Biostatistics (2014), **0**, 0, *pp.* 1–17
doi:10.1093/biostatistics/manuscript

Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses

VEGARD NYGAARD[†], EINAR ANDREAS RØDLAND[†]

Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital HF - Radiumhospitalet, Montebello, 0310 Oslo, Norway

EIVIND HOVIG[*]

Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital HF - Radiumhospitalet, Montebello, 0310 Oslo, Norway

Institute of Cancer Genetics and Informatics, Oslo University Hospital HF - Radiumhospitalet, Montebello, 0310 Oslo, Norway

Department of Informatics, University of Oslo, 0316 OSLO Norway

ehovig@ifi.uio.no

Summary

Removal of, or adjustment for, batch effects or centre differences is generally required when such effects are present in data. In particular, when preparing microarray gene expression data from multiple cohorts, array platforms, or batches for later analyses, batch effects can have confounding effects. Many methods and tools exist for this purpose. One method, ComBat, which is part of the R package sva, is particularly popular due to its ability to remove batch differences even when batches are small and heterogeneous. It also has the option of preserving the difference between study groups, as batch adjustments may otherwise bias, usually deflate, group differences when study groups are not evenly balanced across batches. Using a two-way ANOVA model to simultaneously estimate both group an batch effects is a natural approach in such cases. Unfortunately, this frequently used and recommended approach may systematically induce incorrect group differences in downstream analyses when groups are distributed between the batches in an unbalanced manner. The scientific community seems to be largely unaware of how this approach may lead to false discoveries. *[Har moderert påstanden noe, jfr. reviewer 2, pt. 6.]*

Key words:

[§]Manuscript version: November 17, 2014.
[†]These two authors contributed equally.
[*]To whom correspondence should be addressed.

1. Introduction

Extraneous variables, if left unaccounted for, have the potential to lead an investigator into drawing wrong conclusions. A common example is "batch effects" caused by reagents, microarray chips, and other equipment made in batches that may vary in some way, which often have systematic effects on the measurements. A similar example is "centre effects", when samples or data come from multiple sources. See Luo *and others* (2010) for more examples.

For a typical experiment of comparing differences between study groups, the presence of batch effects will decrease the statistical power, since it adds variation to the data. If the batch–group design is unbalanced, i.e. if the study groups are not equally represented in all batches, batch effects may also act as a confounder and induce false differences between groups (Leek *and others*, 2010).

The standard way to handle an extraneous variable is to include it in the statistical model employed in the inquiry. However, many analysis tools for high throughput data do not cater for this option, and when available, it could still be outside the competence of the investigator. Therefore, an alternative two-step procedure has emerged. First the batch effects are estimated and removed, creating a "batch effect free" data set. Then, the statistical analyses are performed on the adjusted data without further consideration of batch effects. This appealing compartmentalization is also convenient for practical purposes, for example when data-processing and statistical analyses are performed by different personnel. Unfortunately, as we demonstrate in this paper, when the batch–group design is unbalanced, this approach may be unreliable.

A simple removal of batch effects can be achieved by subtracting the mean of the measurements in one batch from all measurements in that batch, i.e mean adjustment or one-way ANOVA adjustment as implemented in the method `pamr.batchadjust` from the `pamr` package in R. When the batch–group design is balanced, mean-adjustment will remove most, but not necessarily all, variance attributed to batch and leave the between group variance, thus increasing the statistical power. However, when the batch–group design is unbalanced, batch differences will in part be influenced by group differences, and thus batch correction will reduce group differences and thereby reduce the statistical power. In very uneven group–batch designs with multiple groups, spurious group differences may even be induced in this way. Figure 1c illustrates both of these effects.

To mitigate the above problems, one may simultaneously estimate batch effects and group differences, e.g. using a two-way ANOVA, and only remove the batch differences from the data. Effectively, group differences are estimated based on within batch comparisons, and applied to the batch adjusted data. In a balanced group–batch design, estimates of group differences and batch effects are independent, and this approach becomes identical to the above described zero-centering per batch. However, if the group–batch design is heavily unbalanced, estimation of group differences and batch effects are interdependent, and when applying the estimated group differences across the entire data set, point estimates are used while the estimation errors are ignored. Subsequent analyses will then systematically underestimate the size of estimation errors and exaggerate the confidence in group differences. Figure 1d illustrates how confidence interval are deflated by this batch adjustment method by comparing them to confidence intervals from the original ANOVA in Figure 1e.

A popular tool for batch adjustment of gene expression data is ComBat (Johnson *and others*, 2007) which is now included in the R bioconductor package `sva` (Leek *and others*, 2012). This allows the inclusion of covariates, e.g. group difference, the effects of which should not be removed in the batch adjustment. In addition, it uses an empirical Bayes approach to avoid over-correction for batch effects, which is critical for small batches. Increasingly, the inclusion of group difference

as a covariate when removing batch effects has been recommended, both in the `sva` tutorial and user fora. ComBat has thus helped popularise this approach. Based on the actual use of batch adjustment methods like ComBat by the authors and others, ourselves included, we suspect that adjusted data are commonly treated as "batch effect free" in subsequent analyses, and as a consequence, confidence in group effects has been overestimated and false results reported. The commercial software Partek [citation or reference needed?], and the R packages `limma` (Smyth and Speed, 2003) and `ber` (Giordan, 2013) also offer batch adjustment with covariates. Seeing that `ber` is sparsly used, and `Partek` and the relevant method "removeBatchEffect" in `limma` note that the this is not intended used prior to linear modeling, we choose to use the more popular and well documented ComBat to demonstrate the problem. We note that the `sva` are now being updated to address these concerns.

When sample or batch sizes are small, statisticians would most likely take extra precautions. However, batch adjustment using two-way ANOVA or ComBat on unbalanced data sets may be just as harmful for large samples and batch sizes as for small. For example, group comparisons using one-way ANOVA on the batch adjusted data will essentially result in F-statistics that are inflated by a fixed factor which depends on the unevenness of the design, rather than the size of the sample or batches. The effect of this may be further exacerbated by running these analyses a large number of times, e.g. on thousands of genes, and use false discovery rate to determine significant cases: an approach that is particularly sensitive to inflated false positive rates.

In conclusion, when study groups are unevenly distributed across batches, batch effects can have confounding effects if not properly handled. This is difficult to do through a two-step procedure as batch adjustment will not be able to produce a "batch effect free" data set, as has also been observed in Buhule *and others* (2014). Estimating batch and group effects simultaneously through two-way ANOVA tries to overcome this, and does succeed at removing systematic biases in the group differences induced by the batch effects, but at the cost of making the observations interdependent to an extent that may invalidate subsequent analyses. Proper statistical analysis of unbalanced designs require models that handle batch effects as part of the analysis. However, the best approach is to ensure a balanced study design from the start, to avoid data analysis problems as well as the loss of statistical power that ensues when batch and group effects need to be disentangled.

