in ranking genes. In most array studies with small sample size, t-based methods suffer from unreliable error estimates, therefore gene ranking substantially varies from experiment to experiment, which causes a low specificity as shown in our simulation, or a low reproducibility in experimental applications. Although increasing sample size would improve the performance of the t-based method as shown in simulation studies, it is uncommon to have large sample size in laboratory biological experiments. The poor reproducibility has been a major concern, discouraging some biologists from trusting the results of array experiments. It is indeed exciting news that combining multiple studies significantly improves reliability (Figs 2 and 3). Importantly, it appears that rank products have consistently the highest reproducibility/specificity in both simulation and experimental data applications, regardless of the scale of heterogeneity among datasets. This observation suggests that gene rankings from the rank product method are more robust against noise and other hidden variables embedded in different datasets.

4.2 Comparison to other meta-analysis approaches

Wang *et al.* (2004) introduced a meta-analysis method from a Bayesian perspective. The basic idea is to treat one dataset as prior knowledge that gives preliminary information about the expression difference for the given gene and then to increase our knowledge by adding another dataset to get an updated posterior assessment of the expression change. Let $D_1 = \bar{T}_1 - \bar{C}_1$ and s_1^2 be the estimate of expression change and its estimated variance from the first dataset. If we assume a normal distribution of the errors in measurement, a well-known Bayesian calculation shows that the best estimate of the true difference after adding the second dataset is given by

$$D_{\text{combined}} = \frac{(D_1/s_1^2) + (D_2/s_2^2)}{(1/s_1^2) + (1/s_2^2)}$$

and the variance is given by

$$\frac{1}{s_{\text{combined}}^2} = \frac{1}{s_1^2} + \frac{1}{s_2^2}$$

In other words, we combine difference measurements from different datasets by weighting them using variance. It is easy to show that this formula generalizes to the scenario of multiple datasets and the final results does not depend on the order in which multiple datasets enter the study, $D_{\text{combined}} = \sum_i (D_i/s_i^2)/\sum_i (1/s_i^2)$ and $1/s_{\text{combined}}^2 = \sum_i (1/s_i^2)$. However, a simple derivation can show that the above method is indeed the maximum likelihood estimator (MLE) for the t-based modeling approach introduced in Section 3 under the fixed-effects model, if we change the effect size d_i to $Y_i = \bar{T}_i - \bar{C}_i$ (Normand, 1999). We consider the standardized (scale-free) statistic d_i more appropriate when datasets are generated from different laboratories, therefore we do not treat the Bayesian approach as a fundamentally different method. Indeed, application to simulated data confirms that the Bayesian approach has very similar outcomes to the t-based method (data not shown).

Ghosh *et al.* (2003) proposed another two t-test based approaches. The first one utilizes a weighted average of the t-statistics T_i from the individual datasets,

$$T = n^{-1} \sum_{i=1}^{I} n_i T_i \tag{4}$$

as the test statistics and obtains statistical significance from permutation. The second algorithm has a more general formula with study effect as main effect as well as interaction with each gene. Let Y_i=$(T_{i1} \ldots T_{in_{iT}}, C_{i1}, \ldots, C_{in_{iC}})$ denote the expression of a given gene in the ith study with a total of $k = 1, \ldots, (n_{iT} + n_{iC})$ samples, the model is written as

$$E[Y_{ik}] = \gamma_{0i} + \gamma_{1i} X_k + \gamma_{2i} Z_i + \gamma_{3i} X_k Z_i \tag{5}$$

where X_k is a covariate for experimental condition (treatment versus control) of the kth sample and Z_i is the study indicator. A likelihood test for testing $H_0 : \gamma_{1i}$=γ_{3i}=0 would yield a set of test statistics (based on least square estimates) for assessing differential expression, and a permutation test is proposed to generate significance measurements. One should notice that model 4 assumes the treatment effect is the same across all the studies, which is similar to the FEM in model 3, while model 5 assumes treatment effect varies between studies, which is the idea of the REM in model 3. As a result, these approaches are so similar to the *t*-based method that they share most of its features illustrated above.

The various rank aggregation approaches of DeConde *et al.* (2006) are based on meta-search methods from computer science, which are used to combine lists of search results. Because the concept works on rank-ordered lists, it will share many of the advantageous performance characteristics of rank products. On test data from five different prostate cancer datasets (DeConde *et al.*, 2006), the performance of rank aggregation and rank products is indistinguishable (not shown). Therefore, rank aggregation can be considered an interesting alternative to rank products and should be further investigated.

5 CONCLUSION

In this article, we compare the performance of three meta-analysis methods for microarray studies, using array data generated on the Affymetrix platform. At the data analysis level, we limited our comparison to the improvement of the detection of differential expression, as this is currently the most common aim of microarray experiments. The heterogeneity among multiple datasets leads to many statistical issues affecting the integration process. Our study shows that meta-analysis achieves increased power and higher reproducibility by integrating multiple datasets. In general, the non-parametric rank product method outperforms the other methods in terms of sensitivity and specificity, especially in a setting of small sample size and large between-study variation. This suggests that rank products should be preferred in such a setting since transferring fold-changes into ranks increases the robustness against variations from different sources. In addition, simulation shows that *P*-values from the rank product approach are a more reliable significance measurement than those from the *t*-based method. Nevertheless, the *t*-based method can achieve dramatic improvements in terms of gene ranking when combining multiple studies together.

