Biostatistics (2014), **0**, 0, *pp.* 1–11
doi:10.1093/biostatistics/diff

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

VEGARD NYGAARD, EINAR ANDREAS RØDLAND, EIVIND HOVIG*

Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital HF - Radiumhospitalet, Montebello, 0310 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, estimated from a two-way ANOVA model, to avoid conflating batch effects and group differences during batch adjustments. 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 this problem, which most likely has contributed to false discoveries being presented in the published literature.

Key words:

1. Introduction

Extraneous variables, if left unaccounted for, have the potential to lead an investigator into drawing wrong conclusions. ~~In molecular biology, extraneous variables are often called "batch effects", probably due to the fact that reagentsand other equipment, for instance~~ A common example is "batch effects" caused by reagents, microarray chips, and other equipment are made in batches, ~~and this is frequently observed as an effect in~~ 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 comparing ~~group differences~~ differences between study groups, the presence of batch effects will decrease the statistical power since it adds variation to the data. If

†Manuscript version ~~June 18, 2014.~~June 26, 2014.

*To whom correspondence should be addressed.

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. ~~In the next step~~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.

Er de neste tre avsnittene nødvendige eller kan de forkortes drastisk siden dette forklares grundigere siden?

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 1 illustrates both 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. ~~If~~ However, if the group–batch design is heavily unbalanced, estimation of group differences and batch effects are interdependent~~. However, in~~ , and when applying the estimated group differences across the entire data set, the uncertainties of these estimates are ignored. ~~If the group–batch design is unbalanced and batch compositions ignored after batch adjustments have been made, later analyses will~~ Subsequent analyses will then systematically underestimate the statistical uncertainties and exaggerate the confidence of group differences. Figure 1 illustrates how statistical uncertainties are deflated by this batch adjustment method by comparing them to the uncertainties from the original ANOVA.

The ComBat method described in Johnson *and others* (2007), and included in the sva package (Leek *and others*, 2012), can use either of the two above described approaches to estimate batch differences~~, but~~. In addition, it uses an empirical Bayes approach to avoid over-correction for batch effects, which is critical for small batches. ~~It has thus improved and popularised the two-way ANOVA procedure for retaining group differences when adjusting for batch effects, and~~ Increasingly, the inclusion of group difference as a covariate when removing batch effects has been recommended, both in the sva tutorial and user fora, and ComBat has thus help popularise this approach. Based on actual use of ComBat by the authors and others, ourselves included, we suspect thus 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.

Statisticians would most likely take extra precautions should the sample or batch sizes be very small. However, the effects of batch adjustment using two-way ANOVA or ComBat on unbalanced

data sets remain even as sample and batch sizes increase. 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.

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 ~~$r = 1, \ldots, n_{ij}$ are the different samples~~ $\epsilon_{ijr} \sim N(0, \sigma^2)$ 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. One such model, used by 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. $\delta_{ig} = 1$. ~~We will discuss the effect of empirical Bayes estimation of~~ Empirical Bayes will tend to shrink the estimates of γ_{ig} ~~later, but our main argument is more easily made in the simpler context. Empirical Bayes may reduce the~~, 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.

The most common method for removing batch effects is to zero-centre each batch:

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

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. $\widetilde{Y}_{ijr} = \Delta\widetilde{Y}_{ijr} + \bar{Y}$. When comparing groups, the common value $\bar{Y}$ has no effect, and so this is equilvalent 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:

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

Thus, batch centering will tend to reduce group differences in an unbalanced design, and ~~thus~~ reduce the power of downstream analyses. By reducing the differences between some groups, i.e. those found together in the same batches, it may also induce false differences between other groups.

Removing batch effects while retaining group differences can be done 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, $\hat{\gamma}_i$:

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

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:

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

In a balanced group–batch design, ~~this has the same effect on~~ the effect is the same all groups and thus does not influence group comparisons. In an unbalanced design, however, the ~~estimated batch effects $\hat{\gamma}_i$ will correlate with the estimated group effects $\hat{\beta}_j$.~~ estimation errors $\hat{\gamma}_i - \gamma_i$ will induce systematic differences between groups which, when ignored in downstream analyses, may lead to over-confidence in estimated group differences.

