Biostatistics (2007), **8**, 1, *pp.* 118–127
doi:10.1093/biostatistics/kxj037
Advance Access publication on April 21, 2006

Adjusting batch effects in microarray expression data using empirical Bayes methods

W. EVAN JOHNSON, CHENG LI*

Department of Biostatistics and Computational Biology,
Dana-Farber Cancer Institute, Boston, MA, USA and Department of Biostatistics,
Harvard School of Public Health, Boston, MA, USA
cli@hsph.harvard.edu

ARIEL RABINOVIC

Department of Genetics and Complex Diseases, Harvard School of Public Health, Boston, MA, USA

Summary

Non-biological experimental variation or "batch effects" are commonly observed across multiple batches of microarray experiments, often rendering the task of combining data from these batches difficult. The ability to combine microarray data sets is advantageous to researchers to increase statistical power to detect biological phenomena from studies where logistical considerations restrict sample size or in studies that require the sequential hybridization of arrays. In general, it is inappropriate to combine data sets without adjusting for batch effects. Methods have been proposed to filter batch effects from data, but these are often complicated and require large batch sizes (>25) to implement. Because the majority of microarray studies are conducted using much smaller sample sizes, existing methods are not sufficient. We propose parametric and non-parametric empirical Bayes frameworks for adjusting data for batch effects that is robust to outliers in small sample sizes and performs comparable to existing methods for large samples. We illustrate our methods using two example data sets and show that our methods are justifiable, easy to apply, and useful in practice. Software for our method is freely available at: http://biosun1.harvard.edu/complab/batch/.

Keywords: Batch effects; Empirical Bayes; Microarrays; Monte Carlo.

1. Introduction

With the many applications of gene expression microarrays, biologists are able to efficiently extract hypotheses that can later be tested experimentally in a lab setting. For example, a microarray experiment might compare the gene expression profile of diseased or treated tissue (treatment) with the profile of normal tissue (controls) to determine which genes are associated with the disease or the presence of the treatment, providing better understanding of disease/gene relationships. However, practical considerations limit the number of samples that can be amplified and hybridized at one time, and replicate samples may

*To whom correspondence should be addressed.

be generated several days or months apart, introducing systematic "batch effects" or non-biological differences that make samples in different batches not directly comparable. Batch effects have been observed from the earliest microarray experiments (Lander, 1999), and can be caused by many factors including the batch of amplification reagent used, the time of day when an assay is done, or even the atmospheric ozone level (Fare *and others*, 2003). Batch effects are also inevitable when new samples or replicates are incrementally added to an existing array data set or in a meta-analysis of multiple studies that pools microarray data across different labs, array types, or platforms (Rhodes *and others*, 2004). Some researchers have presented methods for adjusting for batch effects (Benito *and others*, 2004; Alter *and others*, 2000), but these methods require many samples (>25) in each batch for best performance and may remove real biological variation from the data. In this paper, we develop an empirical Bayes (EB) method that is robust for adjusting for batch effects in data whose batch sizes are small.

1.1 *Microarray data with batch effects*

Data set 1 resulted from an oligonucleotide microarray (Affymetrix HG-U133A) experiment on human lung fibroblast cells (IMR90) designed to reveal whether exposing mammalian cells to nitric oxide (NO) stabilizes mRNAs. Control samples and samples exposed to NO for 1 h were then transcription inhibited for 7.5 h. Microarray data were collected at baseline (0 h, just before transcription inhibition) and at the end of the experiment (after 7.5 h) for both the control and the NO-treated group. It was hypothesized that NO will induce or inhibit the expression of some genes, but would also stabilize the mRNA of many genes, preventing them from being degraded after 7.5 h. One sample per treatment combination was hybridized, resulting in four arrays. This experiment was repeated at three different times or in three batches (totaling 12 samples). The batches in this data set were identical experiments using the same cell source, and were conducted by the same researcher in the same lab using the same equipment. Figure 1(a) contains a heat map of data set 1 using a standard hierarchical clustering algorithm produced using the dChip software (Li and Wong, 2003). This heat map exhibits characteristics commonly seen by researchers attempting to combine multiple batches of microarray data. All four samples in the second batch cluster together, indicating that the clustering algorithm recognized the batch-to-batch variation as the most significant source of variation within this data set. We give another example data set with batch effects, denoted throughout this paper as data set 2, in the online supplementary materials available at *Biostatistics* online.

