Comparison of meta-analysis to combined analysis of a replicated microarray study

Darlene R. Goldstein[1], Mauro Delorenzi[2], Ruth Luthi-Carter[3]
and Thierry Sengstag[2]

[1]École Polytechnique Fédérale de Lausanne (EPFL)
Institut de mathématiques
CH-1015 Lausanne, Switzerland

[2]Bioinformatics Core Facility
Institut Suisse de Recherche Expérimentale sur le Cancer (ISREC)
and Swiss Institute of Bioinformatics
CH-1066 Epalinges, Switzerland

[3]École Polytechnique Fédérale de Lausanne (EPFL)
Laboratoire de neurogénomique fonctionnelle
CH-1015 Lausanne, Switzerland

1 Introduction

Microarray technologies measure mRNA abundance for thousands of genes in parallel. The high throughput nature of microarrays has contributed to their rise in importance for studying the molecular basis of fundamental biological processes and complex disease traits. Whereas only a few years ago microarray experiments were uncommon, they are now regularly used in a great variety of biological and medical studies.

Several different types of microarray platforms are available. Those currently in common use include high-density short oligonucleotide arrays, such as Affymetrix GeneChip ® arrays; long oligonucleotide arrays, such as those produced by Agilent; and cDNA arrays, fabricated in laboratories on site at many academic and commercial institutions.

The widespread use of microarrays has resulted in a large-scale, rapid expansion of data. Many research groups throughout the world are engaged in gene expression studies of the same or similar conditions – specific cancers, for example. Data from many microarray

studies are deposited in publicly available databases such as Gene Expression Omnibus (GEO) [3, 13]. It is hoped that ready access to the data will facilitate the integration of information across different studies.

Each microarray study gives rise to its own list of 'interesting' genes. The lists from different studies, however, may not exhibit substantial concordance. Discordant results may produce scientific confusion or disagreement regarding the underlying biology, as well as lost time and misused resources. Consequently, the ability to synthesize information across studies is essential.

Meta-analysis consists of statistical methods for combining results of independent studies addressing related questions. One aim of combining results is to obtain increased power – studies with small sample sizes are less likely to find effects even when they exist. Putting results together increases the effective sample size, thereby allowing more precise effect estimation and increasing power. The uncovering of a significant effect from a combined analysis, where individual studies do not make positive findings at the same significance level, has been referred to in the microarray meta-analysis literature as 'integration-driven discovery' (IDD) [7].

Given the limited size of most microarray studies to date, meta-analysis thus seems a natural approach to the problem of integrating conclusions from different microarray studies. Indeed, there is a recent and increasing literature for meta-analysis of microarray studies [7, 18, 36, 37, 40]. Meta-analysis is not without problems, however.

A major difficulty with synthesizing results is the occurrence of study heterogeneity. Studies which are apparently similar may in fact differ in many ways, some of which may be quite subtle [43]. In general, studies carried out by different research groups may vary in:

- scientific research goals
- population of interest
- design
- quality of implementation
- subject inclusion and exclusion criteria
- baseline status of subjects (even with the same selection criteria)
- treatment dosage and timing
- management of study subjects
- outcome definition or measures
- statistical methods of analysis.

Additional issues more specific to the microarray context include:

- differences in the technology used for the study
- heterogeneity of measured expression from the same probe occurring multiple times on the array
- multiple (different) probes for the same gene

– variability in probes used by different platforms
– differences in quantification of gene expression, even when the same technology is used.

In this chapter, we examine properties of different methods for combining information from what is essentially a replicated experiment carried out with Affymetrix GeneChips. Our aim is to demonstrate that even in this almost ideal situation, several issues concerning appropriate data normalization and combination still arise. We first give some background on the study, then describe the statistical analyses and present results, and conclude with a discussion. Because the focus here is on meta-analysis, we treat only briefly the specifics of microarray data analysis. For further details on these aspects see e.g. Goldstein and Delorenzi [19] for a review or `http://www.nslij-genetics.org/microarray/` for a bibliography of papers on microarray data analysis.

2 Study description

The data were obtained from two experiments on the R6/2 mouse. The R6/2 mouse line is transgenic for exon 1 of the human Huntington's disease (HD) gene, thus serving as an experimental model for the disease [33]. These mice exhibit mRNA changes weeks in advance of neuronal death or gliosis phenotypes [31, 32].

Two separate studies were carried out to investigate the effects on gene expression of different drugs on HD and normal (or wild type (WT)) mice in order to identify genes differentially expressed between HD and WT mice. Each experiment was designed as a 2x2 factorial layout, where one factor is drug/placebo treatment and the other is HD/WT mouse.

We consider only the control groups for the two studies, which received the placebo (injected with normal saline 30-60 minutes prior to sacrifice). In Study I there were 8 control mice, while in Study II there were 6 control mice. In each study, half of the mice were HD and half WT.

The two experiments were carried out by the same laboratory a few months apart. In each experiment, the same protocols were used throughout with regard to mouse breeding, care and sacrifice, mRNA extraction, and hybridization to the microarray. Thus, these data are essentially those of a completely replicated study.

Affymetrix GeneChips contain several (usually 11 – 20) 25-mer oligonucleotides used to measure the abundance of a given target sequence, the perfect match (PM) probes, as well as an equal number of negative controls, the mismatch (MM) probes. The set of probes for a given target sequence is called a probe set. A single fluorescently labeled sample is hybridized to the array which is then scanned with a laser, yielding absolute measures of fluorescence intensity. The intensities are indicative of the amounts of mRNAs containing the target sequence in the sample, and thus provide a means of quantifying levels of gene expression.

The studies were carried out with the Affymetrix MOE 430A (Mouse Expression Array). These chips contain in total 22,690 probe sets, to which, with a slight abuse of

terminology, we refer henceforth as 'genes'. The data are deposited in GEO with series accession number GSE1980, and should be publicly available by the end of 2005 at `http://www.ncbi.nlm.nih.gov/geo/`.

3 Statistical analyses

There are several components of the data analyses to be carried out. First, quality of the hybridizations should be assessed so that low quality chips are removed from further analysis. Before statistical analyses can take place, the primary data obtained from scanning and image analysis of arrays must first be quantified using a measure of gene expression. Once there is a measure of expression of each gene for each individual, we compute for each gene a statistic for assessing genes for differential expression between the HD and WT mice. Some determination of significance should also be made for these statistics, taking into account the multiplicity of hypotheses tested.

We compare analyses carried out under two scenarios: one, where the data are combined and analyzed as a single set; and two, where the two data sets are analyzed separately and their results are combined via meta-analysis. The steps are described in detail below. All analyses reported here were coded in the R statistical programming environment [21, 34], using the following packages from R (2.0.1) and BioConductor (release 1.5) [17]: `affy` [25], `affyPLM` [6], `car` [16], `limma` [39], `qvalue` [12], and `rmeta` [30].

3.1 Chip quality assessment

We assessed all chips for quality with the RMA-QC approach described in Collin [9] and implemented in the BioConductor R package `affyPLM` [6]. In this method, gene expression is modeled as the sum of chip and probe effects, with the model fit by robust regression (i.e. outliers are downweighted; see equation 1 below). Pseudoimages of the robust regression weights or residuals for each probe provide a graphical means to assess chip quality; numerical measures indicative of quality were also computed.

By these criteria, all 14 chips were of similar and suitably high quality that none required exclusion.

3.2 Quantifying gene expression

Several methods of quantifying gene expression from probe fluorescence intensities on Affymetrix GeneChips are in popular use, e.g. MAS5/GCOS [1], the Li-Wong method, implemented in dChip [27], and Robust Multichip Average (RMA) [23], among many others. For a comparison of methods see `http://affycomp.biostat.jhsph.edu/` [11], where it is easily seen that no method is best under every circumstance. We have chosen to use RMA, due to its demonstrated favorable properties [5, 23, 24].

An important yet difficult aspect of gene expression quantification is normalization. (The term 'normalization' as used here is not related to the normal, or Gaussian, distribution.)

The purpose of normalization is remove the effects of systematic variation other than that due to the effect of interest. Examples of such variation include differences in sample preparation, scanning intensities, and variability among chips. Ideally, any observed differences in gene expression remaining after normalization are due to differential expression rather than artifactual differences in measured expression.

