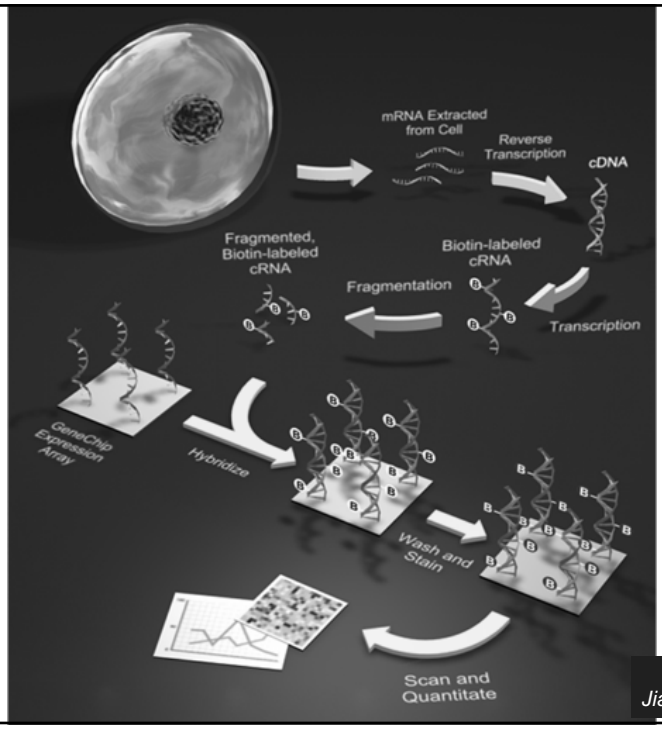
Overview

I. Technology background

II. Experimental design: Hybridization

III. Diferential expression analysis

IV. Experimental design: Power and sample size



I. Technology background

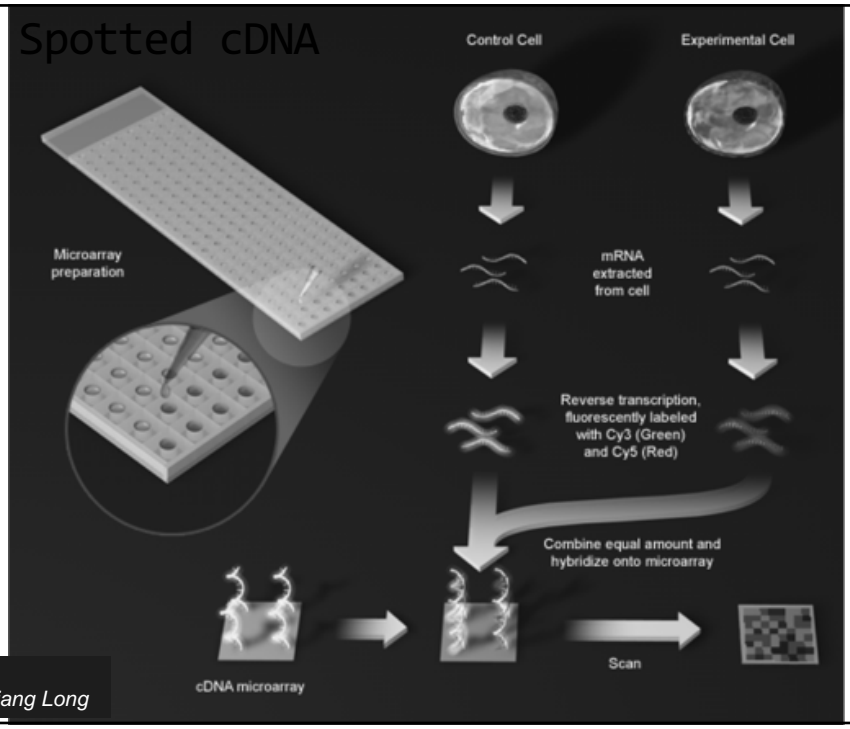
Two-color

- Spotted cDNA
- Agilent spotted probes

One-color

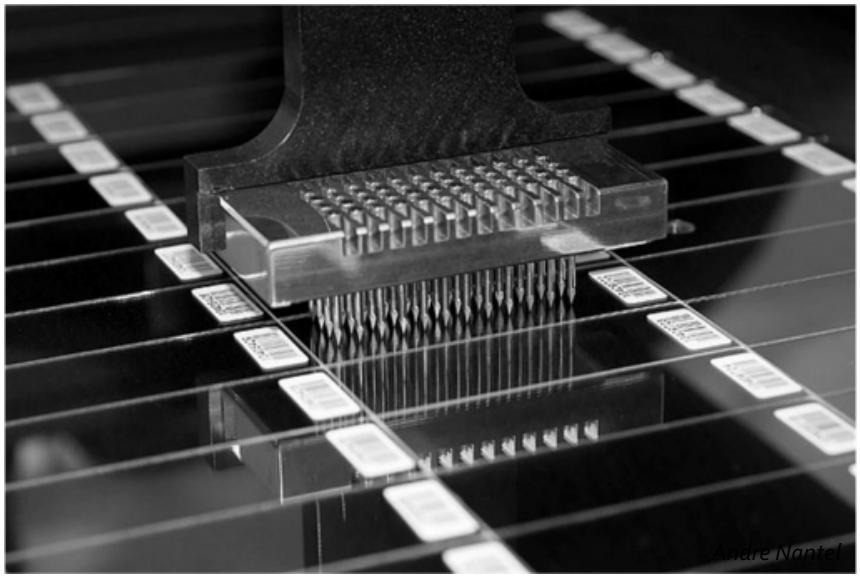
- Affymetrix GeneChip
- NimbleGene
- Illumina BeadChip

