Title:

Covariate balanced allocation of samples to batches to mitigate the impacts of technical variability

Authors

Mulvey John F*, Lundby Alicia

Affiliation: Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

*To whom correspondence should be addressed.

ORCIDs

John Mulvey 0000-0003-3843-2442

Alicia Lundby 0000-0002-1612-6041

Word count

Abstract: 127 Main: 2043

Abstract

Performing assays on large numbers of samples requires their analysis in distinct batches, which commonly affects the measurements made in a systematic way. Analytical approaches can correct for such batch effects, however for this to be possible the batch should not be confounded with either the independent variable or any covariates relevant to the analysis. Thus, how samples are distributed across batches influences subsequent analytic conclusions. We present *SampleAllocateR*, an open source tool that employs optimisation methods from machine learning to optimally allocate a preselected set of samples to experimental batches in a way that statistically balances specified covariates between batches. This results in better statistical estimates of the effect of not only the technical batch effects but also the other specified covariates upon the results of the experiment.

Introduction

Experimental assays, including not only traditional biochemical assays but also more multiplex approaches where many analytes are measured in parallel (exemplified by -omics technologies), require samples to be processed and/or measured in batches. This often

results in the introduction of technical batch effects that should be accounted for during the statistical analysis of the analyte of interest.

Controlled experimental designs seek as far as is possible to hold constant all covariates other than the independent variable. To take a fictitious example: comparing a drug vs vehicle only between co-housed littermates of genetically identical mice with a surgically induced myocardial ischaemia. To conduct such an experiment requires the use of a model system (in our example, a mouse model of a human disease) which necessitates a tradeoff in the relevance to the actual system of interest (a myocardial infarction in a patient). Facilitated by debate around the external validity of some common model systems to the human diseases of interest (Perel *et al.* 2007; Downey and Cohen 2009; Pound and Ritskes-Hoitinga 2018), there has been an increasing amount of interest in analysing clinical samples directly so as to maximise the translational relevance of the research. In contrast to samples from carefully controlled experiments, clinical samples usually differ in not only a single independent variable of interest but also in a number of *covariates* that commonly also impact upon the measured quantity. Examples of such covariates include age and body mass index (both of which have been well described to impact outcomes after myocardial infarction). Further than merely correcting for these covariates, researchers are furthermore also increasingly interested in understanding interactions between these covariates and the outcome of interest (Ferdinandy, Szilvassy and Baxter 1998; Ferdinandy, Schulz and Baxter 2007). Batch effects can be statistically corrected for in the assessment the effect of the independent variable and covariates of interest (Leek *et al.* 2010), but only if care is taken when allocating samples across the different experimental batches to ensure that there is minimal confounding between the (technical) batch variable and these (biological) covariates.

When describing experimental designs, researchers commonly refer to samples being randomised between batches (Forshed 2017; Alseekh *et al.* 2021), or sometimes will describe “blocking” of samples according to a key covariate of interest (Forshed 2017). This follows the adage for the design of clinical trials, usually attributed to George Box, to “block what you can and randomize what you cannot” (Box, Hunter and Hunter 2005). However, measurements of samples that have already been collected differs from the design of clinical trials in that the covariates for all samples are already known at the start of the experiment. Randomisation, either of individual samples or of blocks produces no *statistical bias*: that if the allocation of samples to batches were to be repeated many times, there would be no difference between the expected value of an estimator and the true value of the parameter. However, randomisation does not remove the *statistical variability* of the estimator, and thus even without any *statistical bias* our estimator may be far from the true value. This would correspond to a randomly generated allocation that, by random chance, contains confounding between the batch variable and the covariates of interest. Since we get the opportunity to perform our experiment only once, randomisation therefore represents a poor strategy: the allocation of samples between batches that we randomly draw will, depending on our luck, have a greater or lesser degree of confounding between batches and variables and covariates of interest.

We therefore advocate that samples should be, *non*-randomly and very deliberately, allocated across batches in order to minimise the degree of confounding by experimental batch in the single layout used for the experiment. Sample allocation across batches is often done *ad hoc* by researchers where the resultant layouts may not be close to the layout which minimises the degree of confounding. This approach may be due in part to the lack of well performing tools to help in the design of the experiment (Table S1).

