Differential meta-analysis of RNA-seq data from multiple studies

Andrea Rau, Guillemette Marot, Florence Jaffrezic

Differential meta-analysis of RNA-seq data from multiple studies

Andrea Rau[1,2] Guillemette Marot[3,4] Florence Jaffrézic[1,2]

June 13, 2013

1. INRA, UMR1313 Génétique animale et biologie intégrative, 78352 Jouy-enJosas, France
2. AgroParisTech, UMR1313 Génétique animale et biologie intégrative, 75231 Paris 05, France
3. Université Lille Nord de France, UDSL, EA2694 Biostatistics
4. Inria Lille Nord Europe, MODAL

ABSTRACT

Background: High-throughput sequencing is now regularly used for studies of the transcriptome (RNA-seq), particularly for comparisons among experimental conditions. For the time being, a limited number of biological replicates are typically considered in such experiments, leading to low detection power for differential expression. As their cost continues to decrease, it is likely that additional follow-up studies will be conducted to re-address the same biological question.
Results: We demonstrate how p-value combination techniques previously used for microarray meta-analyses can be used for the differential analysis of RNA-seq data from multiple related studies. These techniques are compared to a negative binomial generalized linear model (GLM) including a fixed study effect on simulated data and real data on human melanoma cell lines. The GLM with fixed study effect performed well for low inter-study variation and small numbers of studies, but was outperformed by the meta-analysis methods for moderate to large inter-study variability and larger numbers of studies.
Conclusions: The p-value combination techniques illustrated here are a valuable tool to perform differential meta-analyses of RNA-seq data by appropriately accounting for biological and technical variability within studies as well as additional study-specific effects. An R package `metaRNASeq` is available on the R Forge.
Keywords: meta-analysis, RNA-seq, differential expression, p-value combination

1 Background

Studies of gene expression have come to rely increasingly on the use of high-throughput sequencing (HTS) techniques to directly sequence libraries of reads (i.e., nucleotide sequences) arising from the transcriptome (RNA-seq), yielding counts of the number of reads arising from each gene. Although the costs of HTS experiments continue to decrease, for the time being RNA-seq experiments are typically performed on very few biological replicates, and therefore analyses to detect differential expression between two experimental conditions tend to lack detection power. However, as costs continue to decrease, it is likely that additional follow-up experiments will be conducted to re-address some biological questions, suggesting a future need for methods able to jointly analyze data from multiple studies. In particular, such methods must be able to appropriately account for the biological and technical variability among samples within a given study as well as for the additional variability due to study-specific effects. Such inter-study variability may arise due to technical differences among studies (e.g., sample preparation, library protocols, batch effects) as well as additional biological variability.

In recent years, several methods have been proposed to analyze microarray data arising from multiple independent but related studies; these meta-analysis techniques have the advantage of increasing the available sample size by integrating related datasets, subsequently increasing the power to detect differential expression. Such meta-analyses include, for example, methods to combine p-values [14], estimate and combine effect sizes [6], and rank genes within each study [4]; see [10] and [9] for a review and comparison of such methods. In particular, [14] showed that the inverse normal p-value combination technique outperformed effect size combination methods or moderated t-tests [18] obtained from a linear model with a fixed study effect on several criteria, including sensitivity, area under the Receiver Operating Characteristic (ROC) curve, and gene ranking.

In many cases the meta-analysis techniques previously used for microarray data are not directly applicable for RNA-seq data. In particular, differential analyses of microarray data, whether for one or multiple studies, typically make use of a standard or moderated t-test [11, 18], as such data are continuous and may be roughly approximated by a Gaussian distribution after log-transformation. On the other hand, the growing body of work concerning the differential analysis of RNA-seq data has primarily focused on the use of overdispersed Poisson [2] or negative binomial models [1, 17] in order to account for their highly heterogeneous and discrete nature. Under these models, the calculation and interpretation of effect sizes is not straightforward. [12] recently proposed an effect size combination method for Poisson-distributed data, based on an Anscombe transformation, but this method is not well-adapted to RNA-seq data due to the presence of over-dispersion among biological replicates as well as zero-inflation. To our knowledge, no other transformation has been proposed to obtain effect sizes for over-dispersed Poisson or negative binomial data.

In this paper, we consider several methods for the integrated analysis of RNA-seq data arising from multiple related studies, including two p-value combination methods as well as a model fitted over the full data with a fixed study effect. We first demonstrate how the inverse normal and Fisher p-value combination methods can be adapted to the differential meta-analysis of RNA-seq data. Then we compare these two methods to the results of independent per-study analyses and a negative binomial generalized linear model (GLM) with a fixed study effect as implemented in the `DESeq` Bioconductor package [1]. All methods are compared on real data from two related studies on human melanoma cell lines, as well as in an extensive set of simulations varying the inter-study variability, number of studies, and biological replicates per study.