Spotted cDNA

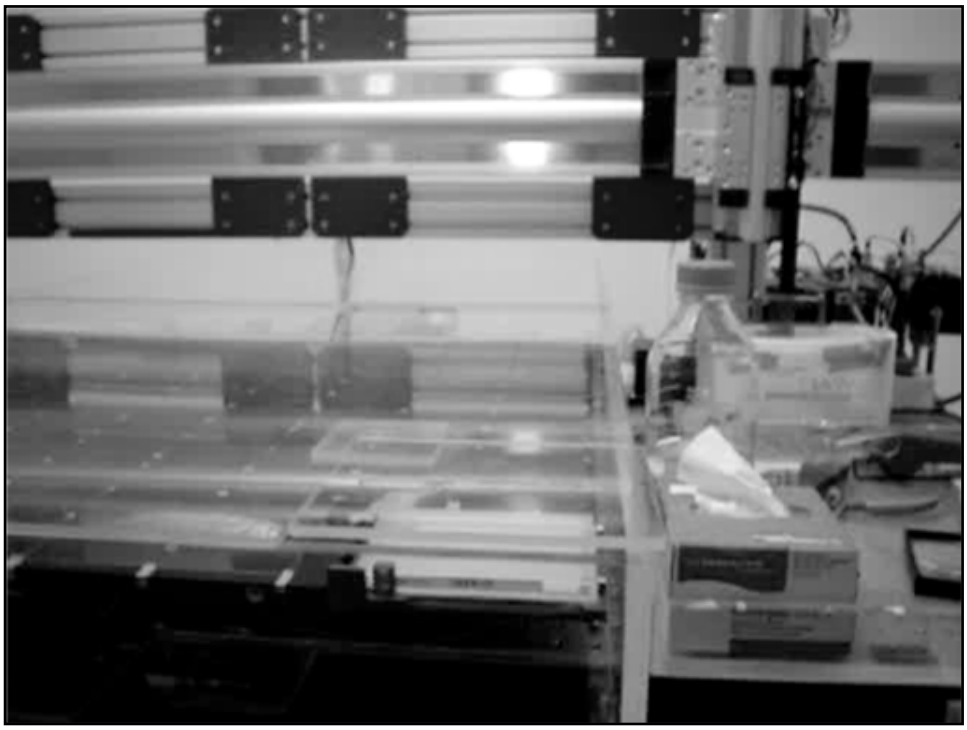


Jiang Long

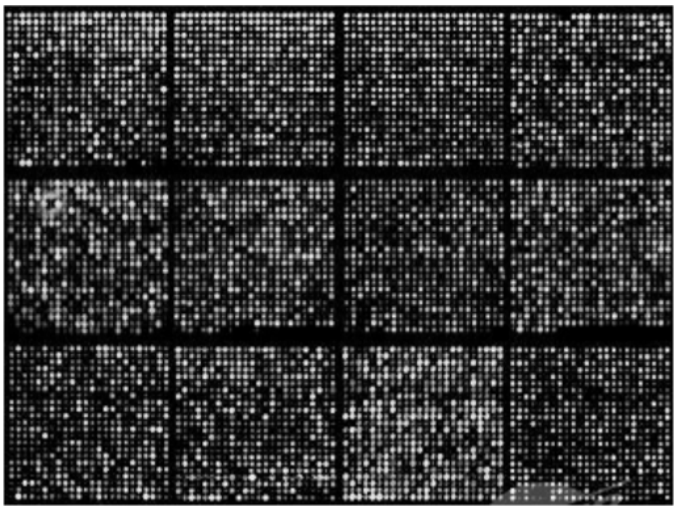
Printer catridge



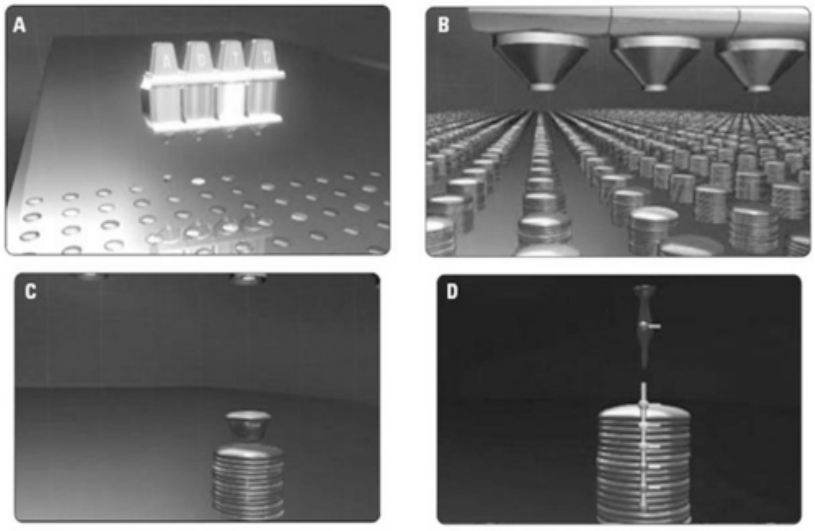
Andre Nantel



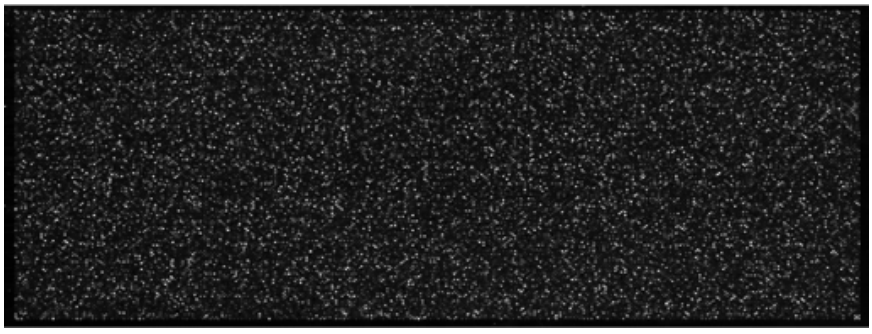
Spot separated by block associated to pins



