Published in final edited form as:

J Am Stat Assoc. 2009 December ; 104(488): 1318–1323. doi:10.1198/jasa.2009.ap09575.

Rejoinder

Robert B. Scharpf[Postdoctoral Fellow],

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205

Håkon Tjelmeland[Professor],

Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

Giovanni Parmigiani[Professor], and

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health and Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205

Andrew B. Nobel[Professor]

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599

Robert B. Scharpf: rscharpf@jhsph.edu; Håkon Tjelmeland: haakont@stat.ntnu.no; Giovanni Parmigiani: gp@jimmy.harvard.edu; Andrew B. Nobel: nobel@email.unc.edu

We thank the editor for facilitating a discussion that addresses the challenges of integrating gene expression data from multiple studies, and the discussants for their insightful comments. Several interesting topics regarding the estimation of cross-platform differential gene expression in our Bayesian model have been raised, including the problem of discordance in multiple studies, the assessment of goodness of fit, other functionals of the posterior, and recent advances in Metropolis–Hastings proposals that may improve mixing. Our rejoinder is organized as follows. We begin with a general discussion of discordant differential gene expression in multiple studies, followed by more detailed responses to the discussants' main points. Our remarks include recent progress with the theoretical model and improvements to software implementation in the R package *XDE* available from Bioconductor (<http://www.bioconductor.org>). We close with a few remarks regarding the problem of simultaneously estimating the proportion of differentially expressed genes and the magnitude of their differential expression in cross-study analyses of gene expression data.

DISCORDANCE

In the context of our Bayesian model, we have defined discordance of differential gene expression in multiple studies as (a) a gene that is differentially expressed in at least one pair of studies, with (b) the difference across the binary phenotype in one or more of the differentially expressed study pairs of opposite sign. An indicator for discordant differential expression takes the value 1 when both of these conditions hold and 0 otherwise. Recall that two implementations of the Bayesian model are available. In prior model A, we allow the gene-specific indicators (δ) for differential expression to be study-specific, $\delta_{g1}, \dots, \delta_{gP}$, for studies $p \in \{1, \dots, P\}$. In prior model B, we set the restriction that $\delta_{g1} = \dots = \delta_{gP} = \delta_g$ (see Scharpf et al. 2009 for details). To illustrate how the posterior mean of the discordant indicator measures the strength of discordance, consider an analysis involving only two studies and the posteriors that are available from prior model A. Using the foregoing criteria, for each Markov chain Monte Carlo (MCMC) sample, a gene is discordant if (a) the indicator for differential expression is 1 in both studies *and* (ii) the signs of the estimated differences in expression with respect to the binary outcome are (+,−) or (−,+) in the pair of studies. Note that (i) implies that a gene is *not* considered discordant if the indicator for

differential expression is (0, 1) or (1, 0); therefore, the posterior mean for the indicator of discordant differential expression is at least as small as the minimum of the study-specific posterior means of differential expression indicators. Typically, small posterior means of the differential expression indicator occur for genes with small differences in the average expression with respect to the outcome variable. Thus the indicator for discordant differential expression directly weighs the evidence of differential expression in each study and indirectly weighs the magnitude of the estimated offsets. The posterior mean of the discordant indicator provides a continuous, gene-specific measure of the strength of discordance. In the setting of three or more studies, values of the posterior mean near 1 can be interpreted as strong discordance in at least one pair of studies.

The discussants provide complementary perspectives on discordant differential gene expression in multiple studies. Ghosh and Choi question whether the mechanisms for discordance proposed in Scharpf et al. (2009) could explain discrepant findings in studies using the same technology (and the same set of probes) for measuring gene expression. Having observed discordance in their research with yeast cell cycle experiments, Fan and Liu discuss approaches for identifying concordant changes in gene expression in the presence of interstudy discordance and within the general framework of the two implementations of the Bayesian model via an illustrative simulation.

