

Power and Sample Size Estimation for Microarray Studies

Maarten van Iterson

Microarray Analysis Group
Human Genetics
Leiden University Medical Center

MGC course 2011

Introduction: Sample size determination

Microarray: Sample size determination

Some examples

Other approaches to sample size determination

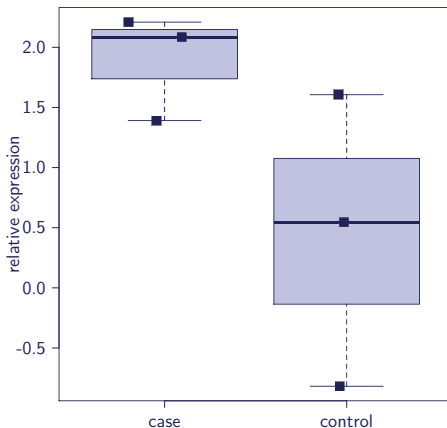
Summary

qPCR for a single gene

Is there expression difference between case and control for a specific gene of interest?

The null hypothesis H_0 :
There is no difference between case and control in expression for this specific gene.

(H_A : *There is differential expression between case and control.*)



Two-sample Student's t -test

means the same or difference close to zero ($1.90 - 0.445 = 1.45$)
spread of the data (0.441, 1.22)
sample sizes (3, 3)

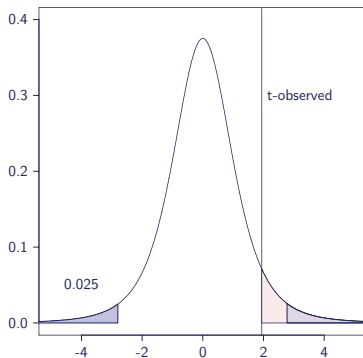
Two-sample Student's t -test:

t -observed = 1.94

- the test statistics measures discrepancy between data and H_0
- large values of t -observed (either positive or negative) is evidence against H_0
- the level of evidence against H_0 is measured by the p -value
- the *smaller* the p -value the *stronger* evidence against H_0

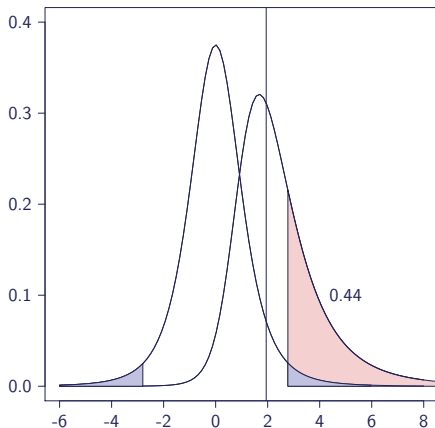
What is the evidence against H_0 given the data

- Under H_0 test statistics has a certain *null* distribution
- Using the *null* distribution calculate p-value = 0.12 (two-sided, $\alpha = 0.05$)
- Is there evidence for a certain alternative e.g. the difference is 2 (effect size)



What is the evidence for H_A

- power = 0.44 given the effect size 2 and significance level 0.05
- To have 50% power means that, if the gene is differentially expressed, we have on average 50% chance for finding it.
- There are many different alternatives thus many *alternative* distributions!

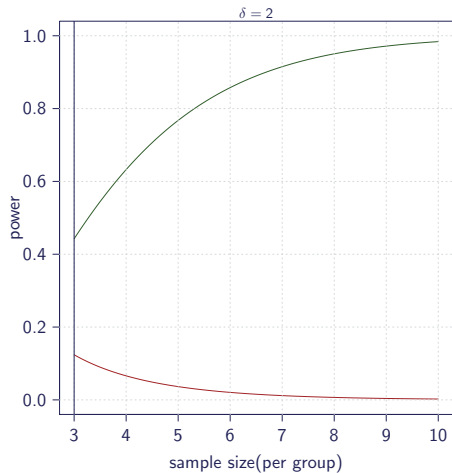
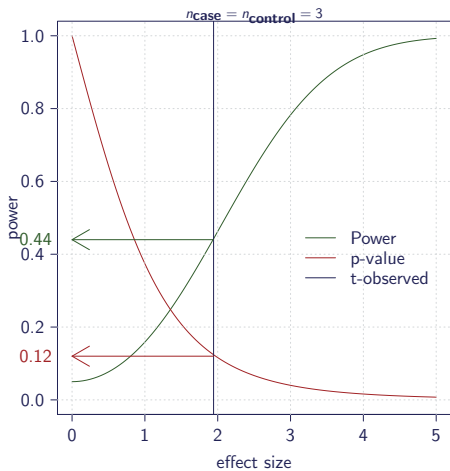


power is related to

- sample size
- effect size
- standard deviation
- significance level(α)

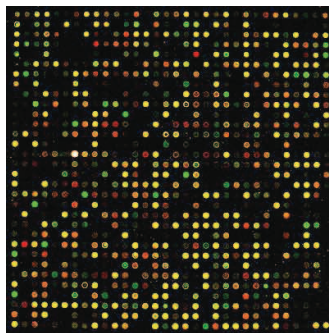
Knowing 4 of these the 5th can be determined. Or knowing 3 of them (say **standard deviation**, **significance level** and **effect size**) a 4th can be chosen (**sample size**) so as to maximize **power**.