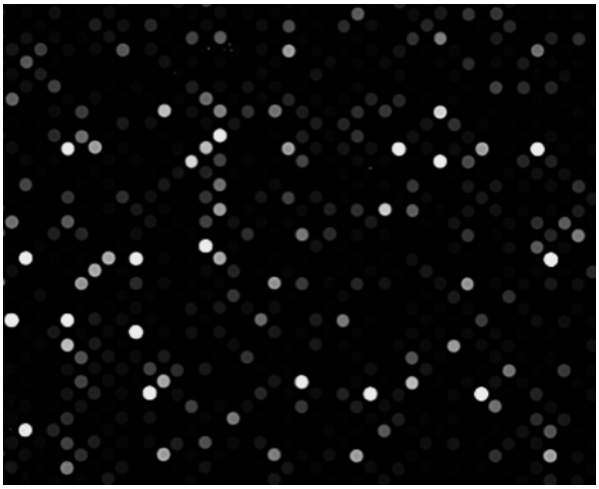
Agilent printed oligo arrays



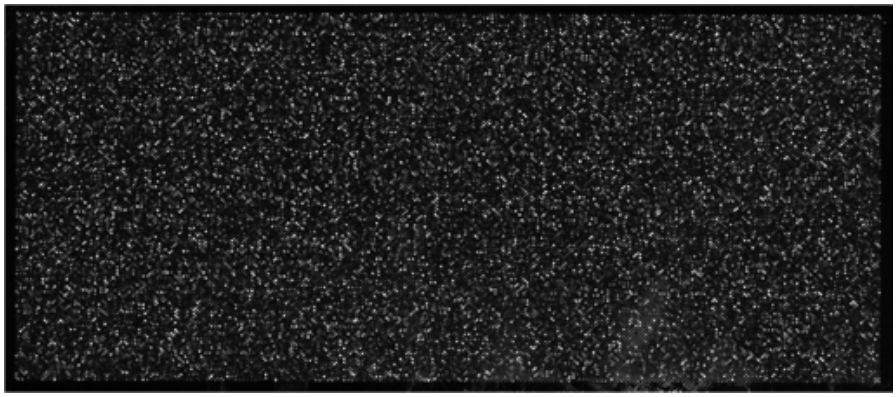
Agilent Mouse Whole-Genome Oligo Array

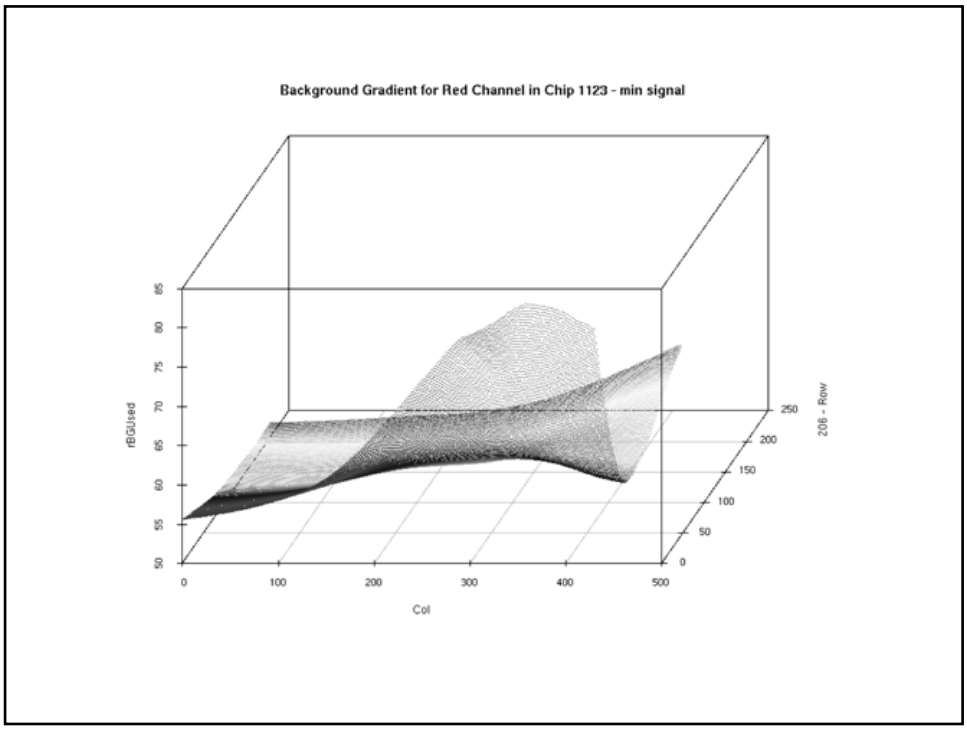


Feature Quality

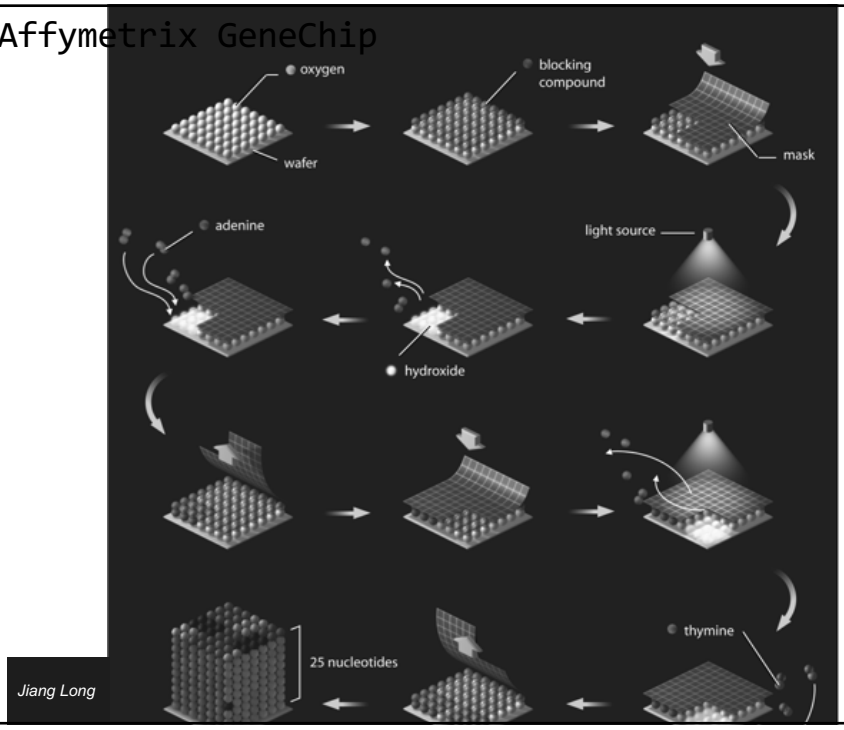


Agilent Mouse Whole Oligo Array

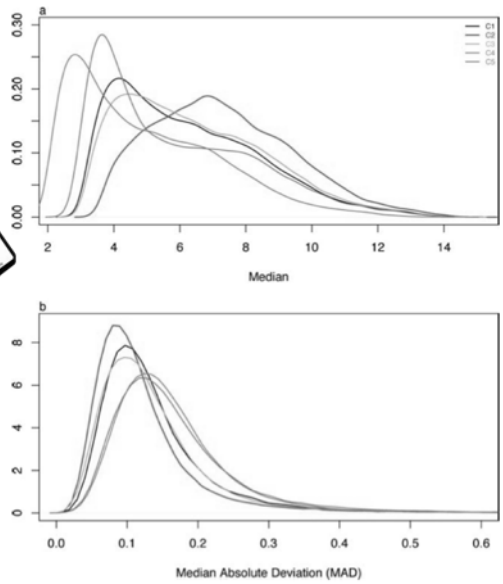




Affymetrix GeneChip

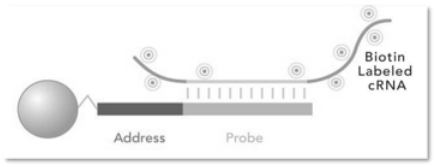


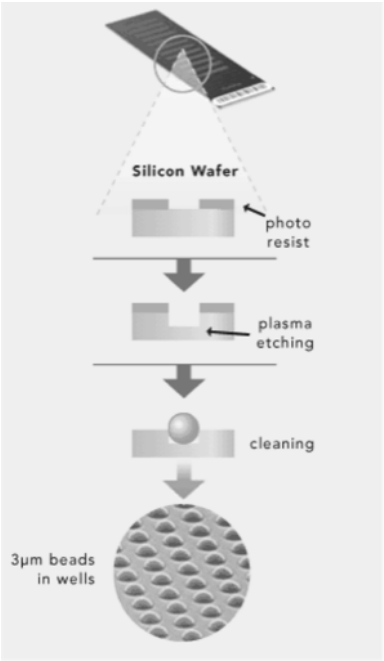
Laboratory effects in Affy data



Yang 2009, PlosOne 3:e3724

Illumina BeadChip

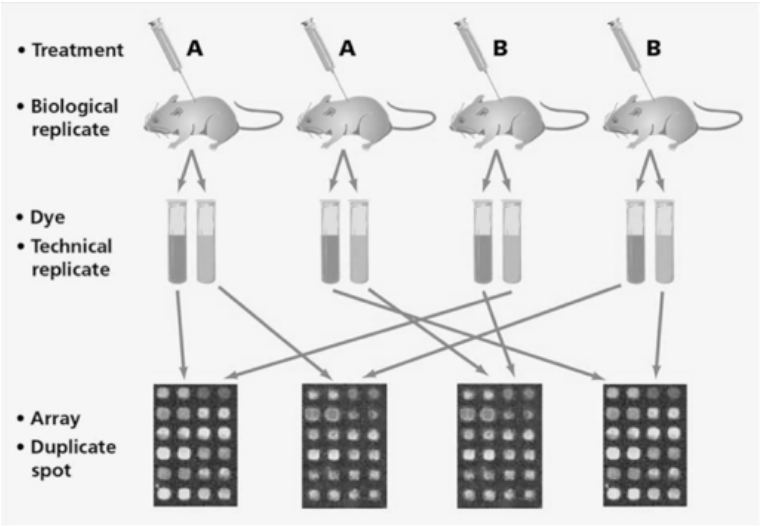




Multiple arrays per slide

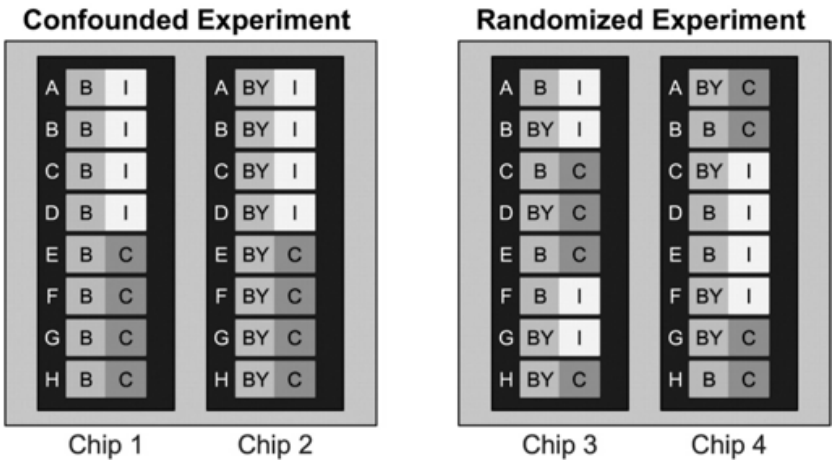
HumanHT-12 HumanRef-8 and HumanWG-6

II. Experimental design: Hybridization



