UNIVERSITY OF WISCONSIN DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS

Technical Report # 166, February 2002

Slightly modified version to appear in Statistics in Medicine, 2003

On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles

C.M. Kendziorski, Ph.D.
M.A. Newton, Ph.D.
H. Lan, Ph.D.
M.N. Gould, Ph.D.

UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS
AND MEDICAL INFORMATICS
K6/446 Clinical Science Center
600 Highland Avenue
Madison, WI 53792-4675
608-263-1706

Summary

DNA microarrays provide for unprecedented, large-scale views of gene expression and, as a result, have emerged as a fundamental measurement tool in the study of diverse biological systems. Statistical questions abound; but many traditional data analytic approaches do not apply, in large part because thousands of individual genes are measured with relatively little replication. Empirical Bayes methods provide a natural approach to microarray data analysis because they can significantly reduce the dimensionality of an inference problem while compensating for relatively few replicates by using information across the array. We propose a general empirical Bayes modeling approach which allows for replicate expression profiles in multiple conditions. The hierarchical mixture model accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations. Two distinct parameterizations are considered: a model based on Gamma distributed measurements and one based on log-normally distributed measurements. False detection rate and related operating characteristics of the methodology are assessed in a simulation study. We also show how the posterior odds of differential expression in one version of the model is related to the ratio of the arithmetic mean to the geometric mean of the two sample means. The methodology is used in a study of mammary cancer in the rat, where four distinct patterns of expression are possible.

1 Introduction

Enabled by resources created from genome sequencing projects, DNA microarray technology has emerged as a fundamental measurement tool in the study of diverse biological systems. Microarrays offer an unprecedented ability to perform large-scale studies of gene expression. As a result, the focus of many research efforts has shifted from individual genes to multiple genes and the complicated and orchestrated ways in which they interact to maintain life.

With the shift from individual to integrated analysis in molecular biology comes a shift in the related statistical questions posed and methods required. The number of measurements of distinct genes across an array greatly exceeds that for any individual gene. Thus, we as statisticians are faced with the "large p, small n" paradigm (West *et al.* 2000a, 2000b). Empirical Bayes methods provide a natural approach to microarray data analysis because they can reduce significantly the dimensionality of an inference problem involving many unknown parameters (e.g., Efron and Morris 1973, 1977). Our earlier work described a version of parametric empirical Bayes analysis for spotted microarrays and was restricted to so-called "single-slide" data in which each gene produces two measurements, one from each cell condition (Newton *et al.* 2001). The methodology provides improved estimation of expression fold-change and allows for the assessment of differential expression by the calculation of a posterior odds. In spite of there being very little data per gene, the methodology works because inference about a given gene uses information on the fluctuations of expression measurements from all genes. One goal of the present paper is to extend the parametric empirical Bayes calculations beyond the single-slide case, and thus to allow replicate expression profiles in multiple cell conditions.

The methodological work presented here is motivated in part by an experiment to study gene expression in a rat model of breast cancer. Microarray data were obtained from mammary epithelial cells harvested from 12 week old females representing four distinct inbred lines (two parentals and two congenic lines; see Section 5). The parental strains

differ in their susceptibility to breast cancer and identifying differentially expressed genes could provide insight into the genetic basis of this difference. Expression measurements were obtained using the Affymetrix GeneChip technology, and probe-pair-level data were processed by DNA-Chip Analyzer (Li and Wong, 2001) to produce a quantitative expression index for each gene and for each RNA sample. More details are given in Section 5. An interesting feature of the present study is the presence of four inter-related groupings (the four inbred lines). For each gene, we are not simply asking if there is differential expression or not, but we are asking something about the pattern of differential expression among the four groups.

The development of statistical methods to identify differentially expressed genes has recently received much attention, especially methods to identify genes that are differentially expressed between two conditions. A general approach to this problem is to conduct a hypothesis test at each gene and then correct for multiple comparisons. Most of the test statistics currently used are t (or t-like) and differ primarily in the estimation of the variance. Dudoit *et al.* (2001) use a t-statistic with variance estimated by the within gene sample variance and go on to address the multiple comparisons problem extensively using permutation analysis. Tusher *et al.* (2001) also use the within gene sample variance, but adjust the denominator of their test statistic by adding a constant to account for the dependence between the relative difference in expression and absolute intensity; they address the multiple comparisons problem using the method of false discovery rates. Baldi and Long (2001) use the posterior variance derived from a Bayesian analysis and do not consider the multiple comparisons problem. Methods such as these which treat the genes as separate fixed effects may have reduced efficiency when compared to methods which treat the genes as arising from some population, and thus which allow a level of information sharing amongst genes.

Information sharing is consequence of the empirical Bayes approach. The particular

method proposed by Newton *et al.* (2001) amounts to a simple two group mixture-model calculation. Stochastically, each gene is either differentially expressed or not. Those genes which are not present data according to some background distribution, and those which are present data according to a different distribution. The specific forms of these distributions arise by another layer of mixing over the latent mean expression level for each gene. In that work, the expression measurements are independent and follow a Gamma distribution conditional upon the latent mean expression level. The Gamma model is convenient numerically and analytically, but also has some justification in the modeling of abundances in a large population. The latent mean values are treated not as fixed effects (as they would be in the standard analyses outlined above) but follow the conjugate, inverse-Gamma distribution. Two measurements that happen to have the same latent mean value represent equivalent expression; otherwise there is differential expression. With these components in place, inference about differential expression amounts to computing the odds of that event, conditional on the measurements. The analysis is *empirical* Bayes because the small number of unknown parameters which index the component distributions are estimated from the data. In Section 2 we describe an extension of this approach to replicate profiles in multiple conditions.

There are other mixture-modeling approaches to expression data analysis. Working with a specific experimental design, Efron *et al.* (2000, 2001) describe empirical Bayesian calculations which relax the parametric assumptions. After a long series of pre-processing steps, each gene yields a one-dimensional test statistic whose marginal distribution turns out to be known and whose null distribution (i.e., on equivalent expression) can be nonparametrically estimated. Lee *al.* 2000 also use the idea of a two group mixture model for expression analysis; their calculations were in a slightly different context and were applied to parameter estimates from a first-stage analysis. Here we do not endeavor to extend either of these approaches to the case of multiple conditions, but in Section 4 we do offer some

numerical comparisons of false detection rate between our proposal and the nonparametric method in the context of two conditions.

Our proposed hierarchical mixture model accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations. We consider two distinct families: a model based on Gamma distributed measurements and one based on log-normally distributed measurements. As it is often observed (e.g. Chen, *et al.*, 1997), a constant coefficient of variation is built in to both models. The models also account for differential variation in apparent fold change (e.g. Dudoit *et al.* 2001, Newton *et al.* 2001, Tusher *et al.* 2001).

Using the proposed hierarchical mixture model, we derive expressions for the posterior probability of differential expression in the case of two conditions, and we summarize the general case of multiple conditions (Section 2). Two specific parametric formulations are given in Section 3. We report a simulation study to assess the sampling properties of the resulting approach to identify differentially expressed genes. Operating characteristics including the sensitivity, specificity, and false detection rate are evaluated by simulation under a range of conditions (Section 4). We use the methodology to analyze rat mammary epithelium expression profiles in Section 5. Throughout we compare results from the two models. In Section 6 we discuss additional features of the modeling framework, including some simple statistics which arise from the posterior odds formulas.

2 Hierarchical modeling framework

Our models attempt to describe the probability distribution of a set of expression measurements taken on a gene g. The measurements may arise from cells under different conditions, and there may be replicate measurements in each condition. We assume that some preprocessing technique has been used to adequately normalize the data so that the measurements can be viewed as *bona fide* approximations of relative gene expression

in the sampled cells. Let us initially consider comparing two conditions, with data $\mathbf{x}_g = (x_{g,1}, x_{g,2}, \ldots, x_{g,n_1})$ from the n_1 replicate measurements in the first condition and $\mathbf{y}_g = (y_{g,1}, y_{g,2}, \ldots, y_{g,n_2})$ from the second condition. Two levels of mixing characterize the distribution of these data.

