12-2-2004

Cross-study Validation and Combined Analysis of Gene Expression Microarray Data

Elizabeth Garrett-Mayer
Johns Hopkins University, esg@jhu.edu

Giovanni Parmigiani
Division of Biostatistics, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, gp@jimmy.harvard.edu

Xiaogang Zhong
Department of Mathematics and Applied Statistics, Johns Hopkins University

Leslie Cope
Division of Biostatistics, The Sidney Kimmel Comprehensive Cancer Center

Edward Gabrielson
Department of Pathology, Johns Hopkins School of Medicine, egabriel@jhmi.edu

Cross-study validation and combined analysis of gene expression microarray data

Elizabeth Garrett-Mayer
Division of Biostatistics, The Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins School of Medicine, Baltimore, MD 21205

Giovanni Parmigiani Division of Biostatistics, The Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins School of Medicine, Baltimore, MD 21205

Xiaogang Zhong
Department of Mathematics and Applied Statistics,
Johns Hopkins University, Baltimore, MD 21218

Leslie Cope
Division of Biostatistics, The Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins School of Medicine, Baltimore, MD 21205

Edward Gabrielson
Department of Pathology,
Johns Hopkins School of Medicine, Baltimore, MD 21205

Abstract

Investigations of transcript levels on a genomic scale using hybridization-based arrays led to formidable advances in our understanding of the biology of many human illnesses. At the same time, these investigations have generated controversy, because of the probabilistic nature of the conclusions, and the surfacing of noticeable discrepancies between the results of studies addressing the same biological question. In this article we present simple and effective data analysis and visualization tools for gauging the degree to which the finding of one study are reproduced by others, and for integrating multiple studies in a single analysis.

We describe these approaches in the context of studies of breast cancer, and illustrate that it is possible to identify a substantial, biologically relevant subset of the human genome within which hybridization results are reproducible. The subset generally varies with the platforms used, the tissues studied, and the populations being sampled. Despite important differences, it is also possible to develop simple expression measures that allow comparison across platforms, studies, labs and populations. Important biological signal is often preserved or enhanced. Cross-study validation and combination of microarray results requires careful, but not overly complex, statistical thinking, and can become a routine component of genomic analysis.

1 Introduction

Microarray experiments measure simultaneously the transcriptional activity of a large number of genes. In recent years, hundreds of these experiments have been performed, providing important insight on gene regulation, and revealing some interesting relationships between genes and disease phenotypes. Yet, because of cost and other practical limitations, most microarray studies have used a relatively small number of biological samples. As a result, cross-referencing lists of genes found to be associated with disease phenotypes in two separate studies usually produces relatively few genes in common (Parmigiani *et al.*, 2004), even when one restricts attention to genes measured in both experiments. While an incomplete overlap is to be expected given the small samples typically used and the large number of comparisons made, discrepancies have generated skepticism of this type of investigations.

In this scenario, three related statistical questions are important to making progress towards an objective assessment of the worthiness of microarray analysis results: 1) reproducibility, that is whether different measuring techniques are capturing the same biological variation; 2) validation, that is whether the conclusions of a study are supported by other similar studies and 3) combination, that is whether more reliable conclusions can be reached by jointly analyzing multiple studies. In this paper we develop and illustrate simple and effective statistical approaches to address these three questions.

Variation in measurements of gene expression includes "technological" variation, associated with limitations of the measuring technologies, and "biological" variation, due to the phenotype or experimental condition being studied, as well as natural variation of levels of gene expression in different samples of the same type (Pritchard *et al.*, 2001; Oleksiak *et al.*, 2002; Enard *et al.*, 2002). Because the determinants of both technological and biological variation tend to vary from study to study and from lab to lab, study- and lab-specific effects are inherent in most gene expression array datasets. Study-specific conditions may affect different genes differently, generating study "signatures." Overall, study effects can dominate the biological signal of interest (Aach *et al.*, 2000) in a large number of genes.

There are several microarray technologies currently in use (Schena, 2000; Southern, 2001; Hardiman, 2002). Although all exploit hybridization, they differ in how DNA sequences are laid on the array, in the length of these sequences and in the number of samples measured in each hybridization. As a result, an important source of technological variability in gene expression measurements is the platform used. Several studies have compared measurements across platform. Kuo *et al.* (2002) compared mRNA measurements from Stanford-type cDNA microarrays and Affy oligonucleotide chips using the so-called "NCI 60" set of cancer cell lines (Ross *et al.*, 2000; Scherf *et al.*, 2000). Based on correlation between matched measurements and concordance between clusters, they concluded that correlation can be poor and clusters of genes and cell lines can be discordant between the two technologies. They also provide evidence to indicate that sequence-specific factors influence reproducibility. Similar caveats about cross-platform variability have been raised by other analyses comparing Affy to custom-made cDNA arrays (Yuen *et al.*, 2002) and Affy to IncyteGenomics arrays (Kothapalli *et al.*, 2002).

To compare experiments that are performed on different gene expression platforms, oligonucleotide probe sets, spotted sequences, and other microarray features need to be linked. Expressed sequence tag (EST) sequencing projects have generated cDNA sequences for human, mouse and other organisms. These are identified by an accession number in databases such as GenBank. Extensive efforts have been devoted to grouping these sequences into clusters representing a single transcript (Boguski and Schuler, 1995; Miller *et al.*, 1999; Quackenbush *et al.*, 2001). UniGene, developed at the National Center for Biotechnology Information (NCBI) (National Center for Biotechnology Information, 2003), partitions ESTs derived from one organism into mutually exclusive clusters based on sequence homology (Boguski and Schuler, 1995). As GenBank is growing, UniGene clustering is performed periodically, resulting in new clusters. Typically, a sequence-specific identifier (GenBank accession number) serves as a reference to the array probe sequences. A subset of the UniGene clusters is reliably linked to genes of known function such as those catalogues in LocusLink (National Center for Biotechnology Information, 2004; Pruitt and Maglott, 2001).

Because of these challenges, the first and most critical step in cross-study analysis of gene expression is to identify a subset of genes that are consistently measured across platforms. Even after two features are mapped to the same Unigene cluster of LocusLink ID, inconsistencies across platforms can still be substantial because of differences in hybridization efficiencies, limitations of linkage databases, isoform variation, and a number of other factors. To this end, we propose a tool, termed integrative correlation, that can be used to investigate reproducibility and to isolate a subset of reproducible genes for further analysis. Integrative correlation was previously illustrated, though not described in any statistical detail, in Parmigiani *et al.* (2004) for the two study case. Here we provide a rigorous discussion in the general case of an arbitrary number of studies, and describe how to choose reproducibility cutoffs using false discovery rates.

We then turn to the assessment of reproducibility of gene selection in class comparison analyses, whose goal is to identify the genes that are differentially expressed across a given set of conditions or phenotypes. We propose exploratory analysis techniques based on visualizing suitably chosen standardized effect sizes. Such visualizations make it simple to place the study-specific effect sizes and their discrepancies across studies in the context of the variation across the genome. We also build on these visualization to propose simple meta-analytic methods for selecting genes that are reproducibly associated with a phenotype of interest, and for assessing the false discovery rates associated with this selection. While typical meta-analytic approaches focus on combination, our genome-wide implementation focuses on reproducible selection, where reproducibility is refined both in terms of integrative correlation and consistency of effect sizes.

To illustrate our statistical methods for assessing reproducibility, and comparing results across studies, we will use three breast cancer datasets (Hedenfalk *et al.*, 2001; Van't Veer *et al.*, 2002; Huang *et al.*, 2003). Two of these studies provide information about BRCA1 status of cancers (Hedenfalk *et al.*, 2001; Van't Veer *et al.*, 2002) which we will use to determine which genes show evidence of differential expression in BRCA1 and sporadic breast cancers. BRCA1 positive tumors have shown in some studies to be associated with decreased survival as compared to sporadic cancers (Foulkes *et al.*, 1997; Robson *et al.*, 2004; Stoppa-Lyonnet *et al.*, 2000; Moller *et al.*, 2002), while other studies show no association (Verhoog *et al.*, 1998; Pierce *et al.*, 1998).

