2

Hierarchical Mixture Models for Expression Profiles

MICHAEL A. NEWTON, PING WANG, AND CHRISTINA KENDZIORSKI
University of Wisconsin at Madison

Abstract

A class of probability models for inference about alterations in gene expression is reviewed. The class entails discrete mixing over patterns of equivalent and differential expression among different mRNA populations, continuous mixing over latent mean expression values conditional on each pattern, and variation of data conditional on latent means. An R package `EBarrays` implements inference calculations derived within this model class. The role of gene-specific probabilities of differential expression in the formation of calibrated gene lists is emphasized. In the context of the model class, differential expression is shown to be not just a shift in expected expression levels, but also an assertion about statistical independence of measurements from different mRNA populations. From this latter perspective, `EBarrays` is shown to be conservative in its assessment of differential expression.

2.1 Introduction

Technological advances and resources created by genome sequencing projects have enabled biomedical scientists to measure precisely and simultaneously the abundance of thousands of molecular targets in living systems. The effect has been dramatic, not only for biology, where now the cellular role for all genes may be investigated, or for medicine, where new drug targets may be found and new approaches discovered for characterizing and treating complex diseases, the effect has also been dramatic for statistical science. Many statistical methods have been proposed to deal with problems caused by technical and biological sources of variation, to address questions of coordinated expression and differential expression, and to deal with the high dimension of expression profiles compared to the number of profiles. Our interest is in the question of differential expression. We do not attempt to review the considerable body of

statistical research that addresses this question; we focus here on methods for this problem that are related to a class of *hierarchical mixture models*.

A model is hierarchical if it describes observed variation using both latent random variables and the conditional variation of data given realizations of these latent quantities. In our work, the latent random variables include gene- and condition-specific expected values, these being the target quantities that one would measure in the absence of either biological or technical variation. Hierarchical models naturally incorporate multiple sources of variation, and they have an important role in the analysis of experiments with few microarrays because they can channel relevant information from other genes into gene-specific calculations, thus improving sensitivity.

The term mixture model can be used in a very broad sense to describe distributions; however, in expression work it has the following narrow interpretation: gene-specific hypotheses about differential expression are treated as latent discrete random variables. In comparing two mRNA populations, for example, it is as if a gene tosses a coin to decide whether or not it is differentially expressed, and then produces data distributed according to the particular outcome. Mixture models are convenient in structuring high-dimensional inference; genes become apportioned to different components of the mixture model. Often this modeling is done late in the data analysis stream: a mixture is fit to one-dimensional gene-specific summary measures (e.g., p values) rather than to the full data, and thus it may be unable to recover information lost by forming these summaries. Another problem is that some mixture methods rely on permutation to develop null distributions. This can be effective but it can fail when there is limited replication, as is often the case.

The first empirical Bayesian analysis of expression data was published in 2001. Focusing on preprocessed, two-channel microarray data, our group noted an inefficiency of the naive fold change estimator R/G, obtained from each gene's intensity measurements R and G in the two color channels on a spotted cDNA microarray (Newton et al. 2001). Our model-based estimate of fold change was $(R + c)/(G + c)$ for a statistic c which depends on sources of variation affecting the intensity measurements and which is computed from data on all genes. This modified fold-change estimator emerged as an intermediate between the posterior mode and posterior mean of the true fold change in the context of a specific Gamma-Gamma (GG) hierarchical model. We showed by simulation how this estimator has reduced mean squared error (log scale) and also how the gene ranking is improved. In addition, this 2001 *JCB* paper addressed the question of testing for differential expression in the context of a parametric hierarchical mixture model, and gave formulas for the posterior probability and odds of differential expression. The paper also

noted a statistical curiosity of testing in this mixture model context, namely, that the number of genes that may be confidently declared to be differentially expressed may be much smaller than the estimated proportion of genes that are truly differentially expressed. This concept is helpful in formalizing power calculations. Further, in spite of improvements in statistical computing, we also recognized in this first paper the importance of computationally efficient methods in the domain of high-throughput data; our models were sufficiently simple that Markov chain Monte Carlo methods could be safely avoided.