Here we present a tool to allocate samples to batches in a manner that maximises the *balance* of specified covariates, either continuous or categorical, between batches. This increases the accuracy of estimates made when correcting for the technical batch effects, and hence improves the statistical estimate of the predictor variables relevant to the experiment.

Methods

We specify that the problem is to allocate a set of preselected samples with a number of known covariates into batches of a specified size, so as to maximise the balance of the given covariates. This is summarised in Figure 1A.

We use the balance of covariates to refer to the degree to which the distribution of covariates is similar across different batches (Austin 2009). We measure the balance of a covariate by calculating an appropriate statistical test for continuous or categorical covariates using the batch as the independent variable (Kruskal-Wallis test or Fisher's exact test, respectively). Where multiple covariates are present, we define a "balance score" as the harmonic mean of the *p*-values of all of these tests: allowing their combination into a single metric without making strict assumptions of their independence (Wilson 2019). The harmonic mean has additional advantages in that it is sensitive to relatively small changes in balance, ensures that every covariate must be meaningfully balanced rather than allowing one to compensate for another, and in our case is easily interpretable on the scale [0,1].

We begin by initially distributing samples at random across batches. We define the balance across batches as the objective function that we seek to maximise. We use a simulated annealing algorithm to approximate the global optimum of the objective function (Pincus 1970; Kirkpatrick, Gelatt and Vecchi 1983). Briefly, at each of a specified number of steps we randomly select samples to exchange between batches to generate a neighbouring layout and recalculate the objective function. This neighbouring state is probabilistically accepted depending upon the "energy" of the system, which gradually decreases over time, facilitating both early broad exploration of the solution space and later climbing to the maxima of the local environment. This is computationally efficient enough to run quickly on a personal computer and does not require external dependencies.

We additionally provide functionality to either allocate individual samples to batches, or to block together samples if required. Blocking is preferred if possible as it eliminates the impact of batch effects upon the estimation of the variable that is blocked, rather than merely mitigating against it. To do this, we perform the same procedure described above on blocks rather than individual samples.

Results

We construct a fictious experiment, consisting of allocating 98 samples to a batch size of 13 using simulated covariate data. We demonstrate that when considering a single covariate, randomly allocating samples to batches will as expected result in a uniform distribution of *p* values when testing for differences in a given covariate between batches (Figure 1B). However, when multiple covariates are considered simultaneously, there were few ways to allocate samples into a highly balanced layout indicated by the sparsity of high balance scores: a metric composite metric constructed to combine *p* values across all covariates (Figure 1C; 3 covariates). This makes balanced layouts hard to find by "random search" methods, which is compounded by the fact that there are large number of possible ways to partition even a modest number of samples into distinct batches. For this reason, we apply a iterative optimisation method, which achieve a higher level of covariate balance between technical batches than an alternative random search methods that simulate a large number of random layouts and select the best one (Figure 1D). This is achieved despite a decrease in the computational time required (Figure 1E).

Discussion

We present a tool and description of the underlying method to allow scientists to easily allocate their samples across technical batches whilst minimising the impact of the resulting technical variation upon the conclusions drawn from downstream data analysis. We make it freely and openly available as an R package (Mulvey 2024), with a simple interface and sufficient documentation for it to be used not only by bioinformaticians but also by nonexpert users.

We design this tool for the proximal use case in the design of -omics experiments analysing clinical samples. By allowing the user to freely specify the size of the batches, it can be applied across diverse -omics methodologies where batch effects have been described including but not limited to RNAseq (Marioni *et al*. 2008), proteomics (label free quantification, isobaric mass tags) (Hu 2005), metabolomics (Dunn *et al*. 2011), genomics (Kircher, Heyn and Kelso 2011; Schirmer *et al*. 2015) and epigenomics. It is applicable also for variations thereof where the measurements of these features is performed in single cells, nuclei (Lun and Marioni 2017) or resolved in space, and equally to traditional assays without multiplexing such as those based upon ELISA.