2 Data

The datasets that we have chosen are three publicly available breast cancer microarray gene expression studies, of which two had information about BRCA1 status. The first study is by Van't Veer *et al.* (2002) who selected 98 primary breast cancers (18 BRCA1 and 80 sporadic). From each patient, total RNA was isolated and used to obtain complementary RNA (cRNA). For each sample, two hybridizations were performed using fluorescent dye reversal on oligonucleotide microarrays containing about 25,000 genes. A reference pool of cRNA was created by pooling equal amounts of cRNA from each of the sporadic tumor samples. Intensities were then quantified as log-ratios (as compared to the reference pool), and normalized. The data from this experiment was downloaded from `http://www.nature.com`.

The second data was first published by Hedenfalk *et al.* (2001). Twenty-two breast cancers were analyzed of which seven are BRCA1 tumors. Complementary DNA (cDNA) was obtained from each tumor sample and hybridized to two channel cDNA arrays. The reference sample was cell-line MCF-10, a nontumorigenic breast-cell line. Data from this study was downloaded from `http://www.nejm.org`, and can also be accessed at `http://www.nhgri.nih.gov/DIR/Microarray`.

Huang *et al.* (2003) analyzed 89 heterogeneous breast tumors which were obtained at biopsy of primary tumor and banked between 1991 and 2001 and chosen based on clinical parameters. Total RNA was extracted and synthesized to cDNA. Affymetrix arrays were used for hybridization, arrays were scanned using Affymetrix GeneArray scanner. We obtained the orignal CEL files from scanned chips and gene expression was quantified from probe-level information using RMA (Irizarry *et al.*, 2003). BRCA1 information was not available for this dataset, but we included it in our analysis to demonstrate assessing gene reproducibility across more than two datasets. The Huang data can be accessed at `www.thelancet.com`.

Each dataset was preprocessed before being made publicly available.

3 Methods

The methods can be divided into three areas: evaluating reproducibility and reliability of gene expression across studies, comparing strength of evidence of gene-phenotype associations across studies, and combining effects across studies. For each dataset, the only genes of interest are the ones which are common across the studies being compared. For ease of computation, the datasets are hence subsetted to include only the common genes based on Unigene ID and are then sorted so that the order of the genes is identical across studies.

The tools for performing the analyses we demonstrate in this paper are available as an R libary called `MergeMaid` (Cope *et al.*, 2004) which can be downloaded at `http://astor.som.jhmi.edu/MergeMaid`. Functions are available for merging data, estimating correlations within and across studies, performing comparative analyses, and validating gene sets.

3.1 Notation

We use f and h to index studies (i.e., datasets), j and k refer to genes within studies, and n_f refers to the number of samples within study f. The genes which are common across two studies f and h (i.e., the intersection of genes) is denoted by $\mathcal{G}_{fh}$ whereas $\mathcal{G}_f$ is the set of all genes in study f.

3.2 Reproducibility of genes: Integrative correlation

3.2.1 Integrative correlation of two studies

When considering just one gene expression dataset, it can be difficult or impossible to determine whether or not expression levels are reliably measured. If the spot on the chip is incorrect (i.e. the sequence spotted does not correspond to the gene that is assumed to be spotted), which is not an uncommon occurence, it will usually be consistently incorrect for all of the genes in one study because the same type of chips are generally used within one study. By comparing patterns of expressions across two studies which use different types of chips (e.g. Affymetrix oligonucleotide chips versus spotted cDNA glass arrays), we may be able to determine if there are inconsistencies by looking at how the genes are regulated in relation to other genes. Additionally, if a gene shows relatively little variability across samples, we would expect it to have relatively low correlations with other genes. To assess which genes lack reproducibility and lack variability, we consider correlations between genes.

Specifically, gene reproducibility is assessed by looking at the correlation structure across studies. For a dataset with G genes, the $G \times G$ correlation matrix describes how each gene is correlated to every other gene in the study. If we calculate the correlation matrices for two studies using the same set of G genes and these datasets both have comparable and reliable gene expression measures across samples, then we would expect that the two correlation matrices would be similar. In other words, for gene j, the correlation between the j^{th} row of the correlation matrix in study f and the j^{th} row of the correlation matrix in study h would be high. We term this the "integrative correlation" for gene j, denoted by r_j^{fh}

$$r_j^{fh} = \frac{\sum_{g=1,g\neq j}^{G}(\rho_{fjg} - \bar{\rho}_{fj})(\rho_{hjg} - \bar{\rho}_{hj})}{\sqrt{\sum_{g=1,g\neq j}^{G}(\rho_{fjg} - \bar{\rho}_{fj})^2 \sum_{g=1,g\neq j}^{G}(\rho_{hjg} - \bar{\rho}_{hj})^2}} \tag{1}$$

where ρ_{fjg} is the correlation between genes g and j in study f, and $\bar{\rho}_{fj}$ is the average correlation between gene j and all of the other G genes being assessed. This gene-specific measure can tell us which genes tend to be measured consistently and with agreement across studies. In general, we do not necessarily expect there to be high correlations, but we do expect that we will see overall positive trends. Genes showing negative trends or no trend suggest that the gene signals across the two studies are different for that particular gene, which may be due to mislabeled spots on the arrays, other chip-specific problems, or artifacts of the experimental conditions. Genes showing no trend (i.e. correlation close to zero), may lack variability in one or both studies. Variability can be measured by looking at a gene's correlation with other genes within the study: if the variability in the gene is due to signal as opposed to noise, we would expect that the gene would show a wide range of correlations with other genes. If the range of correlations for a given gene is narrow, then we conclude that the gene shows little variability. As a result, genes with a narrow range of correlation with other genes will not be useful in our meta-analysis: the variation in these genes is likely random (i.e. not related to phenotype) and they will not be helpful is distinguishing between phenotypes.

For any given gene comparison across two studies, we highlight five common scenarios for the resulting integrative correlation: (1) r_j^{fh} is positive and gene j shows variability in both studies f and h, (2) r_j^{fh} is negative and shows variability in both studies f and h, (3) r_j^{fh} is low (i.e. close to zero) and variability of gene j is high in both studies f and h, (4) r_j^{fh} shows low correlation and variability of gene j is high in only one of the studies, and (5) r_j^{fh} is low and variability of gene j is low in both studies. These are exhibited in Figure 1, where the five scenarios are shown via genes 1 through 5, respectively. For each panel, the correlations of gene j with each of the other genes in $\mathcal{G}_{fh}$ in study f is plotted versus the correlations of gene j with each of the other genes in $\mathcal{G}_{fh}$ in study h, with the integrative correlations shown for each panel.

Of the five types of genes shown in figure 1, only genes with patterns similar to gene 1 are of interest. Genes 2 and 3 show different patterns of expression in the two studies, suggesting lack of reproducibility. For Gene 5, we see low variation in both studies so that the gene is not likely to be informative for distinguishing between phenotypes. Gene 4 only shows variability in one of the two studies: this suggests that gene 4 could either be (a) poorly or incorrectly measured in one of the studies, or (b) one of the studies lacks heterogeneity of samples for gene 4 while the other study shows signficant variability across samples. Regardless of whether (a) or (b) is true, the gene does not show a comparable pattern across the two studies and cannot be considered consistently measured. Hence, in addition to identifying genes with inconsistent patterns in the two studies (i.e. genes like genes 2 and 3), the integrative correlation of genes also identifies genes which do not show sufficient variability in co-regulation across samples (i.e., like genes 4 and 5). For simplicity, we refer to genes similar to gene 1 as “reproducible” and genes like genes 2, 3, 4, and 5 as “non-reproducible.” We consider the integrative correlation a gene’s “reproducibility score.”

Plots like those shown in figure 1 can be made rather easily to see patterns of reproducibility across datasets. However, $\mathcal{G}_{fh}$ is usually large so that a high throughput method is necessary for determining which genes are reproducible or not via a cutoff for reproducibility. An attractive approach due to its ease of implementation and its popularity in gene expression analyses is to use a permutation method. The data within the rows of the original data matrices are randomly permuted, and the gene-specific correlations recalculated (i.e. the $G \times G$ correlation matrix is recalculated for study f and study h). The reproducibility score based on the permuted versions of datasets can be used to estimate the null distribution of reproducibility scores: the scores we would expect to see if the gene expression dataset were based purely on random cross-linkage of array features. By comparing the null distribution of reproducibility scores to the distribution based on the original datasets, we can determine a cutoff for reproducibility for which it is expected that only a small fraction of the genes with scores higher than the cutoff are “false positives” for reproducibility. This is akin to the false discovery rate (Tusher *et al.*, 2001), but instead of using a cutoff to control the false positive rate for genes that show differential expression, here we are establishing a cutoff to control the percentage of genes retained for analysis that are not reproducible across studies. Once the cutoff is chosen, we can proceed in the analysis using only the genes that have been deemed reproducible.

