Gene expression

Moderated effect size and P-value combinations for microarray meta-analyses

Guillemette Marot1,*, Jean-Louis Foulley1, Claus-Dieter Mayer2 and Florence Jaffrézic1

1 INRA, UMR 1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, F-78350, France and 2 Biomathematics and Statistics, Rowett Institute of Nutrition and Health, University of Aberdeen, Scotland, UK

Received on March 6, 2009; revised on June 25, 2009; accepted on July 7, 2009

Advance Access publication July 23, 2009

ABSTRACT

Motivation: With the proliferation of microarray experiments and their availability in the public domain, the use of meta-analysis methods to combine results from different studies increases. In microarray experiments, where the sample size is often limited, meta-analysis offers the possibility to considerably increase the statistical power and give more accurate results.

Results: A moderated effect size combination method was proposed and compared with other meta-analysis approaches. All methods were applied to real publicly available datasets on prostate cancer, and were compared in an extensive simulation study for various amounts of inter-study variability. Although the proposed moderated effect size combination improved already existing effect size approaches, the P-value combination was found to provide a better sensitivity and a better gene ranking than the other meta-analysis methods, while effect size methods were more conservative.

Availability: An R package metaMA is available on the CRAN.

Contact: guillemette.marot@jouy.inra.fr

1 INTRODUCTION

Meta-analysis, which consists in combining data or results from different studies, has been widely used in medicine and health policy to interpret contradictory results from various studies or overcome the problem of reduced statistical power in studies with small sample sizes. Hedges and Olkin (1985) and Stangl and Berry (2000) provide good reviews of meta-analysis techniques.

Microarray experiments are a typical example of small sample size designs. These experiments enable us to study gene expressions from thousands of genes at a time, often relying on small sample sizes. Hedges and Olkin (1985) and Stangl and Berry (2000) provide good reviews of meta-analysis techniques.

Motivation: With the proliferation of microarray experiments and their availability in the public domain, the use of meta-analysis methods to combine results from different studies increases. In microarray experiments, where the sample size is often limited, meta-analysis offers the possibility to considerably increase the statistical power and give more accurate results.

Results: A moderated effect size combination method was proposed and compared with other meta-analysis approaches. All methods were applied to real publicly available datasets on prostate cancer, and were compared in an extensive simulation study for various amounts of inter-study variability. Although the proposed moderated effect size combination improved already existing effect size approaches, the P-value combination was found to provide a better sensitivity and a better gene ranking than the other meta-analysis methods, while effect size methods were more conservative.

Availability: An R package metaMA is available on the CRAN.

Contact: guillemette.marot@jouy.inra.fr

1 INTRODUCTION

Meta-analysis, which consists in combining data or results from different studies, has been widely used in medicine and health policy to interpret contradictory results from various studies or overcome the problem of reduced statistical power in studies with small sample sizes. Hedges and Olkin (1985) and Stangl and Berry (2000) provide good reviews of meta-analysis techniques.

Microarray experiments are a typical example of small sample size designs. These experiments enable us to study gene expressions from thousands of genes at a time, often relying on small sample sizes. Hedges and Olkin (1985) and Stangl and Berry (2000) provide good reviews of meta-analysis techniques.

Motivation: With the proliferation of microarray experiments and their availability in the public domain, the use of meta-analysis methods to combine results from different studies increases. In microarray experiments, where the sample size is often limited, meta-analysis offers the possibility to considerably increase the statistical power and give more accurate results.

Results: A moderated effect size combination method was proposed and compared with other meta-analysis approaches. All methods were applied to real publicly available datasets on prostate cancer, and were compared in an extensive simulation study for various amounts of inter-study variability. Although the proposed moderated effect size combination improved already existing effect size approaches, the P-value combination was found to provide a better sensitivity and a better gene ranking than the other meta-analysis methods, while effect size methods were more conservative.

Availability: An R package metaMA is available on the CRAN.

Contact: guillemette.marot@jouy.inra.fr

2 METHODS

2.1 Effect size calculation

Let Ysg and Ysjg be the expression levels for gene g in conditions i and j for study s and replicate r. The data are assumed to be normally distributed as Ysg ~ N(µsg, σsg2) and Ysjg ~ N(µsjg, σsjg2). A simple effect size is the standardized difference:

\[ \theta_g = \frac{\mu_{sg} - \mu_{sjg}}{\sigma_g} \]  

(1)

For effect size calculations, the procedure described by Choi et al. (2003) was applied to estimate the study effect and obtain a test statistic for differential expression. The corresponding hierarchical model used was therefore:

\[ dsg = \theta_g + \epsilon_{sg}, \quad \epsilon_{sg} \sim N(0, \tau^2_{sg}) \]

\[ \theta_g = \mu_t + \epsilon_{sg}, \quad \epsilon_{sg} \sim N(0, \tau^2_{\theta_g}) \]  

(2)

where \( d_{sg} \) is the estimation of the effect size for study s and gene g, \( \tau^2_{sg} \) is the within-study variance, \( \tau^2_{\theta_g} \) is the between-study variance, \( \mu_t \) is the total parameter to estimate, this increases sensitivity, i.e. the proportion of true positives among the truly DE genes.

