Bioinformatics, 37(18), 2021, 3051–3052
doi: 10.1093/bioinformatics/btab159
Advance Access Publication Date: 8 March 2021
Applications Note

Gene expression

Omixer: multivariate and reproducible sample randomization to proactively counter batch effects in omics studies

Lucy Sinke *, Davy Cats and Bastiaan T. Heijmans

Molecular Epidemiology, Department of Biomedical Data Science, Leiden University Medical Centre, Leiden 2333 ZC, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on September 29, 2020; revised on February 2, 2021; editorial decision on February 27, 2021; accepted on March 4, 2021

Abstract

Motivation: Batch effects heavily impact results in omics studies, causing bias and false positive results, but software to control them preemptively is lacking. Sample randomization prior to measurement is vital for minimizing these effects, but current approaches are often ad hoc, poorly documented and ill-equipped to handle multiple batches and outcomes.

Results: We developed Omixer—a Bioconductor package implementing multivariate and reproducible sample randomization for omics studies. It proactively counters correlations between technical factors and biological variables of interest by optimizing sample distribution across batches.

Availabilityand implementation: Omixer is available from Bioconductor at http://bioconductor.org/packages/release/bioc/html/Omixer.html. Scripts and data used to generate figures available upon request.

Contact: l.j.sinke@lumc.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Batch effects can overshadow biological differences in size (Baggerly *et al.*, 2008) and critically influence the results of omics studies (Harper *et al.*, 2013; Lambert and Black, 2012). Even in benign cases, they decrease power to detect a true biological effect or contaminate results with false positives (Leek *et al.*, 2010). Despite the numerous statistical methods developed to adjust for batch effects (Espin-Perez *et al.*, 2018; Johnson *et al.*, 2007; van Iterson *et al.*, 2017), a reactive approach is often insufficient. In fact, when technical variables are confounded with experimental factors of interest, batch effect correction will mask the underlying biological signal (Goh *et al.*, 2017).

Sample randomization is a proactive, and arguably more impactful, method for obtaining reproducible results in high-throughput experiments (Yang *et al.*, 2008). However, its implementation suffers from several key issues. Particularly where there are numerous or nested batches each composed of a limited number of samples, such as separate microarrays or sequencing lanes, single random draws can inadvertently result in high correlations between technical covariates and biological factors. This is further complicated by an often poorly documented randomization process that is not necessarily reproducible. Although stratified randomization has been shown to effectively remove chip effects in microarray experiments (Buhule *et al.*, 2014), it does not address all relevant biological variables. Therefore, to adequately combat bias in results, employing methods capable of handling a wider array of research setups is imperative.

We developed Omixer—an R package for multivariate and reproducible randomization in omics studies. From a diverse range of randomized sample layouts, it selects the one that optimally balances biological variables across batches. Omixer offers the flexibility required to perform randomization effectively in a variety of study designs and experimental setups.

2 Materials and methods

To optimize distribution of biological variables across batches, sample randomization is performed multiple times (default: 1000; see Supplementary Fig. S1 for more information). After combining resulting lists with the user-specified plate layout, statistically robust tests of correlation determine the optimal setup, where the absolute sum of correlations between biological and technical factors is minimized. As a precautionary step, layouts with evidence for any tested batch associations are excluded ($P < 0.05$ for any batch-outcome correlation), although in practice this will not change the resulting layout given suitably large iteration numbers (see Fig. 1A).

To reserve wells for control samples or other studies, a mask can be specified in the options, and paired samples such as those from