2. Methods for batch effect correction

2.1 *Model for data with batch effects*

We will base our discussion on a simple model for data with batch effects:

$$Y_{ijr} = \alpha + \beta_j + \gamma_i + \epsilon_{ijr} \tag{2.1}$$

where $i = 1, \ldots, m$ are the different batches, $j = 1, \ldots, M$ are different study groups that we wish to compare, and $\epsilon_{ijr} \sim N(0, \sigma^2)$ are the error terms for samples $r = 1, \ldots, n_{ij}$ within batch i and group j.

When combining data from more diverse data sources, e.g. microarray data from different platforms, a more general model is required. The model used by ComBat (Johnson *and others*, 2007) is

$$Y_{ijgr} = \alpha_g + X_r\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijgr} \tag{2.2}$$

where $g = 1, \ldots, G$ are different measurements, e.g. genes, performed for each sample, and X is the design matrix which in our case will indicate the study group. This permits independent

rescaling of data from different batches. In addition, Johnson *and others* (2007) uses an empirical Bayes approach to estimate γ_{ig} and δ_{ig} to stabilise estimates, which is critical for use with small batches.

For simplicity, we consider the case with a single gene and constant scale, i.e. $\delta_{ig} = 1$. Empirical Bayes will tend to shrink the estimates of γ_{ig}, and thus the amount of batch adjustment, in cases where batch effects are small or cannot be accurately estimated, e.g. for small batches. However, in cases with large batches or substantial batch effects, it should differ little from the two-way ANOVA approach. We will therefore explain the problem in the simpler case where no empirical Bayes is applied, and later demonstrate that it also affects the ComBat approach.

2.2 *Standard batch correction methods*

A common ambition of batch effect adjustments is to remove batch differences in such a way that downstream analyses of the adjusted data may be done without further corrections for batches. We illustrate this using Figure 1, where the first frame contains the "true" values with no batch differences, and the remaining frames show values with various levels of batch effects and batch effect corrections.

2.2.1 *Zero-centering batches:* The most common method for removing batch effects is to zero-centre each batch:

$$\tilde{Y}^0_{ijr} = Y_{ijr} - \bar{Y}_i \quad \text{where} \quad \bar{Y}_i = \frac{1}{n_{i\text{-}}} \sum_{j=1}^{M} \sum_{r=1}^{n_{ij}} Y_{ijr},\ n_{i\text{-}} = \sum_{j=1}^{M} n_{ij}. \tag{2.3}$$

An alternative is to centre each batch to the common average by adding the average value $\bar{Y}$ across the entire data set: i.e. $\tilde{Y}^{\text{avg}}_{ijr} = \tilde{Y}^0_{ijr} + \bar{Y}$. When comparing groups, the common value $\bar{Y}$ has no effect, and so this is equivalent to zero-centring each batch. If the groups are unevenly represented in the different batches, the batch average $\bar{Y}_i$ will tend to capture, through $\bar{\beta}_i$, group differences, as well as batch effects:

$$\tilde{Y}^0_{ijr} = \beta_j - \bar{\beta}_i + \epsilon_{ijr} - \bar{\epsilon}_i \quad \text{where} \quad \bar{\beta}_i = \sum_{j=1}^{M} \frac{n_{ij}}{n_{i\text{-}}} \beta_j,\ \bar{\epsilon}_i = \frac{1}{n_{i\text{-}}} \sum_{j=1}^{M} \sum_{r=1}^{n_{ij}} \epsilon_{ijr}. \tag{2.4}$$

Thus, batch centering will tend to reduce group differences in an unbalanced design, and reduce the power of downstream analyses. By reducing the differences between some groups, i.e. those found together in the same batches, one may also induce false differences between other groups, as is demonstrated in Figure 1c.

2.2.2 *Batch adjustment using two-way ANOVA to estimate batch effects:* Removing batch effects while retaining group differences can be achieved through a two-way ANOVA, in which group effects, β_j, and batch effects, γ_i, are estimated simultaneously. Batch adjusted values may then be obtained by subtracting the estimated batch effects, $\hat{\gamma}_i$:

$$\tilde{Y}^{\text{cov}}_{ijr} = Y_{ijr} - \hat{\gamma}_i = \alpha + \beta_j + (\gamma_i - \hat{\gamma}_i) + \epsilon_{ijr}. \tag{2.5}$$

This will yield batch adjusted values, where any systematic bias induced by the batch differences has been removed, while the group differences are retained.

The estimation error $\hat{\gamma}_i - \gamma_i$ affects all values within the same batch in the same manner. Thus, while the aim is to remove spurious dependencies within batches, it may also induce new dependencies. The batch effect estimation errors will influence group effects in proportion to how well the group is represented in each batch:

$$\tilde{Y}_{\cdot j}^{\text{cov}} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{m} \sum_{r=1}^{n_{ij}} \tilde{Y}_{ijr} = \alpha + \beta_j + \bar{\epsilon}_{\cdot j} - \sum_{i=1}^{m} \frac{n_{ij}}{n_{\cdot j}} (\hat{\gamma}_i - \gamma_i) \quad \text{where} \quad n_{\cdot j} = \sum_{i=1}^{m} n_{ij} \tag{2.6}$$

so that

$$\tilde{Y}_{\cdot j}^{\text{cov}} - \tilde{Y}_{\cdot j'}^{\text{cov}} = (\beta_j - \beta_{j'}) + (\bar{\epsilon}_{\cdot j} - \bar{\epsilon}_{\cdot j'}) - \sum_{i=1}^{m} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}} \right) (\hat{\gamma}_i - \gamma_i). \tag{2.7}$$

In a balanced group–batch design, the estimation error $\hat{\gamma}_i - \gamma_i$ has the same effect for all groups, and thus does not influence group comparisons. In an unbalanced design, however, it will induce systematic differences between groups which, when ignored in downstream analyses, may lead to biases or over-confidence in estimated group differences. In Figure 1d, the effect of batch correction using group as covariate is shown with confidence intervals of the corrected data. For comparison, least square means estimates (R package `lsmeans`) are used in Figure 1e to illustrate more appropriate confidence intervals that incorporate the uncertainties of the batch effect estimates.