In the previously mentioned meta-analyses studies, authors based the calculation of the P-values or effect sizes to be combined on standard t-tests, i.e. on gene-by-gene analyses. They therefore gained sensitivity for gene detection by combining different studies, but it is expected that even more sensitivity could be obtained using shrinkage approaches. The aim of this article is to propose a method to calculate moderated effect sizes and to compare their performance with the combination of standard effect sizes or of P-values from standard and moderated t-tests. These methods were applied to publicly available datasets on prostate cancer and compared in an extensive simulation study.
a linear mixed model. An estimation of the between-study variances $\tau^2$ can be obtained using the method of moments as suggested by Choi et al. (2003). Parameter $t_{d}$ is estimated as in the generalized least squares method: $t_{d} = \frac{\sum_{i=1}^{n}(x_{i}^2 + r_{i}^2)}{\sqrt{\sum_{i=1}^{n}(x_{i}^2 + r_{i}^2)}}$. With $Var(t_{d}) = \frac{1}{\sum_{i=1}^{n}(x_{i}^2 + r_{i}^2)}$. A t-score to test for DE genes is then constructed as follows:

$$ z_{i} = \frac{t_{d}}{\sqrt{Var(t_{d})}} $$

(3)

Although Choi et al. (2003) advise permutations to calculate $P$-values and estimate the false discovery rate (FDR), a faster solution is suggested in the Bioconductor package GeneMeta, which assumes a normal distribution on the $z$-scores after checking the reliability of this hypothesis by a $q$-$q$ plot.

### 2.1.1 Moderated effect sizes for unpaired data

To estimate the effect size defined in Equation (1) for unpaired data, Choi et al. (2003) considered the unbiased estimator of the standardized mean difference (for more clarity, indices $g$ and $s$ are omitted in this section):

$$ d' = \frac{1 - \frac{3}{4(n-23-1)}}{(4)$$

(4)

with $d = (\bar{Y}_i - \bar{Y}_j)/s_p$ for conditions $i$ and $j$, and $S_p$ is a pooled SD. These $d$ effect sizes can easily be linked to student’s $t$-tests via the relationship:

$$ t_d = d' \cdot \sqrt{n} $$

(5)

We propose to extend these effect sizes to account for moderated $t$-tests.

First we consider the popular shrinkage approach proposed by Smyth (2004) and implemented in the Bioconductor R package limma. We will also accommodate the effect size calculation to another shrinkage approach proposed by Jaffrézic et al. (2007), which allows us to analyse data with heterogeneous variances between conditions. This method is implemented in the R package SMVar available on the CRAN.

As the same variance is assumed for both conditions in limma, in this case the moderated effect size can be estimated as:

$$ d_{extera} = \frac{\bar{Y}_i - \bar{Y}_j}{2} $$

(6)

For SMVar, different variances are assumed in each condition $i$ and $j$ such that $Y_i \sim N(\mu_i, \sigma_i^2)$ and $Y_j \sim N(\mu_j, \sigma_j^2)$. In this case, we rely on the effect size definition proposed by Kulinskaya and Staudte (2007) where the denominator $\sigma$ is:

$$ \sigma = \sqrt{\frac{\sigma_i^2 + \frac{1}{n_i} + \sigma_j^2 + \frac{1}{n_j}}{2}} $$

(7)

with $q = n_i/n$ and $n = n_i + n_j$. This parameter can be rewritten as $\sigma^2 = (\sigma_i^2 + \frac{1}{n_i} + \sigma_j^2 + \frac{1}{n_j})/2$ so that $\sigma^2 = \sigma_i^2 + \sigma_j^2/2$. This effect size can, therefore, be linked to the Welch statistic as:

$$ d_{Welch} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\sigma^2}} $$

(8)

As SMVar relies on a Welch statistic, a natural moderated effect size would be:

$$ d_{SMVar} = \frac{d_{Welch}}{\sqrt{\sigma^2}} $$

(9)

To apply the meta-analysis procedure described in the previous paragraph, variances of effect sizes are also needed. The estimator of the variance $Var(d) = (n_i^{-1} + n_j^{-1} + d^22z/n_i + n_j)^{-1}$ given in Choi et al. (2003) is, however, an asymptotic estimator. As the number of replicates is often limited in microarray experiments, we decided to compute the exact form of the variances for moderated effect sizes. Using the distribution of effect sizes provided by Hedges (1981), it can be shown that:

$$ \text{Var}(d) = m \left[1 + \hat{\sigma}^2\right] - \frac{\hat{\sigma}^2}{c(m)} $$

