

Meta-Analysis Combines Affymetrix Microarray Results Across Laboratories

John R. Stevens¹ & R.W. Doerge^{1,2}

¹ Department of Statistics, ² Department of Agronomy

Purdue University, West Lafayette, IN 47907

Email: jrsteven@stat.purdue.edu, doerge@purdue.edu

Introduction

Microarray Data

Microarrays are small chips that allow for the simultaneous monitoring of expression of thousands of genes. Currently the most popular microarray platforms are cDNA and oligonucleotide arrays. Affymetrix manufactures the GeneChip oligonucleotide microarray, a small chip containing a grid of hundreds of thousands of features or probes. Each gene is represented by 14-20 probe pairs fixed to the array

- A probe pair consists of a perfect match (PM) and mismatch (MM) probe, both of which are 25-mer sequences
- PM probes match a segment of the gene, while MM probes differ at 13th position, with millions of copies of segment in each spot of array
- Tissue sample is prepared and washed across array, with mRNA in sample hybridizing to representative probes, and array is scanned
- Genes expressed in tissue will have an abundance of corresponding mRNA, resulting in higher intensities recorded for related spots on array

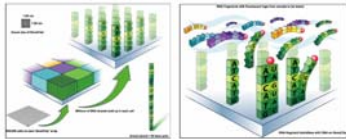


Fig. 1. Summary of Affymetrix technology (images courtesy Affymetrix, www.affymetrix.com).

Affymetrix Algorithms for Differential Expression

Affymetrix has developed statistical algorithms [1] to report a signal log ratio (SLR) estimate comparing the gene expression levels in two different tissue samples to identify differentially expressed genes. The SLR uses PM-MM differences and is related to the fold change (FC):

$$FC = \begin{cases} 2^{SLR} & \text{if } SLR > 0 \\ (2^{-SLR})^{-1} & \text{otherwise} \end{cases}$$

The SLR estimate is derived by use of Tukey's bweight algorithm [2] and is reported by Affymetrix's commercial software MAS 5.0, along with 95 percent confidence bounds.

For gene k, let $\hat{\theta}_k$ be the reported SLR estimate and $\hat{\theta}_k^{(95)}$ be the upper bound of the 95 percent confidence interval for the true SLR θ_k

- The estimated variance for $\hat{\theta}_k$ is $v_k = (\hat{\theta}_k^{(95)} - \hat{\theta}_k) / t_{k,0.975}$, where $t_{k,0.975}$ is the upper .025 critical value of the t distribution with $df_k = \max(7(n_k - 1), 1)$ d.f., where n_k is the number of probe pairs used for gene k

- Under $H_0: \theta_k = 0$, $\sqrt{v_k} / \sqrt{v_k} \square t_{k,0.975}$, so this distribution can be used to test whether the gene is significantly differentially expressed between the two tissues

Motivation: Differing Results from Multiple Laboratories

With microarray technology becoming more prevalent, it is now not unusual to find several different laboratories employing the same microarray technology to identify genes related to the same condition in the same species. Although the experimental specifics are similar, a different list of significant genes may result from the data analysis in each laboratory. We propose a statistically-based meta-analytic approach to microarray analysis for the purpose of systematically combining these results from the different laboratories.

Methods

Meta-Analysis Basics

The general meta-analytic framework [3] assumes that a measurable relationship exists between certain quantities of interest, and n independent studies have produced an estimate of the relationship. Certain kinds of standardized estimates are called effect size estimates and can be combined across studies.

Effect size estimates must satisfy three conditions [4]:

- Comparability: Address the same measure or quantity in each study
- Standardization: Be standardized to the same scale $= (\bar{X}_i - \bar{X}_j) / S_j$
- Variability: Include measure of variability

Types of effect size estimates include standardized difference estimates, standardized relation estimates, and measures of significance [5]. Fixed effects, random effects, and Bayesian models can be applied to effect size results [6]. Prior applications of meta-analysis to microarray results [7-9] have used only P-values or have sought to combine results across platforms without accounting for fundamental technological differences.

