Supplementary Materials for “Adjusting batch effects in microarray data using Empirical Bayes methods.”

This material contains additional information on this research including additional data examples, detailed on derivation of results, and additional discussion topics.

A.1: Description of Data Set 2

Data set 2 originates from an RNAi knockout experiment that studied the biological effects of inhibiting the TAL1 oncogene in human cells using an RNAi technique. This experiment was conducted in three separate batches. The first batch contained 6 RNAi samples and only 2 controls. After the first batch of the experiment was finished, the researchers determined that the sample sizes, particularly the number of controls, were not sufficient and subsequently added 3 RNAi samples and 4 controls. The time delay between the two experiments forced the researchers to hybridize the arrays at different times. Although the arrays were processed at the same facility, strong batch effects were exhibited between the two batches. Therefore the researchers repeated the experiment the third time in one batch with 9 RNAi and 6 control samples. Figure A. 1 contains a heatmap clustering of the samples in all the three batches. The right sample branch mainly consists of samples in batch 3, and the left sample branch mainly consists of samples in batch 1 and 2.

The researchers were able to derive some interesting results from the third batch of data set 2 (unpublished data), but the individual sample sizes in the first two batches were considered too small to be useful individually. Using data set 2, we consider whether it is possible to reproduce the results derived solely from the third batch by using only the first two batches of data. Also, we consider the advantage of combining all three batches of data to detect interesting genes.

A.2: Results for Data Set 2

The nonparametric EB (described below) adjustment was applied to data set 2, by combining the first two batches (call it EB2) and then combining all three batches (EB3). The third batch was used for comparison against the EB2 analysis results because it was an identical experiment to EB2 other than the fact that it was conducted in a single batch. Treatment status (RNAi or control) was included in the EB procedure for these adjustments. For EB3 the samples clustered well into small groups based on treatment status and 14 out of the 15 samples from the third experiment clustered next to their experimental counterpart from the first two studies (see Figure A. 2).

Differential expression was assessed using Welch's t-test to determine the differential expression of RNAi versus control samples. EB2 produced at list of 86 significant genes at a false discovery (q-value) threshold of 0.05 (Storey and Tibshirani, 2003). The third batch alone produced a list of 37 significant genes using the same threshold. Crossing the significant gene lists, we observed 13 genes common in both lists (Fisher's exact p-value < 1e-15). Furthermore, the dChip software was used to find significant Gene Ontology (GO) (http://www.geneontology.org) clusters. To find clusters, we first filtered out the genes that do not satisfy the criteria c<sd(gene)/mean(gene)<10, where c set is low enough to allow for about 1000 genes to remain after filtering. This filtering method selects genes with possible differential expression between all conditions, and enrichment of the gene's corresponding GO terms is defined as the overrepresentation of GO terms in the list as compared to the number of occurrences in the data before filtering in all the genes in the array (based on a p-value<.05). 53 GO clusters were significant in EB2 and 36 significant clusters were found in the third batch. These lists had 32 GO clusters in common.

Without any adjustment, combining these two batches produced a list of only 9 genes a q-value cutoff of 0.05 (only 4 of these appeared in the 37 batch 3 genes), which is substantially less than using the EB adjustment at the same cutoff. Using the L/S method, we adjusted the first two batches of data set 2 (call these data LS2) and found 113 significant genes at a q-value cutoff of 0.05, including 80 of the 86 genes found in EB2. We expected the LS2 data to produce a larger gene list due to the reduction in variance caused by data over-adjustment, so we believe the 86 EB12 gene list to be much more conservative. Looking at these lists closer, we found that 85 of the 86 EB12 genes were in the first 128 LS12 genes, and found 11 LS12 genes were not in the top 150 EB12 genes; of these 11 genes, only 1 appears in the list of 37 genes from batch 3 and the remaining 10 are ranked between 593 and 24,460 in the batch 3 list, so these are presumably false positives. These results indicate the following: 1) the batches need to be adjusted because not adjusting the data produces few to many genes to be useful. The short unadjusted list likely caused by increased variability introduced by batch effects. 2) Although the LS method produces the largest list of genes, the EB list appears to be more robust. Additionally, in this case the EB list is about 80% of the size of the LS list, so not much information is lost in the more conservative list. Also note that a slight improvement over the unadjusted can be made using standard ANOVA techniques (which is really just an additive only LS adjustment). This produces a slightly larger list than the unadjusted (but smaller than the EB), but this also suffers from a lack of robustness in small sample sizes.