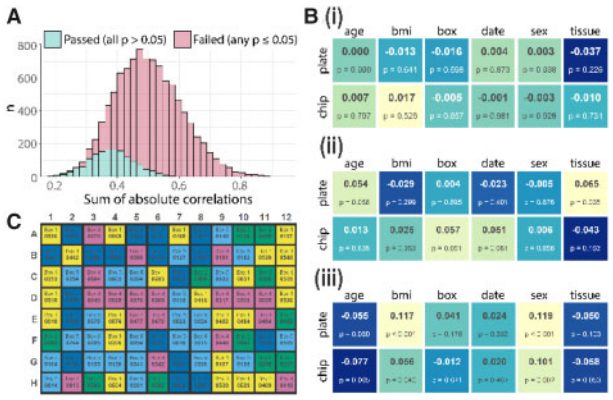

Fig. 1. Overview of Omixer functionality and graphical output with (**A**) distribution of the sum of absolute correlations from 10 000 randomized layouts, coloured by filtering step outcome (**B**) resulting correlation matrices from the (i) optimal Omixer layout, (ii) median result and (iii) worst case scenario after simple randomization and (**C**) lab-friendly sample sheets created by Omixer as a PDF, showing the first plate colour coded by box number

twin studies can be blocked so they remain together in the same batch. Non-standard plate layouts can be specified, but Omixer will automatically generate the most commonly used plate and chip combinations. Previously generated layouts can quickly be reproduced, and lab-friendly sample sheets reduce the risk of mixups when manually pipetting samples.

2.1 Multivariate and reproducible randomization

The main function, omixerRand, takes a sample list and plate layout as input and optimizes distribution of specified biological variables across batches. Resulting correlations are visually displayed and the optimal seed is saved locally. By loading this seed, previously generated layouts can be reproduced quickly and efficiently with the omixerSpecific function.

2.2 Lab-friendly sample sheets

The omixerSheet function converts the output of previous Omixer functions into lab-friendly sample sheets, saving these in the working directory as a printable PDF. Wells can be coloured by other variables, such as box number (see Fig. 1C) or tissue, to further smooth transition into the wet lab.

2.3 Omixer outperforms simple randomization

Particularly when multiple batch types and outcomes are present, a single randomization is likely to result in significant correlations. As an example, we randomized 672 samples across 2 levels of batches, as described in the Omixer vignette. Following 10 000 simple randomizations, 85% of the resulting layouts have at least one *P*-value under 0.05. The distribution of the sum of absolute correlations for the resulting 10 000 layouts (Fig. 1A) suggests that the expected sum of correlations between batches and outcomes following a single randomization is 0.5. The correlations present in an average selection (Fig. 1B.ii) are small on the whole (0.004 to 0.065), but significant associations ($P < 0.05$) still exist.

Looking at the worst case scenario following simple randomization (Fig. 1B.iii), we see that simple randomization has the potential to choose layouts with multiple significant associations ($P < 0.05$ for 5 comparisons), resulting in large batch effects that will bias results. By contrast, Omixer would reject all layouts with significant correlations, and instead return an optimal layout from the 15% remaining (blue in Fig. 1A). In this example, the optimal layout's correlations (Fig. 1B.i) are all under 0.037 and none are significant.

3 Conclusions

In conclusion, Omixer offers an intuitive, reproducible alternative to current randomization practices in omics research. Its implementation is a key step in combatting batch effects preemptively and reducing the risks of sample mixups in the wet lab.

Funding

This work was supported by the Joint Programming Initiative 'a Healthy Diet for a Healthy Life' (JPI-HDHL) DIMENSION project [ZonMW project number: 529051021].

Conflict of Interest: none declared.

Data availability

Scripts and data used to generate the images is available upon request. Otherwise, Omixer is a software tool that uses user input data.

References

Baggerly,K. *et al.* (2008) Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J. Clin. Oncol.*, **26**, 1186–1187.

Buhule,O. *et al.* (2014) Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Front. Genet.*, **5**, 354.

Espin-Perez,A. *et al.* (2018) Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptomic data. *PLoS One*, **13**, e0202947.

Goh,W.W.B. *et al.* (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.*, **35**, 498–507.

Harper,K. *et al.* (2013) Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol. Biomarkers Prev.*, **22**, 1052–1060.

Johnson,W. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Lambert,C. and Black,L. (2012) Learning from our GWAS mistakes: from experimental design to scientific method. *Biostatistics*, **13**, 195–203.

Leek,J. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

van Iterson,M. *et al.*; BIOS Consortium. (2017) Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.*, **18**, 19.

Yang,H. *et al.* (2008) Randomization in laboratory procedure is key to obtaining reproducible microarray results. *PLoS One*, **3**, e3724.