The 2001 *JCB* paper concerned both testing and estimation for high-dimensional microarray data based on novel hierarchical and mixture-modeling structures. However, the delivered methodology remained rather limited; it handled single-slide spotted-array data comparing expression profiles in two conditions. There was nothing intrinsic to the model development that forced such restrictions, and so we pursued extensions that allowed replicate expression profiles in multiple mRNA populations (Kendziorski et al. 2003). There, we extended the GG calculations to this setting and we also developed parallel calculations based on a log-normal-normal (LNN) hierarchical specification. Emphasis was taken away from estimation of fold change and was transferred to computing posterior probabilities for various patterns of equality among gene- and condition-specific expected values. This has more relevance for inference with multiple mRNA populations. Tools to implement the multigroup inference calculations were offered in the Bioconductor package `EBarrays`.

Data analysts tend to favor methods that are simply structured and that have little reliance on modeling assumptions. A popular approach to differential expression, for example, is to apply *ordinary* statistical procedures (such as the *t*-test) separately to each gene, and then to paste the inferences together in some reasoned way (e.g., Dudoit et al. 2002). Although often effective, this approach usually rests on implicit assumptions about variation and it can suffer inefficiencies when shared properties of genes are not well accommodated. `EBarrays`, on the other hand, delivers inference summaries by attempting to capture the relevant sources of variation of the entire high-dimensional expression profile. It is explicit about the underlying assumptions:

(i) Parametric observation component (log-normal or Gamma)
(ii) Parametric mean component (conjugate to observation component)
(iii) Constant coefficient of variation
(iv) Only marginal information (rather than among-gene dependence) is relevant

Much experience with the package indicates good operating characteristics, especially when the number of replicate chips per condition is low. In examples where the parametric fit is poor it is beneficial to have more flexible methods. Work since Kendziorski et al. (2003) has investigated these assumptions, examined their significance, and generalized the methodology.

Adopting our proposed mixture structure, but not the hierarchical modeling elements, Efron et al. (2001) described a nonparametric empirical Bayesian analysis for assessing differential expression. The nonparametric nature of the analysis is appealing, since it seems to alleviate parametric constraints and may thus be favored in routine data analysis. However, the flexibility is somewhat illusory; it enters mainly in estimation of a one-dimensional distribution of gene-specific summary measures. The proposed method relies on permutation to assess a common null distribution (so it can fail when the number of replicate microarrays is low), and takes advantage of the large number of genes to develop the nonparametric density estimate. Further, assumptions about the suitability of the proposed gene-specific summary statistic are left implicit. Importantly, the Efron et al. paper may have been the first to relate gene-specific posterior probabilities of equivalent expression to rates of false detection in a reported list of genes. Much subsequent research on the control of the false discovery rate (FDR) seems to stem from this observation.

In the following sections we visit a few topics relevant to inference about expression alterations that seem to be notable developments since our first work in the area.

2.2 Dual Character of Posterior Probabilities

In the context of multiple simultaneous hypothesis testing, posterior probabilities have a curious dual character that other testing summaries lack. The duality is almost transparant once stated, but we think it is worth noting here because it simplifies the interpretation of gene lists.

Each gene j from a large set of J genes may or may not be differentially expressed between two mRNA populations. We say it is equivalently expressed, EE_j, if it is not differentially expressed; data are analyzed to assess this null hypothesis. A Bayesian (or empirical Bayesian) analysis yields the posterior probability $e_j = P(EE_j|\text{data})$; a non-Bayesian analysis might yield a p value or some other gene-specific summary statistic.

Genes exhibiting the strongest evidence for differential expression will be those with the smallest e_j, and one could naturally consider forming a list of *discoveries*, $\mathcal{D} = \{j : e_j \leq \tau\}$, for some threshhold τ. The duality is this: gene j gets to be in $\mathcal{D}$ by virtue of the small magnitude of e_j. At the same time, e_j

is the probability (conditional on the data) that this assignment is a mistake. In other words, it is the probability of a type I error; that gene j should not have been placed on the list of differentially expressed genes. The magnitude e_j conveys both a decision about j and the conditional probability of a faulty decision. Other gene-specific summaries, like p values, do not have this dual character.

The property is useful for multiple simultaneous inference because the expected number of false discoveries (conditional on the data) is simply

$$cFD(\tau) = \sum_j \underbrace{e_j}_{\text{error rate}} \underbrace{1[e_j \leq \tau]}_{\text{discovery}} \tag{2.1}$$

and the conditional false discovery rate is $cFDR(\tau) = cFD(\tau)/N(\tau)$, where $N(\tau) = \sum_j 1[e_j \leq \tau]$ is the size of the list. A list $\mathcal{D}$ formed from all genes for which $e_j < 5\%$, for example, has $cFDR$ less than 5%. A more refined usage tunes τ to set the conditional false discovery rate at some value like 5%, and has been called the *direct posterior probability approach* to controlling this rate (Newton et al. 2004).