3.2.2 Integrative correlation of two or more studies

If we are interested in finding out which genes are reproducible across more than two studies, the integrative correlation as defined above is not directly applicable. As an extension, we consider using an approach based on the eigenvalues and eigenvectors of the correlation matrices, and using a principal components analysis approach. Consider first the situation where we are evaluating the reproducibility of gene j in just two studies and we calculate the correlation between column j in each of correlation matrices of studies f and h. The two eigenvalues associated with the comparison of these two vectors of correlations are $1 + r_j^{fh}$ and $1 - r_j^{fh}$, corresponding to the first and second principal components (Johnson and Wichern, 1999; Everitt and Dunn, 2001). When $r_j^{fh} > 0$, then the $1 + r_j^{fh}$ is the eigenvalue associated with the first principal component, and the second principal component when $r_j^{fh} < 0$. Although it is not obvious by looking at the two eigenvalues whether or not $r_j^{fh} > 0$, by looking at the signs of the elements in the eigenvector of the first principal component (i.e. by looking at the “loadings” of the first principal component), we can tell if the correlation is positive or negative. That is, if the two loadings have opposite signs, then $r_j^{fh} < 0$. So, instead of calculating the integrative correlation, an equivalent approach would be calculate the eigenvalues,

Table 1: Comparison of average correlations to λ_1^* for three studies

gene	pairwise correlations 1vs2 1vs3 2vs3	average of correlations	λ_1^*
A	0.44 0.16 0.32	0.30	0.54
B	0.71 0.05 -0.01	0.25	-0.57

λ_1 and λ_2, where $\lambda_1 > \lambda_2$ and the eigenvector of the first principal component is $\{a_{11}, a_{12}\}$. We then define

$$\begin{aligned} \lambda_1^* &= \lambda_1/2 \quad \text{if } a_{11} \times a_{12} > 0 \\ &= -\lambda_1/2 \quad \text{if } a_{11} \times a_{12} < 0. \end{aligned}$$

In the above equation, the integrative correlation r_j^{fh} and λ_1^* are equivalent. Additionally, we can use a permutation approach as described in the previous section to estimate the null distribution of λ_1^* and we obtain the same set of reproducible genes.

Using the eigenvalue approach, we can now extend integrative correlation to three studies. Recall that in a principal components analysis of k variables, the sum of the eigenvalues $\sum \lambda_i = k$, and λ_i/k is the proportion of variability in data that is described by the i^{th} principal component. As such, if the eigenvalue of the first principal component is high, it indicates that there is a strong linear relationship for gene j across all three studies. Note, however, that the eigenvalue does not distinguish between the direction of the associations: it could be that two studies have strong agreement for gene j and the third study has strong by negative correlation of gene j with the other two studies. In that case, the eigenvalue would be large but the gene is not reproducible across the three studies. However, just as we can use information in the eigenvectors in the case of two studies, we can use an analogous approach for $k = 3$ study comparisons:

$$\begin{aligned} \lambda_1^* &= \lambda_1/3 \quad \text{if } a_{11} \times a_{12} > 0 \text{ and } a_{11} * a_{13} > 0 \\ &= -\lambda_1/3 \quad \text{if } a_{11} \times a_{12} < 0 \text{ or } a_{11} * a_{13} < 0. \end{aligned}$$

Note that the λ_1^* statistic takes the same range as a correlation and has a similar interpretation. In the above equation, only if all three components of the first eigenvector have the same sign will λ_1^* have a positive sign. Similarly, a permutation distribution can be determined as above but in this case all three studies are required to be randomly permuted before re-analyzing the data. The principal components approach can clearly be extended to assess reproducibility between any number of studies.

The above approach is similar, yet superior to estimating the pairwise integrative correlation between each pair of studies and averaging the pairwise correlations. In table 1, there are two different correlation scenarios that produce similar average correlations, but quite different values of λ_1^*. Gene A shows moderate correlation across the three studies, resulting in an average correlation of 0.30 and a value of λ_1^* of 0.54. For gene B, it appears that the gene is measured consistently in two of the studies, with a high correlation between two studies of 0.71, but the gene is poorly measured in study 3, with correlations between studies 1 and 3 of -0.05 and between studies 2 and 3 of -0.01. For gene B, the average correlation is 0.25, while $\lambda_1^* = -0.57$. So, if using the average correlation, we cannot distinguish between the types of correlation patterns in genes A and B, while there is a very distinct difference in λ_1^*.

Depending on the ultimate goal of these analyses, the λ_1^* value can be used for different purposes. Using a permutation approach, we can identify all genes which show agreement above that which would be expected due to chance. Specifically, we randomly permute the gene expression values (by gene) in the three studies and repeat the estimation of λ_1^* using the now permuted datasets. We can then compare the empirical distribution of λ_1^* based on the permuted and non-permuted datasets. Some cutoff s_α is chosen to control the rate of unreproducible genes included in the analysis. In the example shown in table 1, gene A would likely be chosen as reproducible while gene B would not.

Note that the range of values of λ_1^* can also be used for identifying genes which show strong evidence of poor measurement in one of the studies. Genes with large negative values of λ_1^* such as gene B clearly show some agreement (in gene B, there is strong agreement between studies 1 and 2), but the negative sign indicates that one study does not agree with the other studies. So, comparing the pairwise correlation values (i.e. the integrative correlation based on just two studies) to λ_1^* is helpful for determining which genes appear to be poorly measured within each study. This can be a very useful result for future analyses of the datasets under investigation.

3.3 Comparative Analyses

3.4 Binary Phenotypes

After choosing genes to include in the analysis based on their reproducibility, the association between gene and phenotype can be compared across two studies. First, however, summary statistics need to be calculated in each study, quantifying the association between gene and phenotype for each of the reproducible genes. We find this preferable to combining the raw data values across studies. In our experience, gene expression data cannot be easily combined just as in many meta-analysis settings. A few reasons for this include that often times the available data has been preprocessed using algorithms so that the original data cannot be recovered (as is the case with the Vant' Veer data in our study), the experimental designs of the studies including sample selection may be different, and the measurement of expression might be different (e.g. log-ratio of expression versus absolute expression). Others have integrated datasets, but using studies with comparable chip measurement (Jiang *et al.*, 2004; Shedden and Taylor, 2004; Morris *et al.*, 2004) unlike our situation.

In the case of a binary phenotype variable, we can detect which genes are associated with phenotype in several ways. Logistic regression can be used, where phenotype is regressed on gene expression for each gene, and the log odds ratio quantifies the relationship. While naturally appealing, because this model implies that gene expression is "predictive" of phenotype, this approach runs into problems when gene expression perfectly predicts phenotype. That is, when there exists a gene whose range of expressions for the two phenotypes do not overlap, the odds ratio is not estimable using a standard logistic regression approach.

As an alternative, a difference in means and a t-ratio approach can be used, where the average expression in the two phenotypes are compared. To provide a common metric across studies, for each gene in each study, we divide the gene expression data by the standard deviation of gene expression data in the reference type before estimating the difference in means (e.g., in the BRCA1 versus sporadic example, we would consider sporadic as our reference phenotype). This transforms the data so that, regardless of platform, effect sizes are comparable across studies, representing the number of standard deviation units that the two phenotypes differ per gene. As a result, our effect size, w_j, for gene j is measured by the differences in the means in the two groups being compared, divided by the standard deviation of the reference group:

$$w_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{SD_{1j}}. \tag{2}$$

We use the w_j value to tell us about the "effect size" for the association between phenotype and gene expression for gene j.

