On the Synthesis of Microarray Experiments

R. Gentleman
Division of Public Health Sciences,
Fred Hutchinson Cancer Research Center,

M. Ruschhaupt
Division of Molecular Genome Analysis,
German Cancer Research Center,
Heidelberg, Germany
and Department of Medical Informatics, Biometrics and Epidemiology (IBE),
Ludwig-Maximilians-University, Munich, Germany

W. Huber
Group Leader, European Bioinformatics Institute,
European Molecular Biology Laboratory
Cambridge CB10 1SD, England

September 30, 2005

1 Introduction

DNA microarray technology takes advantage of hybridization properties of nucleic acid and uses complementary molecules attached to a solid surface, referred to as *probes*, to measure the quantity of specific nucleic acid transcripts of interest that are present in a sample, referred to as the *target*. Microarrays provide a rich source of data on the molecular working of cells. Each microarray reports on the abundance of tens of thousands of mRNAs. Virtually every human disease is being studied using microarrays with the hope of finding the molecular mechanisms of disease. There are a number of different platforms available, some from commercial vendors and others essentially home made. The efficacy of the assay, as well as the effects of non-specific signal and crosstalk, differ across experiments and technologies. In practice the raw intensity data are heavily manipulated before one obtains the expression values that most statisticians, biologists and clinicians use in their research.

With many different investigators studying the same disease and with a strong commitment to publish supporting data in the scientific community, there are often many different datasets available for any given disease. Hence there is substantial interest in finding methods for combining these datasets to provide better and more detailed analysis of all available data. In this paper we review some of the methods that have been proposed, and explore these and other alternative methods for combining the data with a view to obtaining more precise information about changes in gene expression that relate to different disease phenotypes.

Choi et al. (2003) proposed the use of meta-analytic tools and argued in favor of the synthesis of experiments on the basis of estimated effects. While we agree in principle with that approach, we note that there is, in fact, a more general approach that should be considered. A succinct discussion is given in Cox and Solomon (2003), in particular in Chapter 4. We note that standard, and general statistical models can be employed to address many of the questions that arise, and we provide some translations between these two references in the remainder of this paper. In addition we present outputs from both approaches and compare them on two examples.

The usual application of meta-analysis is to analyze a single outcome, or finding, using published data where typically only summary statistics are available. With microarray experiments, we are often in the more fortuitous situation of having the complete set of primary data available, not just the summary statistics. There are usually thousands of genes that were measured in each experiment and it is unlikely that all will be implicated in the disease process, so some reduction is needed. By phrasing the synthesis in terms of standard statistical models many of the recently developed p-value adjustment methods for multiple comparisons can be applied directly. Finally, we emphasize that even when synthesizing studies, it will be important to ask specific direct questions; undirected searches are unlikely to be enlightening.

We first consider exploratory methods with the intention of ascertaining whether, in broad terms, the experiments are similar enough to warrant combining. Once that question has been answered we next turn our attention to the development of a suitable model. One of the fundamental requirements for combining data is that all of the studies have been carried out in such a way that the treatment effects of interest were measured in all studies. One of the major accomplishments of meta-analysis was the realization that the scale need not be the same in all studies, but rather that some transformation of all effects to a similar scale was sufficient. We consider such an analysis both in the classical meta-analytic framework and in the form of linear mixed effects models (Pinheiro and Bates, 2000). We follow some of the development of Cox and Solomon (2003), who discuss this application in a similar context.

When combining experiments there are generally two problems that must be dealt with. One is the matching problem, and the second is the problem of how to combine the estimated effects in each experiment into a reasonable overall estimate. The matching problem comes in two parts, first the matching of probes and second the matching of samples and experiments. The first matching problem exists whenever a different set of probes is used in the experiments. The intention to assay the same mRNA in all experiments does not always eventuate, due to mismatching sequences, complex gene structure, or other problems. Further, the measured intensity at any probe on an array is a function not just of the target abundance, but also of the sequence that was used for the probe and the technology used for the assays. Hence, two arrays that use different probe sequences may not be directly comparable. Parmigiani et al. (2004) have proposed methodology to deal with the gene matching problem and have provided software in the form of an R package MergeMaid. In this paper we do not address this question and simply resort to matching genes on the basis of their GenBank and Unigene identifiers. A more complete analysis would consider these and other issues and would almost surely focus more directly on ensuring that the probes used in the different experiments were measuring in fact the same thing. However, that is a separate and substantial area of research, and in a sense is orthogonal to the problem of synthesizing experiments – which is our goal here.

The second matching problem must be addressed when selecting samples or experimental conditions to be used in the analysis. Some of the issues relate to whether similar quantities can be estimated from the experiments and we discuss that problem in more detail subsequently. In one of our examples we consider the synthesis of two experiments with the goal of assessing whether we can detect a gene expression signature that is associated with the presence of lymph nodes that carry metastases, for two quite different primary tumor tissues. However, we cannot easily tell if the definitions of lymph node positivity used in the two experiments were sufficiently similar to warrant integration of the analyses. This is not an isolated problem: different investigators use the same term for different conditions or different terms for the same condition. Again, we confine ourselves to a cursory discussion of the issues since our main focus is on the statistical models that may be used to combine the estimated effects.

It is also important that some assessment of the additional information that has been gained by combining the data be made. One measure is the number of features that have a significant treatment effect in the combined analysis that did not have a significant effect in either experiment alone. These are the new discoveries, or as Choi et al. (2003) called them, "the integration-driven discoveries".

We have provided all data sets and software used to carry out the analyses reported here in the form of an R package (or compendium) called GeneMetaEx that is available from the Bioconductor project. Readers are able to ascertain the exact details of every computation reported and of every figure produced in this paper. Further, they can extend and explore our analysis according to their own interests.

2 Materials and Methods

2.1 The experimental data

In our first example we combine two data sets that both report on the estrogen receptor (ER) status of women with breast cancer. The goal of that analysis is to find those genes which show differential expression between ER positive and ER negative tumors. As we shall see the signal is quite strong, and rather remarkably ER status effects the expression of a very large number of mRNA species. However, some thought should be paid to the fact that the two samples could be quite different, could represent different populations and hence observed differences may simply reflect the sampling heterogeneity.

In the second example we compare patients for which positive lymph nodes were detected to patients where no lymph node involvement was detected. One data set is the same as was used for the breast cancer example described above, but we use a different covariate. The second experiment has patients with head and neck cancer. In this case we are interested in finding common signatures of lymph node metastasis, across tissues. The differences in the patient populations are quite large. Further, since different tissues are involved in the lymph node comparison it is unlikely that the same set of genes is expressed, let alone differentially expressed. Even though thousands of genes were measured we anticipate that relatively few will be involved in a –at this point hypothetical– common molecular basis for lymph node metastasis. Hence some form of reduction in the genes analyzed should be considered.