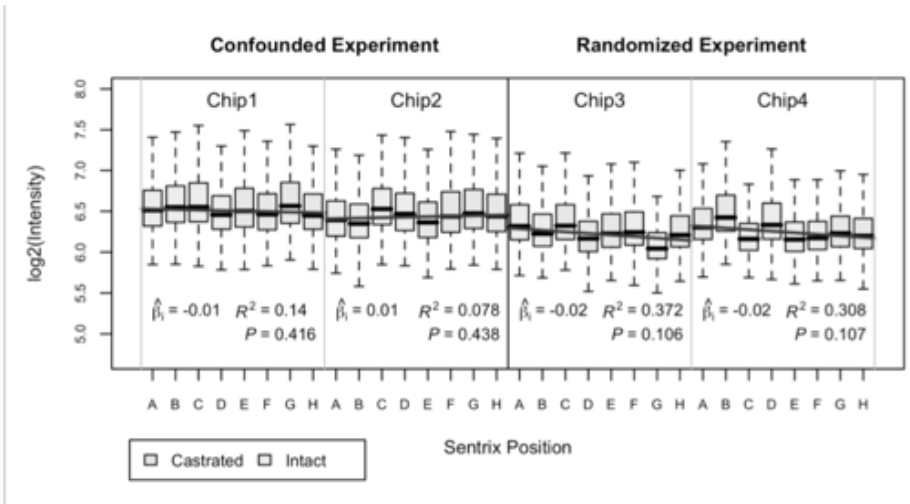
Churchill 2002, Nat Genet 32:490

Alternative hybridization layouts



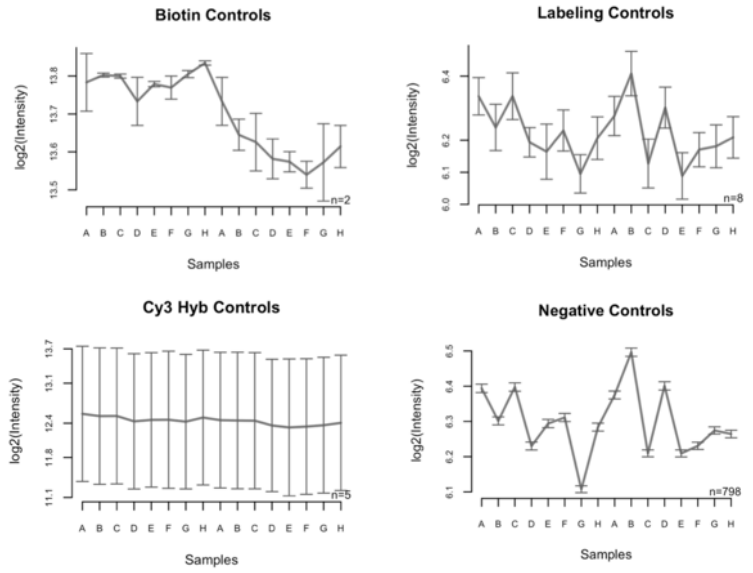
Verdugo 2009, Nucl. Acids Res. 37:5610

Effect of array position in Illumina BeadChips

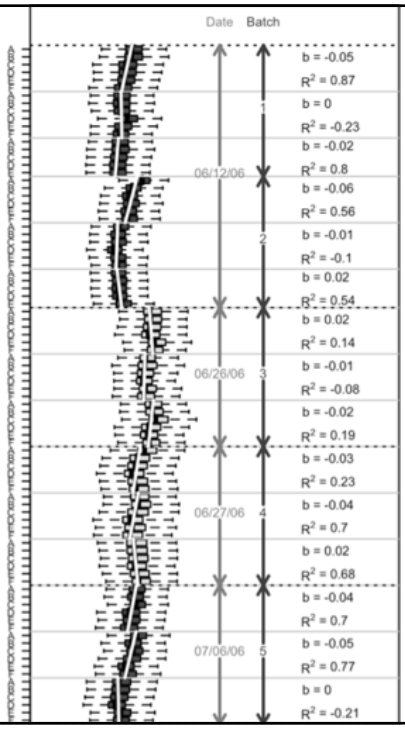


Verdugo 2009, Nucl. Acids Res. 37:5610

Effects control probes

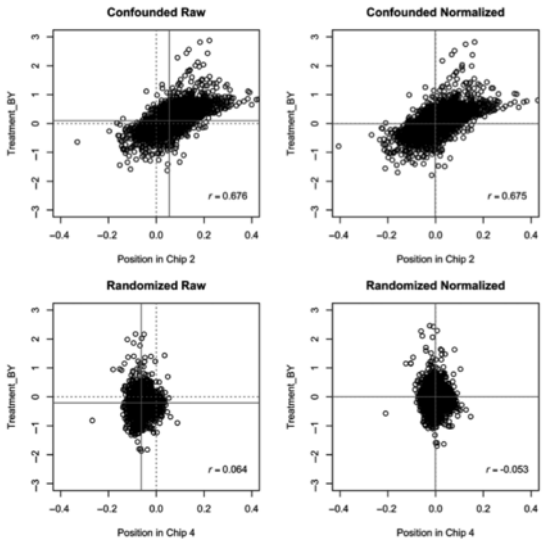


It may not be position, but time



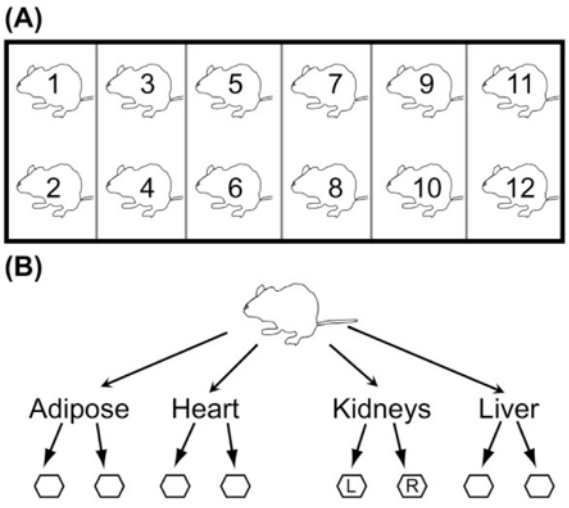
Verdugo 2009, Nucl. Acids Res. 37:5610

Randomization is critical



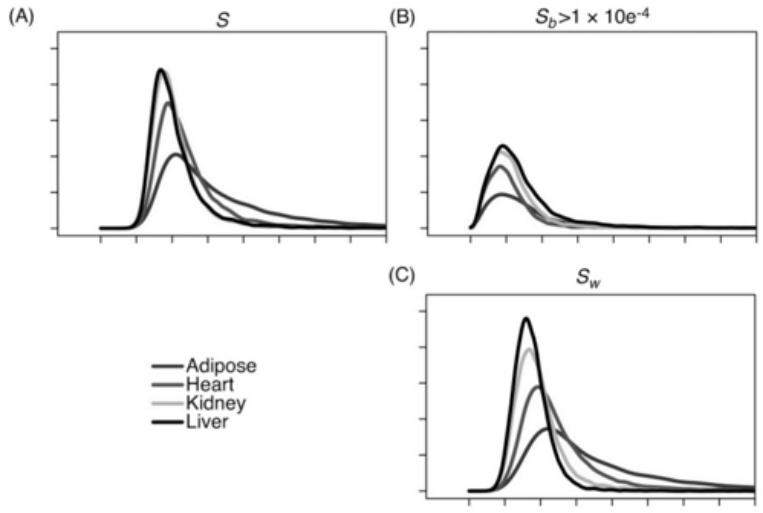
Verdugo 2009, Nucl. Acids Res. 37:5610

Technical variation may not come from microarray effects



Vedell 2011, BMC Genomics 12: 167.