RMA consists of three steps: a background adjustment, quantile normalization and probe set summarization. Background is estimated assuming that the observed signal is the convolution of an exponential signal with Gaussian background (noise). Quantile normalization forces equality of quantiles across samples. Such a normalization is appropriate assuming that the true distributions of intensities are the same in all samples (of course, the same probe may occur at different quantiles across samples). For each probe set on the chip, the $\log_2$ background-corrected normalized signal $\log_2 b(PM_{ij})$ is modeled as

$$\log_2 b(PM_{ij}) = \mu_i + \alpha_j + \epsilon_{ij}, \tag{1}$$

where μ_i is the summary measure of expression for the given probe set on chip i, α_j is a probe-specific effect, and ϵ_{ij} are independently and identically distributed mean 0 errors [24]. For parameter identifiability, it is assumed that $\sum_j \alpha_j = 0$. The model is fit via median polish [44]; the estimated chip effect μ_i is the RMA value of the probe set for chip i. RMA values were computed with the `affy` package.

3.3 Identifying differential expression

A commonly addressed problem in microarray experiments is detection of genes differentially expressed under two or more conditions. A substantial number of statistical papers propose methods for this purpose, with new ones still being introduced (for an overview see Goldstein and Delorenzi [19]). The high dimensionality of microarray data has also brought to the fore multiple hypothesis testing issues. The approach we adopt is described here.

3.3.1 Moderated t-statistic

Perhaps the most readily interpretable measure of differential expression is given by the fold change (ratio) in expression of a given gene between two types of samples (HD and WT here). It is more convenient to consider fold change on the logarithmic scale, M = (average) $\log_2$(fold change).

The measure M has the shortcoming of not taking into account differing variability of different genes. The variability of M, though, is not the same across the range of signal intensities. In particular, genes with larger variance across arrays are likely to produce large values of M even when they are not truly differentially expressed between the two sample types.

An obvious way to deal with differing variability is by standardization. Here M is divided by its standard error, which is estimated based on expression measures of the corresponding gene. Thus, the difference in average expression between sample types is quantified with a t-statistic. However, a problem here is that the t-statistic performs very poorly at identifying true differential expression with the small sample sizes found in typical microarray studies.

Bayesian and empirical Bayes methods have been proposed as a compromise between single gene estimates of variability and no estimate of variability at all. These use data from all genes to improve estimation of differential expression for single genes [28, 39]. These methods have been shown to perform well, in terms of true and false positive and negative rates, at identifying differential expression. In addition, the methods have been extended to be applied to a large variety of experimental designs through a linear modeling approach [29, 39].

We follow the linear modeling approach here. For each gene g in a given study, the measured gene expression vector Y_g across samples is modeled as

$$Y_g = X\beta_g + \epsilon_g,$$

where X is the design matrix, β_g is a vector of coefficients, and ϵ_g is a vector of error terms. The design matrix X is the same for all genes within a study.

The moderated t-statistic for coefficient j and gene g is given by

$$\text{mod } t_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g\sqrt{v_{gj}}},$$

where $\hat{\beta}_{gj}$ is the estimate of coefficient j for gene g, $\tilde{s}_g$ is the square root of the empirical Bayes shrinkage estimated variance, and v_{gj} is the scaling for the variance, reflecting sample size. That is, mod t is the ratio of M to its standard error, which has now been estimated taking into account expression levels not only of gene g but of all genes. It is similar to the ordinary t-statistic, but with a moderated standard error estimate and correspondingly an increased number of degrees of freedom. For a detailed explanation, refer to Smyth [39]. We base inference about effects on mod t.

3.3.2 Multiple hypothesis testing

The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of identical mean expression in the two sample types.

A multiplicity problem arises when attempting to assess the statistical significance of the results on tests carried out on several thousands of genes simultaneously. With whole genome coverage arrays consisting of probes for many thousands of genes, most genes will not be differentially expressed between the conditions under investigation. Thus, even a nominal p-value of say 0.01 cannot be characterized as 'significant', since such small p-values will occur by chance when such a large number of tests are made.

Classical approaches to correction for multiple testing focus on control of the family-wise error rate (FWER), or probability of at least one false positive result in all tested hypotheses. The resulting procedures tend to be depressingly conservative though. Recent developments in controlling the false discovery rate (FDR), or expected proportion of false positive findings among the rejected hypotheses, appear to provide a promising way to come

up with meaningful significance measures among thousands of genes [4, 35]. In general, procedures controlling the FDR are typically less conservative than those controlling the FWER. FDR control thus seems well suited for microarray studies.

The (nominal, unadjusted) p-value of a test reflects significance only for a single gene considered in isolation. The q-value of a test measures the proportion of false positives (FDR) incurred among rejected nulls when that test is called significant. It has been described as the expected proportion of false positives among all test results as or more extreme than the one obtained [41, 42].

We make use of q-values to take into account the large number of individual hypotheses tested. The mod t p-values from a set of single gene tests are transformed to q-values with the `qvalue` package [12]. We call a test result 'significant' by fixing a q-value (or FDR) threshold, usually at 0.05. In many microarray studies a higher threshold may be more relevant. For example, a FDR of 0.25 still suggests that three of four significant findings are real. This may be all that is attainable with study sizes available in practice.

3.4 Combined data analysis

In the combined data analysis, we consider all 14 chips as a single data set from the same experiment. This is not a completely artificial treatment, as most large experiments take place over a period of time and include hybridizing groups of chips at different times. RMA measures are obtained by quantifying expression, including normalization, on all chips together.

The linear modeling approach is used to identify genes differentially expressed between HD and WT mice. We consider a series of models, each of which includes an effect on gene expression of HD over WT. There might also be additional variability due to study, so we also allow for a study, or 'batch', effect as well as entertain the possibility of an HD by study interaction.

The design matrices are set up using treatment contrasts so that the effects of interest are included as coefficients in the models. Thus, for each gene g, the three combined data linear models are given by (the subscript g is suppressed):

Model A: $y = \beta_{HD} + \epsilon$
Model B: $y = \beta_{HD} + \beta_{batch} + \epsilon$
Model C: $y = \beta_{HD} + \beta_{batch} + \beta_{HD \times batch} + \epsilon$.

The coefficients are estimated by ordinary least squares. The `limma` package is used to compute for each gene under each model the statistic mod t and corresponding p-values as a prelude to obtaining q-values.

3.5 Meta-analysis

In the meta-analyses, each experiment is first analyzed as a separate study. After heterogeneity analysis, results from the two studies are combined under three meta-analytic techniques:

fixed effects meta-analysis, random effects meta-analysis, and Fisher p-value combination. Computations were done with the R package `rmeta` [30].

In the separate study analyses, gene expression is again quantified with RMA, but values are computed using only chips from the same study (8 chips for Study I or 6 chips for Study II). Linear modeling is carried out as above, but because each study is analyzed individually the model includes only the HD effect (Model A). Fitting the model produces effect estimates (coefficients) for each gene, while the empirical Bayes procedure produces shrinkage estimates of variance, moderated t-statistics and p-values, which in turn yield q-values and gene rankings based on evidence in favor of differential expression between HD and WT mice.

3.5.1 Heterogeneity analysis

Prior to combining effect sizes from different studies, it is important to verify that they are homogeneous – that is, that they all seem to be estimating the same underlying population parameter. Existing graphical methods for assessing inter-study heterogeneity, such as forest plots of individual study confidence intervals, seem of limited usefulness in the microarray setting, as one such plot would be required for each individual gene. We thus depend on numerical assessments to screen genes for heterogeneous treatment effects across studies.

The standard test of homogeneity [8] tests, for each gene g, the null hypothesis of homogeneity of treatment effects β_i in k studies (the subscript g is suppressed)

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k$$

against the general alternative that at least one β_i is different. The test statistic Q is given by

$$Q = \sum_{i=1}^{k} w_i(\hat{\beta}_i - \bar{\beta}_\cdot)^2, \tag{2}$$

where $\hat{\beta}_i$ estimates the treatment effect (the HD coefficient in the linear model for a given gene) in study i, w_i is the weight given to study i (most commonly taken as the reciprocal of the variance of the outcome estimate), and $\bar{\beta}_\cdot$ is the weighted average treatment effect

$$\bar{\beta}_\cdot = \frac{\sum_i w_i \hat{\beta}_i}{\sum_i w_i}. \tag{3}$$

Under the null hypothesis, the distribution of Q is approximately χ^2_{k-1}.