In Efron et al. (2001), e_j was called the local FDR because it measured the conditional type I error rate for that specific gene. Storey (2002) criticized the unmodified use of e_j's for inference because they lack error rate control simultaneously for a list of discovered genes. Averages of e_j values over a short list of reported genes convey a more useful multiple-testing quantity. Storey (2002) introduced the q value as a gene-specific inference measure that carries a multiple-testing interpretation. In our notation, the q value for gene j is $q_j = cFDR(e_j)$. This is the expected proportion of type I errors among those genes k with e_k no larger than that of the input gene j. The procedure that rejects all null hypotheses EE_j for which $q_j \leq 5\%$ targets a *marginal* FDR of 5%. Literature on q values centers on the analysis not of raw data (or e_j's) but of p values derived from separate gene-specific hypothesis tests. Since p values do not have the dual character described above, their distribution needs to be modeled as a mixture in order that q values can be derived. In this way, modeling is transferred from the full data down to gene-specific p values. An advantage is that it is easier to be nonparametric with one-dimensional statistics; a disadvantage is that information may have been lost in first producing the gene-specific p values.

The dual character of posterior probabilities was pointed out in Newton et al. (2004), though the issue is understood in other work (Genovese and Wasserman 2002; Storey 2003; Müller et al. 2005). Notably, Müller et al. (2005) tackle the issue from a more formal Bayesian position, and study list-making inference as a general decision problem.

2.3 Differential Expression as Independence

Consider replicate profiles available from two mRNA populations, and allow that preprocessing has removed systematic sources of variation. Gene j provides measurements $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,m})$ in one condition and $y_j = (y_{j,1}, y_{j,2}, \ldots, y_{j,n})$ in the second condition. The concept of equivalent expression, EE_j, and its counterpart differential expression, DE_j, are hypotheses that require some definition in terms of their effect on the probability density $p(x_j, y_j)$. Most studies focus modeling on the null hypothesis EE_j. One could state this in terms of a common expectation $\mu_j = E(x_{j,i}) = E(y_{j,k})$ (for any chips i, k) that the measurements are targeting, or one could state it in terms of exchangeability of all the measurements. Then under the null hypothesis, permutation of microarray labels would be valid, and this could be used to generate a null distribution of a test statistic (e.g., Dudoit et al. 2002). Such an approach can be effective when the number of microarrays is large, but notice that the approach avoids defining differential expression as anything more than the opposite of EE_j. The approach adopted in `EBarrays` does not require permutation and can be applied when there are very few replicate microarrays. In it, DE_j is defined as independence of x_j and y_j. This independence is marginal with respect to any gene level parameters and is conditional on genomic-level hyperparameters that are not specific to gene j. That is, gene j is differentially expressed if measurements from one condition are not useful predictors of measurements in the second condition. By contrast, all measurements x_j and y_j on a gene j that is equivalently expressed are correlated by virtue of having a shared, latent, random mean.

For the sake of demonstration, consider a comparison in which x_j and y_j are univariate ($m = n = 1$). The calculations embodied in the LNN model of `EBarrays` consider that these (log) expression values are conditionally independent normally distributed variables with means (μ_1, μ_2) and with a common variance σ^2. Further, the means (μ_1, μ_2) are random effects (suppressing the gene dependence); their marginal distribution is conjugate, being normal centered at a genomic mean μ_0 and having variance τ_0^2. Thus x_j and y_j have equal marginal distributions obtained after integrating the latent means: Normal$(\mu_0, \sigma^2 + \tau_0^2)$. The issue of EE_j or DE_j enters into the dependence between x_j and y_j. We assert that on EE_j, $\mu_1 = \mu_2$ with probability 1, and, further, that on DE_j, the component means are independent. Upon integrating the latent means, we have (1) exchangeability of gene-level measurements on the null EE_j, and (2) independence between x_j and y_j on the alternative DE_j. Mechanistically, we can imagine that on EE_j a single mean value is realized for the gene j, and then all the observations are generated as a random sample

under that parameter setting. Alternatively, on DE_j, each mRNA population selects its mean value independently of the others, from the same distribution, and measurements arise conditionally on these different means.