In addition to looking at effect sizes, to assess statistical significance of the observed differences in expression across the two groups, we can use a significance analysis of microarrays (SAM) approach, an extenstion of the t-test. The SAM statistic (d_j) is calculated for each gene j making an adjustment for variance stabilization (Tusher *et al.*, 2001). Note that there are many different statistics that can be chosen: we have used the SAM statistic for illustration. The SAM statistic (d_j) has been widely used for detecting differentially expressed genes in gene expression microarray studies and takes the form

$$d_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{s_j + s_0} \tag{3}$$

where the standardized mean expression in the two phenotypes for gene j are denoted by $\bar{x}_{1j}$ and $\bar{x}_{2j}$, and s_j is a pooled estimate of the standard error of the difference. The way in which d_j differs from the t-statistic is the inclusion of s_0, which is chosen to minimize the coefficient of variation of d_j, where the coefficient of variation of d_j is computed as a function of s_j in moving windows across the range of s_j. When the t-ratio is used in gene expression analyses, many genes tend to be deemed significant that have very small variation in expression across phenotype, but also have small standard errors. The SAM statistic adjusts for that with the addition of s_0 in the denominator, picking up more genes with larger effect sizes and fewer genes with inconsequential differences.

Notice the difference between w_j and d_j: w_j is simply a scaled version of the effect size (similar to a standardized regression coefficient), while d_j is more comparable to a t-statistic, which directly provides information about statistical significance. The major computational difference is that d_j is sample size adjusted because it uses a standard error in the denominator, whereas w_j is not sample size adjusted, using a standard deviation in the denominator. One reason to prefer w_j to d_j is that when considering studies of different sizes, the values of d_j are not directly comparable while the w_j statistics are. After these computations have been made for each of the reproducible genes in each of the studies, we can compare the results using simple scatterplots, including statistical significance (determined by SAM or another approach) as part of the display.

3.5 Survival and Continuous Phenotypes

In many of the oncologic gene expression data that are collected, time to death or relapse is of primary interest. Researchers are intent on finding genes which may be predictive of good or poor prognosis. The approach used above for comparing results for binary phenotypes can be extended. Just as the logistic regression approach was mentioned for looking at binary phenotypes, a Cox proportional hazards model approach can be used for assessing associations between time to death and gene expression, where survival time or relapse time is regressed on gene expression. Unlike the estimation problem that may occur in logistic models, the Cox model will accomodate any gene expression data that shows some variation across samples and has at least several failures.

Log hazard ratios and their statistical significance can be compared across studies. However, the data needs to be standardized first in order that the units of the log hazard ratios are comparable. A logical approach is to subtract the row mean (or median) and then divide by the row standard deviation (or another measure of variability) for each row of the gene expression matrix. Additionally, we can use the SAM-type adjustment: we can find the log hazard ratio (h_j) and its standard error (s_j) for each gene, to calculate the Z-score ($Z_j = h_j/s_j$) for each gene. Then we can estimate the value of s_0 that stabilizes the variability of the Z-scores and recalculate the Z-score such that

$$Z_j = \frac{h_j}{s_j + s_0}. \tag{4}$$

The SAM approach as applied to Cox regression coefficients was used by Bullinger *et al.* (2004) for finding gene expression profiles associated with survival in acute myeloid leukemia patients.

Continuous outcomes can be handled analogously, using linear regression where continuous phenotype is regressed on (standardized) gene expression for each gene. In the case of linear regression, it is suggested that some time be spent ensuring that a linear model is appropriate: by studying the linear regression model on several randomly chosen genes and several genes which show strong association with phenotype, the assumptions of linearity and constant variance can be explored.

3.6 Displaying Results Comparing Effect Sizes Across Studies

An intuitive way to look at the agreement between effect sizes for many genes across studies is to create a scatterplot, with effect sizes for one study on the x-axis and for the other study on the y-axis. Instead of plotting a least squares regression line (i.e. a "best-fit" line determined by minimizing the sum of the squared vertical residuals), we advocate plotting the best-fit line that is found by minimizing the sum of the squared perpendicular residuals. This line can be found by estimating the first principal component ($\{a_1, a_2\}$) associated with the covariance between the effect sizes. The best-fit line has slope $\frac{a_2}{a_1}$ and intersects the point of the mean effect size of both studies. This line is more appropriate than a linear regression line because it is invariant to which study is plotted on the x-axis and which on the y-axis unlike the least squares line.

3.7 Combining Estimates Across Studies

The techniques we propose can generally be considered meta-analytic approaches to analyzing gene expression data. Meta-analysis is a broad area consisting of techniques for analyzing data obtained from different studies. In our methods, we do not actually combine the data across studies, but instead perform comparative analyses, making inferences based on consistency across studies, and estimate combined inferential statistics. A clear understanding of the basic tenets of meta-analysis are critical to make logicial comparisons. General reviews of meta-analysis include Normand (1999); Hedges and Olkin (1985); Cook et al. (1995). Ghosh et al. (2003); Moreau et al. (2003) provide discussions of some of the statistical issues that arise when performing meta-analysis in gene expression arrays, including combining measures from different platforms, complex data structures, multiple comparisons, and duplicate spots within arrays.

Just as in standard meta-analytic approaches, we can go one step further and estimate "pooled" estimates of effect sizes. For example, when looking at overall survival, we could estimate a pooled hazard ratio estimate across studies for each gene in common across a set of studies. Variation in estimates might be assumed to arise from a fixed effects models, such that, for each gene, the log hazard ratios across studies are realizations from a large population of estimates with a common mean, θ_j. On the other hand, a random effects approach could be assumed, where for each gene and each study, we assume a different mean (e.g. θ_{fj}). The latter is a more reasonable approach in general for combining the results from gene expression array experiments due to the many sources of heterogeneity across studies, including the variety of tissues and differences in experimental methods.

However, in the gene microarray setting, we are often in the situation of comparing the results of just a few studies. In our applied example, we have only two microarry studies of breast cancer with information on BRCA1 status. As a result, a random effects approach is not feasible and alternative methods for combining effects need to be utilized. We outline an approach for combining effects across two studies which can be extended for $k > 2$ studies. For $k >> 2$, we recommend considering a random effects approach as described above.

To combine results, consider first the situation where the studies to be combined have equal sample sizes and the average effect size in each study is 0. In this case, one logical approach would be to simply average the effect sizes in the two studies. Another approach would be to use the best-fit line defined in the previous section, project each point to the best-fit line, and use the distance from the origin to the projected point (divided by $\sqrt{2}$ to preserve the metric) to represent the combined effect size. Using the definitions of a_1 and a_2 from the section 3.6, to combine effect size w_1 from study 1 and w_2 from study 2, we would estimate the combined effect w_c as $w_c = \frac{a_1 w_1 + a_2 w_2}{a_1 + a_2}$. This is analogous to using the "fitted" value from linear regression as a best estimate of y for a given value of x. A major difference with this approach is that x and y are treated symmetrically, unlike a linear regression approach. In linear regression, a horizontal line (with slope=0) indicates no linear association between x and y. When using our approach, with the line fitted based on perpendicular residuals, either a horizontal line with slope of 0 or a vertical line with slope of ∞ indicate no linear association between x and y.

In practice, we need to make some modifications to the above approach because (a) sample sizes are generally not the same and we would like to give more weight to estimates from studies with larger sample sizes, and (b) the average effect size in each study is generally not equal to zero (although they are often very close to zero). Our method for combining results across two studies is described in box 1 below:

1 Define the two vectors w_1 and w_2 to be the vectors representing effect sizes in the two studies.

2 Center w_1 and w_2 by their means.

3 Fit principal components to w_1 and w_2 using covariance, saving a_1 and a_2, the loadings of the first principal component.

4 Estimate the combined effect: $w_c = \frac{a_1 \sqrt{n_1} w_1 + a_2 \sqrt{(n_2)} w_2}{a_1 \sqrt{n_1} + a_2 \sqrt{n_2}}$.

Step 2 ensures that the fitted line intersects the origin. Step 3 uses covariance (instead of correlation) which uses variation in estimates to determine which study should be weighted more heavily: studies with more variation in estimates will tend to be favored. Step 4 does two things: (a) it sample size adjusts the fitted line, and (b) it rescales the metric to preserve the original scale of the effect sizes.

The vector w_c can now be used for determining which genes are associated with phenotype using the combination of information from the two studies. This can be done using a permutation approach: for each study, the gene expression data is randomly permuted by row, effect sizes are recalculated, and the combined estimates are then recalculated. The distribution of combined effect sizes is compared to the combined effect sizes based on the the permutation approach.

4 Results

The three studies described in section 2 (referred to as Hedenfalk, Huang, and vant' Veer) were analyzed for reproducibility, and we assessed the comparability of results of the Hedenfalk and vant' Veer in regards to the genes associated with BRCA1 status.