Welch's t-test was also applied to EB3 to find differential expressed genes; yielding 1599 genes significant at a q-value cutoff of 0.05, including 34 of the 37 batch 3 genes. Reducing the q-value threshold to 0.01 yielded 488 significant genes (including 32 of the 37 batch 3 genes), and decreasing the threshold further to 0.001 yielded 161 significant genes (including 28 of the 37). The GO cluster analysis for EB3 yielded 85 significant clusters including 33 of the 36 batch 3 clusters. These results indicated a clear increase in power to detect biological differences and the benefit of combining batches was clearly evident.

A.3 Derivation for Parametric EB Batch Adjustments

A.3.1 Empirical Hyperprior Estimation

In section 3.1, we assumed that the parametric forms for prior distributions on the batch effect parameters to be

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

The hyper-parameters γ_i, τ_i^2, λ_i, θ_i are estimated empirically from standardized data using the Method of Moments (MM), and Letting $\hat{\gamma}_{ig} = \frac{1}{n_i}\sum_j Z_{ijg}$ (batch *i* sample mean for gene *g*), estimates of γ_i and τ_i^2 are given (respectively) by

$$\bar{\gamma}_i = \frac{1}{G}\sum_g \hat{\gamma}_{ig} \text{ , and } \bar{\tau}_i^2 = \frac{1}{G-1}\sum_g \left(\hat{\gamma}_{ig} - \bar{\gamma}_i\right)^2 .$$

Additionally, letting $\hat{\delta}_{ig}^{\,2} = \frac{1}{n_i-1}\sum_j \left(Z_{ijg} - \hat{\gamma}_{ig}\right)^2$ (batch *i* sample variance for gene *g*) we calculate $\bar{V}_i = \frac{1}{G}\sum_g \hat{\delta}_{ig}^{\,2}$ (mean of the $\hat{\delta}_{ig}^{\,2}$) and $\bar{S}_i^2 = \frac{1}{G-1}\sum_g \left(\hat{\delta}_{ig}^{\,2} - \bar{V}_i\right)^2$ (variance of the $\hat{\delta}_{ig}^{\,2}$). Setting the sample moments $\bar{V}_i$, $\bar{S}_i^2$ equal to the theoretical moments of the *Inverse Gamma* distribution, namely $\theta_i/(\lambda_i\text{-}1)$ (mean) and $\theta_i^2/[(\lambda_i\text{-}1)^2(\lambda_i\text{-}2)]$ (variance), and then solving this system yields estimates for λ_i and θ_i as follows:

$$\bar{\lambda}_i = \frac{\bar{V}_i + 2\bar{S}_i^2}{\bar{S}_i^2} \text{ and } \bar{\theta}_i = \frac{\bar{V}_i^3 + \bar{V}_i\bar{S}_i^2}{\bar{S}_i^2}.$$

A.3.1 Obtaining the Parametric Batch Effect Adjustments

After the standardization procedure in Step 1 of Section 3.1, and assuming $\gamma_{ig} \sim N(\gamma_i, \tau_i^2)$, we apply Bayes' Theorem to find the conditional (posterior) distribution of γ_{ig}, denoted $\pi(\gamma_{ig} \mid \boldsymbol{Z}_{ig}, \delta_{ig}^2)$, which satisfies

$$
\begin{aligned}
\pi(\gamma_{ig} \mid \mathbf{Z}_{ig}, \delta^2_{ig}) &\propto L(\mathbf{Z}_{ig} \mid \gamma_{ig}, \delta^2_{ig})\pi(\gamma_{ig}) \\
&\propto \exp\left\{-\tfrac{1}{2\delta^2_{ig}}\sum\nolimits_j (Z_{ijg}-\gamma_{ig})^2\right\}\exp\left\{-\tfrac{1}{2\tau_i^2}(\gamma_{ig}-\gamma_i)^2\right\} \\
&= \exp\left\{-\tfrac{1}{2\delta^2_{ig}}\left[\sum\nolimits_j {Z_{ijg}}^2 - 2\sum\nolimits_j Z_{ijg}\gamma_{ig} + n_i{\gamma_{ig}}^2\right] - \tfrac{1}{2\tau_i^2}\left[{\gamma_{ig}}^2 - 2\gamma_{ig}\gamma_i + {\gamma_i}^2\right]\right\} \\
&\propto \exp\left\{-\tfrac{1}{2}\left(\frac{n_i\tau_i^2+\delta^2_{ig}}{\delta^2_{ig}\tau_i^2}\right)\left[{\gamma_{ig}}^2 - 2\left(\frac{\tau_i^2\sum\nolimits_j Z_{ijg} + \delta^2_{ig}\gamma_i}{n_i\tau_i^2+\delta^2_{ig}}\right)\gamma_{ig}\right]\right\}.
\end{aligned}
$$