In response to Ghosh and Choi, we argue that discordance attributable to genetic heterogeneity can occur whether one platform or several platforms are used to measure gene expression. Scharpf et al. (2009) provided a simplifying illustration in which different strains of mice respond differently to a given treatment. For an example involving human populations, consider that adjusting for population substructure is an important step in genome-wide association studies, because admixed populations differ by minor allele frequencies at many polymorphic loci (i.e., single-nucleotide polymorphisms) and have been shown to have differing susceptibility to common diseases (Cheng et al. 2009). While studies assessing the level of contribution of DNA-level variation to gene expression are ongoing, gene expression as a trait is linked more closely to sequence-level variation than to complex traits, such as obesity and diabetes (Schadt et al. 2008). If the genetic context differs across studies, then discordance in measures of gene expression may be seen, regardless of whether or not the same instrument was used to measure gene expression. Similarly, heterogeneity of epigenetic mechanisms that regulate gene expression and environmental differences between study sites also can contribute to interstudy discordance. In contrast to a genetic or epigenetic mechanism of discordance, Fan and Liu have observed discordance in differential gene expression in yeast experiments due to temporal differences in expression that arise from the experimental procedure used to synchronize yeast cell cycles. Such disparate sources of discordance are often latent, are potentially correlated with the outcome of interest, and can be indistinguishable in their effect on the data.

Through simulation, Fan and Liu evaluate criteria by which a model that assumes concordance performs better in terms of area under the receiver operator characteristic curve (AUC) for detecting differential expression compared with a model that allows discordance. They assess the two approaches over varying levels of simulated discordance. Intuitively, their simulation results suggest that when the number of discordant pairs in two studies is small, a model that assumes concordance has a higher AUC than a model that allows discordant differential expression. Their results are consistent with the more elaborate simulation setting discussed in Section 6.1 of Scharpf et al. (2009), in which background noise was obtained from an experimental data set (see Figure 4). In row 1 of Figure 4, no discordant pairs were simulated, and prior model B performed equivalently or outperformed prior model A in terms of detecting concordant (panel 1) or discordant (panel 2) differential expression. In row 2, the discordant pairs were allowed by simulating studyspecific offsets

($\delta_{gp}\Delta_{gp}$). Here, compared with prior model B, the AUC in prior model A was higher for concordant differential expression (panel 1) and approximately equivalent for discordant differential expression (panel 2). The results of Fan and Liu suggest that prior model A would become more effective as the parameter for the correlation of the offsets approached 0 in our simulation.

In summary, genetic, environmental, and experimental sources of heterogeneity across studies can contribute to differences in differential gene expression irrespective of the technology used to measure expression. In practice, we assess the overall reproducibility across studies using simple exploratory plots to determine the usefulness of any integrative method for cross-study differential gene expression. When discordance does not emerge as a global pattern between two studies, the simulation experiments of Scharpf et al. (2009) and Fan and Liu suggest data characteristics that can be useful in choosing among the two implementations of the Bayesian model. Specifically, the single-indicator model for differential expression is more useful for measuring concordant and discordant differential expression if t -statistics across studies are highly correlated, while the multiple-indicator implementation of the Bayesian model is more likely to be effective when a substantial proportion of the gene pairs are discordant. The posterior mean of the discordant indicator may be helpful for discriminating between *weak* and *strong* discordance, and thus potentially useful in prioritizing subsequent experiments to determine whether there is a biological basis for the observed discordance.

OTHER FUNCTIONALS OF THE POSTERIOR

Ghosh and Choi note that functionals of the posterior distribution besides concordant or discordant differential expression may be of interest. In particular, they discuss a pattern observed by Tomlins et al. (2005) in the joint distribution of two genes involved in a gene fusion event (see Ghosh and Choi's Figure 1). Tomlins et al. (2005) developed a nonparametric statistic for cancer outlier profile analysis (COPA) to capture a gene fusion event that appears as an outlier in the joint distribution of two genes for a subset of the samples. Ghosh and Choi propose an extension based on the Mahalanobis distance that can be estimated from the posteriors in our Bayesian model. Two considerations in the context of multiple studies are relevant to this idea.