Within Mouse Variance can be larger than between mice



Experimental design also important for: RNA sequencing

8 lanes in one flow cell!

Conclusions

- Technical artifacts tend to have larger effects than biological factors (genotype, treatment).
- Technical variation may arise from array technology or lab protocols.
- Confounding of experimental and technical factors creates false positives.
- Statistical modeling cannot correct for confounding.
- Randomization is critical, regardless of technology.

III. Diferential expression analysis

Ajuste de un modelo estadístico

- Gene by gene One-way ANOVA

$y_{ij} = \mu + A_i + \varepsilon_{ij}$

where,

y_{ij}	general logarithm of the gene expression in i^{th} treatment group of the j^{th} replicate
μ	mean
A_i	effect of the i^{th} treatment (i=1->5)
ε_{ij}	residual effect

Sums of squares in a One-way ANOVA

$$SS(A) = r\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2$$

$$SSE = \sum_{j=1}^{r}\sum_{i=1}^{a}(y_{ij} - \overline{y}_{i.})^2$$

$$SS(Total) = \sum_{j=1}^{r}\sum_{i=1}^{a}(y_{ij} - \overline{y}_{..})^2$$

SS(Total)=SS(A)+SSE

Statistical model in a Factorial design

- Gene by gene ANOVA

$$y_{ij} = \mu + A_i + B_j + AxB_i + \varepsilon_{ijk}$$

where,

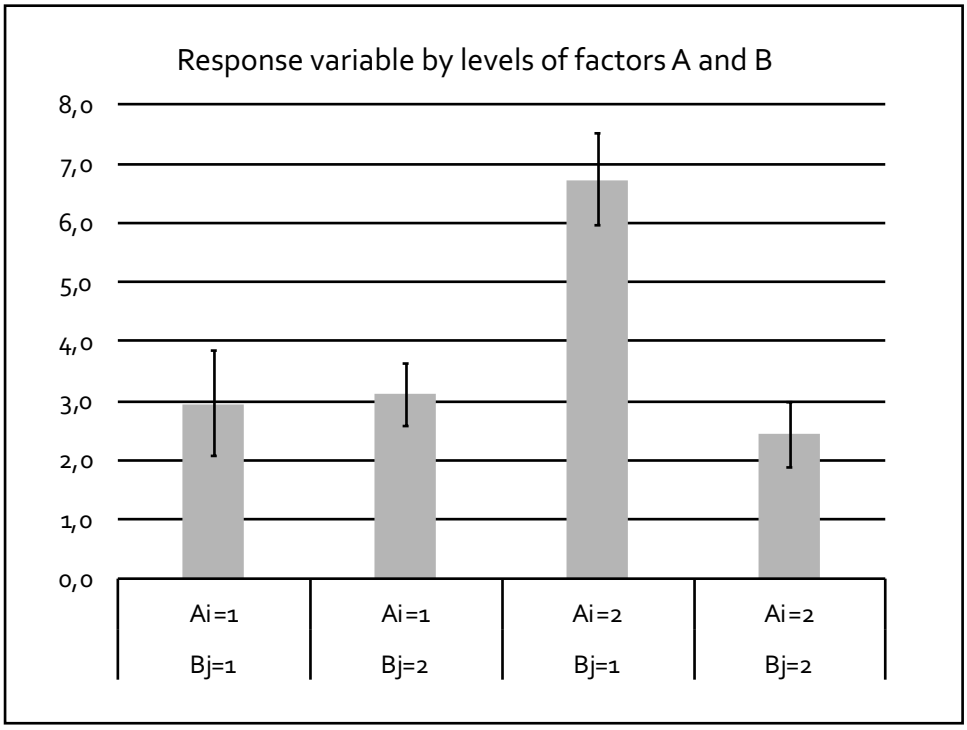
y_{ij}	general logarithm of the gene expression in i^{th} treatment group of the j^{th} replicate
μ	mean
A_i	effect of the i^{th} treatment (i=1->2)
B_j	effect of the j^{th} treatment (i=1->2)
AxB_{ij}	effect of the ij^{th} treatment
ε_{ijk}	residual effect

Gene expression for one gene

		Factor A	
		i=1	i=2
Factor B	j=1	3,0	6,7
		5,4	8,3
		2,0	5,2
		1,4	9.1
	j=2	1.5	2,3
		2,0	1,5
		3,2	4,0
		4,1	1,9

Gene expression for one gene

		Factor A	
		i=1	i=2
Factor B	j=1	y_{111}	y_{121}
		y_{112}	y_{122}
		y_{113}	y_{123}
		y_{114}	y_{124}
	j=2	y_{121}	y_{221}
		y_{122}	y_{222}
		y_{123}	y_{223}
		y_{124}	y_{224}



Sums of squares in a factorial design

$$SS(A) = rb\sum_{i=1}^{a}(\overline{y}_{i..} - \overline{y}_{...})^2$$

$$SS(B) = ra\sum_{j=1}^{b}(\overline{y}_{.j.} - \overline{y}_{...})^2$$

$$SS(AB) = r\sum_{j=1}^{b}\sum_{i=1}^{a}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$$

$$SSE = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \overline{y}_{ij.})^2$$

$$SS(Total) = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \overline{y}_{...})^2$$

$SS(total)=SS(A)+SS(B)+SS(AB)+SSE$

ANOVA table for an a x b factorial design

Source	SS	df	Mean Square
Factor A	SS(A)	(a−1)	SS(A)/(a−1)
Factor B	SS(B)	(b−1)	SS(B)/(b−1)
Interaction	SS(AB)	(a−1)(b−1)	SS(AB)/((a−1)(b−1))
Error	SSE	(N−ab)	SSE/(N−ab)
Total (Corrected)	SS(Total)	(N−1)	

Hypothesis testing

- Ho : Effect of A = 0
- H1 : Effect of A ≠ 0

$$F(A) = \frac{MS(A)}{MSE}$$

- Ho : Effect of B = 0
- H1 : Effect of B ≠ 0

$$F(B) = \frac{MS(B)}{MSE}$$

- Ho : Effect of AxB = 0
- H1 : Effect of AxB ≠ 0

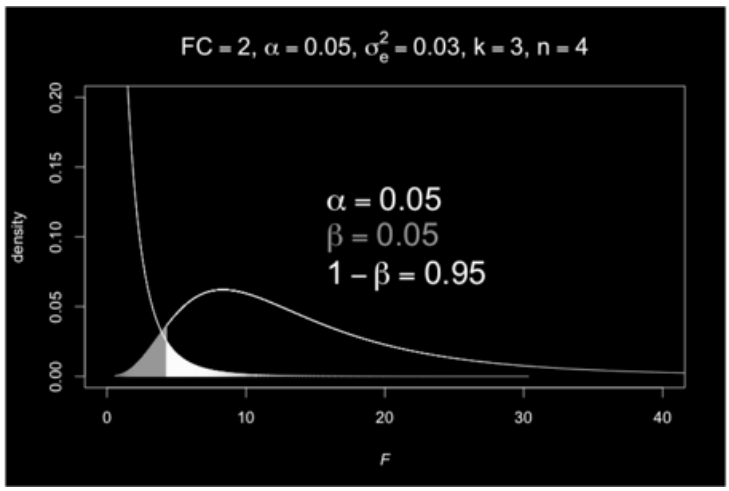
$$F(AxB) = \frac{MS(AxB)}{MSE}$$

Reject Ho if $P<\alpha$ Reject Ho if $fdr<FDR$

IV. Experimental design: Power and sample size

- What is appropriate *n* to detect DE?
- Find the power of a test at a given *n*
- $F = \frac{MS_{genotype}}{MS_{error}} = \frac{SS_{genotype} / df1}{SS_{error} / df2}$
- $ß = Pr(F_{H1} \mid df1, df2, \lambda)$
- $\lambda = \frac{(k-1)\, n\, \sigma^2_{genotype}}{\sigma^2_{error}}$