One stage of mixing is discrete and captures the different patterns of differential expression. On a null hypothesis there is equivalent expression (EE) between the two conditions, and the data arise from a joint probability density (pdf) $f_0(\mathbf{x}_g, \mathbf{y}_g)$. Alternatively there is differential expression (DE), and the joint pdf is $f_1(\mathbf{x}_g, \mathbf{y}_g)$. *A priori* we do not know which situation manifests itself for gene g, and so we introduce the discrete mixing parameter p to denote the unknown probability of differential expression. Thus, the marginal distribution of the data is

$$p f_1(\mathbf{x}_g, \mathbf{y}_g) + (1-p) f_0(\mathbf{x}_g, \mathbf{y}_g). \tag{1}$$

If we know the parameter p and the form of f_0 and f_1 then by Bayes rule, the posterior probability of differential expression may be calculated:

$$\frac{p \;\; f_1(\mathbf{x}_g, \mathbf{y}_g)}{p \; f_1(\mathbf{x}_g, \mathbf{y}_g) + (1-p) \; f_0(\mathbf{x}_g, \mathbf{y}_g)}.$$

Equivalently, one could consider the posterior odds of DE:

$$\text{odds}_g = \frac{P\left(\text{DE}|\mathbf{x}_g, \mathbf{y}_g\right)}{P\left(\text{EE}|\mathbf{x}_g, \mathbf{y}_g\right)} = \frac{p}{1-p} \;\; \frac{f_1(\mathbf{x}_g, \mathbf{y}_g)}{f_0(\mathbf{x}_g, \mathbf{y}_g)}$$

We use these summary measures to enable gene-specific inferences concerning differential expression. (Though it is masked in the notation, events DE and EE are specific to the gene g.)

The calculations are further specified by a second stage of mixing so that we obtain particular forms for the pdfs f_0 and f_1. Intuitively, the measurements $\mathbf{x}_g$ and $\mathbf{y}_g$ should tend to be closer together in f_0 since there the variations do not include any systematic shifts between the cell types. The situation is naturally characterized by mixing over

a latent gene-specific mean expression level. Consider the case of equivalent expression (EE). We say that all $N = n_1 + n_2$ measurements arise independently and identically from an observation distribution $f_{\text{obs}}(\cdot|\mu_g)$. The parametric forms we consider are the Gamma distribution and the log-Normal distribution; they both capture evident features in real array data parsimoniously. Were we to treat the μ_g as a fixed effect, we would not take advantage of information sharing. Our approach is thus to consider μ_g as arising from some genome-wide distribution $\pi(\mu_g)$, which represents fluctuations in mean expression levels among genes. On the null hypothesis EE, the marginal probability of data becomes

$$f_0(\mathbf{x}_g, \mathbf{y}_g) = \int \left(\prod_{i=1}^{n_1} f_{\text{obs}}(x_{g,i}|\mu_g) \right) \left(\prod_{j=1}^{n_2} f_{\text{obs}}(y_{g,j}|\mu_g) \right) \pi(\mu_g)\, d\mu_g.$$

On the alternative hypothesis of differential expression DE, the latent mean value μ_g, which governs the sample of $x_{g,i}$'s , is different from that value μ'_g for the $y_{g,i}$'s. Naturally, each of these values arises independently from $\pi(\mu_g)$ and so we have the representation

$$f_1(\mathbf{x}_g, \mathbf{y}_g) = f_0(\mathbf{x}_g)\, f_0(\mathbf{y}_g). \tag{2}$$

In the notation here we allow that f_0 is simply the predictive density of the argument having integrated away a common mean value: i.e.

$$f_0(z_1, \ldots, z_n) = \int \left(\prod_{i=1}^{n} f_{\text{obs}}(z_i|\mu_g) \right) \pi(\mu_g)\, d\mu_g. \tag{3}$$

In summary, the odds of differential expression based on replicates $\mathbf{x_g} = (x_{g,1}, \ldots, x_{g,n_1})$ in one condition and replicates $\mathbf{y_g} = (y_{g,1}, \ldots, y_{g,n_2})$ in the second condition is

$$\text{odds}_g = \frac{p}{1-p} \frac{f_0(\mathbf{x}_g)\, f_0(\mathbf{y}_g)}{f_0(\mathbf{x}_g, \mathbf{y}_g)}. \tag{4}$$

We parameterize the component distributions in the next section to complete the mixture formulation. Inference proceeds by estimating the parameters in f_0 via marginal maximum likelihood and then computing the posterior odds of differential expression for each gene g.

A nice feature of the mixture modeling framework is that it carries over transparently to comparisons among multiple conditions (beyond DE and EE). For example, given three

conditions, $\binom{3}{0} + \binom{3}{2} + \binom{3}{3} = 5$ expression patterns are possible. These include equivalent expression across the three conditions, altered expression in just one condition, and distinct expression in each condition. With microarrays from four cell conditions there are 15 different patterns. (The total number of patterns is equal to the Bell exponential number of possible set partitions, in fact.) As we see in the rat mammary study which has four cell conditions, we can sometimes reduce the total number of patterns to a more manageable level, and in that case we reduce to four interesting patterns (Section 5).

Suppose that $m+1$ distinct patterns of expression are possible for a data vector $\mathbf{d}_g = (d_{g,1}, \ldots, d_{g,N})$ on some gene g. Then, generalizing (1), $\mathbf{d}_g$ is governed by a mixture of the form

$$\sum_{k=0}^{m} p_k f_k(\mathbf{d}_g), \tag{5}$$

where $\{p_k\}$ are mixing proportions and component densities $\{f_k\}$ give the predictive distribution of measurements for each pattern of expression. Consequently, the posterior probability of expression pattern k is

$$P(k|\mathbf{d}_g) \propto p_k f_k(\mathbf{d}_g). \tag{6}$$

Furthermore, generalizing (2), the pattern-specific predictive density $f_k(\mathbf{d}_g)$, for $k > 0$, will be a product across subsets of the data vector, say,

$$f_k(\mathbf{d}_g) = \prod_{s \in S} f_0(\mathbf{d}_{g,s})$$

where S is a set partition of $\{1, 2, \ldots, N\}$ constructed in such a way that any measurements contained in a component subset s in S share a common mean value, μ_s.

The posterior probabilities summarize our inference about expression patterns at each gene. They can be used to identify genes with altered expression in at least one group, to classify genes into distinct expression groups, or to order genes within groups. Before posterior summaries can be evaluated, however, we must first specify distributional forms for the components of the hierarchical mixture model.

3 The Gamma-Gamma and Lognormal-Normal models

The general mixture model in Section 2 is specified by an observation component $f_{\text{obs}}(\cdot|\mu_g)$ which characterizes fluctuations in repeated measurements from a gene having latent mean expression level μ_g, and a second component $\pi(\mu_g)$ which describes fluctuations in these means among genes. Here we describe two particular versions of the general mixture formulation.

In the Gamma-Gamma (GG) model, the observation component is a Gamma distribution having shape parameter $\alpha > 0$ and a mean value μ_g; thus, with *scale* parameter $\lambda = \alpha/\mu_g$,

$$f_{\text{obs}}(z|\mu_g) = \frac{\lambda^\alpha z^{\alpha-1} \exp\{-\lambda z\}}{\Gamma(\alpha)}$$

for measurements $z > 0$. Note that the coefficient of variation in this distribution is $1/\sqrt{\alpha}$, taken to be constant across genes g. Matched to this observation component is a marginal distribution $\pi(\mu_g)$ which we take to be an inverse Gamma. More specifically, fixing α, the quantity $\lambda = \alpha/\mu_g$ has a Gamma distribution with shape parameter α_0 and scale parameter ν. Thus three parameters are involved, $\theta = (\alpha, \alpha_0, \nu)$, and, upon integration, the joint predictive density corresponding to (3) has the form

$$f_0(z_1, z_2, \ldots, z_n) = K \, \frac{\left(\prod_{i=1}^n z_i\right)^{\alpha-1}}{\left(\nu + \sum_{i=1}^n z_i\right)^{n\alpha+\alpha_0}} \tag{7}$$

where

$$K = \frac{\nu^{\alpha_0} \, \Gamma(n\alpha + \alpha_0)}{\Gamma^n(\alpha) \, \Gamma(\alpha_0)}$$

