Biostatistics (2014), **0**, 0, *pp.* 1–10
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, EIVIND HOVIG*

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

ehovig@radium.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, recommended by the authors [Anbefalt av dem alle eller kun av Johnson?], 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 recommended and frequently used 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 reagents and other equipment, for instance microarray chips, are made in batches, and this is frequently observed as an effect in the measurements. See **?** for more examples.

For a typical experiment comparing group differences, 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 act as a confounder and induce false differences between groups (**?**).

*To whom correspondence should be addressed.

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, 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.

The first step 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 influences 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, this means group differences are estimated based on within batch comparison, and applied to the batch adjusted data. In a balanced group–batch design, group differences and batch effects are independent, and this approach becomes identical to the above described zero-centering per batch. If the group–batch design is heavily unbalanced, estimation of group differences and batch effects are interdependent. Unfortunately, when group differences estimated from an unbalanced group–batch design are applied to the entire data set, if this batch adjusted data set is later analysed for group differences, the confidence will be exaggerated as the estimation of the group differences in the unbalanced design is less accurate than in the corresponding balanced design. 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 **?**, and included in the sva package (**?**), can use either of the two above described approaches to estimate batch differences, but uses an empirical Bayes approach to avoid over-correction for batch effects for small batches. It has popularised the two-way ANOVA procedure for retaining group differences when adjusting for batch effects, and the inclusion of group difference as a covariate when removing batch effects has been recommended [reference]. Based on actual use of ComBat by the authors and others, we suspect thus adjusted data are commonly treated as "batch effect free" in subsequent analysis. And as a consequence, confidence in group effects has been overestimated and false results reported.

Kort oppsummere behovet for batch-korrigering: generelt og i microarraydata.

Kort beskrive metoder som gjoer dette (batch-sentrering, ANOVA, ComBat).

Forklare tilsiktet effekt disse paa batch+gruppe-forskjeller (balansert versus ikke-balansert).

Forklare kort hva faktisk effekt er: ANOVA gir redusert gruppe-forskjeller, mens for ComBat overestimeres den statistiske styrken/sikkerheten av forskjellen.

2. Methods for batch effect correction

2.1 *Model for data with batch effects*

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

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

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 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 **?**, 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, **?** 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 γ_{ig} later, but out main argument is more easily made in the simpler context.

2.2 *Standard batch correction methods*

The main ambition of batch effect adjustments is to be able to remove batch differences in such a way that downstream analyses of the adjusted data may be done without further batch adjustments.

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

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

An alternative is to centre each batch to the common average by adding the average value $\overline{Y}$ across the entire data set: i.e. $\widetilde{Y}_{ijr} = \Delta\widetilde{Y}_{ijr} + \overline{Y}$. When comparing groups, the common value $\overline{Y}$ has no effect, and so this is equilvalent to zero-centring each batch. However, if the different groups are unevenly represented in the different batches, the batch average $\overline{Y}_i$ will tend to capture group differences as well as batch effects. Thus, batch centering may reduce group differences, and thus reduce the power of downstream analyses.

Removing batch effects while retaining group differences can be done through an ANOVA analysis in which the group effects, β_j, and the batch effects, γ_i, are estimated simultaneously. Batch adjusted values may then be obtained by subtracting the estimated batch effects, $\hat{\beta}_j$:

$$\widetilde{Y}_{ijr} = Y_{ijr} - \hat{\beta}_j = \alpha + (\beta_j - \hat{\beta}_j) + \gamma_i + \epsilon_{ijr}. \tag{2.4}$$

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

Unfortunately, the estimation error $\hat{\beta}_j - \beta_j$ affects all values within the same batch in the same manner, inducing a dependency between the values. If the study groups are evenly represented in all batches, this will not influence estimated group differences as all groups are equally affected. However, if groups are unevenly represented, this induced dependency can have severe impact on downstream analyses.