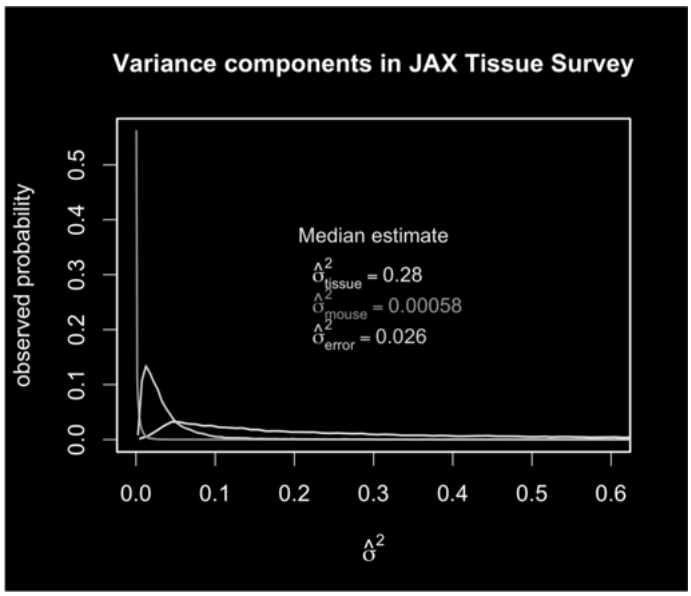
Power from an *F* test



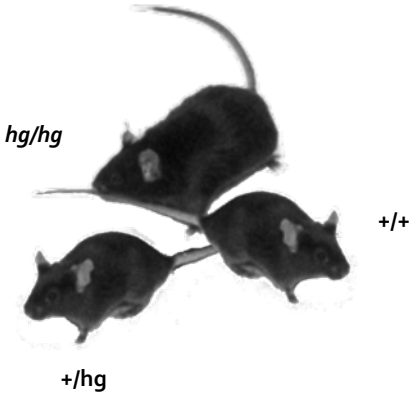
Mouse Illumina Tissue Survey

- Illumina Sentrix Mouse-6 V1.1 R1 (46,643 probes + 14 labeling controls)
- C57BL/6J
- sample 1: pancreas, brain, duodenum, jejunum, ileum, colon, stomach, spleen. Two technical replicates were hybridized for the kidney
- sample 2: liver, heart, lung, kidney, adrenals, muscle, testes, gonadal fat, and brown fat
- five biological replicates

Variance of gene expression in Mice



Differential expression between genotypes



Theoretical Genetic Variance (treatment effect)

FC	τ_1 (a)	τ_2 (d)	τ_3 (-a)	$\sigma^2_{genotype}$
1.2	0.09	0	-0.09	0.006
1.5	0.20	0	-0.20	0.027
2.0	0.35	0	-0.35	0.080
4.0	0.69	0	-0.69	0.320
6.0	0.90	0	-0.90	0.535
8.0	1.05	0	-1.05	0.721

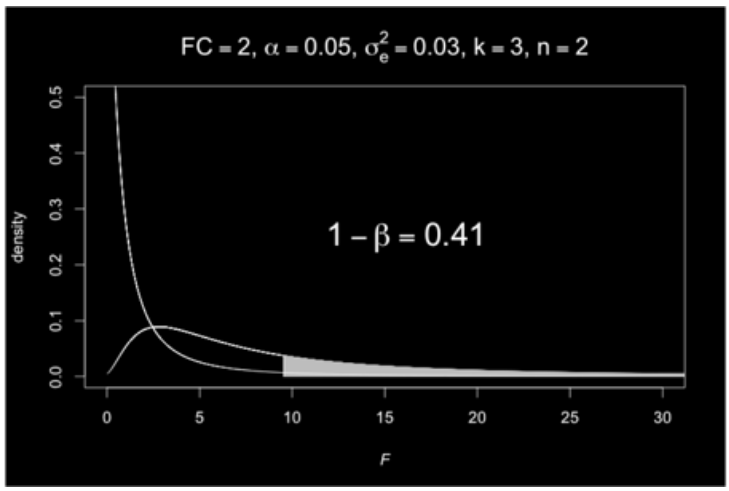
$$\tau = \log(FC)/2$$

$$\sigma^2_{genotype} = \frac{\sum_k \tau_i^2}{k}$$

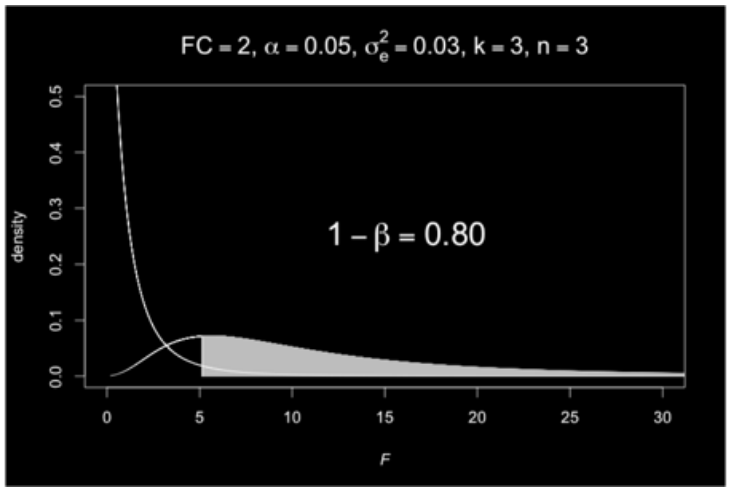
Power calculation for ANOVA

- k = 3, n = 4, α = 0.05
- $\sigma^2_{genotype(FC2)}$= 0.08
- σ^2_{error}= median( ) + 4*median( ) = 0.029
- $\lambda = \frac{(k-1)\, n\, \sigma^2_{genotype}}{\sigma^2_{error}} = \frac{2 * 4 * 0.08}{0.029} = 22.07$
- F_o=*qf(1-0.05, 3-1, 3(4-1)) = 4.3*
- *ß = Pr(F_{H1} | df1, df2,* λ)
- *= pf(F_o, 3-1, 3(4-1), 22.07) = 0.05*
- *Power = 1 - ß = 0.95*