Finally, we note that our focus is on RNA-seq data arising from two or more studies in which all experimental conditions under consideration are included in every study (with potentially different numbers of biological replicates); differential analyses among conditions that are not studied in the same experiment are typically limited, or even compromised, due to the confounding of condition and study effects.

2 Methods

Let y_{gcrs} be the observed count for gene g ($g = 1, \ldots, G$), condition c ($c = 1, 2$), biological replicate r ($r = 1, \ldots R_{cs}$), and study s ($s = 1, \ldots, S$). Note that the number of biological replicates R_{cs} may vary between conditions and among studies. Let μ_{gcs} be the mean expression level for gene g in condition c and study s. For an integrated differential analysis of gene expression across all studies, two approaches can be envisaged: the combination of p-values from per-study differential analyses, and a global differential analysis. We illustrate both using the default methods and parameters of the `DESeq` (v1.10.1) analysis pipeline [1], although other popular methods, e.g., `edgeR` [17], could also be used.

2.1 P-value combination from independent analyses

For the differential analysis of gene expression within a given study s, we assume that gene counts y_{gcrs} follow a negative binomial distribution parameterized by its mean $\ell_{crs}\mu_{gcs}$ and dispersion α_{gs}, where ℓ_{crs} is a normalization factor to correct for differences in library size. A comparison of different methods to estimate ℓ_{crs} may be found in [7]. We are interested in testing the per-gene null hypothesis

$$H_{0,gs} : \mu_{g1s} = \mu_{g2s} \quad \text{vs} \quad H_{1,gs} : \mu_{g1s} \neq \mu_{g2s}. \tag{1}$$

After obtaining per-gene mean and dispersion parameter estimates, a parametric gamma regression is used to obtain fitted dispersion estimates by pooling information from genes with similar expression strengths. Raw per-gene p-values p_{gs} are subsequently computed using a conditioned test analogous to Fisher's exact test. Additional details may be found in [1] and the `DESeq` package vignette. Once these vectors of raw p-values have been obtained, we consider two possible approaches to combine them across studies: the inverse normal and the Fisher combination methods. We note that both of these approaches assume that under the null hypothesis, each vector of p-values is assumed to be uniformly distributed.

2.1.1 Inverse normal method

For each gene g, we define

$$N_g = \sum_{s=1}^{S} w_s \Phi^{-1}(1 - p_{gs}) \tag{2}$$

where p_{gs} corresponds to the raw p-value obtained for gene g in a differential analysis for study s, Φ the cumulative distribution function of the standard normal distribution, and w_s a set of

weights [13]. We propose here to define the study-specific weights w_s as in [15]:

$$w_s = \sqrt{\frac{\sum_c R_{cs}}{\sum_\ell \sum_c R_{c\ell}}},$$

where $\sum_c R_{cs}$ is the total number of biological replicates in study s. This allows studies with large numbers of biological replicates to be attributed a larger weight than smaller studies. We note that other weights may also be defined by the user depending on the quality of the data in each study, if this information is available.

Under the null hypothesis, the test statistic N_g in Equation (2) follows a $\mathcal{N}(0,1)$ distribution. A unilateral test on the right-hand tail of the distribution may then be performed, and classical procedures for the correction of multiple testing such as [3] may subsequently be applied to the obtained p-values to control the false discovery rate at a desired level α.

2.1.2 Fisher combination method

For the Fisher combination method [8], the test statistic for each gene g may be defined as

$$F_g = -2 \sum_{s=1}^{S} \ln(p_{gs}), \tag{3}$$

where as before p_{gs} corresponds to the raw p-value obtained for gene g in a differential analysis for study s. Under the null hypothesis, the test statistic F_g in Equation (3) follows a χ^2 distribution with $2S$ degrees of freedom. As with the inverse normal p-value combination method, classical procedures for the correction of multiple testing such as [3] may be applied to the obtained p-values to control the false discovery rate at a desired level α.

2.1.3 Additional considerations for p-value combination

We note that the implementation of the previously described p-value combination techniques requires two additional considerations to be taken into account.