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 a runnable example demonstrating how to adjust a data set with ComBat followed by ~~a~~ an F-test. ~~Swapping~~ Replacing the real data with random numbers from a standard normal distribution (mean=0, sd=1), but otherwise following the instructions, will generate the p-value distribution shown in Figure ~~??. The skewed distribution is a indication that this approach may have a unintentional adverse effect.~~ 2.

As can be seen from the Q-Q plot, 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 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 ~~random data with different means in each batch , but no difference betwene groups, are drawn~~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 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 containing n_{iA} and n_{iB} samples from each of the two groups for $i = 1, \ldots, m$.

If we estimate the group difference within batch i, we get

$$\widehat{\Delta\beta}_i = \bar{Y}_{iA} - \bar{Y}_{iB} \sim \mathrm{N}\left(\Delta\beta, \frac{\sigma^2}{\nu_i}\right) \quad \text{where} \quad \bar{Y}_{ij} = \frac{\sum_{r=1}^{n_{ij}} Y_{ijr}}{n_{ij}},\ \nu_i = \frac{1}{\frac{1}{n_{iA}} + \frac{1}{n_{iB}}}, \tag{3.7}$$

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

$$\widehat{\Delta\beta} = \frac{\sum_{i=1}^{m} \nu_i \widehat{\Delta\beta}_i}{\nu} \sim \mathrm{N}\left(\Delta\beta, \frac{\sigma^2}{\nu}\right) \quad \text{where} \quad \nu = \sum_{i=1}^{m} \nu_i. \tag{3.8}$$

If batch and group effects are estimated using a two-way ANOVA, the estimate $\widehat{\Delta\beta}$ 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 $\widehat{\Delta\beta}$ unchanged by the batch adjustment. However, if this batch adjusted data set is analysed without considering batch effects, the variance of $\widehat{\Delta\beta}$ 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

$$\widehat{\Delta\beta} \sim \mathrm{N}\left(\Delta\beta, \sigma^2 \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\right). \tag{3.9}$$

Using Jensen's inequality, we may then derive

$$\frac{1}{\frac{1}{n_A} + \frac{1}{n_B}} = n \cdot \frac{n_A}{n} \cdot \frac{n_B}{n} \leqslant \nu = \sum_{i=1}^{m} \frac{1}{\frac{1}{n_{iA}} + \frac{1}{n_{iB}}} = \sum_{i=1}^{m} n_i \cdot \frac{n_{iA}}{n_i} \cdot \frac{n_{iB}}{n_i} \tag{3.10}$$

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

3.3 *Examples of undesired consequences*

As the amount of false positive results when trying to retain group differences depends on the batch–group balance, we will show two examples with varying degree of unbalancedness.

3.3.1 *Experiment 1*

~~In the first experiment (~~An experiment by Towfic *and others*, 2014 ~~), cells were treated with~~ treated cells with either glatiramer acetate (a medicine for multiple sclerosis) or a generic drug, and mRNA was measured using microarrays alongside control samples. A batch effect correlating to the chip (Illumina WG-6_V2, six samples per chip) was observed and adjusted

for with ComBat, whereafter the data was tested for differentially expressed genes, yielding a list of 1000 genes (Table S5, Towfic *and others*, 2014). Unfortunately the batch–treatment design was unbalanced with several batches having only one of the main treatments of interest. When we re-analyzed their data without using ComBat, but instead blocked for batch effect in limma, only 9 genes were found (FDR < 0.05). The above outlined sanity check with random numbers was also carried out. The distribution of p-values for different settings are shown in Figure ~~??~~3a. Our conclusion is that most of the genes reported as differentially expressed in (Towfic *and others*, 2014) are false positives. This example is a sort of "worst case" scenario for applying ComBat, since it both has a very unbalanced batch–group design and a a priori assumption of no difference. The R-code for our analysis and a more extensive report can be downloaded from Github *[Update reference]*.

