UNIVERSITY of PISA

PhD COURSE in

Molecular, metabolic and functional exploration

of the nervous system and the sense organs

(SSD BIO/12)

A COMPARISON OF META-ANALYSIS METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES IN MICROARRAY EXPERIMENTS: AN APPLICATION TO MALIGNANT PLEURAL MESOTHELIOMA DATA

Candidate

Dr. Manuela Di Russo

Supervisor

Dr. Silvia Pellegrini

A.A. 2012/2013

ABSTRACT

The proliferation of microarray experiments and the increasing availability of relevant amount of data in public repositories have created a need for meta-analysis methods to efficiently integrate and validate microarray results from independent but related studies.

Despite its increasing popularity, meta-analysis of microarray data is not without problems. In fact, although it shares many features with traditional meta-analysis, most classical meta-analysis methods cannot be directly applied to microarray experiments because of their unique issues.

Several meta-analysis techniques have been proposed in the context of microarrays. However, only recently a comprehensive framework to carry out microarray data meta-analysis has been proposed. Moreover very few software packages for microarray meta-analysis implementation exist and most of them either have unclear manuals or are not easy to apply.

We applied four meta-analysis methods, the Stouffer's method, the moderated effect size combination approach, the t-based hierarchical modeling and the rank product method, to a set of three microarray

studies on malignant pleural mesothelioma. We focused on differential expression analysis between normal and malignant mesothelioma pleural tissues. Both unfiltered and filtered data were analyzed. The lists of differentially expressed genes provided by each method for either kind of data were compared, also by pathway analysis. These comparisons highlighted a poor overlap between the lists of differentially expressed genes and the related pathways obtained using the unfiltered data. Conversely, a higher concordance of the results, both at the gene and the pathway level, was observed when filtered data were considered. The fact that a significant number of genes were identified by only one of the tested methods shows that the gene ranking is based on different perspectives. In fact, the analyzed methods are based on different assumptions and focus on diverse aspects in selecting significant genes. Since so far there is no consensus on what is (are) the 'best' meta-analysis method(s), it may be useful to select candidate genes for further analysis using a combination of different meta-analysis methods. In particular, differentially expressed genes detected by more than one method may be considered as the most reliable ones while genes identified by only a single method may be further explored to expand the knowledge of the biological phenomenon of interest.

1 INTRODUCTION

Microarray technology simultaneously measures the mRNA of tens of thousands of genes in biological samples in a high-throughput and cost-effective manner. Since its introduction in 1995 [1], microarray technology has improved dramatically and became a widely used tool to study the whole transcriptome of many organisms. It has been adopted to explore the molecular basis of fundamental biological processes and complex diseases [2, 3], to improve the disease taxonomy [4, 5], to classify patients into known disease subclasses [6], to analyze the response to drug administration [7], and to predict disease outcomes [8, 9].

Enhancements in microarray technology and its widespread use have led to the generation of a relevant amount of data and resulted in several large public data repositories such as Gene Expression Omnibus (GEO) [10] (http://www.ncbi.nlm.nih.gov/geo/) from NCBI, ArrayExpress [11] (http://www.ebi.ac.uk/arrayexpress/) from EBI and CIBEX (Center for Information Biology gene EXpression database) [12] (http://cibex.nig.ac.jp/).

It is not uncommon to find multiple microarray gene expression studies performed by different research groups worldwide addressing the

same or similar biological questions. Hence there has been a growing interest in developing methods to efficiently integrate microarray data from independent studies with the aim of fully exploiting the rich information produced. Meta-analysis appears to be an effective solution to this pressing issue [13].

As stated by Hedges, “meta-analysis consists of statistical methods for combining results from independent but related studies” [14]. However the term meta-analysis is also widely used in a broader sense, as we do here, to indicate the whole process of identification, selection, assessment and quantitative synthesis of several studies concerning a well-defined research question [15]. Many people use the term meta-analysis interchangeably with systematic review, however not all the systematic reviews are meta-analyses. In fact a meta-analysis is a systematic review which provides a statistical synthesis of the results and produces an overall estimate of the effect of interest.

Meta-analysis offers several practical advantages.

First of all, meta-analysis represents an inexpensive solution to overcome the problem of reduced statistical power of microarray experiments and to reveal true effects of interest [16]. Typically, in microarray experiments many probes are investigated in few samples due to the high cost of this technology or the lack of biological replicates available. The straight consequence is that studies with small sample sizes are less likely to detect true effects and more prone to false positive and false negative results. Putting results together, therefore, increases the sample size and the statistical power of the study. It also allows a more accurate estimation of the effect, even if derived from small but consistent variations.

Moreover, meta-analysis has the potential to strengthen and extend the results obtained by individual studies and to increase their reliability. Indeed, it has been shown that microarray studies are poorly reproducible across platforms and/or laboratories [17, 18]. Technological differences among different microarray platforms [19], large variations in biological and experimental settings, small sample sizes and inappropriate statistical methods [20, 21] have been pointed out as the major sources that contribute to the inconsistency of microarray results. Many of these can be assessed and controlled or overcome by the use of standard reporting methods and the careful application of large-scale meta-analysis techniques with an appropriate statistical modeling of the inter-study variation [16].

Meta-analysis has been widely used in the area of medical and epidemiological research as well as in the sociological and behavioral sciences [22]. The applicability of meta-analysis methods to microarray datasets was demonstrated for the first time in 2002 by Rhodes who combined four datasets on prostate cancer to determine genes that were differentially expressed between clinically localized prostate tumor and benign prostate tissue samples [23]. Since then, several applications of meta-analysis to microarray data appeared in the literature [24-26].

Through a systematic search on PubMed, Tseng and colleagues [27] found that 333 microarray meta-analysis papers (including reviews, biological applications, methodological articles and database/software description papers) were published until December 2010, thus confirming the relevant interest of the scientific community in this challenging task. In more than half of the above mentioned publications, meta-analysis was applied to identify Differentially Expressed Genes (DEGs) between two or more conditions [28-30]. However microarray studies have also been

combined for classification analysis [31], to identify co-expressed genes or to build gene networks [32-34], to evaluate reproducibility and bias across studies [35-37]. Figure 1.1 illustrates a microarray meta-analyses summary performed by Tseng and colleagues [27].

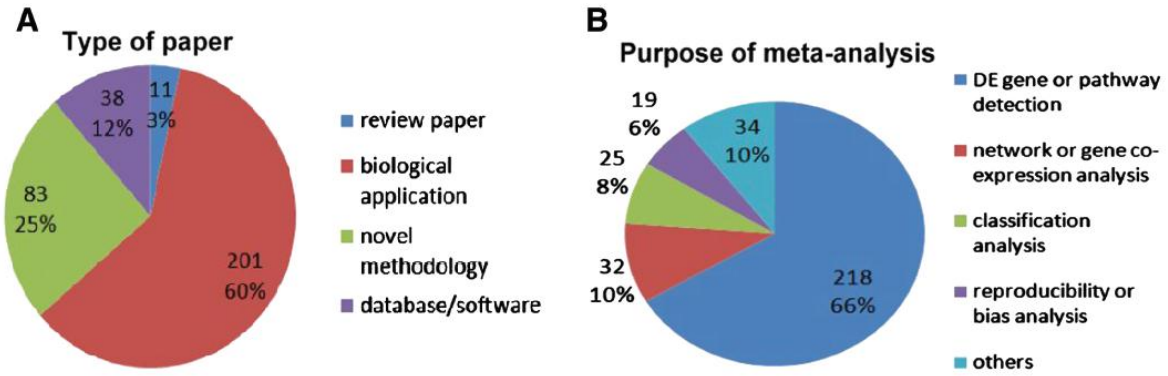

Figure 1.1: Classification of the 333 microarray meta-analysis papers reviewed by Tseng based on the type of paper (A) and the purpose of meta-analysis (B) (image modified from [27])

Despite its increasing popularity, however, meta-analysis of microarray data is not without problems. In fact, although it shares many features with traditional meta-analysis, most classical meta-analysis methods cannot be directly applied to microarray experiments because of their unique issues such as the large number of variables involved and the technical complexities of combining data across different experimental platforms (e.g. gene nomenclatures, species and analytical methods) [38].

1.1 AIM OF THE STUDY

In this study, we focused on the application of meta-analysis to the two-class comparison microarray experiments. The objective of this kind of studies is to identify DEGs between two well-defined conditions, namely cases and controls. Four statistical approaches were comparatively evaluated: the weighted version of the inverse normal method by Marot and Mayer [39] and the moderated effect size combination approach

proposed by Marot [40], both implemented in the R (http://www.r-project.org/) package metaMA, the t-based hierarchical modeling described in Choi et al. [16] and implemented in the Bioconductor (http://www.bioconductor.org/) package GeneMeta [41] and the rank product method with the RankProd Bioconductor package [42]. These methods were applied to a set of three publicly available microarray studies on malignant pleural mesothelioma to identify DEGs between normal and malignant mesothelioma pleural tissues. Since it is not yet clear if filtering is beneficial from a meta-analysis perspective, both unfiltered and filtered data were analyzed to evaluate the impact of a common filtering strategy on meta-analysis results.

2 BACKGROUND

A considerable literature has been published to guide the whole review process and the meta-analysis for medical and epidemiological studies [43-45]. Moreover, some guidelines for the reporting of systematic reviews and meta-analyses, outlined in the Quality of Reporting of Meta-Analyses statement for randomized trials by QUORUM group [46] and its evolution into PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [47], are universally accepted.

On the contrary, there is little guidance to carry out a meta-analysis of microarray datasets. The first attempt in this direction was represented by the paper of Ramasamy and colleagues who proposed a seven-step practical approach to conduct a meta-analysis of microarray datasets: "(1) Identify suitable microarray studies; (2) Extract the data from studies; (3) Prepare the individual datasets; (4) Annotate the individual datasets; (5) Resolve the many-to-many relationship between probes and genes; (6) Combine the study-specific estimates; (7) Analyze, present, and interpret results" [48]. Steps from 2 to 5 apply separately to the individual datasets.

Each step, in turn, consists of several critical points that will be highlighted and examined in detail in the following paragraphs.

2.1 IDENTIFICATION OF SUITABLE MICROARRAY STUDIES

A meta-analysis begins with a well-formulated objective. As highlighted in the introduction, meta-analysis of microarray studies can be used for several purposes, for example to identify DEGs between two or more groups, to identify co-expressed genes, to build gene networks or to evaluate reproducibility and bias across studies. In the following we will focus on meta-analysis for DEGs detection, however most of the considerations apply regardless of the specific topic.

The study selection process is guided by the definition of the inclusion/exclusion criteria. These criteria should be *a priori* established and should derive immediately from the objective(s) of the study. They can be based on biological (e.g. specific disease, type of outcome, type of tissues, organism) or technical issues (e.g., density of array, minimum number of arrays). A clear, detailed and unambiguous formulation of inclusion/exclusion criteria, possibly in the form of a real protocol, is essential to avoid the most frequent criticism of the meta-analysis, that is "mixing apples and oranges" [49].

Locating the studies is by far the most difficult and the most frustrating aspect of any meta-analysis but it is the most important and critical step. Many meta-analyses begin with a systematic literature search. Keywords concerning the research question and their synonyms are typically used to identify studies for inclusion in the review. In order to retrieve all the relevant studies on a given topic, the search should be as comprehensive as possible, therefore it is recommended to search all the

main electronic databases of abstracts listed in Table 2.1. Reading the latest review articles and contacting specific investigators that are known to be active in the area can help to identify additional studies missed by automated search and ongoing research efforts with unpublished data.

Database	Web site
Online repositories of abstracts	
PubMed	http://www.pubmed.gov/
Google Scholar	http://scholar.google.com/
Web of Science (requires subscription)	http://wos.mimas.ac.uk/
SCOPUS (requires subscription)	http://www.scopus.com/
Microarray repositories recommended by MIAME for mandatory data deposition	
Array Express	http://www.ebi.ac.uk/arrayexpress/
CIBEX	http://cibex.nig.ac.jp/
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
Other useful sites for data identification	
ONCOMINE	http://www.oncomine.org/
Stanford Microarray Database (SMD)	http://smd.stanford.edu/

Table 2.1: Useful web resources to identify suitable studies for microarray meta-analysis (modified from [48])

Concerning microarrays, it is appropriate to extend the search to public microarray data repositories, as well as to a few more specialized databases, listed in Table 2.1. A quick review of the abstracts and experiments description is essential to eliminate those studies that are clearly not relevant to the meta-analysis or do not meet the specified selection criteria.

After the identification of candidate studies from abstracts, the articles or inherent information from authors, where available, have to be retrieved to confirm their eligibility. To limit the risk of compromising the quality of meta-analysis results, the included studies should undergo a quality assessment, that is an accurate evaluation of the study characteristics in terms of the study design, implementation and analysis [49]. In fact, if a meta-analysis includes many low-quality studies, then the

errors in the primary studies will be carried over to the meta-analysis, where they may be harder to identify, and the obtained result will be biased ("garbage in, garbage out"). Regarding microarray studies, the quality assessment should be performed at the study-level as well as at the data-level, as will be extensively described in the following paragraphs.

2.2 EXTRACTION OF THE DATA FROM STUDIES

As illustrated in Figure 2.1, there are four levels of data arising from microarray analysis: (1) the scanned images, (2) the raw data or FLEO (Feature-Level Extraction Output) files [48], such as Affymetrix CEL and GenePix GPR files, that is the quantitative outputs from the image analysis software, (3) the Gene Expression Data Matrix (GEDM) arising from the application to raw data of preprocessing algorithms, which represents the gene expression summary for every probe and sample and (4) the list of genes that are declared as differentially expressed in the study.

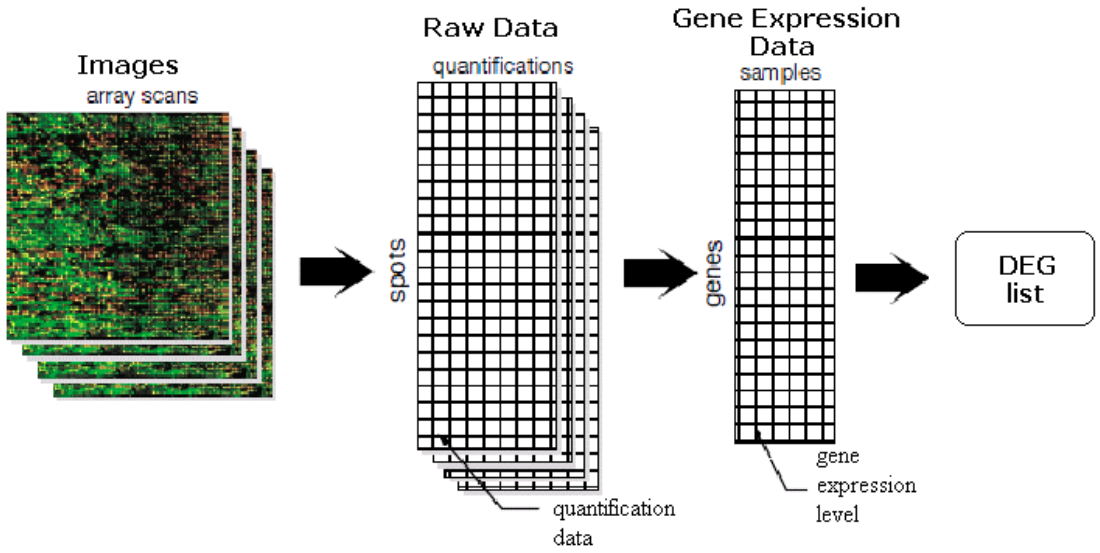

Figure 2.1: Types of data relevant to a microarray experiment (image modified from [50])

According to the conclusions of the study of Suarez-Farinas [51] and the recommendations of Ramasamy [48], raw data represent the ideal input for meta-analysis because they are independent of the specific preprocessing algorithms used and can be converted to GEDMs in a consistent manner thus producing more comparable data. By contrast, using GEDMs as input for meta-analysis is unsuitable because they considerably depend on the choice of the preprocessing algorithms, which may produce non-combinable results. The same considerations apply to the lists of DEGs. In fact, even if DEGs lists are easier to obtain since they are often included in the main text or supplementary data of published microarray studies, they heavily depend on the preprocessing algorithms, the statistical methods and cutoffs, and the annotation system adopted in the original study.

In relation to the data retrieval phase there are three major problems: (1) the efficient access to microarray data, (2) their standardization, and (3) the comparability across platforms.

2.2.1 MICROARRAY STANDARDS AND REPOSITORIES

In the past years, most of the publicly available microarray data produced by different research groups worldwide were scattered in the web both as supplementary data of a published article and as links to the authors web pages. Consequently it was very difficult for the researchers to locate and systematically collect the relevant data available. This problem has been addressed and partially solved through the development of several public repositories. Today, many web databases exist. ArrayExpress from EBI and GEO from NCBI are the two largest ones: on 24 January 2013, GEO contained 35618 experiments and 870318 samples while 35035 experiments and 1009648 assays were available in

ArrayExpress. Several other microarray databases are housed in specific universities or groups, including Stanford Microarray Database (SMD) and RNA Abundance Database (RAD; http://www.cbil.upenn.edu/RAD [52]) from University of Pennsylvania, or are focused on particular organisms (e.g. yeast Microarray Global Viewer; http://www.transcriptome.ens.fr/ymgv/ [53]) or diseases (e.g. ONCOMINE and Cancer Genome Workbench (CGWB) [54]) [27, 55].

At the beginning the effectiveness and the use of these public databases were severely limited by two factors: (1) the incompleteness or the lack of experimental information needed to assess the quality of the data, to repeat a study or to reanalyze the data, and (2) the lack of standards for presenting and exchanging such data. A considerable improvement occurred with the publication of the Minimum Information About a Microarray Experiment (MIAME) [50] standard by the Microarray Gene Expression Data Society (MGED) (http://www.mged.org). MIAME guidelines describe the minimum information that has to be provided to enable the comprehension of the results of a microarray experiment and their validation by independent researchers. The information required by MIAME standard includes the experimental design, array design (e.g. platform type and provider, gene identifiers, probe oligonucleotides), details on samples and treatments applied (e.g. laboratory protocols for sample treatments, extraction and labeling), hybridization conditions, measurements and normalization controls (e.g., normalization techniques applied and control elements used to obtain the final processed data). The current MIAME standard requires the submission to public databases of both the FLEO and GEDM files [56].

Since the MIAME publication in 2001, the major data repositories are supporting the archiving of MIAME-compliant data, and most peer-review

journals have adopted MIAME guidelines as a requirement for the paper publication [57]. The availability of standardized microarray gene expression data in these public repositories: (1) greatly enhanced the accessibility, the retrieval and the sharing of the data; (2) increased the reliability of the data quality and (3) improved the comparability and integration of data from different laboratories in a meta-analysis perspective.

Despite the wide adoption of MIAME standard by public microarray repositories and scientific journals, only about one-third of published studies have their raw data deposited in public databases [58]. Moreover, even when data are available, the incomplete annotation and/or the lack of data processing and analysis description limit their usefulness for further analyses [59].

2.2.2 CROSS-PLATFORM COMPARABILITY

One major issue in meta-analysis of microarray datasets concerns the possibility of combining raw measurements from different microarray technologies.

Although all DNA microarrays are based on the hybridization of complementary nucleic acid strands, the available platforms differ in the manufacturing process, hybridization protocols, image and data analysis, making comparison of the data across platforms very difficult.

Based on the length of the probes, microarrays can be classified as: (a) cDNA arrays, using probes constructed with PCR products of up to a few thousands base pairs, (b) short oligonucleotide arrays, using short probes (25-30 mer), such as Affymetrix GeneChip® arrays (Santa Clara, CA, USA), and (c) long oligonucleotide arrays, such as those produced by Agilent® (Palo Alto, CA, USA), using 60-70 mer long probes. Probe design

varies among microarrays. Long oligonucleotides are thought to mimic the properties of cDNA probes offering high sensitivity and good specificity, while giving better probe homogeneity. For both cDNA and long oligonucleotide arrays, typically one probe is designed for each gene that is to be probed [60]. In Affymetrix arrays, for each gene, a unique region is identified, then a set of 11–20 complementary probes spanning this region is synthesized. These complementary probes are referred to as 'Perfect Match' probes (PM). Each PM probe is then paired with a 'Mismatch' probe (MM), which has the same sequence as the PM except the central base replaced with a mismatched nucleotide. The complete set of PM and MM probe pairs for each gene is referred to as a 'probe set' [61].

Short oligonucleotides showed a higher specificity in target identification compared to long cDNA clones that were more prone to cross-hybridization [62].

Gene annotation can also contribute to platform differences. Gene expression values can be compared effectively across platforms only if genes are accurately identified on all platforms. Unfortunately, the lack of standardized annotation methods and of a regular update of annotations severely affect the cross-platform comparability. Moreover, the presence of poorly annotated and/or not specific probes on some arrays contribute to increase misalignments among platforms [63]. In any case, even if an accurate translation between different nomenclatures is achieved, the differences in how different platforms measure specific transcripts still remain and could have important impact on any attempt to conduct effective microarray data meta-analysis by increasing the false negative rate [38].

Based on the expression measurement techniques, microarray technologies can be classified as: one-color or single-channel and two-

color or two-channel. In one-color microarrays, such as Affymetrix arrays, a single labeled RNA sample is hybridized on a chip thus providing an absolute measurement of expression in the given sample (absolute quantification). By contrast, expression levels measured by cDNA microarrays and long oligonucleotide platforms, using two-channel detection, are usually reported as a ratio of the signal from a target RNA sample relative to one from a co-hybridized sample (relative quantification) [1]. These different measurement strategies result in diverse experimental designs which complicate the direct comparison and integration of the data. The use of a common reference design for the two-channel platforms, where each experimental RNA sample is co-hybridized with a reference RNA sample, represents a valid solution as it closely reproduces the single-channel approach.