From this result one can calculate the posterior probability of any given expression pattern following the prescription in Section 2. In the special case of two conditions, the posterior odds for differential expression (4) simplify to

$$\text{odds}_g = K' \, \frac{\left(\sum_{i=1}^{n_1} x_{g,i} + \sum_{i=1}^{n_2} y_{g,i} + \nu\right)^{N\alpha+\alpha_0}}{\left(\sum_{i=1}^{n_1} x_{g,i} + \nu\right)^{n_1\alpha+\alpha_0} \left(\sum_{i=1}^{n_2} y_{g,i} + \nu\right)^{n_2\alpha+\alpha_0}} \tag{8}$$

where

$$K' = \frac{\nu_0^\alpha \, \Gamma(n_1\alpha + \alpha_0) \, \Gamma(n_2\alpha + \alpha_0)}{\Gamma(\alpha_0) \, \Gamma(N\alpha + \alpha_0)}.$$

and recall that $N = n_1 + n_2$ is the total number of observations on gene g. The odds may be computed as soon as we have estimates in hand for $\theta = (\alpha, \alpha_0, \nu)$. In Section 6 we point out an interesting connection between these posterior odds and the arithmetic-geometric mean inequality.

The GG calculations derived above extend those presented in Newton *et al.* 2001 to replicates and multiple conditions. Many investigators would consider as reasonable a different model for the array measurements – one in which the log-transformed measurements have a Gaussian observation component. We may use this in our hierarchical mixture model as follows. Let us say the natural logarithms of the measurements are denoted $\tilde{\mathbf{x}}_g$ and $\tilde{\mathbf{y}}_g$. The latent gene-specific mean μ_g is now a mean for the log-transformed measurements, and these measurements have a sampling variance σ^2 which we treat as common to all genes. Note that the coefficient of variation for the original measurements becomes $\sqrt{\exp(\sigma^2) - 1}$ in this model. A conjugate prior for the μ_g is normal with some underlying mean μ_0 and variance τ_0^2. Integrating as in (3), the joint predictive density f_0 for an n-dimensional input becomes Gaussian with mean vector $\underline{\mu}_0 = (\mu_0, \mu_0, \ldots, \mu_0)^t$ and exchangeable covariance matrix

$$\mathbf{\Sigma}_n = \left(\sigma^2\right) \mathbf{I_n} + \left({\tau_0}^2\right) \mathbf{M_n}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix and M_n is an $n \times n$ matrix of ones. This basic formulation has been well studied (e.g., Carlin and Louis, 1996). In our context there is an additional layer of discrete mixing, and we may derive the posterior probability of different expression patterns following (6). For the special case of two conditions, the odds of differential expression (4) may be written in terms of quadratic forms. Let $\delta_g = (\tilde{\mathbf{x}}_g, \tilde{\mathbf{y}}_g)^t - \underline{\mu}_0$ denote the centered, transformed full data vector for gene g,

$$\text{odds}_g = \sqrt{\frac{|\Sigma_N|}{|\Sigma_*|}} \exp\left\{-\frac{1}{2}\delta_g^t(\Sigma_*^{-1} - \Sigma_N^{-1})\delta_g\right\}$$

where Σ_* is the $N \times N$ block-diagonal matrix with Σ_{n_1} in the upper left block and Σ_{n_2} in the lower right block.

For either the lognormal-normal (LNN) model or GG model, we can use the method of maximum (marginal) likelihood to obtain estimates of the small set of unknown parameters. (In the GG model, $\theta = (\alpha, \alpha_0, \nu)$ and in LNN, $\theta = (\mu_0, \sigma^2, \tau_0)$.) The mixing proportion p is a fourth parameter. The marginal loglikelihood is a sum over genes g of terms (1):

$$l(\theta) = \sum_g \log \left[p\, f_1(\mathbf{x}_g, \mathbf{y}_g) + (1-p)\, f_0(\mathbf{x}_g, \mathbf{y}_g) \right] \tag{9}$$

and this may be optimized by various methods. We use the Splus program *nlminb* (Statistical Sciences, 1993). In the case of three or more conditions, we use the EM algorithm to handle the vector of mixing proportions (Dempster *et al.* 1977) (see Appendix).

The classification of genes as differentially expressed (DE) or equivalently expressed (EE) according to the state favored by the posterior odds is an optimal procedure in the context of the mixture model. It minimizes the expected number of mistakes. Interestingly, this goal is different from the the goal in classical testing which is to bound the type I error rate and then aim to maximize the power. To reduce type I errors we could make a more stringent decision rule and assume EE unless the odds favoring DE are much larger than 1:1. Any particular decision rule may be evaluated in terms of its sampling properties, and in the next section we do so via simulation for a few model settings.

4 Simulations

The proposed methodology provides a way to infer patterns of differential expression among two or more conditions, but it relies on parametric model assumptions and the implementation of numerical optimization methods. To assess the methodology we performed a small set of simulation studies. These provide some insight into whether or not the parameters are well estimated, how much inference is affected by fitting a model

different from the one which generated the data, and perhaps most importantly, they provide information on error rates in the inference of differential expression.

First, we simulated the GG model with 10,000 genes in two conditions, having three replicates in each condition. We took model parameters similar to those obtained in Newton *et al.* 2001 ($\alpha = 10$, $\alpha_0 = 0.9$, and $\nu = 0.5$). The prior probability that a gene is differentially expressed was set to $p = 0.2$. The GG and LNN mixture models described in Section 3 were each fit to these simulated data. Histograms of the simulated data along with the fitted marginal densities are shown in the left panel of Figure 1. As expected, the fitted GG marginal density more closely describes the simulated data.

Next, we simulated a similar data set under the LNN model ($\mu_0 = 2.3$, $\sigma = 0.3$, and $\tau_0 = 1.39$); each mixture model was again fit to the simulated data. As shown in the right panel of Figure 1, the simulated data is better described by the LNN density. Although expected, this result illustrates that comparing the marginal densities to the empirical distribution can give insight into which model assumptions are more appropriate.

We did a more formal comparison of GG and LNN by calculating a log Bayes factor to measure the relative fit of these models (Kass and Raftery 1995). We take this simply as the difference of the log predictive densities (given by equation 9) calculated under GG or LNN assumptions. For each simulated data set, the Bayes factor correctly identified the model generating the simulated data. Considering the success of either approach in identifying the underlying model, one might think that the parametric assumptions have a substantial effect on which genes are identified as differentially expressed. We find that this is not the case.

The differences in the simulated data which allow for model identification do not seem to impact the mixture model's ability to identify differentially expressed genes. Figure 2 shows the average intensities (across replicates) for spots identified as changed (odds > 1) using the GG or LNN model applied to GG or LNN data. The odds plots look similar within simulated data set, independent of model assumption. The numerical results are in fact similar (see

Table 1). For this simulation, 1968 (1952) genes in the GG (LNN) data happened to be differentially expressed . Each method applied to each data set identified about 1470 genes as differentially expressed (i.e., odds > 1). Out of those identified, approximately 95% were correct.

Simulations were repeated to assess the sensitivity, specificity, positive and negative predictive values, and false detection rates of the methodology. We varied the proportion p of differentially expressed genes from 0.1 to 0.5 (in increments of 0.1). For each fixed proportion, 100 sets having 6 arrays each (3 replicates in 2 conditions, as above) were simulated. Parameter values were defined as above. Odds were calculated using both GG and LNN models. Sensitivity is calculated as the average (over the 100 simulations) of the fraction of differentially expressed genes correctly identified by the method (odds > 1); specificity is the average of the fraction of equivalently expressed genes correctly identified (odds ≤ 1). The positive predictive value (PPV) is defined as the average of the fraction of genes with odds > 1 that are truly differentially expressed; the negative predictive value (NPV) is the average of the fraction of genes with odds ≤ 1 that are equivalently expressed. The false detection rate (FDR) is the average of the ratio of the number of false positives to the number of genes identified as differentially expressed.