Effect Sizes from Microarray Results

The results of an Affymetrix microarray experiment are naturally suited for a meta-analysis because the SLR estimates satisfy the conditions of effect size estimates:

- Comparability: The same conditions are measured in each study
- Standardization: The same algorithm standardizes all SLR estimates
- Variability: The variance of the estimates can be derived from the reported confidence intervals in each study

Fixed Effects Meta-Analysis

Assume that n independent studies have provided SLR estimates comparing the same two conditions, with $\hat{\theta}_k$ representing the SLR estimate for gene k in laboratory i, and $v_{i,k}$ the corresponding variance estimate, $i=1, \dots, n$. Then the most general meta-analytic approach proceeds as follows:

- Assume the model

$$\hat{\theta}_k = \theta_k + \epsilon_{i,k} + \delta_{i,k} + \sigma_{i,k}$$

where θ_k is the true fixed underlying SLR, and $\sigma_{i,k}$ is normal sampling error

- The assumption $H_0^k: \theta_k = \dots = \theta_{i,k}$ is referred to as the homogeneity assumption and can be interpreted as assuming that any differences between laboratories are due only to sampling error

- Using the weights $w_{i,k} = 1/v_{i,k}$, the fixed effects meta-analysis estimates the common value parameter $\hat{\theta}_k$ and its variance by

$$\hat{\theta}_k = \frac{\sum w_{i,k} \hat{\theta}_{i,k}}{\sum w_{i,k}}, \quad \hat{v}_k = \frac{1}{\sum w_{i,k}}$$

- Test the homogeneity assumption H_0^k by use of the test statistic

$$Q_k = \sum w_{i,k} (\hat{\theta}_{i,k} - \hat{\theta}_k)^2$$

- Under $H_0^k: Q_k \square \chi_{n-1}^2$

- The test of significance considering $H_0^k: \theta_k = 0$ can be considered by use of the test statistic

$$Z_k = \hat{\theta}_k / \sqrt{\hat{v}_k}$$

- Under $H_0^k: Z_k \square N(0,1)$, and so the significance P-value $P_{2,k}$ for gene k is the value such that $|Z_k|$ is the upper $P_{2,k} / 2$ critical value of the $N(0,1)$ distribution.

Significant differences in gene expression patterns between labs may be detected by testing the homogeneity assumption, and significant differential expression can be detected by the test of significance.

Random Effects Meta-Analysis

Similar to the fixed effects model, but instead of homogeneity, assume

$$\hat{\theta}_{i,k} = \theta_k + \epsilon_{i,k} + \delta_{i,k} + \sigma_{i,k}$$

where $\delta_{i,k}$ is the random deviation of θ_k from $\hat{\theta}_k$, with $\text{Var}(\delta_{i,k}) = \Delta_k^2$. Note that Δ_k^2 is a measure of the amount of variation between labs for a given gene, and that $H_0^k: \Delta_k^2 = 0$. Then proceed as follows:

- Estimate Δ_k^2 with

$$\hat{\Delta}_k^2 = \max(0, Q_k / (n - 1) / (\sum w_{i,k} - \sum w_{i,k} / \sum w_{i,k}))$$

using $w_{i,k}$ and Q_k as in the fixed effects model

- Compute adjusted weights

$$w_{i,k} = 1 / (v_{i,k} + \hat{\Delta}_k^2)$$

- The estimate of the "population mean" SLR and its variance is given by

$$\hat{\theta}_k^{(re)} = \frac{\sum w_{i,k} \hat{\theta}_{i,k}}{\sum w_{i,k}}, \quad \hat{v}_k^{(re)} = \frac{1}{\sum w_{i,k}}$$

- Perform the test of significance as in the fixed effects model, considering $H_0^k: \theta_k = 0$ by use of the test statistic

$$Z_k^{(re)} = \hat{\theta}_k^{(re)} / \sqrt{\hat{v}_k^{(re)}}$$

Under $H_0^k: Z_k^{(re)} \square N(0,1)$.