4.1 Merging datasets

The bioconductor library `MergeMaid` was used for merging the three datasets. The `MergeMaid` library combines phenotype and gene expression data from multiple studies, averages multiple occurences of genes within a study before merging, keeps track of which genes are in common across studies, and sorts genes consistently across studies. We found that the Hedenfalk and vant' Veer study had 1121 genes in common, Hedenfalk and Huang had 1668 in common, and vant' Veer and Huang had 5108 in common. Across all three studies, there were 941 common genes.

4.2 Reproducibility

Reproducibility was assessed for each pair of studies by calculating the pairwise and the overall integrative correlations. The distributions of the pairwise integrative correlations and their null distributions (as determined by permutation) are shown in figures 2A, 2B, and 2C. The s_α cutoff was set to 1%, meaning that 1% of genes deemed reproducible will not be reproducible. We can see the effect that sample size has on these null distributions: the null integrative distribution for the Huang (n=89) vs. vant' Veer (n=98) study is very narrow as compared to the null distributions involving the Hedenfalk study (n=22). From these plots, approximately 80% of the genes are reproducible between vant' Veer and Huang, 60% between vant' Veer and Hedenfalk, and 60% between Huang and Hedenfalk.

The overall integrative correlation (i.e. λ_1^*) is shown in figure 2D. Notice a small bump in both distributions to the left of the mode. This can be explained as representing genes that have high correlation between two of the three studies, but are such that the third study has negative correlation with the other two studies. In these cases, it is likely that the gene is incorrectly measured in the discordant study. The overall integrative correlation for the three studies is less conservative than the pairwise integrative correlations: approximately 85% of genes are defined as reproducible (797 of the 941 common genes). Note that in figure 2D, there is flattening of the null distribution at a value for the integrative correlation of 0. This is due to the nature of the principal components analysis. With three variables included in the analysis, the chance that the pairwise correlations between all three variables is zero is virtually zero. As such, the probability of obtaining an integrative correlation of 0 when more than two studies are being compared is essentially 0.

In Figure 3, we compare the behavior of the overall integrative correlation (λ_1^*) with the average of the pairwise correlations. For both λ_1^* and the average pairwise correlations, permutation test were performed, where gene reproducibility was defined by the upper 5% of the null distribution. That is, any gene with reproducibility scores larger than the 95th quantile of the null distribution is considered to be reproducible. Note that this cutoff is somewhat arbitrary–other cutoffs can be chosen depending on user preference. The integrative correlation is shown on the x-axis, the average pairwise correlation on the y-axis, and the points are coded to identify which genes are deemed reproducible by each method. Notice the linearity in the upper right of the figure. However, in the upper and lower left, there are marked departures from linearity. Overall, there is good agreement: the majority of the genes are blue (reproducible by both methods), or green (unreproducible by both methods). In this analysis, there are no genes which are found to be reproducible by integrative correlation but not by the average correlation , but there are a handful of genes (shown in red) which are reproducible by the average correlation but not by the integrative correlation, λ_1^*. These are the genes that we saw in the little bump in figure 2D: they are genes which agree strongly in two studies, but not in the third. Genes such as these are interesting because they should be considered in any pairwise comparisons between the studies in agreement. However, when trying to obtain genes which are consistently measured in all three studies, these genes show relatively strong evidence of discordance.

4.3 Comparing and combining gene associations for BRCA1

In order to investigate the comparability of results across the Hedenfalk and vant' Veer studies, we restrict our attention to genes which were determined to be reproducible across the two studies, as defined above. For each of the two studies, effect sizes, w_j, are estimated for the reproducible genes and plotted versus each other, with the best-fit line included, plotted using a solid line (figure 4A). Recall that the best fit line is determined by minimizing the sum of the squared perpendicular residuals and its interpretation should not be confused with the interpretation of the standard least squares regression line. In this case, the slope represents the ratio of variability explained by the variable measured on the y-axis as compared to the one on the x-axis. Hence, a slope of 0 or ∞ means that there is no association between the effects in the two studies. In figure 4A, the slope is close to 1, suggesting comparable explanatory power for the two studies. As a contrast, the effect sizes and the best fit line (in solid) is shown for the unreproducible genes in figure 4B, which is close to vertical, suggesting almost no association between the two studies among the unreproducible genes.

The dotted lines on figures 4A and 4B represent the sample size adjusted combined effect size. Notice that both of these are shifted downward toward the x-axis, favoring the estimates in the vant' Veer study.

We next combined scores, weighting by the principal components loadings and study sample sizes, as described in box 1. The distribution of combined effect sizes is shown in figure 5, along with its null distribution. Genes were considered significantly associated with BRCA1 status if they were beyond the 2.5th or the 97.5th percentile of the null distribution. Out of the XXX reproducible genes analyzed, a total of 226 genes were found to be significantly associated with BRCA1 using the combined results of the two studies. An additional quantity of interest is the combined d_j value, which is essentially the SAM statistic averaged across studies. This is shown as compared with our combined effect size in figure 6. As can be clearly seen, these two statistics are highly correlated, although the ranges are rather different due to the sample size adjustment in d_j.

5 Discussion

Given the multitude of gene expression studies that are currently available, it is clearly of interest to make inferences based on their collective evidence. Additionally, given the concerns that gene expression data is often poorly reproducible, it is important to be able to confirm results found in one study by looking at other similar studies. When looking at a single study, we usually do not have the ability to statistically evaluate the reliability of gene measurement and instead usually rely on gene-by-gene validations using methods such as RT-PCR among a small subset of genes found to be interesting. When multiple studies are available, statistical cross-study validation offers an effective, high-throughput alternative that should be exploited more routinely in genomic analyses.

In this article we describe simple tools for both the comparison of results and the combination of association measures across two or more studies. We focused on studies of gene expression using micorarrays, because they are common, expensive, and controversy exists on their validity. Our analysis plans can be extended to other high throughput techniques for measuring the transcriptome and the proteome. We focused on class comparison issues, though reproducibility filters would be advisable in class prediction and class discovery settings as well.

In practice, we suggest that reproducibility should always be considered before comparing results across studies. By filtering out genes that appear to be mismeasured or incorrectly linked, we avoid making comparisons that are not biologically relevant, and we will likely substantially reduce both the number of discordant findings and the number of falsely concordant findings. This allows us, to some extent, to separate the technology-specific and study-specific variation from the biological variation: for example, while strong biological variation can lead to small integrative correlations, it is far less likely to lead to negative integrative correlations. The latter are more likely the result of inaccurate cross-linkage or strong differences in the hybridization behavior of the sequences used as probes in different platforms.

A novel addition in this paper is the extension of the reproducibility assessment to more than two studies. In the case of two studies, calculating a measure to describe the overall agreement between two studies is straightforward. However, when looking at more than two studies, finding a one-number summary of the overall agreement led us to principal components analysis, and a statistic based on the value of the eigenvalue associated with the first principal component, with sign based on the signs of the elements of the eigenvector. In practice, this approach has been useful for finding genes which are measured consistently across all studies and also for finding genes which are measured consistently in only two of three studies, suggesting mismeasurement in the third study.

While individual studies may not provide convincing evidence of gene-phenotype relationships, collectively analyzing gene expression datasets with the same phenotypes may provide substantially more information. By looking at a collection of datasets, we can assess which genes appear to be reproducible across studies and which show consistent trends in their association with phenotype. Many gene expression datasets are publicly available so that now there are collections of datasets, each with the goal of finding associations between phenotype and genes.

To date, while many authors have compared platforms using the same samples in controlled experiments (Kuo *et al.*, 2002; Yuen *et al.*, 2002; Barczak *et al.*, 2003; Culhane *et al.*, 2003; Mecham *et al.*, 2004; Kothapalli *et al.*, 2002) there have been few researchers who have published approaches for combining information across samples and platforms from different studies. Grigoryev *et al.* (2004) combined gene expression across different Affymetrix chip types (U47A, U34A, U95A, and U133A) for different animal species. Their approach included matching genes using orthologs (i.e. matching genes across species) and directly combining normalized gene expression data before analysis to obtain overall measures of "statistical significance" of orthologs. While this may be an appropriate method across matched Affymetrix chip technologies, the applicability of this approach for different platforms is unlikely. Rhodes *et al.* (2002) meta-analyzed four prostate cancer gene expression studies, comparing prostate tumor tissue to benign prostate tissue. For each gene, they first computed the p-value to the test for the null hypothesis of no difference across the two groups. They then ranked these p-values within each study and combined the resulting ranks. Combining p-values according to their ranks is consistent in this case with combining t-statistics according to their ranks. Test statistics and their significance values depend both on information about the magnitude of the fold-change across the two classes considered, and on information about the precision with which this fold change can be measured in the experiment at hand. The latter reflects within-class variability for a gene, but also the study sample size. Therefore, it would not have been appropriate to combine p-values or t-statistics directly across studies (although their rank can be combined). Because of this limitation, most of the combination approaches developed in medicine (Hedges and Olkin, 1985; Hasselblad and McCrory, 1995) are based on standardized effects (SE). In the genomic context a SE for a two-class comparison could be defined as the fold-change divided by the within-class standard deviation. SEs have the added advantage of allowing for direct combination across studies. This approach is likely to use information more efficiently and also leads to a combined estimate of direct biological interpretation.