(10)

with

$$ c(m) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{\frac{m}{2} \Gamma\left(\frac{m-1}{2}\right)}} $$

(11)

In these formulae, $\delta$ is the effect size defined in Equation (1), $\hat{\sigma}$ is equal to $n_i^{-1}n_j(\bar{Y}_i - \bar{Y}_j)$ for limma and $n = n_i + n_j$ for SMVar and $m$ is the number of degrees of freedom. Note that for both estimators given in Equations (6) and (9), the calculation of this variance is possible using Equation (10) thanks to the degrees of freedom of the moderated $t$-statistics provided in both procedures. For limma (Smyth, 2004), $m$ equals to the sum of prior degrees of freedom and residual degrees of freedom for the linear model for gene $g$.

For SMVar, degrees of freedom are calculated by Satterthwaite’s approach as:

$$ m = \frac{\sigma_i^2 + \frac{1}{n_i}}{\sigma^2} + \frac{\sigma_j^2 + \frac{1}{n_j}}{\sigma^2} $$

(12)

This generalizes the formula given in Jaffrézic et al. (2007) for the case where the number of replicates is the same for both conditions.

### 2.1.2 Unbiased estimators of effect sizes

Using the distribution of effect sizes provided by Hedges (1981), unbiased estimators can be defined from the previously proposed moderated effect sizes as:

$$ d'_{unbiased} = \frac{c(m)\delta_{moderated}}{\sqrt{\text{Var}(d)_{moderated}}} $$

(13)

with $c(m)$ given in Equation (11). Equation (13) can be seen as an extension of Equation (4) with $d' = c(m)d$ where $c(m) = 1 - 3/(4m - 1)$ and $m = n - 2$. Assuming that $\text{Var}(c(m)) = 0$, which holds exactly for standard effect sizes and works quite well in practice for moderated effect sizes, the variance of the unbiased effect sizes is computed as $c(m)$ times the variance of the biased estimators given in Equation (10). Since $c(m) = 1$, unbiased estimators have a smaller variance than biased ones.

### 2.1.3 Moderated effect sizes for paired data

For both moderated $t$-tests with limma and SMVar, the unbiased effect size for paired data is obtained via the relationship:

$$ d_{paired} = \frac{c(m)\delta_{moderated}}{\sqrt{\text{Var}(d)_{paired}}} $$

(14)

with $c(m)$ the number of replicates.

### 2.2 P-value combination

Many authors such as Rhodes et al. (2002) and Hu et al. (2006) use Fisher’s combined probability test to combine $P$-values across studies. The main disadvantage of this approach is that, as pointed out by Hong and Bretting (2008), it requires to treat over- and under-expressed genes separately. In this study, we therefore suggest to use the inverse normal method, which is symmetric in the sense that $P$-values near zero are accumulated in the same way as $P$-values near unity (Hedges and Olkin, 1985), and is therefore suitable for combining results for DE genes when the direction of deviation from the null hypothesis is not known. Loughin (2004) compared different $P$-value combination methods in a simulation study and stated that the (unweighted) inverse normal method worked well in cases where the evidence against the null-hypothesis is spread equally across the different studies. The inverse normal method, so-called by Hedges and Olkin (1985), refers to the averaging of transformed individual $P$-values to normal scores.

This procedure was first introduced independently by Snuffer et al. (1949) and (Liptak, 1958). Let $N_i$ be the total number of studies to be combined and $p_i(x)$ the individual $P$-value calculated for study $i$ and gene $g$.

$$ S_g = \frac{1}{N} \sum_{i=1}^{N}(\Phi^{-1}(p_i(x))) $$

(15)

To avoid directional conflicts, it is necessary to use one-sided $P$-values for each study. Under the null hypothesis $S_g$ follows a standard normal
distribution. An overall two-sided P-value can then be obtained by
\[ p_s = 2(1 - \Phi(|S_s|)) \]  \hspace{1cm} (16)

An alternative to (15) is to use the weighted method by Marot and Mayer (2009), which is implemented in the R package metaMA.
\[ S_w = \sum_{i=1}^{N} w_i \Phi^{-1}(1 - p_i(x)) \]  \hspace{1cm} (17)

with
\[ w_i = \frac{n(i)}{\sum_{j=1}^{N} n(j)} \]
where \( n(i) \) is the number of replicates in study \( s \).