First, a crucial underlying assumption for the statistics defined in Equations (2) and (3) is that p-values for all genes arising from the per-study differential analyses are uniformly distributed under the null hypothesis $H_{0,gs}$ defined in Equation (1). This assumption is, however, not always satisfied for RNA-seq data; in particular, a peak is often observed for p-values close to 1 due to the discretization of p-values for very low counts. To circumvent this first difficulty, we propose to filter the weakly expressed genes in each study using the `HTSFilter` Bioconductor package [16], as described in the Supplementary Materials. We note that in so doing, it is possible for a gene to be filtered from one study and not from another. As will be seen in the following, this approach appears to effectively filter those genes contributing to a peak of large p-values, resulting in p-values that appear to be roughly uniformly distributed under the null hypothesis.

Second, for the two p-value combination methods described above, unlike microarray data, under- and over-expressed genes are analyzed together for RNA-seq data. As such, some care must be taken to identify genes exhibiting conflicting expression patterns (i.e., under-expression when comparing one condition to another in one study, and over-expression for the same comparison in another study). We suggest that genes exhibiting differential expression conflicts among studies be

identified post hoc, and removed from the list of differentially expressed genes; this step to remove genes with conflicting differential expression from the final list of differentially expressed genes may be performed automatically within the associated R package `metaRNASeq`.

2.2 Global differential analysis

For a global analysis of RNA-seq data arising from multiple studies, we assume that gene counts y_{gcrs} follow a negative binomial distribution parameterized by the mean $\ell_{crs}\mu_{gcs}$ and dispersion α_g, where ℓ_{crs} is the library size normalization factor. We are now interested in testing the global per-gene null hypothesis

$$H_{0,g} : \mu_{g1} = \mu_{g2} \quad \text{vs} \quad H_{1,g} : \mu_{g1} \neq \mu_{g2}.$$

Per-gene mean and fitted dispersion estimates are obtained as before. In order to estimate a possible effect due to study, a full and reduced model are fitted for each gene using negative binomial generalized linear models (GLM); the full model regresses gene expression on the experimental condition and study, while the reduced model regresses gene expression only on the study. The two models are compared using a χ^2 likelihood ratio test to determine whether including the experimental condition significantly improves the model fit. Note that for the global differential analysis we use the `HTSFilter` Bioconductor package [16] to filter the full set of data across studies, resulting in a vector of raw filtered per-gene p-values that may be corrected for multiple testing using classical procedures [3] to control the false discovery rate at a desired level α. Additional details may be found in the `DESeq` package vignette.

3 Results and Discussion

3.1 Application to real data

3.1.1 Presentation of the data

The negative binomial GLM and p-value combination methods were applied to a pair of real RNA-seq studies performed to compare two human melanoma cell lines [19]. Each study compares gene expression in a melanoma cell line expressing the Microphtalmia Transcription Factor (MiTF) to one in which small interfering RNAs (siRNAs) were used to repress MiTF, with three biological replicates per cell line in the first (hereafter referred to as Study A) and two per cell line in the second (Study B). The raw read counts and phenotype tables for Study A are available in the Supplementary Materials of [7], and the data from Study B from [19].

The characteristics of the data from these two studies are summarized in Supplementary Table 2. In particular, we note that the data from Study A tend to have larger total library sizes and a smaller number of unique reads than those from Study B; in addition, Study A appears to exhibit larger overall per-gene variability than does Study B (Supplementary Figure 8). These two points indicate that in this pair of studies, a considerable amount of inter-study variability appears to be present (Supplementary Figure 9).

3.1.2 Results

After performing individual differential analysis for each study using the negative binomial model and exact test as described above, we obtained per-gene p-values for each study (Figure 1, histograms in background). As previously stated, an important underlying assumption of the p-value combination methods is that the p-values are uniformly distributed under the null hypothesis; we note that this is not the case here, especially for the second study, due to a large peak of values close to 1 resulting from the discretization of p-values. In order to remove the weakly expressed genes contributing to this peak in each study, we filtered the data from each study as proposed in [16], resulting in a distribution of raw p-values from each study that appears to satisfy the uniformity assumption under the null hypothesis (Figure 1, histograms in foreground).

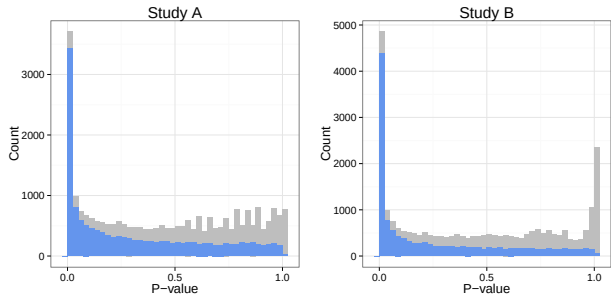

Figure 1: Histograms of raw p-values obtained from per-study differential analyses in the real data: unfiltered (in grey) and filtered (in blue) using the method of [16]. Figure made using the `ggplot2` package [20].