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 F-test. Swapping the real data with random numbers from a normal distribution (mean=0, sd=1), but otherwise following the instructions, will generate the p-value distribution shown in Figure 1. The skewed distribution is a indication that this approach may have a unintentional adverse effect. [Jeg har ikke beskrevet permutatsjon checken fordi jeg er usikker paa om vi skal bruke den.]

Formler og ord: Forklare hva som gaar galt med formler og ord. Overlater til Einar aa skrive dette.

3.2 *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.

1) In the first experiment(**?**), cells were treated with glatiramer acetate (a medicine for multiple sclerosis) or a generic 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 differetially expressed genes, yielding a list of 1000 genes (Table S5, **?**). 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). Additional sanity checks with random numbers or perumted labels were also carried out and the distribution of p-values for differents settings are shown in Fig. 2. Our conclusion is that most of the genes reported as differtially expresssed in (**?**) 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 assumpotion of no difference. The R-code for our analysis and a more extensive report can be downloaded from github ().

2) The second example is taken from the supporting information for the original ComBat article (**?**) where it is denoted "data set 2". Cells inhibited for the expression of the TAL1 gene were compared to controls on a microarray platform. 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 (qj0.05). Then, we analysed their data without batch adjustment in ComBat, but blocking for batch in limma, resulting in 377 probes (qj0.05). In addition the two sanity checks outline above were performed. In contrast to the above results obtained for(**?**), the P-value distributions for the alternative analyses (Fig.3 a,b) does not indicate a huge difference. Nevertheless, we belive that P-values are deflated for the ComBat adjusted analysis. The R-code for our analysis and a more extensive report can be downloaded from github ().

3) [Her kan vi fylle paa med flere eksempler etter samme mal hvis det trengs. Feks en fersk artikkel fra Nature Genetic (**?**). Der har de kombinert sekvenseringsdata (deres data) fra 9 lymfom pasienter med et (eksternt) microarray data set som har 6 lymfom proever og 5 kontroller, dvs

kontroll gruppen er helt fravaerende hos dem. Dvs veldig ubalansert (9/0, 6/5), MEN skille mellom phenotypene er saa stor at de sikkert hadde funnet omtrent de samme genene signifikante med aa bare bruke de balanserte dataene (men det var jo ikke deres.) Dette er ikke en viktig del av paperet, men det er veldig feil selvom konsekvensense ikke blir store. Dog gir det et falskt inntrykk av at dataene deres ligner tidligere data, se heatmap i suppl.. Dessuten lager de GO analyser av dette som egentlig er 7 aar gamle andres data og ikke deres]

4. Discussion

4.1 *Increased emphasis on preserving group difference*

In the original ComBat article (**?**) it is clear that the primary motivation behind ComBat was to employ an Emprical Bayes method for batch-effect removal. The feature of retaining group differences for unbalanced designs is optional and seems to be subordinate, only exemplified in the supplementary information. However, over the years this feature became more important judging from advice given by the author to users ([1]() ,[2]() ,[3]() ,[4]()). [Aa linke direkte er kanskje litt ufint, alternativt kan det linkes til forumet uten konkrete innlegg.] And when ComBat was incorporated in the sva package (**?**), inclusion of group labels was made almost mandatory (an undocumented option of passing a NULL value exists). Thus, this problematic use has likely been common.