Parameter estimates averaged over the 100 simulations are given in Tables 2 and 3. As shown, the parameter estimates are close to the true values, with little standard error. Tables 4 and 5 present operating characteristics of each approach. The characteristics are similar under different simulation assumptions. For each method, the sensitivity ranges from 65% to 80% and is increasing with increasing p. The specificity is at or above 95% for each method and each value of p considered. The positive predictive value ranges between 94% and 95%, while the negative predictive value decreases from near 97% when $p = 0.1$ to near 80% when $p = 0.5$. The average false detection rate (FDR), near 0.05 for all values of p, increased slightly with increasing p. A graphical representation is shown in Figure 3.

The FDR estimates suggest that using an odds value greater than one as a rejection rule results in a type I error rate near 0.05. Interestingly, the estimates of the FDR are similar to those reported by Efron *et al.*, 2000, in assessment of their empirical Bayes approach. A lower bound on p for the data set considered there is estimated to be 0.189. The authors consider the FDR rates using the posterior probability of differential expression at values greater than and equal to 0.9. This corresponds to an odds$>$ 9. They report an FDR of 0.0048 at this level. Our results are similar. For $p = 0.2$ and odds$>$ 9.1, the FDR averaged over 100 simulations was 0.0054 (GG on GG), 0.0052 (GG on LNN), 0.0057 (LNN on GG), and 0.0049 (LNN on LNN); standard errors were all less than 0.0003.

5 Data Analysis

Rat strains vary greatly in their resistance to carcinogen-induced mammary cancer. The inbred Copenhagen (COP) strain is almost completely resistant to carcinogenesis induced by DMBA, while the inbred Wistar Furth (WF) strain is highly susceptible (Gould *et al.* 1989). By careful breeding, intermediate inbred lines can be produced which carry the homozygous WF/WF genotype throughout the genome except on a relatively small and interesting region where the animals are homozygous COP/COP. Such animal populations are referred to as congenic lines (Figure 4). The size of the homozygous COP/COP region is approximately 30 cM in congenic line CI and 1.5 cM in congenic line CII.

In this experiment, we are interested in the identification of genes that are differentially regulated among the parental strains (COP and WF) and the derived congenic lines (CI and CII) in mammary epithelial cells. By a standard protocol, mammary epithelial cells were harvested from untreated 12 week old females. Messenger RNAs were extracted, prepared, and then probed using a set of three Affymetrix Rat Genome U34 chips. In most cases, these mRNAs were pooled from samples of four genetically identical animals to reduce animal to animal variation. Intensity measurements were obtained for 26,379 genes recorded on 10

chip sets: 1 COP, 2 WF, 5 CI and 2 CII lines.

All data was processed through DNA-Chip Analyzer (Li and Wong, 2001). DNA-Chip Analyzer (dChip) uses a statistical model for probe level data to account for artifacts such as probe-specific biases. Corrected and normalized model-based estimates of gene expression were obtained for 25,248 genes (1131 were identified as outliers). A small fraction of the measurements are negative and these cannot be used by the model fitting procedures, so they are omitted for that purpose (796 out of 25,248). These observations can be included in the posterior probability calculations as long as they are set to a boundary value.

Both the GG and LNN models were used to categorize patterns of gene expression across the parental strains and derived congenic lines. For these four conditions, there are 15 possible expression patterns; however, if latent expression in each congenic matches one of the parentals, only four expression patterns are possible (see Figure 4). A null pattern consists of equivalent expression across the four conditions. The other three patterns allow for differential expression between the parental strains, with the congenic lines exhibiting the same mean expression as one of the parentals. Specifically, differential expression of the COP parent only is specified in pattern 1, between the congenics in pattern 2, and of the WF parent only in pattern 3. Note that differences in genotype need not imply differences in expression. Genes classified into the null pattern show equivalent expression across groups, but differ in genotype. Patterns 1 and 2 also allow for distinct genotype and expression patterns. Parameter estimates for each model are given in Table 6.

Under the GG model, 24,795 genes had posterior probability greater than 0.5 of being in the null pattern; 250, 86, and 111 genes were classified into patterns 1, 2, and 3, respectively. We did not classify six genes because for them no pattern had posterior probability greater than 0.5. The LNN model identified slightly more genes as differentially expressed. Specifically, 24,164 were classified into the null pattern; 447, 346, 280, were classified into patterns 1, 2, and 3; and 11 were not classified. Identified under both methods

were 24,119 (null), 217 (pattern 1), 51 (pattern 2), and 78 (pattern 3) genes.

Three genes identified as pattern 3 by the GG model are shown in Table 7. Two of these genes (J00801 and L08100) are known markers of mammary gland differentiation, and a common belief is that differentiation protects against tumor development. For each of these genes, the average intensity in the WF condition is higher than that observed in the COP or congenic lines. This indicates increased expression (and increased differentiation) in the WF, which is unexpected since the WF strain is tumor susceptible. It may be the case that not all forms of differentiation are associated with resistance. Preliminary data in other rat strains and other experiments are supporting this hypothesis (Gould, unpublished data). The third gene (J00772) is rat prostatein. Recent work suggests that this gene, normally associated with the ventral prostate, is strongly expressed in the stromal cells of the rat mammary gland (Watson and Gould, unpublished data). The GG calculations classify this gene as having elevated expression in WF, but the LNN calculations are equivicol, and consider it to be unchanged. Further study of this gene is warranted.

As a separate calculation, we analyzed the data from the WF and COP parentals only, omitting the congenics. Table 6 contains parameter estimates. The odds calculation under GG assumptions estimates 58 genes to be differentially expressed. Of these, 57 are also identified by the LNN model. Figure 5 gives a graph of the average intensities (across replicates) for the spots identified as changed in the GG model. These results are consistent with the multiple group analysis. Each of the 58 genes identified as differentially expressed in the two group analysis is also identified when comparing multiple groups. 48 of the 58 genes have posterior probability larger than 0.5 of being in pattern 1, 4 of the genes are in pattern 2, and 5 of the genes are in pattern 3. One gene was not classified. For both the four and two group analysis, Bayes factors indicated that the GG model fits better than the LLN model.

6 Discussion

We have extended empirical Bayes methodology for gene expression data to account for replicate arrays, multiple conditions, and a range of modeling assumptions. A discrete mixture model is introduced to describe the distribution of intensity measurements across multiple conditions. Using the model and Bayes rule, posterior probabilities are obtained from which inferences regarding differential expression patterns can be made.

The general hierarchical mixture model proposed accounts for differences among genes in their average expression levels, differential expression for a given gene among cell types, and measurement fluctuations. Since properties of individual experiments affect each of these features, the distributions governing them are to some extent experiment dependent. However, there are characteristics inherent to microarray data that are observed in many experiments, such as increasing variation with increasing mean and within-gene correlation. The present approach accounts for these properties and is also flexible in the sense that various choices can be made on the distributional forms at each level of the model. Obviously there are some advantages in the particular choices which we have made because the observation component is naturally matched to the distribution of the latent mean expressions.

We studied operating characteristics of the method to infer differential expression in the two group setting and found that error rates are well controlled. Under Gamma-Gamma model assumptions, the estimated positive predictive value was at least 94%, regardless of the proportion p of differentially expressed genes. The negative predictive value decreased from 97% to 80% with increasing p. This indicates that although the method may be missing genes, a positive identification is most likely an accurate one. Estimates of the sensitivity and specificity reflect this as well. The sensitivity increased from 65% to 80% with increasing p, while the specificity was at or above 95% for each value of p considered. The average false detection rate (FDR) was near 0.05 for all values of p, indicating that control of the type I

error rate is inherent to this empirical Bayes approach. Virtually identical numerical results were obtained under LNN assumptions. In addition, there was much overlap between specific genes identified using either approach.

These results suggest that error rates are reasonably low and that particular modeling assumptions might have only a minimal impact on the accurate identification of differentially expressed genes. We note that such results are preliminary, and further work is required before any such conclusions can be made in general. Only two groups having a fixed number of replicates in each group were considered in our simulation study. The study could be extended to evaluate error rates in the case of multiple conditions for a varying number of replicates. Additional model forms should also be considered, both for data simulation and odds calculations. We are currently investigating the effects of nonparametric assumptions on the latent mean distribution $\pi(\mu_g)$.