The per-study filtered p-values were combined using the test statistics defined in Equations (2) and (3), and the corresponding results were compared to those of the intersection of independent per-study analyses and the global analysis using a negative binomial GLM with a fixed study effect. We note that for the independent per-study differential analyses, a gene is declared to be differentially expressed if identified in both studies with no differential expression conflict (see Section 2.1.3).

The Venn diagram presented in Figure 2 compares the lists of differentially expressed genes found for all methods considered. It may immediately be noticed that the independent per-study analysis approach is very conservative, and both of the p-value combination approaches (Fisher and inverse normal) considerably increase the detection power. In addition, a large number of genes are found in common among the p-value combination methods and the global analysis (3578 compared to only 1583 from the intersection of individual studies). In order to determine whether the genes uniquely identified by a particular method appear to be biologically pertinent, an Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com) was performed to identify functional annotation for the genes uniquely identified by the Fisher p-value combination method with respect to the global analysis, and vice versa. We note that the sets of genes uniquely identified by the Fisher method or the global analysis (Supplementary Tables 2 and 3), as well as the set of genes found in common (Supplementary Table 4), all appear to include genes of potential interest related to cancer or melanoma, which was the main focus of this set of studies. As such, for this pair of studies it appears that the union of genes identified by the two approaches may be

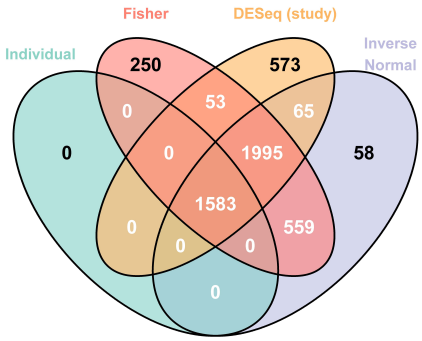

Figure 2: Venn diagram presenting the results of the differential analysis for the real data for the two meta-analysis methods (Fisher and inverse normal), the global analysis (DESeq (study)), and the intersection of individual per-study analyses (Individual). Figure made using the `VennDiagrams` package [5].

of biological interest; to further study the effect of number of studies and inter-study variability on the performance of each method, we investigate an extensive set of simulated data in the following section.

3.2 Simulation study

Data were simulated according to a negative binomial distribution,

$$Y_{gcrs} \sim \mathcal{NB}(\mu_{gcs}, \phi_{gs})$$

where μ_{gcs} and ϕ_{gs} represent the mean and dispersion, respectively, for gene g, condition c and study s, and the mean-variance relationship is defined by

$$\text{Var}(Y_{gcrs}) = \mu_{gcs} + \frac{\mu_{gcs}^2}{\phi_{gs}}.$$

In order to incorporate inter-study variability, we consider the following situation for the mean parameter μ_{gcs}:

$$\log(\mu_{gcs}) = w_{gc} + \varepsilon_{gcs}, \text{ and } \varepsilon_{gcs} \sim \mathcal{N}(0, \sigma^2),$$

where w_{gc} represents the overall mean for gene g in condition c, ε_{gcs} the variability around these means due to a study effect, and σ^2 the size of the inter-study variability. Note that as ε_{gcs} affects μ_{gcs} through a log link, the value of $\exp(\varepsilon_{gcs})$ has a multiplicative effect on the mean.

3.2.1 Parameters for simulations

To fix realistic values for the parameters $\{w_{gc}, \phi_{gs}, \sigma\}$, we first performed individual per-study differential analyses by fitting a negative binomial model with the default methods and parameters of the `DESeq` package (see Section 2.1) on the unfiltered human data presented in Section 3.1. The per-study false discovery rate was subsequently controlled at the $\alpha = 0.05$ level [3]. For the genes identified as differentially expressed in both studies, w_{g1} and w_{g2} were fixed to be the values

Table 1: Simulation settings, including the number of studies and the number of replicates per condition in each study.