4.2 *Motivation for this warning*

Our knowledge of ComBat came through a typical use case when trying to salvage unbalanced data which had batch effects. Upon realizing that the confidence on our group differences were exaggerated, the literature was searched for a better understanding of correct use and potential overseen limitations of ComBat. But the authors of ComBat and the sva package recommended our usage (**?**, **?**). In addition, other works looking into the problem of batch effects were mostly recommending ComBat without much concern (**?**, **?**). A brief inquiry into some of the articles citing ComBat (574 Google Scholar) revealed few problems, and their method descriptions regarding ComBat were mostly sparse, limited to one or two sentences ([24391845], [18414638], [21731603], [23630272], [24584070]). A further indication of their carefree use of this potentially devastating procedure was the frequent neglect to state the program parameters, i.e batch labels ([18414638], [21731603]) or group labels ([24391845], [18414638], [21731603], [23630272], [23482648], [24584070]). Often no effort was done in order to substantiate the existence of batch effects in their data, except for stating the presence of batches ([24391845]?, [18414638], [21731603], [23630272]). In one instance ComBat was even applied on data were effects due to batch were investigated, but not found [husker ikke ref]. 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 that many published results from data adjusted by ComBat are completely or partially false. And knowing that scientists don't give up where one analysis fails, a method that almost ensures a result given a sufficiently unbalanced design will continue to be used.

4.3 *Practical advice*

We have shown that adjusting for batch effects while preserve the group difference may lead to varying degree of false results. Knowing this, to what degree 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. no 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 (**?**, [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.

Reproducible research

[vi boer klare aa tilfredstille kravene i http://biostatistics.oxfordjournals.org/content/10/3/405.full]

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 genomics* **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.

Yoo, Hae Yong, Sung, Min Kyung, Lee, Seung Ho, Kim, Sangok, Lee, Haeseung, Park, Seongjin, Kim, Sang Cheol, Lee, Byungwook, Rho, Kyoohyoung, Lee, Jong-Eun, Cho, Kwang-Hwi, Kim, Wankyu, Ju, Hyunjung, Kim, Jaesang, Kim, Seok Jin, Kim, Won Seog, Lee, Sanghyuk *and others*. (2014, March). A recurrent inactivating mutation in RHOA GTPase in angioimmunoblastic T cell lymphoma. *Nature genetics* **46**(4), 371–375.

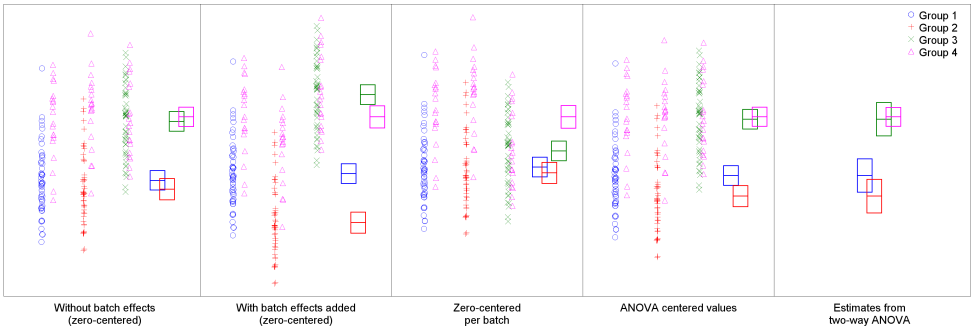

Figure 1. Simulated data from four study groups was generated where groups 1 and 2 have lower means than groups 3 and 4. These were placed in three different batches with batch effect added. Values and boxes showing mean and two standard errors of the mean are displayed for data without batch effects, 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.