3. Results

3.1 *A simple sanity check*

The undesired consequences of preserving group effects when correcting for batch effect is readily illustrated with a sanity check using random numbers. The documentation accompanying the `sva` library has an executable example demonstrating how to adjust a data set with ComBat, followed by an F-test. Replacing the real data with random numbers from a standard normal distribution, $\mathrm{N}(0, 1)$, but otherwise following the instructions, will generate the p-value distribution shown in Figure 2a.

As can be seen from the QQ plot in Figure 2b, the main effect of the procedure is to inflate the F-statistic by a factor. The size of this factor depends on how unbalanced the group–batch design is, rather than on the sample size in itself. Thus, increasing the sample size will not reduce the problem.

If the number of samples is increased and there are no actual batch effects present, the empirical Bayes estimates used by ComBat will shrink the batch effect estimates and thus moderate the batch adjustments. However, if batch differences are added which are not constant across all genes, the problem remains even as the samples size increases.

3.2 *Explanation for the simple two-group comparison*

To explain more clearly what is happening, and to quantify the size of the problem, we may consider the simple case of estimating the difference $\Delta\beta = \beta_A - \beta_B$ between two groups, A and B, when there are m batches with batch $i = 1, \ldots, m$ containing n_{iA} and n_{iB} samples from each of the two groups.

3.2.1 *Group comparison from two-way ANOVA:* If we estimate the group difference within batch i, we get

$$\Delta\hat{\beta}_i = \bar{Y}_{iA} - \bar{Y}_{iB} \sim \mathrm{N}\left(\Delta\beta, \frac{\sigma^2}{\nu_i}\right) \quad \text{where} \quad \bar{Y}_{ij} = \frac{\sum_{r=1}^{n_{ij}} Y_{ijr}}{n_{ij}},\ \nu_i = \frac{1}{\frac{1}{n_{iA}} + \frac{1}{n_{iB}}}, \tag{3.8}$$

from which we may express the overall estimate of $\Delta\beta$

$$\Delta\hat{\beta} = \frac{\sum_{i=1}^{m} \nu_i \Delta\hat{\beta}_i}{\nu} \sim \mathrm{N}\left(\Delta\beta, \frac{\sigma^2}{\nu}\right) \quad \text{where} \quad \nu = \sum_{i=1}^{m} \nu_i. \tag{3.9}$$

This is the same as would be obtained from a two-way ANOVA analysis.

3.2.2 *Group comparison from one-way ANOVA after batch adjustment:* If batch and group effects are estimated using a two-way ANOVA, the estimate $\Delta\hat{\beta}$ will be as stated above, and so the estimated group difference is unaffected. The batch effects are then removed, and the estimated group differences retained, leaving the estimated $\Delta\hat{\beta}$ unchanged by the batch adjustment.

However, if this batch adjusted data set is analysed without considering batch effects, the variance of $\Delta\hat{\beta}$ will be computed under the assumption that it is derived from a comparison of $n_A = \sum_{i=1}^{m} n_{iA}$ versus $n_B = \sum_{i=1}^{m} n_{iB}$ samples, and thus satisfy

$$\Delta\hat{\beta} \sim \mathrm{N}\left(\Delta\beta, \frac{\sigma^2}{\nu_0}\right) \quad \text{where} \quad \nu_0 = \frac{1}{\frac{1}{n_A} + \frac{1}{n_B}}. \tag{3.10}$$

Using Jensen's inequality, we may derive

$$\nu_0 = \frac{1}{\frac{1}{n_A} + \frac{1}{n_B}} = n \cdot \frac{n_A}{n} \cdot \frac{n_B}{n} \geqslant \nu = \sum_{i=1}^{m} \frac{1}{\frac{1}{n_{iA}} + \frac{1}{n_{iB}}} = \sum_{i=1}^{m} n_i \cdot \frac{n_{iA}}{n_i} \cdot \frac{n_{iB}}{n_i} \tag{3.11}$$

with equality if and only if the ratios $n_{iA} : n_{iB} = n_A : n_B$ for all batches $i = 1, \ldots, m$.

In effect, ν represents the effective sample size in the unbalanced group–batch design, while ν_0 represents the nominal sample size when batches are ignored. The ratio $\nu/\nu_0 \leqslant 1$ indicates to what extent the two-step procedure leads to underestimates of the random variability, and hence false or over-confidence in group differences.

3.3 *Distribution of F-statistic in the general case*

The detailed calculations for general linear models are presented in the appendix. Here, we summarise the results for two-way ANOVA based batch adjustments based on (2.1) when there are m batches, M study groups, and n_{ij} samples in batch i and group j.

A one-way ANOVA checking for group differences will assume that the F-statistic follows an $\mathrm{F}_{q,n-q-1}$ distribution. However, if the data have been batch adjusted prior to this analysis, the approximate distribution of this F-statistic, with correct first two momenta, will be

$$F \approx \frac{\tilde{q}\tilde{\sigma}^2}{q\sigma^2} \cdot \left(1 + \frac{r}{n-q-r-1}\right) \cdot \mathrm{F}_{\tilde{q},n-q-r-1} \tag{3.12}$$

where $\tilde{q}$ and $\tilde{\sigma}^2$ are as computed in the appendix and satisfy $\tilde{q}\tilde{\sigma}^2 \geqslant q\sigma^2$ and $\tilde{q} \leqslant q$ with equalities if and only if the groups are evenly represented in all batches.

The two factors both cause a bias in the F-statistic. The first factor, $\tilde{q}\tilde{\sigma}^2/q\sigma^2$, captures the unbalancedness of the design but is otherwise independent of the sample size. The second factor, $q - r/(n - q - r - 1)$ captures loss of variance due to the centering of the data, but is primarily a problem when there are few samples in each batch. Finally, the degrees of freedom, $\tilde{q}$ and $n - q - r - 1$, of the F-distribution are smaller than assumed by the $\mathrm{F}_{q,n-q-1}$ distribution, and so the variance is larger.

As the sample size increases, assuming the distribution between groups and batches remain unchanged, the effect of the second factor and the second degree of freedom will diminish. However, the factor $\tilde{q}\tilde{\sigma}^2/q\sigma^2$ will remain, as will the reduced degree of freedom $\tilde{q}$.

3.4 *Examples of undesired consequences*

The extent to which batch adjustment with group differences retained will confound subsequent analyses depends on the batch–group balance. We have reanalysed two cases with varying degree of unbalance. The first case represents a "worst case" scenario since it both has a very unbalanced batch–group design, and was in part selected since it provided a methods description which allowed us to assess the use of ComBat, and perform a full re-analysis of the data. The second was taken from the original publication of the ComBat method (Johnson *and others*, 2007).