1.2 *EB applications in microarrays*

EB methods have been applied to a large variety of settings in microarray data analysis (Chen *and others*, 1997; Efron *and others*, 2001; Newton *and others*, 2001; Tusher *and others*, 2001; Kendziorski *and others*, 2003; Smyth, 2004; Lönnstedt *and others*, 2005; Pan, 2005; Gottardo *and others*, 2006). EB methods are very appealing in microarray problems because of their ability to robustly handle high-dimensional data when sample sizes are small. EB methods are primarily designed to "borrow information" across genes and experimental conditions in hope that the borrowed information will lead to better estimates or more stable inferences. In the papers mentioned above, EB methods were usually designed to stabilize the expression ratios for genes with very high or very low ratios, stabilize gene variances by shrinking variances across all other genes, possibly protecting their inference from artifacts in the data. In this paper, we extend the EB methods to the problem of adjusting for batch effects in microarray data.

2. Existing methods for adjusting batch effect

2.1 *Microarray data normalization*

Microarray data are often subject to high variability due to noise and artifacts, often attributed to differences in chips, samples, labels, etc. In order to correct these biases caused by non-biological conditions,

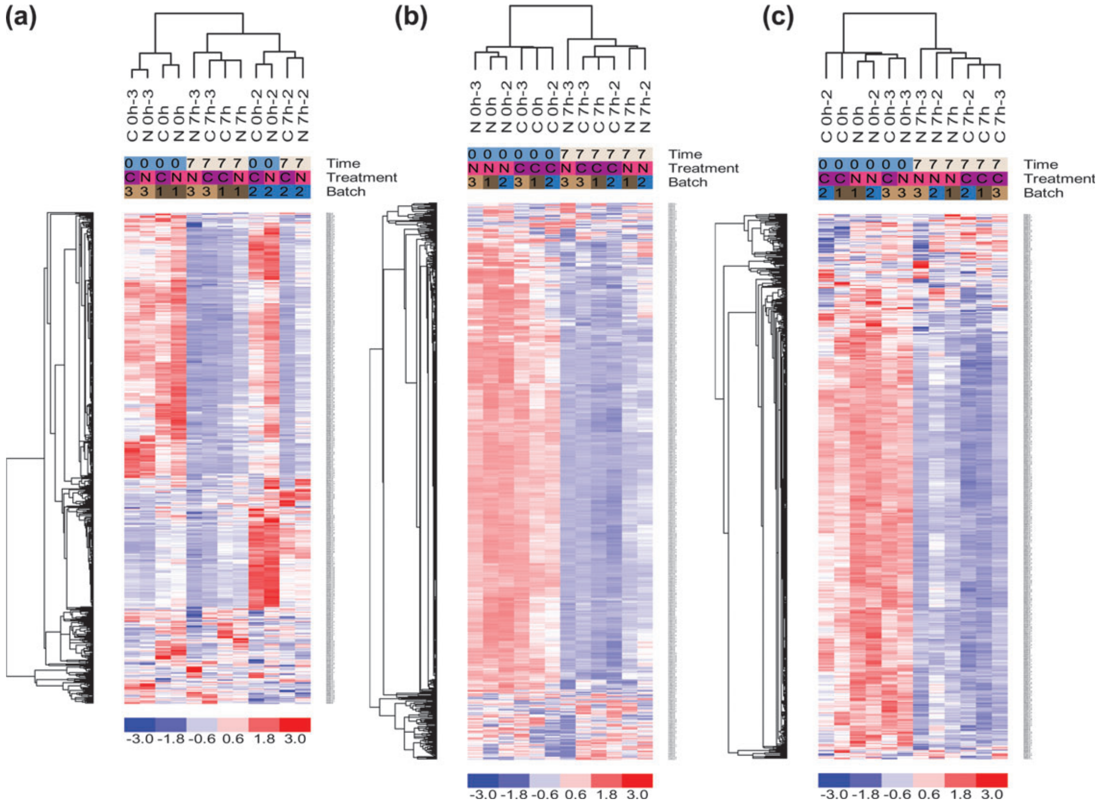

Fig. 1. Heat map clusterings for data set 1. The gene-wise expression values are used to compute gene and sample correlations and displayed in color scale, and the sample legends on the top are 0 (0 h), 7 (7.5 h), C (Control), and N (NO treated). (a) Expression for 628 genes with large variation across all the 12 samples. Note that the samples from the batch 2 cluster together and the baseline (time = 0) samples also cluster by batch 1 and 3; (b) 720 genes after applying "standardized separators" (which standardize each gene within each batch to have a mean 0 and variance of 1) for gene filtering and clustering in the dChip software; (c) 692 genes after applying the EB batch adjustments and then filtered for clustering. Note that there is no strong evidence of batch effects after adjustment in heat maps (b)–(c). The EB adjustment in (c) has the advantage of being robust to outliers in small sample sizes.

researchers have developed "normalization" methods to adjust data for these effects (Schadt *and others*, 2001; Tseng *and others*, 2001; Yang *and others*, 2002; Irizarry *and others*, 2003). However, normalization procedures do not adjust the data for batch effects, so when combining batches of data (particularly batches that contain large batch-to-batch variation), normalization is not sufficient for adjusting for batch effects and other procedures must be applied.

2.2 *Other batch effect adjustment methods*

A few methods for adjusting data for batch effects have been presented in the literature. Alter *and others* (2000) propose a method for adjusting data for batch effects based on a singular-value decomposition (SVD) by adjusting "the data by filtering out those eigengenes (and eigenarrays) that are inferred to represent noise or experimental artifacts." Nielsen *and others* (2002) successfully apply this SVD batch effect adjustment to a microarray meta-analysis. Benito *and others* (2004) use distance weighted discrimination (DWD) to correct for systematic biases across microarray batches by finding a separating hyperplane between the two batches, and adjusting the data by projecting the different batches on the DWD plane, finding the batch mean, and then subtracting out the DWD plane multiplied by this mean.