We now introduce the three data sets that we will use for our examples. One is a study of breast cancer reported by West et al. (2001) in which 46 patients were assayed and two phenotypic conditions were made public, the estrogen receptor (ER) status and the lymph node (LN) status. We will refer to this as the Nevins data in the remainder of the text. The samples were arrayed on Affymetrix HuGeneFL GeneChips. ER status was determined by immunohistochemistry and later by a protein immunoblotting assay. We have used 46 samples, of which 4 gave conflicting evidence of ER status depending on the test used. Lymph node status was determined at the time of diagnosis. Tumors were reported as negative when no positive lymph nodes were discovered and as positive when there were at least three identifiably positive lymph nodes detected.

A second breast cancer data set was made public by van't Veer et al. (2002) in which tumors from 116 patients were assayed on Hu25K long oligomer arrays. Among other covariates the authors published the ER status of the tumors. Their criterion was a negative immunohistochemistry staining, a sample was deemed negative if fewer than 10% of the nuclei showed staining and positive otherwise. We refer to this as the van't Veer data in the remainder of the paper. Our primary example involves combining these two data sets to obtain a better view of the relationship between genes and ER status.

The third experiment we consider is one published by Roepman et al. (2005), which assayed patients with primary head and neck squamous cell carcinoma using long oligomer arrays. Lymph node status of the individuals involved was determined by clinical examination followed by computed tomography and/or magnetic resonance imaging. Any nodes that were suspected of having metastatic involvement were aspirated and a patient was classified as lymph node positive if the aspirate yielded any metastatic tumor cells. For our second comparison, we combine these data, which we call the Holstege data, with the Nevins data on the basis of LN status.

Some of the issues that arise in combining experiments can already be seen. For the comparison on the basis of ER status we see that the two used similar, but different methods for assessing ER status. One might want to revert the Nevins data to the classifications based only on immunohistochemistry staining to increase comparability across the two experiments. This is likely to come at a loss of sensitivity since one presumes that the ultimate (and in four cases different) classification of samples was the correct one.

For the synthesis of experiments on the basis of lymph node status the situation is even more problematic. One might wonder whether approximately the same effort was expended in determining lymph node status in the two experiments. We emphasize that value of any synthesis of experiments will have a substantial dependency on the comparability of the patient classifications. If the classifications of samples across experiments are quite different then it is unlikely that the outputs will be scientifically relevant.

As noted previously probes were matched on the basis of GenBank or UniGene identifiers. For the Nevins – van't Veer synthesis we have 3988 mRNA targets in common, while for the Nevins – Holstege synthesis there are 3786 common mRNA targets.

2.2 Issues and rationale for combining experiments

We remind the reader that one of the most important principles of meta-analysis is that the different data sets should be chosen in such a way that the quantity of interest is comparable across studies. It is not reasonable to assume that all genes on the microarray will be affected in the same way, or even that most will be affected at all. Hence, many of the per gene models are not going to reveal any new information. The hope is that there are a number of important genes whose effects are not obvious from any individual experiment, but when the experiments are analyzed as a whole the effect becomes obvious and can be extracted from the analysis.

Many of the issues were first raised in Glass (1976) who noted that in situations where potentially different scales of measurement have been used it will be necessary to estimate an index of effect magnitude that does not depend on the scaling or units of the variable used. For two-sample problems the scale-free index that is commonly used is the so-called *effect size*, which is the difference in means divided by the pooled estimate of standard deviation (note that this is not the t-statistic, which would use the standard error of the mean difference). Other measures include the correlation coefficient and the log odds ratio, but we do not consider them here.

Hence, a determination of whether the quantities being measured in all experiments are sufficiently similar is needed. We will demonstrate that under fairly general conditions this seems to be true. In fact, methods based on effect size only require that the scales of measurement be linearly related in some manner, while we will show that often it is the same quantity that is being measured; which is a somewhat stronger property.

In two sample experiments it will often be sensible to compare fold-change, or its logarithm, between phenotypes of interest. In this case both two-color arrays (such as spotted cDNA arrays) and single channel arrays (such as those from Affymetrix) may provide estimates on the same scale.

We describe one model for changes in gene expression due to a treatment effect and demonstrate that, under this model, estimated treatment effects between one-color and two-color experiments can be combined. This observation does not preclude the existence of other models which may or may not validate the synthesis of experiments, but rather it simply provides a scenario under which such an analysis is reasonable.

We consider the comparison of two experiments. One experiment is carried out using a one color technology, such as Affymetrix GeneChips and the other experiment was carried out using a two-color technology such as cDNA arrays. For the two-color experiment we presume that a common reference was used for all hybridizations. If this is not the case, then another reasonable scenario is for each chip to represent a pair of matched samples, one treated and one not; or a dye-swap can be accommodated. In our example, we presume that the same set of samples have been hybridized to each of the two types of arrays.

Let $X_{m,i}$ represent the per cell level of abundance for mRNA target m in a sample of interest labeled i. Further we presume a common control C with levels of mRNA abundance given by C_m. We typically drop the m subscript when only one target is being considered. We let $X_i^* = \log_2(X_i)$, $C^* = \log_2(C)$, and also use $X_i^c = \log_2(X_i/C)$. Then let

$$\begin{aligned} \overline{X}_I^* &= \frac{1}{|I|}\sum_{i\in I} X_i^* \\ \overline{X}_I^c &= \frac{1}{|I|}\sum_{i\in I} X_i^c \end{aligned}$$

be the average logarithmic abundances and average log-ratios in a group of samples I, and we see that for the comparison between two different sets of samples I and J,

$$\overline{X}_I^c - \overline{X}_J^c = \overline{X}_I^* - \overline{Y}_J^*, \tag{1}$$

since the terms involving $\log_2(C)$ cancel out. So, the logarithm of the geometric means between the two groups is the same quantity for one-color arrays as for two-color arrays that use a common reference.

Unfortunately, the simple picture of Equation (1) becomes more complicated when we consider the *estimates* of the log-ratios and the logarithmic abundances, rather than these quantities themselves. This is called the problem of attenuation, and in particular, differential attenuation between different experiments complicates the comparison of the estimates between the experiments. The problem may be understood as follows. Using notation analogous to above, let $Y_{m,i}$ be the expected value of the observed intensity of the probe for mRNA target m in sample i, and D_m that for the control. Ideally, $Y_{m,i}$ are proportional to $X_{m,i}$, with a proportionality factor that might not be known, but is constant, so it would cancel out in within-probe comparisons, and an equality analogous to Equation (1) would still hold for the estimates. Unfortunately, current microarray technologies provide intensities that are not strictly proportional to the target abundance, but can be biased by an unspecific background:

$$Y_{m,i} = X_{m,i} + a_{m,i} \qquad \text{and} \qquad D_m = C_m + a_{m,C}, \tag{2}$$

where $a_{m,i}$ and $a_{m,C}$ are numbers that depend in a complicated way on the details of the technology and the experiment. Now for simplicity again dropping the target index m, we get for the logarithmic intensities and the log-ratios

$$Y_i^* = \log_2(X_i + a_i) \qquad \text{and} \qquad Y_i^c = \log_2 \frac{X_i + a_i}{C + a_C}. \tag{3}$$

The values a are called the background and often they are positive. Thus, the estimates of logarithmic fold-changes are typically attenuated towards zero, and the size of the attenuation depends on the abundance of the target. If the abundance is large, X_i is much larger than a_i and the attenuation is negligible. However, there are many genes that are present at low abundance. For these, the attenuation can be strong, and depends on the details of the experiment.