In the event that the null hypothesis is not rejected, any differences between studies are assumed to be due to chance variation, and it is considered appropriate to combine estimates via a fixed effects model. A major limitation of this approach, though, is the low power of the test to detect even substantial heterogeneity due to small sample sizes or a small number of studies. One way to avoid the risk of combining heterogeneous results is to relax the significance criterion from 0.05 to 0.10, say.

If instead the test shows that significant heterogeneity exists between study results, then combination via a random effects model is typically favored. Where possible, heterogeneity should be scrutinized rather than ignored, with an aim toward explaining important study differences [2]. Because our studies were carried out by the same laboratory using identical protocols, tracking down reasons that some genes show heterogeneity across studies while others do not seems an unsolvable problem.

It should also be kept in mind that there is one homogeneity test per gene, so the usual caveats regarding multiple hypothesis testing apply.

3.5.2 Fixed effects meta-analysis

Fixed effect (FE) meta-analysis assumes no heterogeneity between results of the different studies and therefore that a fixed effects model can be used to estimate the assumed common underlying treatment effect. In FE meta-analysis, each individual study estimate receives weight inversely proportional to its variance. The weighted estimates are pooled as above to yield the estimate of the treatment effect given by equation 3, where the weights w_i are inversely proportional to the variances. These weights are used as they minimize the variance of the combined estimate $\bar{\beta}_{\cdot}$ [10]. The variance of the weighted estimator is just $1/\sum_{i=1}^{k} w_i$. Under the assumption of normality of $\bar{\beta}_{\cdot}$, a p-value for each single gene test of HD effect (i.e. differential expression between HD and WT for the given gene) is readily obtained; corresponding q-values are obtained from the set of p-values across all genes.

3.5.3 Random effects meta-analysis

If the study results do exhibit heterogeneity, then there is assumed to be no single underlying value of HD effect but rather a distribution of values. In the presence of heterogeneity, differences among study results are considered to arise from inter-study variation of true effect size as well as chance variation. Use of a FE model understates the true degree of variability of $\bar{\beta}_{\cdot}$, resulting in p-values which are artificially low. A more conservative approach is to use a model which accounts for the additional source of variability due to study.

Random effects (RE) meta-analysis assumes that individual studies may be estimating different treatment effects. The aim is to estimate characteristics of the distribution of effects, particularly the mean population effect size and between study variance of effect sizes. As in the FE case we use weighted estimates, but the weights are adjusted to take into account the additional variability between studies:

$$w_i^* = \frac{1}{(1/w_i) + \hat{\tau}^2},$$

where $\hat{\tau}$ estimates inter-study variability (see Cooper and Hedges [10] for a derivation). The estimated mean treatment effect is given by equation 3, but with w_i^* in place of the w_i. Similarly, the variance of the weighted estimator is now given by $1/\sum_{i=1}^{k} w_i^*$. When the inter-study variance is estimated as 0, the RE model reduces to the FE model.

As for the FE model, single gene p-values from the RE model are obtained assuming normality of the effect distribution; q-values are then computed from the p-values across all genes.

3.5.4 Meta-analysis by Fisher p-value combination

In FE and RE meta-analysis, combined estimates of effect size provide the basis for analysis. Other methods of meta-analysis, dating back to at least the 1930s, are based on combining the p-values from independent studies. Although it is usually preferable to base inference on effect sizes, there are situations for which combining p-values may be considered justified – for example, when only p-values are reported without a corresponding estimate of effect size, or when study characteristics (design, treatment levels) are sufficiently different that combining effect estimates seems unacceptable [20].

Several methods exist for combining p-values. One popular method is due to Fisher [15]. Under the null hypothesis of no treatment effect, the individual study p-values p_i are independent uniformly distributed $U(0,1)$ random variables. Upon rescaling, the Fisher summary test statistic is given by

$$S = -2\sum_{i=1}^{k} \log(p_i). \tag{4}$$

To assess the significance of the Fisher statistic S we need to determine its p-value. The theoretical null distribution of S should be χ^2_{2k} (here $2k = 4$).

We compute p-values for the Fisher combined p-value statistic S in two ways: first, with the χ^2_4 approximation and second, by a resampling procedure proposed in Rhodes et al. [36]. In the resampling procedure, rather than choosing a p-value at random from $U(0,1)$ a p-value is instead chosen at random from each of the sets of p-values from the two studies. These are then combined as in equation 4 into a randomized summary statistic S^R. We obtain an empirical distribution of S^R by repeating the resampling procedure 100,000 times. The p-value for the Fisher S statistic is estimated as the proportion of the resampling-generated statistics S^R_i greater than or equal to the original observed value S. This method yields a more conservative estimate for the p-value of S because the distribution of actual study p-values is not uniform.

4 Results

Here we present detailed results of the analyses outlined above for the combined data set and for meta-analyses of separate experimental study outcomes.

4.1 Combined data

4.1.1 Combined vs. separate gene expression quantification

Because the first step of analysis requires a measure of gene expression, we compared quantification of expression with the combined data (RMA values based on all 14 chips) to individual study quantification (we separately compute RMA values based on the 8 chips from Study I and RMA values from the 6 chips from Study II). Figure 1 contains plots to explore this comparison.

Figures 1(a) and (c) show separate versus combined RMA values for one chip from each study (chip 1 from Study I and chip 1 from Study II, or Chips I-1 and II-1). These chips are representative of all chips in the respective studies – all plots were quite similar within study – so our remarks on the plots apply to all chips. Each gene on the chip is represented by a point in the plot, with the diagonal line representing equal expression by each method.

As it is difficult to detect differences from the line of equality in these scatter plots, we have also plotted the corresponding rotated and rescaled version, the Difference-Average plot [44] (a specific version of this plot is also called an MA plot in the microarray literature; figures 1(b) and (d)). In this representation, the difference between RMA values computed separately and combined is plotted against the average of the two values for the chip. If both RMA values were identical, all points would fall on the horizontal line at 0. Differences are more readily detected in this version of the plot.

It is easily seen that Study I chips tends to have higher RMA values when all chips are combined, while Study II chips have lower RMA values in the combined data set. The tendency persists throughout the range of ($\log_2$) signal intensities.

Many investigators have assumed that normalization of a set of chips together would remove artifacts of this nature. In fact, this does not appear to be the case at all. The persistence of the study batch artifact can be seen, for instance, using cluster analysis [14, 26]. When we cluster samples (chips) based on gene expression, the ones from the same study cluster together. The clustering details (algorithm, dissimilarity measure, number of genes) do not seem to affect the cluster results to any great degree.

Figure 2 shows an example of a dendrogram obtained clustering samples using all genes with Ward's method of clustering and 1 – correlation dissimilarity. The major cluster split occurs between Study I chips and Study II chips. There is a minor, and less clean, split on HD status: Study I samples 1 – 4 and Study II samples 1 – 3 are from the HD mice, the rest are WT mice. Thus, we see that even in what we might expect to be quite homogeneous studies, the most striking difference is in fact purely artifactual and is exactly of the sort that normalization is meant to remove. It is not quite clear why the effects persist, they do not appear to vary systematically with intensity (data not shown). However, quantile normalization of the entire set together does not remove the study batch effect. The batch effect must be removed in another way before reliable inference relating to differential expression can take place.