There are difficulties faced by researchers who try to implement the SVD and DWD batch adjustment methods. These methods are fairly complicated and usually require many samples (>25) per batch to implement. For the SVD adjustment, the eigenvectors in the SVD are all orthogonal to each other, so the method is highly dependent on proper selection of first several eigenvectors, which makes finding the batch effect vector not always clear if it even exists at all. In addition, the SVD approach factors out all variation in the given direction, which may not be completely due to batch effects. The DWD method can only be applied to two batches at a time. In one example, Benito *and others* (2004) use a stepwise approach, first adjusting the two most similar batches, and then comparing the third against the previous (adjusted) two. The stepwise method yields reasonable results in their three-batch case, but this could potentially break down in cases where there are many more batches or when batches are not very similar.

2.3 *Model-based location/scale adjustments*

Location and scale (L/S) adjustments can be defined as a wide family of adjustments in which one assumes a model for the location (mean) and/or scale (variance) of the data within batches and then adjusts the batches to meet assumed model specifications. Therefore, L/S batch adjustments assume that the batch effects can be modeled out by standardizing means and variances across batches. These adjustments can range from simple gene-wise mean and variance standardization to complex linear or non-linear adjustments across the genes.

One straightforward L/S batch adjustment is to mean center and standardize the variance of each batch for each gene independently. Such a method is currently implemented in the dChip software (Li and Wong, 2003), designated as “using standardized separators” (see Figure 1(b)). In more complex situations such as unbalanced designs or when incorporating numerical covariates, a more general L/S framework must be used. For example, let Y_{ijg} represent the expression value for gene g for sample j from batch i. Define an L/S model that assumes

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}, \tag{2.1}$$

where α_g is the overall gene expression, X is a design matrix for sample conditions, and β_g is the vector of regression coefficients corresponding to X. The error terms, ε_{ijg}, can be assumed to follow a Normal distribution with expected value of zero and variance σ_g^2. The γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for gene g, respectively. The batch-adjusted data, Y_{ijg}^*, are given by

$$Y_{ijg}^* = \frac{Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g - \widehat{\gamma}_{ig}}{\widehat{\delta}_{ig}} + \widehat{\alpha}_g + X\widehat{\beta}_g, \tag{2.2}$$

where $\widehat{\alpha}_g$, $\widehat{\beta}_g$, $\widehat{\gamma}_{ig}$, and $\widehat{\delta}_{ig}$ are estimators for the parameters α_g, β_g, γ_{ig}, and δ_{ig} based on the model.

