

Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture


2005 - 17th Annual Conference Proceedings

A BAYESIAN AND COVARIATE APPROACH TO COMBINE RESULTS FROM MULTIPLE MICROARRAY STUDIES

John R. Stevens

R. W. Doerge

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

 Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Stevens, John R. and Doerge, R. W. (2005). "A BAYESIAN AND COVARIATE APPROACH TO COMBINE RESULTS FROM MULTIPLE MICROARRAY STUDIES," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1141>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

A BAYESIAN AND COVARIATE APPROACH TO COMBINE RESULTS FROM MULTIPLE MICROARRAY STUDIES

John R. Stevens^{1,2} and R.W. Doerge^{1,3}

¹ Department of Statistics, Purdue University, West Lafayette, IN

² Current address: Dept. of Mathematics and Statistics, Utah State University, Logan, UT

³ Department of Agronomy, Purdue University, West Lafayette, IN

Abstract

The growing popularity of microarray technology for testing changes in gene expression has resulted in multiple laboratories independently seeking to identify genes related to the same disease in the same organism. Despite the uniform nature of the technology, chance variation and fundamental differences between laboratories can result in considerable disagreement between the lists of significant candidate genes from each laboratory. By adjusting for known differences between laboratories through the use of covariates and employing a Bayesian framework to effectively account for between-laboratory variability, the results of multiple similar studies can be systematically combined via a meta-analysis. Meta-analyses yield additional information not available from any single study and provide a clearer understanding of each gene's true relationship to the disease of interest. A simulation model based on the Barley Affymetrix GeneChip microarray demonstrates the utility of this approach. Further illustration is provided from a mouse model for multiple sclerosis.

Keywords: microarray, meta-analysis, hierarchical Bayes linear model

1 Introduction

The use of microarrays (Lockhart et al. 1996; Craig et al. 2003) has become increasingly common in agricultural research, as evidenced by the recent growth of public repositories of microarray data such as BarleyBase, a “community resource for cereal microarrays” (www.barleybase.org) (Shen et al. 2005). Microarray technology allows researchers to better understand the expression (transcript) levels of individual genes under different conditions by taking advantage of the “Central Dogma” (Crick 1970) of molecular biology. A gene's expression level in a tissue sample can be measured by the level of corresponding mRNA abundance in the sample. Individual laboratories can estimate the level of gene expression in specific tissue samples and compare these expression levels for the purpose of estimating the degree and significance of differential expression. Differences in experimental specifics and chance variation can affect the results from different laboratories even though they are studying the same disease in the same organism.

1.1 Microarray Technology

Affymetrix (www.affymetrix.com) produces the GeneChip microarray which is a small chip containing a grid of hundreds of thousands of features or spots. Each feature, also called a probe, represents a segment of a gene in a species. The length of these segments are typically 25 nucleotides, where nucleotides are the building blocks of DNA (Campbell 1995). There are four possible nucleotides at each position in DNA: cytosine (C), thymine (T), adenine (A), and guanine (G). A single gene may be thousands of nucleotides long. Each probe consists of a specific ordered sequence of 25 nucleotides that is referred to as a 25-mer oligonucleotide sequence, or a 25-mer (“oligo” means “few”). Features come in perfect-match and mismatch pairs, called probe pairs. Each perfect-match (PM) probe contains millions of copies of the same 25-mer segment of a specific gene fixed to the chip, while the corresponding mismatch (MM) probe contains millions of copies of a 25-mer oligonucleotide sequence identical to the perfect-match sequence except for a single substitution at position 13 (i.e., the middle position).

In order to determine which genes are expressed or transcribed in a given sample of tissue, the sample is prepared so that mRNA sequences in the tissue are labelled with fluorescent tags. This prepared sample is then washed across the microarray. Genes that are indeed expressed or transcribed will be represented by mRNA in the tissue sample. This mRNA will hybridize (i.e., find its match) to its corresponding feature(s) on the microarray. The motivation for the use of mismatch (MM) probes is to allow for adjustments for cross-hybridization, i.e., to control for “hybridization specificity” (Lockhart et al. 1996). Once the array has been hybridized, it is scanned, and those features with hybridized mRNA will fluoresce at an intensity proportional to the mRNA abundance level. The image of the scanned array is recorded, and intensities for individual features are used as raw data in subsequent analyses. The raw data consists of the intensities of the individual spots (features) on the array. These intensities come in pairs for each probe, with PM denoting the intensity of a perfect-match probe and MM denoting the intensity of the corresponding mismatch probe.