Finally, different preprocessing steps, such as quality filtering, background correction and normalization, adopted to transform the raw data into the corresponding gene expression values, have substantial influence on the data [64].

All these differences produce qualitatively different data whose comparability has been widely debated. See for references [19, 65-71].

2.3 PREPARATION OF THE INDIVIDUAL DATASETS

Once the raw data from individual studies have been collected, they have to be converted into GEDMs, which can then be used as input for the meta-analysis.

Before the preprocessing or transformation steps, Ramasamy [48] suggests to check the quality of the arrays in the individual studies to identify and remove those of poor quality. Microarray quality is assessed

by comparing suitable numerical summaries (e.g. average background, scale factors, percentage of present calls,) across microarrays, so that outliers and trends can be visualized and poor-quality arrays can be identified. There are many Bioconductor packages for quality assessment including arrayMagic [72] for the two-color technology platform, Simpleaffy [73] and affyPLM [74] for the Affymetrix platform and ArrayQualityMetrics [75] which manages many microarray technologies. Only the arrays that pass the quality check should be included in the meta-analysis.

At this point the data undergo different levels of transformation or preprocessing that are: background or mismatch subtraction, probe set summarization which combines multiple measures of the same transcript, normalization within and between arrays. As it is now widely known [76], using different raw data transformation methods leads to disagreements in the resulting DEGs even within one experiment on a single platform.

It is thus evident the need to consistently process the data to remove any systematic differences. The simplest case is when data from multiple studies the same platform have to be combined. In this case it is, in fact, sufficient to apply the same algorithm to all datasets. Much more often, however, researchers are faced with the problem of combining datasets from different platforms, which may have different designs and thus different preprocessing methods options. In this case, comparable preprocessing algorithms should be applied to the individual datasets. There are very few universally applicable preprocessing algorithms, such as the variance stabilizing normalization [77]. By contrast, it is more common to use different preprocessing methods for each platform. Unfortunately, there is currently no consensus on which preprocessing algorithms produce comparable expression measurements across

different platforms [48]. However, it has been found that while the default procedures suggested by microarray manufacturers result in general slightly better accuracy, the results provided by alternative approaches, like those proposed by Bioconductor packages, are far more precise [70].

The identification and adjustment of any batch effects, especially in large microarray datasets, are also of great importance. Many different experimental features can cause biases including different sources of RNA, different microarrays print batches or platforms, as shown in Figure 2.2. Unsupervised visualization techniques such as Support Vector Machines (SVM) [78], Singular Value Decomposition (SVD) [79], Principal Component Analysis (PCA) [80] and the Distance Weighted Discrimination (DWD) method proposed by Benito and colleagues [81] can help to identify any grouping caused by experimental factors within microarray datasets.

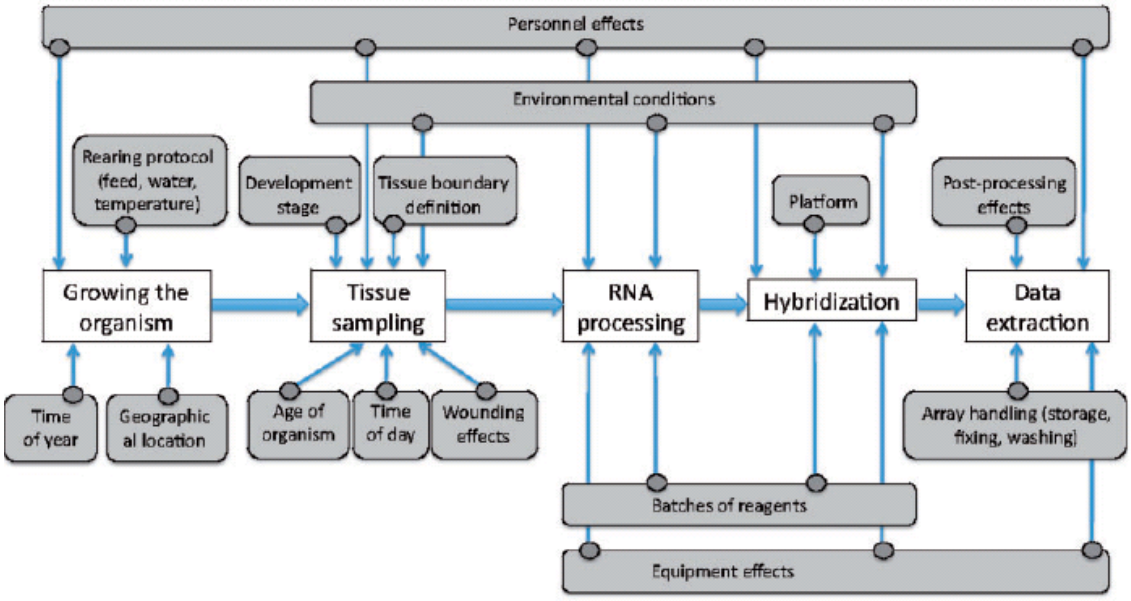

Figure 2.2: A visualization of batch effect sources at each stage of a microarray gene expression experiment (image from [82])

In single-study analysis it is common practice to filter out probes based on different criteria. Probes showing severe manufacturing or hybridization problems or a signal-to-noise ratio below a fixed threshold,

probes marked as ‘absent’ or showing little variation among experimental conditions are usually excluded or under-weighted from the successive analysis. To date, it has been demonstrated that filtering improves cross-platform reproducibility [21, 64, 67, 83] but it is not yet clear whether filtering is beneficial to meta-analysis.

Another problem that may occur in this phase deals with the management of possible within studies technical replicates. In fact, technical replicates cannot be considered as independent observations and should be aggregated taking, for example, the mean or median of the corresponding gene expression measurements.

Finally, one could check that the processed expression values from multiple platforms are comparable. Concerning this topic, one may use visualization techniques such as multidimensional scaling [84] to investigate for any clustering of arrays by studies.

2.4 ANNOTATION OF THE INDIVIDUAL DATASETS

The first step to combine different microarrays datasets is to find genes common to all arrays. The annotation of the individual datasets is a non trivial task because of the lack of a uniform nomenclature system and the many-to-many relationship between probes and genes.

Microarray manufacturers use specific probe-level identifiers (probe IDs) (e.g. Affymetrix probe ID) to identify the probes present on their own arrays. Moreover, different manufacturing techniques lead to the creation of multiple probes for the same gene. Therefore, one needs to identify which probes represent a given gene within and across platforms. In fact, even if the datasets share the same platform, the combination of different array versions creates serious difficulties, since the probe IDs are not

conserved from version to version. In conclusion, to combine microarray datasets across studies a unique nomenclature must be adopted and all the different platform-specific IDs must be translated to a common identifier. There are many different options that could be used to this end. Genbank or RefSeq [63] accession number, Unigene ID [85] and Entrez ID [86] are the most common. Accession numbers are associated with specific transcripts, then there may be multiple per gene. Mapping between platforms on the basis of the accession number could produce an accurate result, as one can be confident that the probes are truly measuring the same entity; however, such an approach would be problematic as there would be many accession numbers for which probes only exist on one platform, greatly diminishing the ability to map between platforms. For this reason, mapping on the gene level is the most common choice. This allows to incorporate the information from many more probes, as it is much more likely to be able to find some probes associated with a gene for each platform than to find a probe associated with a specific accession number. Unigene and Entrez Gene have different strengths and weaknesses. While Unigene IDs may incorporate more information, it is very dynamic and is constantly being revised. Entrez IDs, on the other hand, are very stable and have been well-curated [87].

The problem of matching platform-specific probe IDs can be tackled in three ways. The traditional method is to use the annotation files provided by the manufacturers. The accuracy of these files was long criticized as the knowledge of the transcriptome is constantly growing. However, in recent years more and more manufacturers provide to release updated annotation files with varying degrees of regularity in an attempt to keep these annotations current.

Another option is to align the probe sequences provided by the vendors to a recent revision of either the Genome or the Transcriptome using the BLAST algorithm [88], trying to obtain more up-to-date gene-to-probe associations. It has been shown that cross-platform correlations improved using stringent sequence matching of the probes on the different platforms [89]. However, the probe sequences are not always available and this procedure can be computationally intensive and time-consuming for very large numbers of probes.

Alternatively, one can simply map probe IDs to a gene-level identifier (gene ID) such as Entrez ID or UniGene ID. Many published microarray meta-analyses [24, 26, 51, 90] have relied on UniGene ID to unify the different datasets, across platforms and array versions. The translation of the probe IDs to the corresponding gene IDs can be performed using either some Bioconductor annotation packages (e.g. annotate [Gentleman R. annotate: Annotation for microarrays. R package version 1.36.0.], annotationTools [91]) that aggregate the information from various platform-specific Bioconductor packages, or Web tools such as SOURCE [92] and RESOURCERER [93], MADGene [94], DAVID converter [95] and Onto-Translate [96]. The same mapping build, ideally the most recent, should be used for all datasets to avoid inconsistencies between releases [48].

Allen and colleagues [87] found that a BLAST alignment of the probes to the Transcriptome was more accurate than using the vendor's annotation or Bioconductor packages. They also proposed a combination of all three methods (the "Consensus Annotation") showing that it yielded the most consistent expression measurements across platforms.

The annotation of individual datasets is further complicated by the non univocal relationship between probes and genes, which means that in

some cases a probe could report to more than one gene and *vice versa*. Many probes can map to the same gene ID because of the clustering nature of the UniGene, RefSeq, and BLAST systems involved, or because the microarrays used contain duplicated probes. *Vice versa*, a probe may map to more than one gene ID if the probe sequence is not specific enough. Sometimes, a probe has insufficient information to be mapped to any gene ID. These probes should be removed from further analysis.

The simplest and even most stringent approach to solve these confounding situations is to use only the probes with one-to-one mapping for further analysis, thus excluding probes without a gene ID, probes mapping to multiple gene IDs and probes mapping to the same gene ID. Alternatively, probes with multiple gene IDs may be considered as independent gene expression measurements and be replaced by a new record for each gene, while multiple probes mapping to the same gene ID can be summarized using one of the following options: (1) selecting a probe at random, (2) taking the average of expression values across multiple probe IDs to represent the corresponding gene, (3) choosing the probe ID with the largest Inter Quartile Range (IQR) (or other similar statistics, such as standard deviation or coefficient of variation) of expression values among all multiple probe IDs to represent the gene. Although the option number 2 has been widely used due to its simplicity, IQR method is biologically more reasonable and robust and is highly recommended [97].

Recently, the MicroArray Quality Control (MAQC) project proposed another alternative. A single RefSeq ID was selected for each probe mapping to multiple RefSeq IDs, primarily the one annotated by TaqMan assays, or secondarily the one present in the majority of platforms. When a platform contained multiple probes matching the same RefSeq entry,

only the probe closest to the 3’ end of the RNA sequence was included [71].

The multiple gene expression datasets may not be very well aligned by genes and the number of genes in each study may be different. Therefore the common genes across multiple studies have to be identified and extracted. When a large number of studies were included in the meta-analysis, the number of genes common to all studies may be very small. At this point there are two possibilities: using only genes appearing in all datasets, or including also genes appearing in at least a pre-specified number of studies.

Having solved the many-to-many relationship by expanding and summarizing probes, one summary statistic per gene ID per study is available. The next step will be to combine the summary statistic for each gene ID across the studies using a meta-analysis technique.

3 STATISTICAL METHODS FOR MICROARRAY DATA META-ANALYSIS

This chapter deals with the sixth step of Ramasamy's guidelines for microarray meta-analysis. The choice of a meta-analysis method depends on the type of outcome (e.g. binary, continuous, survival), the objective of the study and the type of available data. As previously illustrated, we focused on the two-class comparison, the most commonly encountered application of meta-analysis to microarray data, whose aim is the detection of DEGs between two experimental groups or conditions.

There are two principal approaches to perform a meta-analysis, the relative and the absolute approach [98]. The relative meta-analysis is the most common one and is based on the calculation of a relative score expressing how each gene correlates to the experimental condition or phenotype of interest in each dataset. These scores are used to quantify the differences or similarities among studies and are integrated to find overall results (see Figure 3.1).

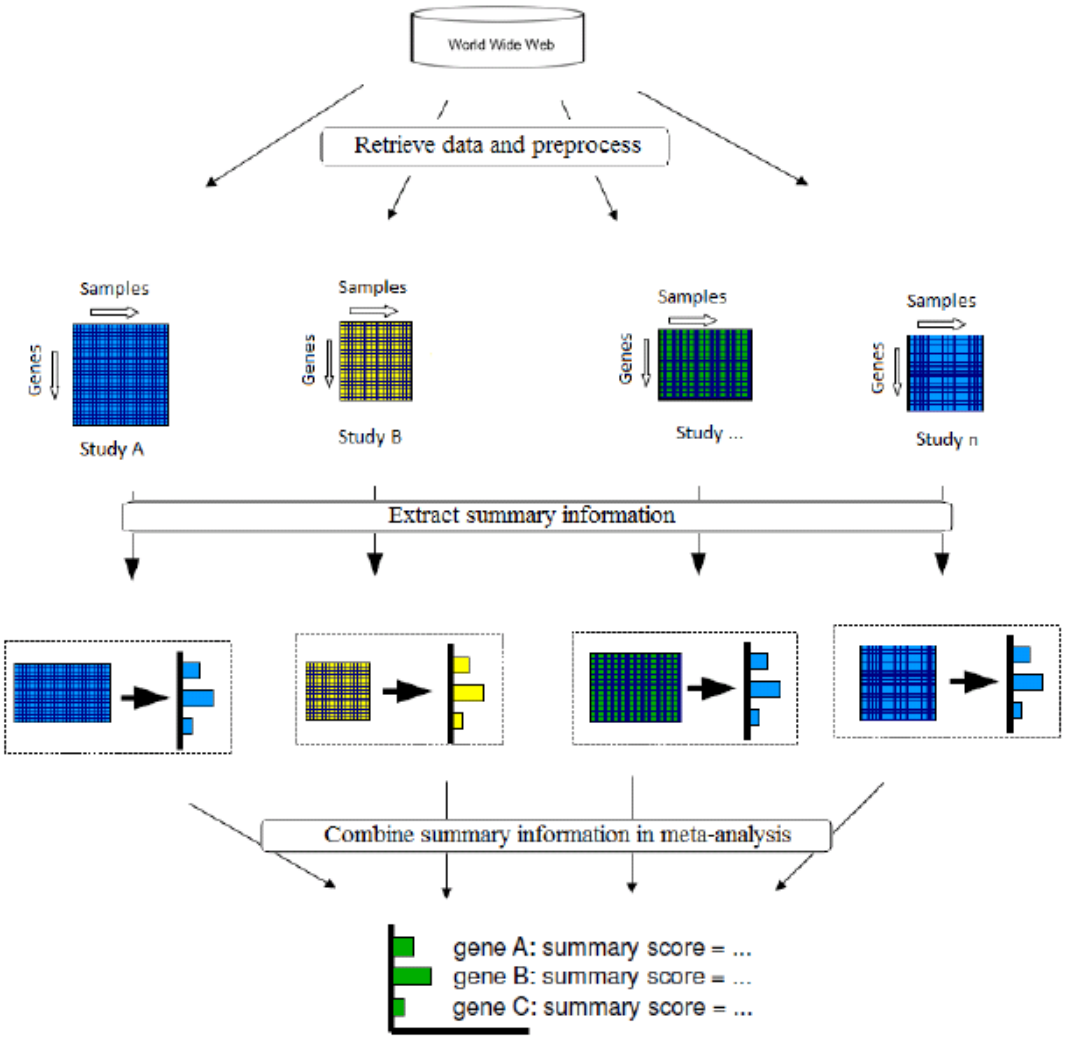

Figure 3.1: Stages of relative meta-analysis of microarray data (image from [13])

In contrast, in the absolute meta-analysis raw data from various microarray studies are integrated after transforming the expression values to numerically comparable measures. The derived data from the individual studies are normalized across studies and subsequently merged, thus enlarging the sample size and increasing the power of statistical tests. Traditional microarray data analysis is then carried out on the new merged dataset (see Figure 3.2) [13, 99].

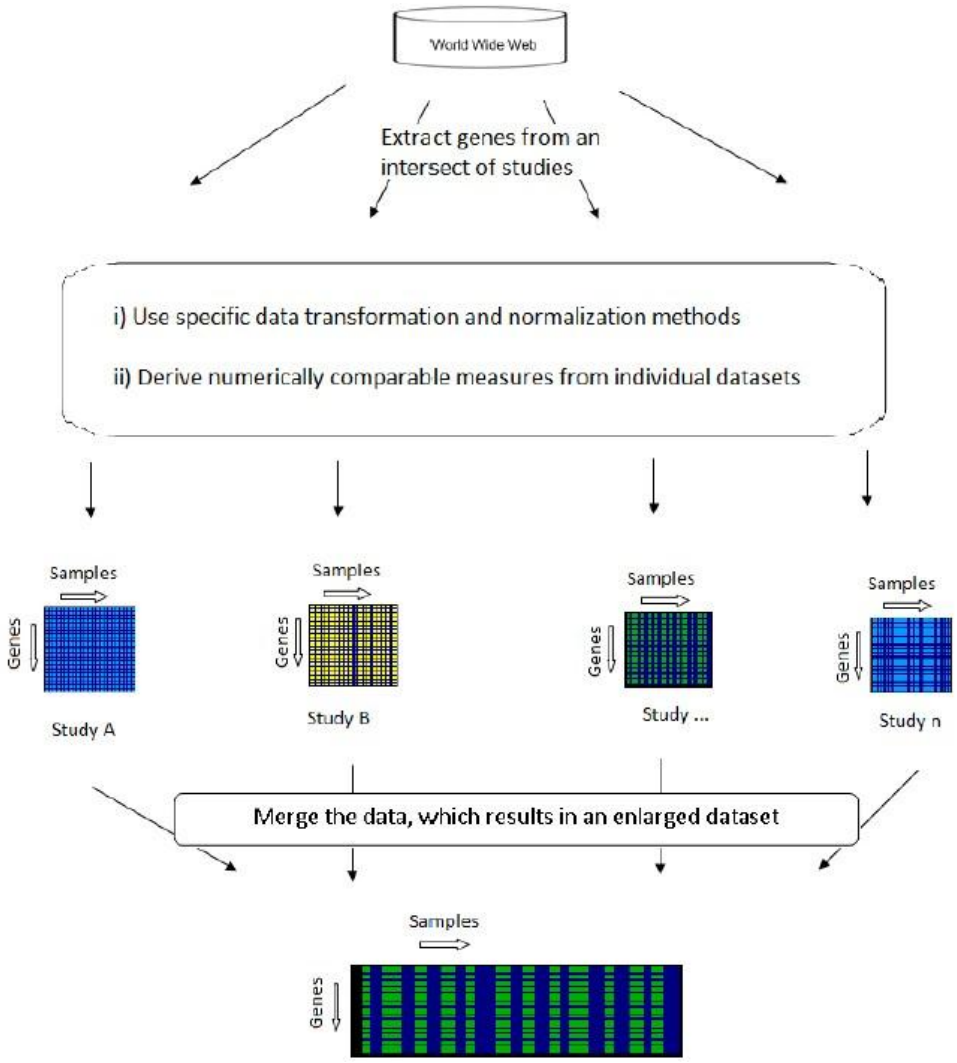

Figure 3.2: Stages of absolute meta-analysis of microarray data (image from [13])

Although merging data can be attractive for its intuitiveness and convenience, cautions have to be taken since normalizations do not guarantee to remove all cross-study differences. There are few examples of studies where the absolute meta-analysis has been applied [31, 100, 101]. Contrary to relative meta-analysis, which is always possible for cross lab, platform and even species comparisons, absolute meta-analysis usually considers studies from the same or similar array platform [102, 103]. The collection of datasets from only one platform allows to pre-process and normalize data using the same method on all samples simultaneously.

In recent years several relative meta-analysis methods have been proposed using different approaches. There are four generic ways of combining information for DEGs detection:

1. Vote counting
2. Combine p-values
3. Combine effect sizes
4. Combine ranks.

They differ in the type of statistics measures proposed to summarize the study results.

3.1 VOTE COUNTING

Vote-counting is the simplest of the above approaches. For each gene, vote counting simply counts the number of studies in which a gene has been claimed significant [104]. To provide a statistical basis to vote counting techniques results, one can either calculate the significance of the overlaps using the normal approximation to binomial as described in Smid and colleagues [105] or calculate the null distribution of votes using random permutations [24]. For very small numbers of studies (usually 2–4), the results can be summarized using a Venn diagram which displays the intersection and union distribution of DEGs lists detected by each individual study. In literature, it is well known that vote counting is statistically inefficient [14]. Moreover, vote counting does not yield an estimate of differential expression extent and the results highly depend on the statistical methods used in individual analyses. On the other hand, vote counting is useful when raw data and/or p-values for all genes are not accessible while only the lists of DEGs are available for each study.

Vote counting in the context of microarrays has been used successfully by Rhodes and colleagues [24], who applied it to identify a shared gene expression signature across cancer subtypes. First t-tests were calculated by comparing the treatment and control group in each study. Then, a binary score was assigned to each gene in each study based on whether its p-value passed a threshold (vote = 1) or not (vote = 0). Finally, a simulation of the likelihood of obtaining k or fewer votes (where k is the number of studies included) was done to estimate a significance level.

3.2 COMBINING P-VALUES

Combining p-values from multiple studies for information integration has long history in statistical science. Methods based on the combination of p-values are easy to use and provide more precise estimates of significance. However these methods do not indicate the direction (e.g. up or down regulation) nor the extent of differential expression. Moreover the results highly depend on the statistical methods used in individual analyses. Nevertheless, integration of p-values does not require that different studies use the same measurement scales therefore it is possible to combine results from studies realized by completely different technologies.