First, the joint analysis of differential gene expression in multiple studies appears to be most useful when the sample size in any individual study is small (Scharpf et al. 2009). In such instances, a model that borrows strength across studies outperforms single-study estimates of differential expression, as assessed through simulations using receiver operator characteristic curves. Observing outliers, such as the gene fusion events in Tomlins et al. (2005), requires a large number of subjects, however. The benefit of combining estimates of differential expression across studies is less compelling when the sample sizes in the individual studies are large. More importantly, it is quite plausible that a random sample of subjects in subsequent studies may differ in terms of the proportion of subjects that have undergone the gene fusion event. When considering approaches for combining the estimated distances across studies, it may be useful to avoid any averaging of the distances that potentially could dampen the signal.

Second, computing a statistic such as the Mahalanobis distance for all possible gene pairs in high-dimensional data sets is not computationally practical, as pointed out by Ghosh and Choi (Section 3). Prioritizing genes based on the posterior mean for the indicator of concordant differential expression assumes that the genes have interesting marginal distributions. Simple statistics for prioritizing gene groups based on their corresponding multivariate distributions are needed. For instance, Dettling, Gabrielson, and Parmigiani

(2005) discussed correlation-based approaches to scoring all pairwise combinations of genes simultaneously to identify specific patterns in the joint distribution. A combination of simple correlation-based approaches to select a subset of genes that satisfy a particular pattern, followed by a model-based statistic from the posterior to prioritize this list, may be the most computationally practical. For example, Figure 1 displays a plot of two genes in the breast cancer data set in which the ER phenotype (plotting symbol “x”) tends to be more prevalent when both genes are turned on. This gene was selected based on the absolute difference in the Spearman correlations for the ER-positive and ER-negative samples in each study (Dettling, Gabrielson, and Parmigiani 2005).

Many other posteriors could be computed from the Bayesian model. Straightforward extensions include posterior means for specific profiles of differential expression, for instance, up-regulated in study one, down-regulated in study two, up-regulated in study three, and so on. Given that it has become common to look for groups of genes that are differentially expressed because of their role in, for instance, the same biological pathway, posterior means could be developed for groups of genes that are differentially expressed. For example, it would be straightforward to evaluate gene sets based on a posterior probability that at least a given fraction of the genes are concordantly differentially expressed.

PRIORITIZING GENES

The implementation of the Bayesian model in R allows the flexible computation of posterior means for differential expression at different levels of stringency for concordance. In particular, the analyst can select the number of studies in which the gene is differentially expressed, m , as well as a cutoff for the posterior mean, k . As Ghosh and Choi note, the price to pay for this flexibility is a more complicated scheme for prioritizing genes. The relationships among m , k , and estimates of the effect size in the individual studies can be easily visualized and plotted to help guide this decision, however. To illustrate, we have calculated posterior means for concordant differential expression in the four breast cancer studies using the multiple indicator implementation of the Bayesian model after a burn-in of 20,000 Markov chain Monte Carlo (MCMC) samples.

Figure 2 plots posterior means for the indicator of concordant differential expression for the four breast cancer studies at different levels of m . The least stringent definition of concordance requires differential expression in one or more studies. For any given MCMC sample, genes satisfying more stringent criteria, $m = 3$ and $m = 4$, are a subset of the $m = 1$ genes. The solid horizontal and vertical lines depict a fixed threshold of the posterior mean, k . An alternative approach is to prioritize genes by an estimate of the Bayesian effect size; however, note that for a fixed k , selecting larger values of m tends to correspond to selecting genes with a larger Bayesian effect size. Compare, for instance, panels (a) and (b) in Figure 3, in which the highlighted genes were obtained using $m = 1$ (a) and $m = 4$ (b).