3. EB method for adjusting batch effect

The most important disadvantage of the SVD, DWD, and L/S methods is that large batch sizes are required for implementation because such methods are not robust to outliers in small sample sizes. In this section, we propose a method that robustly adjusts batches with small sample sizes. This method incorporates systematic batch biases common across genes in making adjustments, assuming that phenomena resulting in batch effects often affect many genes in similar ways (i.e. increased expression, higher variability, etc). Specifically, we estimate the L/S model parameters that represent the batch effects by “pooling information” across genes in each batch to “shrink” the batch effect parameter estimates toward the overall mean of the batch effect estimates (across genes). These EB estimates are then used to adjust the data for batch effects, providing more robust adjustments for the batch effect on each gene. The method is described in three steps below.

3.1 *Parametric shrinkage adjustment*

We assume that the data have been normalized and expression values have been estimated for all genes and samples. We also filter out the genes called as “absent” in more than 80% samples to eliminate noise. Suppose the data contain m batches containing n_i samples within batch i for $i = 1, \dots, m$, for gene $g = 1, \dots, G$. We assume the model specified in (2.1), namely,

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg},$$

and that the errors, ε, are normally distributed with mean zero and variance σ_g^2.

Step 1: Standardize the data
The magnitude of expression values could differ across genes due to mRNA expression level and probe sensitivity. In relation to (2.1), this implies that α_g, β_g, γ_g, and σ_g^2 to differ across genes, and if not accounted for, these differences will bias the EB estimates of the prior distribution of batch effect and reduce the amount of systematic batch information that can be borrowed across genes. To avoid this phenomenon, we first standardize the data gene wise so that genes have similar overall mean and variance. We estimate the model parameters α_g, β_g, γ_{ig} as $\widehat{\alpha}_g$, $\widehat{\beta}_g$, $\widehat{\gamma}_{ig}$ for $i = 1, \dots, m$ and $g = 1, \dots, G$. For our examples from data sets 1 and 2, we use a gene-wise ordinary least-squares approach to do this, constraining $\sum_i n_i \widehat{\gamma}_{ig} = 0$ (for all $g = 1, \dots, G$) to ensure the identifiability of the parameters. We then estimate $\widehat{\sigma}_g^2 = \frac{1}{N}\sum_{ij}(Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g - \widehat{\gamma}_{ig})^2$ (N is the total number of samples). The standardized data, Z_{ijg}, are now calculated by

$$Z_{ijg} = \frac{Y_{ijg} - \widehat{\alpha}_g - X\widehat{\beta}_g}{\widehat{\sigma}_g}.$$

Step 2: EB batch effect parameter estimates using parametric empirical priors
Compared to (2.1), we assume that the standardized data, Z_{ijg}, satisfy the distributional form, $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$. Note that the γ parameters here are not the same as in (2.1). Additionally, we assume the parametric forms for prior distributions on the batch effect parameters to be

$$\gamma_{ig} \sim N(Y_i, \tau_i^2) \quad \text{and} \quad \delta_{ig}^2 \sim \text{Inverse Gamma } (\lambda_i, \theta_i).$$

The hyperparameters γ_i, τ_i^2, λ_i, θ_i are estimated empirically from standardized data using the method of moments, and estimators are derived and given in the supplementary materials available at *Biostatistics* online.

These prior distributions (Normal, Inverse Gamma) were selected due to their conjugacy with the Normal assumption for the standardized data. For data set 2 in the supplementary material available at *Biostatistics* online, these priors did not fit well, so we developed the non-parametric prior method given in the supplementary materials available at *Biostatistics* online. For data set 1, these were moderately reasonable distributions for the priors (Figure 2), as the adjusted data did not differ much from