Several methods exist for combining p-values from independent tests; below, four p-value combination methods used in the context of microarray meta-analysis are briefly described.

3.2.1 FISHER'S METHOD

Fisher's method [106] computes a combined statistics from the log-transformed p-values obtained from the analysis of the individual datasets:

$$(1) \qquad \boldsymbol{S_g = -2log\left(\prod_{i=1}^{k} p_{gi}\right)}$$

where p_{gi} is the unadjusted p-value from one-sided hypothesis testing for gene g and study i and k being the number of individual combined studies.

The meta-analysis null hypothesis is that all the separate null hypotheses, are true, whereas the alternative hypothesis is that at least one of the separate alternative hypotheses is true. Assuming independence among studies and p-values calculated from correct null distributions in each study, S_g follows a chi-square distribution with 2k degrees of freedom under the joint null hypothesis of no differential expression, thus p-values of the combined statistics can be calculated for each S_g. Alternatively, statistical inference can be done non-parametrically using a permutation approach. As there are many genes, p-values of the summary statistics must be corrected for multiple testing using one of the available procedures such as the Bonferroni correction, the false discovery rate (FDR) proposed by Benjamini and Hochberg [107] or its modified version proposed by Storey [108]. Finally, a threshold is chosen and two meta-lists, that are the lists of DEGs resulting from the meta-analysis, of over and under-expressed genes are reported. It is worth pointing out that Fisher's product should be applied to p-values for up and down regulation separately. Using p-values from two-sided testing means ignoring the direction of the significance and may lead one to select genes that are discordant in direction of gene regulation between the studies [109].

Rhodes and colleagues [23] were the first who applied Fisher's method to microarray data. They identified a meta-signature of prostate cancer combining the results of four studies performed on different platforms.

Some variations to Fisher's method have been proposed that give different weights to p-values from each dataset. Weight assignment can depend on the reliability of each p-value based on the data quality. Recently, Li and Tseng [110] introduced an adaptively weighted Fisher's method (AW) where the weights are calculated according to whether or not a study contributes to the statistical significance of a gene. Li and Tseng showed the superior performance, in terms of power, of their AW statistics compared to Fisher's equally weighted and other p-values combination methods, like Tippett's minimum p-value [111] and Pearson's (PR) statistics.

3.2.2 STOUFFER'S METHOD

Instead of log-transformation, Stouffer's method [112] uses the inverse normal transformation. Unlike the Fisher's method, which requires to treat over and under-expressed genes separately, the inverse normal method is symmetric in the sense that p-values near zero are accumulated in the same way as p-values near one [14]. In the Stouffer's method, the one-sided p-values for each gene g from k individual studies are transformed into z scores and then combined using the following expression:

(2) $$\boldsymbol{Z_g = \sum_{i=1}^{k} \Phi^{-1}(1 - p_{gi})/\sqrt{k}}$$

where $\phi^{-1}(\ )$ is the inverse cumulative distribution function of standard normal distribution. Under the null hypothesis, the z statistic follows a normal N(0,1) distribution and therefore a p-value for each Z_g

can be calculated from the theoretical normal distribution. Finally, to take into account multiple comparisons, the FDR or other multiple testing correction methods can be applied. An alternative to (2) is to use the weighted method proposed by Marot and Mayer [39] which is implemented in the R package metaMA. Here:

$$(3) \qquad \boldsymbol{Z_g = \sum_{i=1}^{k} w_i \Phi^{-1}(1 - p_{gi})/\sqrt{k}}$$

and

$$(4) \qquad \boldsymbol{w_i = \sqrt{n_i}/(\sum_{i=1}^{k} n_i)}$$

being n_i the sample size of study i.

3.2.3 MINP AND MAXP METHODS

In the minP [111] and maxP [113] methods, for each gene, the minimum or maximum p-values over different datasets are taken as the test statistics. Smaller minP or maxP statistics reflects stronger differential expression evidence, however while minP declares a gene as differentially expressed if it is in any of the studies, maxP tends to be more conservative considering as differentially expressed only genes that have small p-values in all studies combined.

Combining p-values are techniques that in theory could use the published lists of DEGs, but may not be able to do so in practice. For example, most publications report the significant genes based on two-sided p-values, while the aforementioned methods require one-sided p-values. So it is preferable to use the raw data to minimize the influence of different methods across datasets.

3.3 COMBINING EFFECT SIZES

Methods based on the combination of effect sizes have been the most common approach to the meta-analysis of microarray studies. In statistics, an effect size is a measure of the strength of a phenomenon [114] (e.g. the relationship between two variables in a statistical population) or a sample-based estimate of that quantity. In general, effect sizes can be measured in two ways:

1. as the standardized difference between two means, or
2. as the correlation between the two variables [115].

Standardized Mean Difference (SMD) is the difference between two means, divided by the variability of the measures. Effect sizes based on SMD include Cohen's *d* [116], Hedges' *g* [14], and Glass's *delta*. All three employ the same numerator (i.e. the difference between group means) but different estimates of the variability at the denominator [117].

An effect size approach is effective for microarray data application. First it provides a standardized index. At present, the measure of expression levels is not interchangeable in particular between oligonucleotide arrays and cDNA arrays. cDNA microarrays report only the relative change compared to a reference, which is rarely standardized. Obtaining effect sizes facilitates the combining of signals from one-color and expression ratios from two-color technology platforms. Second, it is based on a well-established statistical framework for the combination of different results. Third, it is superior to other meta-analytic methods in that it has the ability to manage the variability between studies. Moreover, in comparison to the p-values summary approaches, combining effect sizes gives information about the magnitude and direction of the effect.

In meta-analysis, the basic principle is to calculate the effect sizes in individual studies, convert them to a common metric, and then combine them to obtain an average effect size. Once the mean effect size has been calculated it can be expressed in terms of standard normal deviates (Z score) by dividing the mean difference by its standard error. A significance p-value of obtaining the Z score of such magnitude by chance can them be computed.

Without loss of generality, we can assume that we are comparing two groups of samples, such as treatment (t) and control (c) groups, in each study i=1,2,..k. For each study i, let n_{it} and n_{ic} denote the number of samples in treatment and control group, respectively, with $n_i = n_{it} + n_{ic}$. Let x_{gitr} and x_{gicr} represent the raw expression values for gene g in conditions t and c for study i and replicate r and y_{gitr} and y_{gicr} be the corresponding log-transformed values. The data are assumed to be normally distributed as $y_{gitr} \sim N(\mu_{git}, \sigma_{gi}^2)$ and $y_{gicr} \sim N\left(\mu_{gic}, \sigma_{gi}^2\right)$. In a microarray experiment with two groups, the effect size refers to the magnitude of difference between the two groups' means. There are many ways to measure effect size for gene g in any individual study [118]. The SMD proposed by Cohen is defined as:

(5) $$\boldsymbol{d_{gi}} = \left(\boldsymbol{\bar{y}_{git}} - \boldsymbol{\bar{y}_{gic}}\right) / \boldsymbol{s_{gi}^{pool}}$$

where $\bar{y}_{git}$ and $\bar{y}_{gic}$ are the sample means of logged expression values for gene g in treatment (t) and control (c) group, in the i*th* study, respectively and s_{gi}^{pool} is the pooled standard deviation:

(6) $$\boldsymbol{s_{gi}^{pool}} = \sqrt{\frac{(\boldsymbol{n_{it}}-\boldsymbol{1})\boldsymbol{s_{git}^2}+(\boldsymbol{n_{ic}}-\boldsymbol{1})\boldsymbol{s_{gic}^2}}{(\boldsymbol{n_{it}}+\boldsymbol{n_{ic}})}}$$

where s_{git}^2 and s_{gic}^2 denote the sample variances of gene g's expression level in the treatment and control groups, respectively.

Alternatively the SMD proposed by Hedges and Olkin [14] may be used. Hedges and Olkin showed that the classical Hedges'*g* overestimates the effect size for studies with small sample sizes. They proposed a small correction factor to calculate an unbiased estimate of the effect size which is known as the Hedges' adjusted *g* and is given by:

$$(7) \qquad \boldsymbol{d'_{gi} = d_{gi}\left(1 - \frac{3}{4(n_i-2)-1}\right)}$$

The estimated variance s_{gi}^2 of the unbiased effect size is given by:

$$(8) \qquad \boldsymbol{s_{gi}^2 = \left(n_{it}^{-1} + n_{ic}^{-1}\right) + d_{gi}^2\left(2(n_{it} + n_{ic})\right)^{-1}}$$

Then the effect size index d_{gi} (or its unbiased version) across studies is modeled by a hierarchical model:

$$(9) \qquad \begin{cases} \boldsymbol{d_{gi} = \theta_{gi} + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim N(0, s_{gi}^2)} \\ \boldsymbol{\theta_{gi} = \mu_g + \delta_{gi}, \quad \delta_{gi} \sim N(0, \tau_g^2)} \end{cases}$$

where μ_g is the average measure of differential expression across datasets for each gene g, which is typically the parameter of interest, τ_g^2 is the between-study variance, which represents the variability between studies, and s_{gi}^2 is within-study variance, which represents the sampling error conditioned on the i*th* study. The model has two forms: a fixed effect model (FEM) and a random effect model (REM), and the choice depends on whether between-study variation is ignorable. A FEM assumes that there is one true effect μ_g common to all studies included in a meta-analysis and that all differences in observed effect sizes are due to sampling error alone. Thus τ_g^2 = 0 and consequently $d_{gi} \sim$ N(μ_g, s_{gi}^2). By

contrast, in REM each study further contains a random effect that can incorporate unknown cross-study heterogeneity in the model. Thus $d_{gi} \sim$ N(θ_{gi} , s_{gi}^2) and $\theta_{gi} \sim$ N(μ_g, τ_g^2).

To determine whether FEM or REM is most appropriate, the Cochran's Q statistic [119] may be used to test homogeneity of study effect, which is assessing the hypothesis that τ_g^2 is zero. Q statistic is defined as:

$$\boldsymbol{Q_g = \sum_{i=1}^{k} w_{gi}(d_{gi} - \hat{\mu}_g^F)^2} \quad (10)$$

where $w_{gi} = s_{gi}^{-2}$ is the statistical weight and

$$\boldsymbol{\hat{\mu}_g^F = \frac{\sum_{i=1}^{k} w_{gi} d_{gi}}{\sum_{i=1}^{k} w_{gi}}} \quad (11)$$

is the weighted least squares estimator of the average effect size under the FEM which ignore the between-study variance. Under the null hypothesis of homogeneity (i.e. τ_g^2 = 0), Q follows a chi-square distribution with k-1 degree of freedom. A large observed value of the Q statistics relative to this distribution suggests the rejection of the hypothesis of homogeneity, which should indicate the appropriateness of the REM. It must be noted that this homogeneity test has low power [120] and non-significant results do not imply that true homogeneity exists. If the null hypothesis of τ_g^2 = 0 is rejected, one method for estimating τ_g^2 is the method of moments developed by DerSimonian and Laird [121]:

$$\boldsymbol{\hat{\tau}_g^2 = max\left\{0, \frac{Q_g - (k-1)}{\sum_{i=1}^{k} w_{gi} - (\sum_{i=1}^{k} w_{gi}^2 / \sum_{i=1}^{k} w_{gi})}\right\}} \quad (12)$$

Then τ_g^2 is used to estimate the parameter μ_g and its variance by a point estimator defined as in the generalized least squares method:

(13) $$\widehat{\boldsymbol{\mu}}_{\boldsymbol{g}}^{\boldsymbol{R}}(\boldsymbol{\tau}_{\boldsymbol{g}}^{\boldsymbol{2}}) = \frac{\sum_{i=1}^{k} \boldsymbol{w}_{gi}^{*} \boldsymbol{d}_{gi}}{\sum_{i=1}^{k} \boldsymbol{w}_{gi}^{*}}$$

and

(14) $$\boldsymbol{Var}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{g}}^{\boldsymbol{R}}(\boldsymbol{\tau}_{\boldsymbol{g}}^{\boldsymbol{2}})\right) = \frac{\mathbf{1}}{\sum_{i=1}^{k} \boldsymbol{w}_{gi}^{*}}$$

where $w_{gi}^{*} = (s_{gi}^{2} + \hat{\tau}_{g}^{2})^{-1}$ is the statistical weight under REM.

The z statistic to test for DEGs under REM is then constructed as follows:

(15) $$\boldsymbol{Z}_{\boldsymbol{g}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{g}}^{\boldsymbol{R}}(\boldsymbol{\tau}_{\boldsymbol{g}}^{\boldsymbol{2}}) / \sqrt{\boldsymbol{Var}\left(\widehat{\boldsymbol{\mu}}_{\boldsymbol{g}}^{\boldsymbol{R}}(\boldsymbol{\tau}_{\boldsymbol{g}}^{\boldsymbol{2}})\right)}$$

The z statistic for FEM is the same as that for REM except that τ_{g}^{2} = 0. To evaluate the statistical significance of the combined results, the p-values can be obtained from a standard normal distribution N(0,1) using these Z scores. For a two-tailed test, the p-value for each gene is given by:

(16) $$\boldsymbol{p}_{\boldsymbol{g}} = \mathbf{2}\left(\mathbf{1} - \boldsymbol{\Phi}(|\boldsymbol{Z}_{\boldsymbol{g}}|)\right)$$

where $\Phi(Z)$ is the standard normal cumulative distribution. To assess the statistical significance not assuming normal distribution, empirical distributions may be generated by random permutations. In both cases, the p-values obtained are unadjusted values which should be corrected to take into account the multiple comparisons.

Choi and colleagues [16] were among the first who applied these models to microarray meta-analysis. To estimate the effect size they considered the unbiased estimator of the SMD defined in equation (7) where d_{gi} was obtained from the standard t statistics for each gene from each individual dataset via the relationship:

(17) $$\boldsymbol{d_{gi} = t_{gi}/\sqrt{\widetilde{n_i}}}$$

with $\widetilde{n_i} = (n_{it} * n_{ic})/((n_{it} + n_{ic})$.

To estimate the statistical significance simultaneously addressing the multiple testing problem, Choi and colleagues adapted the core algorithm of Significance Analysis of Microarrays by Tusher and colleagues [122]. Column-wise permutations were performed within each dataset to create randomized data and z scores under the null distribution, z_g^{*b} for permutation b = 1, 2, …B. The ordered statistics $z_{(g)}$ ($z_{(1)} \leq \cdots \leq z_{(p)}$) and $z_{(g)}^{*b}$ ($z_{(1)}^{*b} \leq \cdots \leq z_{(p)}^{*b}$) were obtained and the FDR was estimated for a given gene by:

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

where I(·) is the indicator function equal to 1 if the condition in parentheses is true, and 0 otherwise. The denominator represents the number of genes called significant in real data. The numerator is the expected number of falsely significant genes and given by the mean number across B permuted data. Integration of data using this meta-analysis method facilitated the discovery of small but consistent expression changes and increased the sensitivity and reliability of analysis. Later, Hong and Breitling [109] found that this t-based meta-analysis method greatly improved over the individual analysis, however it suffered from potentially large amount of false positives when p-values served as threshold.

The approach of Choi and colleagues has been implemented in the Bioconductor package GeneMeta [Lusa L, Gentleman R and Ruschhaupt M, GeneMeta: MetaAnalysis for High Throughput Experiments. R package

version 1.30.1] where both alternatives to evaluate the statistical significance of the combined results are available.

Different variations of effect size models have also been developed by other research groups. Hu and colleagues [123] presented a measure to quantify data quality for each gene in each study where the quality index measured the performance of each probe set in detecting its intended target. As they used Affymetrix microarrays they exploited the detection p-values provided by Affymetrix MAS 5.0 algorithm [Affymetrix Microarray Suite User's Guide Version 5.0 Affymetrix, Santa Clara, CA; 2001] to define a measure of quality for each gene in each study and incorporated these quality scores as weights into a classical random-effects meta-analysis model. They demonstrated that the proposed quality-weighted strategy produced more meaningful results then the unweighted analysis. In a later paper, Hu and colleagues [124] proposed a re-parameterization of the traditional mean difference based effect size by using the log ratio of means, that is, the log fold-change, as an effect size measure for each gene in each study. They replaced the effect size defined in equation (5) with the following expression:

$$\boldsymbol{d_{gi} = ln\left(\bar{x}_{git}/\bar{x}_{gic}\right)} \quad (19)$$

where $\bar{x}_{git}$ and $\bar{x}_{gic}$ are the sample means of the unlog-transformed gene expression values for gene g in treatment and control group in a given study. The estimated variance s_{gi}^2 of this new effect size can be estimated as follows:

$$\boldsymbol{s_{gi}^2 = \frac{1}{n_{it}}\frac{s_{git}^2}{\bar{x}_{git}^2} + \frac{1}{n_{ic}}\frac{s_{gic}^2}{\bar{x}_{gic}^2}} \quad (20)$$

Redefined d_{gi} and s_{gi}^2 were then placed into the classical hierarchical model (9) using both the quality-weighted and quality-unweighted frameworks. Hu and colleagues' idea comes from two well-known evidences. On the one hand the fact that, with small sample sizes, the traditional standard mean difference estimates are prone to unpredictable changes, since gene-specific variability can easily be underestimated resulting in large statistics values. Many efforts have been made to overcome this problem by estimating a penalty parameter for smoothing the estimates using information from all genes rather than relying solely on the estimates from an individual gene [122]. On the other hand the evidence that DEGs may be best identified using fold-change measures rather than t-like statistics [125]. Hu and colleagues applied their method to simulated datasets and real datasets focusing on the identification of differentially expressed biomarkers and their ability to predict cancer outcome. Their results showed that the proposed effect size measure had better power to identify DEGs and that the detected genes had better performance in predicting cancer outcomes than the commonly used standardized mean difference.

Stevens and Doerge [126] proposed an alternative for the SMD as estimator for differential expression specific for Affymetrix data. It is represented by the signal log ratio (SLR) automatically reported by MAS 5.0 [Affymetrix Microarray Suite User's Guide Version 5.0 Affymetrix, Santa Clara, CA; 2001], defined as the signed $\log_2$ of the signed fold-chance (FC), that is, $FC=2^{SLR}$ if $SLR\geq 0$ and $FC=(-1)2^{-SLR}$ if $SLR<0$. The meta-analytic framework is described in Choi and colleagues [16].

Unlike previously mentioned meta-analysis studies where the p-values or effect sizes to be combined were based on standard t-tests, Marot and colleagues [40] proposed to extend these effect sizes to

account for moderated t-tests. In the last few years, several authors such as Smyth [127] or Jaffrézic and colleagues [128] showed that, in single study analyses, shrinkage approaches leading to moderated t-tests were more powerful to detect DEGs than gene-by-gene methods when small numbers of biological replicates are available. Indeed, shrinkage consists in estimating each individual gene value borrowing information from all the genes involved in the experiment. By decreasing the total number of parameters to estimate, the sensitivity is increased. Marot and colleagues considered two popular shrinkage approaches: that proposed by Smyth [127] and implemented in the Bioconductor package limma and that developed by Jaffrézic and colleagues [128] implemented in the R package SMVar. In the first approach, as the same variance is assumed for both experimental conditions in limma, the moderated effect size $d_{gi}^{(moderated)}$ for a given gene in a given study can be estimated as in (17) where t_{gi} is the limma moderated t-statistics. SMVar assumes different variances for treatment and control groups thus the moderated effect size for a given gene in a given study can be estimated as in(17) where t_{gi} is Welch t-statistics [129] and $\widetilde{n_i} = n_{it} + n_{ic}$. Moreover, the degrees of freedom gained using shrinkage approaches allowed Marot and colleagues to calculate the exact form of the variance for moderated effect sizes instead of the asymptotic estimator used by Choi and colleagues (see Equation (8)). Using the distribution of effect sizes provided by Hedges [130], it can be shown that:

$$(21) \quad \boldsymbol{var(d_{gi}) = s_{gi}^2 = \frac{m_i}{(m_i-2)\widetilde{n_i}}\left(1 + \widetilde{n_i} d_{gi}^2\right) - \frac{d_{gi}^2}{(c(m_i))^2}}$$

with

(22) $$\boldsymbol{c(m_i) = \Gamma\left(\frac{m_i}{2}\right)/\left(\sqrt{m_i/2} * \Gamma\left(\frac{m_i-1}{2}\right)\right)}$$

where $\widetilde{n_\iota} = (n_{it} * n_{ic})/(n_{it} + n_{ic})$ in limma and $\widetilde{n_\iota} = n_{it} + n_{ic}$ in SMVar and m is the number of degrees of freedom. In limma, m equals to the sum of prior degrees of freedom and residual degrees of freedom for the linear model of gene g. In SMVar, degrees of freedom are calculated by Satterthwaite's approach [131]. Then the unbiased estimators can be obtained from the moderated effect sizes as:

(23) $$\boldsymbol{d_{gi}^{\prime(moderated)} = c(m_i) d_{gi}^{(moderated)}}$$

This equation can be seen as an extension of Equation (7) with $c(m_i) = 1 - 3/(4m_i - 1)$ and $m_i = n_i - 2$. Assuming that $var\big(c(m_i)\big) = 0$, 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 $c(m)^2 * s_{gi}^2$. Since c(m)<1, unbiased estimators have a smaller variance than biased ones. The Marot and colleagues' approach has been implemented in the R package metaMA which offers three variants of effect sizes (classical and moderated t-test) and uses explicitly the random effect model. Only the Benjamini and Hochberg [107] multiple testing correction is available.