Power curves



expression array for thousands of genes

- Measurements on several thousands of genes
- For each gene a hypothesis, test-statistic and p-value
- Want a single measure of power for the whole experiment



Ferreira and Zwinderman, SAGMB (2006) and Ferreira and Zwinderman, Int. J. Biostat. (2006)

Factors affecting the power

- sample size
 - number of samples
 - type of design
- fraction of non-differentially expressed genes.
 - How many genes are differentially expressed?
- distribution of effect sizes
 - What are the differences between tumor and normal samples for each gene?
- variation across samples in each condition
 - technical or biological

Variation across samples

Biological

- genetical
- environmental
- age, sex
- human, animal, cell lines

Technical

- platform, protocol, laborant

qPCR vs Microarray

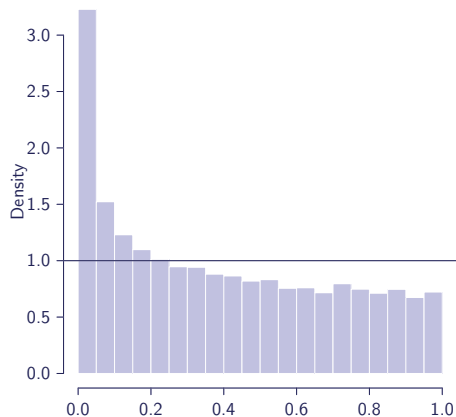
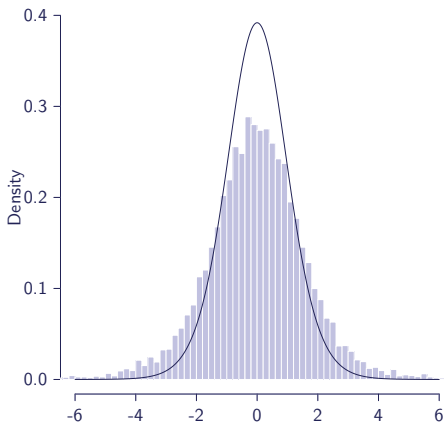
qPCR

- effect size
- sample variability
- significance level ($\alpha = 0.05$)

Microarray

- each gene has an effect size
 - each gene can have different variability
 - significance level for each gene
-
- A multiple testing problem: $\alpha = 0.05$ means that roughly one out of every twenty tests will be false positive
 - Only a proportion of the genes will be differentially expressed
 - Adaptive Benjamini-Hochberg approach for multiple testing correction

Test statistics and p-values



microarray experiment comparing 8 ApoA1 knockout mice with 8 normal C57BL/6

Power and sample size determination

1. Input:

- test statistics and p-values derived from pilot-data
- sample size of the pilot-data and significance level

2. Estimation:

- proportion on non-differentially expressed genes
- density of effect sizes

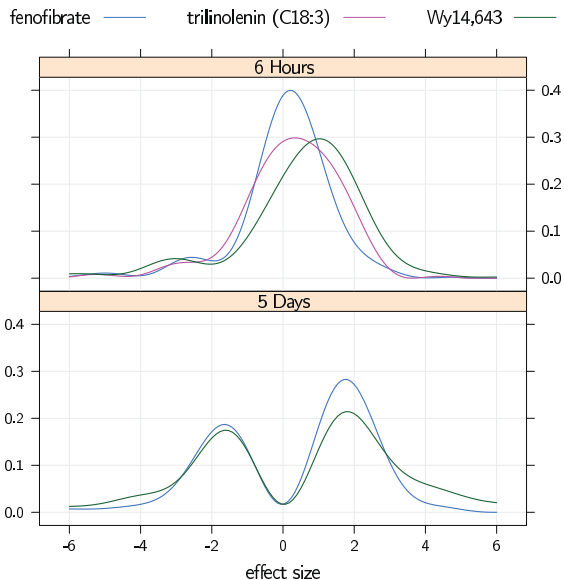
3. Output: power curve as function of the sample size

