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, 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. 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, the uncertainties of these estimates are ignored. Subsequent analyses will then systematically underestimate the statistical uncertainties and exaggerate the

confidence of group differences. Figure 1d illustrates how statistical uncertainties are deflated by this batch adjustment method by comparing them to the uncertainties 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 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 R package `limma` (Smyth and Speed, 2003) also offers batch adjustment with covariates, but notes that this is not intended used prior to linear modeling.

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.

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. 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. $\delta_{ig} = 1$. 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.

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

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

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. $\tilde{Y}_{ijr}^{\text{avg}} = \tilde{Y}_{ijr}^0 + \bar{Y}$. 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:

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

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.

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, $\hat{\gamma}_i$:

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

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:

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

so that

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

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 estimation errors.

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 (mean=0, sd=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.

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

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

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

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

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

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

Using Jensen's inequality, we may derive

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

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 $\nu_0/\nu \geqslant 1$ indicates to what extent the uncertainty of the estimated group difference is deflated by the two-step procedure.

3.3 *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.

3.3.1 *Experiment 1* In an experiment by Towfic *and others* (2014), cells were treated with either glatiramer acetate (a medicine for multiple sclerosis) or a generic drug, and mRNA was measured using microarrays alongside control sample measurements. 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 were 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 the use of ComBat, but instead blocked for batch effect in `limma`, only 9 genes were found (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. 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 an a priori assumption of no difference.

3.3.2 *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 was conducted on three different time points (used as batches) with a total of 30 samples and a fairly balanced batch–treatment set up containing 6+2, 3+4 and 9+6 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 uncertainty estimates and over-confidence in the results. The problem is essentially independent of sample size.

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*

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. 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.

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), which offers ComBat with covariates, as well as TCGA and inSilicoDb (Taminau *and others*, 2011), presently without covariates as an option.

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*

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. 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. However, the two-group assessment should still give a fair idea of the reliability of pairwise group comparisons. 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.

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.

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 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.

References

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.

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.

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., STEENHOFF, D., COLETTA, A., MEGANCK, S., LAZAR, C., DE SCHAETZEN, V., DUQUE, R., MOLTER, C., BERSINI, H., NOWÉ, A. *and others.* (2011, November). in-

SilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. *Bioinformatics* **27**(22), 3204–5.

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.
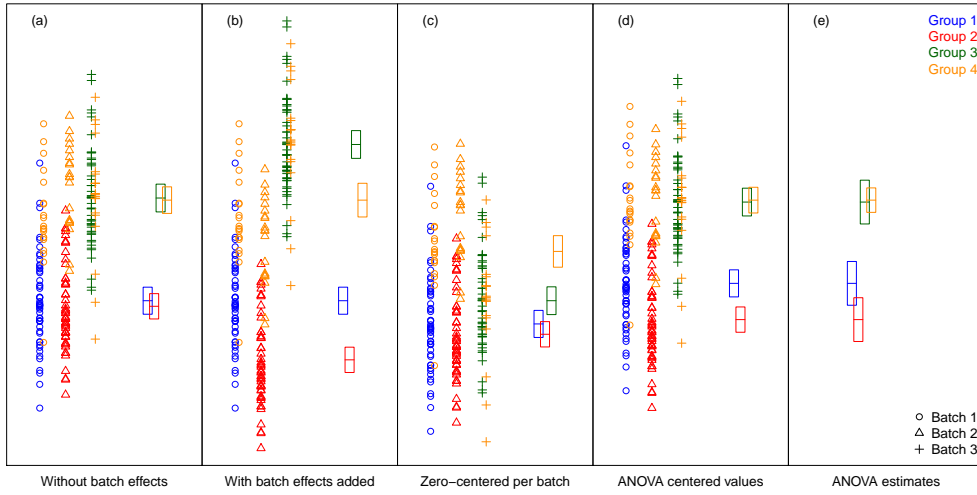
Figure 1. We simulated expression of one gene from four study groups unevenly distributed in three batches containing 50+0+0+20, 0+50+0+20 and 0+0+50+20 samples from groups 1 to 4. This case, design and effects, was selected to illustrate the spurious effects that may arise from different batch adjustments. 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.

(a) P–values, Random numbers
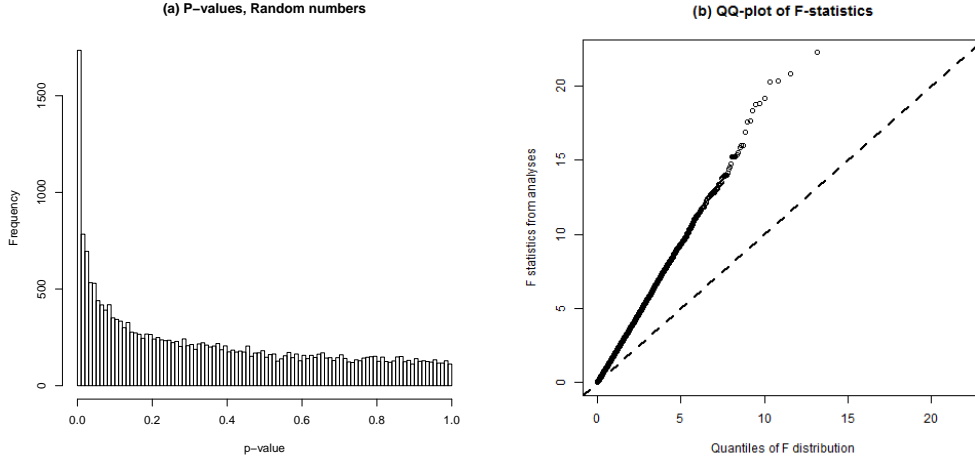
(b) QQ-plot of F-statistics

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

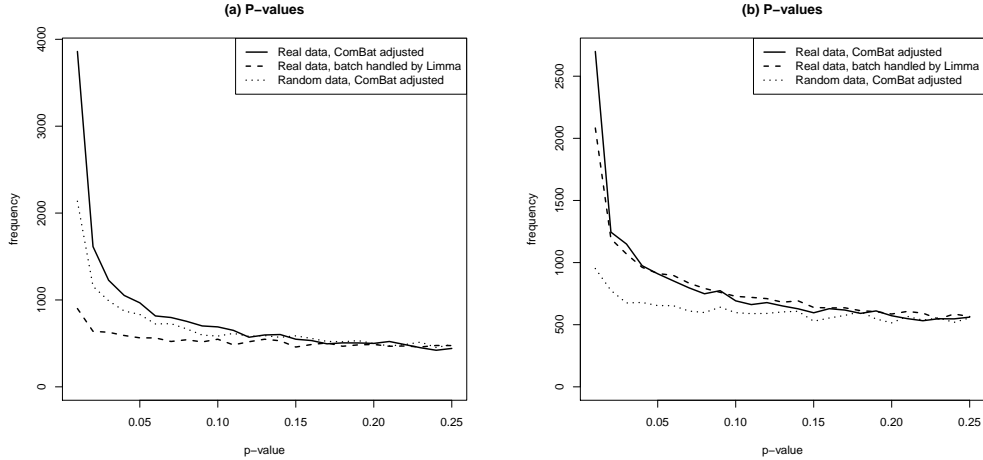**(a) P–values**

(b) P–values

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