We note that *background correction* methods can alleviate the problem, but as they are not perfect, and because of finite sample effects, they do not completely remove it. In some cases, it may be possible to model the differential attenuation explicitly. We return to this idea below when considering general random effects models.

2.3 Exploratory Data Analysis

We propose some simple plots and diagnostics that will help to ascertain whether the same information has been measured and whether there seem to be grounds for combining the data from the different experiments. First, summary statistics can be computed separately for each experiment and compared on a gene by gene basis. Perhaps the most obvious statistic to compute is a t-statistic for each gene, but this is problematic. The t-statistic confounds the effect size and the sample size, in the sense that the same real difference in means will result in a larger t-statistic for larger samples. We propose using the *effect size*, described in Section 2.2, a quantity that is often considered in meta-analysis. In the example below we used the `zScores` function from the GeneMeta package to compute various per experiment and combined summaries.

Once the estimates have been computed they can be plotted against each other. If there are more than two groups then we suggest using a scatterplot matrix to show all pairwise relationships. Boxplots and parallel coordinates plots can also be valuable tools for assessing overall behavior. We anticipate general similarity of the estimated test statistics for all experiments, and gross deviations from this should be cause for concern.

We also suggest looking at the distribution of differences between the estimated effects. Under the belief that the experiments are measuring the same thing then these differences should be centered around zero. If this is not true then there are likely to be fundamental differences between the experiments that cannot be resolved by statistical means, but rather reflect real differences in the populations sampled or the technologies that were used.

We can see in Figure 1 that there are substantial qualitative differences between the two different comparisons. The correlation in estimated effects between the two ER experiments, 0.66, is remarkable. Typically you would not expect to see such a substantial correlation, and it potentially reflects the importance of ER as controller of transcriptional activity.

On the other hand, just because the plot for the lymph node comparisons shows almost no correlation does not mean that a synthesis or meta-analysis is inappropriate. In fact this is likely to be the more common case and it reflects the fact that there are relatively few genes whose expression levels correlate with lymph node status in both experiments. This is to be expected, the two data sets use samples from completely different tissues and we anticipate differences in the mechanisms that control expression as well as in which genes are expressed

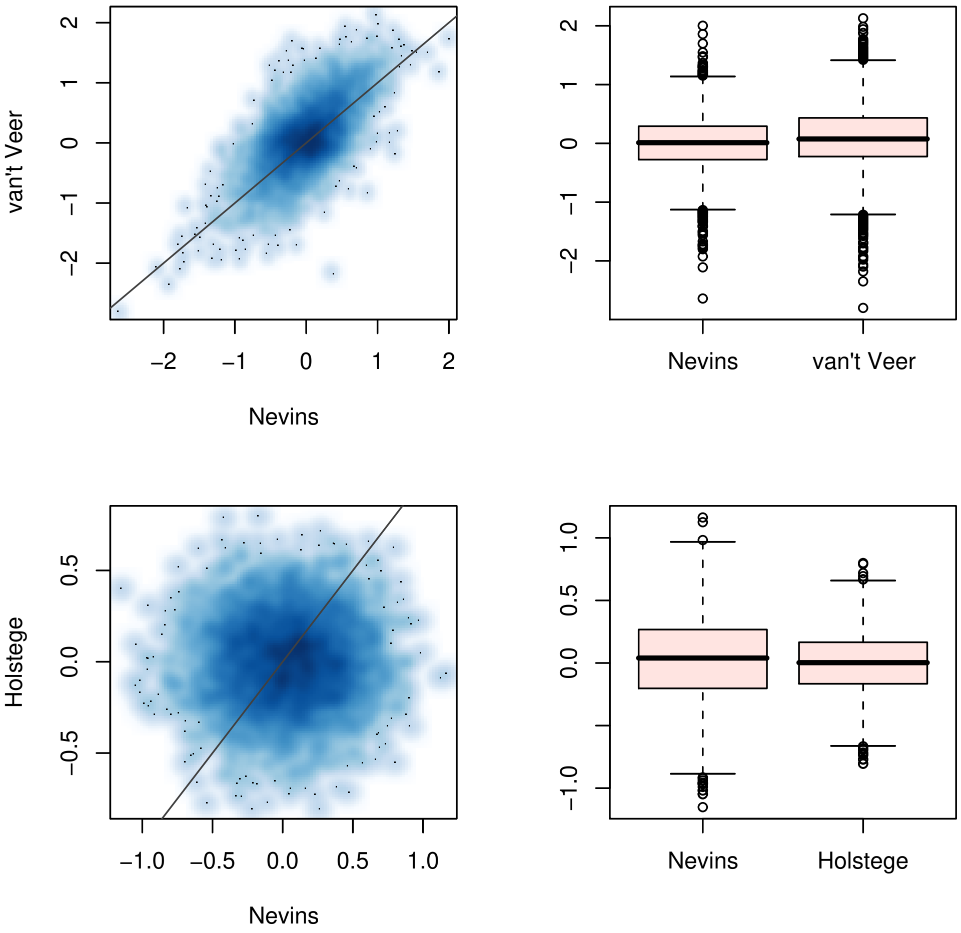

Figure 1: Pairwise plots and boxplots of per gene effect size statistics.

and at what level. But that does not preclude there being a relatively small set of genes that do show similar effects in the two experiments.

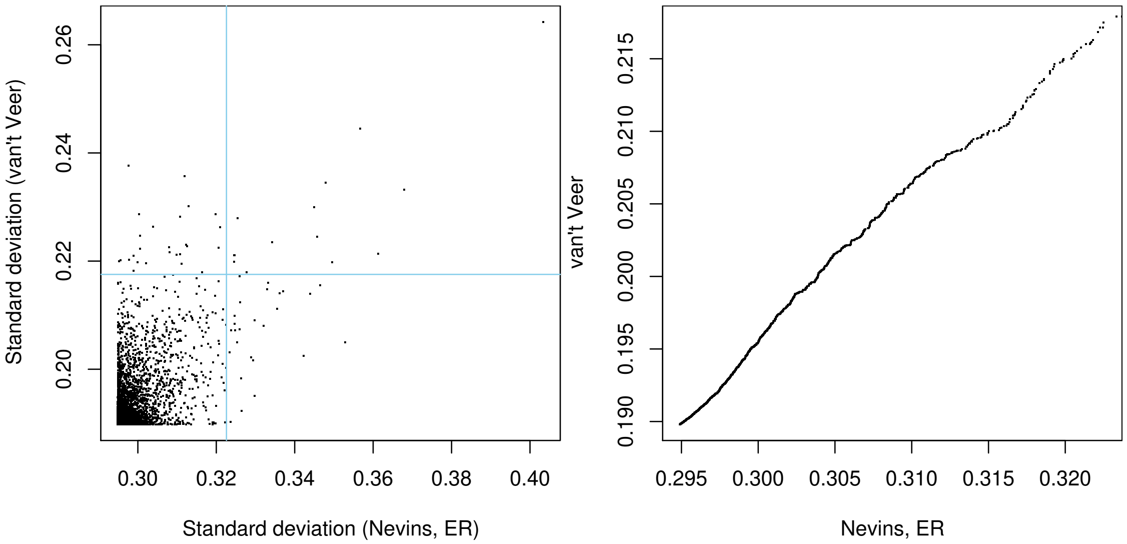

Figure 2: Comparison of the estimated per gene standard deviations from the random effects model for two datasets. Left panel: scatterplot, vertical and horizontal lines indicate the 99%-quantiles. There is a clear enrichment of genes in the upper right sector, that is, of genes that have a high standard deviation in both experiments simultaneously (odds-ratio 80, $p <$ 2.22e-16). Right panel: qq-plot, indicating that the shape of the distributions is about the same, but the standard deviations for the Nevins data are about one and a half times as high as those for the van't Veer data.