3.4.1 *Experiment 1* In an experiment described in Towfic *and others* (2014), the effect of glatiramer acetate (a medicine for multiple sclerosis) was compared to the effect of a generic drug. Cells were treated with glatiramer acetate (34 samples), the generic drug (11 samples), or one of 14 other treatments (64 samples), and mRNA was measured using microarrays. A batch effect correlating to the chip (Illumina WG-6_V2, six samples per chip, 17 chips in total) was observed and adjusted for with ComBat using treatment as a covariate. Batch adjusted data were then tested for differentially expressed genes, yielding hundreds of differentially expressed probes (Table S5, Towfic *and others*, 2014) using a FDR < 0.05 threshold. Unfortunately, the batch–treatment design was highly unbalanced, with several batches having only one of the main treatments of interest. We re-analyzed the cited data (GEO: GSE40566) without using ComBat, but instead blocked for batch effect in `limma`, detecting only 9 differentially expressed genes at FDR < 0.05. The sanity check outlined above with random numbers was also carried out. The distribution of p-values for different settings are shown in Figure 3a.

It should be noted that the original analyses by Towfic *and others* (2014) differed from our re-analysis in one critical manner which was clarified upon contacting the authors. Their analyses used a different preprocessing of the microarray data (GEO: GSE61901) in which two technical replicates of each biological sample was included, corresponding to the two different slots of the bead-microarray. For analyses where the two slots have been combined to provide one set of expression values per sample, they refer to Bakshi *and others* (2013) in which Partek was used for batch adjustment.

3.4.2 *Experiment 2* The supporting information of the original ComBat article (Johnson *and others*, 2007) demonstrates the method on cells inhibited for the expression of the TAL1 gene compared to controls on a microarray platform (denoted "Data set 2"). The experiment consist of 30 samples in 3 batches (batch1:6/2, batch2:3/4 and batch3:9/6 treatment/control samples). ComBat was applied followed by a T-test, in order to identify differentially expressed genes. First, we reproduced their analysis, including the adjustment by ComBat, but using `limma` instead of the T-test, resulting in 1003 probes ($q < 0.05$). Then, we analysed their data without batch

adjustment in ComBat, but blocking for batch in `limma`, resulting in 377 probes ($q < 0.05$). In addition, the sanity checks outlined above were performed. The distribution of P-values for different settings are shown in Figure 3b. In contrast to the results obtained for Towfic *and others* (2014) (Figure 3a), the P-value distributions for the alternative analysis do not indicate a huge difference. Nevertheless, the P-values may still be somewhat deflated in the ComBat adjusted analysis.

4. Discussion

The use of study group, or other form of outcome, as a covariate when estimating and removing batch effects is problematic if the data is treated as "batch effect free" in subsequent analyses. When the group–batch distribution is unbalanced, i.e. where batches do not have the same composition of groups, this will lead to deflated estimates of the estimation errors, and over-confidence in the results. The problem is essentially independent of sample size as illustrated in Figure 3c.

The size and impact of the problem will depend greatly on how unbalanced the group–batch distribution is: if it is only moderately unbalanced, it need not be a concern, whereas in heavily unbalanced cases it may have a huge influence. The impact is also more likely to be notable when used to analyse a large number of features, e.g. a large set of genes, followed by multiple testing corrections such as false discovery rate, as the effect is more pronounced for more extreme values.

4.1 *Increased emphasis on preserving group difference*

Multiple batch adjustment tools exist, most of which have batch centering as the main approach, but also offering the inclusion of covariates. Some of these give qualified warnings against using batch adjusted data in subsequent analyses.

In the original ComBat article (Johnson *and others*, 2007), the main objective was to employ an empirical Bayes method to allow better handling of small batches. The inclusion of covariates, e.g. for retaining group differences for unbalanced designs, was optional and appeared subordinate, only exemplified in the supplementary information. However, over the years, this feature became more important, judging from advice given on user fora. Later, when ComBat was incorporated into the `sva` package (Leek *and others*, 2012), specification of a covariate model became required: batch adjustment without covariates was no longer a default behaviour. Although specification of a null model was still possible, the help file specifies that covariates should contain "outcome of interest and other covariates besides batch". Thus, we expect this usage to have grown more common over time.

4.2 *Motivation for this warning*

Our knowledge of the problem discussed in this article came through a typical use case when trying to adjust for batch effects in an unbalanced data set using ComBat. Upon realizing that the confidence in the estimated group differences was exaggerated, we searched the literature for a better understanding of correct use and the potential limitations of ComBat. However, the authors of ComBat and the `sva` package recommended including study group as a covariate as we had done: this is now changing as the supporting documentation of `sva` is being updated to warn users of the potential dangers. In addition, other studies investigating batch effects were mostly recommending ComBat without much concern (Kupfer *and others*, 2012, Kitchen *and*

others, 2011, Chen *and others*, 2011). A brief inquiry into some of the articles citing ComBat (466 in Web of Science) revealed few reported problems, although their method descriptions regarding ComBat were mostly sparse, limited to one or two sentences. Some used ComBat with covariates, some without, but we did not find any that addressed the potential problems related to use of covariates. However, the inability of batch adjustment methods in dealing with unbalanced designs has been noted (Buhule *and others*, 2014).

A further indication of the carefree use of the procedure was the frequent omittance of program parameters that were applied, i.e batch labels or whether group labels were supplied as covariates. Often, no effort was undertaken in order to substantiate the existence of batch effects, beyond stating the presence of batches. The incorporation of ComBat into various analysis tools appears to make it more accessible, but its use less transparent. Such tools include GenePattern (Reich *and others*, 2006, version 3.9.0), AltAnalyze (Emig *and others*, 2010, version 2.0.8) and SCAN.UPC (Piccolo *and others*, 2013, version 2.6.3), which offers ComBat with covariates, as well as TCGA and inSilicoMerging (Taminau *and others*, 2012, version 1.8.7), presently without covariates as an option. Especially troublesome, are the ComBat implementations in ChAMP (Morris *and others*, 2014, version 1.2.8) and intCor (Fernández-Albert *and others*, 2014, version 1.03), where the use of covariates seems automatic and hidden from the user in the method call, while there is no mention of covariates in the documentation or articles.

These problems are not specific to ComBat, but are shared by other batch adjustment methods: indeed, the empirical Bayes approach used by ComBat will dampen the batch adjustments and may thus reduce the problem slightly relative to a more direct two-way ANOVA approach. Although some of the tools warn against performing subsequent analyses on batch adjusted data, these warnings are often somewhat vague and seem to be primarily concerned about the degrees of freedom in the data set. Batch adjustment will reduce the degrees of freedom in the data which may influence the distribution of the test statistics in subsequent analyses, e.g. F-statistics from ANOVA. For small data sets or where there are many small batches this may be a problem, while it would be less of a concern when there are a fair number of samples in each batch. However, we find that the main problem is not related to the degrees of freedom, but estimation errors during the batch adjustments which get applied across the data, which can induce group differences in the adjusted data set even in the absence of batch effects.