Recently, Bayesian meta-analysis models have also been developed. Choi and colleagues [16] introduced the first Bayesian meta-analysis model for microarray data which integrated standardized gene effects in individual studies into an overall mean effect. Inter-study variability was included as a parameter in the model with an associated uninformative inverse gamma prior distribution. Markov Chain Monte Carlo simulation was used to estimate the underlying effect size. Conlon and colleagues

[132] introduced two Bayesian meta-analysis models for microarray data: the standardized expression integration model and the probability integration model. The first model is similar in approach to that described in Choi and colleagues' study [16], except that standardized gene expression values (i.e. log-expression ratios standardized so that each array within a study has zero mean and unit standard deviation) were combined instead of effect sizes since the analyzed data are assumed to be from the same platform and comparable across studies. Conversely, the second model combines the probabilities of differential expression calculated for each gene in each study. Both models produce the gene-specific posterior probability of differential expression, which is the basis for inference. Since the standardized expression integration model includes inter-study variability, it may improve accuracy of results versus the probability integration model. However, due to the typical small number of studies included in microarray meta-analyses, the variability between studies is difficult to estimate. The probability integration model eliminates the need to specify inter-study variability since each study is modeled separately, and thus its implementation is more straightforward. Conlon and colleagues found that their probability integration model identified more true DEGs and fewer true omitted genes (i.e. genes declared as differentially expressed in individual studies but not in meta-analysis) than combining expression values.

Another meta-analysis method based on the modeling of the effect size within a Bayesian framework is that described by Wang and colleagues [25] and termed posterior mean differential expression. The main idea of their method is that one can use data from one study to construct a prior distribution of differential expression for each gene, whose distribution is then updated using other microarray studies thus

providing the posterior mean differential expression. The z statistics obtained weighting the posterior mean differential expression by individual studies' variances, has a standard normal distribution due to classic Bayesian probability calculation and may be used to test the differential expression. Alternatively random permutations can be used to estimate the distribution of the z scores under the null hypothesis and to determine the significance of the observed statistics.

3.4 COMBINING RANKS

Methods combining robust rank statistics are used to contain the problem of outliers which affect the results obtained using methods combining p-values or effect sizes. This can be a significant problem when thousands of genes are analyzed simultaneously in the noisy nature of microarray experiments. Instead of p-values or effect sizes, the ranks of differentially expressed evidence are calculated for each gene in each study. The product [42], mean [133] or sum [134] of ranks from all studies is then calculated as the test statistics. Permutation analysis can be performed to assess the statistical significance and to control FDR.

Zintzaras and Ioannidis [133] proposed METa-analysis of RAnked DISCovery datasets (METRADISC), which is based on the average of the standardized rank. METRADISC is the only rank-based method that incorporates and estimates the between-study heterogeneity. In addition the method can deal with genes which are measured in only some of the studies. The tested genes in each study are ranked based on the direction in expression change and the level of statistical significance or some other metrics. If there are G genes being tested, the highest rank G is given to the gene that shows the lowest p-value and is over-expressed in

treatment group (t) *vs* control group (c). The lowest rank 1 is given to the gene that shows the lowest p-value and is down-regulated in treatment group *vs* control group. Genes with equal p-values are assigned tied ranks. The average rank R* and the heterogeneity metric Q* for each gene g across studies are defined as:

$$(24) \qquad \boldsymbol{R}_{\boldsymbol{g}}^{*} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \boldsymbol{R}_{\boldsymbol{gi}}/k$$

and

$$(25) \qquad \boldsymbol{Q}_{\boldsymbol{g}}^{*} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \left(\boldsymbol{R}_{\boldsymbol{gi}} - \boldsymbol{R}^{*}\right)^{\boldsymbol{2}}$$

where R_{gi} is the rank of the gene g for study i (i=1 to k studies). The statistical significance for R* and Q* for each gene is assessed against the distributions of the average ranks and heterogeneity metrics under the null hypothesis that ranks are randomly assigned. Null distributions are calculated using non-parametric Monte Carlo permutation method. In this method, in a run, the ranks of each study are randomly permutated and the simulated metrics are calculated. The procedure is repeated a number of times, depending on the required accuracy of the final p-values.

Four statistical significance values are provided for each gene: statistical significance for high average rank, for low average rank, for high heterogeneity and for low heterogeneity. The statistical significance for high average rank is defined as the percentage of simulated metrics that exceed or are equal to the observed R*. The statistical significance for low average rank is the percentage of simulated metrics that are below or equal to the observed R*. Significance of heterogeneity is defined analogously. Interesting genes are those with significant average rank (either low or high) and low heterogeneity which indicates that the results

are consistent among different studies. The desired threshold of statistical significance for the R* and Q* testing should be selected on a case-by-case basis, depending on the desired trade-off between false negatives and false discovery rate. As a default, Zintzaras and Ioannidis recommend a level of 0.05/G, where G is the total number of genes shared by all the datasets, for average rank testing, and a less stringent p-value for heterogeneity-testing.

The original version of METRADISC performs an unweighted analysis giving equal weight to all studies. Alternatively, one may weight each study by its total sample size or other weight functions depending on the type of data to be combined. For two-class comparisons a very common weight function is given by:

(26) $$\boldsymbol{w_i} = (\boldsymbol{n_{it}} * \boldsymbol{n_{ic}})/(\boldsymbol{n_{it}} + \boldsymbol{n_{ic}})$$

where n_{it} and n_{ic} are the number of samples in groups t and c in study i, respectively. Then the weighted average rank for each gene across studies is defined as:

(27) $$\boldsymbol{R^*_{wg}} = (\sum_{i=1}^{k} \boldsymbol{w_i R_{gi}})/\sum_{i=1}^{k} \boldsymbol{w_i}$$

Heterogeneity testing should instead be performed with unweighted analyses, so as small studies are allowed to show their differences against larger ones [135].

Hong and colleagues [42] proposed a modification and extension of the rank product method, which was initially introduced by Breitling and colleagues [136] to detect DEGs between two experimental conditions in a single study. The Fold-Change (FC) is chosen as a selection method to compare and rank the genes within each dataset. These ranks are then combined to produce an overall score for the genes across datasets,

obtaining a ranked gene list. The method focuses on genes which are consistently highly ranked in a number of datasets, for example genes that are regularly found among top up-regulated genes in many microarray studies. In detail, the algorithm of the method consists of five steps.

1. For each gene g (g=1 to G genes), pair-wise ratios or FCs are calculated within each dataset i (i=1 to k studies). Let n_{ti} and n_{ci} be the number of samples in group 1 and 2 in study i, then the total number of pair-wise comparisons is equal to $L_i=n_{it}*n_{ic}$.
2. Ranks are assigned (1 for the highest value) according to fold-change ratio. R_{gil} is the rank of gene g in i*th* study under l*th* comparison, l=1…L_i.
3. RankProduct for each gene g is calculated as:

(28) $$\boldsymbol{RP_g} = \left(\prod_i \prod_l \boldsymbol{R_{gil}}\right)^{\frac{1}{L}}$$

where L is the sum of products of number of samples in groups:

(29) $$\boldsymbol{L} = \sum_{i=1}^{k} (\boldsymbol{n_{it}} * \boldsymbol{n_{ic}})$$

The smaller the RP value the smaller the probability that the observation of the gene at the top of the lists is due to chance. It is equivalent to calculating the geometric mean rank.

4. b permutations of gene expression values within each array are performed and all previous steps repeated in order to obtain the null rank product statistic $RP_g^{*(b)}$.
5. Step 4 is repeated B times to estimate the distribution of $RP_g^{*(b)}$. This distribution is used to calculate p-value and FDR for each gene.

(30) $$\boldsymbol{p_g} = (\mathbf{1}/\boldsymbol{GB}) \sum_{\boldsymbol{b}} \sum_{(\boldsymbol{g})} \boldsymbol{I}\left(|\boldsymbol{RP_g^{*(b)}}| \leq \boldsymbol{RP_{m=g}}\right)$$

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

Converting FCs into ranks increased the results robustness against noise and across-studies heterogeneity as demonstrated by Yuen and colleagues [66]. In fact, they showed that although the FCs of DEGs had poor consistency across platforms, the rank orders were comparable.

In a recent study, Hong and Breitling [109] comparatively evaluated rank product method, Fisher's method and the t-based hierarchical modeling, showing that the rank product outperformed the other methods in terms of sensitivity and specificity, especially in the setting of small sample size and/or large between-study variation.

The rank product method is implemented in the Bioconductor package RankProd [42].

DeConde and colleagues [137] proposed three aggregation approaches based on meta-search methods from computer science, which are used to combine ranked results from multiple internet search engines [138]. Because they rely on rank-ordered gene lists, they share many of the advantageous characteristics of rank products. In particular, two of the algorithms use Markov chains to convert the pair-wise preferences between the gene lists into a stationary distribution, representing an aggregate ranking, while the third algorithm is based on an order-statistics model.

3.5 AVAILABLE SOFTWARE

Despite the availability of many microarray meta-analysis methods, there exist very few software packages for microarray meta-analysis

implementation and most of them either did not have clear manuals or had functions that were not easy to apply.

Compared with popular microarray data analysis packages (e.g. SAM, limma or BRB array tool), existing meta-analysis packages are relatively primitive and difficult to use. In the R and Bioconductor environment, metaGEM (which implements vote counting, Fisher's method and random effects model) [48], GeneMeta (which implements fixed and random effects model), metaMA (which implements random effects model and Stouffer's method), metaArray (which implements three meta-analysis approaches: (1) probability of expression (POE) [90, 139], (2) integrative correlation [140] and (3) posterior mean differential expression [25]) [90], OrderedList (which compares ordered gene lists) [141], RankProd (that implements rank product method) [42] and RankAggreg (that implements various rank aggregation methods) [142] are available. The R package MAMA [Ihnatova I, 2012; MAMA: Meta-Analysis of MicroArray, R package version 2.1.0] was the first tool that implemented many different meta-analysis methods. It uses a common framework to manage and combine the individual datasets. It additionally offers some functionalities to combine and visualize outputs from different methods, allowing a complex view on change in gene expression.

Recently MetaOmics [143], a suite of three R packages MetaQC, MetaDE and MetaPath, for quality control, DEGs identification and enriched pathway detection for microarray meta-analysis, respectively, has been developed. The MetaQC package [144] provides a quantitative and objective tool to determine suitable study inclusion/exclusion criteria for meta-analysis. MetaDE contains many state-of the art meta-analysis methods to detect DEGs (Fisher, Stouffer, adaptively weighted Fisher (AW), minimum p-value, maximum p-value, r*th* ordered p-value (rOP)

[145], fixed and random effects model, rank product, naïve sum of ranks and naïve product of ranks [134]). Finally the MetaPath package [146] provides a unified meta-analysis framework and inference to detect enriched pathways associated with outcome. At present MetaOmics is the most complete software for microarray data meta-analysis and the only tool that provides a systematic pipeline to assist the user in conducting the meta-analysis. Information concerning the R and Bioconductor packages described above is summarized in Table 3.1.

Package	Environment	Reference	Link
metaGEM	R	[Ramasamy A et al., 2008]	http://hdl.handle.net/10044/1/4217
GeneMeta	Bioconductor	[Lusa L et al.,]	http://www.bioconductor.org/packages/release/bioc/html/GeneMeta.html
metaMA	R	[Marot G et al., 2009]	http://cran.r-project.org/web/packages/metaMA/
metaArray	Bioconductor	[Ghosh D and Choi H]	http://www.bioconductor.org/packages/2.11/bioc/html/metaArray.html
RankProd	Bioconductor	[Hong F et al., 2006]	http://www.bioconductor.org/packages/2.11/bioc/html/RankProd.html
OrderedList	Bioconductor	[Lottaz C et al., 2009]	http://www.bioconductor.org/packages/2.11/bioc/html/OrderedList.html
RankAggreg	R	[Pihur H and Datta S, 2009]	http://cran.r-project.org/web/packages/RankAggreg/index.html
MAMA	R	[Ihnatova I, 2010]	http://cran.r-project.org/web/packages/MAMA/index.html
MetaOmics (MetaQC, MetaDE, MetaPath)	R	[Wang X et al., 2012]	http://cran.r-project.org/web/packages/MetaQC/ http://cran.r-project.org/web/packages/MetaDE/ http://cran.r-project.org/web/packages/MetaPath/

Table 3.1: Available R and Bioconductor packages for microarray data meta-analysis

4 APPLICATION TO REAL DATASETS

We applied some of the meta-analysis methods described in the previous chapter to a set of three microarray experiments from four malignant pleural mesothelioma (MPM) studies. We focused on differential expression analysis between normal and malignant mesothelioma pleural tissues.

4.1 MATERIALS AND METHODS

4.1.1 DATA COLLECTION

To systematically collect MPM microarray studies PubMed (http://www.ncbi.nlm.nih.gov/pubmed) was searched with keywords related to the study background. The automatic search covered up to the end of December 2011. The search was extended to the two largest public microarray data repositories: GEO and ArrayExpress. MPM microarray studies suitable for meta-analysis were selected according to the following inclusion/exclusion criteria:

- any human studies investigating at least four patients with MPM and at least four patients with corresponding normal pleural samples using high-density arrays were included,
- any studies using mesothelioma derived cell lines or studies using specialized arrays were excluded,
- any studies where patients with MPM have been exposed to drugs were excluded.

In the following, datasets will be referred to by the name of the first author of the related papers.

4.1.2 DATA PREPROCESSING

The quality assessment of the raw data was performed using affyPLM [74] and arrayQualityMetrics [75] Bioconductor packages with the aim of identifying and possibly removing poor quality arrays and to detect possible systematic effects.

All datasets were preprocessed independently for background correction, normalization and summarization.

According to Ramasamy [48], who stated that the same preprocessing algorithm should be used for multiple studies conducted on the same platform, raw data from Crispi and Røe were preprocessed using the implementation of the Affymetrix MAS 5.0 algorithm provided by the Bioconductor package simpleaffy [73], setting the scale parameter to 100. Gene expression levels for each microarray in the Gordon study were generated and scaled to a target intensity of 100 using Affymetrix Microarray Suite v.5.0 (Santa Clara, USA). The $\log_2$ of the expression values was taken.

Alternatively, in order to assess the overall effect of the preprocessing methods on meta-analysis results, for all the studies the

data preprocessed using the algorithm described by the original authors in their papers were also considered. In particular, probe set intensities were obtained by means of gcrma [147] followed by quantile normalization for Crispi dataset and by means of RMA (Robust Multi-array Average) [148] for Røe dataset. In both cases, the gene expression values of the technical replicates in the Røe dataset have been averaged after the preprocessing step. The assessment of data quality was repeated after the preprocessing step to check the normalization efficiency.

Box plots and density plots for each sample in each dataset were used to evaluate the effect of the preprocessing steps on the data.

4.1.3 DATA ANNOTATION

To combine microarray data across studies a unique gene-level identifier must be adopted. We relied on official Gene Symbols by the HUGO Gene Nomenclature Committee [149] to achieve a uniform annotation across array versions. Affymetrix probe set IDs were mapped to the corresponding Gene Symbol IDs using platform-specific Bioconductor annotation packages such as hgu133a.db [Carlson M, hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a). R package version 2.8.0], hgu133plus2.db [Carlson M, hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 2.8.0] and AnnotationDbi [Pages H, Carlson M et al., AnnotationDbi: Annotation Database Interface. R package version 1.20.3]. Probe set IDs with no Gene Symbol and probe set IDs mapping to more than one Gene Symbol were discarded from each dataset. *Vice versa*, when multiple probes sets mapped to an identical Gene Symbol, the probe set ID with the largest Inter Quartile Range (IQR) of expression values (among all multiple probe

IDs) was selected to represent the gene. Larger IQR represents greater variability (and thus greater information content) in the data and this probe matching method has been recommended in Bioconductor [74]. Only the genes in common between the three datasets were retained for the analysis.

The gene expression values for the genes in common to all studies from the three datasets were bound together to form two final meta-datasets, denoted as A and B, containing MAS 5.0 preprocessed data and data processed according to the original authors' methods, respectively.

Box plots and PCA plots of meta-datasets A and B were used to inspect for any clustering of arrays by studies. affy [150] and EMA (Easy Microarray data Analysis) [151] Bioconductor packages were used to obtain these plots.

4.1.4 GENE FILTERING

As stated in Chapter 2, it is unclear if filtering is beneficial from a meta-analysis perspective. Here we tried to evaluate the effect of a common filtering strategy on meta-analysis results. Two sequential steps of gene filtering were applied to meta-datasets A and B. In the first step, genes with very low expression showing small average expression values across studies were filtered out. Specifically, mean intensities of each gene across all samples in each study were calculated and the corresponding ranks were obtained. The sum of such ranks across the three studies of each gene was calculated and genes with the lowest 30% rank sum were considered unexpressed genes and were filtered out. Similarly, in the second step, genes with small variation between the experimental conditions were filtered out by replacing mean intensity in the first step with standard deviation. Genes with the lowest 30% rank sum of standard

deviations were filtered out. Finally only the genes which passed the two filtering steps were retained for further analysis. We denoted these additional filtered meta-datasets as A1 and B1.

Box plots and PCA plots of datasets A1 and B1 were used to inspect for any clustering of arrays by studies.

4.1.5 DATA META-ANALYSIS

A combination of meta-analysis methods was applied to both unfiltered (A and B) and filtered (A1 and B1) meta-datasets. First of all, a standard limma analysis was performed including a study effect in the linear model. The Benjamini and Hochberg (BH) multiple correction method with a threshold of 5% was applied to identify the DEGs. In addition three meta-analysis approaches were applied to the meta-datasets:

1. the weighted inverse normal p-value combination method proposed by Marot and colleagues [40] and implemented in the R package metaMA. Both the standard t-statistics and the moderated t-statistics were used to calculate the p-values for each gene in the individual studies analyses. We will refer to this method as p-value combination method.
2. The effect size combination approach both in the Choi and colleagues' version [16], implemented in the Bioconductor package GeneMeta, and in the Marot and colleagues' version [40], available in the R package metaMA. We will refer to these methods as the GeneMeta and the metaMA method, respectively.
3. The rank combination method by Hong and colleagues [42], implemented in the Bioconductor package RankProd. We will refer to this methods as RankProd method.

The DEGs were selected at significance level of 0.05 in all methods.

In order to the selected DEGs lists from different methods be more comparable in size, we considered the lists of the top 500 genes with the smallest p-values/q-values for each method.

All analysis was carried out in R version 2.15.2 and Bioconductor release 2.11.

To further assess the DEGs lists produced by the different methods, pathway analysis was done using PathwayExpress [152]. Hypergeometric distribution and FDR were used for the p-value calculation and p-value correction, respectively. The set of genes shared by all the three datasets was selected as reference array for the over-representation analysis. The magnitude of the measured expression changes was not used because this information was not available for all the tested methods.

4.2 RESULTS

Only four microarray studies [153-156] met the inclusion/exclusion criteria. This may depend on the difficulty of collecting a large number of tissue samples, due to the low incidence of MPM. However the two papers by Røe and colleagues referred to the same dataset so there were only three distinct datasets suitable for the meta-analysis.

By querying ArrayExpress and GEO, the raw data (in the form of CEL files) were retrieved for Crispi and Røe studies but not for Gordon study for which only gene expression data have been deposited in GEO database. The authors were asked for the raw data but they did not make them available.

The studies were performed on different versions of the Affymetrix oligonucleotide microarray platform. The Crispi and Røe studies used

Affymetrix Human Genome U133 Plus2.0 GeneChip (Santa Clara, CA, USA), containing 54675 probe sets, each with eleven probe pairs. The Crispi dataset included thirteen samples, four normal pleural samples and nine MPM samples. The Røe dataset consisted of one papillary serous adenocarcinoma pleura sample, one pleural plaque sample with unknown disease condition, seven MPM samples, where two were from the same patient, seven parietal pleural samples, where two were from the same patient, and four visceral pleural samples that were from the same control patients. Finally, the Gordon study was performed on Affymetrix Human Genome U133A GeneChip, containing 22283 probe sets, each with eleven probe pairs. This dataset included forty human MPM tumor specimens, five normal pleura specimens, four normal lung specimens and five mesothelial cell lines, where four are MPM derived and one (Met5a) is a nontumorigenic immortalized mesothelial cell line. The papillary serous adenocarcinoma sample and the pleural plaque sample with unknown disease condition in the Røe dataset were discarded. Moreover the four visceral pleural samples were excluded from the meta-analysis according to the authors. They observed that mesothelioma, parietal and visceral pleural tissues show distinct expression profiles and parietal pleura can be considered as the main reference because mesothelioma usually develops in the parietal pleura, subsequently invading the visceral layer [155]. Similarly, only the forty human MPM tumor specimens and the five normal pleural specimens in the Gordon dataset were retained for the meta-analysis.

The main features of the three datasets included in the meta-analysis after excluding unsuitable samples are summarized in Table 4.1.

Authors	Repository (Accession Number)	Normal	Tumor	Platform	Total Probes	Probes with Gene Symbols
Crispi et al. (2009)	GEO (GSE12345)	4	9	HGU133Plus2	54675	41910
Røe et al. (2010)	ArrayExpress (E-MTAB-47)	7	7	HGU133Plus2	54675	41910
Gordon et al. (2005)	GEO (GSE2549)	5	40	HGU133A	22283	20365

Table 4.1: Microarray datasets and samples included in the meta-analysis

Figure 4.1 and Figure 4.2 show the box plots of MAS 5.0 preprocessed, $\log_2$ transformed gene expression data separately for the three datasets, while box plots of preprocessed data obtained applying different algorithms for each dataset are shown in Figure 4.2 and Figure 4.3.