3.3.2 *Experiment 2* The ~~second example is taken from the supporting information for~~ supporting information of the original ComBat article (Johnson *and others*, 2007) ~~where it is denoted "Data set 2". Cells~~ demonstrates the method on cells inhibited for the expression of the TAL1 gene were compared to controls on a microarray platform (denoted "Data set 2"). The experiment was conducted on three different time points (used as batches) with a total of 30 samples and a fairly balanced batch–treatment set up (6-2, 3-4 and 9-6). 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 does not indicate a huge difference. Nevertheless, ~~we belive that~~ the P-values ~~are deflated for~~ may still be somewhat deflated in the ComBat adjusted analysis. The R-code for our analysis can be downloaded from Github *[Update reference]*.

3.3.3 ~~*More?*~~

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 uncertainty estimates and over-confidence in the results. The problem is essentially independent of sample size, rather than being particular to small samples.

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 cases, e.g. a large set of gene, followed by multiple testing corrections such as false discovery rate.

4.1 *Increased emphasis on preserving group difference*

In the original ComBat article (Johnson *and others*, 2007)~~it is clear that the primary motivation behind ComBat~~ , the main objective was to employ an ~~Emprical Bayes method for batch-effect removal~~ emprical Bayes method to allow better handling of small batches. The ~~feature of~~ inclusion of covariates, e.g. for retaining group differences for unbalanced designs~~is optional and seems to be~~ , 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. ~~When ComBat was later~~ Later, when ComBat was incorporated in 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.

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 on our group differences were exaggerated, ~~the literature was searched~~ we searched the literature for a better understanding of correct use and potential ~~overseen~~ limitations of ComBat. ~~But~~ However, the authors of ComBat and the sva package recommended ~~our usage (,)~~including study group as a covariate as we had done. In addition, other ~~works looking into the problem of~~ studies investigating batch effects were mostly recommending ComBat without much concern (Kupfer *and others*, 2012, Kitchen *and others*, 2011). A brief inquiry into some of the articles citing ComBat (574 on Google Scholar) revealed few ~~problems, and~~ reported problems, although their method descriptions regarding ComBat were mostly sparse, limited to one or two sentences. A further indication of ~~their~~ the carefree use of ~~this potentially devastating~~ the procedure was the frequent omittance of program parameters that were used, i.e batch labels or whether group labels were supplied as covariates. Often no effort was done in order to substantiate the existence of batch effects ~~in their data, except for~~ beyond stating the presence of batches. The incorporation of the method into analysis pipelines ([16642009], TCGA) and other packages ([23452776], [21937664]) could make its ~~usage even more trivial and parameters setting harder to perceive. Taken together we fear~~ even more accessible, but its usage less transparent. We are concerned that many published results from data adjusted by ComBat ~~are completely or partially false~~with study group as covariate may be unreliable in light of our findings, and furthermore that the frequent lack of proper method description accompanying published results make it hard to judge they are affected or not. We hope that our warning will ~~enlighten the community and reduce this unfortunate combination of methods~~help caution the scientific community against this particular approach.

4.3 *Practical advice*

For an investigator facing an unbalanced data set with batch effects, our advice is to adjust for batch inside the statistical test and avoid the two step procedure outlined above. If this is not possible, then applying the method described in this article should only be done with great caution and downstream confidence estimates must be treated with suspicion.

Knowing that adjusting for batch effects while preserving the group difference may lead to varying degree of false results, to what extent can an investigator trust a result from a work

applying such a method? 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, this thoroughness is not feasible if the downstream analysis can not adjust for batch effects by it self. 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 bioinformatic skills and work hours could still be in short supply. For such situations, a superficial assessment can be performed, taking special note of batches were groups of interest are near missing and how likely a group difference is. In essence asking if the balanced parts (effective sample size?) of the data has enough power to detect the presumed effects and if this is the case (Johnson *and others*, 2007, [24584070]), treat the results more like an ordered list with the most likely true positives on top while de-emphasizing the somewhat deflated p-values . In contrast, if biological knowledge suggest that a group effect is unlikely ([24391845], [18414638]?, [21731603], [23630272]), an intermediate lack of batch-group balance could lead to a mostly false result.

5. Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

5.1 *Reproducible research*

The data and scripts used to generate the results in this work, i.e figures, are made available at https://github.com/ous-uio-bioinfo-core/batch-adjust-warning-figures.git. Additional analyses were also performed, but considered excessive for the current article. However, some readers might find this very interesting and we provide it as an extra resource at https://github.com/ous-uio-bioinfo-core/batch-adjust-warning-reports.git

[vi boer klare aa tilfredstille kravene i http://biostatistics.oxfordjournals.org/content/10/3/405.full] [For aa tilfredstille kravene skal data og script vare available. Jeg mener vi klarer dette. paa githuben er ikke data fra towfic, men det blir lastet av scriptet fra GEO. Ikke beskrevet her men i scriptet. Videre i forklaringen til biostatistics staar det at script og data skal bli lagt paa deres sider, som vi kan hvis de vil]

Acknowledgments

[...Acknowledgements...] *Conflict of Interest*: None declared.

References

Johnson, W Evan, Li, Cheng and Rabinovic, Ariel. (2007, January). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* **8**(1), 118–27.

Kitchen, Robert R, Sabine, Vicky S, Simen, Arthur a, Dixon, J Michael, Bartlett, John M S and Sims, Andrew H. (2011, January). Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. *BMC ge-*

nomics **12**(1), 589.

Kupfer, Peter, Guthke, Reinhard, Pohlers, Dirk, Huber, Rene, Koczan, Dirk and Kinne, Raimund 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, Jeffrey T, Johnson, W Evan, Parker, Hilary S, Jaffe, Andrew E and Storey, John D. (2012, March). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)* **28**(6), 882–3.

Leek, Jeffrey T, Scharpf, Robert B, Bravo, Héctor Corrada, Simcha, David, Langmead, Benjamin, Johnson, W Evan, Geman, Donald, Baggerly, Keith and Irizarry, Rafael 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.

Towfic, Fadi, Funt, Jason M, Fowler, Kevin D, Bakshi, Shlomo, Blaugrund, Eran, Artyomov, Maxim N, Hayden, Michael R, Ladkani, David, Schwartz, Rivka and Zeskind, Benjamin. (2014, January). Comparing the biological impact of glatiramer acetate with the biological impact of a generic. *PloS one* **9**(1), e83757.

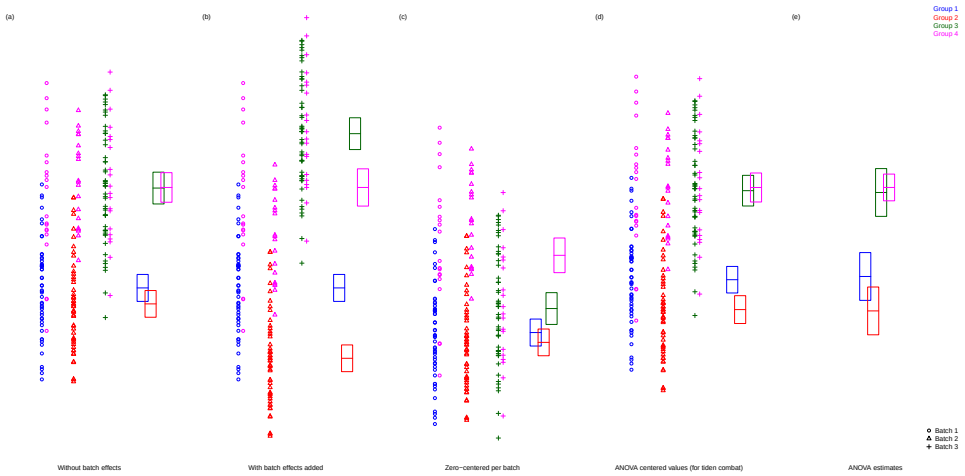

Figure 1. Simulated data from four study groups was generated where groups 1 and 2 have lower means than groups 3 and 4. These were placed in three different batches with batch effect added. Values and boxes showing mean and two standard errors of the mean are displayed for data without batch effects, after adding batch effects, after batch centering, and after ANOVA based batch centering. The last frame shows the least squares estimates of the group means from a two-way ANOVA analysis with 2 standard errors. This case, design and effects, was selected to illustrate the spurious effects that may arise.

