Beyond the CpG: an integrative approach to decoding DNA methylation in immunometabolic health

Sinke, L.J.

CHAPTER FOUR

Omixer

Omixer: multivariate and reproducible sample randomisation to proactively counter batch effects in genomic studies

Lucy Sinke[1], Davy Cats[1], and Bastiaan T. Heijmans[1]

[1] Leiden University Medical Centre, Leiden, The Netherlands

Published in Bioinformatics 37(18), 3051-2 (2021)
DOI: 10.1093/bioinformatics/btab159

Abstract

Batch effects heavily impact results in genomic research, causing bias and false positive results, yet software to control them pre-emptively is lacking. Sample randomisation prior to measurement is vital for minimising these effects. However, current approaches are often *ad hoc*, poorly documented, and ill-equipped to handle multiple batches and outcomes.

We developed *Omixer*: a Bioconductor package implementing multivariate and reproducible sample randomisation for genomic studies. It proactively counters correlations between technical factors and biological variables of interest by optimising sample distribution across and within batches.

Background

Batch effects can overshadow biological differences and critically influence the results of genomic research[1–3]. Even in benign cases, they decrease power to detect a true biological effect and may contaminate results with false positives[4]. Despite numerous statistical methods developed to adjust for batch effects, these frequently prescribed reactive approaches are often insufficient[5–7]. When technical variables are confounded with experimental factors of interest, batch effect correction will likely mask genuine underlying biological signals[8].

Sample randomisation is a proactive, and arguably more impactful, method for obtaining reproducible results in high-throughput experiments[9]. Despite this, its current implementation suffers from several key issues. Particularly where there are numerous or nested batches 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 randomisation process that is not by default reproducible. Although stratified randomisation is capable of effectively removing batch effects in microarray experiments, it cannot address all relevant biological variables[10]. Therefore, to adequately combat bias in results, it is imperative that we employ methods capable of handling a wider array of research setups.

We developed *Omixer*: an R package for multivariate and reproducible randomisation in genomic studies. From a large number of randomised sample layouts, it selects the one that optimally balances biological variables across batches. Omixer offers the flexibility required to perform randomisation effectively and reduces the risk of masked or false positive signals across a range of common experimental setups and study designs.

Keywords: *Bioconductor, genomics, randomisation, software*

Results

Multivariate and reproducible sample randomisation

To optimise distribution of samples across batches, randomisation is performed multiple times. The default number of iterations is 1,000 and this can be adjusted if needed (using option `iterNum`; **Fig. 1**). After combining sample lists with the specified plate layout, statistical tests of correlation determine the optimal setup. This is defined as the layout that minimises the absolute sum of correlations between defined biological and technical factors. As a precautionary step, layouts with evidence for any tested batch associations are excluded ($p < 0.05$), although in practice this will not change the result given suitably large iteration numbers (**Fig. 2a**).

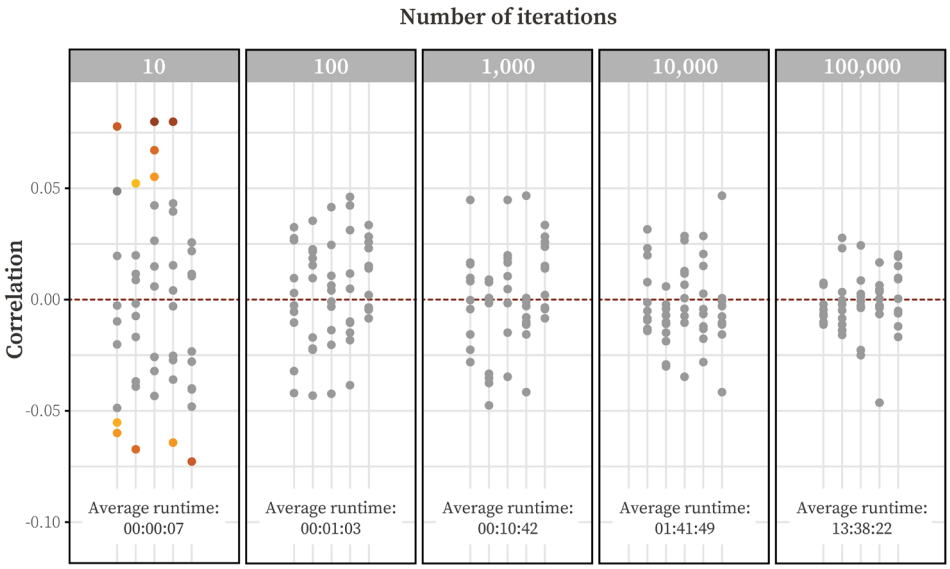

Figure 1 | The default number of iterations Omixer performs is 1,000. Increasing this number may result in smaller overall correlations between technical and biological factors, and complex designs could benefit from such increases above the default. Users should be aware that this choice comes at the cost of increased runtime, exemplified here using 1 CPU with 64 GB RAM to randomise 616 example samples across seven 96-well plates.

To reserve wells for control samples or other studies, a mask can be specified (using option: `mask`), and paired samples such as those from twin studies can be blocked so that they remain together in the same batch (using option: `block`). Non-standard plate layouts can be input, but Omixer also generates commonly used plates (using options: `wells` and `plateNum`). Previously generated layouts can quickly be reproduced using the *omixerSpecific* function and the automatically saved *randomSeed* object.

The main function, *omixerRand*, takes a sample list and plate layout as input and optimises the distribution of specified biological variables (option: `randVars`) across batches (option: `techVars`). The resulting correlations between these two types of variables in the returned layout are then visually displayed as a heatmap for inspection by the user.

