Power and Sample Size Estimation for Microarray Studies

Maarten van Iterson[1], José Ferreira[2], Judith Boer[3] and Renée Menezes[4]

[1]Center for Human and Clinical Genetics, Leiden University Medical Center, the Netherlands
[2]EMI-Stat&Mod, RIVM, the Netherlands
[3]Dept. of Pediatrics - Oncology and Hematology, Erasmus MC - Sophia Children's Hospital, the Netherlands
[4]Dept. of Epidemiology and Biostatistics, VUmc, the Netherlands

UseR! 2011

What is the appropriate sample size when testing many features simultaneously?

For example, measuring gene expression differences between groups using microarray or RNAseq.

Appropriate means: *When desired power is reached*.

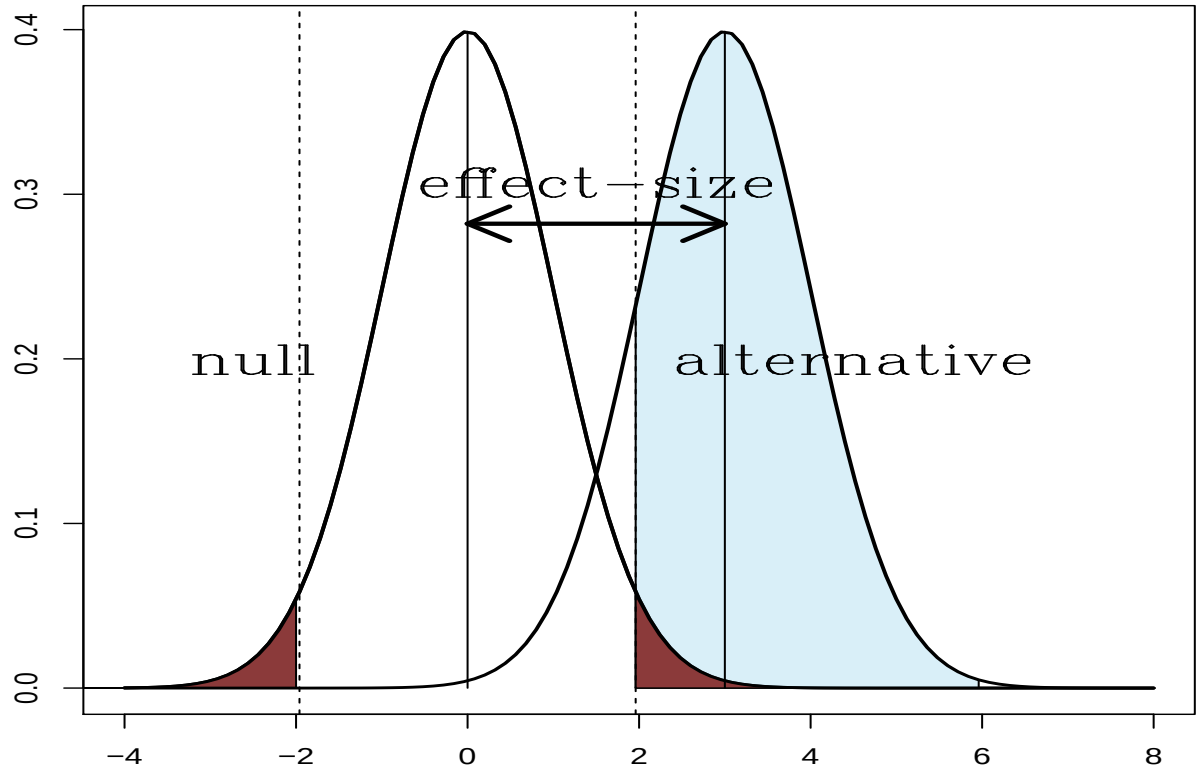
Power does not only depend on **sample size** but also on **effect size**, **sample variability** and **significance level**.

Sample size determination either simulation or **pilot-data** based.

Single hypothesis vs multiple hypotheses testing

- not a *single* rejection region but *many* (multiple testing problem)
- not a *single* effect size but *distribution* of effect sizes
- only a proportion will be rejected

average power: the proportion of correctly rejected observations



Histograms of observed test statistics (**A**) and p-values (**B**).

