# Chapter 2

# Laboratory Experimental Design for a Glycomic Study

Ivo Ugrina, Harry Campbell, and Frano Vučković

## Abstract

Proper attention to study design before, careful conduct of procedures during, and appropriate inference from results after scientific experiments are important in all scientific studies in order to ensure valid and sometimes definitive conclusions can be made. The design of experiments, also called experimental design, addresses the challenge of structuring and conducting experiments to answer the questions of interest as clearly and efficiently as possible.

**Key words** Randomization, Blocking, Replication, Experimental design

## 1 Introduction

The general principles of study design and analysis of 'omics studies (including glycomic studies) in epidemiology research is covered in a number of recent reviews [1]. In addition, these principles have been formulated as reporting guidelines to ensure that key aspects of the study which aid interpretation and review are reported. These also ensure that key data are presented in a standard format in order to promote data synthesis in systematic reviews. Examples of these reporting guidelines include:

1. STROBE ME—STrengthening of Reporting of Observational studies in Epidemiology: Molecular Epidemiology studies [2]
2. STARD—STAndards for Reporting Diagnostic studies [3]
3. REMARK—Guidelines for REporting tumor MARKers [4] and
4. GRIPS—Guidelines for the reporting of Genetic Risk Prediction Studies. [5]

However, much less has been published on the detailed design of laboratory procedures to ensure valid and reliable 'omic data are generated for analysis. This review focuses on this important aspect of glycomic study design.

Gordan Lauc and Manfred Wuhrer (eds.), *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1503, DOI 10.1007/978-1-4939-6493-2_2, 

The main aim of the high-throughput analysis is to analyse very large numbers of samples in a cost-efficient manner and in a relatively short time. Nevertheless, the very large number of samples often necessitates that an experiment lasts several weeks. This can lead to changes over time in the setup of a lab analysis (e.g., change of column in a UPLC machine) of glycans. These changes may distort later results leading to some variables falsely appearing to be correlated, i.e., leading to bias and/or confounding. These are not necessarily an artifact of changes within the laboratory, since samples usually come in batches and these problems may exist even before they enter the lab (e.g., bias introduced due to cases being in one batch, controls in another). Variables affecting results and possibly leading to bias and/or confounding are often called *nuisance factors*. An effective approach to reduce or even eliminate the effects of nuisance factors can be achieved with a proper application of the theory of experimental design.

The "design of experiments" was first described by Ronald A. Fischer in 1920 [6] to improve agricultural work and results. Although it was originally developed for agriculture, the main ideas and methods have since been applied in numerous fields and are therefore called the fundamental principles of the experimental design. The three most important principles for experimental design, relevant to high-throughput glycomics experiments, are:

1. Randomization
2. Blocking
3. Replication

Randomization is a method that guards against unknown nuisance factors affecting the results of the experiment. An example of a bias that can be introduced is a change in an instrument used for chemical analysis (e.g., change of a column in a UPLC machine). If all control samples from a case–control study are analysed first and then all the case samples are analysed subsequently, the observed difference between analytical results could be due to the instrument change. In the worst case scenario, the change would occur during the time between analyses of control and case samples. However, if samples are run in a random case–control order (e.g., case, control, control, case, case, …) then any change in the instrument during the experiment should equally affect both the cases and controls and not lead to bias.

With known nuisance factors the blocking method can be applied to increase the precision of results and aid in future analysis. An obvious example of a possible nuisance factor is the batch proportion of cases and controls in a case–control study. In a blocked design, samples should be measured with the same ratio of cases and controls within every batch as within the whole population involved in the analysis. In experiments where such blocking does not occur, any apparent between-batches changes between controls and cases

could be due to batch effects rather than true differences between cases and controls. Other less obvious nuisance factors have been described in the literature, such as effects of gender [7]. A good rule for the design of a high throughput laboratory experiment is to block what you can and randomize what you cannot block [8]. Therefore, both blocking and randomization should be employed.

Replication is a method which acknowledges that there are sources of variability both between runs and (potentially) within runs and thus that replication is required to account for this. A replicate is a complete repetition of the same experimental conditions, beginning with the initial setup. Replicates in high-throughput glycomics may be achieved through two types of technical replicates: technical replicates of biological samples of interest (in future denoted as just *replicates*) and technical replicates as a special, usually in-house, sample to be used within all batches (in future denoted as *standards*). The importance of replicates comes from the idea that if everything in the experiment went perfectly then values for replicates should be the same. It is important to stress here that replicates are performed within an experiment and do not denote a special type of replication where the whole experiment is replicated in a larger sample size. Changes in results between replicates indicate the level of variability of the instrument (if samples were pooled before entering the instrument) or different internal (procedural) steps (if samples were pooled before a specific step) and can point to non-systematic changes possibly revealing previously unobserved nuisance factors.