Using the approaches that we have developed for comparing the two studies that include BRCA1 information, we were able to show that, in general for the reproducible genes, there was quite good agreeement in terms of magnitude and direction of the associations between genes and BRCA1. And, the results seen for the non-reproducible genes make sense: they have a correlation close to zero. Had these genes been left in the analysis, we would have dampened the estimated association between effect sizes in the two studies. We chose to use the effect size, defined as the difference in mean expression between the BRCA1 and sporadic samples, standardized by the standard deviation in the sporadic samples. However, there are many possible choices for statistics of interest to be compared. SAM statistics can be compared, although the magnitude across studies might vary significantly due to differences in sample sizes.

The R library `MergeMaid` has been developed for performing the analyses that we present, including matching datasets, assessing reproducibility of genes, and performing comparative analyses across datasets and can be accessed freely from `http://astor.som.jhmi.edu/MergeMaid`. These tools can be applied generally to datasets from any gene expression platforms for which there exist a set of overlapping genes.

References

Aach, J., Rindone, W., and Church, G. M. (2000). Systematic management and analysis of yeast gene expression data. *Genome Res*, **10(4)**, 431–45.

Barczak, A., Rodriguez, M., Hanspers, K., Koth, L., Tai, Y., Bolstad, B., Speed, T., and Erle, D. (2003). Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Research*, pages 1775–1785.

Boguski, M. S. and Schuler, G. D. (1995). Establishing a human transcript map. *Nat Genet*, **10**, 369–371.

Bullinger, L., Dohner, K., Bair, E., Frohling, S., Schlenk, R., Tibshirani, R., Dohner, H., and Pollack, J. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine*, **350**, 1605–1615.

Cope, L., Zhong, X., Garrett-Mayer, E. S., and Parmigiani, G. (2004). Mergemaid: R tools for merging and cross-study validation of gene expression data. *Statistical Applications in Genetics and Molecular Biology*, **3**, article 27.

Culhane, A., Perriere, G., and Higgins, D. (2003). Cross-platform comparison and visualization of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4(59)**.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G. M., Bontrop, R. E., and Pääbo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.

Everitt, B. and Dunn, G. (2001). *Applied Multivariate Data Analysis*. Arnold, New York.

Foulkes, W., Wong, N., Brunet, J., Begin, L., Zhang, J., Martinez, J., Rozen, F., Tonin, P., Narod, S., Karp, S., and M., P. (1997). Germ-line brca1 mutation is an adverse prognostic factor in ashkenazi jewish women with breast cancer. *Clinical Cancer Research*, **3**, 2465–2569.

Grigoryev, D., Ma, S., Irizarry, R., Ye, S., Quackenbush, J., and Garcia, J. (2004). Orthologous gene-expression profiling in multi-species models: search for candidate genes. *Genome Biology*, **5(5),R34**.

Hardiman, G. (2002). Microarray technologies—an overview. *Pharmacogenomics*, **3(3)**, 293–7.

Hasselblad, V. and McCrory, D. C. (1995). Meta-analytic tools for medical decision making: a practical guide. *Med Decis Making*, **15**, 81–96.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O., Wilfond, B., Borg, A., and Trent, J. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**, 539–548.

Hedges, L. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, New York.

Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C., Bild, A., Iversen, E., Liao, M., Chen, C., West, M., Nevins, J., and Huang, A. (2003). Gene expression predictors of breast cancer outcomes. *The Lancet*, **361**, 1590–1596.

Irizarry, R. A., M, B. B., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, **31**, e15.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Zhe, Q., Chen, J., Tsai, C., and Zhang, S. (2004). Joint analysis of two microarray gene-expression data sets to select lung cancer adenocarcinoma genes. *BMC Bioinformatics*, **5**.

Johnson, R. and Wichern, D. (1999). *Applied Multivariate Statistical Analysis*. Prentice-Hall, Upper Saddle River, NJ.

Kothapalli, R., Yoder, S., Mane, S., and Loughran, T. (2002). Microarray results: how accurate are they? *BMC Bioinformatics*, **3(22)**.

Kuo, W., Jenssen, T., Butte, A., Ohno-Machado, L., and Kohane, I. (2002). Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.

Mecham, B., Klus, G., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D., Mariani, T., Kohane, I., and Szallasi, Z. (2004). Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nuceic Acids Research*, **32**.

Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., and Hide, W. A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res*, **9**, 1143–1155.

Moller, P., Borg, A., Evans, D., Haites, N., Reis, M., Vasen, H., Anderson, E., Steel, C., Apold, J., Goudie, D., Howell, A., Lalloo, F., Maehle, L., Gregory, H., and Heimdal, K. (2002). Survival in prospectively ascertained familial breast cancer: analysis of a series stratified by tumor characteristics: brca mutations and oopherectomy. *International Journal of Cancer*, **101**, 555–559.

Morris, J., Baggerly, K., Wu, C., and Zhang, L. (2004). Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer. *Methods in Microarray Data Analysis*, **IV**, to appear.

National Center for Biotechnology Information (2003). Ncbi unigene. `http://www.ncbi.nlm.nih.gov/UniGene`.

National Center for Biotechnology Information (2004). Ncbi locuslink. http://www.ncbi.nlm.nih.gov/projects/LocusLink.

Oleksiak, M. F., Churchill, G. A., and Crawford, D. L. (2002). Variation in gene expression within and among natural populations. *Nat Genet*, **32(2)**, 261–266.

Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R., and Gabrielson, E. (2004). Cross-study comparison of gene expression data sets for the molecular classification of lung cancer. *Clinical Cancer Research*, **10**, 2922–2927.

Pierce, L., Straderman, M., Narod, S., Olivetto, I., Eisen, A., Dawson, L., Gaffeny, D., Solin, L., Nixon, A., Garber, J., Berg, C., C., I., Heimann, R., Olopade, O., Haffty, B., and Weber, B. (1998). Effect of radiotherapy after breast-conserving treatment in women with breast cancer and germline brca1/2 mutation. *Journal of Clinical Oncology*, **18**, 3360–3369.

Pritchard, C. C., Hsu, L., Delrow, J., , and Nelson, P. S. (2001). Project normal: Defining normal variance in mouse gene expression. *Proceedings of the National Academy of Science*, **98**, 13266–13271.

Pruitt, K. and Maglott, D. (2001). Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Research*, **29**, 137–140.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. (2001). Tigr gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*, **29**, 159–164.

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

Robson, M., Chappuis, P., Satagopan, J., Wong, N., Boyd, J., Goffin, J., Hudis, C., Roberge, D., Norton, L., Begin, L., Offut, K., and Foulkes, W. (2004). A combined analysis of outcome following breast cancer: differences in survival based on brca1/brca2 mutation status and administration of adjuvant treatment. *Breast Cancer Research*, **6**, R8–R17.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, S., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, **24**, 227–235.

Schena, M. (2000). *Microarray Biochip Technology*. BioTechniques Press, Westborough, MA.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. (2000). A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, **24**, 236–244.

Shedden, K. and Taylor, J. (2004). Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. *Methods of Microarray Data Analysis*, **IV**, to appear.

Southern, E. M. (2001). DNA microarrays. History and overview. *Methods in Molecular Biology*, **170**, 1–15.

Stoppa-Lyonnet, D., Ansquer, Y., Dreyfus, H., Gautier, C., Gauthier-Villars, M., Bourstyn, E., Clough, K., Magdelenat, H., Pouillart, P., Vincent-Salomon, A., and Fourquet, A. (2000). Familial invasive breast cancer: worse outcome related to brca1 mutation. *Journal of Clinical Oncology*, **18**, 4053–4059.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

Van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., and Friend, S. (2002). Gene expression profiling predicts clinical outcome of cancer. *Nature*, **415**, 530–536.