3 APPLICATION TO REAL DATASETS

These different methods were compared on real datasets on prostate cancer. Datasets from Singh et al. (2002), LaTulippe et al. (2002) and Stuart et al. (2004) were downloaded from public web sites. All these experiments were generated from the same Affymetrix HG_U95Av2 platform. Data from CEL files were normalized using robust multichip average (RMA) (Irizarry et al., 2003). In the following, datasets are referred to by the name of the corresponding first author. The Singh dataset contains 102 samples, 50 of which are non-tumour prostate samples, the other 52 being prostate tumours. LaTulippe provides three normal samples and 23 cancer samples, while there are 50 normal and 38 cancer samples in the Stuart dataset. Only the 12,600 genes in common between the three datasets were kept for the analysis. Although these datasets are not representative of prostate cancer, datasets from Singh et al. (2002), LaTulippe et al. (2002) and Stuart et al. (2004) were used to extract real inter-study variation and to illustrate how the methods proposed here can be applied. Simulations with smaller sample size designs are presented in the next section.

We first performed standard limma analyses for each of the three studies and applied a Benjamini–Hochberg (BH) correction to take into account the multiple testing problem. At a 1% BH threshold, 1852 genes were significant, 1142 of which in the Singh study, 423 in the LaTulippe and 287 in the Stuart studies, respectively. As shown in the Venn diagram given in Figure 1, only 111 genes were found in common in the top 1000 gene lists.

When binding all the expression data together and including a study effect in the limma linear model, 2422 genes were found significant at the same BH threshold. We compared this gene list with the ones obtained with (i) effect size combination; (ii) weighted inverse normal P-value combination, both procedures being based on limma moderated t-tests. For the effect size combination, 1487 genes were found to be DE at a 1% BH threshold and for the P-value combination, 2637 genes were found significant. Venn diagram corresponding to the comparison of these methods is given in Figure 2. It was found that 1427 genes were common between the three approaches. It can also be noticed that the P-value combination method detected all the genes found with the effect size combination method and all but 101 with the limma including study effect analysis. On the other hand, 256 genes were detected only by the P-value combination approach.

More results obtained with the standard and proposed moderated effect sizes, P-value combination or a global analysis using standard and moderated t-tests are given in Table 1. Since many replicates were involved, we could not observe on these real datasets the gain of DE genes usually found with shrinkage approaches. We could check that, in this case, using the exact variance for standard effect sizes did not change much the number of DE genes compared with using the asymptotic variance. Indeed, the proposed effect size combination based on usual t-tests and the exact variance detected 1507 DE genes, while the z-score given by the GeneMeta package found 1498 DE genes. Table 1 also points out that P-value combination methods detected more genes than either expression data combination or effect size combination.

As far as gene rankings were concerned, they were very similar. Spearman rank correlations equalled 0.81 between the expression data and effect size combination absolute values of test statistics and equalled 0.92 and 0.91 between the P-value combination and each of the two other methods, respectively. Only 33 genes differed between limma including study effect and P-value combination top 1000 gene lists. Effect size combination appeared to have a slightly different ranking with 353 out of its top 1000 genes not found in the other top gene lists. Between the three methods, 554 DE genes were found in common in the top 1000 gene lists.

Fig. 1. Venn diagram comparing the lists of DE genes at a 1% BH threshold obtained by each individual study.

Fig. 2. Venn diagram comparing the lists of DE genes at a 1% BH threshold obtained by combining P-values, effect sizes or binding expression data together.

<table>
<thead>
<tr>
<th>Effect size combination</th>
<th>P-value combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak 1507</td>
<td>2637</td>
</tr>
<tr>
<td>Limma 1487</td>
<td>2637</td>
</tr>
<tr>
<td>SMVar 1647</td>
<td>2730</td>
</tr>
</tbody>
</table>

Table 1. Number of DE genes for the real dataset provided by the different meta-analysis approaches at a 1% BH threshold.
4 SIMULATIONS

Expression data were simulated using a hierarchical model:

\[ y_{i,g} \sim \theta_{i,g} + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma^2_{i,g}) \]

\[ \theta_{i,g} = \mu_{i,g} + \nu_{i,g}, \quad \nu_{i,g} \sim \mathcal{N}(0, \sigma^2_{i,g}) \]

where \( y_{i,g} \) is the expression level for replicate \( r \) of gene \( g \) in condition \( i \) and study \( r \). Note that the variances specified in (2) are related but not identical to (18) because the equations in (2) model effect sizes while the equations in (18) model expression values. Parameters of simulation were obtained from the three real datasets analysed in the previous section. For \( \mu_{i,g} \), we considered the empirical means of gene expression values observed in each condition (tumour/normal) of these datasets. Mean expression values were supposed to be equal for all genes, but the 1427 genes previously found in common between the limma including variance parameters were calculated from the three real datasets and kept different for each gene. The within-study variances were equal to the gene-by-gene empirical estimations of variances in these datasets and were kept different per gene, condition and study. Between-study variance was simulated as the observed between-study variance averaged over the two conditions.

