1. Study design, resources and cost, variability, confounding.

Case Study 3

Experimental Design 1

This is not a good design. The aim of this study is to identify clinically relevant subgroups of patients (clustering method). Two different dyes, in this case, make it is much more difficult to compare between dye-groups. Also, it is difficult to compare between arrays. This design costs half the number of arrays but has less power. Only unrelated samples on the same array are comparable but has dye bias issue.

Experimental Design 2

This one is an appropriate design and costs 60 arrays. The reference sample on each array can serve as the quality control, and the baseline for normalization. In this way, it is meaningful to carry out comparison and clustering with less confounding and full power.

Experimental Design 3

This one is an appropriate design and costs 60 Affymetrix arrays. The Affymetrix array is designed in the way for comparison between samples (normalization). It also comes with corresponding statistical methods.

Case Study 4

Experimental Design 1

This is not appropriate. The normalization method comes with Affymetrix can not deal with global changes which is expected to see alongside the time course. It may remove all variances come from different time points. Thus, this design can not answer the scientific question.

Experimental Design 2

This is an appropriate and robust design. We have a cross-arrays reference sample for quality control and normalization. The intensity ratio on each array is good to compare between arrays.

Experimental Design 3

The loop design is also appropriate. What's more, each sample is tested twice in this design. However, compared with design 2, this one requires more complex data analysis and the loop is more fragile to one bad array occurred amid.

2.

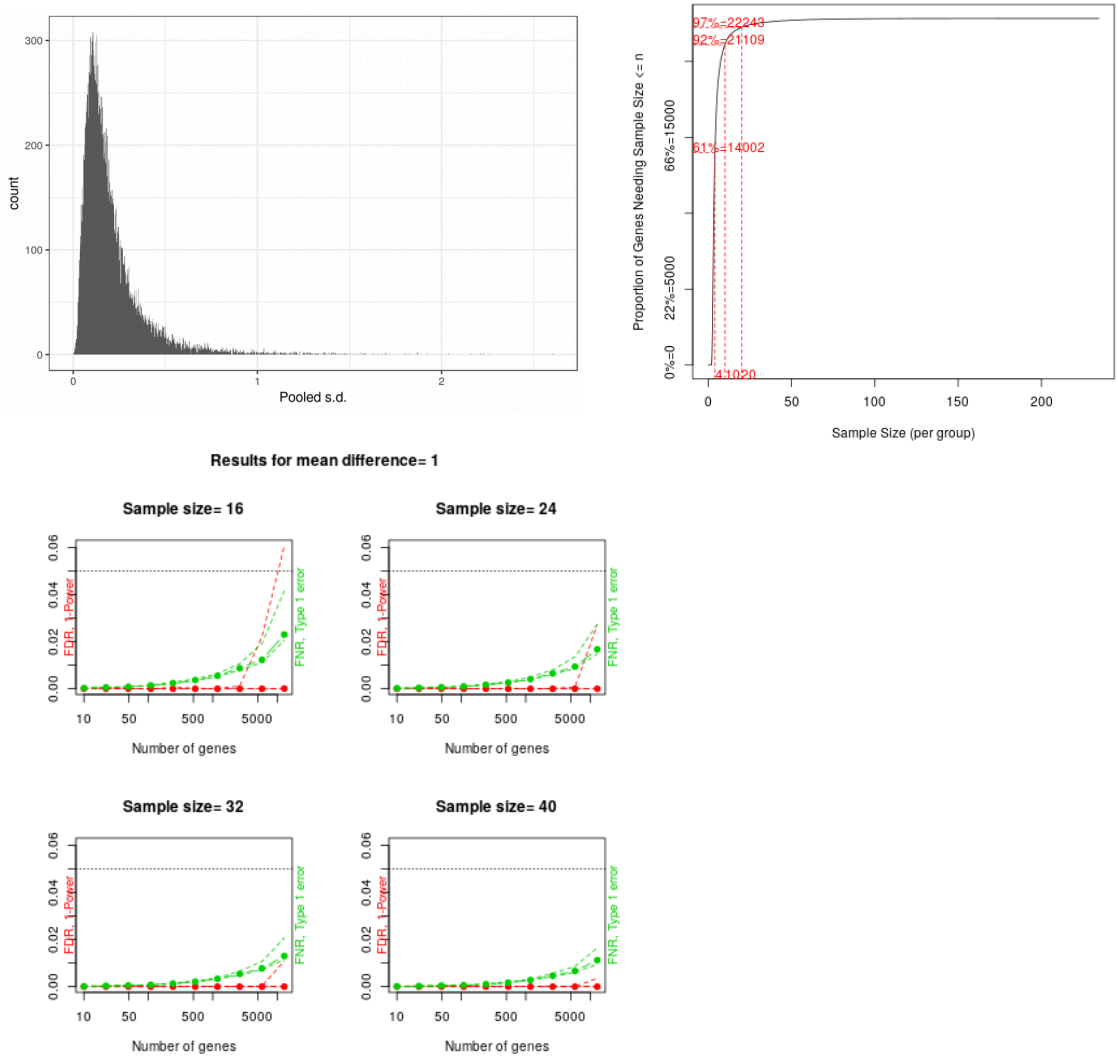
Table 1: Two-group Microarray Sample Size Summary