A model is fully specified when in addition we consider the discrete mixing on EE_j (probability p_0) and DE_j (probability p_1). The marginal distribution of gene-level data is

$$p(x_j, y_j) = p_0 f(x_j, y_j) + p_1 f(x_j) f(y_j), \tag{2.2}$$

where, conveniently, $f(\ \ )$ returns a marginal density of its argument treated as a conditional random sample given a common, latent, random mean. For instance in the case considered,

$$f(x_j, y_j) = p(x_j, y_j | EE_j) = \int p(x_j|\mu)\, p(y_j|\mu)\, \pi(\mu)\, d\mu,$$

where $\pi(\ \ )$ is a normal univariate conjugate prior, $p(x_j, y_j)$ is a normal density with common margins, as above, and with correlation $1/(1 + \sigma^2/\tau_0^2)$ between x_j and y_j owing to them having a common, latent mean. General formulas for this LNN case, and for the GG case, are presented in Kendziorski et al. (2003). Two-group comparisons in `EBarrays` are based on (2.2); the code allows other combinations and user input of the function $f(\ \ )$.

Gene-level inference is based on posterior probabilities, such as

$$e_j = P(EE_j | x_j, y_j) = p_0 f(x_j, y_j)/p(x_j, y_j).$$

Any decision about gene j is based on e_j; in this normal no-replicate case, for example, the odds favor differential expression if

$$\frac{1 - e_j}{e_j} > 1 \qquad \Leftrightarrow \qquad (x_j - y_j)^2 > C \tag{2.3}$$

where, more precisely,

$$C = \frac{4\sigma^2 (a - \mu_0)^2}{\sigma^2 + 2\tau_0^2} + \frac{4\sigma^2 \left(\sigma^2 + \tau_0^2\right)}{\tau_0^2} \left[\log \frac{p_0}{p_1} + \frac{1}{2} \log \frac{\left(\sigma^2 + \tau_0^2\right)^2}{\sigma^2 \left(\sigma^2 + 2\tau_0^2\right)} \right].$$

In other words, we favor DE_j if the measurements in the two conditions are sufficiently far apart, the necessary distance depending on overall expression $a = (x_j + y_j)/2$ and global parameters that delineate the different sources of variation. Ultimately, the analysis is *empirical* Bayes because these global parameters are estimated from the genomic data using an EM algorithm to maximize a marginal likelihood.

We have presented two views about differential expression. The rather direct view considers DE_j as a difference in expected expression measurements between the two mRNA populations, as revealed by a large difference in the observed expression values for gene j (the difference $x_j - y_j$ above, or the

difference of averages in the case of replication). The alternative view holds that differential expression corresponds to independence of measurements between the two conditions; one sample is not a useful predictor of the other. This view, which is not so widely appreciated, has the advantage of supporting a specific alternative hypothesis to EE_j with which we can develop posterior inference. Conveniently, these two seemingly different views are two sides of the same coin.

There is an intermediate ground on which DE_j entails a shift in expected expression without marginal independence between x_j and y_j. However, this formulation is related to a nonidentifiability of the mixture model, and thus is difficult to work with (see Newton et al. 2004). It is possible to establish that inferences derived using the independence view of DE_j (i.e., using `EBarrays`) are conservative if some positive dependence happens to exist between x_j and y_j on DE_j. Wang and Newton (2005) show that when σ^2/τ_0^2 is sufficiently small, then the `EBarrays` even–odds threshold C [see (2.3)] is larger than the threshold C' one would have computed if one were supplied with the correct correlation between x_j and y_j. In other words `EBarrays` is conservative: DE_j is harder to declare using `EBarrays` than if you know the true distribution, and so you make fewer claims of differential expression. It is a rather realistic condition, furthermore, that σ^2/τ_0^2 is small, since we expect variation within a gene (certainly variation of an average in the case of replication) to be small compared to the variation between genes.