Expression data combination, standard and moderated effect size combination as well as \( P \)-value combination based both on standard and moderated \( t \)-tests were compared in a simulation study with 300 runs. For each method, the number of true positives (TP), false positives (FP), false negatives (FN) and sensitivity were calculated. All these criteria were defined as in Marot and Mayer (2009). In particular, sensitivity was defined as follows:

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \]

As a compromise between FDR and sensitivity, we also calculated the area under the receiver operating characteristic (ROC) curve (AUC) for each method. A few plots of ROC curves are given later in the article. To draw ROC curves, the number of FP, TP, FN and true negatives (TN) were computed for all possible cut-offs in the gene list (1–5000). This procedure was repeated for the 300 simulations and the curves describe the dependency between sensitivity \( E(\text{TP}/(\text{TP} + \text{FN})) \) and specificity \( E(\text{TN}/(\text{TN} + \text{FP})) \). The higher the area under the curve is, the better the gene ranking is.

In these simulations, we considered 3 or 5 studies and 6, 8 or 10 replicates for each study. When five studies were simulated, parameters from the third and the fourth study equalled the ones extracted from the Singh and LaTulippe datasets, respectively. For simplicity, the same number of replicates was simulated in each condition for all studies.

It can be seen from Table 2 that the fewer the replicates were, the larger the gain in sensitivity due to the shrinkage of effect sizes was. Indeed, for five studies and six replicates in each condition, the average sensitivity when using classical effect sizes was 14.3% and it increased to 22.9% and 26.7% when shrinking with limma and SMVar, respectively. All FDRs were below the 5% threshold that was required, slightly higher for the SMVar approach and lower for the standard effect size method. Similar ranking of the methods was observed when changing from five to three studies, with an increase in sensitivity for all methods for a given number of replicates. For example, the sensitivity for the standard effect size approach with 10 replicates per condition was 47.2% for five studies, while only 14.2% when based on three studies. This was expected since the total number of replicates used in the meta-analysis was smaller in the latter case. Gene ranking was very slightly improved by using moderated effect sizes. For example, when only six replicates and three studies were considered, the AUC calculated after a gene-by-gene meta-analysis equalled to 0.827 and increased to 0.834 with a moderated effect size combination.

In the second set of simulations, we studied the influence of between-study variability, comparing all the methods either on datasets where no inter-study was simulated or on datasets simulated as previously, accounting for between-study variation. When simulating an homogeneous dataset, among the previous 1427 genes, only the genes significant at a 1% BH threshold in the Singh and LaTulippe datasets, respectively. For five studies and six replicates the AUC calculated after a gene-by-gene meta-analysis equalled to 0.827 and increased to 0.834 with a moderated effect size combination.

Table 2. Influence of the number of studies and replicates (Rep) on the comparison of meta-analysis methods using moderated effect size (ES) estimators for a BH threshold of 5%

<table>
<thead>
<tr>
<th>Rep</th>
<th>Sens (%)</th>
<th>ESSens</th>
<th>ES_Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three studies</td>
<td>6</td>
<td>1.1 (0.5)</td>
<td>4.8 (0.9)</td>
</tr>
<tr>
<td></td>
<td>FDR (%)</td>
<td>0.2 (1)</td>
<td>0.8 (1)</td>
</tr>
<tr>
<td></td>
<td>AUC (%)</td>
<td>82.7 (0.6)</td>
<td>83.3 (0.6)</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>7.0 (1)</td>
<td>11.0 (1.2)</td>
</tr>
<tr>
<td></td>
<td>FDR (%)</td>
<td>0.9 (0.9)</td>
<td>1.2 (0.9)</td>
</tr>
<tr>
<td></td>
<td>AUC (%)</td>
<td>86.4 (0.5)</td>
<td>86.8 (0.5)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>14.2 (1.3)</td>
<td>17.5 (1.3)</td>
</tr>
<tr>
<td></td>
<td>FDR (%)</td>
<td>1.4 (0.8)</td>
<td>1.6 (0.8)</td>
</tr>
<tr>
<td></td>
<td>AUC (%)</td>
<td>89.0 (0.5)</td>
<td>89.3 (0.5)</td>
</tr>
<tr>
<td>Five studies</td>
<td>6</td>
<td>14.3 (1.2)</td>
<td>22.9 (1.4)</td>
</tr>
<tr>
<td></td>
<td>FDR (%)</td>
<td>0.5 (0.4)</td>
<td>1.0 (0.5)</td>
</tr>
<tr>
<td></td>
<td>AUC (%)</td>
<td>91.2 (0.5)</td>
<td>91.6 (0.4)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>47.2 (1.4)</td>
<td>50.4 (1.3)</td>
</tr>
<tr>
<td></td>
<td>FDR (%)</td>
<td>1.6 (0.5)</td>
<td>1.7 (0.5)</td>
</tr>
<tr>
<td></td>
<td>AUC (%)</td>
<td>95.8 (0.3)</td>
<td>95.9 (0.3)</td>
</tr>
</tbody>
</table>

The table shows average estimated sensitivity (Sens), FDR and AUC as well as their estimated SDs into brackets on 300 simulations.
that shrinkage improved the meta-analysis approaches within the three studies with six replicates in each (Fig. 3c). We also tested the effect of replicates per studies and conditions. For example, we simulated methods were obtained with fewer replicates or different number variability was introduced.