We are concerned, in light of our findings, that many published results from batch adjusted data using study group as covariate may be unreliable. Furthermore, the frequent lack of proper method description accompanying published results makes it hard to judge if they are affected or not. We hope that our warning will help caution the scientific community against this particular approach.

4.3 *Practical advice*

The main advice, particularly when batch effects are significant, is to ensure a balanced design in which study groups are evenly distributed across batches.

For an investigator facing an unbalanced data set with batch effects, our primary advice would be to adjust for batch inside the statistical test, and avoid the two-step procedure outlined above. If this is not possible, batch correction using outcome as covariate should only be performed with great caution, and downstream confidence estimates must be treated with suspicion.

Knowing that adjustment for batch effects while preserving the group difference may lead to varying degree of false results, to what extent can an investigator trust published results where such a method has been applied? Essentially, when the batch–group configuration is balanced,

or group difference is ignored (i.e. group labels not given as parameters to ComBat), problems related to preserving group differences will not occur. For other cases, a re-analysis without using this approach is the most rigorous path. However, such a re-analysis is not feasible if the downstream analysis can not independently adjust for batch effects. To reach a reliable result, batch effects need to be handled in some way or another. To make matters worse, a re-analysis relies on the availability of the raw data and a description of processing and analysis steps taken in the original work. Even when this is available, the necessary statistics and bioinformatics skills and work hours could still be in short supply.

A superficial assessment of the reliability of results based on batch corrected data with group differences retained can often be made. A major concern would be batches where either of the groups of interest are missing or strongly under-represented, which would contribute to the nominal sample size, but not to the effective sample size. In section 3.2, the ratio $\nu_0/\nu \geqslant 1$ indicates by how much the effective sample size is overestimated when comparing two groups. A similar computation, although more complicated, can be made for multiple groups *refer to appendix for details*. However, the two-group assessment should still give a fair idea of the reliability of pairwise group comparisons. If the ν_0/ν ratio is close to 1, results should still be reliable, while a ratio much larger than 1 will indicate exaggerated confidence. If the behaviour of the test statistic is well understood, as the F-statistic in ANOVA, one may adjust the statistic according to the ν_0/ν ratio and recompute the p-value. However, this will require deep insight into the statistical behaviour of the model. Down-sampling to the effective samples size, i.e. to a ν/ν_0 portion of the samples, may also be done to see if results persist. However, none of these solutions are ideal.

We did consider if batch adjustment could be modified so as to balance the need for removing batch effects with avoidance of introducing false group differences which would be somewhere between batch centering and two-way ANOVA based adjustment. For two-group comparisons, the batch centering will remove some of the group effects and thus underestimate group differences, while two-way ANOVA based batch adjustment will add random, but unbiased, group effects. Somewhere in between is a balance for which these two balance out. However, the applicability of this approach is very limited and potentially risky since the degree of moderation of the batch adjustment will depend on the subsequent comparison to be made.

Alternatively, one may treat the results more like an ordered list of candidates, with the most likely true positives on top, de-emphasizing the somewhat deflated p-values. This would, however, be hypothesis generating, rather than hypothesis testing. While investigators should always assess the extent to which findings make biological or clinical sense, this is particularly true when the statistical assessment may be unreliable.

Finally, we would like to emphasise the importance of proper description of how data has been prepared for analysis, and what corrections and adjustments have been made (Sandve *and others*, 2013). In cases where data preparation is performed prior to, and separate from, the data analysis, as is now often the case, this is of particular importance, as artifacts may be introduced in the data preparation which could influence the reliability of downstream analyses. When assessing published findings, we frequently found it hard to determine how batch adjustments had been done, and occasionally suspected that study group had been used as covariate without this being stated. Science not only depends on determining sound methods and finding reliable results, but also on communicating sufficient detail to allow readers to assess the extent to which findings can be trusted. Failing to do so may lead cause methodological problems to pass undetected, but also to cast doubt on sound results. An increased focus on making research reproducible is therefore of great importance.

APPENDIX

A. Distribution of error terms after two-way ANOVA batch adjustment

We assume a general linear model with intercept or more general covariates (α), batch effects (β), and study groups or study parameters of interest (γ):

$$Y = \alpha A + \beta B + \gamma C + \epsilon \tag{A.1}$$

where $Y = [Y_1, \ldots, Y_n]$ are the observable random variables, α, β, and γ is a p, q, and r vectors with design matrices A, B, and C, and $\epsilon = [\epsilon_1, \ldots, \epsilon_n] \sim \mathrm{N}(0, \sigma^2 I_n)$ where I_n is the $n \times n$ identity matrix. We also assume the full design matrix, combining A, B, and C, has full rank $p + q + r$: i.e., the covariates are linearly independent.

In general, we wish to assess the explanatory power of B including C as batch effects and A as additional covariates. Often $p = 1$ with α the intercept term, but this formulation allows general covariates to be included in the analyses. The covariates of B which are subject to testing are more commonly specified as contrasts on the full design matrix combining A, B, and C, but in our case this splitting is more convenient.

We can eliminate A from the model, simultaneously removing p degrees of freedom. One way of doing this is to find an $n \times n$ rotation matrix U so that $AU = [0 \mid A']$ with A' a $p \times p$ invertible matrix. Applying this rotation to all elements, we split the n dimensions into $(n-p)+p$: $YU = [Y_0 \mid Y']$, $BU = [B_0 \mid B']$, $BU = [C_0 \mid C']$, and $\epsilon V = [\epsilon_0 \mid \epsilon'] \sim \mathrm{N}(0, \sigma^2 I_n)$. We may then eliminate the degrees of freedom used to estimate α from the model yielding

$$Y_0 = \beta_0 B_0 + \gamma_0 C_0 + \epsilon_0 \tag{A.2}$$

where Y_0 and ϵ_0 are now $n-p$ vectors. If α is the intercept term, this corresponds to centering all variables and covariates, but the rotation also eliminates the corresponding degree of freedom.

If there are no batches, i.e. $r = 0$, we get the common linear model in which we get a decomposition $Y = \hat{\beta}_0 B_0 + \hat{\epsilon}_0$ with $\hat{\gamma}_0 = Y_0 B_0^t (B_0 B_0^t)^{-1}$, where $|\hat{\epsilon}|^2 \sim \sigma^2 \chi_{n-p-q}$ and $|(\hat{\beta}_0 - \beta_0) C_0|^2 \sim \sigma^2 \chi_q$, yielding the common F-statistic

$$F = \frac{|(\hat{\beta}_0 - \beta_0) C_0|^2 / q}{|\hat{\epsilon}_0|^2 / (n - p - q)} \sim \mathrm{F}_{q, n-p-q}. \tag{A.3}$$

If we believe our data are "batch effect free", this is the model we would use, and the distribution of the F-statistic we would assume.