Setting	# of studies	Replicates/study
1	2	(2,3)
2	3	(2,2,3)
3	5	(2,2,3,3,3)

of the empirical means (after normalization for library size differences) for each condition across studies. For the remaining genes, we set $w_{g1} = w_{g2} = w_g$ to be the overall empirical mean (after normalization for library size differences) for gene g across both conditions and studies. Using the gamma-family GLM fitted to the per-gene mean and dispersion parameter estimates for each study (Supplementary Figure 8), we fixed the dispersion parameter ϕ_{gs} to be equal to the fitted values

$$\phi_{gs}^{-1} = \hat{\alpha}_{0s} + \frac{\hat{\alpha}_{1s}}{w_g},$$

where $\hat{\alpha}_{0s}$ and $\hat{\alpha}_{1s}$ are the estimated coefficients from the gamma-family GLM for study s, and w_g is the overall empirical mean for gene g. For weakly expressed genes, it has been observed that little overdispersion is present as biological variation is dominated by shot noise (i.e., the variation inherent to a counting process); for genes with $w_g < 10$, the dispersion parameter is therefore fixed to be $\phi_{gs} = 10^{10}$, which corresponds to nearly zero overdispersion (i.e., mean nearly equal to the variance).

Finally, the parameter σ is chosen to represent a range of values for the amount of inter-study variability. The observed human data exhibit a considerable amount of inter-study variability, corresponding to a value of roughly $\sigma = 0.5$ (see Supplementary Materials, Figure 9). In the following simulations, four values are considered for the parameter σ: $\{0, 0.15, 0.3, 0.5\}$, representing zero, small, medium, and large inter-study variability, respectively. Finally, we note that for genes simulated to be non-differentially expressed, we set $\varepsilon_{g1s} = \varepsilon_{g2s} = \varepsilon_{gs} \sim \mathcal{N}(0, \sigma^2)$.

The simulation settings used for the number of studies and number of replicates per condition in each study are presented in Table 1 and were chosen to reflect the size of real RNA-seq experiments. When more than two studies were simulated, the same simulation parameters were used as for the first two, as determined from the real data. For simplicity, the same number of replicates was simulated in each condition for all studies.

3.2.2 Methods and criteria for comparison

In addition to the intersection of independent per-study analyses (where genes were declared to be differentially expressed if identified in more than half of the studies with no differential conflict), the Fisher and inverse normal p-value combination techniques, and the global analysis with fixed study effect, we also considered a global analysis with no study effect. For each simulation setting and level of inter-study variability σ, 300 independent datasets were simulated, and the filtering method of [16] was applied, either independently to each study (for the independent per-study analyses and p-value combination techniques) or to the full set of data (for the global analysis).

For each method, performance was assessed using the sensitivity, false discovery rate (FDR) and area under the receiver operating characteristic (ROC) curve (AUC). In addition, we also considered