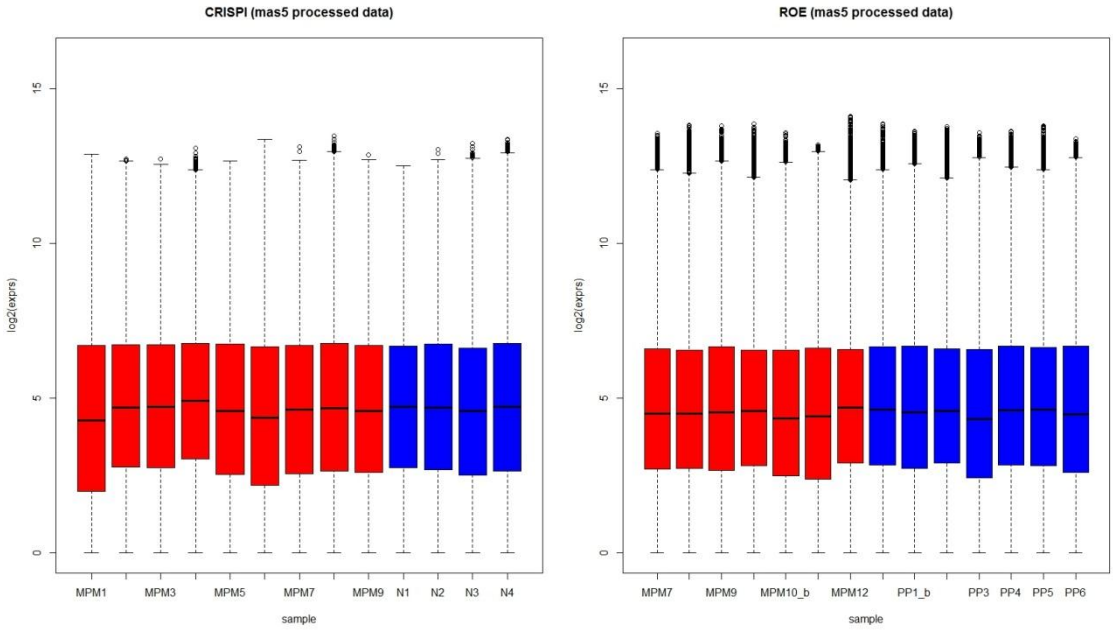

Figure 4.1: Box plots of MAS 5.0 preprocessed data from Crispi (left) and Røe (right) datasets (red=MPM samples, blue=control samples)

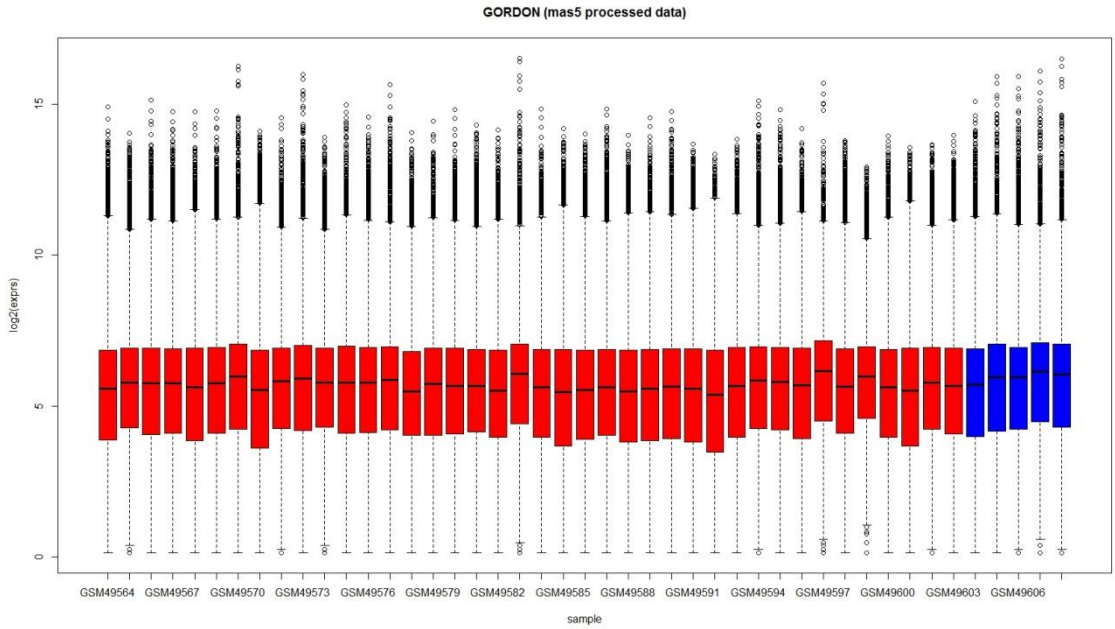

Figure 4.2: Box plots of MAS 5.0 preprocessed data from Gordon dataset (red=MPM samples, blue=control samples)

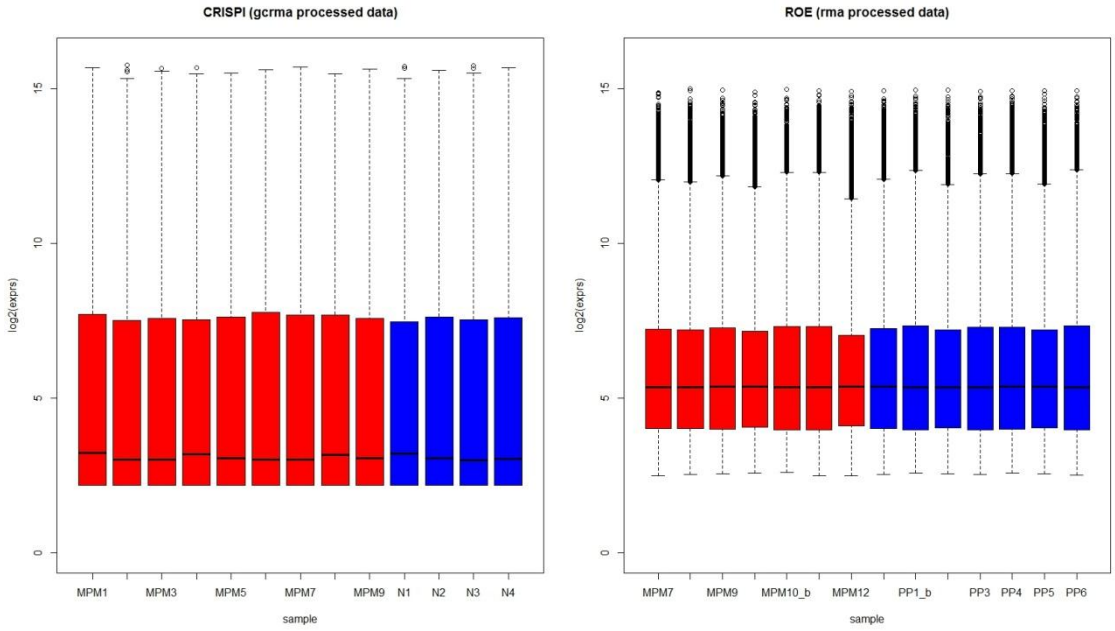

Figure 4.3: Box plots of data preprocessed using the algorithms described by individual authors, gcrma for Crispi dataset (left) and RMA for Røe dataset (right) (red=MPM samples, blue=control samples)

In both cases, the preprocessed data from the three studies were matched using Gene Symbol IDs. By excluding the control probe sets, the 22215 probe sets in the U133A platform mapped to 20365 Gene Symbols, while the 54613 probe sets in the U133 Plus 2.0 platform mapped to

41910 Gene Symbols. After solving the many-to-many relationships, 12701 and 20184 unique Gene Symbols were found in the U133A and U133 Plus 2.0 platform, respectively. Only the 12701 genes in common to all three datasets were kept for the meta-analysis. Consequently the meta-datasets A and B consist of 12701 genes and 70 samples.

After the filtering steps, 6222 matched genes in three studies were analyzed. The meta-datasets A1 and B1 consist of 6222 genes and 70 samples. We note that the somewhat ad hoc gene filtering procedure is necessary and is commonly used in microarray analysis. This procedure can reduce false positives from non-expressed or non-informative genes and increase statistical power in multiple test correction; however important DEGs can be discarded. Figure 4.4 shows the preprocessing diagram and the number of genes selected by each preprocessing step.

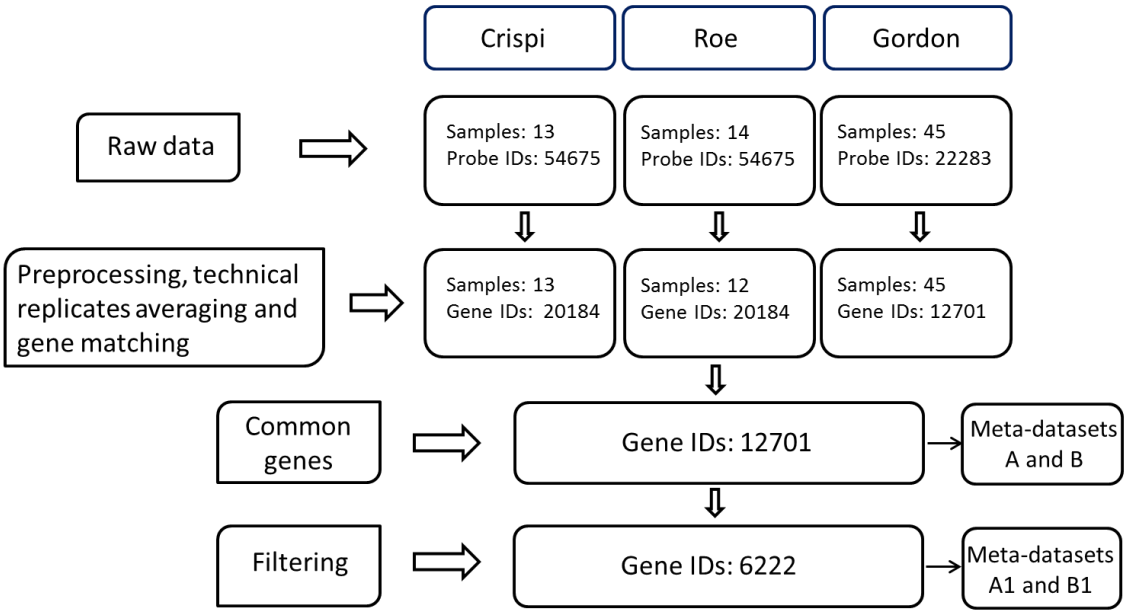

Figure 4.4: A diagram for data preprocessing, gene matching and gene filtering

PCA plots for the meta-datasets A and B are reported in Figure 4.5. In both cases, PCA plots highlight a clustering of arrays by studies.

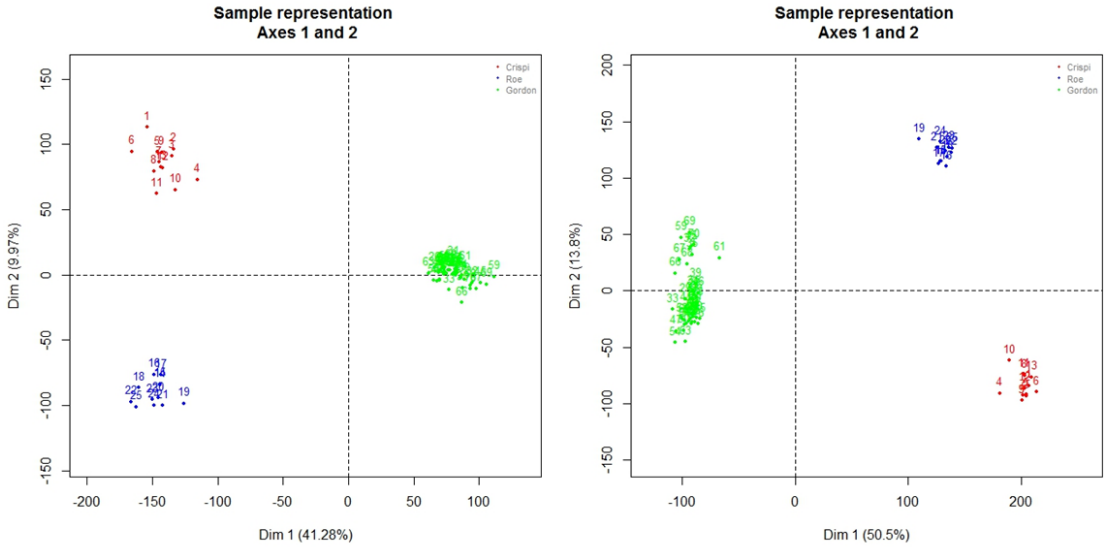

Figure 4.5: PCA plots of MAS 5.0 preprocessed data (left) and data preprocessed using the algorithms described by individual authors (right) (common Gene Symbols only) (red=Crispi dataset, blue= Røe e dataset, green=Gordon dataset)

The same grouping of arrays by studies is evident from the box plots (see Figure 4.6 and Figure 4.7). However the box plots highlight better how the differences between the studies are relevant when the data preprocessed using different algorithms for the three datasets are considered compared to the MAS 5.0 processed data. Figure 4.6 also suggests a possible 'platform effect' as the Crispi and Røe datasets, performed on the same Affymetrix platform, appear much more similar to each other than to the data from the Gordon study performed on a different Affymetrix platform.

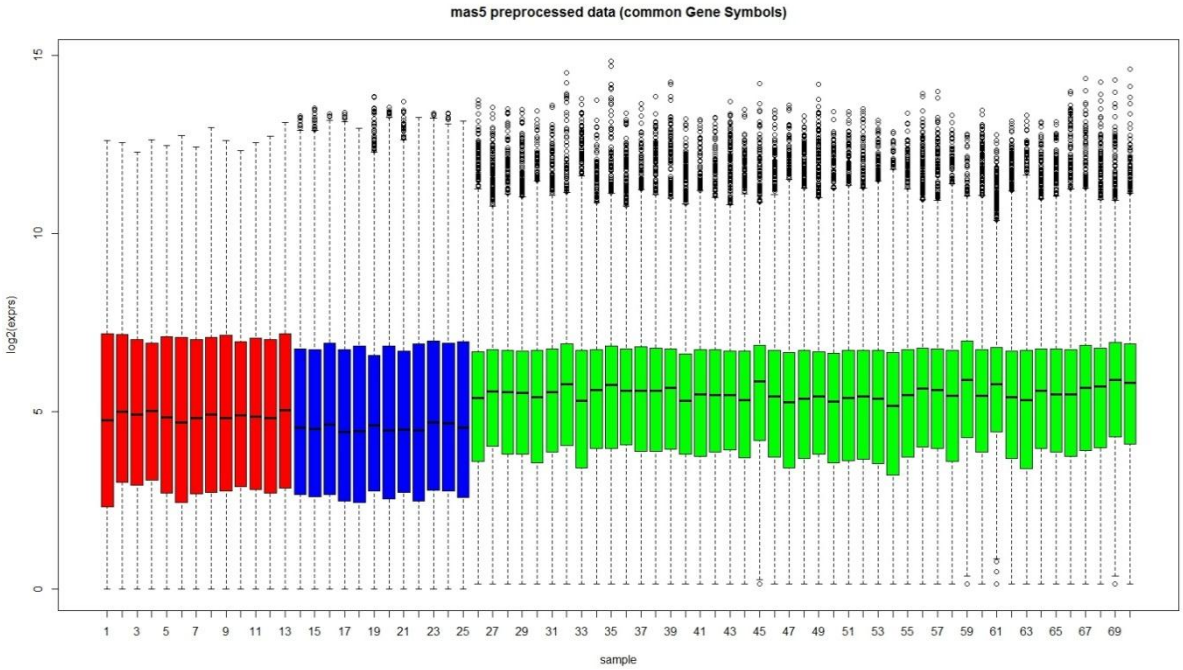

Figure 4.6: Box plots of MAS 5.0 preprocessed data (common Gene Symbols only) (red=Crispi dataset, blue= Røe dataset, green=Gordon dataset)

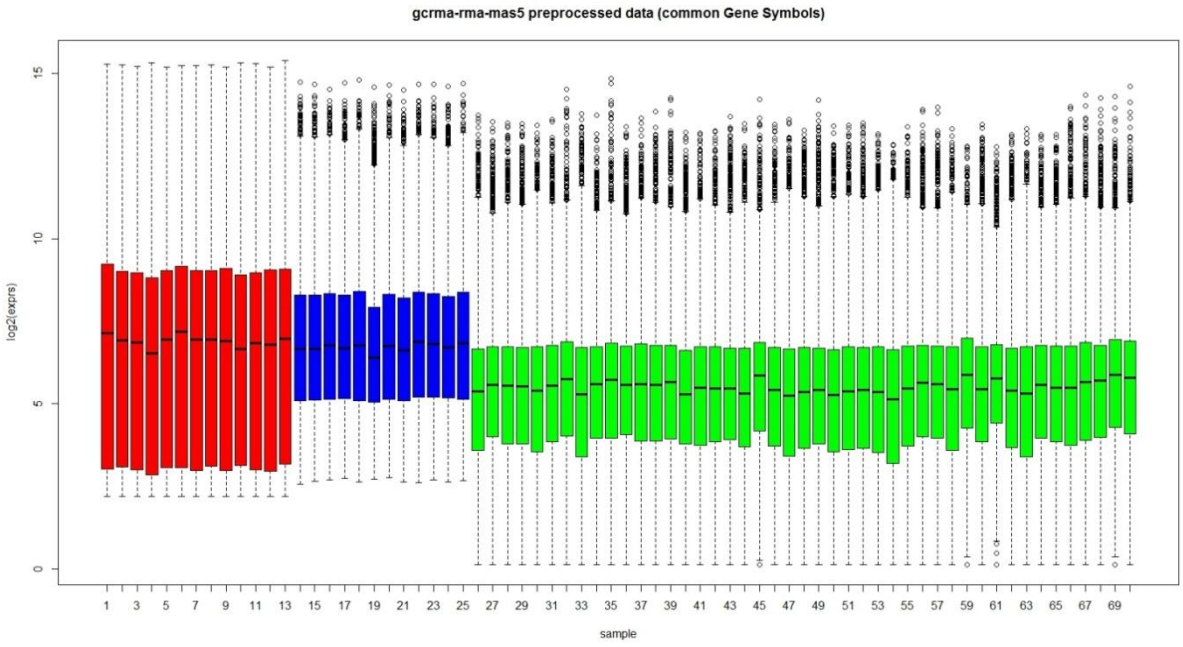

Figure 4.7: Box plots of data preprocessed using the algorithms described by individual authors (common Gene Symbols only) (red=Crispi dataset, blue= Røe dataset, green=Gordon dataset)

The same considerations apply to meta-datasets A1 and B1, whose PCA plots and box plots are shown in Figure 4.8, Figure 4.9 and Figure 4.10.

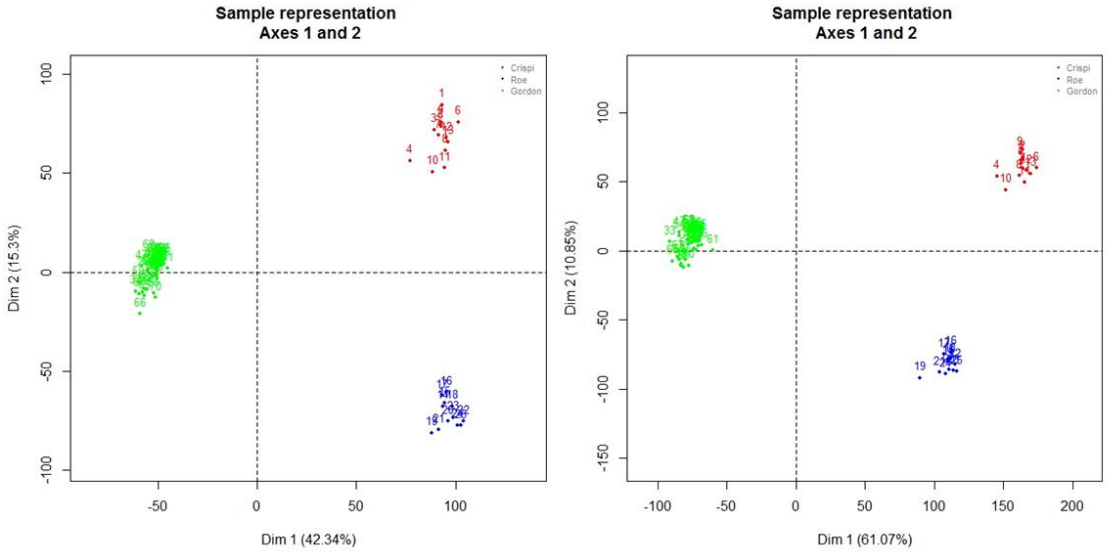

Figure 4.8: PCA plots of MAS 5.0 preprocessed data (left) and data preprocessed using the algorithms described by individual authors (right) after filtering (common Gene Symbols only) (red=Crispi dataset, blue= Røe dataset, green=Gordon dataset)