Historically, we as a community have inadequately accounted for sources of technical variation in our experimental design and there are numerous examples where batch effects have confounded variables of interest. To take one example, Akey and colleagues used sequencing metadata to demonstrate that batch effects in data from the International HapMap consortium (The International HapMap Consortium 2005) were confounded with variables being used by researchers to draw incorrect conclusions about biological differences (Akey *et al*. 2007). It is noteworthy that such metadata has not always been provided by researchers: Baggerly et al (2005) speculate that the serum biomarkers proposed for ovarian cancer diagnosis are likely to be driven by "procedural bias", but the

original authors did not report sufficient metadata for them to demonstrate this. We hope that by highlighting the importance of balancing covariates between technical batches we can also facilitate better reporting of technical metadata. If secondary analyses will be conducted directly upon the covariates of interest, either by the original authors or through the reuse of the data by others, it is critical that they are not confounded by the technical batch effect.

There are some limitations to the method that we have utilised to allocate samples in a manner which balances covariates. Our definition of covariate balance means that it is a function not only of the effect size (for continuous variables) or frequency distributions (for categorical variables) but also the sample size. This means that in cases where the number of samples per batch is very small, our statistical power will be low and that therefore a covariate could appear balanced even whilst other measures (such as the standardised mean difference for continuous variable) could be very different between batches. For this reason we seek to optimise a continuous objective function, and do not recommend performing hypothesis testing to classify layouts as balanced or unbalanced, similar to the discussion present in Imai, King and Stuart (2008) and Ho *et al*. (2007).

We envision that there are several additional features that could be incorporated into our tool in order to improve and extend its functionality. Clinical data often encompass a large number of covariates, and it can be difficult to decide how many of those that are potentially of interest can be accounted for when allocating samples to batches. We currently require that these covariates are specified, but the decision of how many covariates to include could be data-driven, based upon a specified rank order of importance. Secondly, experimental workflows often involve several batch variables, for example sample preparation being performed by different people and then those batches of samples being analysed at different times. We currently only allow a single batch variable to be specified, but it could be useful to allow for multiple nested batch variables. Lastly, some assay technologies also result in positional effects within batches such as those that occur in plate based assays (Kricka *et al*. 1980) but also recently reported to be a potential source of technical variation resulting from sample preparation in proteomics (Maxwell *et al*. 2023). An analogous phenomenon can occur in assays which are not microplate based, such as the “spillover” between different channels in a tandem mass tag proteomics data (Brenes *et al*. 2019). Our current method is agnostic to position within a specified batch. We encourage others in the community to contribute to the open source code in order to implement these features.

Acknowledgements

We would like to thank Konstantin Kahnert for his feedback on an early version of this tool.

Funding

JM is supported by the BRIDGE – Translational Excellence Programme, funded by the Novo Nordisk Foundation (grant agreement no. NNF20SA0064340). AL is funded by the Novo Nordisk Foundation (NNF20OC0059767) and Independent Research Fund Denmark (10.46540/4285-00115B).

Data and code availability

The SampleAllocateR package is freely and openly available under the GPL-3.0 license: doi.org/10.5281/zenodo.14860111. The code necessary to reproduce the analysis in this paper is found at: doi.org/10.5281/zenodo.14859955.

Figure legends

Figure 1. Overview of the strategy employed to allocate samples into batches with covariate balance. A. Schematic of the algorithm. Input data required for a pre-selected set of samples are the set of relevant covariates chosen by the user (categorical and/or continuous), the batch size and optionally a blocking variable. The method is initialised by randomly allocating samples into batches. A balance score is calculated that is indicative of the balance of all covariates, and simulated annealing (a heuristic optimisation method) is used to efficiently search for the allocation with the highest possible balance. **B.** For a single variable, its balance across many random layouts it is expected to be uniformly distributed, meaning that few layouts are imbalanced (i.e. with a low probability that they are distributed evenly across batches) **C.** When considering the balance across many variables, we combine the balance of individual covariates by taking their harmonic mean to combine them without assuming their independence. There are few balanced layouts (with a high *balance score* indicating that they are distributed evenly across batches). **D.** Our method produces an allocation with more covariate balance on an example dataset than "random search" methods that are employed elsewhere. **E.** Computational runtime is reduced by employing simulated annealing, despite the improved performance.

References

Akey JM, Biswas S, Leek JT *et al.* On the design and analysis of gene expression studies in human populations. *Nat Genet* 2007;**39**:807–8.

Alseekh S, Aharoni A, Brotman Y *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* 2021;**18**:747–56.

Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 2009;**28**:3083–107.

Baggerly KA, Morris JS, Edmonson SR *et al*. Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer. *JNCI: Journal of the National Cancer Institute* 2005;**97**:307–9.