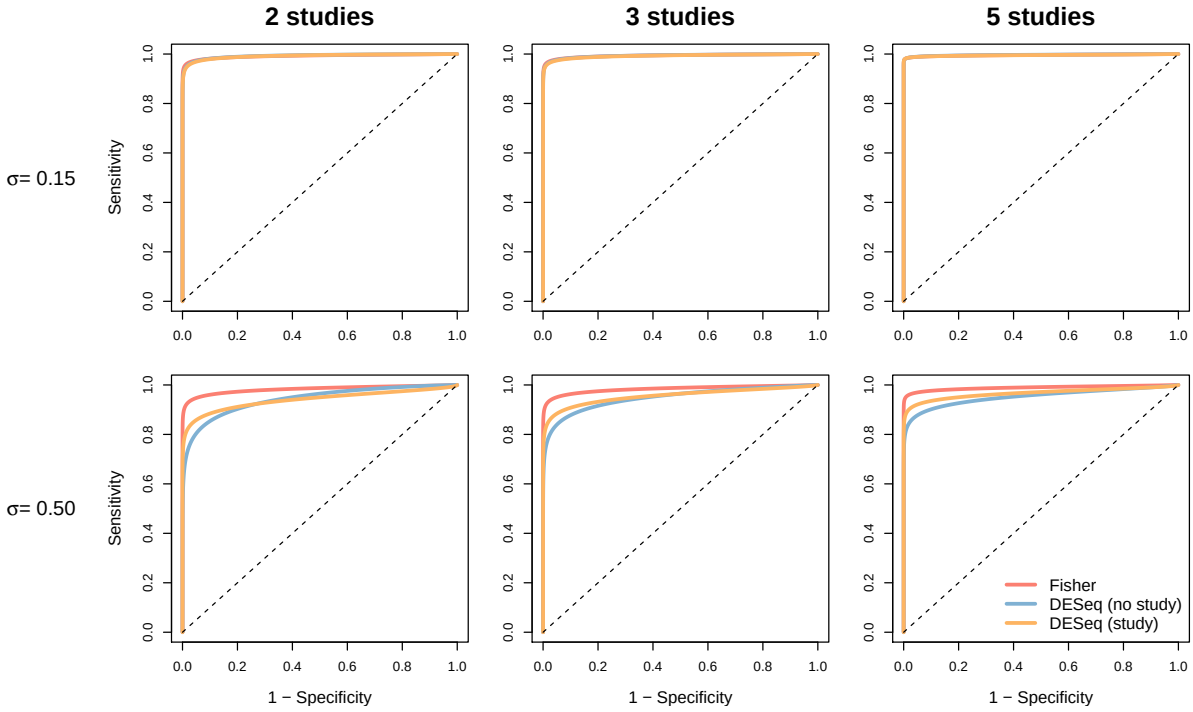

Figure 3: Receiver Operating Characteristic (ROC) curves, averaged over 300 datasets. Each plot represents the results of a particular setting, with columns corresponding (from left to right) to simulations including 2 studies, 3 studies, and 5 studies, and rows corresponding (from top to bottom) to simulations with inter-study variability set to $\sigma = 0.15$ and $\sigma = 0.50$ (low inter-study variability to lareg inter-study variability). Within each plot: Fisher (red lines), global analysis with no fixed study effect (DESeq (no study), blue lines), and global analysis with a fixed study effect (DESeq (study), orange lines). The dotted black line represents the diagonal.

a criterion to assess the "value added" for the p-value combination methods with respect to the global analysis, and vice versa: the proportion of true positives among those uniquely identified by a given method (e.g., the Fisher approach) as compared to another (e.g., the global analysis).

3.2.3 Results

The different methods were first compared with ROC curves, presented in Figure 3 for low and high inter-study variability (results for zero and moderate inter-study variability are shown in Supplementary Figure 5). We note that for clarity, the inverse normal method is not represented on these plots as its performance was found to be equivalent to the Fisher method. It can first be noted that for no or small inter-study variability ($\sigma = 0$ or $\sigma = 0.15$), no practical difference may be observed among the methods. On the other hand, for moderate to large inter-study variability ($\sigma = 0.3$ or $\sigma = 0.5$) differences among the methods become more apparent; this pattern is observed for any number of studies. As expected, including a study effect in the global analysis improves the

performance over a naive global analysis without such an effect. We note that the two proposed meta-analysis methods (inverse normal and Fisher p-value combination) were found to perform very similarly and were able, in the case of large inter-study variability, to outperform the global analysis in terms of AUC (Supplementary Figure 1). In particular, in the presence of large inter-study variability, the naive global analysis without a study effect unsurprisingly has the lowest AUC, and the two meta-analysis methods yield a larger AUC than the global analysis with a study effect.

Considering the sensitivity (Figure 4 and Supplementary Figure 6), the meta-analyses appear to lead to similar, and in some settings considerably higher, detection power compared to the other methods. We note that in all settings, using the intersection of independent analyses leads to much lower sensitivity, even for low or zero inter-study variability. As for the AUC, the sensitivity was found to be considerably improved for the global analysis when including a study effect in the GLM model, particularly for medium to large inter-study variability. The two meta-analysis methods were found to lead to significant improvements in sensitivity as compared to the global analysis in the presence of moderate to large inter-study variability when three or more studies were considered. However, for the setting that most resembles our real data analysis (2 studies, $\sigma = 0.50$), the global analysis with study effect and meta-analyses appear to have similar detection power. Finally, we also note that for all methods the FDR was well controlled below 5% (Supplementary Figure 2).

Based on these criteria, the two proposed meta-analysis methods (inverse normal and Fisher) seem to perform very similarly. In order to more thoroughly investigate the differences between p-value combination methods and the global analysis including a study effect, we calculated the proportion of true positives uniquely detected by the Fisher method as compared to the global analysis with study effect, and vice versa (Figure 5 and Supplementary Figure 7). In the setting closest to the real data analysis presented above (two studies and large inter-study variability), the proportion of true positives found uniquely by either the Fisher approach or the global analysis with fixed study effect are very large (around 80% for both methods). This seems to suggest that the additional genes uniquely found either by the global analysis or Fisher p-value combination method in the real data application may indeed be of great biological interest. For more than two studies, however, as the inter-study variability increases the proportion of truly differentially expressed genes uniquely found by the Fisher method increases compared to the global analysis. For example, for three studies with large inter-study variability ($\sigma = 0.5$), the proportion of truly DE genes uniquely found with the Fisher method was equal to more than 80%, whereas it was only around 40% for the global analysis with a study effect.