We define the batch adjusted data $\tilde{Y}_0 = Y_0 - \hat{\gamma}_0 C_0$ using the estimated batch effect $\hat{\gamma}_0$ from the full model. If we let $X_0 = [B_0^t \mid C_0^t]^t$ be the full design matrix after elimination of covariates A, the parameter estimates are $[\hat{\beta}_0 \mid \hat{\gamma}_0] = Y_0 X_0^t (X_0 X_0^t)^{-1}$. The linear decomposition $Y_0 = \hat{\beta}_0 B_0 + \hat{\gamma}_0 C_0 + \hat{\epsilon}_0$ has $|\hat{\epsilon}_0|^2 \sim \sigma^2 \chi^2_{n-p-q-r}$ and $|(\hat{\beta}_0 - \beta_0) B_0 + (\hat{\gamma}_0 - \gamma_0) C_0|^2 \sim \sigma^2 \chi^2_{q+r}$, and so the variance of the error term is the same in the batch adjusted case $\tilde{Y}_0 = \hat{\beta}_0 C_0 + \hat{\epsilon}_0$. However, the sum of squares $|(\hat{\beta}_0 - \beta_0) B_0|^2$ explained by B is less easily expressed.

Solving $[\hat{\beta}_0 \mid \hat{\gamma}_0] = Y_0 X_0^t (X_0 X_0^t)^{-1}$ and using

$$(X_0 X_0^t)^{-1} \begin{pmatrix} B_0 \\ 0 \end{pmatrix} = \begin{pmatrix} P \\ -(C_0 C_0^t)^{-1} C_0 B_0^t P \end{pmatrix} B_0 \tag{A.4}$$

with $P = (B_0 \rho_{C_0} B_0^t)^{-1}$ and $\rho_{C_0} = I - C_0 (C_0 C_0^t)^{-1} C_0$ yields

$$(\hat{\beta}_0 - \beta_0) B_0 = \epsilon_0 \rho_{C_0} B_0^t P B_0 \sim \mathrm{N}(0, \sigma^2 L_0) \quad \text{with} \quad L_0 = B_0^t P B_0 \rho_{C_0} B_0^t P B_0 \tag{A.5}$$

since $\epsilon_0 \sim \mathrm{N}(0, \sigma^2 I)$. Let $\lambda_1, \ldots, \lambda_q \geqslant 1$ be the non-zero eigenvalues of L_0: this holds since L_0 is a symmetric rank q matrix *complete the proof of this*. Then,

$$|(\hat{\beta}_0 - \beta_0)B_0|^2 = \sigma^2 \sum_{i=1}^{q} \lambda_i U_i \quad \text{where} \quad U_1, \ldots, U_q \sim \chi^2_1. \tag{A.6}$$

Approximating this so that the first two momenta match gives

$$|(\hat{\beta}_0 - \beta_0)B_0|^2 \approx \tilde{\sigma}^2 \chi^2_{\tilde{q}} \quad \text{where} \quad \frac{\tilde{q}\tilde{\sigma}^2}{\sigma^2} = \sum_{i=1}^{q} \lambda_i = \mathrm{tr}(L_0), \; \frac{\tilde{q}\tilde{\sigma}^4}{\sigma^4} = \sum_{i=1}^{q} \lambda_i^2 = \mathrm{tr}(L_0^2), \tag{A.7}$$

from which it follows that $\tilde{\sigma}^2 \geqslant \sigma^2$, $\tilde{q}\tilde{\sigma}^2 \geqslant q\sigma^2$, and $\tilde{q} \leqslant q$, where equality holds if and only if all $\lambda_i = 1$, which happens when B_0 and C_0 are linearly independent, i.e. $B_0 C_0^t = 0$.

We would like to express $\tilde{\sigma}^2$ and $\tilde{q}$ terms of the original design matrices. The rotated design matrices were defined as $B_0 = BU_0$, etc., using the rotation $U = [U_0 \mid U']$ so that $AU_0 = 0$ while AU' is non-singular. Hence, it follows that $V_0 V_0^t = \rho_A = 1 - A^t(AA^t)^{-1}A$ which is the projection onto the $n - p$-hyperplane perpendicular to the p-hyperplane spanned by A. This allows us to express

$$L_0 = U_0^t L U_0 \quad \text{where} \quad L = \tag{A.8}$$

Need to complete this part...

Reproducible research

The data and scripts used to generate the results in this work are available at https://github.com/ous-uio-bioinfo-core/batch-adjust-warning-figures.git. Additional analyses, performed, but not included in the article, may be found at the extended repository https://github.com/ous-uio-bioinfo-core/batch-adjust-warning-reports.git.

Be aware that the different versions of the r-packages could produce different results. In particular, the newest version of `limma` (3.20.8) was needed to produce the plot in Figure 1d.

Acknowledgments

We are grateful to Geir Kjetil Sandve for useful suggestions and discussions.

This work was supported by the EUROCAN platform (VN) and the MetAction project (EAR).

Conflict of Interest: None declared.

References

Bakshi, S., Chalifa-Caspi, V., Plaschkes, I., Perevozkin, I., Gurevich, M. and Schwartz, R. (2013, April). Gene expression analysis reveals functional pathways of glatiramer acetate activation. *Expert opinion on therapeutic targets* **17**(4), 351–62.

Buhule, O. D., Minster, R. L., Hawley, N. L., Medvedovic, M., Sun, G., Viali, S., Deka, R., McGarvey, S. T. and Weeks, D. E. (2014, October). Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Frontiers in Genetics* **5**(October), 1–11.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L. and Liu, C. (2011, January). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one* **6**(2), e17238.

Emig, D., Salomonis, N., Baumbach, J., Lengauer, T., Conklin, B. R. and Albrecht, M. (2010, July). AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic acids research* **38**(Web Server issue), W755–62.

Fernández-Albert, F., Llorach, R., Garcia-Aloy, M., Ziyatdinov, A., Andres-Lacueva, C. and Perera, A. (2014, October). Intensity drift removal in LC/MS metabolomics by common variance compensation. *Bioinformatics (Oxford, England)* **30**(20), 2899–905.

Giordan, M. (2013, February). A Two-Stage Procedure for the Removal of Batch Effects in Microarray Studies. *Statistics in Biosciences* **6**(1), 73–84.