By completing the square, the distribution above can be determined to be the kernel of a normal distribution with expected value

$$
E[\gamma_{ig} \mid \mathbf{Z}_{ig}, \delta^2_{ig}] = \frac{\tau_i^2\sum\nolimits_j Z_{ijg} + \delta^2_{ig}\gamma_i}{n_i\tau_i^2+\delta^2_{ig}}
$$

which, given $\hat{\gamma}_{ig}$, $\hat{\delta}^2_{ig}$, $\bar{\gamma}_i$, and $\bar{\tau}^2_i$ as defined above, can be estimated as

$$
\gamma^*_{ig} = \hat{E}[\gamma_{ig} \mid \mathbf{Z}_{ig}, \delta^{2*}_{ig}] = \frac{n_i\bar{\tau}_i^2\hat{\gamma}_{ig} + \delta^{2*}_{ig}\bar{\gamma}_i}{n_i\bar{\tau}_i^2 + \delta^{2*}_{ig}}.
$$

For the conditional posterior distribution of δ_{ig}^2, given γ_{ig} and *Inverse Gamma*(λ_i, θ_i) prior, we note that

$$
\begin{aligned}
\pi(\delta^2_{ig} \mid \mathbf{Z}_{ig}, \gamma_{ig}) &\propto L(\mathbf{Z}_{ig} \mid \gamma_{ig}, \delta^2_{ig})\pi(\delta^2_{ig}) \\
&\propto \left(\delta^2_{ig}\right)^{-\frac{n_i}{2}}\exp\left\{-\tfrac{1}{2\delta^2_{ig}}\sum\nolimits_j (Z_{ijg}-\gamma_{ig})^2\right\}\left(\delta^2_{ig}\right)^{-(\lambda_i+1)}\exp\left\{-{\theta_i}\big/{\delta^2_{ig}}\right\} \\
&= \left(\delta^2_{ig}\right)^{-\left(\frac{n_i}{2}+\lambda_i\right)-1}\exp\left\{-\frac{\theta_i + \frac{1}{2}\sum\nolimits_j (Z_{ijg}-\gamma_{ig})^2}{\delta^2_{ig}}\right\}.
\end{aligned}
$$

Which can be identified as an *Inverse Gamma* distribution with expected value

$$
E[\delta^2_{ig} \mid \mathbf{Z}_{ig}, \gamma_{ig}] = \frac{\theta_i + \frac{1}{2}\sum\nolimits_j (Z_{ijg}-\gamma_{ig})^2}{\frac{n_i}{2}+\lambda_i - 1}.
$$

Given the MM estimates for $\bar{\theta}_i$ and $\bar{\lambda}_i$ from above, the expectation above can be estimated by

$$\delta_{ig}^{2*} = \hat{E}[\delta_{ig}^2 \mid \mathbf{Z}_{ig}, \gamma_{ig}^*] = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_i}{2} + \bar{\lambda}_i - 1}.$$

Finally, notice that the estimate γ_{ig}* depends on $\delta_{ig}{}^{2}$* and vice versa. There are no closed form solutions for these parameters, and therefore they must be found iteratively. Starting with a reasonable value for $\delta_{ig}{}^{2}$* (e.g. use $\hat{\delta}_{ig}{}^2$), calculate an estimate of γ_{ig}*. Now use the newly found value of γ_{ig}* to estimate $\delta_{ig}{}^{2}$*. Iterate the previous steps until convergence. This can be shown to be a simple case of the EM Algorithm (Dempster et al., 1977), and typically only a few iterations (<30) are necessary to achieve very accurate estimates for the EB batch adjustments.

