BIOS 7659 Homework 1

Tim Vigers

19 September 2020

# Case Studies

Compare resources and cost, variability, confounding, etc.

## Case Study 3

The aim of case study 3 is to identify subgroups of patients using cluster analysis and then build a model to differentiate between the clusters. Samples are taken from 60 patients with B-cell lymphomas.

### Experimental Design 1

Half of the samples are labeled with Cy3 and the other half with Cy5. Then one Cy3-labeled sample and one Cy5-labeled sample are hybridized to 30 different two-color arrays.

The major advantage of this design is reduced cost, because it only requires 30 arrays compared to the other two designs’ 60. However, there are serious issues with this design that outweigh the cost benefits. First, it is not clear how samples will be selected for each label or how they will be distributed across the arrays. There is a possibility for selection-based dye bias here, for example if the investigator tends to favor labeling cases with Cy3 and controls with Cy5.

Assuming that each array has a sample from one case and one control, and that the Cy3-labeled samples are even split between cases and controls, this is a balanced block design. Balanced block designs work well for performing direct comparisons between two groups, but do not work for cluster analyses. Because there is no reference sample, the investigator will not be able to make comparisons between samples on different arrays, due to natural variation between arrays. Without the balanced block design assumption, it might not even be possible to compare cases to controls.

### Experimental Design 2

Samples from each patient are labeled with Cy3 and hybridized to 60 two-color arrays. A universal reference sample labeled with Cy5 is also hybridized to each array.

This is design is obviously more expensive than Design 1, because it require twice the number of arrays and preparation of the global reference sample. However, it has some significant advantages compared to Design 1.

By including the same reference sample on each array, any sample can be compared to any other sample, and the analysis does not depend on comparing two specific groups. This makes the reference design particularly advantageous for cluster analysis and developing models that include information aside from group membership (e.g. disease stage of the patient).

### Experimental Design 3

Samples from each patient are hybridized to 60 different Affymetrix arrays.

This design has many of the same positives as Design 2, although it could potentially be more expensive than using “in-house” arrays made by the lab. Affymetrix arrays are generally comparable between samples, particularly if it is assumed that all the samples were taken at the same timepoint. Any differential hybridization can be dealt with using standard normalization techniques. One potential advantage over Design 2 is that it would be easier to compare results across labs if you want external validation of the model. This could potentially be done using Design 2, but only if both labs use the same reference sample.

## Case Study 4

Yeast was allowed to reproduce across seven timepoints, and the investigators are interested in identifying genes with similar expression profiles across time.

### Experimental Design 1

Each sample is hybridized to a single Affymetrix array.

Affymetrix chips are generally less favorable for time-series designs, especially with only one sample per plate. For example, if one of the arrays fluoresces more than the others, then it would be difficult to tell whether this is due to time-related changes in gene expression or is an artifact of differential hybridization. It’s possible that adjusting the data for background fluorescence would help address this problem, but it would be better to have more replicates if possible,

### Experimental Design 2

Samples from timepoints 1 - 6 are labeled with Cy3 and each is hybridized an array along with a Cy5-labeled reference sample from time 0.

This design is essentially the same as the reference design from the previous section, and would allow for valid comparisons between timepoints. Using time 0 as a reference should also allow for adjustment for differential hybridization (“bright” or “dark” chips).

An additional benefit to this design is that it uses 1 less chip than the other two and will therefore be less expensive. However, using time 0 as a reference an all the arrays requires a large sample at baseline.

### Experimental Design 3

Each sample is labeled once with Cy3 and once with Cy5, and are hybridized to 7 arrays in a loop design.

A major benefit of the loop design is that it only requires n arrays where n = the number of samples, although it does require two aliquots per sample. It allows each sample to be directly compared while adjusting for sample distributions, array effects, etc. However, comparisons between samples that are not adjacent in the loop require modeling the indirect effects of arrays that are in between the arrays of interest, which can be complex.

Compared to Design 2, this approach has the benefit that each sample is measured twice on two different arrays, which helps account for variability between arrays. The major disadvantages are that it requires an additional array (more expensive), and the loop could be broken if one array doesn’t work correctly.

# Sample Size Calculations

The sample size calculations described here assume that gene expression levels are approximately normally distributed within both the treated and non-treated groups. The null hypothesis that a gene is expressed equally in the two groups will be rejected based on a t test at significance level . This level represents the probability of a type 1 error (false positive), and is more strict than the common in order to account for the large number of genes that will be compared. Thus with a 20,000 probe set we would expect approximately 2 genes to be incorrectly identified as differentially expressed.

Statistical power is the probability of correctly rejecting the null hypothesis when the mean expression level of a gene is different between the two groups by . For these calculations we assume that a meaningful difference in standardized expression is on the base 2 logarithmic scale (a twofold difference), and consider two reasonable estimates of the standard deviation (SD) of gene expression for this study, and .

Based on these assumptions, we can solve the equation

for n. Calculations were done using the pwr package (Champely 2020) in the R programming language version 4.0.2. Cost reflects a per-array price of $1,000 and does not include reagents, technician salary, etc.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Power | SD | Total n | n per Group | Cost |
| 0.8 | 0.5 | 30 | 15 | $30,000 |
| 0.85 | 0.5 | 32 | 16 | $32,000 |
| 0.9 | 0.5 | 36 | 18 | $36,000 |
| 0.95 | 0.5 | 40 | 20 | $40,000 |
| 0.8 | 0.25 | 14 | 7 | $14,000 |
| 0.85 | 0.25 | 14 | 7 | $14,000 |
| 0.9 | 0.25 | 16 | 8 | $16,000 |
| 0.95 | 0.25 | 16 | 8 | $16,000 |

# Sample Size Comparisons

## a

Using pwr.t.test, what is the sample size needed based on , fold change of 2 () and standard deviation of to achieve power of at least 0.8 or 0.95?