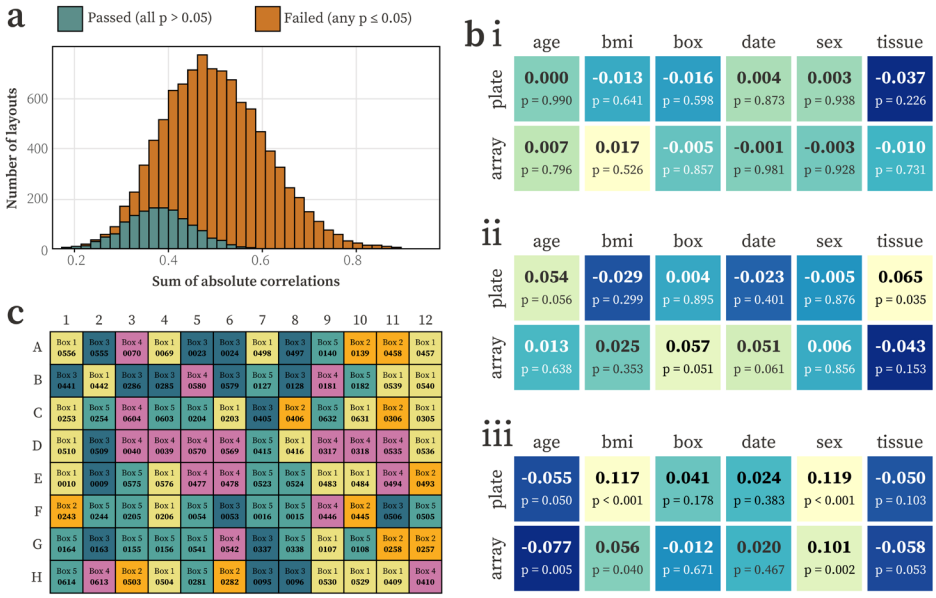

Figure 2 | Overview of Omixer functionality and graphical output. a) Distribution of the sum of absolute correlations from 10,000 randomized layouts, coloured by filtering step outcome. **b)** Resulting correlation matrices from the **i)** optimal Omixer layout, **ii)** median result, and **iii)** worst case scenario after simple randomization. **c)** An example of a lab-friendly sample sheet created by Omixer, showing a 96-well plate coloured by box number.

Omixer outperforms simple randomisation approaches

Particularly when multiple batch types and outcomes are present, a single randomisation is likely to result in significant correlations. To exemplify this, we randomised an example dataset of 616 samples intended to be profiled using the Illumina Infinium MethylationEPIC BeadChip array. Randomisation variables of interest were specified as age, body mass index (BMI), box, isolation date, sex, and tissue. Batches were specified as plate number (plate) and array ID (array). Following 10,000 simple randomizations, 85% of the resulting layouts had at least one significant correlation between a randomisation variable and batch ($p < 0.05$; **Fig. 2a**). The distribution of the sum of absolute correlations in the resulting 10,000 layouts suggested that the expected value for this sum was close to 0.5. Notably, the correlations present in such an average selection

were predominantly small (0.004 to 0.065; **Fig. 2bii**), but significant associations still prevailed ($p < 0.05$).

The worst-case scenario demonstrated that simple randomisation has potential to return layouts with multiple significant associations ($p < 0.05$ for five comparisons in this example; **Fig. 2biii**). This layout could result in substantial batch effects that may bias results. By contrast, *Omixer* would reject all such layouts with significant correlations, and instead return only a layout from the 15% remaining (blue in **Fig. 2a**). In this example, the optimal layout has no significant correlations, and all have absolute values below 0.037 (**Fig. 2bi**).

Omixer produces lab-friendly sample sheets

The *omixerSheet* function converts the output of previous Omixer functions into lab- and colour-blind friendly sample sheets, saving these in the working directory as a printable PDF. Wells can be coloured by specified variables, such as box number or tissue, to further smooth transition into the wet lab (using option: `group`). Such lab-friendly sample sheets improve accessibility of Omixer and reduce the risk of mix-ups when manually pipetting samples (**Fig. 2c**).

Conclusions

In conclusion, Omixer offers an intuitive, reproducible alternative to current randomization practices in genomic research. Its implementation is a key step in combatting batch effects pre-emptively and reducing the risks of sample mix-ups in the wet lab.

Author contributions

L.S. and B.T.H. conceived the project and designed the software. D.C. contributed key expertise in computation and software development.

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].

Declarations of interest

The authors declare no competing interests.

Data & code availability

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

References

1. Baggerly, K., A., Coombes, K. R., *and* Neeley, E. S. Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J Clin Oncol* **26** (7): 1186-1187 (2008).
2. Harper, K. N., Peters, B. A., *and* Gamble, M. V. Batch effects and pathway analysis: Two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol Biomarkers Prev* **22** (6): 1052-1060 (2013).
3. Lambert, C. G. *and* Black, L. J. Learning from our GWAS mistakes: From experimental design to scientific method. *Biostatistics* **13** (2): 195–203 (2012).
4. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11** (10): 733–739 (2010).
5. Espín-Pérez, A. *et al.* Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data. *PLoS One* **13** (8): e0202947 (2018).
6. Johnson, W. E., Li, C., *and* Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** (1): 118-127 (2007).
7. van Iterson, M. *et al.* Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* **18** (1): 19 (2017).
8. Goh, W. W. B., Wang, W., *and* Wong, L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol* **35** (6): 498-507 (2017).
9. Yang, H. *et al.* Randomization in laboratory procedure is key to obtaining reproducible microarray results. *PLoS One* **3** (11): e3724 (2008).
10. Buhule, O. D. *et al.* Stratified randomization controls better for batch effects in 450K methylation analysis: A cautionary tale. *Front Genet* **5:** 354 (2014).