A.4 EB Batch Effect Parameter Estimates using Nonparametric Empirical Priors

The parametric forms for the prior estimates (from Step 2 in Section 3.1) were not satisfactory for data set 2, leading to the need for more flexible options for the prior distributions, so we use a nonparametric empirical prior to accommodate these data. We assume that the data has been standardized as in Step 1 in Section 3.1, and that the standardized data, Z_{ijg}, satisfies the distributional form, $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}{}^2)$. Also let $\hat{\gamma}_{ig} = \frac{1}{n_i}\sum_j Z_{ijg}$ and $\hat{\delta}_{ig}^2 = \frac{1}{n_i - 1}\sum_j \left(Z_{ijg} - \hat{\gamma}_{ig}\right)^2$ as in the previous section. We estimate the batch effect parameters γ_{ig} and $\delta_{ig}{}^2$, using estimates of the posterior expectations of the batch effect parameters, denoted $E[\gamma_{ig}]$ and $E[\delta_{ig}{}^2]$ respectively. Let $\boldsymbol{Z}_{ig}$ be a vector containing Z_{ijg} for j=1,…, n_i. Given the posterior distribution $\pi(\boldsymbol{Z}_{ig}, \gamma_{ig}, \delta_{ig}{}^2)$ of the data $\boldsymbol{Z}_{ig}$ and batch effect parameters $\gamma_{ig}, \delta_{ig}{}^2$, the posterior expectation of γ_{ig} is given by

$$E[\gamma_{ig}] = \int \gamma_{ig} \pi(\mathbf{Z}_{ig}, \gamma_{ig}, \delta_{ig}{}^2) d(\gamma_{ig}, \delta_{ig}{}^2).$$

Now let $\pi(\gamma_{ig}, \delta_{ig}{}^2)$ be the (unspecified) density function for the prior for the parameters $\gamma_{ig}, \delta_{ig}{}^2$, and let $L(\boldsymbol{Z}_{ig} \mid \gamma_{ig}, \delta_{ig}{}^2) = \prod_j \varphi(Z_{ijg}, \gamma_{ig}, \delta_{ig}{}^2)$, where $\varphi(Z_{ijg}, \gamma_{ig}, \delta_{ig}{}^2)$ is the probability density function of a $N(\gamma_{ig}, \delta_{ig}^2)$ random variable evaluated at Z_{ijg}. Using Bayes theorem, the integral above for can be written as

$$E[\gamma_{ig}] = \frac{1}{C(\mathbf{Z}_{ig})} \int \gamma_{ig} L(\mathbf{Z}_{ig} \mid \gamma_{ig}, \delta_{ig}{}^2) \pi(\gamma_{ig}, \delta_{ig}{}^2) d(\gamma_{ig}, \delta_{ig}{}^2) \qquad \textbf{Equation A.1}$$

where $C(\mathbf{Z}_{ig}) = \int L(\mathbf{Z}_{ig} \mid \gamma_{ig}, \delta_{ig}{}^2) \pi(\gamma_{ig}, \delta_{ig}{}^2) d(\gamma_{ig}, \delta_{ig}{}^2)$. We estimate both $C(\mathbf{Z}_{ig})$ and the integral in Equation A.1 using Monte Carlo integration (Liu, 2001, Gilks et al., 1996) over the empirically estimated pairs $(\hat{\gamma}_{ig}, \hat{\delta}_{ig}{}^2)$, which are considered random draws from $\pi(\gamma_{ig}, \delta_{ig}^2)$. Letting $w_{ig''} = L(\mathbf{Z}_{ig} | \hat{\gamma}_{ig''}, \hat{\delta}_{ig''})$ for g'' =1,...,G, $C(\mathbf{Z}_{ig})$ can be estimated by $\hat{C}(\mathbf{Z}_{ig}) = \frac{1}{n} \sum_{g''} w_{ig''}$ and the integral in Equation A.1 can be estimated by

$$\gamma_{ig}^* = \hat{E}(\gamma_{ig}) = \frac{\sum_{g''} w_{ig''} \hat{\gamma}_{ig''}}{n\hat{C}(\mathbf{Z}_{ig})}.$$

The same method is used to find the posterior expectation of δ_{ig}^2, and the nonparametric EB batch adjustments $\gamma_{ig}*$, δ_{ig}^2* are given by

$$\gamma_{ig}^* = \frac{\sum_{g''} w_{ig''} \hat{\gamma}_{ig''}}{\sum_{g''} w_{ig''}} \text{ and } \delta_{ig}^{2*} = \frac{\sum_{g''} w_{ig''} \hat{\delta}_{ig''}^2}{\sum_{g''} w_{ig''}}.$$

Using $\gamma_{ig}*$, δ_{ig}^2* for the batch adjustment estimates, the data are adjusted using the method described in Step 3 in Section 3.1.