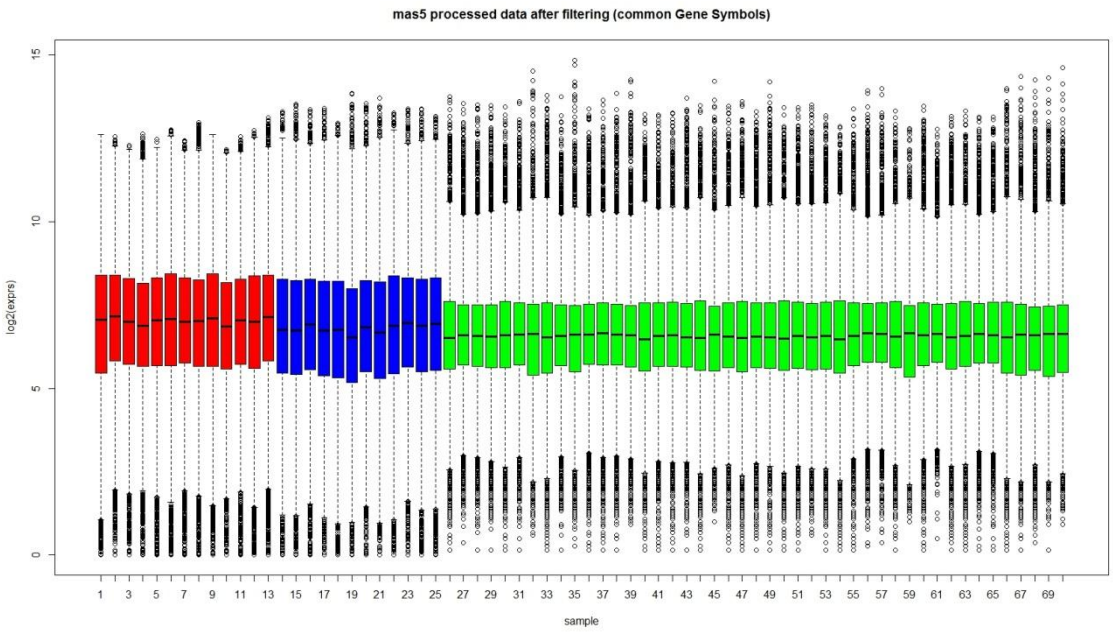

Figure 4.9: Box plots of MAS 5.0 preprocessed data after filtering (common Gene Symbols only) (red=Crispi dataset, blue= Røe dataset, green=Gordon dataset)

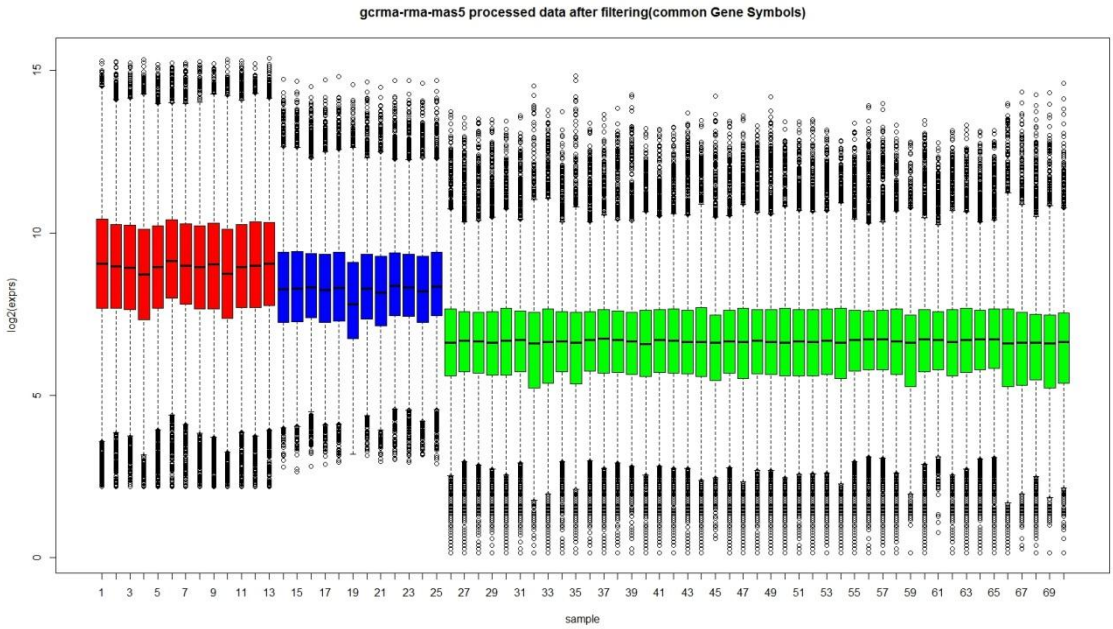

Figure 4.10: Box plots of data preprocessed using the algorithms described by individual authors after filtering (common Gene Symbols only) (red=Crispi dataset, blue= Røe dataset, green=Gordon dataset)

Concerning meta-dataset A, at a 5% BH threshold, using a limma model including a study effect, 1883 genes were found statistically significant. The p-value combination claimed 2420 and 2461 genes as differentially expressed at the same BH threshold using the moderated t-statistics and the standard t-statistics, respectively. For the metaMA method, 698 and 610 genes were found as differentially expressed at a 5% BH threshold using the moderated t-statistics and the standard t-statistics, respectively. The GeneMeta method identified 717 DEGs at a 5% BH threshold using the REM. The Cochran's Q statistics has been used to test the between-study variability and decide between the REM and the FEM. Using the SAM type analysis to estimate the FDR, 1154 DEGs were identified with a FDR<0.05 for the two-sided hypothesis. These 1154 DEGs included the 717 DEGs identified using the BH multiple correction procedure. Finally, the RankProd method detected 1110 DEGs. The number of DEGs identified by each method for meta-datasets A and B is

summarized in Table 4.2, while the meta-analysis results for the filtered meta-datasets A1 and B1 are illustrated in Table 4.3.

	limma with study effect	***P-value combination***		***metaMA approach***		***GeneMeta approach***		***RankProd***
		stand. T	moder. T	stand. t	moder. T	REM	REM_FDR	
dataset A	1883	2461	2420	610	698	717	1154	1110
dataset B	2066	2627	2691	871	947	994	1437	1309
intersection	1446	2024	2002	470	511	546	839	1031
intersection/ (A only+B only - intersection) %	57.77	66.06	64.39	46.49	45.06	46.87	47.89	74.28

Table 4.2: Number of DEGs provided by the different meta-analysis approaches at 5% BH threshold for both datasets A and B and their intersections

	limma with study effect	***P-value combination***		***metaMA approach***		***GeneMeta approach***		***RankProd***
		stand. T	moder. t	stand. t	moder. T	REM	REM_FDR	
dataset A1	1485	1684	1699	559	626	627	912	668
dataset B1	1548	1787	1805	667	742	717	1018	735
intersection	1095	1291	1293	382	428	423	619	511
intersection/ (A1 only+B1 only - intersection) %	56.5	59.22	58.48	45.26	45.53	45.93	47.22	57.29

Table 4.3: Number of DEGs provided by the different meta-analysis approaches at 5% BH threshold for both datasets A1 and B1 and their intersections

Venn diagrams corresponding to the comparisons of these methods applied to meta-dataset A are given in Figure 4.11 and Figure 4.12.

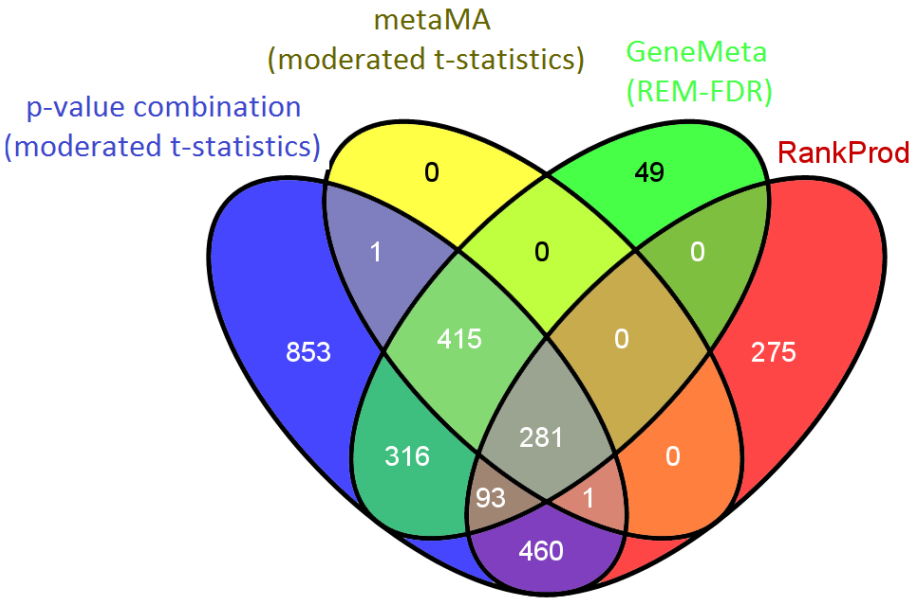

Figure 4.11: Venn diagram for meta-dataset A comparing the DEGs lists at a 5% BH threshold obtained by combining p-values, effect sizes (using both metaMA and GeneMeta approaches) and ranks

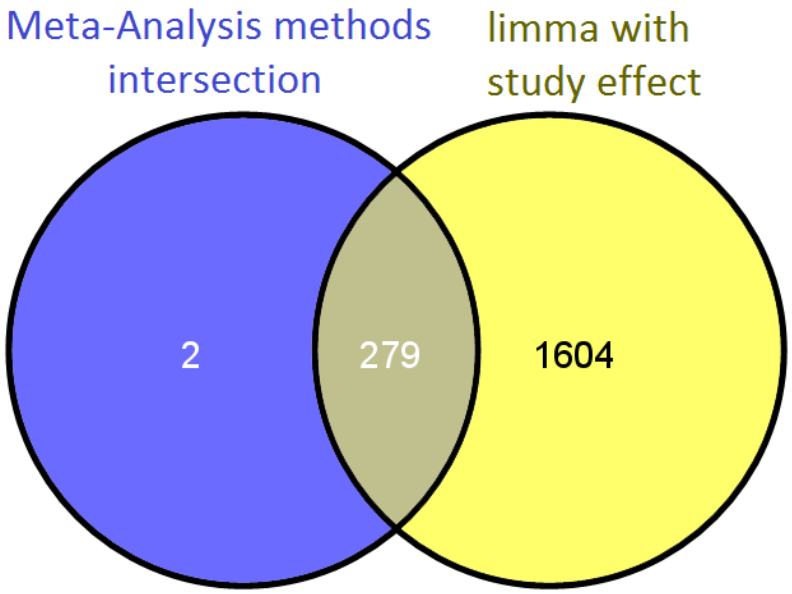

Figure 4.12: Venn diagram for meta-dataset A comparing the DEGs in common between the four meta-analysis methods to the DEGs list provided by limma including a study effect

It was found that 281 DEGs were in common between the four approaches (gray sector in Figure 4.11). This poor overlap is mainly due to the DEGs identified by the RankProd method which look quite different from the DEGs lists generated by the other meta-analysis approaches. In

fact, excluding the RankProd method, the number of DEGs in common between the other three approaches goes up to 696 (see Figure 4.13).

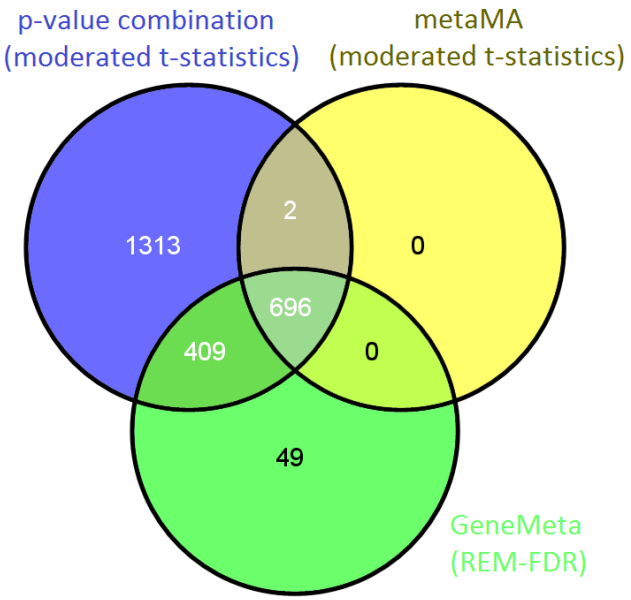

Figure 4.13: Venn diagram comparing the DEGs lists at a 5% BH threshold obtained by combining p-values and effect sizes (using both metaMA and GeneMeta approaches) for meta-dataset A.

It can also be noticed that the p-value combination method detected all the genes found with the metaMA effect size combination method, all but 49 genes found with the GeneMeta method followed by SAM type analysis to estimate the FDR and all but 275 genes identified by the RankProd method. On the other hand, 853 genes were detected only by the p-value combination method. For each method the intersections among the identified DEGs for meta-datasets A and B and A1 and B1 are summarized in Table 4.2 and Table 4.3, respectively.

Since many biological replicates were involved, we could not observe on these datasets the gain of DEGs usually found with shrinkage approaches. Indeed, the effect size combination based on classical t-tests detected 610 DEGs while the effect size combination based on moderated t-tests found 698 DEGs. We could check that, in this case, using the exact variance for standard effect sizes, as done by the metaMA method, the

number of DEGs did not change substantially compared with using the asymptotic variance, as done by GeneMeta. Indeed, the effect size combination based on usual t-tests and the exact variance detected 610 DEGs, while the z-score given by the GeneMeta package found 717 DEGs. Table 4.2 also points out that p-value combination method detected many more genes than the other approaches. The same considerations apply to the meta-datasets B, A1 and B1. Concerning the intersections between the DEGs lists detected by the diverse methods in the meta-datasets A and B, it is worth noting that the greater overlap (74.28 %) was reached by the RankProd method thus confirming the reduced dependence of this approach by the preprocessing methods adopted in single-study analysis and its greater robustness against heterogeneity across studies [66]. The intersections between the DEGs lists detected by the different methods in the meta-datasets A1 and B1 are instead comparable in size.

As far as gene rankings were concerned, they were very similar. The absolute values of test statistics used by the different meta-analysis methods were ranked in descending order and the Spearman rank correlation coefficients were calculated for each pair of methods. Detailed results for meta-datasets A and B are summarized in Table 4.4 and Table 4.5, respectively. As expected, the highest correlation (0.99) was found between GeneMeta and metaMA methods given the similarity between the used summary test statistics. Slightly lower values were obtained for meta-dataset B. Concerning meta-datasets A1 and B1, a slight increase of the correlation coefficients can be observed. Spearman rank correlation coefficients for dataset A1 are summarized in Table 4.8.

In order to make the DEGs lists from different methods more comparable in size, the top 500 DEGs for all the methods were selected.

Figure 4.14 and Figure 4.15 are Venn diagrams displaying the overlap of the top 500 DEGs found by different methods applied to meta-dataset A.

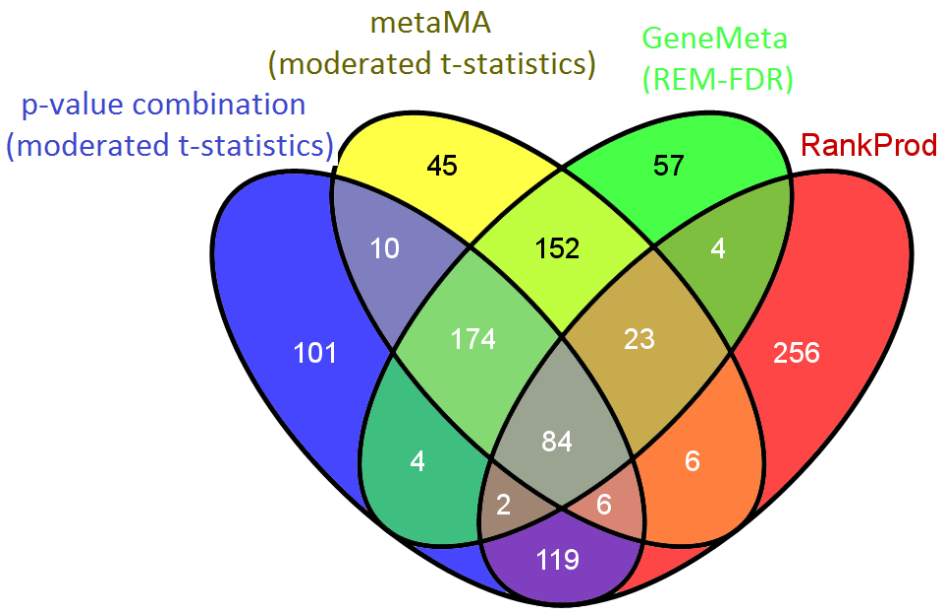

Figure 4.14: Venn diagram showing the overlap of the top 500 DEGs found by different meta-analysis methods applied to meta-dataset A

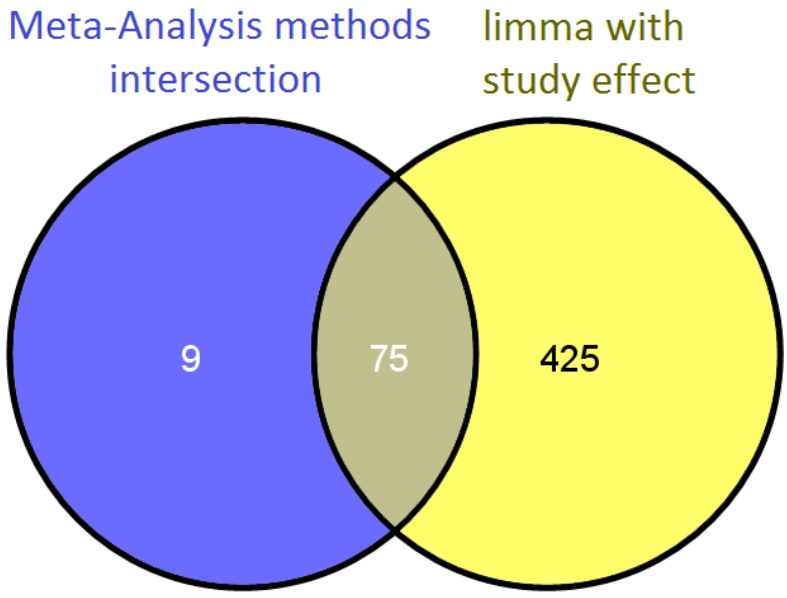

Figure 4.15: Venn diagram showing the overlap of the top 500 DEGs in common between the four meta-analysis methods and the top 500 DEGs found by limma including a study effect applied to meta-dataset A

Detailed information concerning the size of the overlap between the top 500 DEGs lists found using the different meta-analysis methods for

datasets A, B and A1 are presented in Table 4.6, Table 4.7 and Table 4.9, respectively. Also in this case, the use of the filtering process results in a greater overlap of the results provided by the different meta-analysis methods.

The fact that a significant number of genes were selected by only one of the methods shows that the gene ranking is based on different perspectives, thus it may be useful to select candidate genes using a combination of these methods.

		limma with study effect	*P-value combination*		*metaMA approach*		*GeneMeta approach*	*RankProd*
			standard t	moderated t	standard t	moderated t	REM_FDR	
limma with study effect		1						
P-value combination	standard t	0.883	1					
	moderated t	0.897	0.994	1				
metaMA approach	standard t	0.883	0.836		1			
	moderated t	0.898		0.831	0.993	1		
GeneMeta approach	REM_FDR	0.88	0.823	0.82	0.999	0.993	1	
RankProd		0.768	0.687	0.717	0.78	0.81	0.785	1

Table 4.4: Spearman rank correlations for meta-dataset A

		limma with study effect	*P-value combination*		*metaMA approach*		*GeneMeta approach*	*RankProd*
			standard t	moderated t	standard t	moderated t	REM_FDR	
limma with study effect		1						
P-value combination	standard t	0.904	1					
	moderated t	0.92	0.995	1				
metaMA approach	standard t	0.772	0.832		1			
	moderated t	0.783		0.833	0.99	1		
GeneMeta approach	REM_FDR	0.684	0.717	0.717	0.863	0.845	1	
RankProd		0.882	0.754	0.782	0.74	0.761	0.637	1

Table 4.5: Spearman rank correlations for meta-dataset B

		limma with study effect	*P-value combination*		*metaMA approach*		*GeneMeta approach*	*RankProd*
			standard t	moderated t	standard t	moderated t	REM_FDR	
limma with study effect		100						
P-value combination	standard t	76.2	100					
	moderated t	78.6		100				
metaMA approach	standard t	64.4	54.6		100			
	moderated t	65		54.8		100		
GeneMeta approach	REM_FDR	63.4	53.6	52.8	96.4	86.6	100	
RankProd		33.6	37.6	42.2	23.6	23.8	22.6	100

Table 4.6: Overlap between the top 500 DEGs lists identified by various methods for meta-dataset A, expressed as a percentage

		limma with study effect	*P-value combination*		*metaMA approach*		*GeneMeta approach*	*RankProd*
			standard t	moderated t	standard t	moderated t	REM_FDR	
limma with study effect		100						
P-value combination	standard t	75.8	100					
	moderated t	78		100				
metaMA approach	standard t	57.6	54		100			
	moderated t	56.6		56.8		100		
GeneMeta approach	REM_FDR	55.8	52.2	53	95.4	93	100	
RankProd		37.4	40.8	43	25.4	25.8	24.2	100

Table 4.7: Overlap between the top 500 DEGs lists identified by various methods for meta-dataset B, expressed as a percentage

		limma with study effect	*P-value combination*		*metaMA approach*		*GeneMeta approach*	*RankProd*
			standard t	moderated t	standard t	moderated t	REM_FDR	
limma with study effect		1						
P-value combination	standard t	0.922	1					
	moderated t	0.935	0.996	1				
metaMA approach	standard t	0.874	0.848		1			
	moderated t	0.883		0.846	0.996	1		
GeneMeta approach	REM_FDR	0.869	0.837	0.837	0.999	0.996	1	
RankProd		0.892	0.763	0.79	0.806	0.829	0.808	1

Table 4.8: Spearman rank correlations for meta-dataset A1

		limma with study effect	*P-value combination*		*metaMA approach*		*GeneMeta approach*	*RankProd*
			standard t	moderated t	standard t	moderated t	REM_FDR	
limma with study effect		100						
P-value combination	standard t	81.2	100					
	moderated t	84.2		100				
metaMA approach	standard t	64	56		100			
	moderated t	63.2		57.8		100		
GeneMeta approach	REM_FDR	63.2	54.8	56.4	97	94.2	100	
RankProd		47.8	49.4	54	36.4	37.4	35.6	100

Table 4.9: Overlap between the top 500 DEGs lists identified by various methods for meta-dataset A1, expressed as a percentage

4.2.1 COMPARISONS AMONG THE DEGs LISTS AT PATHWAY-LEVEL

To further assess the DEGs lists, a pathway analysis was performed. Table 4.10 shows the top five ranked pathways for the 75 and 79 DEGs (among the top 500 genes) that were selected by all methods simultaneously for meta-datasets A and B, respectively. Table 4.11 shows the top five ranked pathways for the 133 and 129 DEGs (among the top 500 genes) that were selected by all methods simultaneously for meta-datasets A1 and B1, respectively.