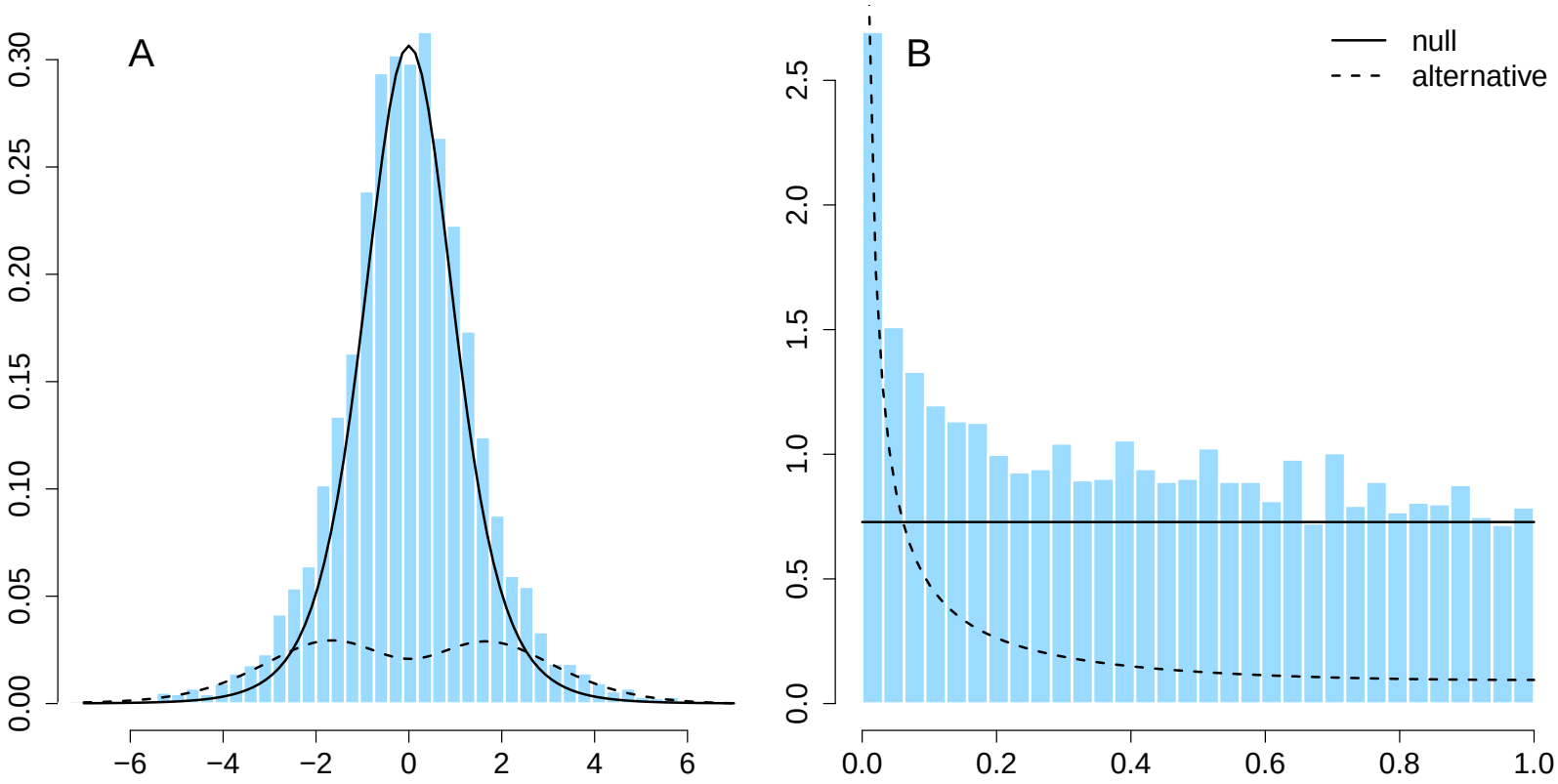


Figure: Parametric null distribution (solid) and estimated alternative distribution (dashed).