A.5 Additional Discussion Topics

A.5.1 Standardization Procedure

The aim of the standardization procedure presented in section 3.1 is to reduce gene-to-gene variation in the data, because genes in the array are expected to have different expression profiles or distributions. However, we do expect that phenomena that cause batch effects to affect many genes in similar ways. To more clearly extract the common batch biases from the data, the standardization procedure standardizes all genes to have the similar overall mean and variance. On this scale, batch effect estimators can be compared and pooled across genes to create robust estimators for batch effects. Without standardization, the gene-specific variation increases the noise in the data and inflates the prior variance, decreasing the amount of shrinkage that occurs. Therefore standardization is crucial for EB shrinkage methods. However this feature is not present in many EB methods for Affymetrix arrays.

A.5.2 Robustness of the EB method

An array with *n* genes can be thought of as a point in *n*-dimensional space, and batch adjustments are moving one cluster of points (one batch) to match another using scale and location shifts. The DWD, SVD and L/S methods don't work well for small samples because they are selecting multi-dimensional adjustments using only a few points (samples) in multidimensional space. In order to visually inspect the effect of batch adjustments, one can consider these data by plotting them on fewer dimensions and still get a reasonable idea of the effect of the batch adjustments. Figure 4 contains two dimensional plots (two genes) from data set 1, selected because both genes contain outlying observations for the same sample. By empirical observation, it seems that there are only small batch effects in the data in these two dimensions (Figure 4 (a)). Additionally, it seems that the outliers are not caused by batch effects, but are truly an outlying observation in both directions. The outliers highly affect the outcome of the L/S estimates, especially the variance (Figure 4(b)) as it appears that the variances of the batches without outliers are over-inflated from the adjustment. Without a variance adjustment, there still would be an over-adjustment for the mean shift. However, the EB adjustment (Figure 4(c)) works very well in these dimensions. Examples without outlying observations typically show only slight difference between the L/S and the EB methods and in larger samples, the L/S methods are more robust and the EB and L/S methods are usually very similar.

A.5.3 Adjustments for balanced designs

In a balanced factorial ANOVA experiment the L/S model covariates (β_g in Equation 2.1) are always orthogonal. As a result, the L/S estimates additive batch effects do not change in the presence of (i.e. when adjusting for) other covariate values. Therefore, in the balanced factorial ANOVA model the L/S adjusted data is often the same whether or not covariates are included in the model. Additionally, the standardization procedure has no effect on these L/S adjustments, and is therefore unnecessary. Note that this does not apply when numerical covariates are included or when sample sizes or conditions are unbalanced across batches. In contrast, the EB adjustments are sensitive to the inclusion of covariates in the L/S model because the residual standard error from the L/S model is an important factor in the shrinkage of the L/S estimates toward the empirical prior. The objective is to find the best possible estimate for the residual error from the L/S model to best estimate the EB batch effect parameters. Therefore if sample sizes are large enough, it is recommended to model all available covariates expected to be significant.

Because the batch sizes from data set 1 are small and because the experimental design is balanced and factorial no covariates are included in the model (no X or βs). For data set 2, X is an indicator vector with elements equal to 1 or 0 if the corresponding sample is an RNAi sample

A.5.4 Sample Size

Finally, note that the EB adjustments are dependent on several factors: the standardized batch mean, the empirical prior distribution, the (residual) variance estimate, and the sample size within the batch. Sample size appears to be a very influential factor in this EB method. As sample sizes increase, the EB adjustment converges to the L/S batch effect parameter estimate and diverges from the empirical prior (see Equation 3.1). In data sets were the sample sizes were relatively moderate (15-25 samples per batch) there is usually not much difference between the EB and the L/S adjustments. However, since the batch sizes of the data from Section 1 were small, some of the adjustments (particularly those with outlying observations) were strongly influenced by the prior, making them more robust. As shown in the examples above, we conclude that the EB method is advantageous for small sample size because is less susceptible to outliers in the data.

A.5.5 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/. Computing times for the batch effect adjustments on data set 1 (3 batches, 12 arrays, 22,283 probes/array) were less than 1 minute for the parametric method and less than 8 minutes for the non-parametric approach on a standard laptop (Windows XP on a 2.13 GHz Intel Pentium M processor). For data set 2 (3 batches, 30 arrays, 54,675 probes/array) the parametric adjustment took less than 3 minutes, and the non-parametric method took just under an hour to complete.