We compare per gene estimated standard deviations in Figure 2. If there are probes that have a high variance in one experiment, but not in the other, one might wonder about the validity of combining those genes. Large differences in variance between the experiments suggests that either there are large differences in the underlying populations or that the measurement device is failing for those particular probes.

3 The Model

3.1 A formalization of the statistical methodology

We now discuss a formal random effects model for each gene comparison. We note that in general the different genes are not independent and hence a *gene at a time* approach will not be optimal. However, in the absence of any knowledge about which genes are correlated with which other genes it is not clear how to approach a genuinely multivariate analysis. Here we describe the gene-at-a-time approach.

Following Cox and Solomon (2003) we write the model for each gene as:

$$Y_{tjs} = \beta_0 + \beta_t + b_j + \xi_{tj} + \epsilon_{tjs}, \tag{4}$$

where Y_{tjs} represents the expression value for the s^{th} sample in the j^{th} experiment, which is on treatment t. Note that we use the term *treatment* interchangeably with what would be called the disease condition or phenotype in the current application. β_0 is the overall mean expression, β_t is the effect for the t^{th} treatment, b_j is a random effect characterizing the j^{th} experiment, ξ_{jt} is a random effect characterizing the treatment by experiment interaction. We assume that the b_j have mean zero and variance τ_b, that the ξ_{jt} have mean zero and variance τ_ξ, and that ϵ_{tjs} are random variables with mean zero and variance τ_ϵ that represent the internal variability.

We can contrast the model in Equation (4) with the model proposed in Choi et al. (2003). Note that the Choi model is based on summary statistics, rather than individual observations. They let μ denote the parameter of interest, for instance, the effect size, possibly on a log scale, between control and treatment. They then let y_j denote the measured effect size for experiment j, with $j = 1, \ldots, k$ and propose the hierarchical model:

$$\begin{aligned} y_j &= \theta_j + \epsilon_j, \qquad \epsilon_j \sim N(0, \sigma_j^2) \\ \theta_j &= \mu + \delta_j, \qquad \delta_j \sim N(0, \tau^2) \end{aligned} \tag{5}$$

where τ^2 represents the between study variability and σ_j^2 denotes the within study variability. This is a random effects model, and the special case where $\tau^2 = 0$ is a fixed effects model, since then $\theta_j = \mu$, almost surely.

This is very similar to the model from Cox and Solomon (2003), albeit with additional distributional assumptions, which we shall ignore. The parameter μ from Equation (5) is related to β_t in Equation (4) if we are considering estimates of the difference in expression on the scale that the Y's are measured on. Since one typically fits the model in Equation (4) to the normalized data then β_t will correspond to a difference in means and not to the effect size. If the latter is wanted then dividing the observed data by the estimated per experiment standard deviations will make β_t and μ correspond to the same quantity.

The parameters δ_j in the Choi model roughly correspond to the ξ_{tj} in Equation (4). They represent terms which account for heterogeneity between the per experiment treatment effects. In Equation (4) they are interactions indicating that β_t is not the same in all experiments.

The parameters b_j from Equation (4) do not appear in the Choi model. They are not relevant, since one wants only to compare the estimated treatment effects, and whether or not there are between experiment differences in the per gene intensities is not relevant. The estimated b_j do have some potential uses.

One simplistic interpretation of the situation is as follows. Each experiment provides us with an independent estimate of the treatment effect and of the variance of that effect. If, across all experiments, the estimated treatment effects are in sufficient agreement (e.g. their confidence intervals are reasonably coincident) then there is no need for a more complex explanation. We can assess that question by testing whether τ^2 in Equation (5) is zero, and hence that $\theta_j = \mu$. The equivalent test, in terms of the model in Equation (4), is to test whether $\tau_\xi = 0$.

If however, the evidence is somewhat disparate then we must presume a more complex model to explain it. The dichotomy proposed by Choi et al. (2003) between fixed and random effect models is essentially that. This same situation is considered by Cox and Solomon (2003) where they indicate that in the first case, one should essentially estimate the overall treatment effect as a weighted average of the per-experiment effects, with weights determined by the per experiment variances, and in the second case as a simple weighted average of the estimates, discarding the per experiment variances. Cox and Solomon (2003) further argue that one can view most estimators used in practice as being some compromise between the two positions.

3.2 Interpretation of parameters

One of the real challenges in such an analysis is to find appropriate methods for making use of all of the data and to move beyond treating such an analysis as being several thousand unrelated analyses. In this section, we consider some of the issues involved and offer some preliminary advice. In Choi et al. (2003) the authors recommend computing Cochran's Q statistic for each gene and then use a qq-plot to compare the estimates to the distribution of Q, which is known under the null hypothesis that a fixed effects model is sufficient. They recommended that if there is significant departure from the presumed χ^2 distribution one should fit a random effects model and if not, then a fixed effects model should be used. Such a procedure presumes that the same model, fixed effects or random effects, is appropriate for all genes. However, that does not always seem to be a reasonable presumption. It seems likely that for any set of experiments there will be many genes for which there is no effect, others for which a fixed effects model is appropriate, and still others for which a random effects model is appropriate. However, the top left panel in our Figure 3 and in Figure 1 b) of Choi et al. (2003) there does seem to be evidence that the quantity being measured by Q is different for all genes.

In most analyses the parameter of interest is β_t, the treatment effect. In the hierarchical model this effect is represented by the parameter μ. As we noted previously the per gene estimates of β_t can be thought of as weighted averages of the per experiment estimates and should be interpreted in that light. Fitting a fixed effects model when a random effects model is appropriate is likely to elevate evidence against the null hypothesis since the quantity used to estimate the variability in the estimates of β_t tends to be too small, as it does not include the between experiment variance.

We next consider the per experiment random effects, namely the b_j. Now, one might anticipate that for each gene they estimate the same quantity, but this, it turns out, is in general not true. There are a number of different effects that are completely aliased with experiment. First there is the usual sampling variability. The samples used in one experiment represent a different sampling of the population than those used for a different experiment and different genes may show quite different levels of within and between sample heterogeneity.

The experiment effect is due to the technician, reagents used, as well as other factors. There is a separate technology effect that arises when different types of microarrays are used such as: short oligomer arrays, long oligomer arrays or cDNA arrays. For an examination of the likely sizes of these effects, in a study using technical replicates, see Irizarry et al. (2005).