When no inter-study variability was simulated, Figure 3a and b, which show the performance in terms of gene artefact of a higher FDR.

Similar results on the relative performance of meta-analysis methods were obtained with fewer replicates or different number of replicates per studies and conditions. For example, we simulated three studies with six replicates in each (Fig. 3e). With these settings, effect size combination also outperformed a joint limma analysis including a study effect, which was already observed in Table 3. It has to be pointed out that this simulation study tends to favour the last method as a simple additive study effect was considered. The $P$-value and effect size combination methods might perform much better than the simple limma linear study effect model in more complicated settings. Nevertheless, we found particularly interesting the fact that when adopting the real application design, that is, to say 50 and 52 replicates in study 1, 3 and 23 in study 2 and 50 and 38 in study 3, the position of the curves was reversed between the last two methods (Fig. 3d). In all cases, the $P$-value combination provided a better gene ranking than the other combination approaches.

To evaluate the performance of meta-analysis methods, Choi et al. (2003) and Conlon et al. (2007) defined the integration-driven discovery rate (IDR) as the proportion of genes that are identified as DE in the meta-analysis that were not identified in any of the individual studies alone. In the same way, Stevens and Doerge (2005) and Conlon et al. (2007) defined the integration-driven revision rate (IRR) as the percentage of genes that are declared DE in individual studies but not in meta-analysis. While IDR represents the information gained by meta-analysis, IRR measures the loss due to it.

Table 3. Influence of the presence of between-study variability (inter) on 300 simulations with 10 replicates for both conditions in each of the five studies

<table>
<thead>
<tr>
<th>Inter</th>
<th>Sens (%)</th>
<th>FDR (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Joint</td>
<td>83.7 (1.4)</td>
<td>4.7 (0.9)</td>
<td>99.4 (0.1)</td>
</tr>
<tr>
<td>JointSMVar</td>
<td>84.8 (1.4)</td>
<td>4.7 (0.9)</td>
<td>99.4 (0.1)</td>
</tr>
<tr>
<td>JointL1</td>
<td>84.2 (1.5)</td>
<td>4.7 (0.9)</td>
<td>99.4 (0.1)</td>
</tr>
<tr>
<td>JointL2</td>
<td>84.1 (1.4)</td>
<td>4.7 (0.9)</td>
<td>99.4 (0.1)</td>
</tr>
<tr>
<td>ES</td>
<td>66.8 (1.9)</td>
<td>1.5 (0.5)</td>
<td>99.2 (0.2)</td>
</tr>
<tr>
<td>ESJoint</td>
<td>71.6 (1.8)</td>
<td>2.2 (0.7)</td>
<td>99.2 (0.2)</td>
</tr>
<tr>
<td>ESSMVar</td>
<td>69.8 (1.8)</td>
<td>1.9 (0.6)</td>
<td>99.2 (0.2)</td>
</tr>
<tr>
<td>pv</td>
<td>82.8 (1.5)</td>
<td>4.8 (0.9)</td>
<td>99.3 (0.1)</td>
</tr>
<tr>
<td>pvSMVar</td>
<td>87.1 (1.3)</td>
<td>6.1 (0.9)</td>
<td>99.5 (0.1)</td>
</tr>
<tr>
<td>pvLimma</td>
<td>84.8 (1.4)</td>
<td>4.9 (0.9)</td>
<td>99.4 (0.1)</td>
</tr>
</tbody>
</table>

Joint and JointSMVar denote the $t$-test and the SMVar global analyses, respectively. JointL1 and JointL2 are the global limma analyses, the first one only gathering the expression data, the second one including a study effect in the linear model. ES stands for effect size combination and pv for $P$-value combination.

of sensitivity: it was around 84% for joint analyses, whereas it was only around 70% for effect size methods. When simulating between-study variability with parameters extracted from the real prostate cancer datasets, the situation was reversed and meta-analysis methods gave better sensitivities and better AUC than joint analyses. The $P$-value combination methods outperformed effect size combination methods with a sensitivity around 69–72% for the first ones and 47–52% for the latter ones. FDRs were a bit higher for $P$-value combination methods, which reflects the fact that the expression value and effect size combinations were much more conservative than the $P$-value combination. These FDRs were, however, still around the required 5% BH threshold. We also noticed that shrinkage improved the meta-analysis approaches within the same scheme (either $P$-value or effect size combination). The AUC results confirmed the good performance of $P$-value combination methods, showing that the gain in sensitivity is not uniquely an artefact of a higher FDR.