Fisher's χ^2 method appears to be highly dependent on the methods used in the individual analysis. It yields rather poor results in combination with *t*-based methods, and performs similar to rank product-based meta-analysis. Therefore, we do not suggest usage of Fisher's method in combining multiple dataset unless only *P*-values are available. Our work also points out that there will be a substantial amount of false positives in the list of genes identified in the individual studies with low sample size and large scale of heterogeneity when a *t*-based method is used, which contributes to the poor consistency among independent studies as reported before by different research groups (Jarvinen *et al.*, 2004; Kothapalli *et al.*, 2002).

The availability of public array repositories opens up a new realm of possibilities for microarray data analysis. An essential challenge is the efficient integration of array data generated by different laboratories and/or different platforms. Currently, there are several (≥ 4) popular array technologies. It is still unclear how measurements from different platforms compare with each other (Moreau *et al.*, 2003), and inconsistencies in gene coverage and annotation make comparisons very difficult. It will be an interesting topic to further explore meta-analysis across different platforms, which is not the scope of this study. As large amounts of data are being produced on a daily basis using a wide variety of experimental designs and technologies, meta-analysis could be beneficial in a much wider range of applications, such as integrating time-course genomic data and proteomic experiments (Varambally *et al.*, 2005). Therefore, we are expecting additional developments of methods for meta-analysis which will significantly enhance our ability to benefit from these powerful high-throughput technologies.

ACKNOWLEDGEMENTS

We thank Dr Jiang Gui for helpful comments and three reviewers for constructive suggestions and comments. This work was supported by the Beckman Research Institute at the City of Hope National Medical Center.

Conflict of Interest: none declared.

REFERENCES

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

Breitling,R. *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Breitling,R. and Herzyk,P. (2005) Rank-based methods as a non-parametric alternative of the *t*-statistic for the analysis of biological microarray data. *J. Bioinf. Comp. Biol.*, **3**, 1171–1189.

Choi,J.K. *et al.* (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, 84–90.

Cochran,B.G. (1954) The combination of estimates from different experiments. *Biometrics*, **10**, 101–129.

DeConde,R.P. *et al.* (2006) Combining results of microarray experiments: a rank aggregation approach. *Stat. Appl. Genet. Mol. Biol.*, **5**, 15.

DerSimonian,R. and Laird,N.M. (1986) Meta-analysis in clinical trials. *Control. Clin. Trials*, **7**, 177–188.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Fisher,R.A. (1925) Statistical Methods for Research Worker. Oliver and Boyd, Edinburg and London.

Gentleman,R.C. *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Ghosh,D. *et al.* (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct. Integr. Genomics*, **3**, 180–188.

Good,I.J. (1955) On the weighted combination of significance tests. *J. R. Stat. Soc.*, **2**, 264–265.

Hedges,L.V. and Olkin,I. (1985) *Statistical Methods For Meta-Analysis*. Academic Press, Burlington, MA.

Hong,F. *et al.* (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825–27.

Irizarry,R.A. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–349.

Jarvinen,A.K. *et al.* (2004) Are data from different gene expression microarray platforms comparable? *Genomics*, **83**, 1164–1168.

Kothapalli,R. *et al.* (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.

Kuo,W.P. *et al.* (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–12.

Newton,M.A. *et al.* (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Moreau,Y. *et al.* (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.*, **19**, 570–577.

Nemhauser,J.L. *et al*. (2004) Interdependency of brassinosteroid and auxin signaling in Arabidopsis. *PLoS Biol*., **2**, E258.

Normand,S.L. (1999) Tutorial in biostatistics-meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med*, **18**, 321–359.

Pan,W. (2002) A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. *Bioinformatics*, **12**, 546–554.

Parmigiani,G. *et al*. (2004) A cross-study comparison of gene expression studies for the molecular classificaiton of lung cancer. *Clin. Cancer Res*., **10**, 2922–2927.

Park,W.D. and Stegall,M.D. (2007) A meta-analysis of kidney microarray datasets: investigation of cytokine gene detection and correlation with RT-PCR and detection thresholds. *BMC Genomics*, **8**, 88.

Pepe,M.S. (2000) Receiver operating characteristic methodology. *J. Am. Stat. Assoc*., **95**, 308–311.

Pyne,S. *et al*. (2006) Meta-analysis based on control of false discovery rate: combining yeast Chip-chip datasets. *Bioinformatics*, **22**, 2516–2522.

Rhodes,D.R. *et al*. (2002) Meta-analysis of microarrays: inter-study validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*., **62**, 4427–4433.

Rosenthal,R. (1991) *Meta-analytic Procedures for Social research*. SAGE Publications.

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

Tippet,L.H.C. (1931) *The Methods of Statistics*. Williams and Norgate, London.

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

Varambally,S. *et al*. (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, **8**, 393–406.

Vert,G. *et al*. (2005) Molecular mechanisms of steroid hormone signaling in plants. *Annu. Rev. Cell Dev. Biol*., **21**, 177–201.

Wang,J. *et al*. (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics*, **20**, 3166–3178.

Wilkinson,B. (1951) A statistical consideration in psychological research. *Psychol. Bull*., **48**, 156–158.

Wu,Z. *et al*. (2004) A model based background adjustement for oligonucleotide expression arrays. *J. Am. Stat. Assoc*., **99**, 909–917.

Yuen,T. *et al*. (2002) Accurancy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res*., **30**, e48.

Zaykin,D.V. *et al*. (2002) Truncated product method for combining *P*-values. *Genetic Epidemiol*., **22**, 170–185.