4 Conclusions

The aim of this paper was to present and compare different strategies for the differential meta-analysis of RNA-seq data arising from multiple, related studies. As expected, naive analyses such as the overlap of lists of differentially expressed genes found by individual studies or a global analysis not accounting for a study effect perform very poorly. On the other hand, the two proposed meta-analysis methods seem to have very similar performances. For low inter-study variability, the results are very close to those of a global GLM analysis including a study effect. When the inter-study variability increases, however, the gains in performance in terms of AUC, sensitivity, and proportion of true positives among uniquely identified genes for the meta-analysis techniques are

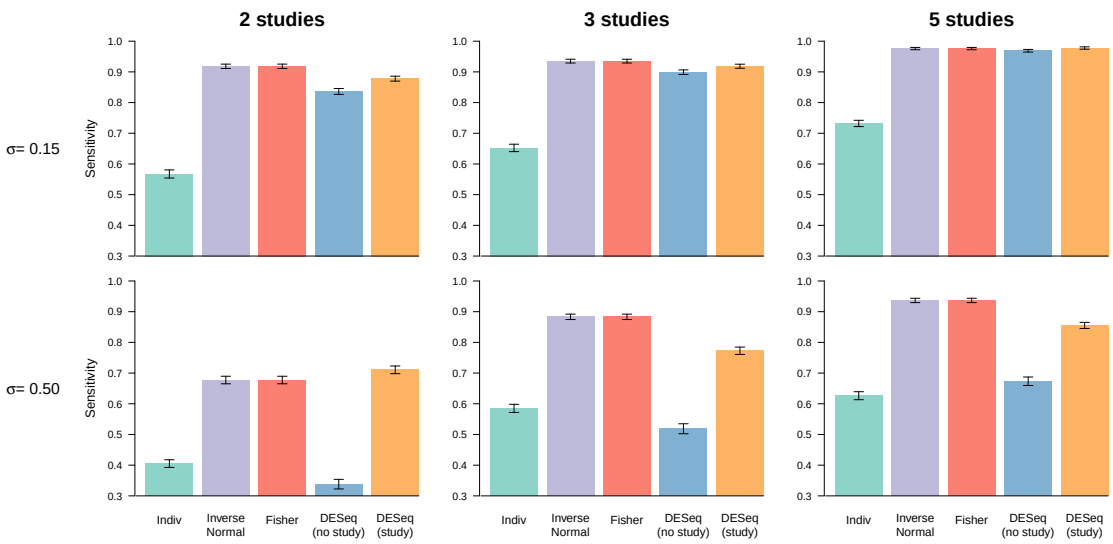

Figure 4: Sensitivity, for the simulation settings corresponding to $\sigma = 0.15$ and $\sigma = 0.50$. Each barplot represents the results of a particular setting, with columns corresponding (from left to right) to simulations including 2 studies, 3 studies, and 5 studies, and rows corresponding (from top to bottom) to simulations with inter-study variability set to $\sigma = 0.15$ and $\sigma = 0.50$ (no inter-study variability to moderate inter-study variability). Within each barplot, from left to right: Individual per-study analyses (green bars), Inverse Normal (purple bars), Fisher (red bars), DESeq with no study effect (blue bars), and DESeq with a fixed study effect (orange bars).

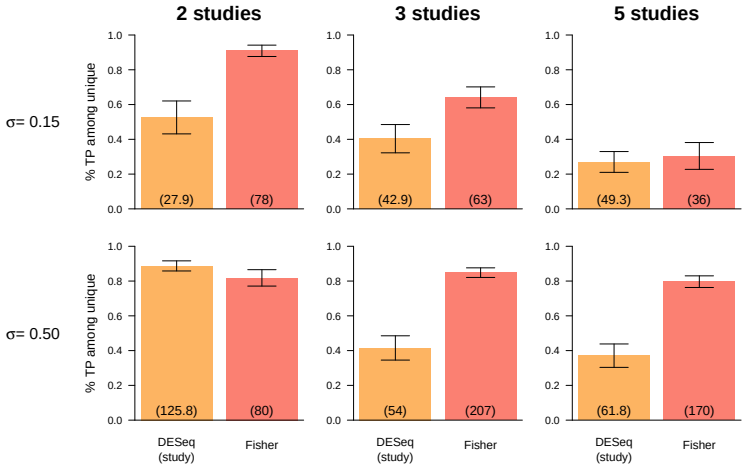

Figure 5: Proportion of true positives among unique discoveries for DESeq with a fixed study effect (orange bars) and Fisher (red bars) for the simulation settings not demonstrated in the main paper. Each barplot represents the results of a particular setting, with columns corresponding (from left to right) to simulations including 2 studies, 3 studies, and 5 studies, and rows corresponding (from top to bottom) to simulations with inter-study variability set to $\sigma = 0.15$ and $\sigma = 0.50$ (no inter-study variability to moderate inter-study variability). Error bars represent one standard deviation, and numbers in parentheses represent the mean total number of unique discoveries for DESeq with study effect as compared to Fisher and vice versa, respectively.

significant as compared to the global analysis, particularly for the analysis of data from more than two studies. We note that both of the proposed p-value combination methods are implemented in an R package called `metaRNASeq`, available on the R-Forge.