MAS 5.0 is the commercial statistical analysis software (Affymetrix 2001; Affymetrix 2002) available from Affymetrix. Individual spot intensities are utilized for the purpose of estimating the true expression levels of individual genes in single samples. These estimated expression levels for each gene are then compared between different samples (or treatment conditions). For example, two separate hybridizations, one for a control (or healthy) condition and one for an experimental (or diseased) condition provide respective estimates for each gene that is represented on the array. A “signal log ratio” (SLR) with 95 percent confidence bounds is reported. The signal log ratio is the signed \log_2 of the signed fold change (FC) familiar to biologists (Affymetrix 2002). That is, $FC = 2^{SLR}$ if $SLR \geq 0$ and $FC = (-1)2^{-SLR}$ if $SLR < 0$. The fold change is a measure of how much a gene’s expression level changes from one condition (e.g., control) to another (e.g., experimental). The method used by Affymetrix to compute the SLR is based on Tukey’s biweight algorithm (Hoaglin et al. 1983). In addition to the SLR measure of differential expression reported by the MAS

5.0 software, a variance estimate of the SLR is obtainable from the MAS 5.0 output (Stevens and Doerge 2005a). In turn, these quantities can be used to test for significant differential expression (between arrays) using a t-statistic (Affymetrix 2001; Stevens and Doerge 2005a).

When multiple laboratories employ this same microarray technology to estimate the magnitude of differential expression, their results will vary due to chance variation and fundamental differences between experimental conditions. A meta-analytic approach can be used to combine results (i.e., the SLR estimates) across experiments in a well-structured manner for the purpose of arriving at a clearer understanding of each gene’s relationship to the condition of interest. Here, we present a Bayesian meta-analysis framework using two examples - a simulated data set based on barley microarray data, and an actual mouse data set.

1.2 Simulation Data

We simulate an example data set that assumes multiple laboratories are studying differential expression between healthy and diseased barley using the barley1 Affymetrix microarray. We also assume that there are two different barley strains that a laboratory could use. The following model can be used to simulate microarray data from these multiple laboratories where there is a single covariate (strain) that differs across laboratories:

$$\begin{aligned}
 Y_{itkl} = & \mu + L_i + G_k + P(G)_{(k)l} + LG_{ik} + \rho_k^{(C)} C_{i(m)} \\
 & + \rho_k [T_t + LT_{it} + TG_{tk} + LTG_{itk} + TP(G)_{t(k)l} \\
 & + \rho_k^{(C)} (CT_{it(m)} + CTG_{itk(m)})] + \epsilon_{(itk)l}.
 \end{aligned} \tag{1}$$

Here Y_{itkl} is the \log_2 of the $PM - MM$ difference for probe l of gene k under treatment t in lab i with covariate m . Six labs were simulated with each lab using the same two treatments (healthy and diseased). The term $\rho_k \sim \text{Bernoulli}(p)$ is 1 if gene k is differentially expressed between conditions $t = 1$ and $t = 2$, and is 0 otherwise. The parameter p corresponds to the percentage of genes that are differentially expressed, with higher values resulting in more differentially expressed genes. Similarly, $\rho_k^{(C)} \sim \text{Bernoulli}(p^{(C)})$ is an indicator variable for whether the expression level of gene k is affected by the covariate level (strain). In this model, L_i is the effect of lab i , T_t is the effect of treatment t , G_k is the effect of gene k , $P(G)_{(k)l}$ is the effect of probe l of gene k , $C_{i(m)}$ is the effect of covariate (strain) level m in lab i , $\epsilon_{(itk)l}$ is a random error term, and the remaining terms are the respective interaction effects. To introduce more between-lab variability, the error variance was allowed to be different in each lab: $\epsilon_{(itk)l} \sim N(0, \sigma_i^2)$ for the error terms in lab i . Each term (X) in the model is assumed to be a random effect from a $N(0, \sigma_X^2)$ distribution, except for the constant μ , the fixed effect T_t , ρ_k , and $\rho_k^{(C)}$. The parameters can be adjusted to introduce varying sources of variability in the “observed” simulated data.

Figure 1 presents a comparison of the SLR estimates from representative of the simulated labs. If the results from the different labs were the same, then the SLR estimates would be

equivalent. However, due to chance variation the labs produce different results, even when they share a common covariate level (barley strain). When there is a covariate difference between simulated labs, then the results are even more disparate. The meta-analytic approach presented later combines the results from each individual laboratory analysis using a Bayesian framework that accounts for such known covariate differences.

1.3 Real Data

One of the many diseases whose genetic basis has been studied in mouse using microarray technology is experimental autoimmune encephalomyelitis (EAE). Mouse is the model organism for most human health studies. EAE is a condition similar to multiple sclerosis in humans, with inflammation in the central nervous system resulting in damage to the myelin covering the nerve fibers (Ibrahim et al. 2001). The effect of this damage is impaired motor skills.

Several laboratories throughout the world have used the Affymetrix technology to study EAE in mouse and have reported their findings (Ibrahim et al. 2001; Carmody et al. 2002; Matejuk et al. 2002; Mix et al. 2002; Matejuk et al. 2003). Some of these laboratories used different strains of mouse and studied gene expression in particular tissue sites. Table 1 summarizes the experimental specifics for these laboratories. Figure 2 compares the SLR estimates from different laboratories. As with the simulated data, the experiments with greater covariate (strain and tissue) differences tend to produce more disparate results. The meta-analytic approach presented here accounts for these differences and combines results in a well-structured manner.