Verhoog, L., Brekelmans, C., Seynaeve, C. van den Bosch, L., Dahmen, G., van Geel, A., Tilanus-Linthorst, M., Bartels, C., Wagner, A., van den Ouweland, A., Devilee, P., Meijers-Heijboer, E., and Klijn, J. (1998). Survival and tumor characteristics of breast-cancer patients with germline mutations of brca1. *Lancet*, **351**, 316–321.

Yuen, T., Wurmbach, E., Pfeffer, R., Ebersole, B., and Sealfon, S. (2002). Accuracy and calibration of commercial oligonucleotide and custom cdna microarrays. *Nucleic Acids Research*, **30**.

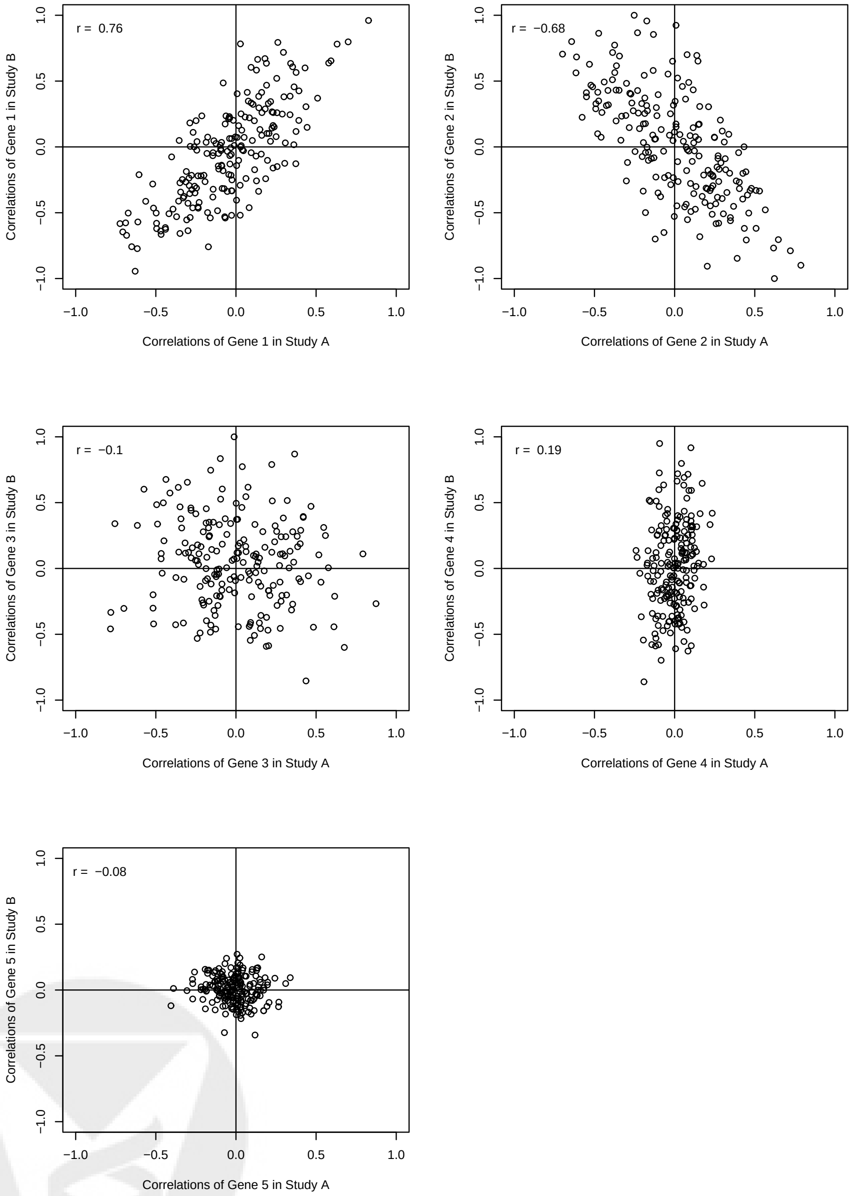

Figure 1: Examples of correlations between correlations of genes.

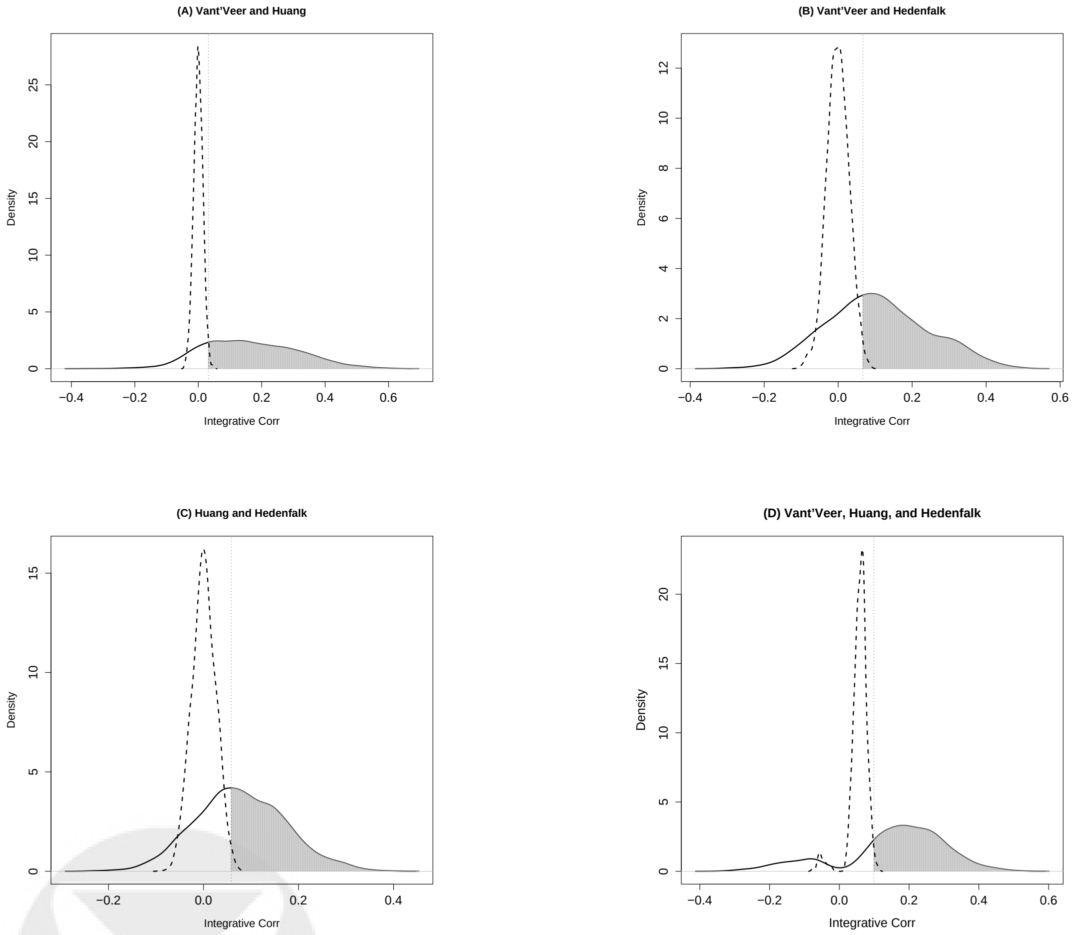

Figure 2: Distribution of integrative correlations across studies. (A) vant' Veer and Huang, (B) vant' Veer and Hedenfalk, (C) Huang and Hedenfalk, (D) All three studies.