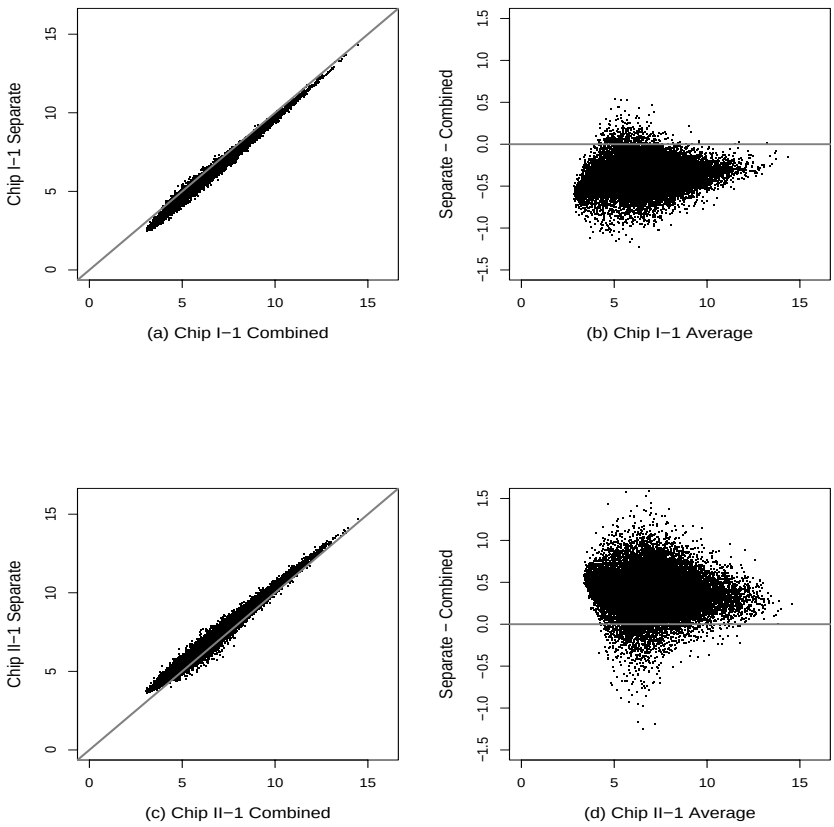

Figure 1: *Comparison of RMA values when studies combined and separate. (a) RMA values for chip I-1 computed within Study I vs. values computed from all chips combined; (b) Difference (Separate – Combined) vs. Average RMA values for chip I-1; (c) RMA values for chip II-1 computed within Study I vs. values computed from all chips combined; (d) Difference vs. Average RMA values for chip II-1. Diagonal and horizontal lines indicate equal values under both methods.*

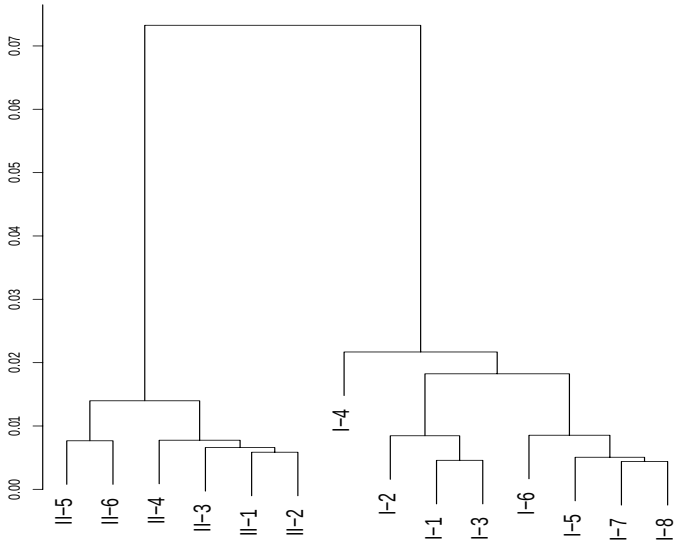

Figure 2: *Cluster dendrogram of combined data RMA values. Samples are clustered using all genes, Ward's method and 1 – correlation dissimilarity.*

4.1.2 HD-study batch interaction

We next turn attention to linear modeling of gene expression in terms of the effects of interest. Here, gene expression is obtained by computing RMA values for the combined set of 14 chips. Although the primary focus is on the HD effect, we must also consider the ramifications of other potential terms for the model. We have just seen the need to include study batch in the model. We now consider Model C to assess the need to include the HD by batch interaction term.

Histograms of p-values and q-values for the estimated interaction effects are shown in figures 3(a) and (b). There are 2242 genes with unadjusted p-values less than 0.05, but only 3 q-values less than 0.10 and thus indication of interaction between HD status and batch for only a few genes. In the face of this mild evidence, we discard Model C and ignore the possibility of interaction in the rest of the analyses.

4.1.3 Detection of differentially expressed genes

On the other hand, we have seen that there is strong evidence of batch effects for many genes. Using Model B to estimate HD and batch effects, we find evidence of significant HD effects for several genes along with a staggering number of genes with strong batch effects (figure 3(b) – (d)). While there are 785 genes with HD q-values < 0.05, nearly one half of the genes (10571 out of 22690) have batch effects with q-values < 0.05. It is not necessary

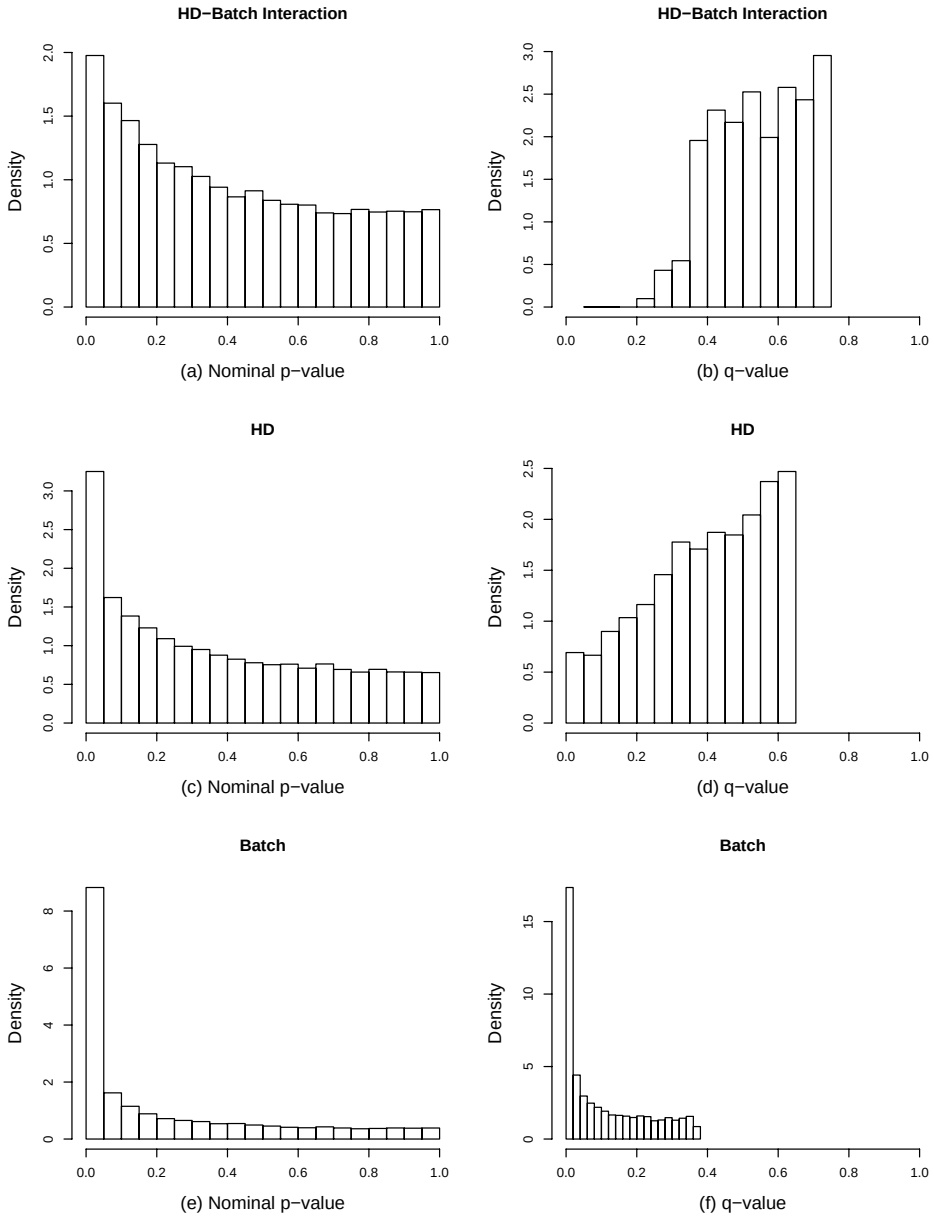

Figure 3: *Histograms of p-values and q-values for Model C interaction term (a, b), and Model B HD (c, d) and study batch (e, f) effects.*

to believe in the exactness of the p- and q-value estimation to conclude that there are many genes with strong batch effects.