2 Methods

The term “analysis” is used to describe the quantitative approaches that are used to draw useful information from raw data. The term “meta-analysis” (Glass 1976) refers to the approaches used to draw useful information from the results of previous analyses. For the current application, meta-analytic approaches can be employed to combine the results (SLR measures of differential expression) from several different labs without having access to the original raw (probe-level) data that yielded the initial results. Such approaches have particular utility with the results of Affymetrix GeneChip microarrays and other fabricated arrays, because the results are given in a uniform format that readily lends itself to comparison between labs and combination across labs. Previous applications of meta-analysis to microarray data (Rhodes et al. 2002; Choi et al. 2003; Moreau et al. 2003; Parmigiani et al. 2004; Rhodes et al. 2004) have focused on combining significance results such as P-values and on combining results across technologies without fully accounting for technological differences. In particular, a previous Bayesian approach (Choi et al. 2003) combined results

from multiple microarray studies that used different microarray platforms. More recent work (Stevens and Doerge 2005a; Stevens and Doerge 2005b) focuses on combining Affymetrix results across laboratories in fixed and random effects meta-analysis models. The statistical method proposed here can be used to account for fundamental differences in experimental conditions by the use covariate information via a Bayesian meta-analysis of results across laboratories.

2.1 Hierarchical Bayes Linear Model

Consider a single gene k , and its corresponding SLR estimate $\tilde{\theta}_{i,k}$ as is available from experiment i (of N_k experiments) along with a variance estimate $v_{i,k}$. Let X_k be the “design matrix” for gene k with N_k rows and m_k columns corresponding to an intercept term and the covariates that are available for gene k . Notice that X_k , N_k , and m_k depend on the gene k since some genes are not represented in every experiment and may have multiple representations in a single experiment. Therefore, X_k is an $N_k \times m_k$ matrix of rank m_k . A Bayesian framework combines results for gene k across experiments in the following way:

$$\begin{aligned}
 \tilde{\theta}_k &= X\beta_k + \delta_k + \epsilon_k \\
 &= \theta_k + \epsilon_k \\
 \delta_k &\sim N(0, \sigma_k^2 I) \\
 \epsilon_k &\sim N(0, V_k) \\
 V_k &= \text{diag}(v_{k,1}, \dots, v_{k,N_k}) \\
 \theta_k | \beta_k, \sigma_k &\sim N(X\beta_k, \sigma_k^2 I) \\
 \beta_k | \sigma_k &\sim N(b_k, D_k) \\
 D_k &= \text{diag}(d_{k,1}^2, \dots, d_{k,m_k}^2) \\
 \sigma_k &\sim \pi(\sigma_k).
 \end{aligned} \tag{2}$$

Here $\tilde{\theta}_k$ is the vector of N_k SLR estimates for gene k , θ_k is the vector of the underlying effect sizes being estimated in each experiment, δ_k is the vector of random deviation of $X_k\beta_k$ from θ_k , and ϵ_k is the vector of sampling error in each lab (Cooper and Hedges 1994). Then δ_k represents between-experiment error, and ϵ_k represents within-experiment error. The level of inter-experiment variability for gene k is σ_k^2 , and $\pi(\sigma_k)$ is the prior distribution of σ_k . This is a hierarchical Bayes linear model, or HBLM (DuMouchel 1994; DuMouchel and Normand 2000). Of particular interest in this model for gene k are the parameter (or covariate) effects β_k .

In the absence of prior knowledge about β_k (i.e., about the true effects of the covariates under consideration), let $d_{k,i} \rightarrow \infty$. This effectively places a diffuse prior on β_k (DuMouchel and Normand 2000). The elements of β_k can be assumed independent; the more general case can be reduced to this independent case (DuMouchel and Normand 2000).

When the non-intercept columns of X_k are centered about their means, the intercept term $\beta_{k,0}$ is referred to as the population mean effect size because it represents the predicted effect size value when each of the covariates are equal to their estimated population means (Cooper and Hedges 1994). In the context of microarray data, centering the columns of X_k allows $\beta_{k,0}$ to be interpreted as the underlying degree of differential expression (the SLR) for gene k after accounting for the effects of the covariates. The magnitude of this SLR ($\beta_{k,0}$) is of particular interest when making statements about the degree of differential expression for a gene (k).

The components of this model (Equation 2) require some additional interpretation. The i^{th} element of the vector $X_k\beta_k$ is the true measure of differential expression for gene k in experiment i . Due to (possibly unknown) factors other than the covariates represented in X_k , experiment i is actually estimating a quantity slightly different, namely $\theta_{k,i}$. The estimate of this quantity is $\tilde{\theta}_{k,i}$. Because there are differences between studies, study i is estimating $\theta_{k,i}$, a random effect size from the population of all possible effect sizes. The difference $\delta_{k,i}$ is experiment-specific random deviation that represents inter-experiment variability not accounted for by the covariates. In addition to this random deviation there is sampling error $\epsilon_{k,i}$ within each study. It is generally assumed that $\epsilon_k \sim N(0, V_k)$ and $\delta_k \sim N(0, \sigma_k^2 I)$ are independent, and $\tilde{\theta}_k \sim N(X\beta_k, V_k + \sigma_k^2 I)$.