In summary, the posterior mean of concordant differential expression captures information regarding the probability of concordant differential expression in multiple studies and, indirectly, the magnitude of the offset. A practical recommendation is simply to rank genes by the posterior mean of concordant differential expression for an arbitrary m . Genes at the top of this list will tend to be the same regardless of m and have large corresponding estimates of the Bayesian effect size that are consistent across studies, a desirable property that is easy to verify. How far down this list to go depends on the estimated proportion of genes differentially expressed in each study. In Concluding Remarks we discuss the challenge of estimating the proportion of differentially expressed genes in multiple studies in more detail.

GOODNESS OF FIT

We agree with Ghosh and Choi that global assessments of goodness of fit are useful, particularly for complex model such as ours with a large number of hyperparameters. For gene-level assessment, Ghosh and Choi follow the approach of Johnson (2004). In brief, expression values are standardized by parameters drawn from the posterior distribution for the mean and variance in equation (1) of Scharpf et al. (2009). The standardized expression values for each gene are binned into equally probable cells based on quantiles of the standard normal distribution. Posterior means of the cell counts are obtained after a large number of MCMC samples. Plotting the chi-squared statistics computed from the posterior mean cell counts provides a global assessment of goodness of fit. Because this is an important component of the data analysis pipeline when assessing the Bayesian model, we have implemented this approach in the XDE software with additional documentation in the accompanying vignette (www.bioconductor.org). In the breast cancer studies, we observed poor goodness of fit for a small percentage of genes in the three largest studies (49+ samples). Further exploration revealed that a small proportion of the samples for each of these genes had more extreme values than would be predicted from a Gaussian distribution. Poor goodness of fit also can arise when estimates of the gene-specific σ s are pulled toward larger values than they would be if estimated independently (Tusher, Tibshirani, and Chu 2001; Lönnstedt and Speed 2002).

POSTERIOR SIMULATION

Problematic in complex Bayesian models are parameters that are highly correlated in their joint posterior distribution. To address this problem, Fan and Liu discuss their recently developed “piloted sequential proposals” method, which proposes new values sequentially for multiple parameters in a group move. We agree that such proposals may be useful in the context of our Bayesian model, and have spent considerable effort developing strategies to improve mixing through block updates of several parameters. In particular, our Bayesian model uses block updates for r and c^2 , ρ and γ^2 , θ and λ , and t and l (Scharpf et al. 2009). In analyses involving more than three studies, the mixing of the correlation matrix for the offsets, Δ , tends to be particularly slow, even with block updates. Recent versions of the software accept or reject proposed values for the elements of the correlation matrix one at a time.

As discussed previously, we expect that discordance in estimates of differential expression for a subset genes will be common. Discordance may be seen even if the overall mean expression values are largely correlated. As described by Scharpf et al. (2009), the priors for the overall means and the estimated offsets share a parameter, τ , in common. To limit the potential influence of the covariance for the gene-specific means on the covariance of the offsets, recent versions of the software have uncoupled the priors for the offsets by specifying separate τ 's for the corresponding covariance matrixes. Together with the block updates outlined earlier, we expect that uncoupling the τ 's in the covariance matrixes of the means and the offsets and updating the elements of the correlation matrixes one at a time will improve the mixing of these parameters.

CONCLUDING REMARKS

Our ability to understand complex traits by characterizing absolute or relative RNA transcript levels through high-throughput technologies depends on how well we can identify the genes that are expressed differentially. Because a large number of differentially expressed genes can be identified in a microarray experiment, and the measurements in such platforms are typically noisy, microarrays often are regarded as a screen for identifying candidate genes and pathways for further study. In principle, one would like to find both the

number of differentially expressed genes and an approach for prioritizing the resulting gene list. In the setting of cross-study differential expression, we suggest an approach for prioritization that identifies genes with large differences in transcript abundance that are concordant across studies.

We conclude by discussing some of the statistical challenges in simultaneously estimating both the proportion of differentially expressed genes and the magnitude of the differences. While framed in the context of the analysis of multiple studies using our Bayesian model, many of the issues are fundamental to the analysis of differential expression in single or multiple studies.