## 2 Materials

For a proper experimental design it is important to obtain as much detailed knowledge of the study and information on known nuisance factors as possible. Thus, experimental design should be derived in collaboration between wet- and dry-labs.

Some of the known *nuisance factors* that are generally applicable to all human glycomics analysis are:

1. Age [7].
2. Gender [9].
3. Geographical location (Continent, State, Region, …).

Other known nuisance factors are more dependent on the underlying study and data on these are often hard to obtain. Examples are:

1. Case–control designation.
2. Batches in sample acquisition (e.g., samples could have been acquired village by village introducing possibly high genetic/location bias).

3. Sample acquisition dates.
4. Number of freeze–thaw cycles (e.g., newly obtained samples vs. old samples thawed many times).
5. Information on sample acquisition centers (e.g., studies combining samples from different hospitals).

## 3 Methods

Although choice of the most appropriate study design is highly dependent on the available data, the main ideas can be presented through four different approaches (with additional information given in **Notes 3–6**).

### *3.1 Cohort Study Where No Additional Information Is Available*

This is an example of a study where a laboratory is asked to analyze glycosylation of a protein in a cohort study where the only data that can be shared are samples and sample names. Since there are no additional data on samples blocking cannot be applied. The following procedure can be used:

1. Decide if replicates are needed based on previous observations (e.g., systematic or non-systematic error).
2. Decide on the number of replicates and standards needed in the study. This decision should be based on cost–benefit analysis taking into account that larger numbers of replicates and standards increase time and budget costs while decreasing error.
3. Randomly assign standards to plates.
4. Select replicates randomly.
5. Randomly assign replicates to plates.
6. Randomly assign other samples to plates (*see* **Note 1**).

### *3.2 Case–Control Study Where No Additional Information Is Available*

This is an example of a study where a laboratory is asked to analyze the glycosylation of a protein for a case–control cohort where the only data that can be shared are samples, sample names and case–control designation. Since there are additional data on samples blocking can be applied. The following procedure can be used:

1. Decide if replicates are needed based on previous observations (e.g., systematic or non-systematic error).
2. Decide on the number of replicates and standards needed in the study. This decision should be based on cost–benefit analysis taking into account that larger numbers of replicates and standards increase time and budget costs while decreasing error.
3. Randomly assign standards to plates.
4. Select replicates randomly with case–control ratio preserved as within the whole cohort (*see* **Note 2**).

5. Randomly assign replicates to plates.
6. Randomly assign other samples to plates with the case–control ratio preserved in plates as within the whole cohort (*see* **Note 2**). This can be achieved by randomly selecting appropriate number of samples from cases first and then appropriate number of samples from controls. The approach should be repeated plate by plate (*see* **Note 1**).

### 3.3 Cohort Study Where Age and Gender Data Is Available

This is an example of a study where a laboratory is asked to analyze the glycosylation of a protein for a cohort where the only data that can be shared are samples and sample names together with age and gender. Since there are additional data on samples blocking can be applied. The following procedure can be used:

1. Decide if replicates are needed based on previous observations (e.g., systematic or non-systematic error).
2. Decide on the number of replicates and standards needed in the study. This decision should be based on cost–benefit analysis taking into account that larger numbers of replicates and standards increase time and budget costs while decreasing error.
3. Randomly assign standards to plates.
4. Select replicates randomly with gender ratio and age distribution preserved as within the whole cohort (*see* **Note 2**).
5. Randomly assign replicates to plates.
6. Randomly assign other samples to plates with the gender ratio and age distribution preserved in plates as within the whole cohort (*see* **Note 2**). This can be achieved by randomly selecting appropriate number of samples from females first and then the appropriate number of samples from males. The approach should be repeated plate by plate (*see* **Note 1**). Since the random selection of samples from males and females could result in different age distributions the procedure can be repeated until more balanced results are obtained.