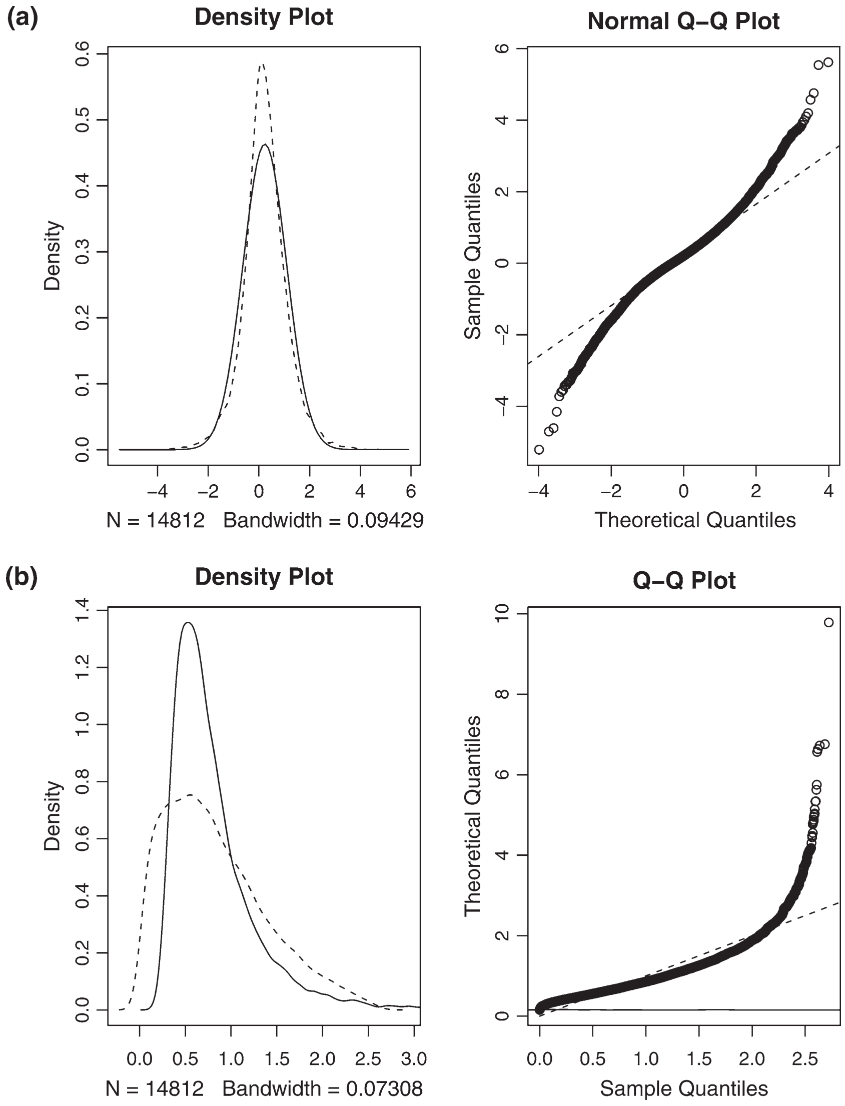

Fig. 2. Checking the prior distribution assumptions of the L/S model batch parameters. (a) The gene-wise estimates of additive batch parameter ($\widehat{\gamma}_{ig}$ of all genes) for data set 1, batch 1. (b) The gene-wise estimates of multiplicative batch parameter ($\widehat{\delta}^2_{ig}$ of all genes) for data set 1, batch 1. Each density plot contains a kernel density estimate of the empirical values (dotted line) and the EB-based prior distribution used in the analysis (solid line). Dotted lines on the quantile–quantile plots correspond to the EB-based Normal (a) or Inverse Gamma (b) distributions.

the data adjusted using a non-parametric prior. Based on the distributional assumptions above, the EB estimates for batch effect parameters, γ_{ig} and δ^2_{ig}, are given (respectively) by the conditional posterior means

$$\gamma^*_{ig} = \frac{n_i \overline{\tau}_i^2 \widehat{\gamma}_{ig} + \delta^{2*}_{ig} \overline{\gamma}_i}{n_i \overline{\tau}_i^2 + \delta^{2*}_{ig}} \quad \text{and} \quad \delta^{2*}_{ig} = \frac{\overline{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma^*_{ig})^2}{\frac{n_j}{2} + \overline{\lambda}_i - 1}. \tag{3.1}$$

Detailed derivations for these estimates for γ_{ig} and δ_{ig}^2 are given in the supplementary materials available at *Biostatistics* online.