AUC numbers were illustrated by the ROC curves plotted in Figure 3a and b, which show the performance in terms of gene ranking. We only focused on meta-analysis methods based on the limma moderated $t$-test, since limma is the most commonly used package for differential expression. While all curves were similar when no inter-study variability was simulated, $P$-value combination slightly outperformed the other approaches as soon as between-study variability was introduced.

In these formulae, MA refers to meta-analysis and IS to individual studies. We found that interpreting IDR was quite misleading since they are highly dependent on the number of DE genes found with each method. Discoveries or Revisions, which correspond to the numerators of the previous quantities are, therefore, given here in addition to these rates.

In this last part of the simulation study, we considered five studies, with 10 replicates per study and between-study variability close to the one observed in the real application. Results presented in Table 4 show that, in this setting, individual studies missed out many genes. Even if all studies had the same number of replicates, large differences could still be observed between the different studies, depending on the within-study variances adopted for each study.

In particular, the first and fourth studies, whose within-study variances had been simulated from the Singh dataset, detected very few genes compared with the other ones, with 16.5 genes DE on average over the 300 simulations, corresponding to a sensitivity of 1.1%. On the other hand, the second and fifth studies, whose within-study variances had been simulated from the LaTulippe dataset, had the highest sensitivity, equal to 18.4%. From the last column ‘summary’ of the table it can be noticed that most of the genes found in the individual analyses were different from one study to the other since there was an average of 469.6 genes found in total when pooling the individual lists. In this pooled list, the FDR was higher than the 5% BH required threshold, which was expected as there was no further correction after combining gene lists.

On the other hand, Table 5 shows that performing meta-analysis considerably increased the number of DE genes and the number of...
Fig. 3. ROC curves comparing gene ranking with various settings for number of replicates and inter-study variability: (a) No inter-study variability (five studies with 10 replicates); (b) inter-study variability (five studies with 10 replicates); (c) inter-study variability (three studies with six replicates); (d) inter-study variability (real application design); (e) inter-study variability [10/10 replicates (conditions 1/2) in study 1, 10/8 and 3/9 in studies 2 and 3].

Table 4. Results with limma analyses for individual studies (10 replicates for both conditions in each study)

<table>
<thead>
<tr>
<th>Study</th>
<th>DE</th>
<th>Sens.</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.5 (8.9)</td>
<td>1.1 (0.6)</td>
<td>4.4 (5.4)</td>
</tr>
<tr>
<td>2</td>
<td>273.7 (16.7)</td>
<td>18.4 (1.1)</td>
<td>4.2 (1.2)</td>
</tr>
<tr>
<td>3</td>
<td>162.3 (9.9)</td>
<td>10.7 (0.6)</td>
<td>6.1 (2)</td>
</tr>
<tr>
<td>4</td>
<td>16.0 (8.7)</td>
<td>1.1 (0.6)</td>
<td>4.2 (1.2)</td>
</tr>
<tr>
<td>5</td>
<td>409.6 (18.8)</td>
<td>30.5 (1.1)</td>
<td>7.4 (1.1)</td>
</tr>
<tr>
<td>Summary</td>
<td>469.6 (18.8)</td>
<td>30.5 (1.1)</td>
<td>7.4 (1.1)</td>
</tr>
</tbody>
</table>

Column ‘summary’ shows the number of genes obtained when pooling the lists of DE genes from individual studies.

true discoveries. As previously, the P-value combination method had the best sensitivity, equal to 71.2% with an FDR of 4.6%, higher than the FDR for effect size combination (1.7%), but still <5%. In terms of gene ranking, the P-value combination also slightly outperformed the other methods with an AUC of 0.966. The limma analysis including a study effect and the moderated effect size approach also performed quite well with AUC of 0.939 and 0.959, respectively. The three methods outperformed a simple limma analysis on the combined expression values, that did not take into account the between-study variability.

Although the IDR criterion has been used by several authors in the literature, it does not check if the additional genes detected with the meta-analyses are actually TP. In order to compare the different methods, we therefore used the number of TP discoveries. Thus, for the effect size combination method, among the 426 genes detected only with the meta-analysis and not with single study analyses, about 414 were TP, whereas there were 589 out of 635 with the P-value combination method. This result confirms, as previously observed on sensitivities, that P-value combination outperforms effect size combination. Note that these discoveries would be even larger if gene-by-gene analyses had been performed for individual studies as in Choi et al. (2003), instead of limma analyses. Concerning the loss of information due to the meta-analysis, among the 470 genes identified by pooling the lists from the individual studies, about 428 genes were dropped on average when jointly analysing the expression data from the five studies in a simple limma analysis, whereas only about 40 (respectively, 164) when combining P-values (respectively, effect sizes). The P-value
Table 5. Comparison of global limma analyses—the first one (JointL1) only gathering the expression data, the second one (JointL2) including a study effect in the linear model—with P-value and effect size combinations