Box GEP, Hunter JS, Hunter WG. *Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition | Wiley*. https://www.wiley.com/en-dk/Statistics+for+Experimenters%3A+Design%2C+Innovation%2C+and+Discovery%2C+2nd+Edition-p-9780471718130 (March 4, 2024, date last accessed)

Brenes A, Hukelmann J, Bensaddek D *et al*. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. *Mol Cell Proteomics* 2019;**18**:1967–80.

Brusco MJ, Cradit JD, Steinley D. Combining diversity and dispersion criteria for anticlustering: A bicriterion approach. *British Journal of Mathematical and Statistical Psychology* 2020;**73**:375–96.

Carry PM, Vigers T, Vanderlinden LA *et al*. Propensity scores as a novel method to guide sample allocation and minimize batch effects during the design of high throughput experiments. *BMC Bioinformatics* 2023;**24**:86.

Cochran WG, Rubin DB. Controlling Bias in Observational Studies: A Review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 1973;**35**:417–46.

Downey JM, Cohen MV. Why Do We Still Not Have Cardioprotective Drugs? *Circulation Journal* 2009;**73**:1171–7.

Dunn WB, Broadhurst D, Begley P *et al*. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* 2011;**6**:1060–83.

Ferdinandy P, Schulz R, Baxter GF. Interaction of Cardiovascular Risk Factors with Myocardial Ischemia/Reperfusion Injury, Preconditioning, and Postconditioning. *Pharmacol Rev* 2007;**59**:418–58.

Ferdinandy P, Szilvassy Z, Baxter GF. Adaptation to myocardial stress in disease states: is preconditioning a healthy heart phenomenon? *Trends in Pharmacological Sciences* 1998;**19**:223–9.

Forshed J. Experimental Design in Clinical ‘Omics Biomarker Discovery. *J Proteome Res* 2017;**16**:3954–60.

Ho DE, Imai K, King G *et al*. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 2007;**15**:199–236.

Hu J. The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales. *Briefings in Functional Genomics and Proteomics* 2005;**3**:322–31.

Imai K, King G, Stuart EA. Misunderstandings Between Experimentalists and Observationalists about Causal Inference. *Journal of the Royal Statistical Society Series A: Statistics in Society* 2008;**171**:481–502.

Kircher M, Heyn P, Kelso J. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* 2011;**12**:382.

Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. *Science* 1983;**220**:671–80.

Kricka LJ, Carter TJ, Burt SM *et al*. Variability in the adsorption properties of microtitre plates used as solid supports in enzyme immunoassay. *Clinical Chemistry* 1980;**26**:741–4.

Leek JT, Scharpf RB, Bravo HC *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.

Lun ATL, Marioni JC. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* 2017;**18**:451–64.

Marioni JC, Mason CE, Mane SM *et al*. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17.

Maxwell CB, Sandhu JK, Cao TH *et al*. The Edge Effect in High-Throughput Proteomics: A Cautionary Tale. *J Am Soc Mass Spectrom* 2023;**34**:1065–72.

Mulvey J. SampleAllocateR. 2024.

Papenberg M. K-Plus anticlustering: An improved k-means criterion for maximizing between-group similarity. *British Journal of Mathematical and Statistical Psychology* 2024;**77**:80–102.

Papenberg M, algorithm) MM (centroid based clustering, Klau GW *et al*. anticlust: Subset Partitioning via Anticlustering. 2023.

Perel P, Roberts I, Sena E *et al*. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ* 2007;**334**:197.

Pincus M. Letter to the Editor—A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems. *Operations Research* 1970;**18**:1225–8.

Pound P, Ritskes-Hoitinga M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *Journal of Translational Medicine* 2018;**16**:304.

Sancho LR, Lozano J-J, Salas A. experDesign: stratifying samples into batches with minimal bias. *Journal of Open Source Software* 2021;**6**:3358.