Following the derivation in Proposition 2 of DuMouchel and Harris 1983, the posterior distribution of σ_k can be expressed as

$$\pi(\sigma_k | \tilde{\theta}_k) \propto \pi(\sigma_k) \frac{|\Psi|^{1/2}}{|X_k^T \Psi X_k|^{1/2}} \exp\left(-\frac{1}{2} \tilde{\theta}_k^T S \tilde{\theta}_k\right), \quad (3)$$

where $\Psi = \Psi(\sigma_k) = [V_k + \sigma_k^2 I]^{-1}$ and $S = S(\sigma_k) = \Psi - \Psi X_k [X_k^T \Psi X_k]^{-1} X_k^T \Psi$. This Bayesian approach (DuMouchel and Normand 2000) estimates the posterior mean and covariance of β_k conditional on σ_k :

$$\begin{aligned} \beta_k^*(\sigma_k) &= E[\beta_k | \tilde{\theta}_k, \sigma_k] \\ &= [X^T \psi_k^{-1} X + D_k^{-1}]^{-1} (X^T \psi_k^{-1} \tilde{\theta}_k + D_k^{-1} b_k), \\ \beta_k^{**}(\sigma_k) &= Cov[\beta_k | \tilde{\theta}_k, \sigma_k] \\ &= [X^T \psi_k^{-1} X + D_k^{-1}]^{-1}, \end{aligned} \quad (4)$$

where $\psi_k = V_k + \sigma_k^2 I$. Note that allowing the diffuse prior with $d_{k,i} \rightarrow \infty$ will result in $D_k^{-1} \rightarrow 0$. Then the posterior mean and covariance of β_k conditional on σ_k are

$$\begin{aligned} \beta_k^*(\sigma_k) &= [X^T \psi_k^{-1} X]^{-1} X^T \psi_k^{-1} \tilde{\theta}_k, \\ \beta_k^{**}(\sigma_k) &= [X^T \psi_k^{-1} X]^{-1}. \end{aligned} \quad (5)$$

The posterior mean and covariance of the individual components of β_k are estimated as follows:

$$\beta_{k,j}^* = \int_{\sigma_k} \beta_{k,j}^*(\sigma_k) \pi(\sigma_k | \tilde{\theta}_k) d\sigma_k,$$

$$\beta_k^{**} = \int_{\sigma_k} (\beta_k^{**}(\sigma_k) + [\beta_k^*(\sigma_k) - \beta_k^*][\beta_k^*(\sigma_k) - \beta_k^*]^T) \pi(\sigma_k | \tilde{\theta}_k) d\sigma_k. \quad (6)$$

Following this, each covariate under consideration (including the intercept term) can be tested separately for significant differences from zero. The P-value corresponding to the hypothesis $H_0^{k,j}: \beta_{k,j} > 0$ is (DuMouchel and Normand 2000)

$$\begin{aligned} P_{k,j} &= P(\beta_{k,j} > 0 | \tilde{\theta}_k) \\ &= \int_{\sigma_k} \Phi \left(\frac{\beta_{k,j}^*(\sigma_k)}{\sqrt{\beta_{k,j}^{**}(\sigma_k)}} \right) \pi(\sigma_k | \tilde{\theta}_k) d\sigma_k, \end{aligned} \quad (7)$$

where Φ is the cumulative distribution function of the standard normal distribution. In the context of a microarray meta-analysis, this P-value corresponds to a test of whether the covariate j has a significantly positive effect on the reported (or observed) effect size estimates for gene k . For the intercept term ($j = 0$), when the columns of X_k are centered, this corresponds to testing whether the population mean underlying effect size (the SLR) for gene k is positive, i.e., whether gene k is significantly positively differentially expressed between the two conditions of interest. If $P_{k,0}$ is sufficiently large (close to 1), gene k is declared significantly upregulated from the control to the experimental condition under consideration. Conversely, if $P_{k,0}$ is sufficiently small (close to 0), gene k is declared significantly downregulated between the conditions.

2.2 Choice and Justification of Priors

As with any Bayesian work, the choice of priors deserves serious consideration and requires justification in the context of the application. The diffuse prior on β_k is chosen because it is noninformative and computationally convenient. On the other hand, some knowledge regarding the desired prior on σ_k is available. It is assumed that each of the experiments produces effect size estimates relatively close to the “true” effect size estimate for that experiment. Other than the covariates of interest, there is neither substantial nor systematic between-labs variability in the effect size estimates. In terms of the prior on σ_k , this means that σ_k is expected to be close to zero, but could vary substantially from zero in instances where the unexplained between-lab variability is unusually large. One convenient prior that provides such a distribution for σ_k is the log-logistic prior:

$$\begin{aligned} \pi(\sigma_k) &= \frac{c_{k,0}}{(c_{k,0} + \sigma_k)^2}, \\ c_{k,0} &= \sqrt{\frac{N_k}{\text{tr}(\text{diag}(V_k)^{-1})}}, \end{aligned} \quad (8)$$

where $c_{k,0}^2$ is the harmonic mean of the N_k sampling variances. The log-logistic prior on σ_k is chosen for contextual appropriateness and computational convenience. This prior has quartiles $c_{k,0}/3$, $c_{k,0}$, and $3c_{k,0}$, and is highly dispersed (with infinite expected values for both σ_k and σ_k^{-1}) (DuMouchel and Normand 2000).