Johnson, W. E., Li, C. and Rabinovic, A. (2007, January). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1), 118–27.

Kitchen, R. R., Sabine, V. S., Simen, A. a., Dixon, J. M., Bartlett, J. M. S. and Sims, A. H. (2011, January). Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. *BMC genomics* **12**(1), 589.

Kupfer, P., Guthke, R., Pohlers, D., Huber, R., Koczan, D. and Kinne, R. W. (2012, January). Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC medical genomics* **5**(1), 23.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. and Storey, J. D. (2012, March). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**(6), 882–3.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. a. (2010, October). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* **11**(10), 733–9.

Luo, J., Schumacher, M., Scherer, a., Sanoudou, D., Megherbi, D., Davison, T., Shi, T., Tong, W., Shi, L., Hong, H., Zhao, C., Elloumi, F., Shi, W., Thomas, R., Lin, S., Tillinghast, G., Liu, G., Zhou, Y., Herman, D., Li, Y., Deng, Y., Fang, H., Bushel, P., Woods, M. *and others.* (2010, August). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The pharmacogenomics journal* **10**(4), 278–91.

Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K. and Beck, S. (2014, March). ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics (Oxford, England)* **30**(3), 428–30.

Piccolo, S. R., Withers, M. R., Francis, O. E., Bild, A. H. and Johnson, W. E. (2013, October). Multiplatform single-sample estimates of transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America* **110**(44), 17778–83.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J. P. (2006, May). GenePattern 2.0. *Nature genetics* **38**(5), 500–1.

Sandve, G. K., Nekrutenko, A., Taylor, J. and Hovig, E. (2013, October). Ten simple rules for reproducible computational research. *PLoS computational biology* **9**(10), e1003285.

Smyth, G. K. and Speed, T. (2003, December). Normalization of cDNA microarray data. *Methods* **31**(4), 265–273.

Taminau, J., Meganck, S., Lazar, C., Steenhoff, D., Coletta, A., Molter, C., Duque, R., de Schaetzen, V., Weiss Solís, D. Y., Bersini, H. *and others*. (2012, January). Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. *BMC bioinformatics* **13**, 335.

Towfic, F., Funt, J. M., Fowler, K. D., Bakshi, S., Blaugrund, E., Artyomov, M. N., Hayden, M. R., Ladkani, D., Schwartz, R. and Zeskind, B. (2014, January). Comparing the biological impact of glatiramer acetate with the biological impact of a generic. *PloS one* **9**(1), e83757.

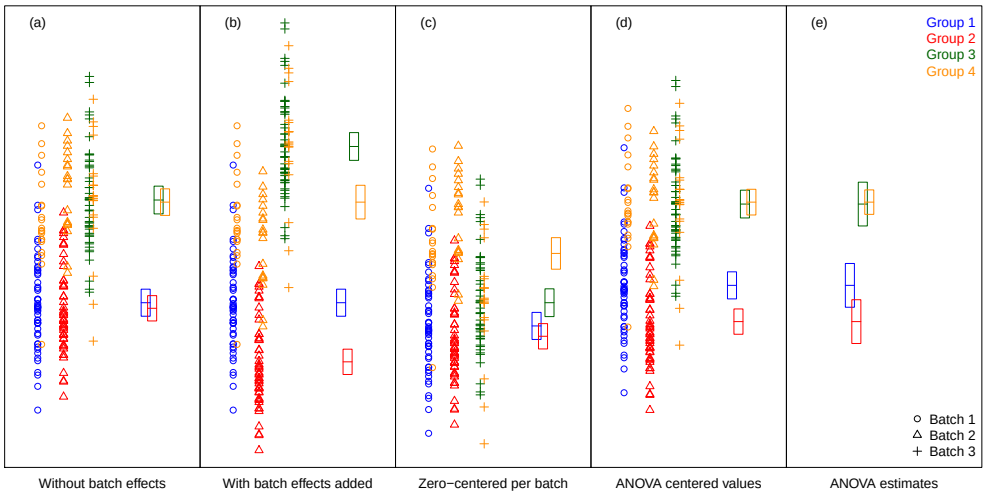

Figure 1. We simulated expression of one gene from four study groups unevenly distributed in three batches containing 50+0+0+20, 0+50+0+20 and 0+0+50+20 samples from groups 1 to 4. This case, design and effects, was selected to illustrate the spurious effects that may arise from different batch adjustments. The Y-axis represent the expression values, while the X-axis is used to visually separate the batches. Circles, triangles, and crosses indicate values from each of the three batches, with colours indicating study groups. Correspondingly coloured boxes to the right of the measurements show group means with 95% confidence intervals. These give some indication as to which group differences would be found significant. a) The "true" values measured in a system without batch effects. Groups 1 and 2 have lower means than groups 3 and 4. b) The same samples as in a) but measured in a system with batch effects. All the groups seem different. c) The measurements in b) are adjusted with batch centering. Group 3 and 4 seem to differ while group 1 and 3 are more similar. d) The measurements in b) are adjusted with two-way ANOVA based batch centering (using `limma`). Group 1 and 2 seem to differ. e) The least squares estimates of the group means from a two-way ANOVA have the same means as in d), but more appropriate confidence intervals.

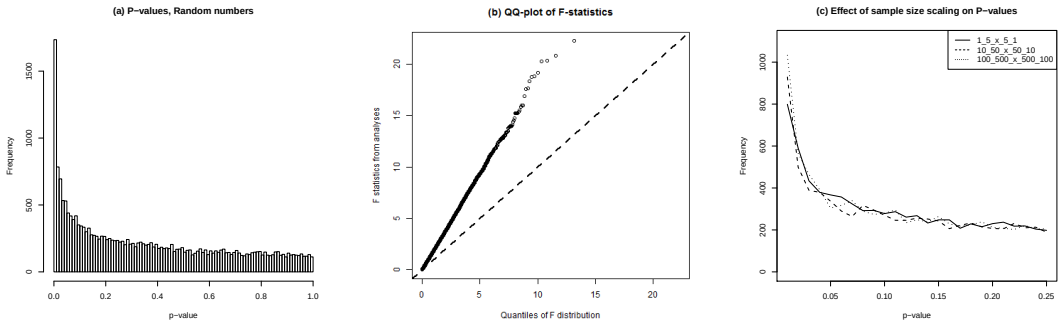

Figure 2. This is a sanity check where the recommended use of ComBat fails, adapted from the user guide in the `sva` package. Real data are substituted with random numbers from a normal distribution, but the batch–group design is retained. ComBat is applied, followed by an F-test. a) P-value distribution b) QQ plot of the F-statistics c) P-value distributions for 3 equally unbalanced random number experiments with different sample sizes, 12,120 and 1200 samples from two study groups with a 1:5 and 5:1 distribution in two batches. A random batch effect is added for 10% of the 20000 genes. This example is not from the `sva` package.