We consider Model A, which contains only the HD term, in order to compare genes identified as differentially expressed between HD and WT mice with and without batch effects. Not surprisingly, the significance of the HD effect is always higher (q-value lower) for Model B, where a study batch effect is included in the model. This is because we have controlled for an important source of variability here by introducing an effective stratification factor (batch). There is within stratum homogeneity but heterogeneity between strata, resulting in increased power to detect HD differences.

Figure 4 displays the q-values for individual gene HD effects both with and without batch effects in the model. This plot shows the importance of the batch effect in uncovering differential expression due to HD status. Use of Model B produces an additional 681 significant genes at the same FDR of 0.05 (points in black to the left of the vertical line and above the horizontal line at 0.05).

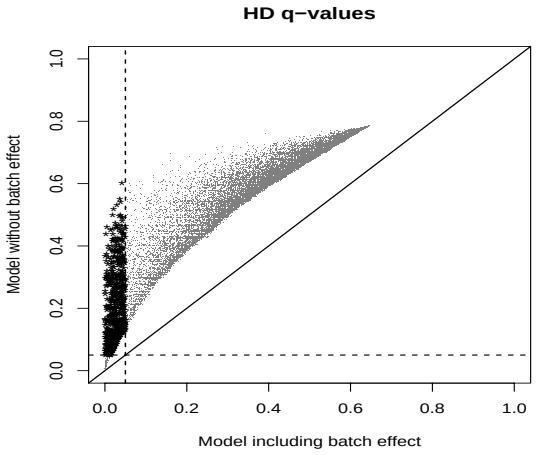

Figure 4: *q-values for HD effects without (Model A) vs. with (Model B) batch effect. Solid diagonal line indicates equal values; dashed vertical and horizontal lines indicate a FDR of 0.05. Highlighted points are genes with significant HD effects for Model B but not for Model A.*

A list of 'interesting' genes, those identified by the analysis as differentially expressed between HD and WT mice, can be produced upon selection of a significance threshold for the HD effect q-value, such as a FDR of 0.05. For comparison with results of meta-analyses of Studies I and II (below), we retain not only the gene list but all of the mod t p-values and corresponding q-values obtained using Model B.

4.2 Meta-analysis of Study I and Study II

Another strategy for dealing with study batch effects is to quantify gene expression separately for each study and then combine the studies by meta-analytic techniques. This is how the problem would necessarily be handled in the case of unrelated studies carried out by different research groups. Here, we are able to examine how meta-analysis would compare with a combined data analysis.

4.2.1 Heterogeneity analysis

We investigate heterogeneity of gene-specific HD estimates from Study I and Study II by computing the statistic Q (equation 2) as well as estimating the inter-study standard deviation (SD) τ for each gene. Characteristics of these are plotted in figure 5.

There is evidence of HD effect heterogeneity for some genes. Several genes have small nominal, unadjusted p-values: 3273 for $p < 0.05$; 5230 for $p < 0.10$ (figure 5(a)). If we choose a more stringent criterion of significance, say a q-value of 0.10, there are still 802 genes with significant heterogeneity. This is substantially larger than the number of genes for which an interaction effect was detected, but also very much smaller than the number with a significant batch effect.

The quantile-quantile plot (figure 5(c)) shows some deviation from the assumed χ^2_1 null distribution. This could be due to inadequacy of the χ^2_1 approximation, but as it is unlikely that all the nulls are in fact true we instead interpret this as indicative of the presence of genes for which the alternative holds, i.e. there is some true heterogeneity. Since the χ^2 test has low power to detect heterogeneity for the small study number and sample sizes that we have, there is likely to be a greater degree of heterogeneity, and for more genes, than suggested here.

The distribution of estimated inter-study SD $\hat{\tau}$ is highly skewed. Over 70% (16428) of genes have $\hat{\tau} < 0.01$ (figure 5(d)). The value of $\hat{\tau}$ corresponding to a FDR of 0.10 is about 0.066. This gives some idea of what a 'large' value of τ is in this context. An example of intensities for a gene displaying heterogeneity is provided in Table 1, which gives the individual study RMA values of each chip for a gene with $\hat{\tau} \approx 0.1$.

Table 1: *Individual Chip RMA Values for a Gene with* $\hat{\tau} \approx 0.1$

	Chip number								Summary	
	1	2	3	4	5	6	7	8	Mean	SD
Study I	8.67	8.88	8.91	8.64	9.08	9.02	9.02	9.27	8.94	0.21
Study II	9.86	9.75	9.96	9.83	9.57	9.73			9.78	0.13

One purpose of testing homogeneity is for deciding between the FE and RE model for combining effect estimates. There will not be a large difference between FE and RE meta-analysis for genes with small τ. A general recommendation which has been made is to carry out both then compare similarity of results. If the results are similar then there is unlikely

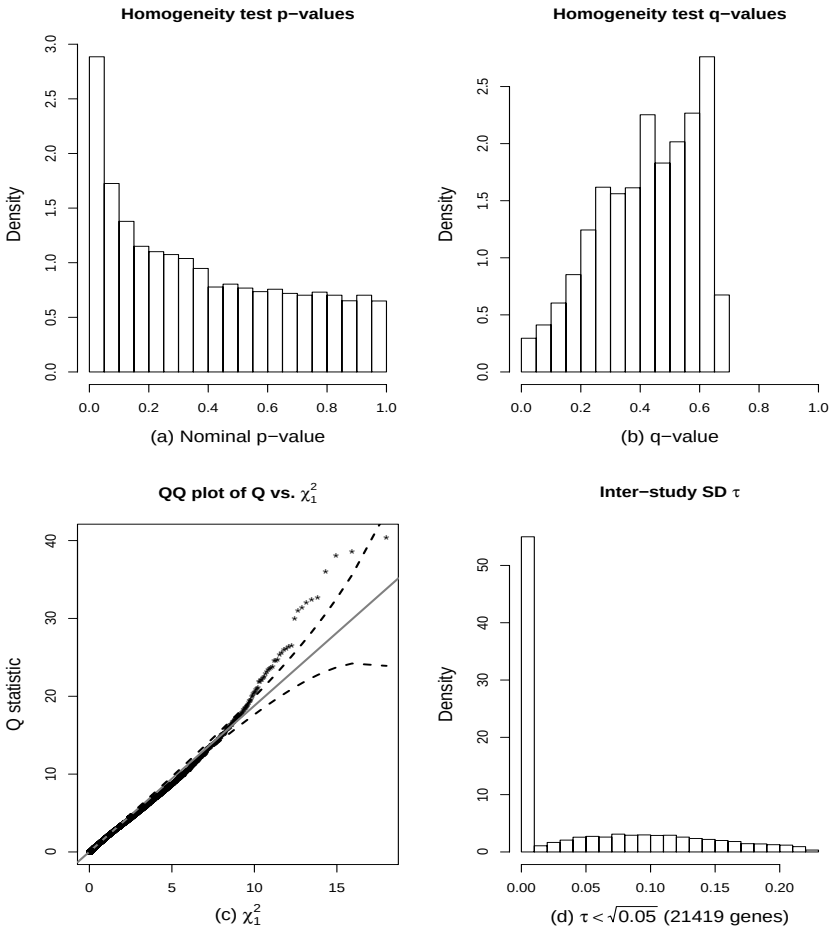

Figure 5: *Plots for heterogeneity analysis: histograms of (a) p-values and (b) q-values for the homogeneity statistic Q; (c) quantile-quantile plot of Q compared to the theoretical χ^2_1 distribution with 95% confidence region (dashed lines); (d) histogram of gene-specific estimated inter-study SD for the 21419 genes for which $\hat{\tau} < \sqrt{.05}$.*

to be important heterogeneity and the FE model would typically be reported. If results are different, it is usually considered preferable to use RE meta-analysis to estimate the mean and SD of the effect size distribution. In the case of extreme, unexplained heterogeneity, it is probably more suitable to avoid combining the study results at all.

4.2.2 Fixed effects and random effects meta-analysis

Based on the results of the heterogeneity analyses, we would choose to adopt the RE model for combining HD effect size estimates. Nevertheless, we investigate both approaches here in order to compare them.

Figure 6(a) compares the combined HD (mean) effect estimated by the RE model versus the FE model combined estimate. With the exception of a few genes, these combined estimates tend to be remarkably similar.