Step 3: Adjust the data for batch effects
After calculating the adjusted batch effect estimators, γ_{ig}^* and δ_{ig}^{2*}, we now adjust the data. The EB batch-adjusted data γ_{ijg}^* can be calculated in a similar way as (2.2), but using EB estimated batch effects

$$\gamma_{ijg}^* = \frac{\widehat{\sigma}_g}{\widehat{\delta}_{ig}^*}(Z_{ijg} - \widehat{\gamma}_{ig}^*) + \widehat{\alpha}_g + X\widehat{\beta}_g.$$

4. RESULTS AND ROBUSTNESS OF THE EB METHOD

4.1 *Results for data set 1*

The parametric EB adjustment was applied to data set 1. Comparing Figure 1(a) to (c) provides evidence that the batch effects were adequately adjusted for in these data. Downstream analyses are now appropriate for the combined data without having to worry about batch effects. Figure 3 illustrates the amount of batch parameter shrinkage that occurred for the adjustments for 200 genes from one of the batches from data set 1. The adjustments viewed in this figure are on a standardized scale, so the magnitude of the actual adjustments also depends on the gene-specific expression characteristics (overall mean and pooled variance from standardization) and may vary significantly from gene to gene.

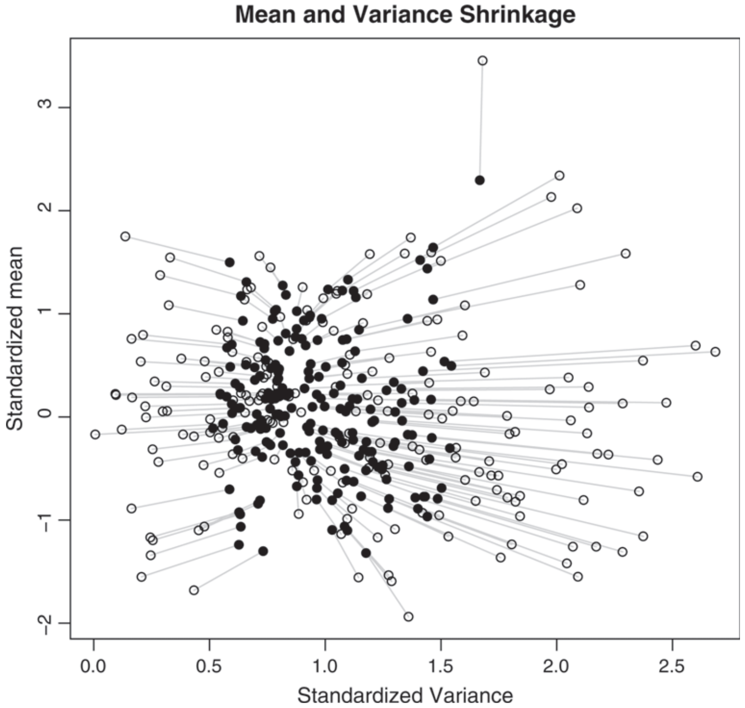

Fig. 3. Shrinkage plot for the first 200 probes from one of the batches in data set 1. The gene-wise and EB estimates of γ_{ig} and δ_{ig}^2 in Section 3.1 are plotted on the Y and X axis. Open circles are the gene-wise values and the solid are after applying the EB shrinkage adjustment.