<table>
<thead>
<tr>
<th></th>
<th>JointL1</th>
<th>JointL2</th>
<th>PvalLimma</th>
<th>ESLimma</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>54.8 (9.3)</td>
<td>853.1 (19.1)</td>
<td>1064.3 (17.7)</td>
<td>732.0 (20.2)</td>
</tr>
<tr>
<td>Sens.</td>
<td>3.5 (0.7)</td>
<td>57.2 (1.2)</td>
<td>71.2 (1)</td>
<td>50.4 (1.3)</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0 (0.3)</td>
<td>4.3 (0.7)</td>
<td>4.6 (0.6)</td>
<td>1.7 (0.5)</td>
</tr>
<tr>
<td>IDR</td>
<td>25.5 (6.2)</td>
<td>54.8 (1.8)</td>
<td>59.7 (1.5)</td>
<td>58.2 (1.8)</td>
</tr>
<tr>
<td>Disc.</td>
<td>14.1 (4.3)</td>
<td>467.2 (21.2)</td>
<td>635.1 (21.8)</td>
<td>426.4 (19.4)</td>
</tr>
<tr>
<td>TP Disc.</td>
<td>14.0 (4.3)</td>
<td>432.7 (18.8)</td>
<td>509.4 (19.7)</td>
<td>338.8 (18.4)</td>
</tr>
<tr>
<td>IRR</td>
<td>91.3 (1.5)</td>
<td>17.8 (1.6)</td>
<td>8.6 (1.2)</td>
<td>34.9 (2.1)</td>
</tr>
<tr>
<td>Revis.</td>
<td>428.8 (18.2)</td>
<td>83.8 (9.4)</td>
<td>40.4 (6.5)</td>
<td>164 (13.2)</td>
</tr>
<tr>
<td>TP Revis.</td>
<td>43.3 (2.5)</td>
<td>82.2 (7.5)</td>
<td>40.2 (2.1)</td>
<td>16.3 (3.6)</td>
</tr>
<tr>
<td>AUC</td>
<td>90.0 (0.4)</td>
<td>93.9 (0.4)</td>
<td>96.6 (0.3)</td>
<td>95.9 (0.3)</td>
</tr>
</tbody>
</table>

The number of (DE) genes, FDR, Sensitivity (Sens.), AUC, IDR, the number of discoveries (Disc.), IRR and the number of revisions (Revis.) are averaged on 300 simulations, 10 replicates were simulated for both conditions in each study.

In the comparison study it was found, however, that the good knowledge of the number of degrees of freedom improved the P-value inverse normal transformations before their combinations and also explains the excellent results of these methods in this article. In a way, the ability of effect sizes to handle variance components was matched by P-value combination using these moderated t-tests.

With the growing amount of publicly available microarray databases, there will be an increasing interest in combining data from different platforms. Technically our metaMA package allows this integration of different platforms since, contrary to the GeneMeta package, it can handle missing data. Thus, genes not spotted onto some arrays could be treated as missing. Moreover, P-value combination methods facilitate cross-platform studies. We would, however, recommend to avoid mixing data from different platforms, if the aim is to increase sensitivity. The minimum we would advise is to only keep genes which could correspond to a common identifier in order to delete missing not at random data. In the case of cross-platform studies, the most difficult job is to match identifiers between platforms; it must be kept in mind that meta-analysis requires a certain data quality.

REFERENCES


Conflict of Interest: none declared.

Funding: Scottish Government Rural and Environment Research and Analysis Directorate (RERAD), to C.-D.M.J.

5 DISCUSSION

Extension of shrinkage approaches from moderated t-tests to effect sizes was a natural way to take into account the sample size in microarray experiments. Thus, not only sensitivity is gained via meta-analysis but also no sensitivity is lost due to the inefficiency of gene-by-gene analyses, especially when there are few replicates. The proposed moderated effect size combination were able to improve traditional effect size meta-analysis approaches. In the comparison study it was found, however, that P-value combination methods usually outperformed effect size combination approaches. The simulation study showed that in various settings, for different numbers of studies, replicates per study and between-study variability close to the one observed between real prostate cancer datasets, the meta-analysis methods outperformed the other meta-analysis methods regarding sensitivity and gene ranking.

It is to be noted that for interpretability reasons P-values have to come from the same statistic, and preferably from moderated t-tests such as limma (Smyth, 2004) or SMVar (Jaffrézic et al., 2007). Effect size combination methods were found to be more conservative and offered more accurate results in terms of P-values.