Due to the additional variability included by the RE model, however, there is a great deal of difference between the standardized estimates (figure 6(b)). Here, we can see that the FE standardized estimates are stochastically larger than the RE ones: about half of the genes have identical results for FE and RE, but for only 949 genes (or 4%) is the RE standardized estimate larger than that of FE. This phenomenon is also clearly reflected by the distributions of the corresponding q-values (figures 6(c) and (d)) – at any FDR, many more genes are called differentially expressed between HD and WT by the FE model.

Figure 7 shows how the methods compare for compare for different degrees of heterogeneity. In figures 7(a), (b) and (c), q-values are transformed by $-\log_{10}$ so that larger values are more significant. We see that significant effects in the RE meta-analysis tend to be for more genes with more homogeneous effects across studies (figures 7(a), (b)); that is, genes for which the inter-study variability does not overwhelm the size of the estimated effect. The q-values from FE and RE are compared directly in figure 7(c), where it is seen that the FE combined HD effect estimate is more significant than that of RE for virtually all genes. Finally, the location (centered at 0) and flatness of the loess curvefigure 7(d) show that heterogeneity is not more frequently found for larger estimated mean HD effect sizes.

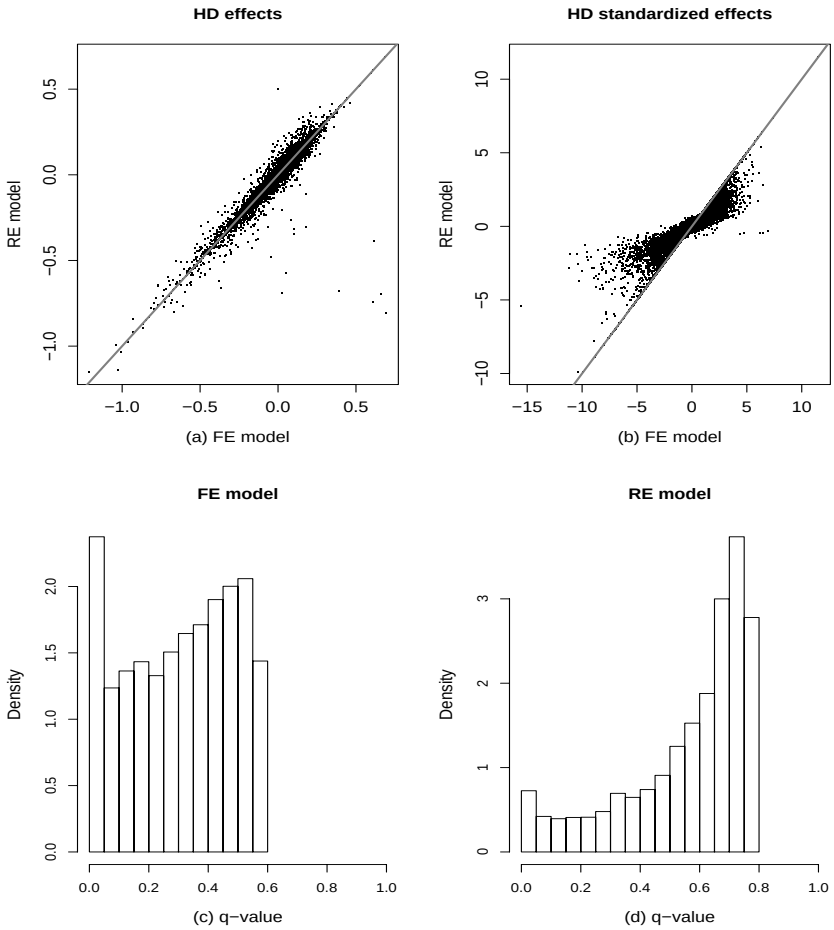

Figure 6: *Comparison of HD effects for FE and RE meta-analysis. (a) HD effects estimated by RE vs. FE; (b) HD standardized effects estimated by RE vs. FE; q-values for HD (standardized) effects estimated by FE (c) and RE (d). Diagonal lines indicate equal values under both methods.*

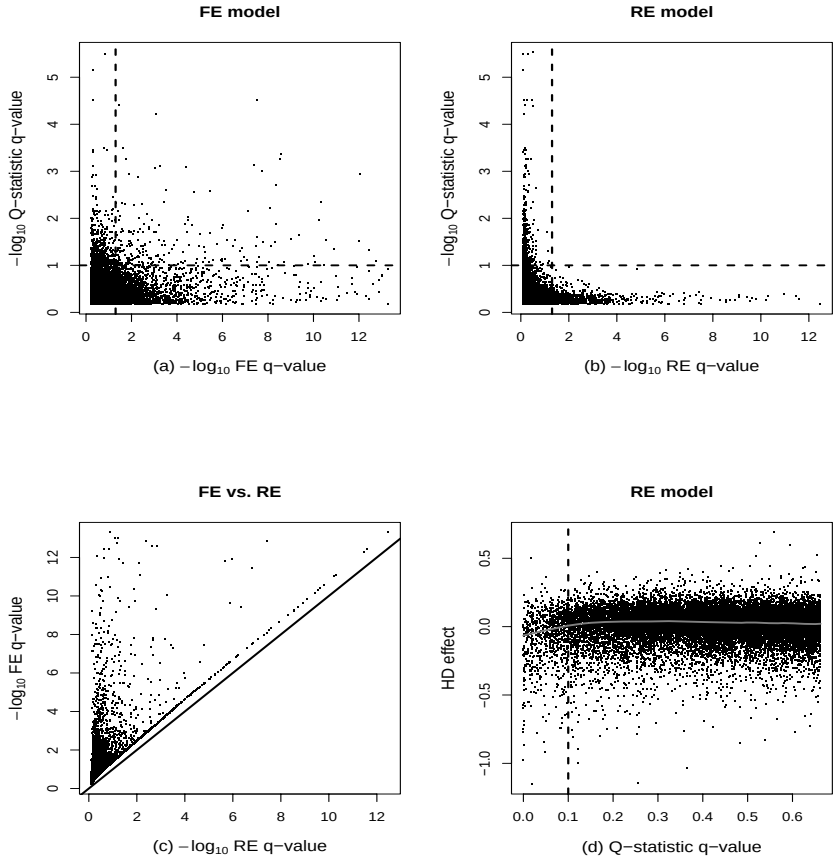

Figure 7: *Characteristics of FE and RE inference for varying heterogeneity.* $-\log_{10}$ *q-value of homogeneity Q-statistic vs.* $-\log_{10}$ *q-value of combined HD effect estimate from FE (a) and RE (b), vertical line indicates HD effect FDR of .05, horizontal line indicates Q-statistic FDR of .10; (c)* $-\log_{10}$ *q-values for FE vs. RE, diagonal line indicates equal values; (d) RE model estimated mean HD effect size vs. q-value of Q-statistic. vertical line indicates Q-statistic FDR of .1, horizontal smooth line is a loess curve.*

4.2.3 Fisher p-value meta-analysis

Figure 8 displays results obtained by combining for each gene the HD mod t p-values from Study I and Study II. The distribution of q-values obtained from p-values derived from the χ^2_4 distribution is compressed downward toward significance (a). Resampling p-values are exceedingly conservative compared to χ^2_4-derived p-values (b). Compared to RE model p-values the χ^2_4 p-values are liberal (c), while the resampling p-values are again conservative, although somewhat less than in comparison to the χ^2_4 p-values (d).