A mixture model for the probability distribution

$$m(t) = \pi_0 f_0(t) + (1 - \pi_0) \int f_1(t, \theta; N) \lambda(\theta) d\theta. \qquad (1)$$

- $m(t)$: observed test statistics (given)

A mixture model for the probability distribution

$$m(t) = \pi_0 f_0(t) + (1 - \pi_0) \int f_1(t, \theta; N) \lambda(\theta) d\theta. \qquad (1)$$

- $m(t)$: observed test statistics (given)
- π_0: indicates the proportion of non-differentially expressed genes (unknown)

A mixture model for the probability distribution

$$m(t) = \pi_0 f_0(t) + (1 - \pi_0) \int f_1(t, \theta; N) \lambda(\theta) d\theta. \qquad (1)$$

- $m(t)$: observed test statistics (given)
- π_0: indicates the proportion of non-differentially expressed genes (unknown)
- $f_0(t)$: *Normal* or a *Student's t* distribution (known)

A mixture model for the probability distribution

$$m(t) = \pi_0 f_0(t) + (1 - \pi_0) \int f_1(t, \theta; N) \lambda(\theta) d\theta. \qquad (1)$$

- $m(t)$: observed test statistics (given)
- π_0: indicates the proportion of non-differentially expressed genes (unknown)
- $f_0(t)$: *Normal* or a *Student's t* distribution (known)
- $f_1(t, \theta; N)$: *Normal* with mean $\neq 0$ or non central t (known)
- $\lambda(\theta)$: density of effect sizes (unknown)

A mixture model for the probability distribution

$$m(t) = \pi_0 f_0(t) + (1 - \pi_0) \int f_1(t, \theta; N) \lambda(\theta) d\theta. \qquad (1)$$

- $m(t)$: observed test statistics (given)
- π_0: indicates the proportion of non-differentially expressed genes (unknown)
- $f_0(t)$: *Normal* or a *Student's t* distribution (known)
- $f_1(t, \theta; N)$: *Normal* with mean $\neq 0$ or non central t (known)
- $\lambda(\theta)$: density of effect sizes (unknown)
- N : represents the effective sample size; $(1/n_A + 1/n_B)^{-1}$ (given)

Estimation of the density of effect sizes (analytically)

$f_1(t, \theta; N)$ normally distributed leads to the following convolution

$$\int \Phi(t - \theta\sqrt{N})\lambda(\theta)d\theta \tag{2}$$

which can be solved using a kernel deconvolution estimator[1]

$$\lambda(\theta) = \frac{1}{2\pi}\int e^{-is\theta\sqrt{N}}\frac{\psi_w(s)\psi_{m_n}(s)}{\psi_{f_0}(s)}\,ds \tag{3}$$

- numerical approximation to the real-part(very time-consuming)
- using `fft`-function like implementation of the `density`-function(really fast)

[1] Ferreira and Zwinderman, SAGMB, (2006).

Generalization to any kind of statistics

approximate the integral by a summation:

$$m_n(t_i) = \pi_0 f_0(t_i) + (1 - \pi_0) \sum_{j=1}^{M} f_1(t_i, \theta_j) \lambda(\theta_j) \Delta\theta. \quad (4)$$

express the density of effect sizes as a sum of B-splines:

$$m_n(t_i) = \pi_0 f_0(t_i) + (1 - \pi_0) \sum_{j=1}^{M} f_1(t_i, \theta_j) \sum_{k=1}^{K} \alpha_k b_k(\theta_j) \, \Delta\theta. \quad (5)$$

Estimation of the density of effect sizes

the discretization transforms the integral equation to matrix equation: $y = X\beta$

X ill-conditioned - **no OLS-solution**
need regularization e.g. minimize $||y - X\beta||^2 + \lambda W(\beta)$

- constrained optimization[2,3] ($\int \lambda(\theta) d\theta = 1$ and $\lambda(\theta) > 0$).
- **ridge regression**

[2] Ruppert *et al.*, Biometrics, (2007)
[3] van de Wiel and In Kim, Biometrics, (2007)