The method assumes that intensity measurements approximate some true underlying expression level. Thus, expression profiles must be normalized in such a way so that any systematic sources of variation have been removed. DNA Chip Analyzer (Li and Wong, 2001) was used here, but many other methods are available. We also note that mRNA samples were pooled across subjects. Of course, under some conditions, this can decrease measurement variability, thereby reducing the number of replicates required. However, owing to array specific effects, pooling does not eliminate the need for replication. Both Kerr *et al.* 2000 and Lee *et al.* 2000 stress the importance of replication in microarray studies. In addition to array effects, if outliers (e.g. contaminated samples) are present, pooling can lead to biased estimates of underlying expression. Optimal experimental designs which provide for maximum measurement accuracy using a minimum number of arrays have yet to be developed. This is an area that requires further investigation.

Finally, we note an interesting statistic which emerges from the odds of differential expression in the GG model (8) when the number of replicates $n_1 = n_2 = n$ per group

is relatively large compared to α_0 and ν. Up to a power and a proportionality constant, the odds favoring DE are

$$\frac{(\bar{x}_g + \bar{y}_g)/2}{\sqrt{\bar{x}_g\,\bar{y}_g}}$$

where $\bar{x}_g$ and $\bar{y}_g$ are respective sample means (on the raw scale) of expression measurements in the two groups. I.e., the odds are related to the ratio of the arithmetic to the geometric mean of the sample means. Considering the arithmetic-geometric mean inequality, this seems to be an interesting measure of the difference between the two samples. A similar analysis of the LNN odds shows that that one is related to the more familiar difference $\bar{\bar{x}}_g - \bar{\bar{y}}_g$; i.e the difference between the arithmetic means of the log-transformed responses (a t-like statistic). We think these facts give some credence to the model-based formulation and also suggest directions that the models could be extended.

Appendix: Estimation in the multiple group case: With data $\mathbf{d}_g$ governed by a mixture of the form (5), we introduce missing pattern indicators $z_{g,l}$ defined as one if the expression pattern of gene g is pattern l and zero otherwise. The *complete* data log likelihood is

$$l_c(\theta) = \sum_g \left\{ \sum_{k=0}^{m} z_{g,k} \left[\log f_k(\mathbf{d}_g) + \log(p_k) \right] \right\}$$

For θ fixed at θ_0, calculation of the expectation conditional on the observed data and θ_0 (E-step) gives

$$\hat{l}_c(\theta) = \sum_g \left\{ \sum_{k=0}^{m} \hat{z}_{g,k} \left[\log f_k(\mathbf{d}_g) + \log(p_k) \right] \right\}$$

$\hat{z}_{g,l}$ is the posterior probability of expression pattern l for gene g:

$$P(l|\mathbf{d}_g) = \frac{p_l f_l(\mathbf{d}_g)}{\sum_{k=0}^{m} p_k f_k(\mathbf{d}_g)}$$

where θ_0 parameterizes the densities f_k. We use the arithmetic mean of $\hat{z_{.,k}}$ to estimate p_k; **nlminb** in Splus provides estimates of θ (M-step). This process is repeated until there is convergence in the estimates. Results are checked from various starting configurations.

Acknowledgements: Code and data used in this article are available at the first author's web site `www.biostat.wisc.edu/~kendzior/`. This research was funded in part by training grant TA-CA 09565 for CMK, the NIH R01 grants CA64364 to MAN, and CA28954 and CA77494 to MNG.

References

1. Baldi, P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 2001; **17**(6): 509-519.

2. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, New York, 1996.

3. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statistical Society, Series B* 1977; **39**: 1-38.

4. Dudoit S, Yang YH, Speed TP, Callow MJ. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2001; to appear.

5. Efron B, Tibshirani R, Goss V, Chu G. Microarrays and their use in a comparative experiment. Technical Report # 37B/213 2000, Stanford University Department of Statistics.

6. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* 2001; **96**(456): 1151-1160.

7. Efron B, Morris C. Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society, Series B,* 1973; **35**: 379-421.

8. Efron B, Morris C. Stein's paradox in statistics. *Scientific American* 1977; **236**: 119-127.

9. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *PNAS* 1998; **95**(25): 14863-8.

10. Gould MN, Wang B, Moore CJ. Modulation of mammary carcinogenesis by enhancer and suppressor genes. In *Genes and Signal Transduction in Multistage Carcinogenesis*, ed. N.H. Colburn, Marcel Dekker, New York, 1989; pp 19-38.

11. Kass RE, Raftery AE. Bayes Factors. *Journal of the American Statistical Association* 1995; **90**(430): 773-795.

12. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J. Comp. Biol.* 2000; **7**: 819-837.

13. Lazzeroni L, Owen A. Plaid models for gene expression data. Department of Statistics Technical Report #211, 2000, Stanford University.

14. Lee MLT, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences* 2000; **97**(18): 9834-9839.

15. Lee MLT, Weining L, Whitmore GA, Beier D. Models for microarray gene expression data. *Proceedings of the ASA Joint Meetings, Atlanta, GA*, 2001.

16. Li C, Wong W. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS* 2001; **98**(1): 31-36.

17. Newton MA, Kendziorski CM, Richmond CS, and Blattner FR. On differential variability of expression ratios: Improving statistical inference about gene expression

changes from microarray data. *Journal of Computational Biology* 2001; **8**: 37-52.

18. Statistical Sciences, *S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.2*, Seattle: StatSci, a division of MathSoft, Inc., 1993.

19. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS* 1999; **96**: 2907-2912.

20. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 2001; **98**: 5116 - 5121.

21. West, M, Nevins JR, Marks JR, Spang R, Blanchette C, Zuzan H. DNA microarray data analysis and regression modeling for genetic expression profiling. Institute of Statistics and Decision Sciences Working Paper #15, 2000a.

22. West M. Bayesian regression analysis in the "large p, small n" paradigm. Institute of Statistics and Decision Sciences Working Paper #22, 2000b.

Table 1: Summary of the GG and LNN models applied data simulated under GG and LNN assumptions. For GG-simulated data 1968 genes (out of 10000) are truly differentially expressed. The number is 1952 for the LNN-simulated data. The number of genes identified as differentially expressed (odds > 1) by each model (GG, LNN) is indicated by (I); the number correctly identified is indicated by (C).

Model	GG data	LNN data
GG (I)	1473	1469
LNN (I)	1448	1488
Both (I)	1430	1445
GG (C)	1396	1389
LNN (C)	1379	1409
Both (C)	1373	1380

Table 2: Summary of parameter estimates for GG model applied to GG simulated data. Parameter estimates are averaged over 100 simulations; standard error is shown in parentheses. For each simulation,
$(\alpha, \alpha_0, \nu) = (10, 0.9, 0.5)$.

p	0.1	0.2	0.3	0.4	0.5
$\hat{\alpha}$	10.001 (0.0098)	9.997 (0.0087)	9.995 (0.0104)	9.993 (0.0099)	10.009 (0.01)
$\hat{\alpha_0}$	0.900(0.0016)	0.900(0.0015)	0.897(0.0015)	0.900(0.0014)	0.901(0.0013)
$\hat{\nu}$	0.499(0.0012)	0.500(0.0011)	0.500(0.0011)	0.500(0.0012)	0.500(0.0011)
$\hat{p}$	0.101(0.0005)	0.201(0.0007)	0.298(0.0008)	0.401(0.0008)	0.501(0.0009)

Table 3: Summary of parameter estimates for LNN model applied to LNN simulated data. Parameter estimates are averaged over 100 simulations; standard error is shown in parentheses. For each simulation, $(\mu_0, \sigma, \tau_0) = (2.33, 0.33, 1.39)$.