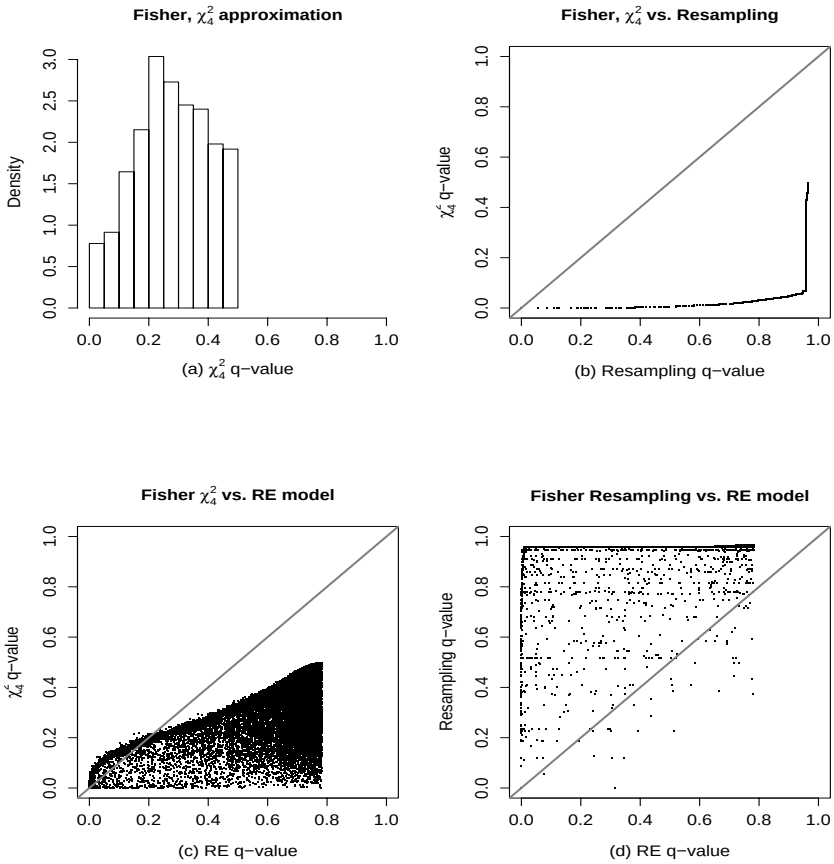

Figure 8: *Comparison of Fisher S statistic q-values. (a) Histogram of χ^2_4 q-values; scatter plot of q-values obtained by χ^2_4 vs. resampling (b), χ^2_4 vs. RE (c), resampling vs. RE (d). Diagonal lines indicate equal values under both methods.*

The lack of agreement indicates the degree to which inference depends on the specific method chosen. Even where the same statistic is used, there is a real problem determining its p-value – the χ^2_4 assumption results in very different p-values from the resampling-based ones. Caution must therefore be exercised in choosing a method and interpreting results.

4.2.4 Comparison of results stratified by heterogeneity status

The findings presented thus far consider the entire set of genes in aggregate. However, the set of all genes can be viewed as a mixture of two types: genes for which the HD effects are homogeneous and genes for which the effects are heterogeneous across studies. It is therefore worth looking at characteristics of the analyses when genes are stratified by heterogeneity status.

Defining heterogeneity status requires a criterion for significance. In the microarray context, one must consider its impact on the subsequent identification of differential expression. To be more conservative in calling a gene differentially expressed, a fairly liberal heterogeneity criterion would seem in order. Taking into consideration the outcome of the heterogeneity analysis above, we decided on a FDR cut-off of 0.10 for Q. For this threshold, the number of genes for which studies are heterogeneous is 802; there are thus 21888 homogeneous ones.

Table 2 gives the proportions of genes with significant HD effects for the four meta-analysis methods, as well as the combined data, for all genes together and also stratified by heterogeneity status (Hom. or Het.) at varying FDR for the HD effect. The methods are: C = combined data, FE = fixed effects model, RE = random effects model, FX = Fisher p-value combination method, χ^2 p-values, and FR = Fisher p-value combination method, resampling method. The Fisher resampling method proportions are extremely low, so the numbers of genes are also reported. For RE, proportions given as zero are actual zeros.

Table 2: *Significance Proportions for Meta-analysis Methods*

	Sig. at FDR = .10			Sig. at FDR = .05			Sig. at FDR = .01		
Method	All	Hom.	Het.	All	Hom.	Het.	All	Hom.	Het.
C	0.07	0.06	0.19	0.03	0.03	0.12	0.01	0.01	0.05
FE	0.18	0.17	0.38	0.12	0.11	0.30	0.06	0.05	0.21
RE	0.06	0.06	0.01	0.04	0.04	0.00	0.02	0.02	0.00
FX	0.08	0.06	0.70	0.04	0.03	0.29	0.01	0.01	0.10
FR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FR Number	(5)	(3)	(2)	(3)	(2)	(1)	(3)	(2)	(1)

FE finds the most significant effects, followed by Fisher χ^2 (FX). These two methods as well as the combined data method also find drastically higher rates of significant effects for studies which are heterogeneous, pointing to the need for caution when combining information. In contrast, RE finds lower rates of significant effects under heterogeneity.

4.2.5 Pairwise agreement of meta-analysis results

Lastly, we look at agreement for pairs of methods stratified by heterogeneity status, varying the FDR for calling a gene differentially expressed between HD and WT (figure 9). The simple agreement rate is just the proportion of genes for which both methods agree on

whether or not the HD effect is significant. The correspondence between plotting symbol and comparison pair is given in table 3.

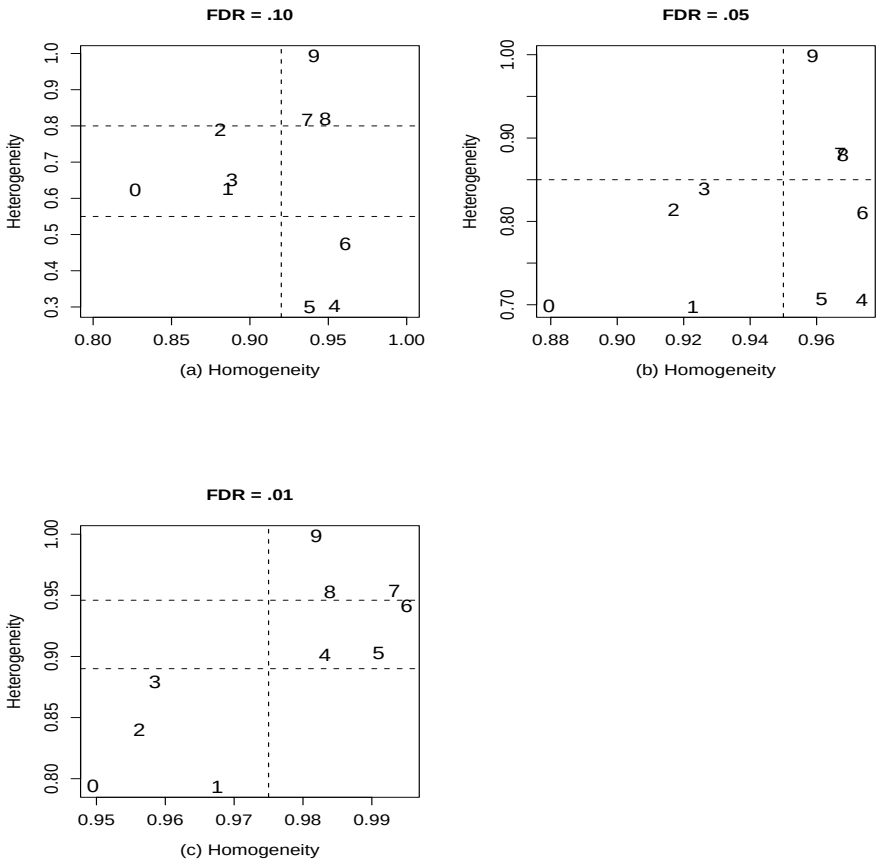

Figure 9: *Simple agreement rates of pairs of methods for varying FDR. Agreement rate under heterogeneity vs. agreement rate under homogeneity for (a) FDR = 0.10, (b) FDR =0.05 and (c) FDR = 0.01. Dashed vertical and horizontal lines separate groups of pairs.*

Table 3: *Correspondence of Plotting Symbol and Pair*

Symbol	0	1	2	3	4	5	6	7	8	9
Pair	FE FR	FE RE	FE C	FE FX	FX RE	FX FR	FX C	C FR	C RE	RE FR

The pairs appear to fall roughly into three groups when homogeneity and heterogeneity rates are considered jointly. Pairs 0 – 3 form one group. This group consists of all pairwise comparisons with FE. These pairs have the lowest agreement under homogeneity. Pairs 4,

5 and 6 have high homogeneity agreement and lower, but increasing with decreasing FDR, heterogeneity agreement. FX appears in each of these pairs. Finally, pairs 7, 8 and 9 have highest agreement under both homogeneity and heterogeneity. These are all pairs are formed from C, RE, FR.

5 Discussion

Pooling raw data from different studies for analysis is not always possible; even when it is possible it might not be recommended (e.g. to avoid Simpson's paradox [38]). However, in the simple setup we have described here, where a single lab has carried out the same experiment twice, one would think that combining the raw data should be a fundamentally sound approach. In particular, carrying out the normalization step on the aggregated data would seem not only desirable but also necessary.