Additional References

Dempster, A. P., Laird, N. and Rubin, D. B. (1977) *Journal of the Royal Statistical Society, Series B (Methodological),* 39, 1-38.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice,* Chapman & Hall/CRC, London.

Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing,* Springer-Verlag New York, Inc., New York.

Storey, J. D. and Tibshirani, R. (2003) *Proc Natl Acad Sci U S A,* 100, 9440-5.

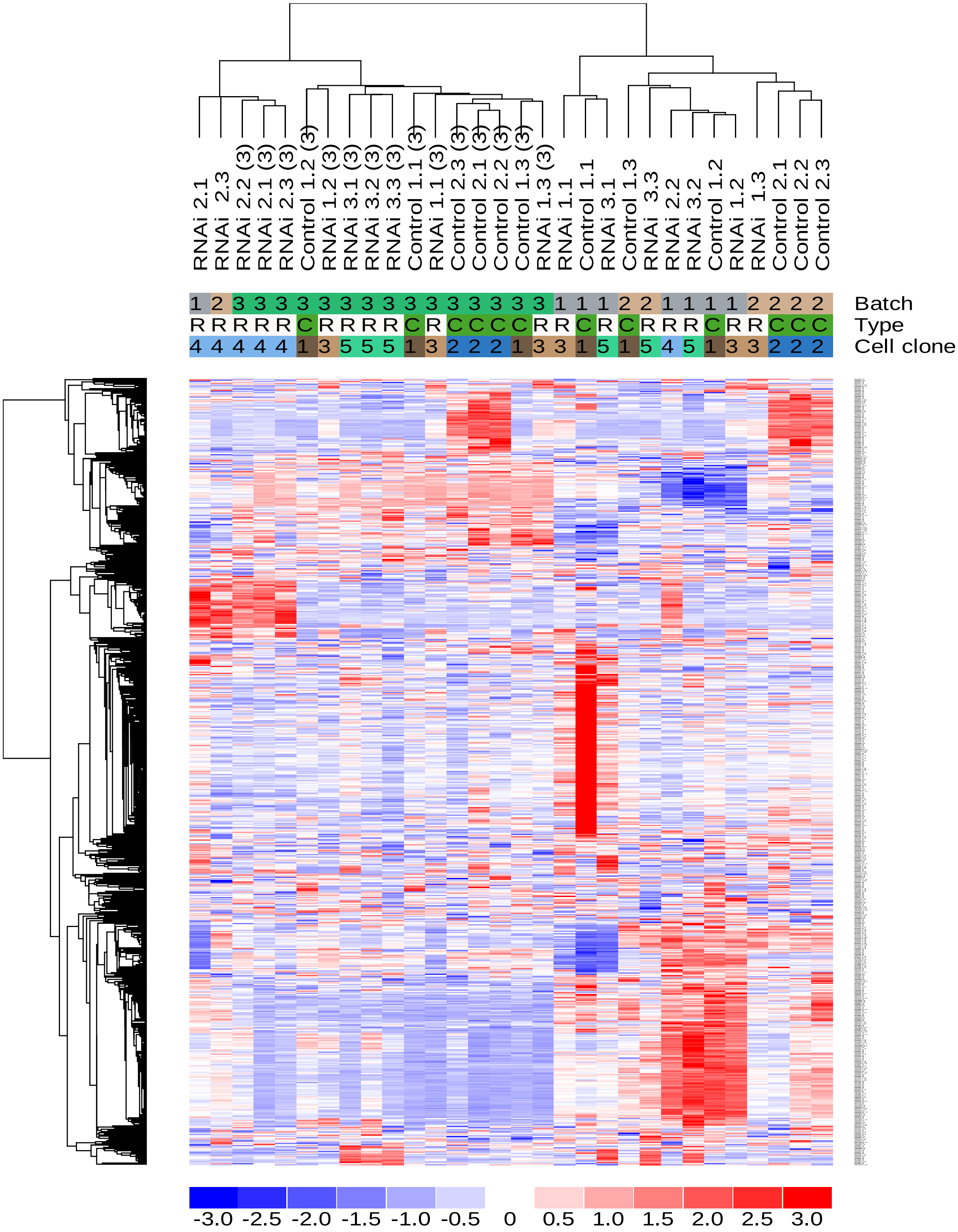

Figure A. 1: A heatmap clustering of data set 2. 698 genes with large variation across all the samples are clustered as in Figure 1. The sample legends on the top are: C (Control), R (RNAi). There appears to be very prevalent batch effects in these data, particularly for the third batch.