On a per gene basis the b_j may reflect differences in labeling efficiencies and background between the two experiments. They may also reflect differences in absolute abundance, the experiments may have been carried out with more or less RNA. The b_j also capture differences in control of expression in cases where different tissues or different organisms are being compared. In this case, the expression of an mRNA may be up-regulated overall in one tissue, but the treatment effect could remain essentially the same. Unless the same type and version of microarrays were used in all experiments, the microarray probes will be different in different experiments, and that effect is also confounded with the experiment random effect. Any misassociation or mismatching of probes is likely to be manifested by an increased experiment effect. We also anticipate that in some circumstances the effect of mismatching could be manifested in the interaction between treatment and experiment.

We expect that some effects will be common to all genes, while others, such as that due to the probe sequence used, will be different for each gene. Hence, we propose examining the estimated per gene experiment random effects and suggest that those with particularly large estimated variances be examined to determine whether the large estimated effect can be explained. However we will report on that analysis elsewhere.

We next consider effects that are confounded with the σ_j. These are within experiment estimates of variability and they will be affected by any real differences in variability between the experiments. The estimates of σ_j will also capture inherent differences in the quality of the arrays and system used. They will also reflect sampling variability. It may be helpful to regress the σ_j on each other, large residuals suggest mRNAs which have different sources of variation in the two experiments; which may indicate genes controlled by different cellular mechanisms, or it may merely reflect a bad set of probes in one of the experiments.

Now we turn our attention to the interaction terms. As is often the case these are the most difficult to interpret, but also often the most valuable as they can indicate failures in the model - and potentially ways in which this failure can be remedied. There are two different sorts of interactions, and to some extent they should be considered separately. First, there is the sort of interaction where the effect in one study is in the same direction (either increased expression or decreased expression) but the size of the effect is substantially different. A cause for this could be differential attenuation as described in Section 2.2. The second type of interaction is the one where the direction of the effect is different. In this case the studies are presenting contradictory evidence for some genes. There can be many reasons for this and the predominant cause is likely to be that the mRNAs are not differentially expressed under the conditions being studied, and hence are not of interest. Other causes include misidentification of probe with target gene for some experiments, different regulation of expression in different tissues, as well as spurious results in one or more experiments.

3.3 Hypothesis Testing

The most common, and in some sense most important hypothesis test is determining whether, in the parlance of meta-analysis, to fit a fixed effects model or a random effects model.

Depending on the approach the test is whether $\tau^2 = 0$, for the meta-analysis formulation, or whether $\tau_\xi = 0$ for the random effects model. We note that there are, in general, problems when testing hypotheses of the form $H_0 : \tau_\xi = 0$, since the hypothesized value is on the boundary of the parameter space; variances must be positive. Such tests are problematic since the usual distribution theory does not apply. An examination of the p-values in Figure 4 clearly shows what can happen. There are a number of results in this general area (Self and Liang, 1987; Stram and Lee, 1994; Crainiceanu and Ruppert, 2004), but their application is not straight forward. The results of Self and Liang (1987) are only valid under a presumption of independent identically distributed data, and this presumption is not generally valid for random effects models and hence the results of Crainiceanu and Ruppert (2004) should be consulted.

4 Two Examples

We now return to the two examples we described previously and will demonstrate some simple analyses of these data using different software tools. We first use the GeneMeta package, which contains functions for carrying out most of the analyses described in Choi et al. (2003). Then we use the nlme package to fit the more general random effects models proposed in Cox and Solomon (2003) and examine some of the output of those analyses. We cannot directly compare the estimates obtained for two reasons. First, the meta-analysis approach is based on effect sizes while the random effects model is based on a difference in means. Second, even if we adjusted the data so that the parameters were the same the software used to fit the random effects model does not allow for fitting the same model as was used by Choi et al. (2003). Basically the way that the random effects are estimated using `lme` results in different weights being used to combine the per experiment effects than would be used for the classical meta-analysis.

4.1 Classical Meta-analysis

Most of the summary statistics mentioned by Choi et al. (2003) have been programmed as part of the `zScores` function in the GeneMeta package. In this section we use it to carry out a standard meta-analysis of the two different comparisons. The usual procedure is to first assess which of the two models, random effects or fixed effects, is appropriate and to then subsequently fit that model. The determination of which model is appropriate is often based on Cochran's Q statistic, if the value of this statistic is large then the hypothesis that the per-study measured effects are homogeneous is rejected and a random effects model is needed. In that case the most common approach to estimating the overall effect is to estimate τ^2 using the DerSimonian–Laird (DerSimonian and Laird, 1986) estimate. Both Brockwell and Gordon (2001) and Böning et al. (2002) raise concerns with respect to the use of the DerSimonian–Laird estimate. In particular if there are many small studies, then the DerSimonian–Laird estimate of τ^2 can be quite biased and this will be reflected in the estimate of the overall effect.

The value returned by the `zScores` function is a matrix that contains many of the statistics described in the Choi et al. (2003) paper, in particular per experiment estimates of effect and of variance, the DerSimonian–Laird estimate of τ^2, and Q. To determine whether to fit a fixed effect model or a random effects model Choi et al. (2003) propose using Cochran's Q statistic. Under the null hypothesis that the variance of the random effect is zero, this statistic have a χ^2_{k-1} distribution, where k is the number of experiments. Thus, comparing the estimates to the quantiles from a χ^2_{k-1} distribution provides a visual assessment of whether a fixed effects model may be tenable. The qq-plot is provide in Figure 3, and there seems to be substantial deviation from a χ^2_1 distribution. This indicates that a random effects model is prefered.

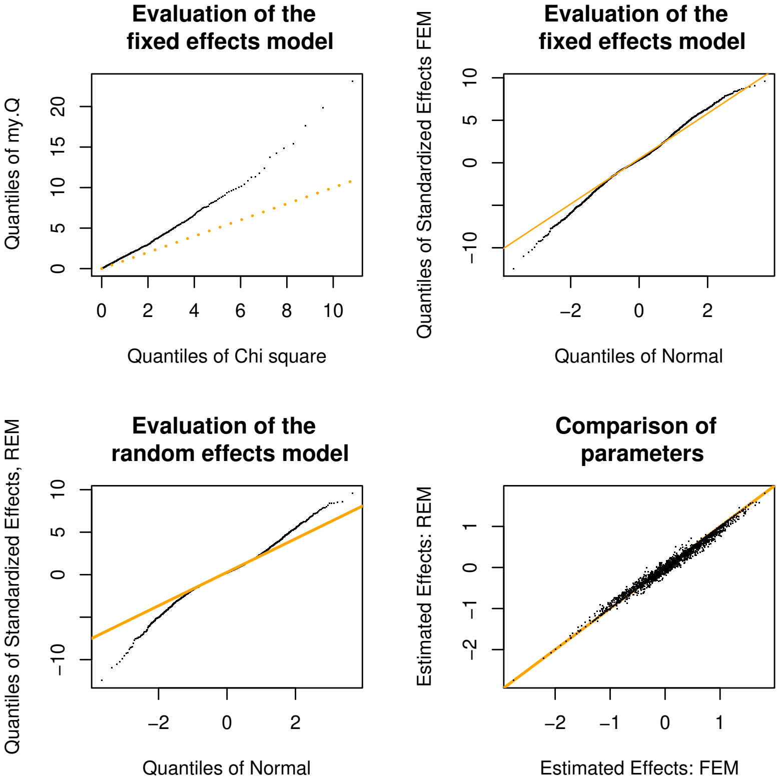

Figure 3: Plots evaluating and comparing a fixed effects model with a random effects model. In the first plot the empirical quantiles of Q are compared to a χ^2 distribution. In the second and third plots the standardized effect estimates are compared to the Normal distribution. In the fourth frame the two different estimates (fixed effects and random effects) are compared.