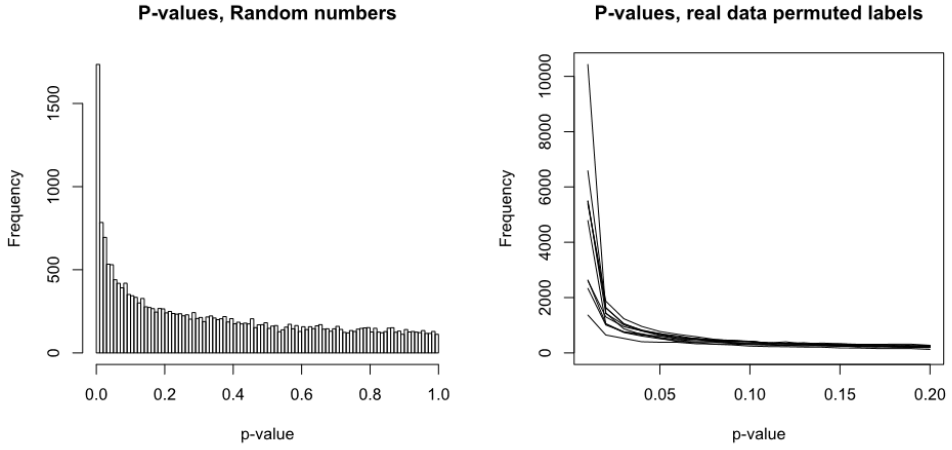

Figure 2. Two sanity checks were the recommended use of ComBat fails. Adapted from the user guide in the sva package. a) Real data is substituted with random numbers from a normal distribution (mean=0, sd=1), but the batch/group design is retained, followed by batch adjustment in ComBat and a F-test. b) 10 runs of real data with the "cancer" labels permuted within batches, followed by ComBat adjustment and a F-test. All permutations produces a skewed p-value distribution.