Our implementation of a Bayesian model parameterizes differential expression as a product of an indicator for differential expression and a scalar quantifying the difference in expression across a binary covariate. The shrunken differences in expression with respect to the outcome variable are referred to as the *offsets*, and the mean of the gene-specific indicators of differential expression is given by the parameter ξ . These parameters have hyperparameters that are shared by genes and studies. Theoretically, the posterior mean of ξ is the proportion of differentially expressed genes. In practice, values for the posterior mean of $\xi > 0.5$ are common and, when observed, deserve further scrutiny.

Interpretation of ξ

As is the case for most complex multilevel models, here there is an increased burden of proof that the modeling assumptions are tenable. In our formulation, true differences in the average transcript levels with respect to the outcome are modeled as a mixture of a point mass at 0 (no differential expression) and a multivariate normal (differential expression). If in fact the distribution of the offsets is multivariate normal and the offsets are large relative to the noise of the platforms, then a model such as ours can help discriminate between the two components of the mixture, as we learned from extensive simulation exercises. But we are concerned about the extent to which departures from the multivariate normal assumption may compromise the performance and, in particular, lead to a small estimated variance of the offsets, a large estimated ξ , and overshrinkage of the offsets—symptoms that we sometimes encounter in practice. The goodness-of-fit statistics proposed by Johnson (2004) in conjunction with scatterplots of the empirical differences versus the offsets from the posterior may be helpful for evaluating the appropriateness of the multivariate normal distribution. Transformations such as the logarithm may help in achieving nearly Gaussian empirical differences. Nevertheless, it is plausible that such transformations may be more effective for obtaining differences in mean expression levels that are multivariate normal for the set of non-differentially expressed genes, while the (unknown) set of differentially expressed genes remains non-Gaussian. Alternative priors, such as a multivariate t , may help reduce the problem of overshrinking the offsets and afford some protection against ξ -inflation. These issues notwithstanding, the overall ranking of genes appears to be unaffected by large ξ (Scharpf et al. 2009).

Apart from modeling assumptions, experimental and biological sources of heterogeneity also can contribute to ξ -inflation. In particular, any source of heterogeneity that is correlated with the outcome variable can affect the classification of differential expression. Whether observed in multiple studies and concordant or in a subset of studies and discordant, such heterogeneity causes inflated estimates of ξ . Consider, for instance, a batch effect that arises as a result of changes to laboratory reagents over time. If more cases than controls were processed toward the end of the experiment, then estimates of differential expression might reflect differences in the experimental conditions, as opposed to biological differences in cases and controls. In the worst-case scenario, cases and controls are processed as separate batches, and inference regarding differential expression is completely confounded by batch.

Whether induced experimentally or through factors relating to genetic, epigenetic, or environmental sources, such heterogeneity can give rise to systematic differences in the expression measures, are indistinguishable in their effects on the data, and often are not known a priori. Approaches that estimate and adjust for latent covariates in the individual studies, through, for instance, tools such as surrogate variable analysis (Leek and Storey 2008), may be helpful and can be applied to the individual studies before the joint analysis.

We interpret ξ as the sum total of all genes that are affected by the outcome variable, as well as biological, environmental, and technological sources of heterogeneity that are correlated with the outcome. We have suggested approaches for evaluating model assumptions that may contribute to ξ inflation, and recommend exploring latent sources of heterogeneity in the individual studies that could contribute to systematic differences in the expression levels. Simultaneously quantifying the number of differentially expressed genes and the magnitude of differential expression in multiple studies remains a difficult problem, which complex multilevel models are poised to address.

Acknowledgments

Scharpf's work was supported by U.S. National Institute of Environmental Health Sciences training grant 5T32ES012871, National Heart, Lung, and Blood Institute training grant 5T32HL007024, and National Science Foundation grant DMS 034211. Nobel's research was supported in part by National Science Foundation grant DMS 0406361 and U.S. Environmental Protection Agency grant RD-83272001.