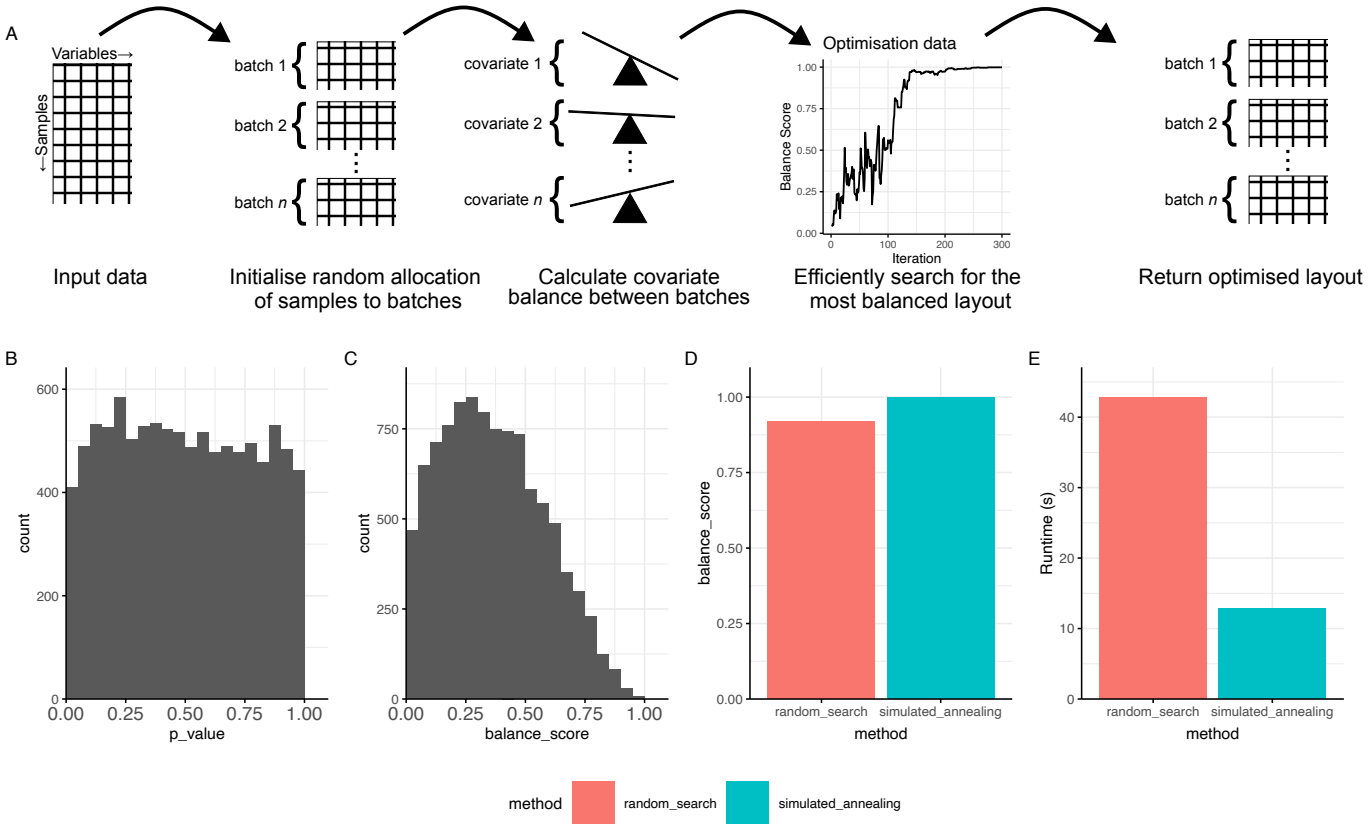
Schirmer M, Ijaz UZ, D'Amore R *et al*. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* 2015;**43**:e37.

Sinke L, Cats D, Heijmans BT. Omixer: multivariate and reproducible sample randomization to proactively counter batch effects in omics studies. *Bioinformatics* 2021, DOI: 10.1093/bioinformatics/btab159.

The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;**437**:1299–320.

Wilson DJ. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* 2019;**116**:1195–200.

Yan L, Ma C, Wang D *et al*. OSAT: a tool for sample-to-batch allocations in genomics experiments. *BMC Genomics* 2012;**13**:689.


A
Variables→
←Samples
Input data
batch 1
batch 2
batch n
Initialise random allocation of samples to batches
covariate 1
covariate 2
covariate n
Calculate covariate balance between batches
Optimisation data
Balance Score
Iteration
Efficiently search for the most balanced layout
batch 1
batch 2
batch n
Return optimised layout
B
count
p_value
C
count
balance_score
D
balance_score
random_search
simulated_annealing
method
E
Runtime (s)
random_search
simulated_annealing
method
method
random_search
simulated_annealing