p	0.1	0.2	0.3	0.4	0.5
$\hat{\mu_0}$	2.328(0.0018)	2.333(0.0021)	2.329(0.0015)	2.332(0.0017)	2.33(0.0015)
$\hat{\sigma}$	0.332(0.0001)	0.331(0.0001)	0.332(0.0002)	0.331(0.0002)	0.332(0.0002)
$\hat{\tau_0}$	1.390(0.0013)	1.386(0.0014)	1.390(0.0011)	1.392(0.0012)	1.391(0.0012)
$\hat{p}$	0.1(0.0005)	0.2(0.0007)	0.3(0.0009)	0.399(0.0009)	0.501(0.0008)

Table 4: Operating characteristics of GG model evaluated on GG (upper) and LNN (lower) simulated data. Estimates are averaged over 100 simulations; standard error is shown in parentheses.

p	0.1	0.2	0.3	0.4	0.5
Sens	0.670(0.002)	0.703(0.002)	0.728(0.001)	0.753(0.001)	0.778(0.001)
Spec	0.996(0.000)	0.990(0.000)	0.982(0.000)	0.968(0.000)	0.948(0.001)
PPV	0.950(0.001)	0.947(0.001)	0.945(0.001)	0.940(0.001)	0.938(0.001)
NPV	0.964(0.000)	0.930(0.000)	0.894(0.001)	0.855(0.001)	0.810(0.001)
FDR	0.050(0.001)	0.053(0.001)	0.055(0.001)	0.060(0.001)	0.062(0.001)

p	0.1	0.2	0.3	0.4	0.5
Sens	0.688(0.002)	0.718(0.002)	0.742(0.001)	0.764(0.001)	0.785(0.001)
Spec	0.996(0.000)	0.991(0.000)	0.983(0.000)	0.972(0.000)	0.955(0.000)
PPV	0.953(0.001)	0.950(0.001)	0.950(0.001)	0.948(0.001)	0.946(0.000)
NPV	0.966(0.000)	0.934(0.000)	0.899(0.000)	0.861(0.001)	0.816(0.001)
FDR	0.047(0.001)	0.050(0.001)	0.050(0.001)	0.052(0.001)	0.054(0.000)

Table 5: Operating characteristics of LNN model evaluated on GG (upper) and LNN (lower) simulated data. Estimates are averaged over 100 simulations; standard error is shown in parentheses.

p	0.1	0.2	0.3	0.4	0.5
Sens	0.661(0.002)	0.693(0.002)	0.716(0.001)	0.741(0.001)	0.764(0.001)
Spec	0.996(0.000)	0.991(0.000)	0.984(0.000)	0.973(0.000)	0.957(0.000)
PPV	0.954(0.001)	0.951(0.001)	0.951(0.001)	0.948(0.001)	0.946(0.001)
NPV	0.963(0.000)	0.928(0.000)	0.891(0.001)	0.849(0.001)	0.802(0.001)
FDR	0.046(0.001)	0.049(0.001)	0.049(0.001)	0.052(0.001)	0.054(0.001)

p	0.1	0.2	0.3	0.4	0.5
Sens	0.693(0.002)	0.721(0.002)	0.747(0.001)	0.769(0.001)	0.791(0.001)
Spec	0.997(0.000)	0.991(0.000)	0.984(0.000)	0.972(0.000)	0.955(0.000)
PPV	0.957(0.001)	0.953(0.001)	0.952(0.001)	0.949(0.001)	0.946(0.000)
NPV	0.967(0.000)	0.934(0.000)	0.901(0.000)	0.864(0.001)	0.821(0.001)
FDR	0.043(0.001)	0.047(0.001)	0.048(0.001)	0.051(0.001)	0.054(0.000)

Table 6: Parameter estimates for GG ($\theta = (\alpha, \alpha_0, \nu)$) and LNN ($\theta = (\mu_0, \sigma, \tau_0)$) models used in two group comparisons between parentals and in four group comparisons among the parentals and derived inbred lines.

Model	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{p_0}$	$\hat{p_1}$	$\hat{p_2}$	$\hat{p_3}$
GG (2 groups)	12.490	0.919	35.842	0.998	0.002	NA	NA
LNN (2 groups)	6.775	0.292	1.193	0.993	0.007	NA	NA
GG (4 groups)	16.738	0.883	24.398	0.985	0.012	0.002	0.001
LNN (4 groups)	6.741	0.257	1.221	0.975	0.017	0.004	0.004

Table 7: Expression averages (left) and posterior pattern probabilities (right) for several genes classified as having expression pattern 3 by the GG model (see Fig. 4). For each gene, the probability vector from the GG model is in the first row and the one from the LNN model is in the second row.

	Group				Expression Pattern			
Gene ID	Cop	CI	CII	WF	null	P1	P2	P3
J00801	3066.3	4777.0	995.3	9082.9	0.05	0	0	0.95
					0.04	0	0	0.96
L08100	4367.5	4002.6	1278.3	14162.3	0	0	0	1
					0	0	0	1
J00772	392.0	325.8	121.7	678.9	0.04	0	0	0.96
					0.97	0.01	0.00	0.02

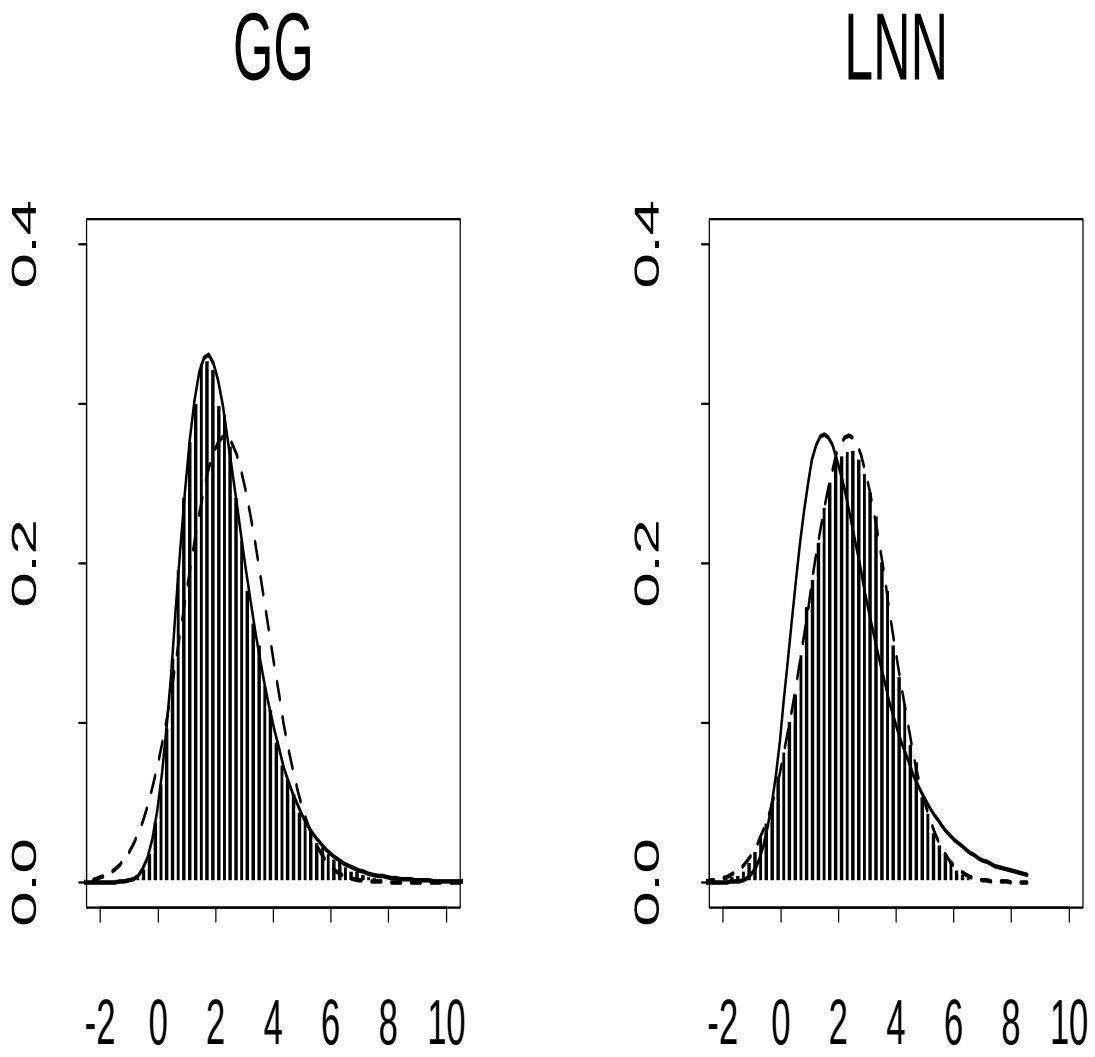

Figure 1: Histograms are of intensities (log scale) simulated under the GG (left) or LNN (right) model. Solid line is fitted marginal density from the GG model and dashed line is fitted marginal density from the LNN model.