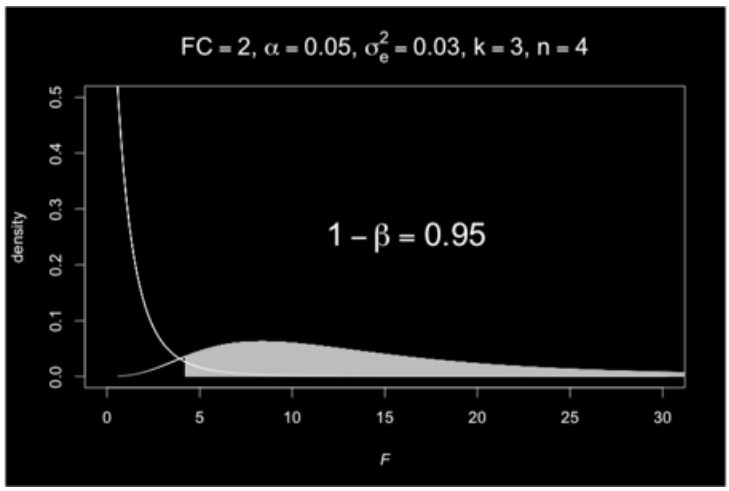
Power from *F* test by n



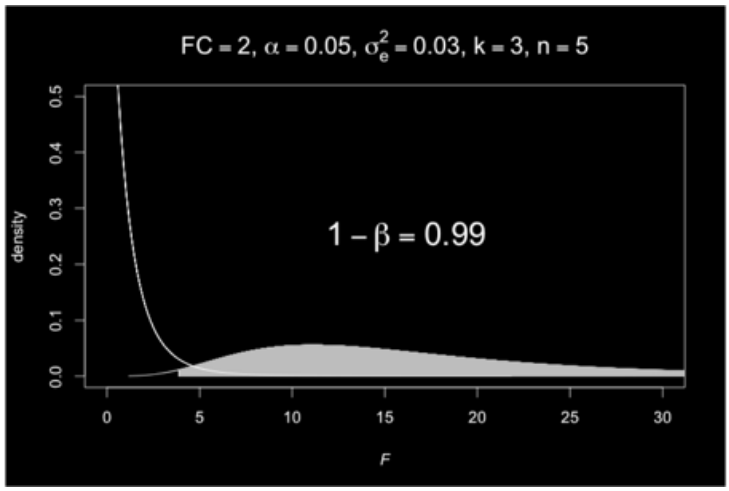
Power from *F* test by n



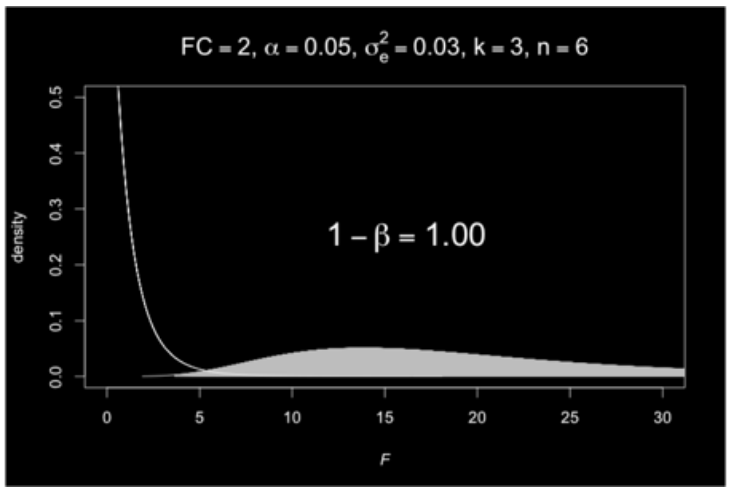
Power from *F* test by n



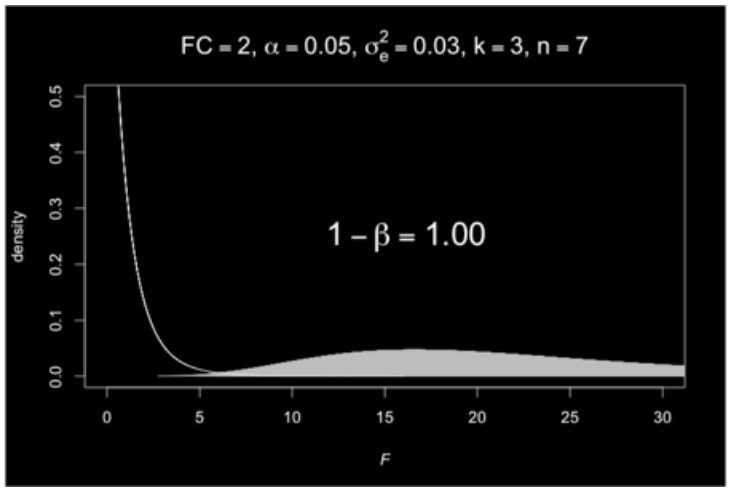
Power from *F* test by n



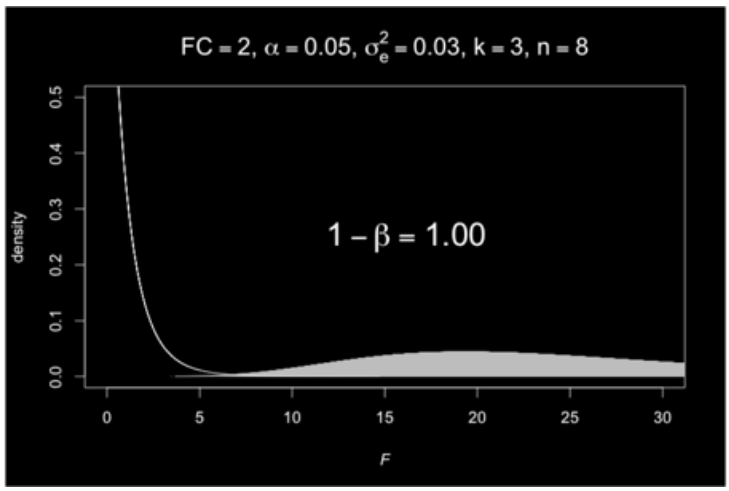
Power from *F* test by n



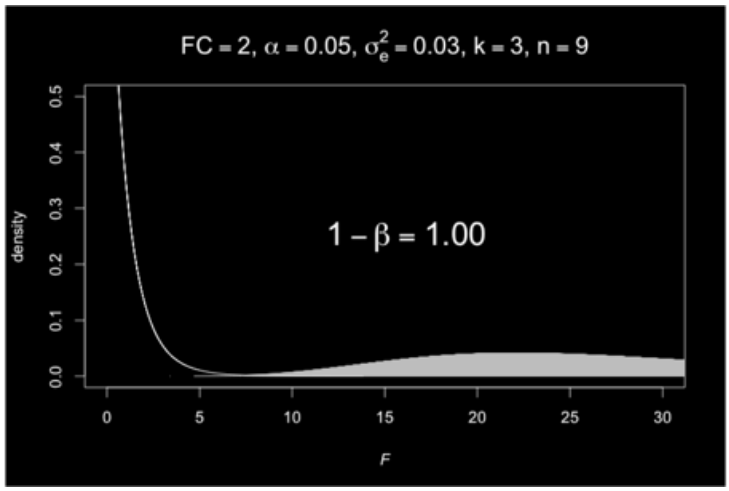
Power from *F* test by n



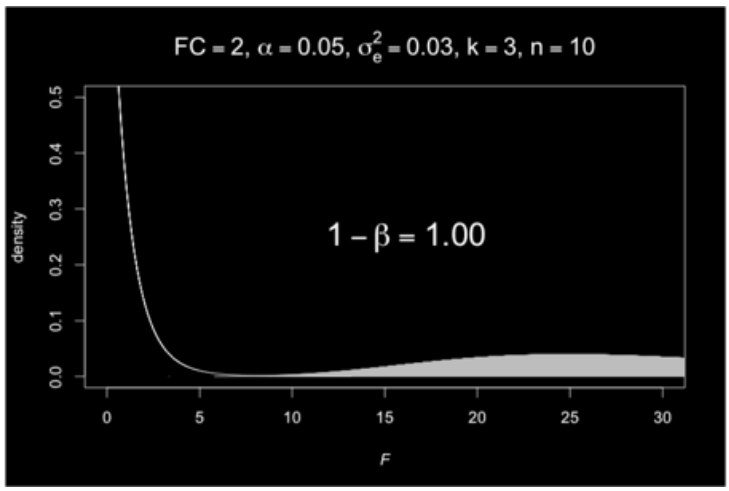
Power from *F* test by n



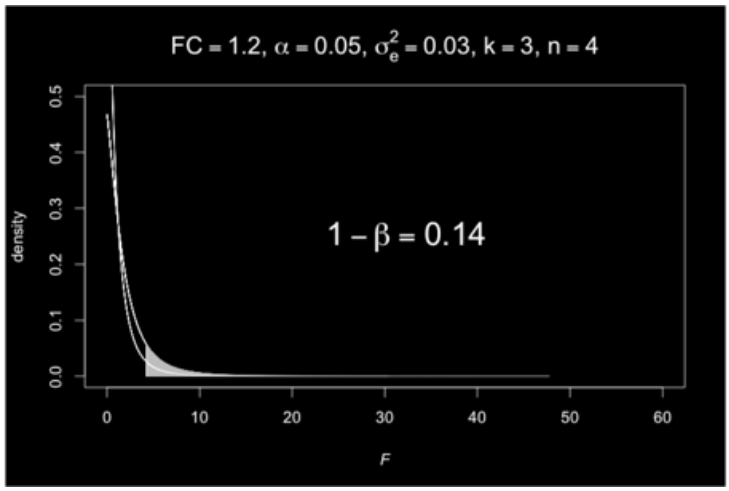
Power from *F* test by n



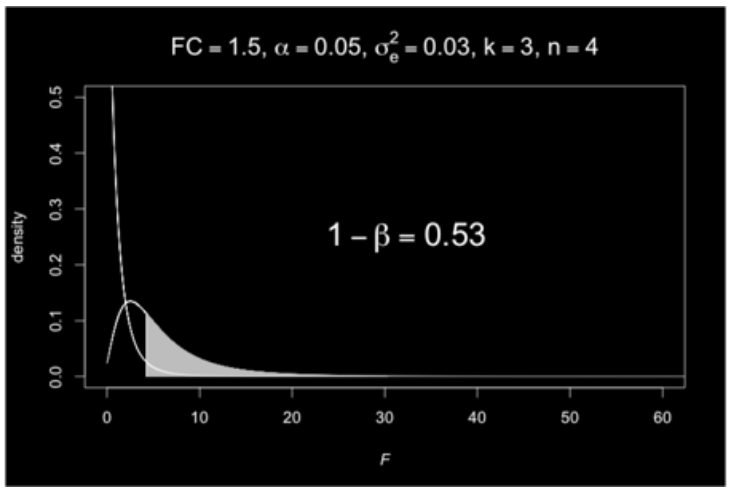
Power from *F* test by n



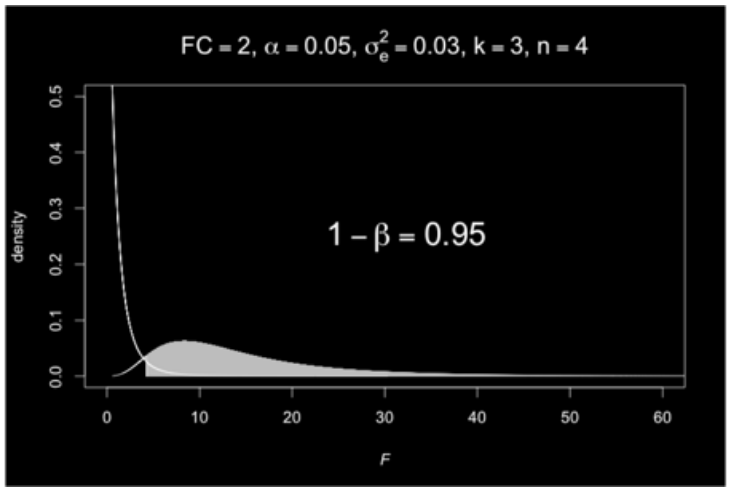
Power from *F* test by FC



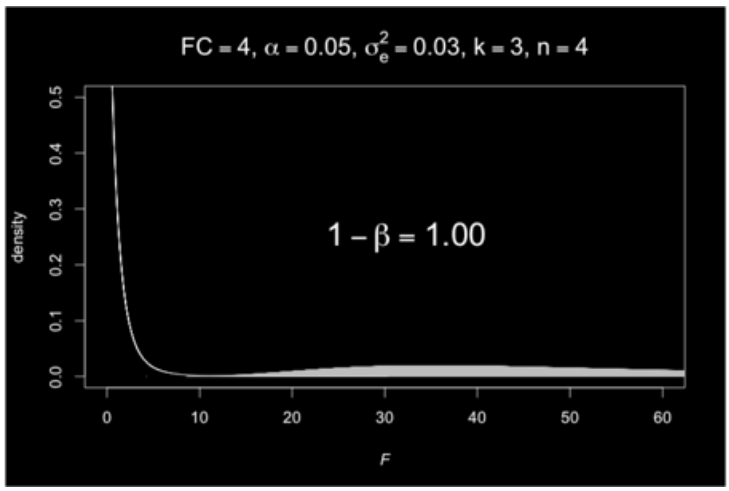
Power from *F* test by FC



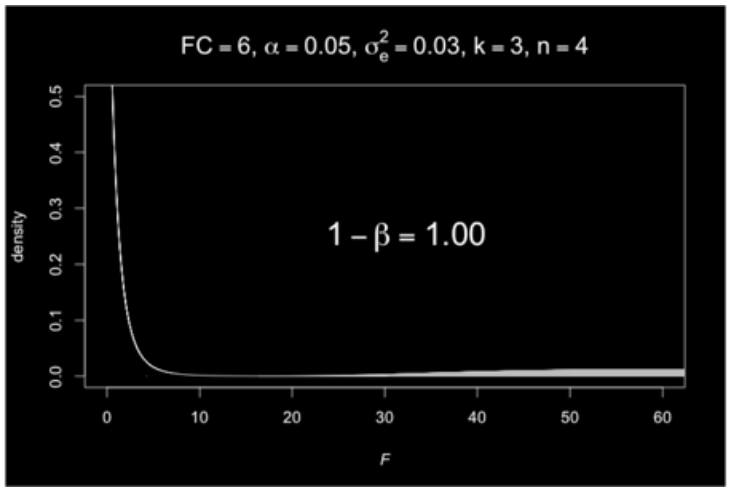
Power from *F* test by FC