2.3 Efficient Implementation

For the implementation of the methods proposed here, it was necessary to analyze Affymetrix data outside the commercial version MAS 5.0 (Affymetrix 2001) in such a way as to duplicate the Signal Log Ratio (SLR) results reported by MAS 5.0. The Bioconductor project (www.bioconductor.org) (Ihaka and Gentleman 1996) makes available several free packages for the R environment (<http://cran.r-project.org>) (Ihaka and Gentleman 1996) for the analysis of genomic data such as Affymetrix microarray data. One of these R packages, the affy package (Gautier et al. 2004), makes a concerted effort to allow for a variety of analyses of Affymetrix probe-level data. As such the Bioconductor resources were used in conjunction with separately developed R code that, given the appropriate .cel Affymetrix data files, will duplicate the SLR estimates and associated confidence intervals reported by MAS 5.0. This leads to the SLR variance estimate (Stevens and Doerge 2005a) necessary for the meta-analysis.

The basic details of the methods of numerical integration employed in the implementation of the hierarchical Bayes meta-analysis are included here. Recall that the prior π on σ_k for gene k is the log-logistic prior (Equation 8) and that the quartiles of π are $\frac{c_{k,0}}{3}$, $c_{k,0}$, and $3c_{k,0}$. Therefore, there are equal probability masses for σ_k in each of the four intervals $(0, \frac{c_{k,0}}{3})$, $(\frac{c_{k,0}}{3}, c_{k,0})$, $(c_{k,0}, 3c_{k,0})$, and $(3c_{k,0}, \infty)$. The basic strategy followed is to evaluate the integrand at n steps within each of these intervals; that is, every integration requires the integrand to be evaluated at $4n$ points. These n points within each interval are equally spaced. The fourth interval is taken from $3c_{k,0}$ to $3c_{k,0} + nc_{k,0}$; note that $\frac{3+n}{4+n}$ 100% of the area under the curve of π is in the interval $(0, 3c_{k,0} + nc_{k,0})$. A Simpsons Rule polynomial approximation (Fleming and Kaput 1979; Monahan 2001) was used for the numerical integration within each of the four intervals. This approach considers the integral of a function $f(x)$ from x_0 to x_n (n even) by equal step sizes Δx , and estimates the integral as

$$\int_{x_0}^{x_n} f(x) \approx \frac{\Delta x}{3} (f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \dots + 4f(x_{n-1}) + f(x_n)). \quad (9)$$

This can be applied to the necessary integrals for the Bayesian meta-analysis (for example, Equation 7) by considering them integrals of a function of σ_k .

3 Results

Figure 3 summarizes the results of the Bayesian meta-analysis of the simulated barley data. A claim of significant differential expression in this Bayesian meta-analysis was based on the intercept P-value from the centered-columns analysis, with the Bayesian posterior probability converted to a two-sided probability (Cooper and Hedges 1994), and the false discovery rate, or FDR (Benjamini and Hochberg 1995), controlled at 0.05. Of the 22,840 genes on

the barley1 Affymetrix microarray, the Bayesian meta-analysis declared 688 genes to be significantly differentially expressed between healthy and diseased barley, and 118 genes to have significant strain effects.

Figure 4 compares the SLR estimates from an individual lab and the Bayesian meta-analysis with the truth underlying the simulation. Of the 688 genes declared significantly differentially expressed by the meta-analysis of the simulated data, 687 were truly differentially expressed (as known from the simulation settings). There were 466 genes that were truly differentially expressed in the simulation, however the Bayesian meta-analysis failed to identify them as differentially expressed. Therefore, with these simulated data, the Bayesian meta-analysis made 1 false positive (type I error) and 466 false negatives (type II errors). Based on these simulations, the meta-analysis tended to provide results that were closer to the true degree of differential expression than did the results from any of the individual labs. Furthermore, the meta-analysis tended to have fewer type I and type II errors than did any single lab.

The Bayesian meta-analysis was also applied to the EAE data that was described earlier. Figure 5 summarizes the results. Only the 9,948 genes that appeared in more than one experiment were used in this meta-analysis. Based on the intercept P-value from the centered-columns meta-analysis, the Bayesian meta-analysis identified 1,051 genes as statistically significantly differentially expressed (controlling the FDR at 0.05). Additionally, there was 1 gene with a significant strain effect and 67 genes with a significant tissue effect.

It is particularly interesting to note that the gene with the largest SLR estimate in the Bayesian meta-analysis is an *Arginase* gene. This gene was recently identified as being strongly up-regulated in the EAE condition (Xu et al. 2003). By pooling results across multiple experiments, there appears to be substantial additional evidence that suggests *Arginase* is strongly related to EAE.