Our focus in this work is on differential analyses between two experimental conditions, but can readily be extended to multi-group comparisons. However, as previously noted, the methods presented here are intended for the analysis of data in which all experimental conditions under consideration are included in every study, thus avoiding problems due to the confounding of condition and study effects. As with all meta-analyses, the p-value combination techniques presented here must overcome differences in experimental objectives, design, and populations of interest, as well as differences in sequencing technology, library preparation, and laboratory-specific effects.

The differential meta-analyses presented here concern expression studies based on RNA-seq data. However, other genomic data are generated by high-throughput sequencing techniques, including chromatin immunoprecipitation sequencing (CHIP-seq) and DNA methylation sequencing (methyl-seq), and the proposed techniques could potentially be extended to these other kinds of data. However, in order to be biologically relevant, the p-value combination methods rely on the fact that the same test statistics, or in the case of RNA-seq data conditioned tests, are used to obtain p-values for each study. An important challenge for the future will be to propose methods able to jointly analyze related heterogeneous data, such as microarray and RNA-seq data, or other kinds of genomic data. This is not straightforward in a meta-analysis framework and remains an open research question.

Authors' contributions

AR participated in the design of the study, performed simulations and data analyses, and helped draft the manuscript. GM designed the study, wrote the associated R package, and helped draft the manuscript. FJ conceived of the study, participated in its design, and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Thomas Strub, Irwin Davidson, Céline Keime and the IGBMC sequencing platform for providing the RNA-seq data.

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(R106), 2010.

[2] P. Auer and R. Doerge. A two-stage Poisson model for testing RNA-seq data. *Stat Appl Genet Mol Biol*, 10(26):1–26, 2011.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B*, 57(1):289–300, 1995.

[4] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: A simple yet powerful new method to detect differential regulated genes in replicated microarray experiments. *FEBS Letters*, 573:83–92, 2004.

[5] H. Chen and P. C. Boutros. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, 12(35), 2011.

[6] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and model interstudy variation. *Bioinformatics*, 19(Suppl 1):84–90, 2003.

[7] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, and F. Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, 2012. doi: 10.1093/bib/bbs046.

[8] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 1932.

[9] F. Hong and R. Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, 2008.

[10] P. Hu, C. M. Greenwood, and J. Beyene. Statistical methods for meta-analysis of microarray data: A comparative study. *Inf Syst Front*, 8:9–20, 2006.

[11] F. Jaffrézic, G. Marot, S. Degrelle, I. Hue, and J.-L. Foulley. A structural mixed model for variances in differential gene expression studies. *Genet Res*, 89:19–25, 2007.

[12] E. Kulinskaya, S. Morgenthaler, and R. G. Staudte. *Meta Analysis: A guide to calibrating and combining statistical evidence*, volume 756 of Wiley Series in Probability and Statistics. John Wiley & Sons, West Sussex, England, 2008.

[13] T. Liptak. On the combination of independent tests. *Magyar Tudomanyos. Akademia Matematikai Kutato Intezetenek Kozlemenyei*, 3:171–197, 1958.

[14] G. Marot, J.-L. Foulley, C.-D. Mayer, and F. Jaffrézic. Moderated effect size and p-value combinations for microarray meta-analyses. *Bioinformatics*, 25(20):2692–2699, 2009.

[15] G. Marot and C.-D. Mayer. Sequential analysis for microarray data based on sensitivity and meta-analysis. *Stat Appl Genet Mol Biol*, 8(Article 3), 2009.

[16] A. Rau, M. Gallopin, G. Celeux, and F. Jaffrézic. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *(submitted)*, 2013.

[17] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

[18] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(Article 3), 2004.

[19] T. Strub, S. Giuliano, T. Ye, C. Bonet, C. Keime, D. Kobi, S. L. Gras, M. Cormont, R. Ballotti, C. Bertolotto, and I. Davidson. Essential role of microphthalmia transcription factor for DNA replication, mitosis and genomic stability in melanoma. *Oncogene*, 30:2319–2332, 2011.

[20] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer, New York, 2009.