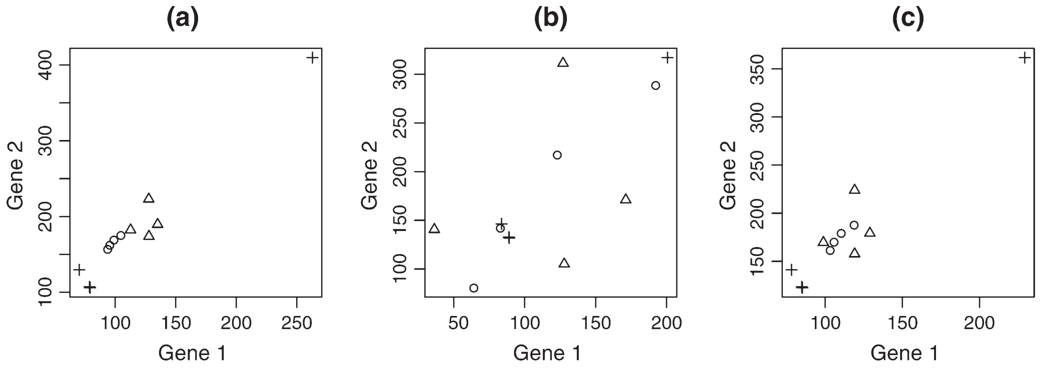

Fig. 4. Plots (a)–(c) illustrate the robustness of the EB adjustments compared to the L/S adjustments. The symbols signify batch membership. Data in (a) are unadjusted expression values for two genes from data set 1; (b) are the L/S-adjusted data. Note that the locations and scales of the batches in (b) have been over adjusted because of outliers in the unadjusted data, and that these outliers disappear in the L/S data; (c) are the EB-adjusted data. The outliers remain in these data and the batch with outliers is barely adjusted. The batches without outliers are adjusted correctly in the EB data.

4.2 *Robustness of the EB method*

The robustness of the EB method results from the shrinkage of the L/S batch estimates by pooling information across genes. If the observed expression values for a given gene are highly variable across samples in one batch, the EB batch adjustment is more like prior and less like the gene's gene-wise estimates, and becomes more robust to outlying observations. This phenomenon is illustrated in Figure 4.

4.3 *Software implementation*

The EB batch effect adjustment method described here is implemented in the R software package (http://www.r-project.org) and is freely available for download at http://biosun1.harvard.edu/complab/batch/. Detailed information on computing times required to adjust the example data sets are given in the supplementary materials available at *Biostatistics* online.

5. Discussion

Batch effects are a very common problem faced by researchers in the area of microarray studies, particularly when combining multiple batches of data from different experiments or if an experiment cannot be conducted all at once. We have reviewed and discussed the advantages and disadvantages of the existing batch effect adjustments. Notably, none of these methods are appropriate when batch sizes are small (<10), which is often the case. In order to account for this situation, we have presented a very flexible EB framework for adjusting for additive, multiplicative, and exponential (when data have been log transformed) batch effects.

We have shown that the EB adjustments allow for the combination of multiple data sets and are robust to small sample sizes. We illustrated the usefulness of our EB adjustments by combining two example data sets containing multiple batches with small batch size, and obtained consistent results from downstream analyses (clusterings and analysis as compared to similar single-batch data) while robustly dealing with outliers in these data. Additional discussion topics are included in the supplementary materials available at *Biostatistics* online.

Acknowledgments

We thank Wing Hung Wong, Donna Neuberg, and Yu Guo for helpful discussions, and Jennifer O'Neil for providing data set 2. This work is supported by grants from the National Institutes of Health (2R01 HG02341 and T32 CA009337) and the Claudia Adams Barr Program in Innovative Basic Cancer Research. Ariel Rabinovic was supported by grants RO1-CA82737 (Bruce Demple) and T32-CA09078. *Conflict of Interest:* None declared.

References

Alter, O., Brown, P. O. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 10101–6.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M. and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–14.

Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-based decisions and the quantitive analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–74.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–60.

Fare, T. L., Coffey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., LeProust, E., Marton, M. J., Meyer, M. R. *and others* (2003). Effects of atmospheric ozone on microarray data quality. *Analytical Chemistry* **75**, 4672–5.

Gottardo, R., Raftery, A. E., Yee Yeung, K. and Bumgarner, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62**, 10–8.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–64.

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**, 3899–914.

Lander, E. S. (1999). Array of hope. *Nature Genetics* **21**, 3–4.

Li, C. and Wong, W. H. (2003). DNA-Chip Analyzer (dChip). In: Parmigiani, G., Garrett, E. S., Irizarry, R. and Zeger, S. L. (editors), *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer. pp 120–41.

Lönnstedt, I., Rimini, R. and Nilsson, P. (2005). Empirical Bayes microarray ANOVA and grouping cell lines by equal expression levels. *Statistical Applications in Genetics and Molecular Biology* **4**.

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.

Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O'Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R. *and others* (2002). Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* **359**, 1301–7.

Pan, W. (2005). Incorporating biological information as a prior in an empirical Bayes approach to analyzing microarray data. *Statistical Applications in Genetics and Molecular Biology* **4**.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9309–14.

Schadt, E. E., Li, C., Ellis, B. and Wong, W. H. (2001). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry Supplement* **37**, 120–5.

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

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. and Wong, W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29**, 2549–57.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–21.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

[*Received February 28, 2006; revised April 14, 2006; accepted for publication April 14, 2006*]