We interpret the plots in Figure 3 as follows. The first frame shows a qq-plot comparing Cochran's Q statistic to quantiles from the appropriate χ^2 distribution and there is a substantial deviation - the observed values are too large. However the two qq-plots that compare standardized effect estimates to the Normal distribution show that those from the random effects model deviate more from Normality. Finally in the fourth frame we demonstrate that the differences in the estimates from the two models are not really that large.

4.2 Estimation via linear models

We fit the model in Equation (4) where the treatment effect is a fixed effect, the experiment is considered to be a random effect and we include a treatment by experiment interaction. In the example below we take two different approaches to estimating this effect. We use both a fixed effects approach and a random effects approach. We note that the second of these is more appropriate since the wrong estimate of variance is used for testing the fixed effects estimate when a random effects model is appropriate.

We first test the hypothesis that no interaction term is needed. As noted above there are essentially two ways in which the interaction could be important. In one situation the treatment has an opposite effect in the two experiments, we can also detect this by simply comparing the estimated effects for each experiment estimated separately. For such probes, or genes, it would not be appropriate to combine estimates. In the other case, the interaction suggests that the magnitude of the effect is different in one experiment, versus the other. For these probes it may simply be the case that the model is incorrect. For example, we might be looking for a change in mean abundance while the magnitude of the effect is a function of the abundance, and hence in samples where the abundance of mRNA transcript is larger a larger effect is observed.

In Figure 4 we present a comparison of the p-values for different models fit to the breast cancer data sets. The first three frames compare p-values computed under two different assumptions on the interaction between experiment and ER status. Software in the nlme package was used to fit both models and the p-values are based on the likelihood ratio. In one model we presume that the interaction is a fixed effect and in the other we presume that it is a random effect. Perhaps the most striking feature in these histograms is the very large number of p-values around 1 for those computed assuming a random effect. This is a reflection of the fact that the hypothesis test here is being performed under non-standard conditions. The test is that the variance of the random effect is zero, and hence is on the boundary of the parameter space. In this case the asymptotics can be delicate (Crainiceanu and Ruppert, 2004) and further study is needed to fully interpret the output.

For those genes which did not exhibit a significant interaction effect, under the assumption that the interaction is a fixed effect, the lower right panel of Figure 4 compares the results from the joint analysis (column labeled C) versus that from individual analyses of the van't Veer (column labeled V) and Nevins (column labeled N) data sets. The rows in this panel correspond to the genes that have a significant treatment effect in at least one of the analyses, and a red area indicates that the effect is positive, blue that it is negative, and white that it is not significant. We can see that the overlap between the genes from the combined analysis is much larger with the van't Veer data than with the Nevins data.

In Figure 5 we repeat the computations with the lymph node status data. Some features in this second set of plots are similar to the plots based on ER status. First, the large peak

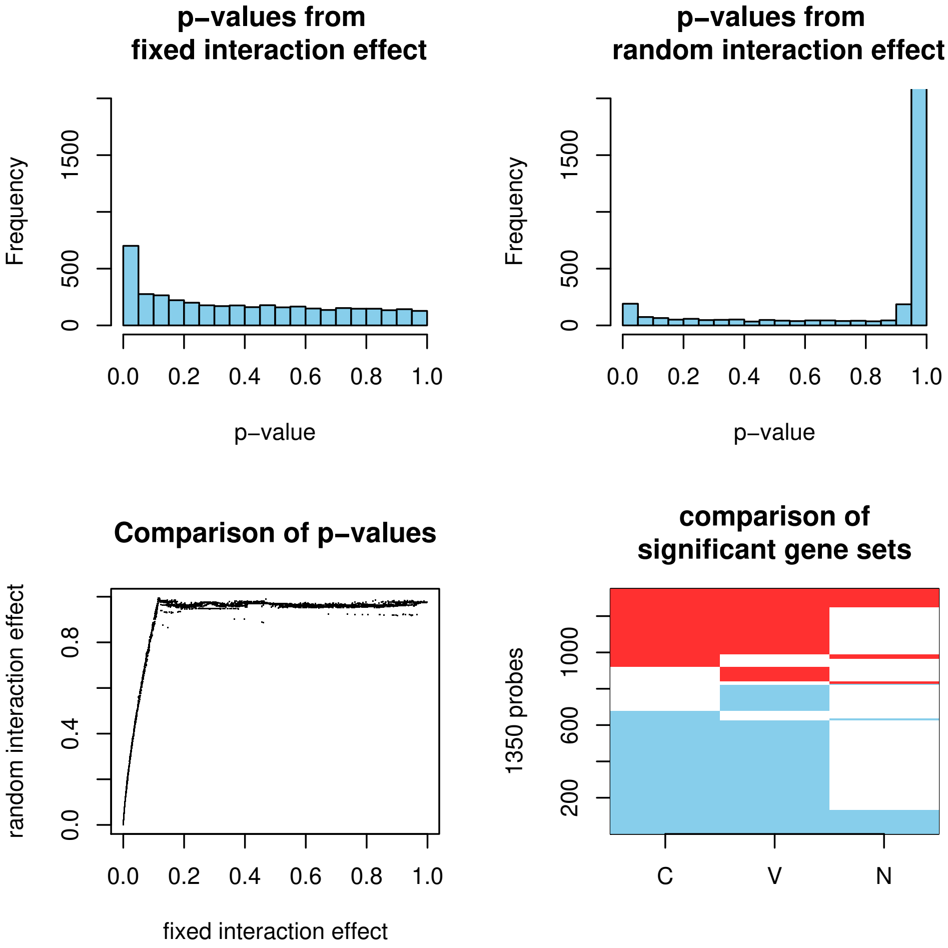

Figure 4: Histograms of p-values for the treatment–experiment interaction effect, estimated under two different models: fixed effects (upper left) and random effects (upper right). The bottom left panel shows the scatterplot of the two p-values for each gene, obtained by the two methods. The bottom right panel compares p-values for the combined analysis with those from the two separate analyses.

of p-values near 1 is also observed for the likelihood ratio statistics based on the random effects model. We also see that the fixed effects model is more likely to suggest a significant interaction effect, again this seems to reflect the fact that the variance being used to assess significance is too small. The bottom right panel suggests that there were relatively few genes for which there was evidence of an effect in both experiments. However, there are sizeable numbers for which neither experiment suggested an effect, but that the combined analysis did. Further exploration of these genes and their potential involvement in lymph node metastasis seems appropriate.