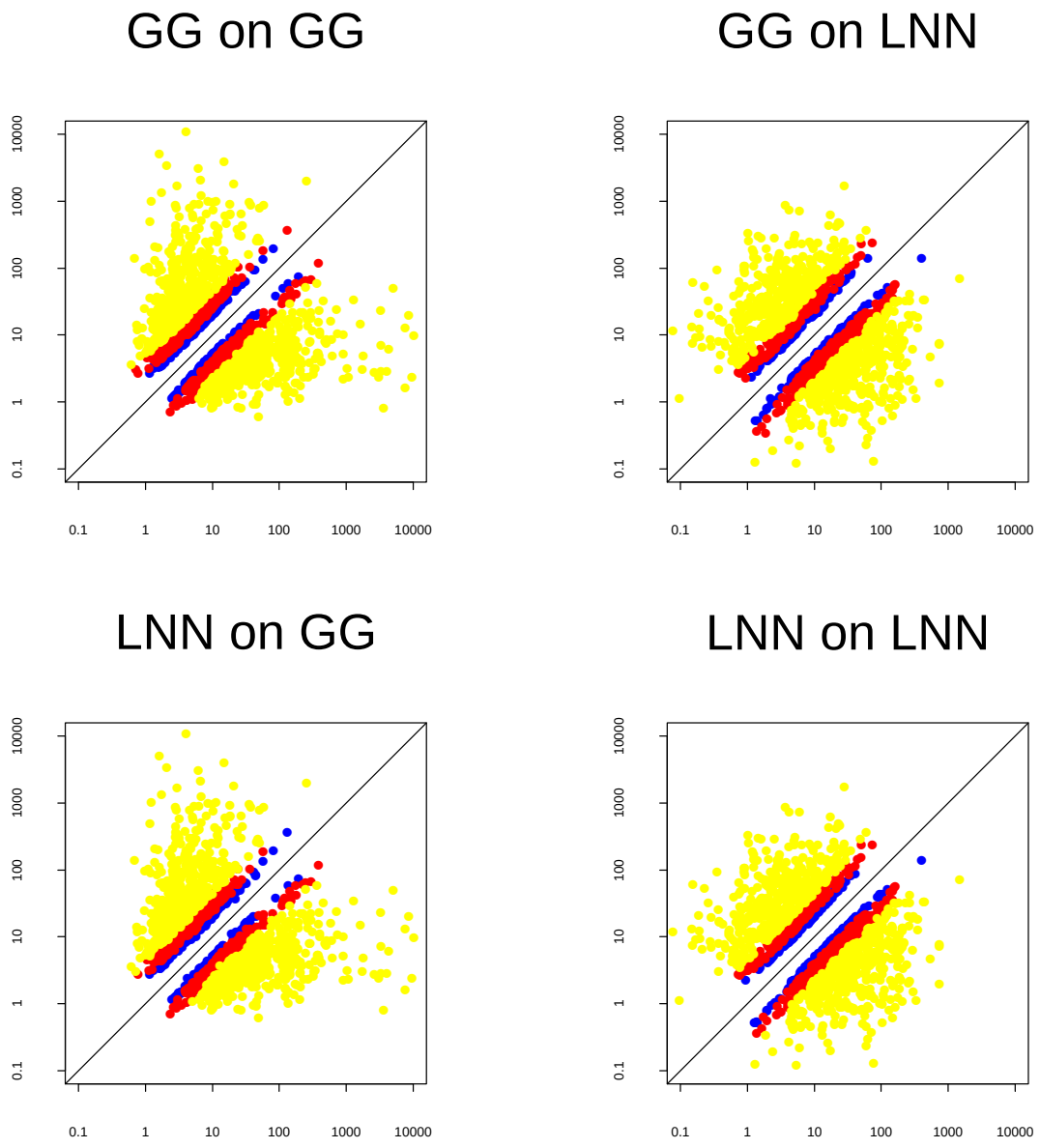

Figure 2: Shown are intensity values averaged across replicates for data simulated under GG (left) and LNN (right) assumptions. Only spots which exhibit significant differential expression as determined by the GG (upper) or LNN (lower) model are shown. Spots are colored by magnitude of $\rho = \log_{10}(\text{odds})$. Blue corresponds to $0 < \rho \leq 1$, red to $1 < \rho \leq 5$, and yellow to $\rho > 5$.

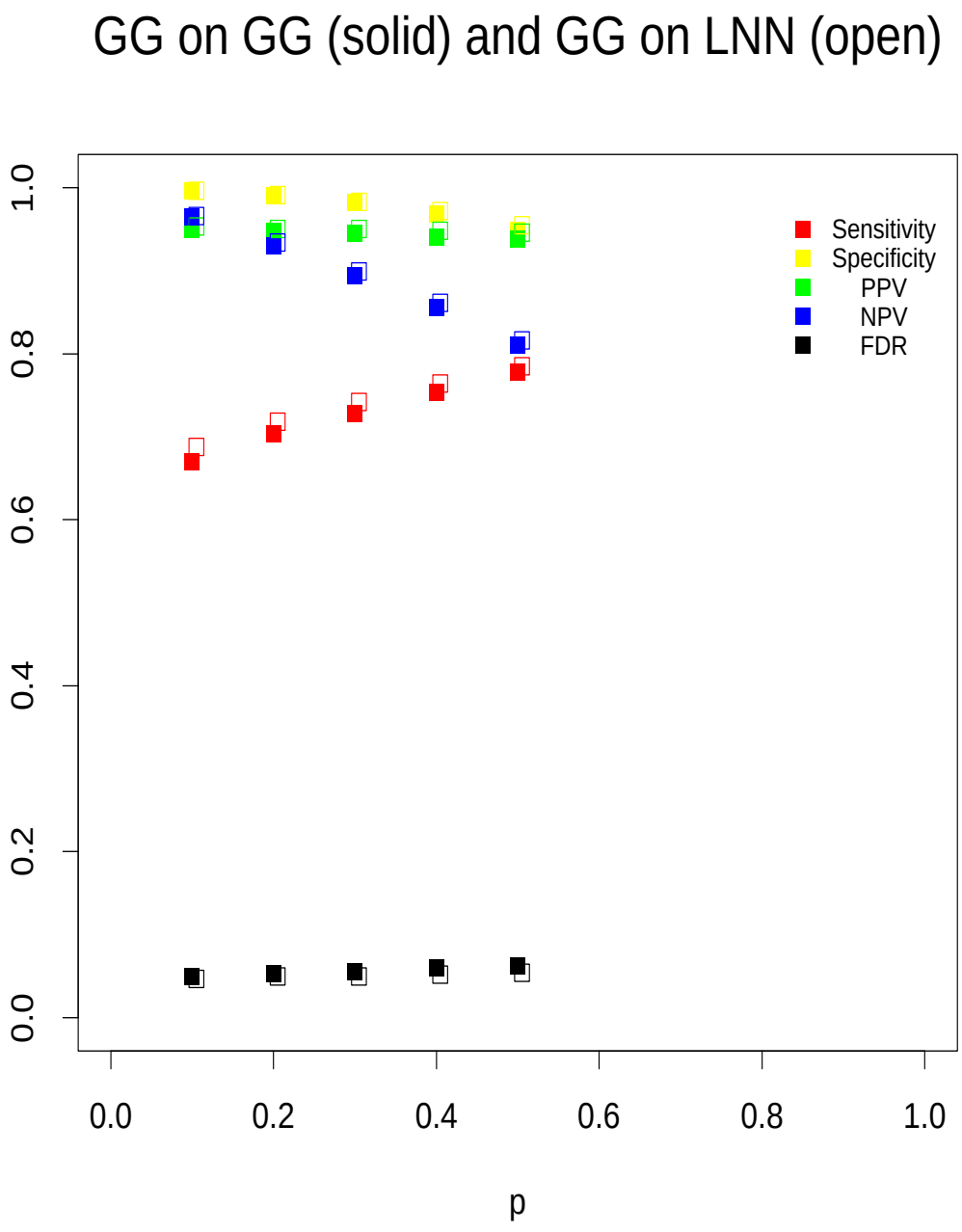

Figure 3: Operating characteristics: Results are shown for the GG model applied to simulated data. To minimize overlap, we jittered the horizontal component. Closed characters imply identical model and simulation assumptions (GG model applied to GG data); open characters imply the opposite (GG model applied to LNN data).