4 Summary and Future Work

When multiple laboratories use the same Affymetrix GeneChip technology to quantify differential gene expression for the same condition, their results will vary due to both chance variation and fundamental differences between experimental conditions. The Bayesian meta-analysis presented in this paper provides a framework to account for both inter-laboratory variability and known differences between experiments when combining microarray results across laboratories through the use of a prior distribution on the level of inter-laboratory variability and the use of covariate information. This approach allows for a clearer understanding of each gene's relationship to the condition (or disease) of interest. The simulation example given here demonstrated that the meta-analytic results tend to better approximate the underlying true degree of differential expression than do the results from any single laboratory. The application of this Bayesian meta-analysis to an actual mouse EAE microarray data set supported the fundamental claim of one of the contributing laboratories, namely

that the *Arginase* gene is strongly related to EAE.

As microarray technology is applied on a larger scale in a greater variety of agricultural research laboratories, it will become increasingly important to have robust meta-analytic methods that are able to combine results across experiments in order to derive the maximum amount of information possible about each gene's relationship to the conditions being studied. Accounting for chance variation and fundamental differences between experimental conditions is an important step in developing and refining these meta-analytic methods. As such it will be necessary to develop robust measures of differential expression that can be meaningfully combined and compared across microarray platforms so that the approach is not limited to a single platform, such as the Affymetrix GeneChip.

One of the fundamental assumptions in a typical meta-analysis is the independence of estimates from the several studies. However, when multiple SLR estimates are reported for the same gene in the same laboratory, the estimates are not necessarily independent, especially when the estimates are based on comparisons involving the same array. This dependence structure can be estimated and accounted for in a meta-analysis of results across laboratories (Stevens 2005). It will be of great interest to account for various covariance structures not only among estimates for an individual gene, but also among groups or networks of genes. With the emergence of other technologies such as protein arrays, similar meta-analytic approaches will need to be developed and refined to combine results across technologies to better understand fundamental underlying pathways. This can lead to great advancements in applied statistics in agriculture as well as other biological sciences.

Acknowledgments

We thank Drs. Halina Offner (Oregon Health Sciences University), Ruaidhrí J. Carmody (University of Pennsylvania School of Medicine), and Saleh M. Ibrahim (University of Ros-tock), as well as their colleagues, for providing access to their raw Affymetrix data.

References

- Affymetrix (2001). *Affymetrix Microarray Suite User's Guide Version 5.0*. Affymetrix, Santa Clara, CA.
- Affymetrix (2002). *Statistical Algorithms Description Document*. Affymetrix, Santa Clara, CA.
- Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57(1), 289–300.

- Campbell, M. K. (1995). *Biochemistry* (2nd ed.). Saunders College Publishing, Philadelphia, PA.
- Carmody, R. J., B. Hilliard, K. Maguschak, L. A. Chodosh, and Y. H. Chen (2002). Genomic scale profiling of autoimmune inflammation in the central nervous system: the nervous response to inflammation. *Journal of Neuroimmunology* 133, 95–107.
- Choi, J. K., U. Yu, S. Kim, and O. J. Yoo (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19(Suppl. 1), i84–i90.
- Cooper, H. and L. V. Hedges (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York, NY.
- Craig, B. A., M. A. Black, and R. W. Doerge (2003). Gene expression data: The technology and statistical analysis. *Journal of Agricultural, Biological, and Environmental Statistics* 8(1).
- Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563.
- DuMouchel, W. and S.-L. Normand (2000). Computer-modeling and graphical strategies for meta-analysis. In D. K. Stangl and D. A. Berry (Eds.), *Meta-Analysis in Medicine and Health Policy*, pp. 127–178. Marcel Dekker.
- DuMouchel, W. H. (1994). hblm. Available at <http://www.research.att.com/~dumouchel/bsoft.html>.
- Fleming, D. J. and J. K. Kaput (1979). *Calculus with Analytic Geometry*. Harper and Row, Publishers, New York, NY.
- Gautier, L., L. Cope, B. M. Bolstad, and R. A. Irizarry (2004). affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3), 307–315.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research* 5(10), 3–8.
- Hoaglin, D. C., F. Mosteller, and J. Tukey (1983). *Understanding Robust and Exploratory Data Analysis*. John Wiley and Sons, New York.
- Ibrahim, S. M., E. Mix, T. Bottcher, D. Koczan, R. Gold, A. Rolfs, and H.-J. Thiesen (2001). Gene expression profiling of the nervous system in murine experimental autoimmune encephalomyelitis. *Brain* 124, 1927–1938.
- Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5(3), 299–314.
- Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* 14.
- Matejuk, A., J. Dwyer, C. Hopke, A. A. Vandenbark, and H. Offner (2003). 17β -estradiol treatment profoundly down-regulates gene expression in spinal cord tissue in mice