Estimation of the proportion of non-differentially expressed genes

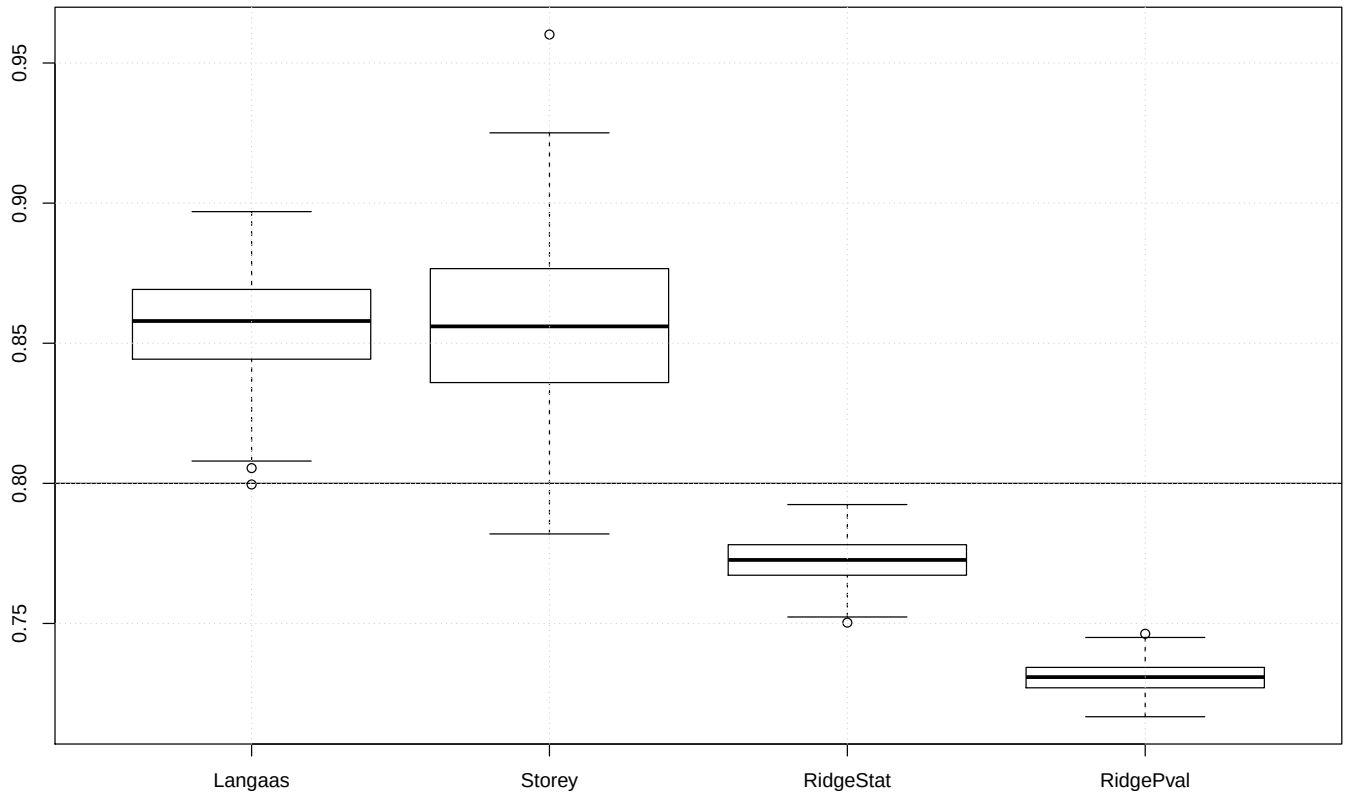


Figure: Boxplots of π_0 estimates with method of Langaas (JRSS, 2005), Storey (JRSS, 2002) or as part of ridge regression estimation of $\lambda(\theta)$ on 250 simulated datasets.