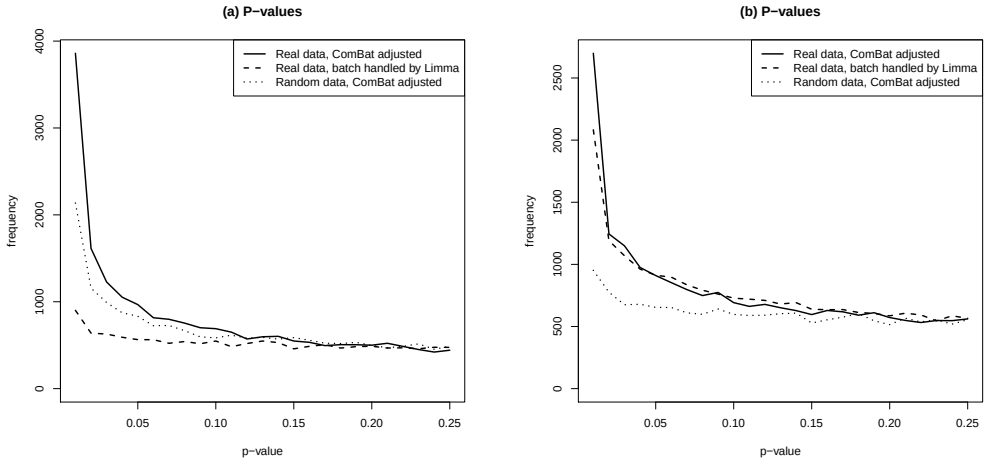

Figure 3. Three analyses of two published data sets where batch effects were adjusted for with ComBat. First, analysed as described on the real data using ComBat. Secondly, with ComBat, but with random numbers instead of real data. Thirdly, instead of ComBat, analyses of the real data with `limma` blocking by batch. a) Re-analysis of Towfic *and others* (2014), glatiramer acetate vs. generic b) Re-analysis of "Data set 2" Johnson *and others* (2007), TAL1 inhibition vs. control

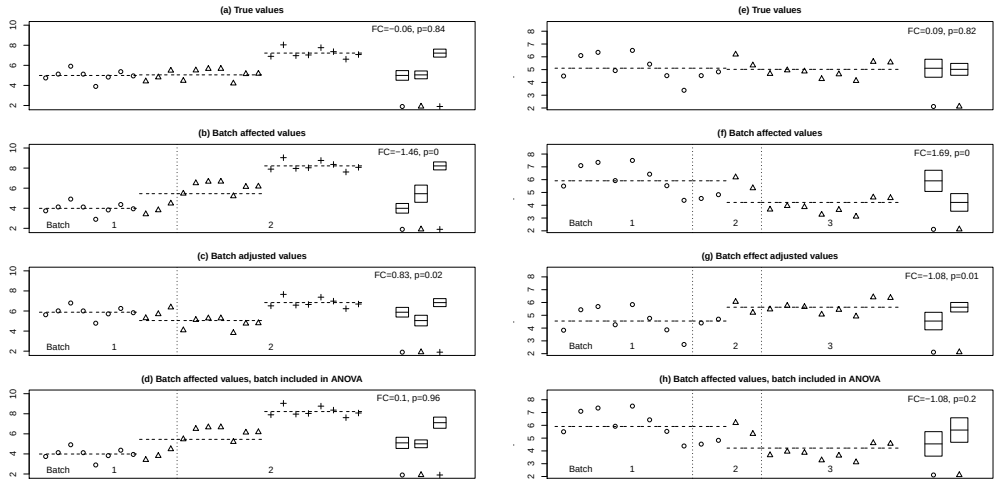

Figure 4. We simulated expression of one gene from two and three study groups. These two(a-d and e-h) cases, designs and effects, were selected to illustrate the spurious effects that may arise from different batch adjustments. The Y-axis represents the log expression values. while the X-axis is used to separate the samples. Circles, triangles, and crosses indicate sample measurements from separate groups. Horizontal dashed lines indicates the mean of the expression values for that groups samples. Vertical dotted lines separates batches with the batch numbers on the lower half of the plot. To the far right of the plots, a typical differential expression test is performed with estimated fold change and a p-value. The mean expression for each group and 95% confidence interval is estimated and depicted as boxes. a) Simulated expression of one gene in 26 samples from 3 groups (circles, triangles and crosses). The fold change and p-value presented are from the circles vs. triangles comparison and calculated with a one-way ANOVA. The confidence intervals indicate that the crosses are different from the other two groups. b) The values from (a) is divided into 2 batches with batch 1 containing circles and triangles, and batch 2 containing triangles and crosses. Batch effects were added, -1 for batch 1 and +1 for batch 2. The ANOVA test now results in a p¡.01 between the circles and triangles. This plot illustrates the problem of disregarding batch effects. c) The values in (b) were batch adjusted using the mean adjustment method. In order to make the means of the two batches equal, some of the group-effect for the crosses were transferred over to the triangels within the same batch. As a result the ANOVA test now indicate a difference between the circle-group and the triangle-group (p¡.01). d) The same values as in (b) are shown and used in a two-way ANOVA with group and batch included in the model. The estimated fold change and corresponding p-value is in effect calculated from the samples in batch 1. e) Simulated expression of one gene in 20 samples from two groups (circles and triangles). The t-test indicates no significant group differences(p=.82). f) The values from (e) are divided into 3 batches in an unbalanced way, with only batch 2 containing samples from both groups. Batch effects were added, +1 for batch 1 and -1 for batch 3. The t-test now results in p¡.01. This plot illustrates the problem with disregarding batch effects. g) The values in (f) were batch effect adjusted while retaining the group difference (using removeBatchEffects), thus creating a “batch effect free” data set. Notice that all the values in the ”circle only” batch 1 are adjusted to make the group mean coincide with the mean of the two circle-samples in batch 2. The triangle-samples in batch 3 are adjusted similarly to match the mean of the two triangle-sampels in batch 2.The t-test now results in a significant group difference (p=.01) and a fold change equal to a fold change calculated from the batch 2 samples only. Thus, the fold change is in effect estimated from only 2+2 samples, but the t-test assumes 10+10 samples with the low p-value as a result. h) The same values as in (f) are shown and used in a two-way ANOVA with group and batch included in the model. The estimated fold change is again in effect only based on samples from batch 2 and is exactly the same as for (g), but the p-value is much higher and the confidence intervals are more appropriate.