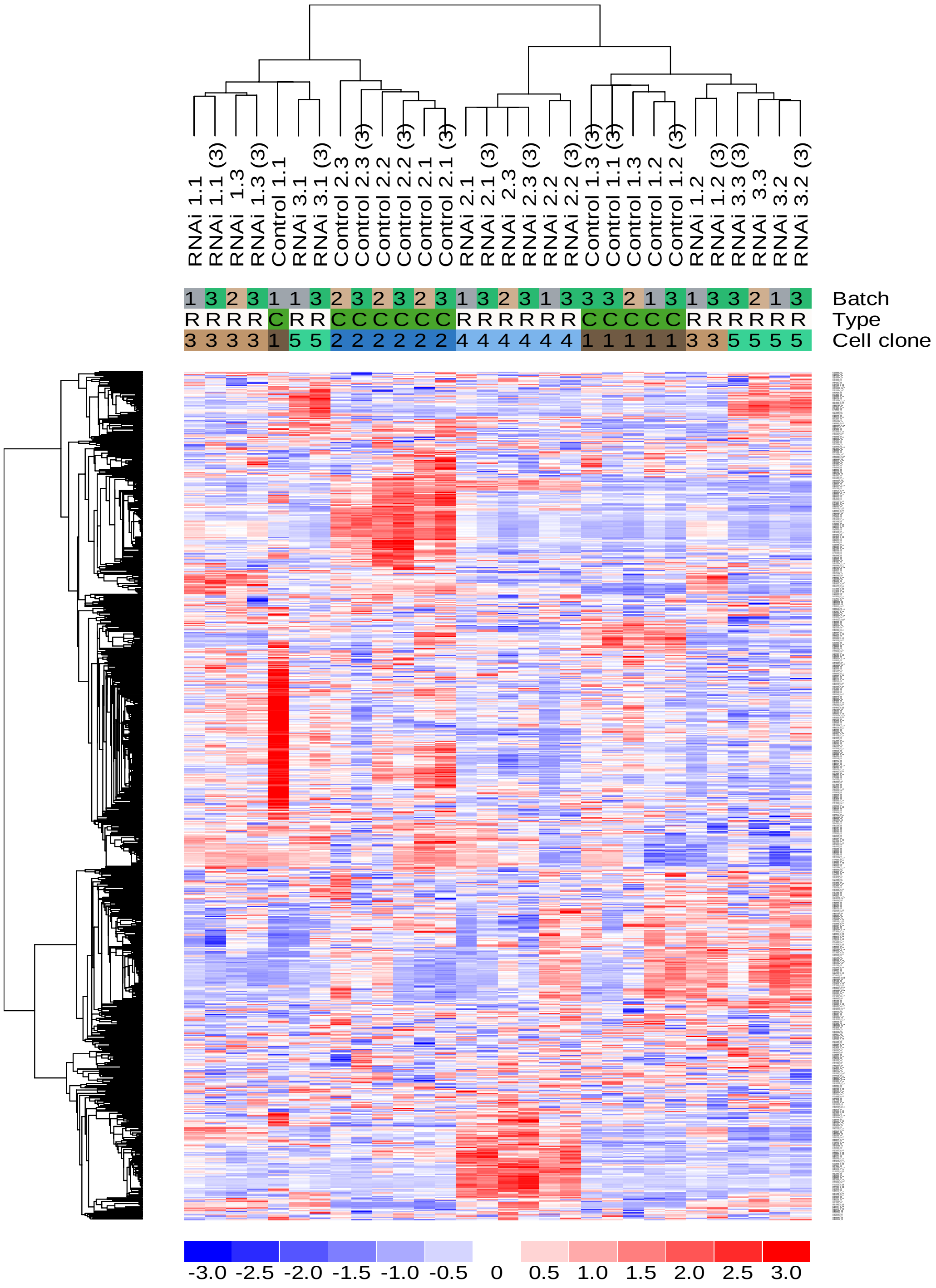

Figure A. 2: A heatmap diagram of 770 genes from data set 2 after applying the EB batch adjustments. After adjustment there is no evidence of batch effects. The samples cluster in small groups based on treatment status (RNAi or control) and cell clone.