Estimation of the average power using Bisection method

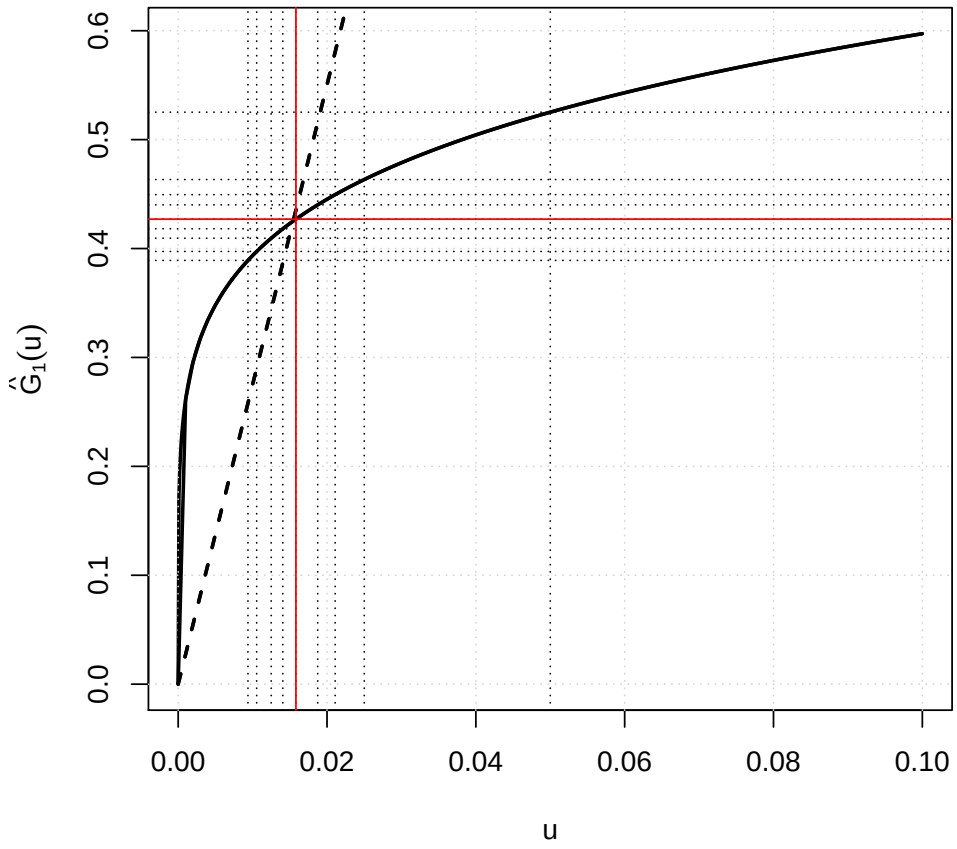


Figure: Ferreira and Zwinderman, Int. J Biostat, (2006) showed that, u^*, the solution to $\hat{G}_1(u; N) = \int H_1(u, \theta; N)\hat{\lambda}(\theta)d\theta = u\frac{\alpha(1-\hat{\pi}_0)}{\hat{\pi}_0(1-\alpha)}$ gives the average power, where α is the desired False Discovery Rate.

Sample size determination

- given pilot-data
- calculate test statistics and p-values
- assume parametric form for the null and alternative
- estimate π_0 and density of effect sizes, $\lambda(\theta)$
- estimate the power of the pilot-data
- or predict power at sample sizes larger than the pilot-data

$$\hat{G}_1(u^*; N') = \int H_1(u^*, \theta; N')\hat{\lambda}(\theta)d\theta = u^* \frac{\alpha(1-\hat{\pi}_0)}{\hat{\pi}_0(1-\alpha)} \qquad (6)$$