Simulation Example

In order to evaluate the usefulness of this meta-analytic approach for microarray data, a simulation study was conducted. "Raw" data for probe l of gene k under condition j in lab i were generated from the following model:

$$MM_{i,j,k} = MM_{i,j} \square \text{Gamma}(\alpha, \beta)$$

$$P_k \square \text{Bernoulli}(p)$$

$$\log(PM_{i,j,k} - MM_{i,j,k}) = \mu + L_i + G_k + P(G)_{i,j,k} + LG_{i,k} + \rho_k (T_i + LT_i + TG_{i,k} + LTG_{i,k} + TP(G)_{i,j,k}) + \epsilon_{i,j,k}$$

All terms in the linear model were random normal effects with mean zero, and the error variance was allowed to differ between labs. The parameters μ , α , and β and the variances of the random effects terms in the linear model can be adjusted to introduce various sources of variability in the "observed" simulated data. Six labs were simulated on the RN_U34 Affymetrix chip, and the data are summarized in Fig. 2 below.

Future research interests will introduce other features to this simulation model, including accounting for "known" relationships between genes.

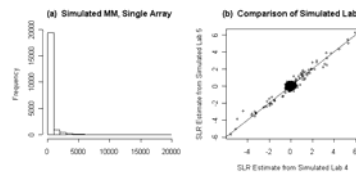


Fig. 2. Plots representing simulated data on the RN_U34 Affymetrix microarray chip. (a) Simulation model parameters were selected such that the distribution of MM (and PM) intensities resembled the distributions found in real data. (b) Simulated results from multiple labs were similar but still exhibited obvious differences.

Results

Both fixed effects and random effects models were applied to the simulated data, and Fig. 3 summarizes the results of the meta-analysis. In all significance tests, the false discovery rate was controlled at 0.05. Of the 1322 genes on the chip, the fixed effects model declared 169 significantly differentially expressed, while the random effects model declared 82 significantly differentially expressed. The difference in the results of the two models is due to down-weighting (see Fig. 3a,b). The random effects model declared 19 genes significantly differentially expressed that were not declared significant by any of the six simulated labs, and these are called Integration-Driven Discoveries (IDDs; see Fig. 3c). IDDs tend to be observed when small but consistent effect sizes are combined. The same model declared 6 genes not significantly differentially expressed that were declared significant by more than one of the simulated labs, and these can be called Integration-Driven Revisions (IDRs; see Fig. 3d). IDRs tend to be observed when large but inconsistent effect sizes are combined. This simulation study demonstrates the ability of a meta-analysis to systematically combine results from multiple laboratories. Future work will incorporate real data from multiple laboratories and discuss the Bayesian model and the use of covariates to account for known differences between labs.

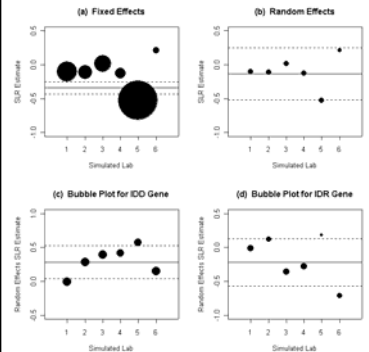


Fig. 3. Summary of Meta-Analysis Results. The bubble areas in these bubble plots are proportional to the weights used in the meta-analysis, and the final estimate and its 95 percent confidence interval (appropriately adjusted to control the FDR) are represented by horizontal lines. (a,b) Estimates for the same gene by the fixed-effects and random-effects models. (c) Estimate for one of nineteen IDD genes. (d) Estimate for one of six IDR genes.

References

1. Statistical Algorithms Description Document. (2002) Affymetrix.
2. Hoaglin, Mosteller, Tukey. (1983) Understanding Robust and Exploratory Data Analysis.
3. Hedges, Olkin. (1985) Statistical Methods for Meta-Analysis.
4. Glass, (1978) Review of Research in Education. 5, 351-379.
5. Cooper, Hedges. (1994) The Handbook of Research Synthesis.
6. Stang, Berry. (2000) Meta-Analysis in Medicine and Health Policy.
7. Rhodes, Barnette, Rubin, Ghosh, Chinnayyan. (2002) Cancer Research. 62, 4427-4433.
8. Cho, Yu, Kim, Yoo. (2003) Bioinformatics. 19(Suppl. 1), 84-90.
9. Moreau, Aerts, Moor, Strooper, Dabrowski. (2003) Trends in Genetics. 19(10).