### 3.4 Case–Control Study Where Age and Gender Data Is Available

This is an example of a study where a laboratory is asked to analyze the glycosylation of a protein for a cohort where the only data that can be shared are samples, sample names, and case–control designation together with age and gender. Since there are additional data on samples blocking can be applied. The following procedure can be used:

1. Decide if replicates are needed based on previous observations (e.g., systematic or non-systematic error).
2. Decide on the number of replicates and standards needed in the study. This decision should be based on cost–benefit analysis taking into account that larger numbers of replicates and standards increase time and budget costs while decreasing error.

3. Randomly assign standards to plates.
4. Select replicates randomly with gender ratio, case–control ratio, and age distribution preserved as within the whole cohort (*see* **Note 2**).
5. Randomly assign replicates to plates.
6. Randomly assign other samples to plates with gender ratio, case–control ratio, and age distribution preserved in plates as within the whole cohort (*see* **Note 2**). This can be achieved by randomly selecting appropriate number of samples from joint distributions of male/case, male/control, female/case, and female/control groups. The approach should be repeated plate by plate (*see* **Note 1**). Since the random selection of samples from aforementioned four groups could result in different age distributions, the procedure can be repeated until more balanced results are obtained.

## 4 Notes

1. If a change in experimental design (plate layout) happens for a reason (e.g., not enough sample in a vial) consult the person who has derived the initial plate/experimental design. In the case of a missing sample a new one can sometimes be found conforming to the current design (blocking, randomization).
2. A perfect (equal) distribution between plates is hard to achieve when controlling (blocking) many factors. Sometimes it is even impossible to achieve it. Therefore, "good enough" (in an expert view) designs should be used.
3. Appropriate software tools are of great use in deriving experimental designs since designs derived by hand can be quite time consuming.
4. If there is a plate with many samples missing or not measured well enough (seen from the consequent quality control) this plate should be taken into consideration for exclusion from the study since its distribution (case–control, gender, age) of nuisance factors could be different from the rest of the experiment.
5. Try to avoid repeating samples that did not pass quality control on a new plate without consulting the person who derived the initial design or at least looking at the distribution of nuisance factors of the failed samples. It could be that these samples could have a completely different distribution from the initial design and could therefore introduce problems in later data analysis.
6. More information on the theory of Experimental Design can be found in books specialized for the topic [8].

## References


1. Tzoulaki I, Ebbels TM, Valdes A, Elliott P, Ioannidis JP (2014) Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. Am J Epidemiol 180:129–139
2. Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, Kirsch-Volders M, Matullo G, Phillips DH, Schoket B, Stromberg U, Vermeulen R, Wild C, Porta M, Vineis P, STROBE Statement (2011) STrengthening the Reporting of OBservational studies in Epidemiology—Molecular Epidemiology (STROBE-ME): an extension of the STROBE Statement. PLoS Med 8:e1001117
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, STARD Group (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 351:h5527
4. Altman DG, McShane LM, Sauerbrei W, Taube SE (2012) Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. PLoS Med 9:e1001216
5. Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ, GRIPS Group (2011) Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Clin Invest 41:1004–1009
6. Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
7. Krištić J, Vučković F, Menni C, Klarić L, Keser T, Beceheli I, Pučić-Baković M, Novokmet M, Mangino M, Thaqi K, Rudan P, Novokmet N, Šarac J, Missoni S, Kolčić I, Polašek O, Rudan I, Campbell H, Hayward C, Aulchenko Y, Valdes A, Wilson JF, Gornik O, Primorac D, Zoldoš V, Spector T, Lauc G (2014) Glycans are a novel biomarker of chronological and biological ages. J Gerontol A Biol Sci Med Sci 69(7):779–789
8. Anderson MJ, Whitcomb PJ (2015) DOE simplified: practical tools for effective experimentation. CRC Press, New York
9. Knezevic A, Gornik O, Polasek O, Pucic M, Redzic I, Novokmet M, Rudd PM, Wright AF, Campbell H, Rudan I, Lauc G (2010) Effects of aging, body mass index, plasma lipid profiles, and smoking on human plasma N-glycans. Glycobiology 20:959–969

http://www.springer.com/978-1-4939-6491-8

High-Throughput Glycomics and Glycoproteomics
Methods and Protocols
Lauc, G.; Wuhrer, M. (Eds.)
2017, XI, 274 p. 51 illus., 42 illus. in color. With online files/update., Hardcover
ISBN: 978-1-4939-6491-8
A product of Humana Press