When comparing the bottom right panels in the two figures we note that the concordance between the analyses in the ER case is larger than that for the LN case. And that for the ER example relatively few genes were found to be significant in the combined analysis but not significant in either. For the LN analysis there were sizable numbers of genes that were found in the combined analysis but not in the separate analyses. Whether or not these observations are important will rely on a further and more detailed exploration of the underlying biology.

4.3 Benefits of integration

So far we have concentrated on a descriptive approach but there does remain a substantive question about what additional information is learned by synthesizing the data sets. A sensible assessment of this question is not simple. We will make some simplifying assumptions that will allow us to make a more specific comparison which can be assessed.

The question we want to address is whether or not the synthesis of two (or more) experiments provides better information than any single experiment. Direct assessment of this question requires data on which the truth is known; and we do not have that. So we will need to examine indirect methods. To simplify matters we first excluded those genes for which the interaction effect was deemed to be significant (here we used the fixed effect model since it was more aggressive). Then, for those that remained we fit the joint model, as well as models to both experiments individually.

A more substantive comparison would incorporate p-value correction methods (Dudoit et al., 2003; Reiner et al., 2003) but in the interest of simplicity we do not address that important question here. We remind the reader that precisely the same number of tests were performed in all cases that we are comparing, so any effects of p-value correction are limited to corrections for strength of evidence and not for different numbers of tests.

We use the unadjusted p-values and find those genes which have a p-value less than 0.01 in the combined analysis but for which the p-value exceeded 0.01 in each separate analysis. For the ER comparison we found 88 genes which had a p-value of less than 0.01 in the combined analysis and values larger than that in each experiment. In Figure 6 we provide a scatterplot matrix comparing those estimates for the different analyses. For comparison there were 235 genes that were significant at the 0.01 level in all three analyses.

For LN status we find that there are 30 genes with p-values less than 0.01 in the combined analysis but with p-values that exceed 0.01 in both experiments. There was 1 gene that was significant at the 0.01 level in all three analyses. There are many fewer genes that are significant in all experiments and fewer that are significant in the combined analysis but not

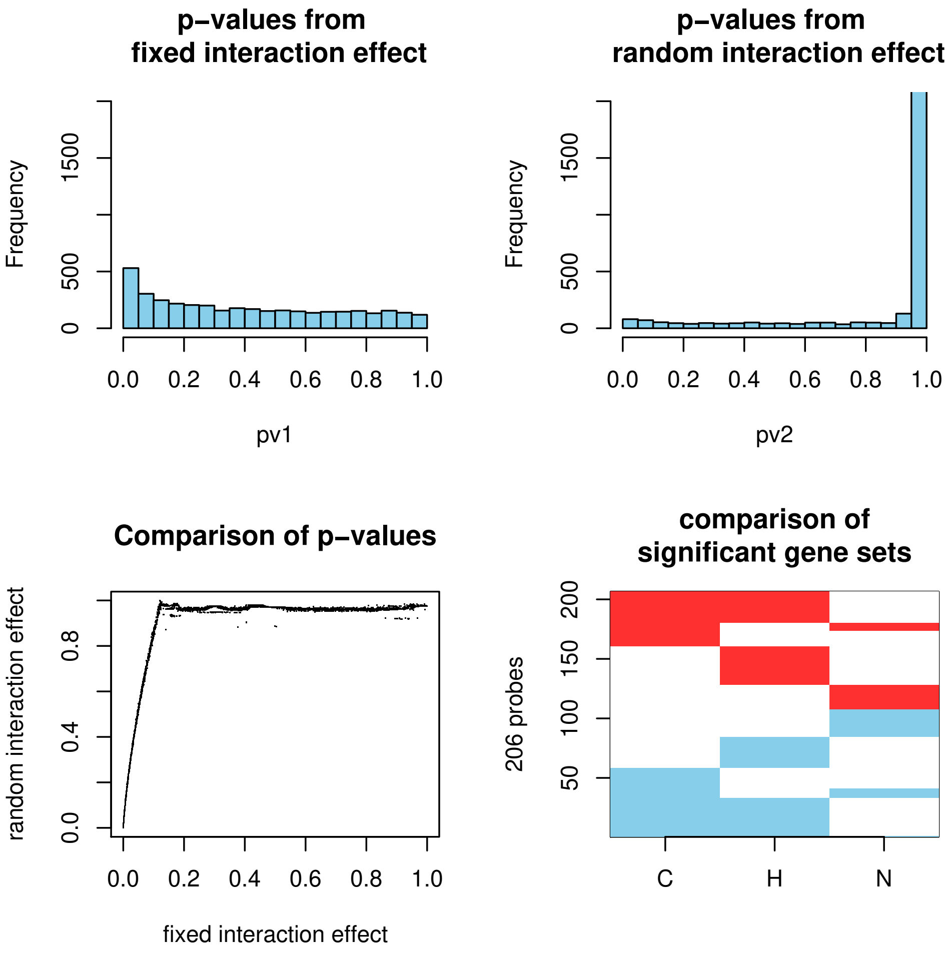

Figure 5: In analogy to Figure 4, p-values for treatment–experiment interaction effects and comparison of the genes selected for treatment main effect for the lymph node data sets. In the lower right panel, C refers to the combined analysis, H to the Holstege data, and N to the Nevins data.

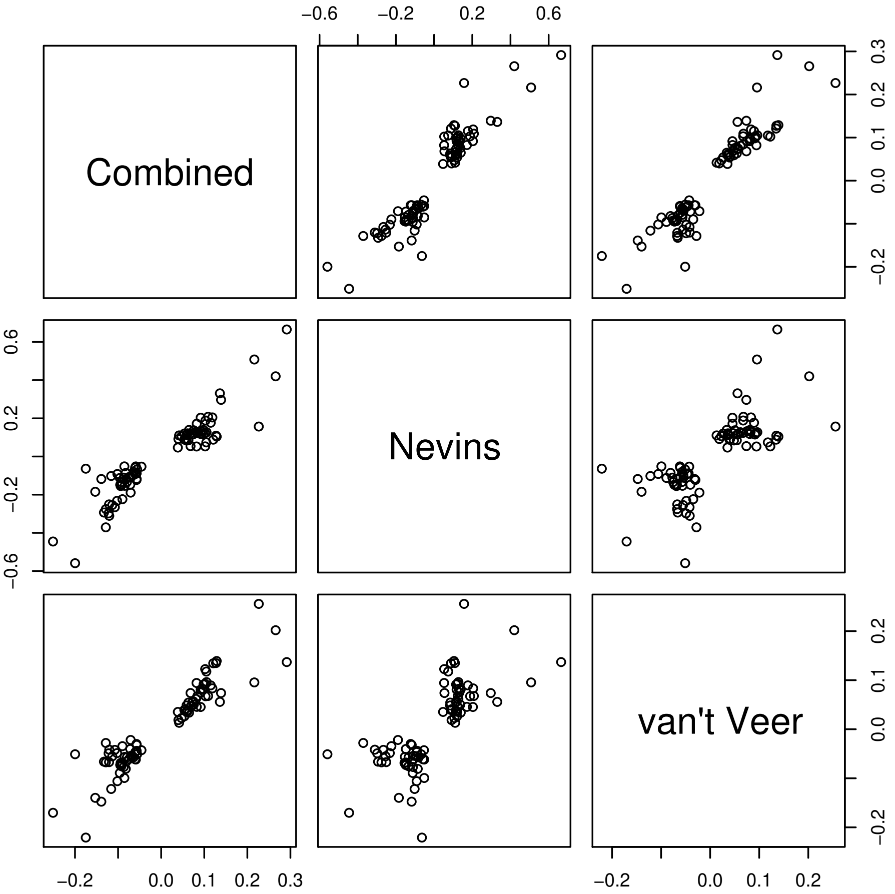

Figure 6: Comparison of estimated coefficients for the effect due to ER status for a combined model and two models fit to individual data sets.