- protected from experimental autoimmune encephalomyelitis. *Archivum Immunologiae et Therapiae Experimentalis* 51, 185–193.
- Matejuk, A., J. Dwyer, A. Zamora, A. A. Vandenbark, and H. Offner (2002). Evaluation of the effects of 17β -estradiol (17β -e2) on gene expression in experimental autoimmune encephalomyelitis using DNA microarray. *Endocrinology* 143(1), 313–319.
- Mix, E., J. Pahnke, and S. M. Ibrahim (2002). Gene-expression profiling of experimental autoimmune encephalomyelitis. *Neurochemical Research* 27(10), 1157–1163.
- Monahan, J. F. (2001). *Numerical Methods of Statistics*. Cambridge University Press, New York, New York, USA.
- Moreau, Y., S. Aerts, B. D. Moor, B. D. Strooper, and M. Dabrowski (2003). Comparison and meta-analysis of microarray data: From the bench to the computer desk. *Trends in Genetics* 19(10), 570–577.
- Parmigiani, G., E. S. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research* 10, 2922–2927.
- Rhodes, D. R., T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* 62, 4427–4433.
- Rhodes, D. R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences* 101, 9309–9314.
- Shen, L., J. Gong, R. A. Caldo, D. Nettleton, D. Cook, R. P. Wise, and J. A. Dickerson (2005). BarleyBase - an expression profiling database for plant genomics. *Nucleic Acids Research* 33, Database issue, D614–D618.
- Stevens, J. R. (2005). *Meta-analytic Approaches for Microarray Data*. Ph. D. thesis, Purdue University, West Lafayette, Indiana.
- Stevens, J. R. and R. W. Doerge (2005a). Combining Affymetrix microarray results. *BMC Bioinformatics* 6:57.
- Stevens, J. R. and R. W. Doerge (2005b). Meta-analysis combines Affymetrix microarray results across laboratories. *Comparative and Functional Genomics* 6, 116–122.
- Xu, L., B. Hilliard, R. J. Carmody, G. Tsabary, H. Shin, D. W. Christianson, and Y. H. Chen (2003). Ariginase and autoimmune inflammation in the central nervous system. *Immunology* 110, 141–148.

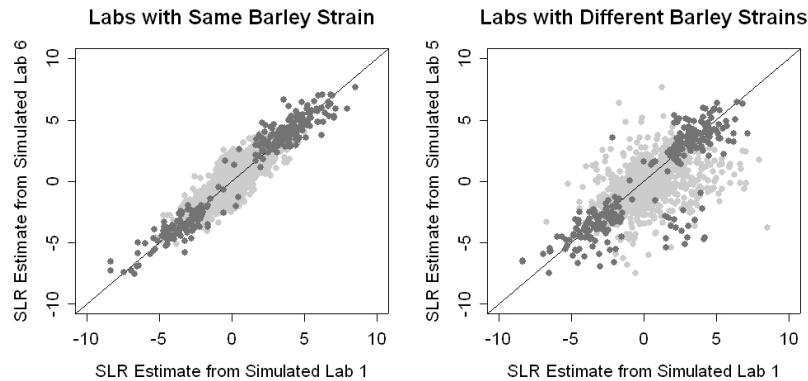


Figure 1: Comparison of SLR estimates from simulated barley laboratories. As an example, specific to lab 1 and lab6, each lab estimated the degree of differential expression for each gene between healthy and diseased barley in a particular barley strain. The dark gray points correspond to genes declared significantly differentially expressed (controlling the FDR at 0.05) by both labs. The results from labs with the same covariate level (strain) tend to agree better.

Table 1: Summary of EAE microarray experiments in mouse. The number of experiments refers to the number of independent array-to-array comparisons (from healthy to diseased tissue) that could be made in each laboratory.

Lab	Strain	Tissue	Number of Experiments
Ibrahim	C57BL/6	spinal cord	1
Chen	C57BL/6	spinal cord	2
Offner(1)	BV8S2/Av4	spinal cord	1
Offner(2)	BV8S2/Av4	spleen	2

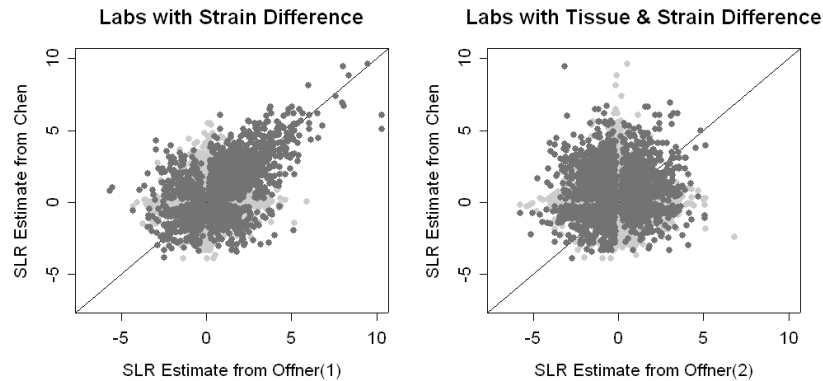


Figure 2: Comparison of SLR estimates from EAE mouse laboratories. Each lab estimated each gene's degree of differential expression between healthy and diseased tissue. The dark gray points correspond to genes declared significantly differentially expressed (controlling the FDR at 0.05) by both labs. The lack of agreement between labs provides motivation for the meta-analytic approach to combine results across laboratories while allowing for inter-laboratory variability and covariate (such as tissue and strain) differences.