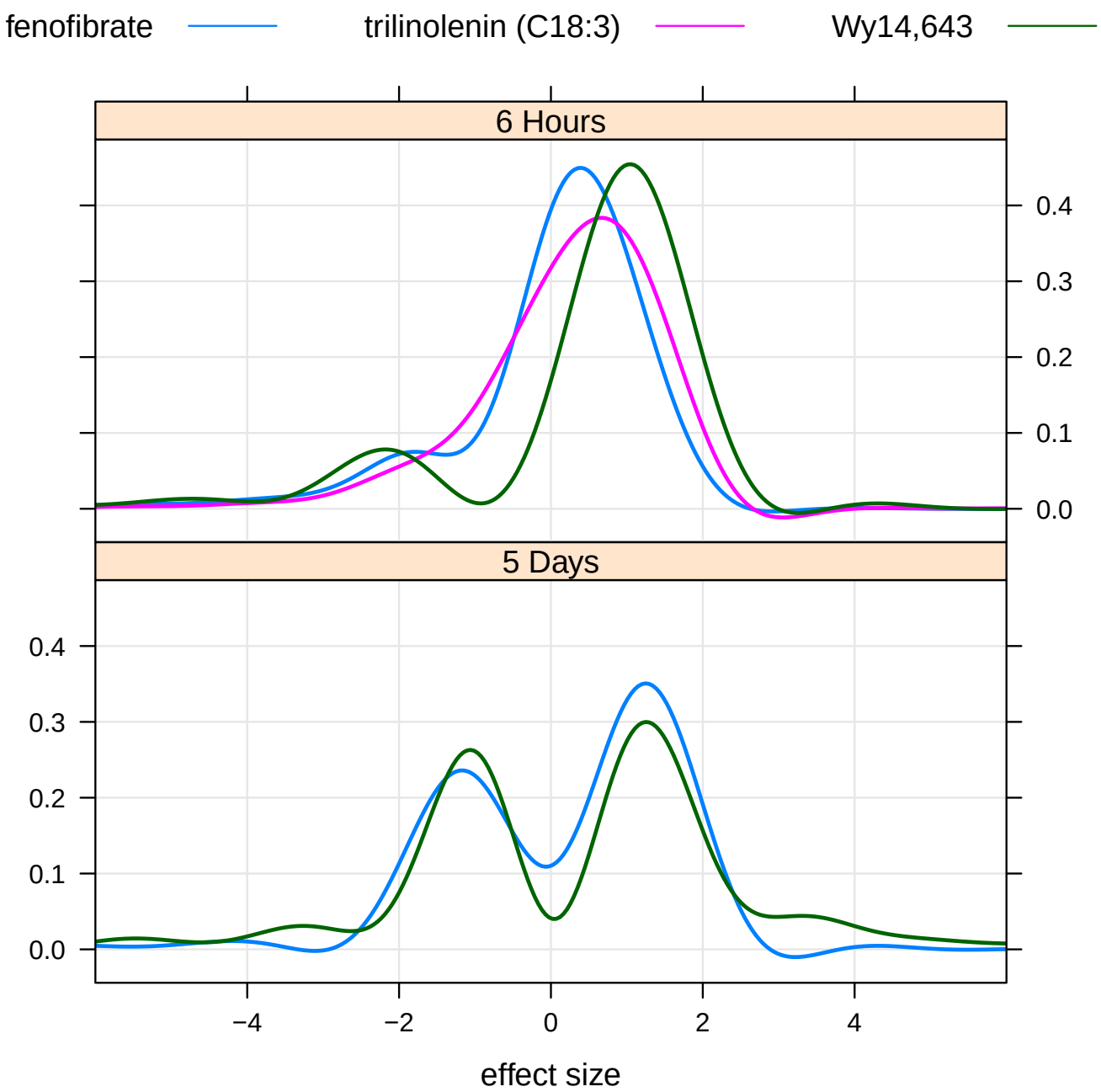
Nutrigenomics example

- 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

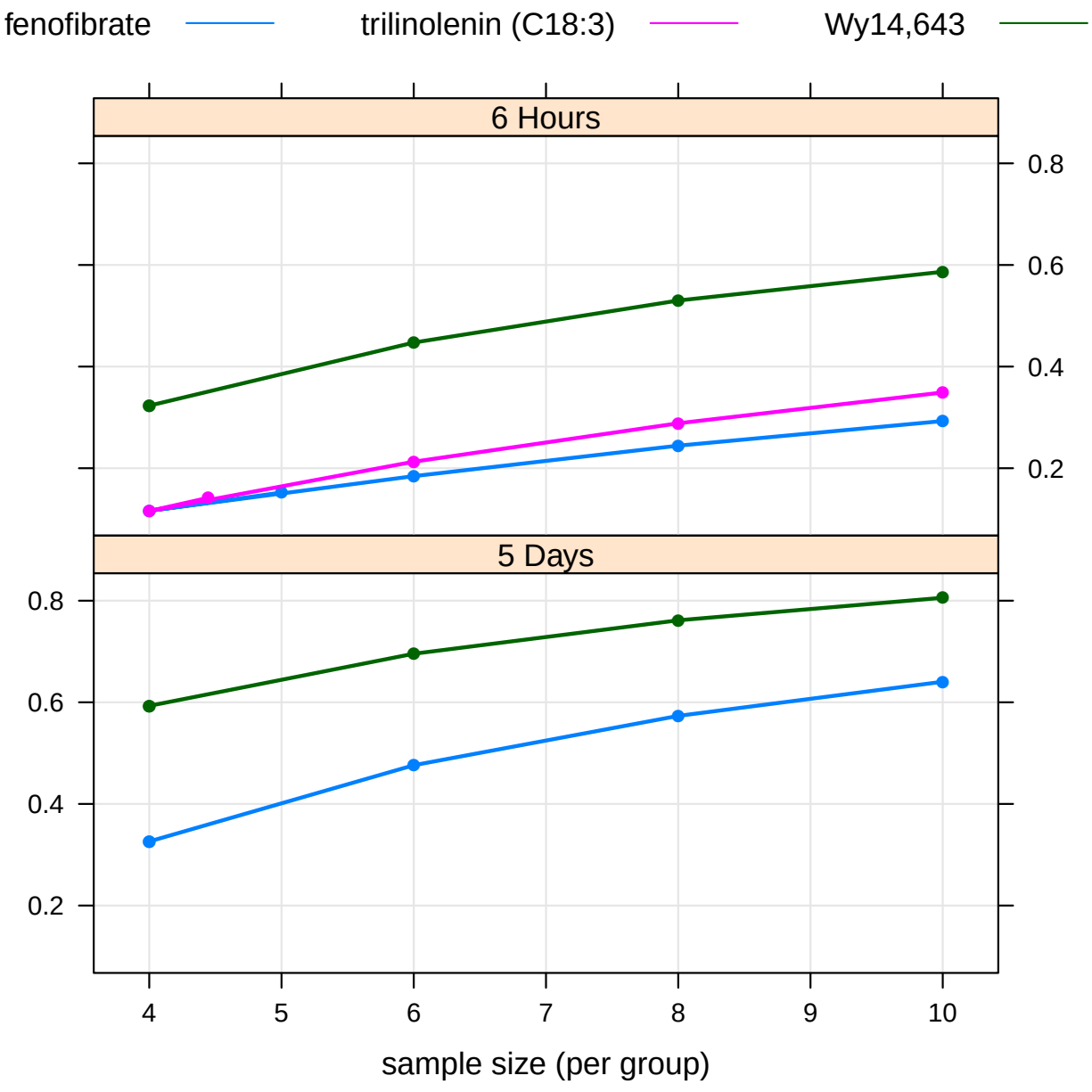
	probe-sets	group A	group B	experiment
1	16539	4 (wild-type)	4 (knock-out)	high, 6 hours
2	16539	4 (wild-type)	5 (knock-out)	intermediate, 6 hours
3	16539	5 (wild-type)	5 (knock-out)	low, 6 hours
4	16539	4 (wild-type)	4 (knock-out)	high, 5 days
5	16539	4 (wild-type)	4 (knock-out)	low, 5 days

van Iterson *et al.* BMC Genomics (2009).

Nutrigenomics example: density of effect sizes



Nutrigenomics example: power curves



Conclusion/Future Plans

General method for sample size determination for high-dimensional data with control of the FDR.

- likelihood ratio statistics (χ^2 and non-central χ^2) or F-statistics
- nonparametric null and assume location-model for the alternative

References

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.

J.A. Ferreia, A. Zwinderman.
Approximate sample size calculations with microarray data: an illustration.
Statistical application in genetics and molecular biology, 5, 1, 2006.

`SSPA`:
http://bioconductor.org/packages/release/bioc/html/SSPA.html.

Other cran and BioConductor packages: `OCplus`, `sizepower`, `ssize`, `ssize.fdr`