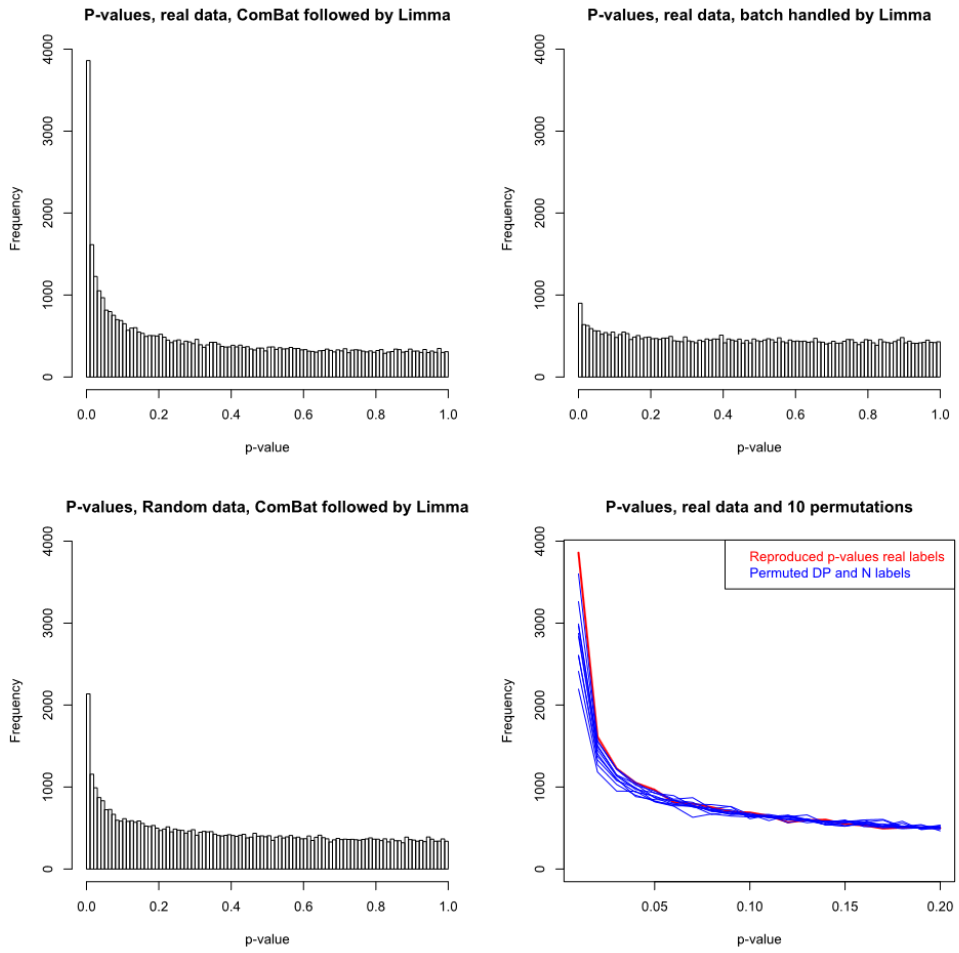

Figure 3. P-value distribution for the main comparison in (data set 2 24421904) glatiramer acetate vs. generic. a) Real data adjusted by ComBat as described in the article followed by a significance test in limma. b) Real data not adjusted by ComBat, tested for significance in limma using batch as a blocking factor. c) Real data is substituted with random numbers from a normal distribution (mean=0, sd=1), but the batch/group design is retained, followed by batch adjustment in ComBat and a significance test in limma. d) 10 runs of real data with the glatiramer acetate("DP") and generic("N") labels permuted within batches, followed by ComBat adjustment and significance test in limma. [jeg tok med begge sanitycheckene her. Er d) slik einar foreslo?. Merk jeg har ikke permutert alle labeler, bare de mellom DP og N. Det er en del andre.]