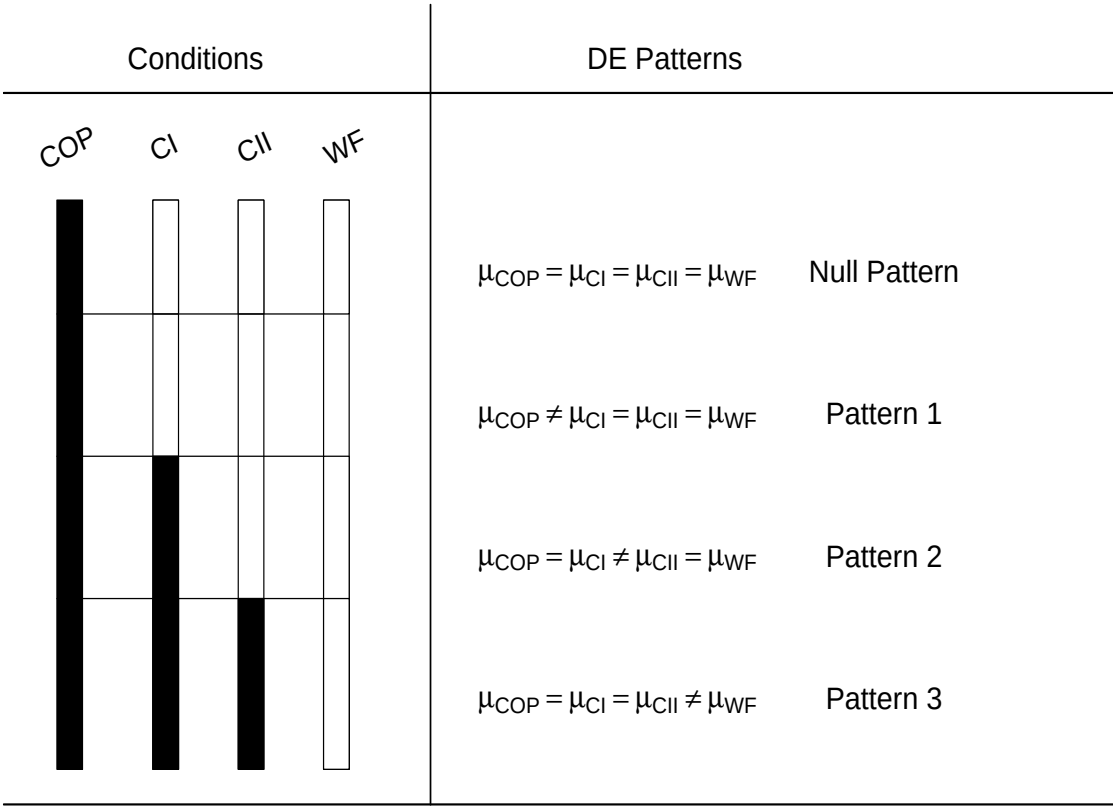

Figure 4: Schematic diagram showing animal lines (conditions) from which mRNAs were obtained (left) along with differential expression patterns (right). Genotypes shown in black (COP/COP) and white (WF/WF) are not drawn to scale (the homozygous COP/COP region is approximately 30 cM in congenic line CI and 1.5 cM in CII). True expression intensities for each group are denoted by μ. Note that differences in genotype do not imply differences in expression.

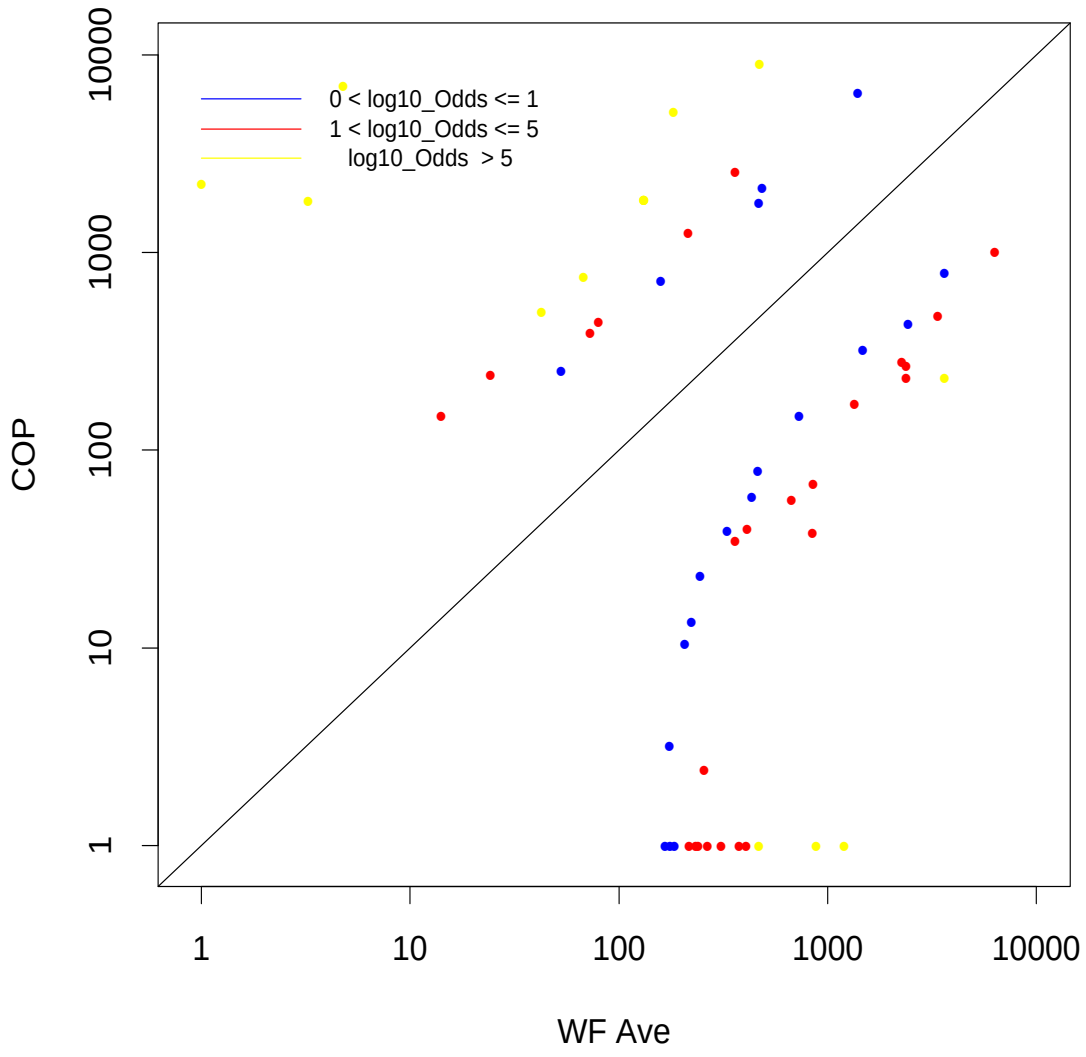

Figure 5: Average intensities across replicates for the WF and COP data. Only spots which exhibit significant differential expression (as determined by the GG model) are shown.