	Power Levels	Sample Size per Group
alpha = 0.001	0.8	11
delta = 1 (log2)	0.85	12
sigma = 0.5 (log2)	0.9	13
	0.95	14
	False Positive (No.)	20

$$n = \frac{4\left(z_{\alpha/2} + z_{\beta}\right)^2}{(\delta/\sigma)^2},$$

This study is to detect a two-fold difference (1 in log2 scale) between treated and untreated mice with a significance level of 0.001 through the 20,000 probe sets array method. The appropriate estimates of variability in log2 scale is 0.5. To achieve power levels 0.8, 0.85, 0.9 and 0.95 respectively, sample sizes per group 11, 12, 13 and 14 are required, respectively. The expected number of false positives is 20 per array.

The formula to calculate the sample size is from Design and Analysis of DNA Microarray Investigations, chapter 3, 3.2. This calculation is based on the assumption that the expression measurements are approximately normally distributed among samples of the same class in the unpaired t test framework. The z distribution here is the approximation of t distribution with the assumption that the n is sufficiently large.



The null hypothesis: log2(Untreated) – log2(treated) = 1
The alternative hypothesis: log2(Untreated) – log2(treated) != 1
The analysis is carried out with R (R version 3.4.4) and Rstudio (Version 1.1.456).

3. With a significance level of 0.001, a two-fold difference, a variance in log2 scale is 0.5

(a) Using pwr package, with a significance level of 0.001, a two-fold difference, a variance in log2 scale is 0.5, to achieve power levels 0.8 and 0.95, sample sizes per group 12 and 15 are required, respectively. There is no π_0 required in the unpaired two-sample t test framework.

(b) The π_0 here is the proportion of non-differentially expressed genes, relative conservative genes across groups. Another parameter is specific for high dimensional data is FDR. A FDR level, 0.05, instead of significant level, is specified in this method. To achieve power levels 0.8 and 0.95, with 80% unchanged genes, sample sizes per group 8 and 11 are required. When there are 90% unchanged genes, sample sizes per group increase to 9 and 12.

(c) The distribution of s.d. is centered roughly at 0.15, and a bell shaped curve with right skewed. This sample size calculation method does not use a single estimation of s.d. but a list of s.d. for all genes in an array. The s.d. in this data set is lower than the estimated s.d. in (a) and (b). In this case, a sample size of 10 per group is sufficient to carry out the study depends on the plots.

(e) The samr package estimates the sample size based on a pilot study of 4 samples per group in this case. From the plots, 8 samples per group is sufficient (power level > 0.95 and significant level < 0. 001) to detect about 1000 truly differential expressed genes.

Appendix