2.4 The Multigroup Mixture Model

Pairwise comparisons are the bread and butter of statistics, but they may not be suitable when analyzing data from more than two mRNA populations. Extending (2.2) to three groups by the inclusion of data z_j, we mix over $\nu = 4$ possible discrete patterns of differential expression and one pattern of equivalent expression:

$$\begin{aligned} p(x_j, y_j, z_j) = {} & p_0 f(x_j, y_j, z_j) + p_1 f(x_j) f(y_j, z_j) \\ & + p_2 f(x_j, z_j) f(y_j) + p_3 f(x_j, y_j) f(z_j) \\ & + p_4 f(x_j) f(y_j) f(z_j). \end{aligned} \tag{2.4}$$

For instance, p_3 is the proportion of genes for which x_j and y_j are equivalently expressed while being differentially expressed from z_j, and p_4 is the proportion of genes that are differentially expressed among all three conditions. More generally, let $\mathbf{d}_j = (d_{j,1}, \ldots, d_{j,N})$ denote the vector holding all measurements on gene j taken across all conditions. We mix over equivalent expression and ν patterns of differential expression so that the joint distribution

$p(\mathbf{d}_j) = \sum_{k=0}^{\nu} p_k f_k(\mathbf{d}_j)$, where p_k is the overall proportion of genes governed by the kth pattern and f_k is the distribution of data conditional on that pattern. The patterns are hypotheses about possible clustering of the expected expression levels across the N measurements, and so, like the case above with $\nu = 4$, each f_k becomes a product of contributions from each component of the clustering. The null pattern $k = 0$ corresponds to $\mu_j = E(d_{j,s})$ being the same for all samples $s \in S = \{1, 2, \ldots, N\}$. Any pattern k partitions S into $r(k)$ mutually exclusive and exhaustive subsets $\{S_{k,i} : i = 1, 2, \ldots, r(k)\}$ on each of which the expected expression level is constant. To complete the specification, we write

$$f_k(\mathbf{d}_j) = \prod_{i=1}^{r(k)} f(\mathbf{d}_{j,S_{k,i}}) = \prod_{i=1}^{r(k)} \int \left(\prod_{s \in S_{k,i}} f_{\text{obs}}(\mathbf{d}_{j,s}|\mu) \right) \pi(\mu)\, d\mu, \qquad (2.5)$$

where $\pi(\mu)$ is a random effects distribution governing the latent, gene-specific expression means and f_{obs} is the observation component of the hierarchical model. Model fitting amounts to estimating the mixing proportions p_k, parameters of the observation component, and parameters of the mean component $\pi(\mu)$.

As a brief illustration, we reconsider data on gene expression in mammary epithelial tissue from a rat model of breast cancer. Each of 10 pools of mRNA was probed with an Affymetrix U34 chip set having 26, 379 distinct probe sets; the 10 pools represent rats of four different genetic strains (1 Copenhagan; 5 Wistar Furth; 2 Congenic I; 2 Congenic II) where each congenic strain was genetically identical to the Wistar Furth parental strain except for a small genomic region in which the genome is homozygous for Copenhagan alleles, at least one of which confers resistance to the development of breast cancer (Shepel et al. 1998; Kendziorski et al. 2003). Expression alterations among these groups are relevant to understanding the Copenhagan strain's resistance to breast cancer.

Table 2.1 shows the $\nu = 14$ patterns of differential expression among the 4 mRNA populations (strains), and the overall equivalent expression pattern. Previous analysis of these data (Kendziorski et al. 2003) was restricted to a subset of four patterns, as code at that point was not sufficiently flexible to handle arbitrary sets of patterns. Figure 2.1 shows the proportions of genes satisfying each pattern based on fitting the LNN model in `EBarrays`.

Detectable differential expression is rather limited in this example, as an estimated 92.7% of genes are equivalently expressed among the four rat strains. DE pattern $k = 4$ represents one case of interest as it concerns genes that may be altered by the process of congenic formation. Filtering by gene-specific posterior probabilities of this pattern $P(\mu_{j,1} = \mu_{j,2} \neq \mu_{j,3} = \mu_{j,4}|\text{data}) =: 1 - e_j$,

Table 2.1. *Patterns of DE Among Four Rat Strains*

k	Mean pattern[a]	k	Mean pattern
0	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	8	$\mu_1 = \mu_2 = \mu_4 \neq \mu_3$
1	$\mu_1 \neq \mu_2 = \mu_3 = \mu_4$	9	$\mu_1 = \mu_2 \neq \mu_3 \neq \mu_4$
2	$\mu_1 = \mu_4 \neq \mu_2 = \mu_3$	10	$\mu_1 = \mu_3 \neq \mu_2 \neq \mu_4$
3	$\mu_1 = \mu_3 = \mu_4 \neq \mu_2$	11	$\mu_1 = \mu_4 \neq \mu_2 \neq \mu_3$
4	$\mu_1 = \mu_2 \neq \mu_3 = \mu_4$	12	$\mu_1 \neq \mu_2 = \mu_4 \neq \mu_3$
5	$\mu_1 = \mu_2 = \mu_3 \neq \mu_4$	13	$\mu_1 \neq \mu_2 \neq \mu_3 = \mu_4$
6	$\mu_1 \neq \mu_2 = \mu_3 \neq \mu_4$	14	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$
7	$\mu_1 = \mu_3 \neq \mu_2 = \mu_4$		