Dataset A

Pathway Name	Impact Factor	p-value	corrected p-value
Asthma	4,59	0,01	0,14
Hematopoietic cell lineage	4,35	0,01	0,14
Apoptosis	4,31	0,01	0,14
Toll-like receptor signaling pathway	4,04	0,02	0,14
Graft-versus-host disease	4,02	0,02	0,14

Dataset B

Pathway Name	Impact Factor	p-value	corrected p-value
DNA replication	3,97	0,02	0,32
Tight junction	3,27	0,04	0,32
Epithelial cell signaling in Helicobacter pylori infection	2,93	0,05	0,32
Non-homologous end-joining	2,55	0,08	0,32
ECM-receptor interaction	2,40	0,09	0,32

Table 4.10: Pathway analysis results for meta-datasets A and B considering the top 500 DEGs selected by all the methods simultaneously

Dataset A1

Pathway Name	Impact Factor	p-value	corrected p-value
Cell adhesion molecules (CAMs)	8,92	1,33E-04	0,01
Neuroactive ligand-receptor interaction	6,28	0,00	0,04
Renin-angiotensin system	5,05	0,01	0,09
ECM-receptor interaction	4,27	0,01	0,14
Tight junction	3,77	0,02	0,17

Dataset B1

Pathway Name	Impact Factor	p-value	corrected p-value
Neuroactive ligand-receptor interaction	8,91	1,35E-04	0,00
ECM-receptor interaction	6,21	0,00	0,03
Cell adhesion molecules (CAMs)	5,00	0,01	0,07
Tight junction	3,29	0,04	0,29
Non-homologous end-joining	2,52	0,08	0,44

Table 4.11: Pathway analysis results for meta-datasets A1 and B1 considering the top 500 DEGs selected by all the methods simultaneously

Table 4.12 and Table 4.13 show the most perturbed KEGG pathways for the top 500 DEGs that were selected by each method, applied to meta-datasets A and B, respectively. Analogous information for datasets A1 and B1 is summarized in Table 4.14 and Table 4.15. A better agreement between the top pathways for the four meta-analysis approaches and between them and the top pathways provided by the intersection of all the methods was observed in the case of filtered datasets A1 and B1 compared to datasets A and B.

Despite the poor overlap of the results at the gene-level, pathway analysis showed a higher concordance. Moreover the differences among the DEGs lists for meta-datasets A and B were confirmed at the pathway-level. In fact, although some pathways such as PPAR signaling pathway, complement and coagulation cascades and cytokine-cytokine receptor interaction were identified in both datasets by two or more meta-analysis methods, the

ranking is quite different and there are also some pathways unique for each dataset (e.g. apoptosis pathway detected only in dataset A and neuroactive ligand-receptor interaction and cell adhesion molecules (CAMs) pathways retrieved only in dataset B). Furthermore it is worth noting that the impact factors (the FDR-corrected p-values) for the top pathways by the RankProd method are higher (respectively, lower) compared to all the other methods in all meta-datasets. A higher impact factor indicates that the top genes lists contain genes that aggregate into certain functions as opposed to individual genes that are unrelated. Thus results with higher impact factor may make more sense and be more easily interpretable.

DEG_limma_with_study_effect

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	8,88	1,39E-04	0,01
Complement and coagulation cascades	5,47	0,00	0,17
Apoptosis	4,05	0,02	0,35
Renin-angiotensin system	3,77	0,02	0,35
Thyroid cancer	3,66	0,03	0,35

DEGs_p-value combination (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	8,88	1,39E-04	0,01
Hematopoietic cell lineage	6,50	0,00	0,06
Adipocytokine signaling pathway	4,54	0,01	0,28
ECM-receptor interaction	4,26	0,01	0,28
Cytokine-cytokine receptor interaction	3,85	0,02	0,29

DEGs_metaMA (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
Complement and coagulation cascades	9,95	4,76E-05	0,00
Renin-angiotensin system	8,17	2,82E-04	0,01
Hematopoietic cell lineage	5,25	0,01	0,14
Cell adhesion molecules (CAMs)	4,93	0,01	0,15
Apoptosis	4,05	0,02	0,28

DEGs_GeneMeta (FDR)

Pathway Name	Impact Factor	p-value	corrected p-value
Renin-angiotensin system	5,82	0,00	0,15
Complement and coagulation cascades	5,47	0,00	0,15
Apoptosis	5,17	0,01	0,15
DNA replication	4,60	0,01	0,16
Cytokine-cytokine receptor interaction	4,57	0,01	0,16

DEGs_RankProd

Pathway Name	Impact Factor	p-value	corrected p-value
ECM-receptor interaction	11,10	1,51E-05	0,00
PPAR signaling pathway	8,88	1,39E-04	0,01
Focal adhesion	6,48	0,00	0,04
Adipocytokine signaling pathway	5,86	0,00	0,06
Systemic lupus erythematosus	4,92	0,01	0,11

Table 4.12: Pathway analysis results for meta-dataset A

DEGs_limma_with_study_effect

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	8,88	1,39E-04	0,01
Cell adhesion molecules (CAMs)	5,99	0,00	0,07
Adipocytokine signaling pathway	5,86	0,00	0,07
Hematopoietic cell lineage	5,25	0,01	0,10
ECM-receptor interaction	4,26	0,01	0,19

DEGs_p-value combination (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	8,88	1,39E-04	0,01
Cell adhesion molecules (CAMs)	5,99	0,00	0,07
Adipocytokine signaling pathway	5,86	0,00	0,07
Hematopoietic cell lineage	5,25	0,01	0,10
ECM-receptor interaction	4,26	0,01	0,19

DEGs_metaMA (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
Complement and coagulation cascades	8,35	2,37E-04	0,02
Cell adhesion molecules (CAMs)	5,99	0,00	0,10
PPAR signaling pathway	4,54	0,01	0,28
ECM-receptor interaction	4,26	0,01	0,28
Vibrio cholerae infection	3,89	0,02	0,30

DEGs_GeneMeta (FDR)

Pathway Name	Impact Factor	p-value	corrected p-value
Complement and coagulation cascades	8,35	2,37E-04	0,02
Cell adhesion molecules (CAMs)	5,99	0,00	0,10
Vibrio cholerae infection	5,19	0,01	0,14
ABC transporters	4,36	0,01	0,22
ECM-receptor interaction	4,26	0,01	0,22

DEGs_RankProd

Pathway Name	Impact Factor	p-value	corrected p-value
ECM-receptor interaction	11,10	1,51E-05	0,00
PPAR signaling pathway	7,31	0,00	0,03
Focal adhesion	6,48	0,00	0,04
Cell adhesion molecules (CAMs)	4,93	0,01	0,11
Systemic lupus erythematosus	4,92	0,01	0,11

Table 4.13: Pathway analysis results for meta-dataset B

DEGs_limma_with_study_effect

Pathway Name	Impact Factor	p-value	corrected p-value
Renin-angiotensin system	7,52	5,42E-04	0,04
Neuroactive ligand-receptor interaction	6,55	0,00	0,05
Cell adhesion molecules (CAMs)	5,74	0,00	0,07
Cytokine-cytokine receptor interaction	5,59	0,00	0,07
PPAR signaling pathway	4,79	0,01	0,12

DEGs_p-value combination (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	6,23	0,00	0,13
Cell adhesion molecules (CAMs)	5,74	0,00	0,13
Renin-angiotensin system	4,76	0,01	0,19
Cytokine-cytokine receptor interaction	4,65	0,01	0,19
Neuroactive ligand-receptor interaction	4,08	0,02	0,23

DEGs_metaMA (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
Neuroactive ligand-receptor interaction	6,55	0,00	0,09
Cell adhesion molecules (CAMs)	5,74	0,00	0,09
Cytokine-cytokine receptor interaction	5,59	0,00	0,09
Complement and coagulation cascades	5,38	0,00	0,09
Renin-angiotensin system	4,76	0,01	0,14

DEGs_GeneMeta (FDR)

Pathway Name	Impact Factor	p-value	corrected p-value
Cell adhesion molecules (CAMs)	5,74	0,00	0,14
Complement and coagulation cascades	5,38	0,00	0,14
Neuroactive ligand-receptor interaction	5,25	0,01	0,14
Renin-angiotensin system	4,76	0,01	0,17
Cytokine-cytokine receptor interaction	3,78	0,02	0,36

DEGs_RankProd

Pathway Name	Impact Factor	p-value	corrected p-value
ECM-receptor interaction	14,05	7,89E-07	5,84E-05
Focal adhesion	11,35	1,18E-05	4,35E-04
Cytokine-cytokine receptor interaction	10,16	3,87E-05	9,56E-04
PPAR signaling pathway	9,56	7,07E-05	0,00
Neuroactive ligand-receptor interaction	6,55	0,00	0,02

Table 4.14: Pathway analysis results for meta-dataset A1

DEGs_limma_with_study_effect

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	9,89	5,08E-05	0,00
Cell adhesion molecules (CAMs)	6,29	0,00	0,07
Neuroactive ligand-receptor interaction	5,89	0,00	0,07
Adipocytokine signaling pathway	4,26	0,01	0,28
ECM-receptor interaction	3,81	0,02	0,35

DEGs_p-value combination (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
PPAR signaling pathway	8,10	3,03E-04	0,02
Cell adhesion molecules (CAMs)	6,29	0,00	0,07
Neuroactive ligand-receptor interaction	4,39	0,01	0,28
Adipocytokine signaling pathway	4,26	0,01	0,28
ECM-receptor interaction	3,81	0,02	0,35

DEGs_metaMA (moderated t-statistics)

Pathway Name	Impact Factor	p-value	corrected p-value
Neuroactive ligand-receptor interaction	9,40	8,28E-05	0,01
Leukocyte transendothelial migration	6,14	0,00	0,08
Cell adhesion molecules (CAMs)	5,16	0,01	0,15
Complement and coagulation cascades	4,26	0,01	0,28
DNA replication	3,59	0,03	0,43

DEGs_GeneMeta (FDR)

Pathway Name	Impact Factor	p-value	corrected p-value
Neuroactive ligand-receptor interaction	9,40	8,28E-05	0,01
DNA replication	6,71	0,00	0,05
Complement and coagulation cascades	5,58	0,00	0,10
Cell adhesion molecules (CAMs)	4,13	0,02	0,29
Leukocyte transendothelial migration	4,02	0,02	0,29

DEGs_RankProd

Pathway Name	Impact Factor	p-value	corrected p-value
ECM-receptor interaction	15,93	1,20E-07	8,90E-06
Focal adhesion	10,19	3,77E-05	0,00
PPAR signaling pathway	9,89	5,08E-05	0,00
Neuroactive ligand-receptor interaction	9,40	8,28E-05	0,00
Cell adhesion molecules (CAMs)	8,84	1,46E-04	0,00

Table 4.15: Pathway analysis results for meta-dataset B1

4.3 DISCUSSION

The increasing availability and maturity of DNA microarray technology has led to an explosion of profiling studies, especially in cancer research. To extract maximum value from the accumulating mass of publicly available cancer gene expression data, methods are needed to evaluate, integrate, and inter-validate multiple datasets. Therefore, we applied a combination of statistical methods to perform a meta-analysis of independent microarray datasets in MPM tumor.

A relevant number of studies on gene expression in MPM have been published so far. Gene expression profiling microarrays have been widely used in mesothelioma research to improve histological classifications [157], to identify predictive or prognostic biomarkers [158, 159] or to examine response to therapy [160]. One of the problems identified with the use of microarray technology and particularly in experiments involving MPM samples, has been the lack of concordance between the several studies. From this, one can conclude that microarray analysis of MPM would appear to be very 'noisy', and the differential expression of the identified genes should be confirmed independently at the RNA (RT-qPCR) or protein level (immunohistochemistry) [161]. There are several potential reasons for the low concordance of these studies such as different sample types (e.g. mesothelioma derived cell lines or mesothelioma patient samples) and/or histological subtypes, array platforms and number of samples. Therefore, a comparison of the results of individual analyses is not enough to evaluate the available gene expression data. Gene expression data often lack statistical power especially due to low sample size, as is the case in most of MPM microarray studies. This might depend on the low incidence of MPM in addition to the costs of the technology

and generally leads to underestimation of variances, which inflates the false-positive rate. Meta-analysis thus represents a good solution to overcome the problem of reduced statistical power of MPM microarray experiments and a valuable alternative for cross-study validation.

The quality of the meta-analysis benefits from the number and the quality of single datasets analyzed. Focusing on differential expression analysis between normal and malignant pleural mesothelioma tissues, only three microarray studies compliant with our inclusion/exclusion criteria were identified and included in the meta-analysis. All the selected studies were performed on two different array versions of the same Affymetrix platform. We tried to obtain raw data for all of them but only the two most recent studies (Crispi and Røe) have the CEL files stored in public databases. For the Gordon study only MAS 5.0 preprocessed gene expression data were publicly available and this might heavily affect the down-stream analysis. In fact, to obtain consistently preprocessed data, the implementation of the MAS 5.0 algorithm provided by the Bioconductor package simpleaffy was used to process Crispi and Røe datasets. This choice has two limitations. On the one hand, there may be some not negligible differences between the expression data provided by the original MAS 5.0 algorithm and its re-implementation. On the other hand, in the last years new preprocessing methods, such as RMA and gcrma, have been developed that outperform MAS 5.0 in terms of sensitivity and specificity (i.e. the true and false detection rate) [162], especially for the detection of DEGs [163]. Despite the use of the same preprocessing method, our data showed a strong between-study variability. Gene expression data obtained applying to each dataset the preprocessing algorithms suggested by the original authors were also considered for the meta-analysis with the aim to evaluate the impact of

the preprocessing methods on the meta-analysis results. In addition, the number of genes evaluated was reduced applying an intensity filter followed by an IQR filter in order to eliminate non-expressed or non-informative genes and increase statistical power in multiple comparison procedure.

To date only two papers by Hong and colleagues [109] and Campain and colleagues [99] performed a systematic comparative analysis on microarray meta-analysis methods performances, in terms of sensitivity and specificity. Although the two studies provided insightful conclusions, the number of methods compared (three and five methods, respectively) and the number of case studies examined (two and three case studies, respectively with each case study combining only 2–5 microarray datasets) were very limited. In addition, some key conclusions from the two papers were even contradictory. Therefore practical guidelines for choosing the 'best' meta-analysis method(s) still lack.

Four meta-analysis methods corresponding to the three most common relative meta-analysis approaches (i.e. combining p-values, combining effect size and combining ranks) were applied to the selected MPM datasets using three R and Bioconductor packages (metaMA, GeneMeta and RankProd). In addition the preprocessed data from the three studies have been directly combined and analyzed using limma, including a study effect in the linear model. This approach can be viewed as an alternative meta-analysis method.

The different methods resulted in significant gene lists of different sizes. The highest number of DEGs was detected by the weighted inverse normal p-value combination method. This high proportion of significant genes may be due to the fact that p-value combination approaches are in general prone to be driven by significant results of individual studies. This

drawback may be partly reduced by the introduction of study-specific weights as demonstrated by Li and Ghosh [164]. On the other hand, effect size combination methods were found to be the most conservative. Taking into account the between-study heterogeneity, which is particularly evident in our data, they lead to a more reliable and perhaps more meaningful set of commonly DEGs. The RankProd approach identified an intermediate number of DEGs. Although it does not incorporate the between-study variability, it has been widely shown that gene rankings from the RankProd method are more robust against noise and other hidden variables embedded in different datasets [109]. Finally, limma analysis including a study effect in the linear model also appeared to be a valuable alternative for meta-analysis.

A poor overlap between the DEGs lists provided by each method was observed, both considering the complete lists and the top 500 most significant DEGs. The overlap increases when filtered data are considered indicating a possible beneficial effect of filtering on our data. The fact that a significant number of genes were only detected by one of the methods stems from the different assumptions and ranking criteria on which the various methods are based. Therefore, when doing meta-analysis on real data, it might be useful to select candidate genes using a combination of methods, so as to capture genes that are interesting from different aspects. In particular, DEGs detected by more than one meta-analysis method may be considered as the most reliable ones while DEGs identified by only one method may be further explored to enrich the knowledge of the biological phenomenon of interest.

Despite the poor overlap of the results at the gene-level, the pathway analysis showed a higher concordance between the different methods, in particular when the filtered data were considered.

The present meta-analysis and the final DEGs lists may have some potential value as regards MPM, since this is the first attempt to integrate microarray datasets in this area. However such genes lists should be taken with caution due to the limitations of this microarray meta-analysis, first of all the small number of studies included. Regarding meta-analysis methodology with microarray data, some limitations emerge from this work. The fact that several microarray studies share the same hypothesis would not be sufficient to successfully integrate results; such studies should use the same sample sources (biological equivalency of cases and controls across studies), similar sample processing protocols and the same microarray platform, with identical probes in the chips. Otherwise, the search for genes in common among platforms and the precision of data could reduce the power of individual studies instead of increasing it. Furthermore, if DEGs do not present large differences among the groups compared, the results of the meta-analyses could be strongly affected by experimental error and patient variability. Finally, the sensitivity of the results from meta-analysis should be tested before a final conclusion is reached. We could not perform any sensitivity analysis because of the small number of included studies and the lack of suitable packages/tools. In fact, the sensitivity analysis has so far been largely neglected in the meta-analysis of microarray data.

5 CONCLUSIONS

Despite the increasing popularity of microarray meta-analysis, many issues remain unsolved that can hinder the effectiveness of its application.

Although many methods have been proposed and used in published applications, a detailed workflow to perform microarray data meta-analysis does not exist yet.

Although many grant agencies and peer-review journals now require to make data available, many old studies or new studies funded by private foundations are still not publicly accessible. Studies with censored or incomplete information can be an obstacle for meta-analysis.

It is still unclear how measurements from different platforms compare with each other and inconsistencies in gene coverage and annotation make comparison much more difficult.

Several microarray meta-analysis methods have been developed. The selection of a suitable meta-analysis method depends on the type of analysis desired and the hypothesis setting behind each method. Ramasamy and colleagues [48] recommend effect size combination

methods as the most comprehensive approach for meta-analysis of two-class gene expression microarrays due to its known advantages However there is no consensus on what is(are) the ‘best’ meta-analysis method(s). A large-scale comparative study and simulation study with adequate evaluation measures are needed to provide insights and practical guidelines for choosing the ‘best’ meta-analysis method(s) in practice.

Only a few of the proposed microarray meta-analysis methods are developed in easy to use software packages. In addition to the scarcity of software packages in the field and the lack of a regular update of the existing ones, most of the available packages either did not have clear manuals or had functions that were not easy to apply. Efforts to provide high-quality documentation of programs in order to make them more reliable and easier to comprehend are well summarized by the concept of ‘literate programming’ and its implementation developed by Knuth [Knuth DE, 1984]. The package Sweave [165] is an example of use of the noweb-like literate programming tool [166] inside the R language for creation of dynamic statistical reports. It provides a flexible framework for mixing text and R code for automatic document generation. A single source file contains both documentation text and R code, which are then embedded into a final document containing the documentation text together with the R code and/or the output of the code (e.g. text, graphs, tables) by running the code through R. The report can be automatically updated if data or analysis change, which allows for truly reproducible research.

All packages available by Bioconductor now should meet this requirement and, in fact, the most recent packages contain one or more ‘vignettes’, that is documents providing a textual, task-oriented description of the package's functionality. Due to its widely recognized

benefits, literate programming practice should be promoted in future software development.

Heterogeneities caused by demographic, clinical and technical variables often exist within and across studies. Failure to consider these potential confounding variables in the statistical models and meta-analysis can result in reduced statistical power or false positives. Meta-analyses of clinical and epidemiological studies use regression modeling to adjust for the confounding effects. Only recently similar techniques have been extended to microarray data meta-analysis [167] however further efforts are needed in this area.

BIBLIOGRAPHY

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.
2. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer**. *Nat Genet* 1996, **14**(4):457-460.
3. Dumitriu A, Latourelle JC, Hadzi TC, Pankratz N, Garza D, Miller JP, Vance JM, Foroud T, Beach TG, Myers RH: **Gene expression profiles in Parkinson disease prefrontal cortex implicate FOXO1 and genes under its transcriptional regulation**. *PLoS Genet* 2012, **8**(6):e1002794.
4. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M *et al*: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* 2001, **98**(24):13790-13795.
5. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**(6769):503-511.
6. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**(5439):531-537.
7. Ohno R, Nakamura Y: **Prediction of response to imatinib by cDNA microarray analysis**. *Semin Hematol* 2003, **40**(2 Suppl 2):42-49.
8. van't Veer LJ, Dai H, van de Vijver MJ,He YD, Hart AA, Mao M, Peterse HL,van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
9. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS *et al*: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning**. *Nat Med* 2002, **8**(1):68-74.

10. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.
11. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG *et al*: **ArrayExpress--a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31**(1):68-71.
12. Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y: **CIBEX: center for information biology gene expression database**. *C R Biol* 2003, **326**(10-11):1079-1082.
13. Sarmah CK, Samarasinghe S: **Microarray Data Integration: Frameworks and a List of Underlying Issues**. *Current Bioinformatics* 2010, **5**:280-289.
14. Hedges LV, Olkin I: **Statistical methods for meta-analysis**: Academic Press, Orlando; 1985.
15. Normand SL: **Meta-analysis: formulating, evaluating, combining, and reporting**. *Stat Med* 1999, **18**(3):321-359.
16. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling interstudy variation**. *Bioinformatics* 2003, **19 Suppl 1**:i84-90.
17. Frantz S: **An array of problems**. *Nat Rev Drug Discov* 2005, **4**(5):362-363.
18. Miklos GL, Maleszka R: **Microarray reality checks in the context of a complex disease**. *Nat Biotechnol* 2004, **22**(5):615-621.
19. Tan PK, Downey TJ, Spitznagel EL, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: **Evaluation of gene expression measurements from commercial microarray platforms**. *Nucleic Acids Res* 2003, **31**(19):5676-5684.
20. Klebanov L, Qiu X, Welle S, Yakovlev A: **Statistical methods and microarray data**. *Nat Biotechnol* 2007, **25**(1):25-26; author reply 26-27.
21. Shi L, Tong W, Fang H, Scherf U, Han J, Puri RK, Frueh FW, Goodsaid FM, Guo L, Su Z *et al*: **Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential**. *BMC Bioinformatics* 2005, **6 Suppl 2**:S12.
22. Olkin I: **Meta-Analysis: Current Issues In Research Synthesis**. *Statistics in medicine* 1996, **15**(12):1253-1257.
23. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression**

profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 2002, **62**(15):4427-4433.

24. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci U S A* 2004, **101**(25):9309-9314.
25. Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV: **Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies**. *Bioinformatics* 2004, **20**(17):3166-3178.
26. Grützmann R, Boriss H, Ammerpohl O, Lüttges J, Kalthoff H, Schackert HK, Klöppel G, Saeger HD, Pilarsky C: **Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes.** *Oncogene* 2005, **24**(32):5079-5088.
27. Tseng GC, Ghosh D, Feingold E: **Comprehensive literature review and statistical considerations for microarray meta-analysis**. *Nucleic Acids Res* 2012, **40**(9):3785-3799.
28. Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ, Kim S: **Integrative analysis of multiple gene expression profiles applied to liver cancer study**. *FEBS Lett* 2004, **565**(1-3):93-100.
29. Griffith OL, Melck A, Jones SJ, Wiseman SM: **Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers.** *J Clin Oncol* 2006, **24**(31):5043-5051.
30. Burguillo FJ, Martin J, Barrera I, Bardsley WG: **Meta-analysis of microarray data: The case of imatinib resistance in chronic myelogenous leukemia.** *Comput Biol Chem* 2010, **34**(3):184-192.
31. Warnat P, Eils R, Brors B: **Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes**. *BMC Bioinformatics* 2005, **6**:265.
32. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer**. *Nat Genet* 2004, **36**(10):1090-1098.
33. Wang Y, Joshi T, Zhang XS, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets**. *Bioinformatics* 2006, **22**(19):2413-2420.
34. Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH: **Functional**

annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol* 2005, **23**(2):238-243.

35. Garrett-Mayer E, Parmigiani G, Zhong X, Cope L, Gabrielson E: **Cross-study validation and combined analysis of gene expression microarray data.** *Biostatistics* 2008, **9**(2):333-354.
36. Yang X, Bentink S, Scheid S, Spang R: **Similarities of ordered gene lists**. *J Bioinform Comput Biol* 2006, **4**(3):693-708.
37. Yang X, Sun X: **Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers**. *BMC Bioinformatics* 2007, **8**:118.
38. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G, McCaffrey TA: **Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization.** *Gene* 2007, **401**(1-2):12-18.
39. Marot G, Mayer CD: **Sequential analysis for microarray data based on sensitivity and meta-analysis**. *Stat Appl Genet Mol Biol* 2009, **8**:Article 3.
40. Marot G, Foulley JL, Mayer CD, Jaffrézic F: **Moderated effect size and P-value combinations for microarray meta-analyses.** *Bioinformatics* 2009, **25**(20):2692-2699.
41. Gentleman R, Ruschhaupt M, Huber W, Lusa L: **Meta-analysis for Microarray Experiments**. In.; 2012.
42. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J: **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.** *Bioinformatics* 2006, **22**(22):2825-2827.
43. Sutton A, Abrams K, Jones D, Sheldon T, Song F: **Methods for meta-analysis in medical research** New York: Wiley; 2000.
44. Deeks JJ, Altman DG, Bradburn MJ: **Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis**. In: *Systematic reviews in health care: Meta-analysis in context* Edited by Egger M DSG, Altman D, editors. London: BMJ Publishing Group 2001: 285-312.
45. Whitehead A: **Meta-analysis of controlled clinical trials**, 1st edition edn. Chichester (United Kingdom): Wiley; 2002.
46. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF: **Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses**. *Lancet* 1999, **354**(9193):1896-1900.

47. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P: **Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement**. *J Clin Epidemiol* 2009, **62**(10):1006-1012.
48. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key issues in conducting a meta-analysis of gene expression microarray datasets.** *PLoS Med* 2008, **5**(9):e184.
49. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR: **Introduction to meta-analysis**. In.: Wiley; 2009.
50. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC *et al*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**. *Nat Genet* 2001, **29**(4):365-371.
51. Suárez-Fariñas M, Noggle S, Heke M, Hemmati-Brivanlou A, Magnasco MO: **Comparing independent microarray studies: the case of human embryonic stem cells**. *BMC Genomics* 2005, **6**:99.
52. Manduchi E, Grant GR, He H, Liu J, Mailman MD, Pizarro AD, Whetzel PL, Stoeckert CJ: **RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies**. *Bioinformatics* 2004, **20**(4):452-459.
53. Marc P, Devaux F, Jacq C: **yMGV: a database for visualization and data mining of published genome-wide yeast expression data**. *Nucleic Acids Res* 2001, **29**(13):E63-63.
54. Zhang J, Finney RP, Rowe W, Edmonson M, Yang SH, Dracheva T, Jen J, Struewing JP, Buetow KH: **Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB)**. *Genome Res* 2007, **17**(7):1111-1117.
55. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M: **Comparison and meta-analysis of microarray data: from the bench to the computer desk.** *Trends Genet* 2003, **19**(10):570-577.
56. Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M *et al*: **Submission of microarray data to public repositories**. *PLoS Biol* 2004, **2**(9):E317.
57. Brazma A: **Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges**. *ScientificWorldJournal* 2009, **9**:420-423.
58. Larsson O, Sandberg R: **Lack of correct data format and comparability limits future integrative microarray research**. *Nat Biotechnol* 2006, **24**(11):1322-1323.

59. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G *et al*: **Repeatability of published microarray gene expression analyses**. *Nat Genet* 2009, **41**(2):149-155.
60. Wang HY, Malek RL, Kwitek AE, Greene AS, Luu TV, Behbahani B, Frank B, Quackenbush J, Lee NH: **Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays**. *Genome Biol* 2003, **4**(1):R5.
61. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al*: **Expression monitoring by hybridization to high-density oligonucleotide arrays**. *Nat Biotechnol* 1996, **14**(13):1675-1680.
62. Koltai H, Weingarten-Baror C: **Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction**. *Nucleic Acids Res* 2008, **36**(7):2395-2405.
63. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33**(Database issue):D501-504.
64. Järvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O: **Are data from different gene expression microarray platforms comparable?** *Genomics* 2004, **83**(6):1164-1168.
65. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies**. *Bioinformatics* 2002, **18**(3):405-412.
66. Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC: **Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays**. *Nucleic Acids Res* 2002, **30**(10):e48.
67. Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ: **Spotted long oligonucleotide arrays for human gene expression analysis**. *Genome Res* 2003, **13**(7):1775-1785.
68. Severgnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, Ghidoni R, Peano C, Bonnal R, Viti F *et al*: **Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment**. *Anal Biochem* 2006, **353**(1):43-56.
69. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms**. *Nat Methods* 2005, **2**(5):337-344.

70. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G *et al*: **Multiple-laboratory comparison of microarray platforms**. *Nat Methods* 2005, **2**(5):345-350.
71. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY *et al*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nat Biotechnol* 2006, **24**(9):1151-1161.
72. Buness A, Huber W, Steiner K, Sültmann H, Poustka A: **arrayMagic: two-colour cDNA microarray quality control and preprocessing**. *Bioinformatics* 2005, **21**(4):554-556.
73. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis**. *Bioinformatics* 2005, **21**(18):3683-3685.
74. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S: **Bioinformatics and Computational Biology Solutions Using R and Bioconductor**: Springer; 2005.
75. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data**. *Bioinformatics* 2009, **25**(3):415-416.
76. Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, Pine PS, Boysen C, Guo X, Chudin E *et al*: **Using RNA sample titrations to assess microarray platform performance and normalization techniques**. *Nat Biotechnol* 2006, **24**(9):1123-1131.
77. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18 Suppl 1**:S96-104.
78. Burges CJC: **A Tutorial on Support Vector Machines for Pattern Recognition** *Data Mining and Knowledge Discovery* 1998, **2**:121-167.
79. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proc Natl Acad Sci U S A* 2000, **97**(18):10101-10106.
80. Jolliffe IT: **Principal Component Analysis** Springer; 2002.
81. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS: **Adjustment of systematic microarray data biases**. *Bioinformatics* 2004, **20**(1):105-114.
82. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solís DY, Duque R, Bersini H, Nowé A: **Batch effect removal**

methods for microarray gene expression data integration: a survey. *Brief Bioinform* 2012.

83. Shippy R, Sendera TJ, Lockner R, Palaniappan C, Kaysser-Kranich T, Watts G, Alsobrook J: **Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations**. *BMC Genomics* 2004, **5**:61.
84. Kruskal JB, Wish M: **Multidimensional scaling**. Beverly Hills: SAGE Publications; 1978.
85. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA *et al*: **Database resources of the National Center for Biotechnology**. *Nucleic Acids Res* 2003, **31**(1):28-33.
86. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2007, **35**(Database issue):D26-31.
87. Allen JD, Wang S, Chen M, Girard L, Minna JD, Xie Y, Xiao G: **Probe mapping across multiple microarray platforms**. *Brief Bioinform* 2012, **13**(5):547-554.
88. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
89. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z: **Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements**. *BMC Bioinformatics* 2005, **6**:107.
90. Choi H, Shen R, Chinnaiyan AM, Ghosh D: **A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments**. *BMC Bioinformatics* 2007, **8**:364.
91. Kuhn A, Luthi-Carter R, Delorenzi M: **Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'**. *BMC Bioinformatics* 2008, **9**:26.
92. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO *et al*: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data**. *Nucleic Acids Res* 2003, **31**(1):219-223.
93. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J: **RESOURCERER: a database for annotating and linking microarray resources within and across species**. *Genome Biol* 2001, **2**(11):SOFTWARE0002.

94. Baron D, Bihouée A, Teusan R, Dubois E, Savagner F, Steenman M, Houlgatte R, Ramstein G: **MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets**. *Bioinformatics* 2011, **27**(5):725-726.
95. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
96. Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA: **Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate**. *Nucleic Acids Res* 2003, **31**(13):3775-3781.
97. Hahne F, Huber W, Gentleman R, Falcon S: **Bioconductor Case Studies**: Springer; 2008.
98. Larsson O, Wennmalm K, Sandberg R: **Comparative microarray analysis**. *OMICS* 2006, **10**(3):381-397.
99. Campain A, Yang YH: **Comparison study of microarray meta-analysis methods.** *BMC Bioinformatics* 2010, **11**:408.
100. Hwang KB, Kong SW, Greenberg SA, Park PJ: **Combining gene expression data from different generations of oligonucleotide arrays**. *BMC Bioinformatics* 2004, **5**:159.
101. Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, Tsai CJ, Zhang S: **Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes**. *BMC Bioinformatics* 2004, **5**:81.
102. Lee Y, Scheck AC, Cloughesy TF, Lai A, Dong J, Farooqi HK, Liau LM, Horvath S, Mischel PS, Nelson SF: **Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age**. *BMC Med Genomics* 2008, **1**:52.
103. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, Miller CJ, Clarke RB: **The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis**. *BMC Med Genomics* 2008, **1**:42.
104. Bushman BJ, Cooper H, Hedges LV: **Vote counting methods in meta-analysis** In: *The handbook ofresearch synthesis* New York: Russel Sage Foundation Publications; 1994: 193-214.
105. Smid M, Dorssers LC, Jenster G: **Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes.** *Bioinformatics* 2003, **19**(16):2065-2071.

106. Fisher R: **Statistical methods for research workers**, 4th edition edn. London: Oliver and Boyd; 1932.
107. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society* 1995, **57**:289-300.
108. Storey JD: **A direct approach to false discovery rates**. *Journal of the Royal Statistical Society* 2002, **64**(3):479-498.
109. Hong F, Breitling R: **A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.** *Bioinformatics* 2008, **24**(3):374-382.
110. Li J, Tseng GC: **An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies**. *Ann Appl Stat* 2011, **5**:994-1019.
111. Tippet LHC: **The methods of statistics**. London: William and Norgate Ltd.; 1931.
112. Stouffer SA, Suchman EA, DeVinnery L, Star S, Williams RMJ: **The American Soldier, Volume I: Adjustement during Army Life** Princeton, NJ: Princeton University Press; 1949.
113. Wilkinson B: **A statistical consideration in psychological research**. *Psychol Bull* 1951, **48**(3):156-158.
114. Kelley K, Preacher KJ: **On effect size**. *Psychological Methods* 2012, **17**:137-152.
115. Rosnow RL, Rosenthal R: **Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers** *Psychological Methods* 1996, **1**:331-340.
116. Cohen J: **Statistical power analysis for behavioral sciences**. New Jersey: Lawrence Erbaum; 1988.
117. Rosenthal R, DiMatteo MR: **Meta-analysis: recent developments in quantitative methods for literature reviews**. *Annu Rev Psychol* 2001, **52**:59-82.
118. Cooper H, Hedges LV: **The handbook of research synthesis**. New York: Russell Sage Foundation; 1994.
119. Cochran BG: **The combination of estimates from different experiments**. *Biometrics* 1954, **10**:101-129.
120. Jones MP, O'Gorman TW, Lemke JH, Woolson RF: **A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations**. *Biometrics* 1989, **45**(1):171-181.
121. DerSimonian R, Laird N: **Meta-analysis in clinical trials**. *Control Clin Trials* 1986, **7**(3):177-188.

122. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.
123. Hu P, Greenwood CM, Beyene J: **Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models**. *BMC Bioinformatics* 2005, **6**:128.
124. Hu P, Greenwood CM, Beyene J: **Using the ratio of means as the effect size measure in combining results of microarray experiments**. *BMC Syst Biol* 2009, **3**:106.
125. Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM *et al*: **Rat toxicogenomic study reveals analytical consistency across microarray platforms**. *Nat Biotechnol* 2006, **24**(9):1162-1169.
126. Stevens JR, Doerge RW: **Combining Affymetrix microarray results.** *BMC Bioinformatics* 2005, **6**:57.
127. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
128. Jaffrézic F, Marot G, Degrelle S, Hue I, Foulley JL: **A structural mixed model for variances in differential gene expression studies**. *Genet Res* 2007, **89**(1):19-25.
129. Welch BL: **The Generalization of "Student's" problem when several different population variances are involved**. *Biometrika* 1947, **34**(1/2):28-35.
130. Hedges LV: **Distribution theory for Glass's estimator of effect size and related estimators**. *Journal of Educational Statistics* 1981, **6**:107-128.
131. Satterthwaite FE: **Synthesis of variance**. *Psychometrika* 1941, **6**(5):309-316.
132. Conlon EM, Song JJ, Liu A: **Bayesian meta-analysis models for microarray data: a comparative study**. *BMC Bioinformatics* 2007, **8**:80.
133. Zintzaras E, Ioannidis JP: **Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays**. *Comput Biol Chem* 2008, **32**(1):38-46.
134. Dreyfuss JM, Johnson MD, Park PJ: **Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers**. *Mol Cancer* 2009, **8**:71.
135. Zintzaras E, Ioannidis JP: **Heterogeneity testing in meta-analysis of genome searches**. *Genet Epidemiol* 2005, **28**(2):123-137.

136. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments**. *FEBS Lett* 2004, **573**(1-3):83-92.
137. DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R: **Combining results of microarray experiments: a rank aggregation approach**. *Stat Appl Genet Mol Biol* 2006, **5**:Article15.
138. Dwork C, Kumar R, Naor M, Sivakumar D: **Rank aggegation methods for the web** *Procs* 2001:613-622.
139. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E: **A statistical framework for expression-based molecular classification in cancer** *JRSS* 2002:717-736.
140. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E: **A cross-study comparison of gene expression studies for the molecular classification of lung cancer.** *Clin Cancer Res* 2004, **10**(9):2922-2927.
141. Lottaz C, Yang X, Sheid S, Spang R: ***OrderedList*—a bioconductor package for detecting similarity in ordered gene lists**. *Bioinformatics* 2006, **22**(18):2315-2316.
142. Pihur H, Datta S, Datta S: **RankAggreg, an R package for weighted rank aggregation**. *BMC Bioinformatics* 2009, **10**(62).
143. Wang X, Kang DD, Shen K, Song C, Lu S, Chang LC, Liao SG, Huo Z, Tang S, Ding Y *et al*: **An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection**. *Bioinformatics* 2012, **28**(19):2534-2536.
144. Kang DD, Sibille E, Kaminski N, Tseng GC: **MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis**. *Nucleic Acids Res* 2012, **40**(2):e15.
145. Song C, Tseng GC: **Hypothesis setting and order statistic for robust genomic meta-analysis.** . *Annals of Applied Statistics* 2012.
146. Shen K, Tseng GC: **Meta-analysis for pathway enrichment analysis when combining multiple genomic studies**. *Bioinformatics* 2010, **26**(10):1316-1323.
147. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays**. *Journal of the American Statistical Association* 2004, **99**(468):909-917.
148. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of**

high density oligonucleotide array probe level data. *Biostatistics* 2003, **4**(2):249-264.

149. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA: **genenames.org: the HGNC resources in 2011**. *Nucleic Acids Res* 2011, **39**(Database issue):D514-519.
150. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**(3):307-315.
151. Servant N, Gravier E, Gestraud P, Laurent C, Paccard C, Biton A, Brito I, Mandel J, Asselain B, Barillot E *et al*: **EMA - A R package for Easy Microarray data analysis**. *BMC Res Notes* 2010, **3**:277.
152. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis**. *Genome Res* 2007, **17**(10):1537-1545.
153. Gordon GJ, Rockwell GN, Jensen RV, Rheinwald JG, Glickman JN, Aronson JP, Pottorf BJ, Nitz MD, Richards WG, Sugarbaker DJ *et al*: **Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling**. *Am J Pathol* 2005, **166**(6):1827-1840.
154. Crispi S, Calogero RA, Santini M, Mellone P, Vincenzi B, Citro G, Vicidomini G, Fasano S, Meccariello R, Cobellis G *et al*: **Global gene expression profiling of human pleural mesotheliomas: identification of matrix metalloproteinase 14 (MMP-14) as potential tumour target**. *PLoS One* 2009, **4**(9):e7016.
155. Røe OD, Anderssen E, Sandeck H, Christensen T, Larsson E, Lundgren S: **Malignant pleural mesothelioma: genome-wide expression patterns reflecting general resistance mechanisms and a proposal of novel targets**. *Lung Cancer* 2010, **67**(1):57-68.
156. Røe OD, Anderssen E, Helge E, Pettersen CH, Olsen KS, Sandeck H, Haaverstad R, Lundgren S, Larsson E: **Genome-wide profile of pleural mesothelioma versus parietal and visceral pleura: the emerging gene portrait of the mesothelioma phenotype**. *PLoS One* 2009, **4**(8):e6554.
157. Hoang CD, D'Cunha J, Kratzke MG, Casmey CE, Frizelle SP, Maddaus MA, Kratzke RA: **Gene expression profiling identifies matriptase overexpression in malignant mesothelioma**. *Chest* 2004, **125**(5):1843-1852.
158. Pass HI, Liu Z, Wali A, Bueno R, Land S, Lott D, Siddiq F, Lonardo F, Carbone M, Draghici S: **Gene expression profiles predict survival and progression of pleural mesothelioma**. *Clin Cancer Res* 2004, **10**(3):849-859.

159. López-Ríos F, Chuai S, Flores R, Shimizu S, Ohno T, Wakahara K, Illei PB, Hussain S, Krug L, Zakowski MF *et al*: **Global gene expression profiling of pleural mesotheliomas: overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction**. *Cancer Res* 2006, **66**(6):2970-2979.
160. Schrump DS, Fischette MR, Nguyen DM, Zhao M, Li X, Kunst TF, Hancox A, Hong JA, Chen GA, Pishchik V *et al*: **Phase I study of decitabine-mediated gene expression in patients with cancers involving the lungs, esophagus, or pleura**. *Clin Cancer Res* 2006, **12**(19):5777-5785.
161. Gray SG, Fennell DA, Mutti L, O'Byrne KJ: **In arrayed ranks: array technology in the study of mesothelioma**. *J Thorac Oncol* 2009, **4**(3):411-425.
162. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Res* 2003, **31**(4):e15.
163. Harr B, Schlötterer C: **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons**. *Nucleic Acids Res* 2006, **34**(2):e8.
164. Li Y, Ghosh D: **Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data**. *Bioinformatics* 2012, **28**(6):807-814.
165. Leisch F: **Sweave: Dynamic generation of statistical reports using literate data analysis**. *Compstat 2002 - Proceedings in Computational Statistics* 2002(69):575-580.
166. Johnson AL, Johnson BC: **Literate programming using noweb**. *Linux Journal* 1997:64-69.
167. Wang X, Lin Y, Song C, Sibille E, Tseng GC: **Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder**. *BMC Bioinformatics* 2012, **13**:52.

INDEX