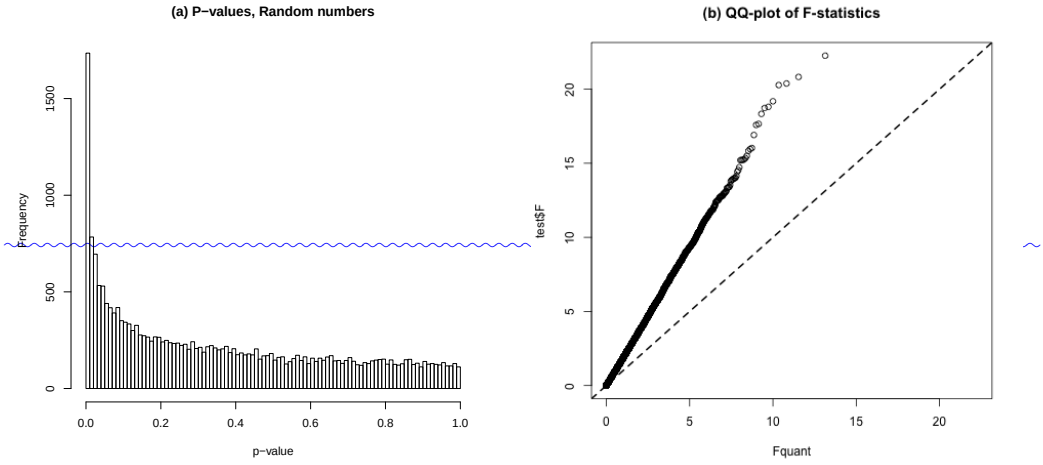

Figure 2. A sanity checks where the recommended use of ComBat fails. Adapted from the user guide in the sva package. Real data are substituted with random numbers from a normal distribution, but the batch-group design is retained. ComBat is applied followed by a F-test. a) P-value distribution b) QQ plot of the F-statistics

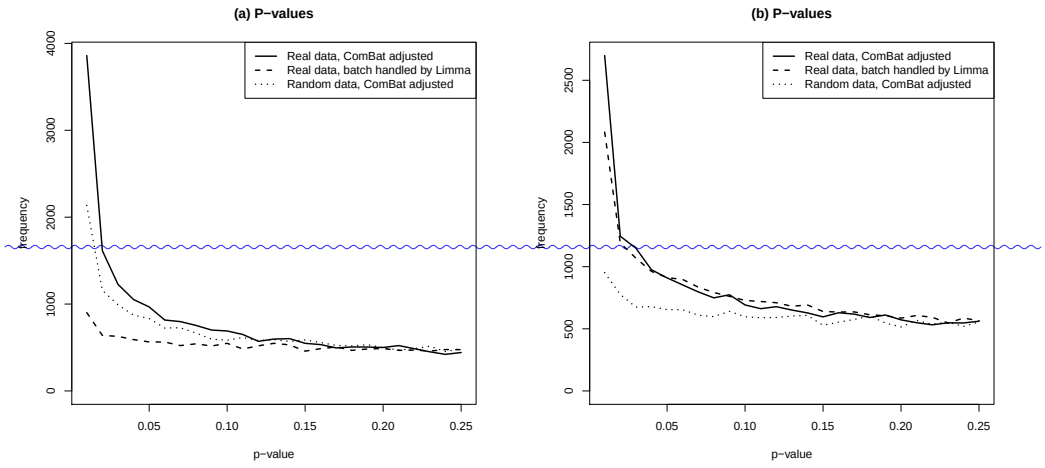

Figure 3. Three analyses of two works where batch effects were adjusted for with ComBat. First as described in the original works with ComBat on real data. Secondly, with ComBat but with random numbers instead of real data. Thirdly without ComBat and instead blocking batch with limma on real data. a) 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