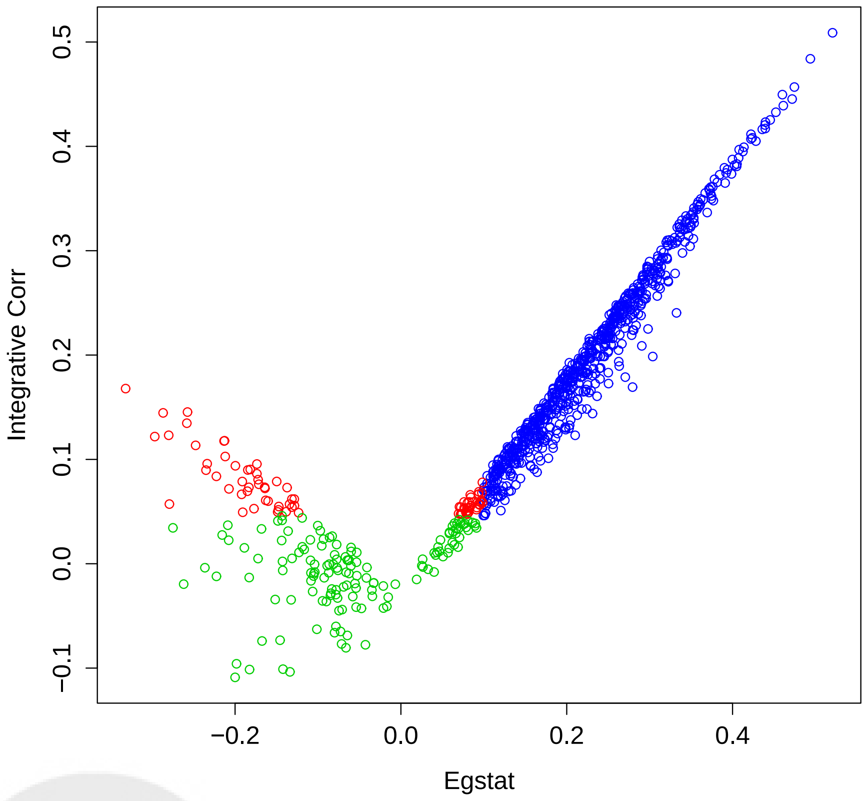

Figure 3: Comparison of using the integrative correlation based on principal components approach (x-axis) and the average of the pairwise correlations (y-axis). Blue points are deemed reproducible by both approaches, green points non-reproducible by both approaches, and red points as reproducible by the average of pairwise correlations but not by the principal components based integrative correlation.

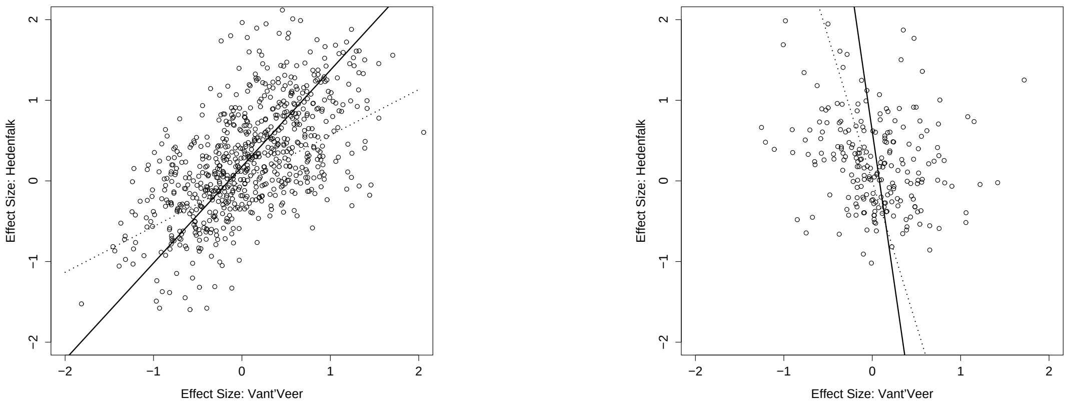

Figure 4: Scatterplots of effect sizes in two studies. Solid line represents the best fit line based on perpendicular residuals. Dotted line is the sample size adjusted line (which favors vant' Veer which is the larger study). Recall that perfectly vertical or horizontal lines indicate no correlation. (A) Reproducible genes, (B) Unreproducible genes.

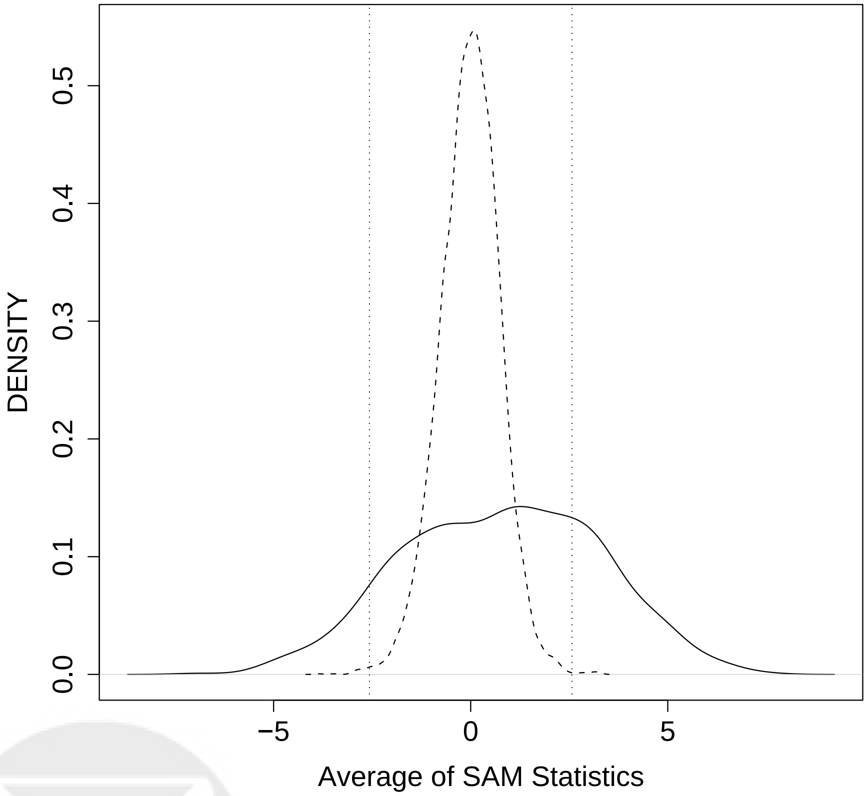

Figure 5: Distribution of combined effect sizes and the null distribution of effect sizes.

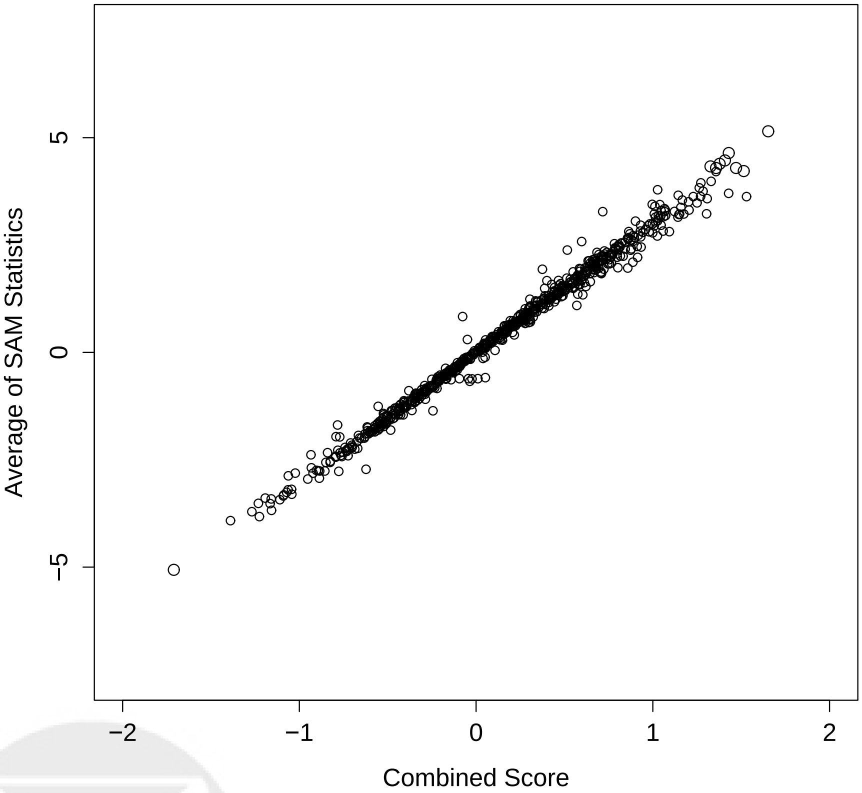

Figure 6: Scatterplot of the combined effect size versus the combined SAM statistic.