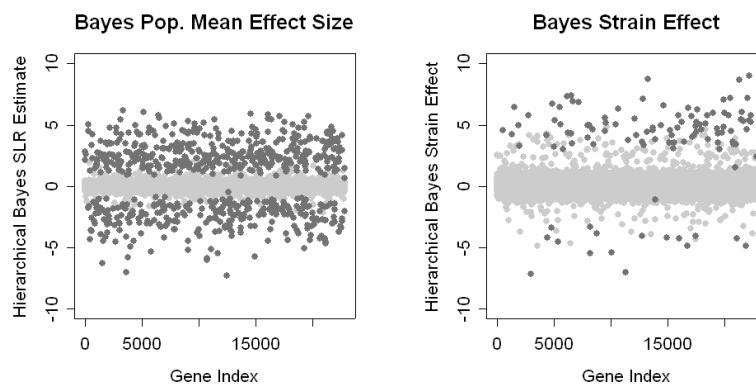


Figure 3: Summary of Bayesian meta-analysis results for simulated barley data. The meta-analysis provided estimates of both the population mean effect size (the SLR) and strain effect for each gene. The dark gray points represent genes where the effect (population mean effect size or strain) was declared statistically significant, controlling the FDR at 0.05. There were 688 genes (of 22,840) declared significantly differentially expressed (based on the SLR), and 118 genes found to have significant strain effects.

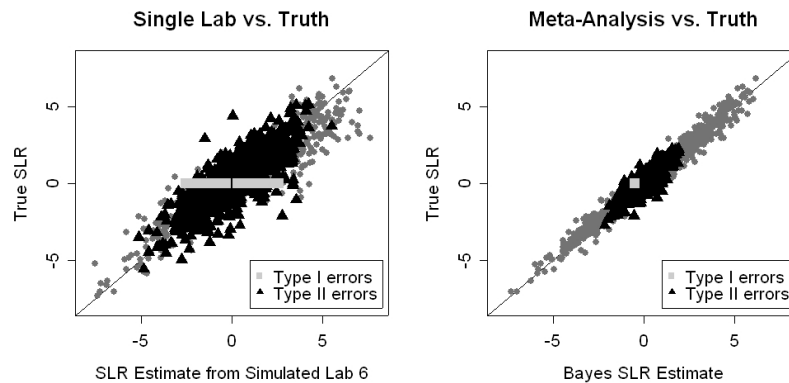


Figure 4: Comparison of estimates with truth underlying simulated barley data. The results of the Bayesian meta-analysis tended to better approximate the true degree of differential expression than did the results from any individual lab. The meta-analysis also tended to have fewer type I and type II errors.

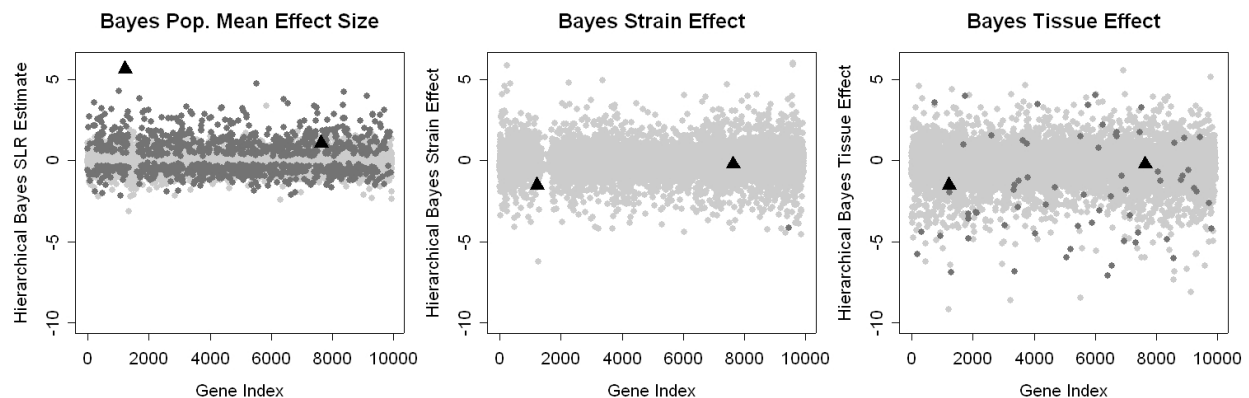


Figure 5: Summary of Bayesian meta-analysis results for observed EAE mouse data. The meta-analysis provided estimates of the population mean effect size (the SLR), strain effect, and tissue effect. The dark gray points represent genes where the effect (population mean effect size, strain, or tissue) was declared statistically significant, controlling the FDR at 0.05. The large black triangles represent *Arginase* genes. Of the 9,948 genes considered, 1,051 genes were declared significantly differentially expressed (based on the SLR), with 1 gene having significant strain effect and 67 genes having significant tissue effect. The largest SLR (population mean effect size) estimate corresponds to an *Arginase* gene, recently found to be highly related to the EAE.