[a] (1) Copenhagan, (2) Wistar Furth, (3) Congenic I, (4) Congenic II. Here, μ_i refers to the expected expression level for mRNA population i.

we can apply the direct posterior probability approach (2.1) to controlling FDR. We find that five probe sets constitute a 5% $cFDR$ short list of genes satisfying this DE pattern. These probe sets have $e_j \leq 0.013$. One of the interesting ones, `rc_AI105022_at`, corresponds to Cullin-3, a gene involved in the ubiquitin cycle and related to breast cancer tumor suppression (Fay et al. 2003). Investigating the biological significance of altered genes such as this is part of ongoing research; it is important to have tools like `EBarrays` which can efficiently sort and calibrate genes by alterations of interest.

2.5 Improving Flexibility

Utility of results from the hierarchical mixture model analysis, as obtained from `EBarrays`, is limited by the suitability of the four structural modeling assumptions described in the introduction. Each of these has been the subject of analysis, and we find that certain assumptions seem to be more important than others. For example, the use of a parametric observation component is often innocuous. Tools in `EBarrays` provide diagnostic *qq*-plots for this component; both Gamma and log-normal distributions often fit well, though a search for improved robust alternatives would be valuable. Calculations in Gottardo et al. (in press) allow log-t errors, and thus are less susceptible to heavy-tailed observations.

The diagnostic plots often indicate suitability of the observation component; however, marginal diagnostics can suggest an overall poor fit from `EBarrays`. This has to do with inflexibility of the distribution $\pi(\mu)$ of latent means. The issue was studied in Newton et al. (2004), and there a nonparametric mean component was proposed. A nonparametric version of the EM algorithm enabled model fit. Comparisons indicated improvements in terms of error rates

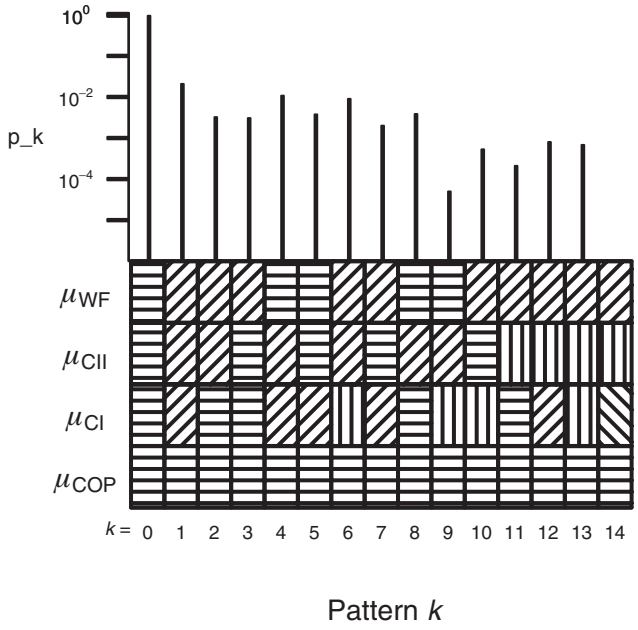

Fig. 2.1. Estimated mixing proportions for 15 patterns of mean expression among four mRNA populations in a rat breast cancer study: Equality of angles in a column stands for equality of the means. Above, height of lines indicates estimated proportion of genes in each pattern.

on short lists compared to the parametric model, gene-specific *t*-testing, and the method of Efron et al. (2001). That paper also showed how to formulate the mixture model in terms of directional alternatives, which can further improve flexibility, but it left unaddressed an extension beyond two-group comparisons to multigroup comparisons.