ADDITIONAL REFERENCES

- Cheng C-Y, Kao WHL, Patterson N, Tandon A, Haiman CA, Harris TB, Xing C, John EM, Ambrosone CB, Brancati FL, Coresh J, Press MF, Parekh RS, Klag MJ, Meoni LA, Hsueh W-C, Fejerman L, Pawlikowska L, Freedman ML, Jandorf LH, Bandera EV, Ciupak GL, Nalls MA, Akylbekova EL, Orwoll ES, Leak TS, Miljkovic I, Li R, Ursin G, Bernstein L, Ardlie K, Taylor HA, Boerwinckle E, Zmuda JM, Henderson BE, Wilson JG, Reich D. Admixture Mapping of 15,280 African Americans Identifies Obesity Susceptibility Loci on Chromosomes 5 and X. *PLoS Genetics*. 2009; 5(5) e1000490. Available at <http://dx.doi.org/10.1371/journal.pgen.1000490>.
- Detting M, Gabrielson E, Parmigiani G. Searching for Differentially Expressed Gene Combinations. *Genome Biology*. 2005; 6(10):R88. Available at <http://dx.doi.org/10.1186/gb-2005-6-10-r88>. [PubMed: 16207359]
- Johnson VE. A Bayesian Chi² Test for Goodness-of-Fit. *The Annals of Statistics*. 2004; 32(6):2361–2384.
- Leek JT, Storey JD. A General Framework for Multiple Testing Dependence. *Proceedings of the National Academy of Sciences of the USA*. 2008; 105(48):18718–18723. Available at <http://dx.doi.org/10.1073/pnas.0808709105>. [PubMed: 19033188]
- Lönnstedt I, Speed T. Replicated Microarray Data. *Statistica Sinica*. 2002; 12:31–46.
- Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. Mapping the Genetic Architecture of Gene Expression in Human Liver. *PLoS Biology*. 2008; 6(5):e107. Available at <http://dx.doi.org/10.1371/journal.pbio.0060107>. [PubMed: 18462017]
- Scharpf RB, Tjelmeland H, Parmigiani G, Nobel A. A Bayesian Model for Cross-Study Differential Gene Expression. *Journal of the American Statistical Association*. 2009; 104:1295–1310. [PubMed: 21127725]

Figure 1.

Scatterplots of expression measurements for two genes in the four breast cancer studies that show the “on/off” characteristic described by Dettling, Gabrielson, and Parmigiani (2005). ER-positive samples are denoted by the plotting symbol “x.” The ER-positive samples tend to be more prevalent when both genes are turned on in the Farmer and Huang studies. In contrast, the ER-negative phenotype of Farmer and Huang depends on the which of the two genes is overexpressed.

Figure 2.

(a) Scatterplot of posterior mean for concordant differential expression for $m = 1$ (horizontal axis) versus $m = 3$ (vertical axis). The points to the right of the vertical lines represent $a = 0.95$ versus $m = 3$. There is but a small difference in selecting 1, 2, or 3 studies as concordantly differentially expressed; the Farmer, Hedenfalk, and Sorlie studies show great similarity. The genes have a much lower posterior probability of concordant expression in all four studies, because the genes differentially expressed in the Huang study are largely independent of the remaining studies.

Figure 3.

Scatterplots of the estimates of Bayesian effect size for the four breast cancer studies. Darker plotting symbols highlight genes with a posterior mean of concordant differential expression > 0.95 . Panels (a) and (b) differ in terms of the stringency of the definition of concordance. In (a), we require that the gene be differentially expressed in at least one study ($m = 1$) and that the direction of differential expression be the same in all studies for which the gene is differentially expressed. In contrast, in (b), we require that the gene be differentially expressed and the direction be the same in all studies ($m = 4$). Increasing the stringency of the definition of concordance at a fixed threshold for the posterior mean tends to select genes with larger estimates of the Bayesian effect size.