We have illustrated, however, that even in such an uncomplicated scenario, without issues of different platforms or experimental designs and protocols, integrating the available information might not be completely straightforward. We have seen persistent batch effects that must be taken into account. We would recommend that new methods developed for more complex situations also be tested in simpler cases so that the properties of the methods may be better understood.

The importance of considering variability across different labs has been noted in the literature [22]; our work here suggests that within lab variability may also need to be considered.

Our results also have substantial implications for large single studies, where patients are recruited over time and arrays are not all hybridized at the same time. Avoidance of problems before they arise calls for careful study design in advance. In addition, comprehensive exploratory data analyses are required once data are collected, to identify and adjust for sources of variability which could obscure the underlying biology. The presence of strong batch effects may be indicative of aspects of laboratory practice in need of improvement. Analysis of data for batch effects can reveal such problems and could therefore help in their rectification.

In this work we can compare results from different methods of analysis, but we are unable to rigorously assess method performance or robustness because it is not feasible at present to determine the truth of the findings. To date we can identify as true positives only a subset of genes that are likely to be differentially expressed in R6/2 mice [31, 32], and we also do not yet know which identified genes are not (false positives), or which genes missed are in fact differentially expressed (false negatives). It is advisable to build some truth into the experiment where feasible, for example by using spike-in controls (specific RNAs added to the sample in known quantities). Nevertheless, we hope that this survey of single methods and method agreement can provide some guidance to investigators in selecting appropriate procedures.

In a similar investigation, [40] use a model to generate a known truth and simulate data from their model to examine properties of meta-analysis of microarray studies. Although this approach may provide some useful broad guidelines, further empirical evidence is required

to gain more refined insight into the sources and magnitudes of variability and their effects on properties of meta-analysis of microarray studies.

In most studies, researchers have the resources for further investigation into only a few of the findings. A typical validation study consists of following up on a few genes, often in the range of $5 - 20$. The research community would benefit from larger scale follow-up studies to enable the properties of different methodologies for synthesis to be judged more critically.

Although a number of intriguing methods have been introduced for meta-analysis of microarray data, the literature in this field is not yet fully developed. Clearly there is a need for further empirical and theoretical research in this challenging area. Large scale validation studies would provide a welcome opportunity to advance both biological and methodological knowledge.

6 Acknowledgements

This work was supported by funds from the Swiss National Science Foundation to the National Centers for Competence in Research (NCCR) in Plant Survival (DRG) and Molecular Oncology (MD, TS), and also by the US National Institutes of Health (RL-C). The authors also acknowledge R. J. Ferrante and J. K. Kubilus for the samples used to generate the HD data set for this analysis.

References

[1] Affymetrix. *Affymetrix: Microarray Suite User's Guide, version 5.0*. Santa Clara, CA, 2001.

[2] K. R. Bailey. Inter-study differences: how should they influence the interpretation and analysis of results? *Statistics in Medicine*, 6:351–358, 1987.

[3] Tanya Barrett, Tugba O. Suzek, Dennis B. Troup, Stephen E. Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E. Lash, Wataru Fujibuchi, and Ron Edgar. NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Research*, 33:D562–D566, 2005.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57: 289–300, 1995.

[5] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19:185–193, 2003.

[6] Ben Bolstad. *affyPLM: affyPLM - Probe Level Models*, 2004. URL `http://www.stat.berkeley.edu/users/bolstad/AffyExtensions`. R package version 1.2.5.

[7] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:i84–i90, 2003.

[8] W. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10:101–129, 1954.

[9] Francois Collin. *Analysis of oligonucleotide data with a view to data quality assessment*. Ph.D. thesis, Department of Statistics, University of California, Berkeley, 2004.

[10] H. M. Cooper and L. V. Hedges. *The Handbook of Research Synthesis*. Russell Sage Foundation, 1994.

[11] Leslie M. Cope, Rafael A. Irizarry, Harris A. Jaffee, Zhijin Wu, and Terence P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20:323–331, 2004.

[12] Alan Dabney and John D. Storey with assistance from Gregory R. Warnes. *qvalue: Q-value estimation for false discovery rate control*. R package version 1.1.

[13] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30: 207–210, 2002.

[14] B. S. Everitt, L. Landau, and M. Leese. *Cluster Analysis*. Oxford University Press, 4th edition, 2001.

[15] R. A. Fisher. *Statistical Methods for Research Workers*. Oxford University Press, 4th edition, 1932.

[16] John Fox. *car: Companion to Applied Regression*, 2005. URL `http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/`. R package version 1.0-17.

[17] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. BioConductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL `http://genomebiology.com/2004/5/10/R80`.

[18] D. Ghosh, T. R. Barette, D. Rhodes, and A. M. Chinnaiyan. Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional and Integrative Genomics*, 3:180–188, 2003.

[19] Darlene R. Goldstein and Mauro Delorenzi. Statistical design and data analysis for microarray experiments. In Alvin Berger and Matthew A. Roberts, editors, *Unravelling Lipid Metabolism with Microarrays*. Dekker, New York, 2004.

[20] V. Hasselblad. Meta-analysis of environmental health data. *Science of the Total Environment*, 160–161:545–558, 1995.

[21] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

[22] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. Garcia, J. Geoghegan, G. Germinoand C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2:345–50, 2005.

[23] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31:e15, 2003.

[24] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Y. D. Beazer-Barclay, K. J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[25] Rafael A. Irizarry, Laurent Gautier, Benjamin Milo Bolstad, Crispin Miller with contributions from Magnus Astrand, Leslie M. Cope, Robert Gentleman, Jeff Gentry, Wolfgang Huber, James MacDonald, Benjamin I. P. Rubinstein, Christopher Workman, and John Zhang. *affy: Methods for Affymetrix Oligonucleotide Arrays*, 2004. R package version 1.5.8.

[26] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[27] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences USA*, 98:31–36, 2001.

[28] I. Lönnstedt and T. P. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

[29] I. Lönnstedt, S. Grant, G. Begley, and T. P. Speed. 'Microarray analysis of two interacting treatments: a linear model'. Technical report, Uppsala University, Sweden, Department of Mathematics, 2001.

[30] Thomas Lumley. *rmeta: Meta-analysis*, 2004. R package version 2.12.

[31] Ruth Luthi-Carter, Andrew Strand, Nikki L. Peters, Steven M. Solano, Zane R. Hollingsworth, Anil S. Menon, Ariel S. Frey, Boris S. Spektor, Ellen B. Penney, Gabriele Schilling, Christopher A. Ross, David R. Borchelt, Stephen J. Tapscott, Anne B. Young, Jang-Ho J. Cha, and James M. Olson. Decreased expression of striatal signaling genes in a mouse model of Huntington´s disease. *Human Molecular Genetics*, 9:1259–1271, 2000.

[32] Ruth Luthi-Carter, Sarah A. Hanson, Andrew D. Strand, Donald A. Bergstrom, Wanjoo Chun, Nikki L. Peters, Annette M. Woods, Edmond Y. Chan, Charles Kooperberg, Dimitri Krainc, Anne B. Young, Stephen J. Tapscott, and James M. Olson. Dysregulation of gene expression in the R6/2 model of polyglutamine disease: parallel changes in muscle and brain. *Human Molecular Genetics*, 11:1911–1926, 2002.

[33] L. Mangiarini, K. Sathasivam, M. Seller, B. Cozens, A. Harper, C. Hetherington, M. Lawton, Y. Trottier, H. Lehrach, S. W. Davies, and G. P. Bates. Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, 87:493–506, 1996.

[34] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

[35] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19:368–375, 2003.

[36] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62:4427–4433, 2002.

[37] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences USA*, 101:9309–9314, 2004.

[38] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241, 1951.

[39] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3, 2004.

[40] John R. Stevens and Rebecca W. Doerge. Combining Affymetrix microarray results. *BMC Bioinformatics*, 6:57, 2005.

[41] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–498, 2002.

[42] John D. Storey and Robert Tibshirani. Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences USA*, 100:9440–9445, 2003.

[43] Alexander J. Sutton, Keith R. Abrams, David R. Jones, Trevor A. Sheldon, and Fujian Song. *Methods for Meta-Analysis in Medical Research*. John Wiley & Sons, 2000.

[44] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.