It may be that improvements obtainable by nonparametric analysis of the mean component are modest compared to improvements that would be possible through a more effective modeling of gene-specific variances. Advances in this direction by Lonnstedt and Speed (2002) and Smyth (2004) are significant, though their empirical Bayesian formulation is rather different than the one described here underlying `EBarrays`. In that work, expression shifts have to do with nonzero contributions in a linear model for expected expression, rather than separately realized mean values. The relative merits of the two forms of mean modeling remain to be worked out (e.g., the role played by discrete mixing proportions is present but less prominent in the linear-model formulation). With regard to variances, Lonnstedt and Speed (2002) and Smyth (2004) put a prior on gene-specific variances, and this provides some flexibility beyond

the constant variance assumption in the LNN version of `EBarrays`. The idea was also used by Baldi and Long (2001) and also by Ibrahim et al. (2002). Kendziorski and Wang (2005) investigate flexible variance modeling in the context of `EBarrays`.

Among-gene dependence is an ever-present concern, though it is difficult to handle owing to the dimensions involved. Permutation-based methods are helpful in guarding against ill-effects of dependence, but they are not always effective. Note that FDR controlling procedures are popular in part because they are fairly robust to among-gene dependencies compared to other multiplicity-adjustment methods. Dahl (2004), generalizing Medvedovic and Sivaganesan (2002), investigates methodology that directly models dependence among genes using a Dirichlet process mixture (DPM) formulation. In the Bayesian effects model for microarrays (BEMMA), different genes share parameters in much the same way that different mRNA populations share mean parameters on a given gene in `EBarrays`. Thus, correlation among genes is explained in terms of shared, latent parameter values. The grouping of genes into clusters where sharing occurs is mediated by the discrete clustering distribution inherent in the DPM model, and is assessed by posterior sampling via Markov chain Monte Carlo. Dahl (2004) shows improvements in the assessment of differential expression when one accommodates coordinated expression by this BEMMA approach. We note that BEMMA uses DPMs in a different way than Do, Müller, and Tang (2005), which used them to improve nonparametric inference based on one-dimensional reductions of the gene-level data. Using a novel mixture formulation, Yuan and Kendziorski (in press) offer another approach for using between-gene dependencies to improve differential expression analysis.

In summary, we see rapid development of methodology for altered gene expression based on the flexible class of hierarchical mixture models reviewed here. As new data analysis and data integration problems emerge in genomics, there will be further demand for such modeling in order to organize variation and to provide effective analysis of data.

Bibliography

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 1–11.

Dahl, D. Conjugate Dirichlet process mixture models: Gene expression, clustering, and efficient sampling, PhD thesis, University of Wisconsin, 2004.

Do, K., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society C*, **54**, 627–644.

Dudoit, S., Yang, Y.H., Speed, T.P., and Callow, M.J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a micro array experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

Fay, M.J., Longo, K.A., Karathanasis, G.A., Shope, D.M., Mandernach, C.J., Leong, J.R., Hicks, A., Pherson, K., and Husain, A. (2003). Analysis of *CUL-5* expression in breast epithelial cells, breast cancer cell lines, normal tissues and tumor tissues. *Molecular Cancer*, **2**, 40.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B*, **64**, 499–518.

Gottardo, R., Raftery, A.E., Yeung, K.Y., Bumgarner, R. (2006). Bayesian robust inference for differential gene expressionin cDNA microarrays with multiple samples. *Biometrics*, **62**, pp. xx–xx.

Ibrahim, J.G., Chen, M.-H., and Gray, R.J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, **97**, 88–99.

Kendziorski, C.M., Newton, M.A., Lan H., and Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22**, 3899–3914.

Kendziorski, C.M. and Wang, P. (2005). Flexible variance modeling with `EBarrays`. Technical Report # 192, UW Department of Biostatistics and Medical Informatics.

Lonnstedt, I. and Speed, T.P. (2002). Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2005). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*, **99**, 990–1001.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8**, 37–52.

Newton, M.A., Noueiry, D., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Shepel, L.A., Lan, H., Haag, J.D., Brasic, G.M., Gheen, M.E., Simon, J.S., Hoff, P., Newton, M.A., and Gould, M.N. (1998). Genetic identification of multiple loci that control breast cancer susceptibility in the rat. *Genetics*, **149**, 289–299.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiements. *Statistical Applications in Genetics and Molecular Biology*, **3**(1) 3.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Storey, J.D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035.

Wang, P. and Newton, M.A. (2005). Robustness of EBarrays to one form of dependence. Technical Report #1114, UW Department of Statistics.

Yuan, M. and Kendziorski C. (in press). A unified approach for simultaneous gene clustering and differential expression. *Biometrics*.