Designing a Metabolomics Experiment

Xiangqin Cui, PhD

Experimental Design

• Experimental design: is a term used about efficient methods for planning the collection of data, in order to obtain the maximum amount of information for the least amount of work. Anyone collecting and analyzing data, be it in the lab, the field or the production plant, can benefit from knowledge about experimental design. http://www.stat.sdu.dk/matstat/Design/index.html

• Good experimental design is the foundation for valid answer to research questions
Key Questions for Experimental Design

• What are the research questions to be answered?
• How will the data be analyzed?

• What is the best design of the experiment to answer the questions using the analysis methods?

General Statistical Principals of Experimental Design

• Replication
• Randomization
• Blocking
• Use of factorial experiments instead of the one-factor-at-a-time methods.
• Orthogonality
Replication

• **Replication** is repeating the creation of a phenomenon (or redo your experiment), so that the variability associated with the phenomenon can be estimated. (no replication, no way to know the variability)

Replications should not be confused with repeated measurements which refer to taking several measurements of a single occurrence of a phenomenon.
More terms saying the same things

• What to replicate?
  • Biological replicates (replicates at the experimental unit level, e.g. mouse, plant, pot of plants...)
    • Experimental unit is the unit that the experiment treatment or condition is directly applied to, e.g. a plant if hormone is sprayed to individual plants; a pot of seedlings if different fertilizers are applied to different pots.
  • Technical replicates
    • Any replicates below the experimental unit, e.g. different leaves from the same plant sprayed with one hormone level; different seedlings from the same pot; Different aliquots of the same RNA extraction; multiple arrays hybridized to the same RNA; multiple spots on the same array.
Randomization

• The experimental treatments are assigned to the experimental units (subjects) in a random fashion. It helps to eliminate effect of "lurking variables", uncontrolled factors which might vary over the length of the experiment. Randomization is essential for making causal inferences.

• How do you look for information on randomization in a paper?

Commonly used randomization method

• Number the objects to be randomized and then randomly draw the numbers using paper pieces in a hat or computer random number generator, such as the one at https://www.random.org/.

Example: Assign two treatments, Hormone and control, to 6 plants

Hormone treatment: (1,3,4); (1,2,6)
Control: (2,5,6); (3,4,5)
Another Example

- Number the objects to be randomized and then randomly draw the numbers.

Example: Assign two treatments, Special Diet and control, to 6 mice

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Special Diet : 1, 3, 4
Control : 2, 5, 6

Blocking

- Some identified uninteresting but varying factors can be controlled through blocking.

- COMPLETELY RANDOMIZED DESIGN
- COMPLETE BLOCK DESIGN
- INCOMPLETELY BLOCK DESIGNS
Completely Randomized Design

There is no blocking

Example

- Compare two hormone treatments (trt and control) using 6 Arabidopsis plants (or mice or human).

Complete Block Design

- There is blocking and the block size is equal to the number of treatments.

Example:

- Compare two hormone treatments (trt and control) using 6 Arabidopsis plants. For some reason plant 1 and 2 are taller, plant 5 and 6 are thinner.

Randomization within blocks
Incomplete Block Design

- There is blocking and the block size is smaller than the number of treatments.

Example:

- Compare three hormone treatments (hormone level 1, hormone level 2, and control) using 6 Arabidopsis plants. For some reason plant 1 and 2 are taller, plant 5 and 6 are thinner.

```
Hormone level 1:   (1,4); (2,4)
Hormone level 2:   (2,5); (1,6)
Control:          (3,6); (3,5)
```

⇒ Randomization within blocks

Let’s look at the applications of these principals in metabolomics studies
General Statistical Principals of Experimental Design

- Replication
- Randomization
- Blocking
- Use of factorial experiments instead of the one-factor-at-a-time methods.
- Orthogonality

Replication in Metabolomics Experiments

- Looks for “replicates”, “sample size”, “samples per group” in publications.
- Number of replication can go from a few to tens, but rarely hundreds.
- The larger the number of replicates the better, but budget is always limited.
- One sample per treatment/condition is not OK. What is wrong?
Replication in Metabolomics Experiments

• Biological replicates are typically more important than technical replicates unless estimating the variation at different levels is the purpose of the experiment in evaluating the technology.

• Biological replicates are often more effective in increasing the power for detecting differential metabolites/genes.

• Technical replicates are useful when technical variability is large and technical replicates are cheap.

Sample Size and Statistical Power Calculation

• Sample size for a general two group comparison

\[
n = \frac{2(z_{(1-\alpha/2)} + z_{(1-\beta)})^2}{(\delta / \sigma)^2}
\]

- \( n \) increases as error, \( \sigma \), increases.
- \( n \) increases as the difference between two means, \( \delta \), decreases.
- \( n \) increases as the significant level of the test, \( \alpha \), decreases.
- \( n \) increases as the power of the test, \( 1-\beta \), increases.
General Statistical Principals of Experimental Design

- Replication
- Randomization
- Blocking
- Use of factorial experiments instead of the one-factor-at-a-time methods.
- Orthogonality
UMSA Analysis

Insulin Resistant

Insulin Sensitive

Components: ONE TWO THTREE

UMSA Analysis

Day 1

Day 2

Components: ONE TWO THTREE
How to Solve This Problem?

• Process all samples in the same day.
• Process half sensitive samples and half resistant samples in each day (balance sample groups against days—“treat each day as a block” in statistical terms).

Strategy 1

All samples

1 Day

Strategy 2

½ Resistant ½ Sensitive ½ Resistant ½ Sensitive

Day 1 Day 2

Known sources of non-biological biases (not exhaustive) that must be addressed

• Technician / post-doc
• Reagent lot
• Temperature
• Protocol
• Date
• Location
• Cage/ Field positions
Too Many factors to balance? -- Randomize

• Number the objects to be randomized and then randomly draw the numbers.

Example: Assign two treatments, Special Diet and control, to 6 mice

1              2 3 4 5 6

Special Diet : 1, 3, 4
Control : 2, 5, 6

Too Many factors to balance? -- Randomize

• Randomize samples in respect to treatments
• Randomize the order of handling samples.
• Randomize batches/runs/days in respect to samples
• Randomize over any other variable procedures.
Can I pool my treatment samples?

- It is rarely recommended unless it is necessary, e.g., working with fruit flies.
- It has potential benefits (reduce biological variability) and drawbacks (lack of measure of variability across individuals).
- Definitely not pooling all your treatment samples into one big pool and your control samples into one big pool.

References


Thank you – Questions?