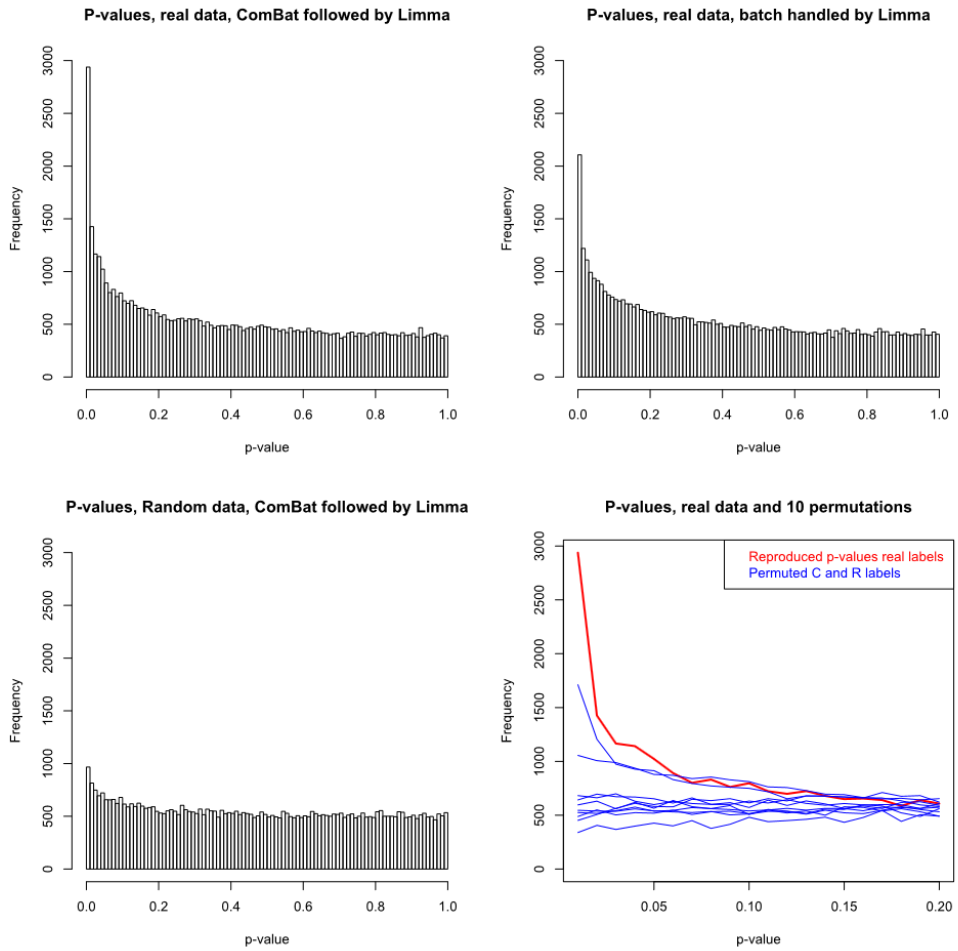

Figure 4. P-value distribution for the main comparison in (data set 2, **?**) for a) Real data adjusted by ComBat as described in the article followed by a significance test in limma. b) Real data not adjusted by ComBat, tested for significance in limma using batch as a blocking factor. c) Real data is substituted with random numbers from a normal distribution (mean=0, sd=1), but the batch/group design is retained, followed by batch adjustment in ComBat and a significance test in limma. d) 10 runs of real data with the group labels permuted within batches, followed by ComBat adjustment and significance test in limma. [Problemet med permutatsjons sanitychechen slik jeg har utfoert og plottet den her er at den neppe kan sies aa feile. Dessuten kan de 3 foerst plottene bli plottet i ett som vist under i en alternativ version jeg synes er finere, men dette gaar ikke for permutasjonstesten tror jeg]

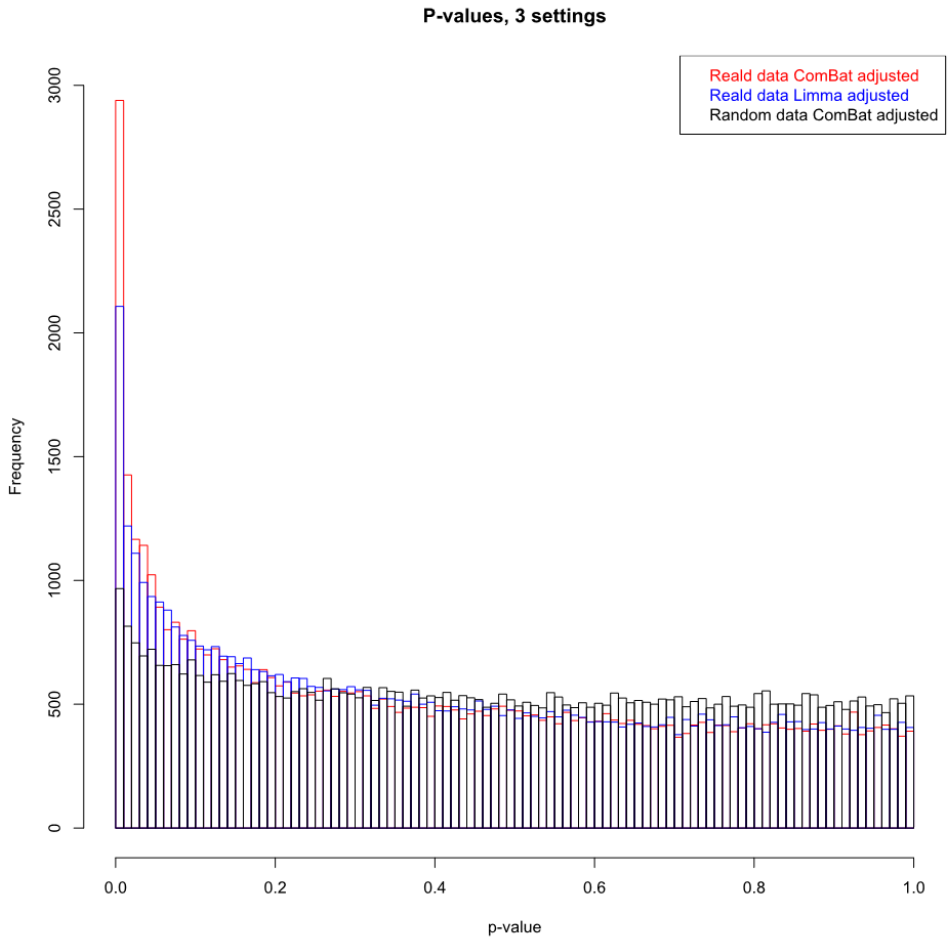

Figure 5. [Alternativ illustrasjon av de tre p-value distribusjonene]