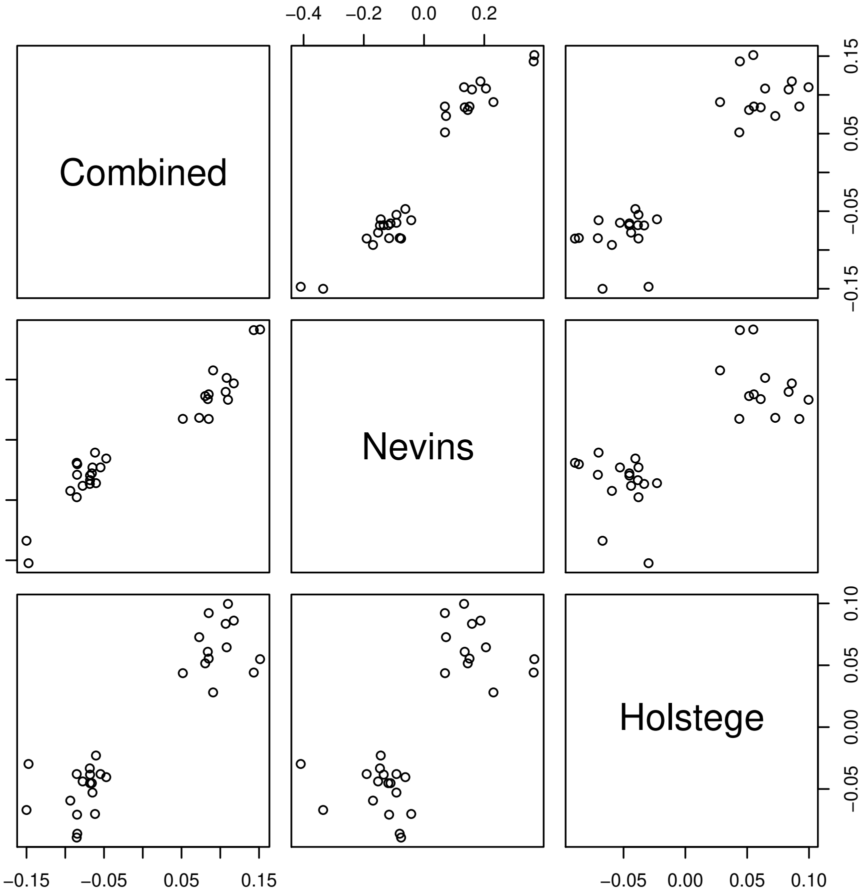

Figure 7: Comparison of estimated coefficients for the effect due to LN status for a combined model and two models fit to individual data sets.

in the individual analyses. We conjecture that this is likely due to the fact that different tissues are involved in the LN comparison. The scatterplot matrix comparing these estimates is provided in Figure 7.

Whether, in fact the combined analysis is beneficial of course rests on the identities and roles of the genes that have been identified in the combined analysis. That of course relies on further biological research and experimentation.

5 Discussion

We have re-examined the methodology proposed in Choi et al. (2003) and contrasted it with an analysis based on mixed effects models (Cox and Solomon, 2003; Pinheiro and Bates, 2000). The two approaches are similar and both can be extended to deal with more complex models and adjustments for other covariates can be included – the computations are of course not the same.

We note that for any specific mRNA we are proposing that a simple random effects model be fit and there are many well-known diagnostic plots that can be usefully applied to assess the appropriateness of the model. Once a set of interesting genes has been identified it will be prudent to explore the models for these genes in more detail. We also suggest that diagnostic plots considering results across genes may be quite informative.

Here, as in other modeling situations with genomic data, simply fitting models for every gene available is likely to lead to substantial inefficiencies due to the need for multiple testing corrections. Even though we are increasing the sample size, there are still far too many probes to test without regard to the penalty for multiple comparisons. Examining small sets of genes that are of known interest will help to ensure that correct inferences are made.

What we have only begun to address is the challenge of comparing thousands of related model fits. We proposed some simple exploratory tools and suggested a number of avenues that may bear fruit, however much remains to be done. Our investigations here have revealed more questions than they have provided answers and our treatment is more in the nature of an introduction to what promises to be an interesting aspect of statistical research. Meta-analysis has been widely use in medical applications where typically only summary statistics are available. The challenges with genomic data are much larger since the available data are much richer and more complex. Combining different forms of experimental data such as microarray data, comparative genomic hybridization data, SNP data, etc., will raise many other problems.

Acknowledgments

The authors would like to thank Ting-Yuan Liu for his help in assembling the data used in the examples. They would also like to thank Doug Bates for help with nlme and Thomas Lumley for discussions on meta-analysis.

References

D. Böning, Malzhan U., E. Dietz, et al. Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics*, 3:445–457, 2002.

S. E. Brockwell and I. R. Gordon. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, 20:825–840, 2001.

J. K. Choi, U. Yu, S. Kim, et al. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19, Suppl. 1:i84–i90, 2003.

D.R. Cox and P. J. Solomon. *Components of Variance*. Chapman and Hall, New York, 2003.

C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *JRSS, B*, 66:165–185, 2004.

R. DerSimonian and N. M. Laird. Meta-Analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188, 1986.

S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

G. V. Glass. Primary, secondary and meta-analysis of research. *Educational Researcher*, 5: 3–8, 1976.

R.A. Irizarry, D. Warren, F. Spencer, et al. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2:345–350, 2005.

G. Parmigiani, E. Garrett-Mayer, R. Anbazhagan, et al. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clincal Cancer Research*, 10:2922–2927, 2004.

J. C. Pinheiro and D. M. Bates. *Mixed-effects models in S and S-PLUS*. Springer, New York, 2000.

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

P. Roepman, LFA. Wessels, N. Kettelarij, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat Genet*, 37 (2):182–186, Feb 2005.

S. G. Self and K-Y Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610, 1987.

D. O. Stram and J. W. Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50:1171–1177, 1994.

L. van't Veer, H. Dai, MJ. van de Vijver, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

M. West, C. Blanchette, H. Dressman, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, 98(20):11462–11467, 2001.