Nutrigenomics example

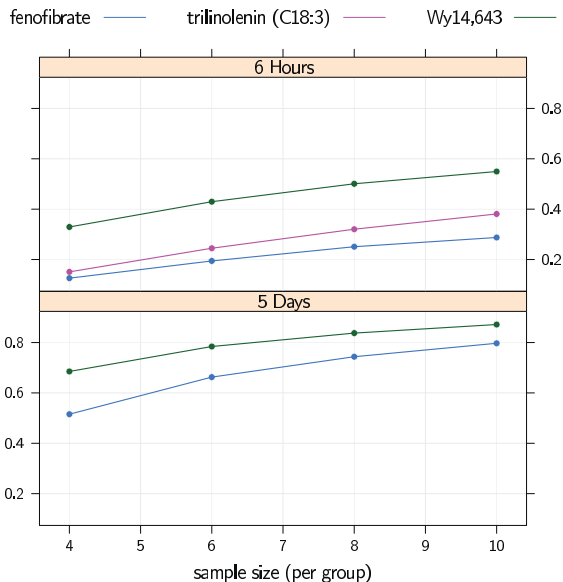
- PPAR- α activation in small intestine
- wild-type and PPAR- α knock out mice
- different PPAR- α agonist: high (Wy14,643), intermediate (trilinolenin or C18:3) and low (fenofibrate) potency
- different exposure times (6 hours and 5 days)
- Affymetrix GeneChip Mouse 430 2.0 arrays

	probe-sets	group A	group B	experiment
1	16539	4 (wild-type)	4 (knock-out)	high, 5 days
2	16539	4 (wild-type)	5 (knock-out)	intermediate, 6 hours
3	16539	5 (wild-type)	5 (knock-out)	low, 6 hours
4	16539	4 (wild-type)	4 (knock-out)	high, 5 days
5	16539	4 (wild-type)	4 (knock-out)	low, 5 days

Nutrigenomics example: density of effect sizes



Nutrigenomics example: power curves

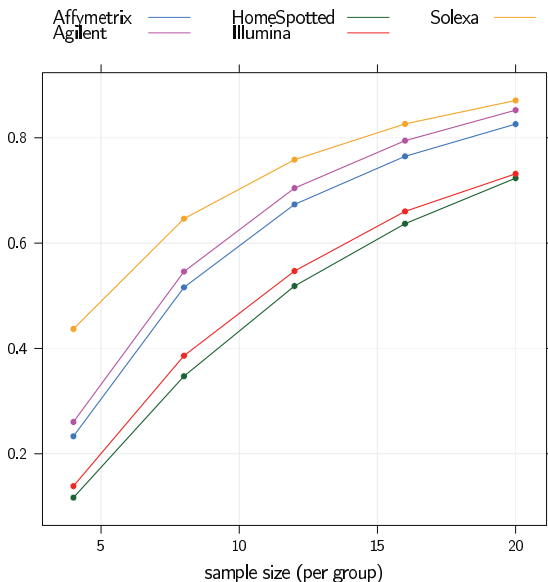


Platform comparison example

- gene expression profiles of hippocampi of δ C-doublecortin-like kinase transgenic mice were compared to wild-type mice
- four microarray technologies 2 two-colour and 2 single-colour and one next-generation sequencing technology

	platform	probe-sets	group A	group B
1	Affymetrix	45101	5 (wild-type)	5 (transgenic)
2	Agilent	41232	5 (wild-type)	5 (transgenic)
3	Illumina	46120	5 (wild-type)	5 (transgenic)
4	home-spotted	21771	5 (wild-type)	5 (transgenic)
5	Solexa	34477	4 (wild-type)	4 (transgenic)

Platform comparison example: power curves



Pilot data vs Simulation

Pilot data

- estimates from pilot-data
- effect sizes
- sample variability
- number of differentially expressed genes

Simulation

- prior knowlegde or guess
- effect sizes
- sample variability
- number of differentially expressed genes

Simulation based method implemented in `ssize`

Remember!

- Increasing the sample size, the power increases. Higher power requires larger sample sizes.
- As the effect-size grows bigger, the power increases.
- As the significance level gets smaller, the power decreases.
- If the standard deviation decreases the power increases.

References



M. Orr, P. Liu.

Sample size estimation while controlling false discovery rate for microarray experiments using `ssize.fdr` package.

The R Journal, 1, 1, May 2009.



van Iterson, M. 't Hoen, P.A.C. Pedotti, P. Hooiveld, G.J.E.J. den Dunnen, J.T. van Ommen, G.J.B. Boer, J.M. Menezes, R.X.

Relative power and sample size analysis on gene expression profiling data.

BMC Genomics, 2009.



SSPA:

<http://bioconductor.org/packages/release/bioc/html/SSPA.html>.