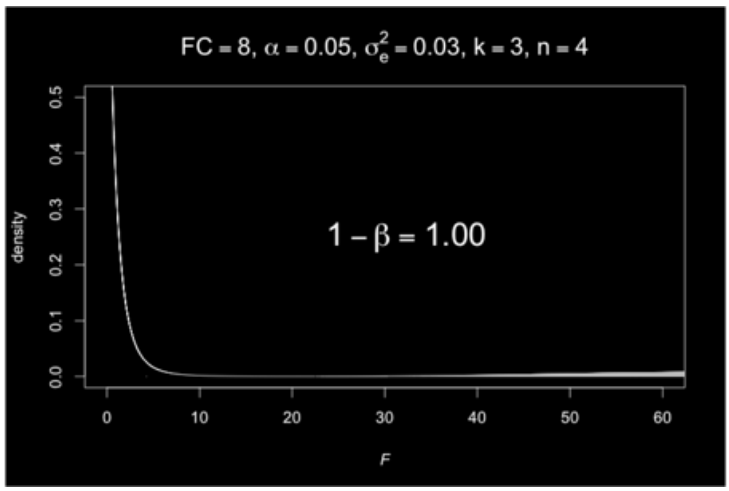
Power from *F* test by FC



Power from *F* test by FC



Power from *F* test by FC



Accounting for multiple comparisons: Liu and Hwang 2007

$$\left(\frac{FDR}{1-FDR}\right)\left(\frac{1-\pi_0}{\pi_0}\right) = \frac{\alpha}{1-\beta}$$

π_o is the proportion of non-differentially expressed genes

ssize.F function from the *ssize.fdr* R package

Bioinformatics **23**: 739

Erratum: *Bioinformatics* 24: 149

Accounting for multiple comparisons: Pounds and Cheng 2005

- Alternative method, equivalent results
- It can calculate FC from data
- Bioinformatics (2005) 21(23): 4263
- Erratum: Bioinformatics (2009) 25(5): 698

Sample size in mouse Illumina experiments

- m = 28000 (number of genes)
- FDR = 0.1
- π_0 = 0.90, 0.95, 0.99
- k = 3
- σ^2_{error}= 0.026
- FC = 2

ssize.fdr R code

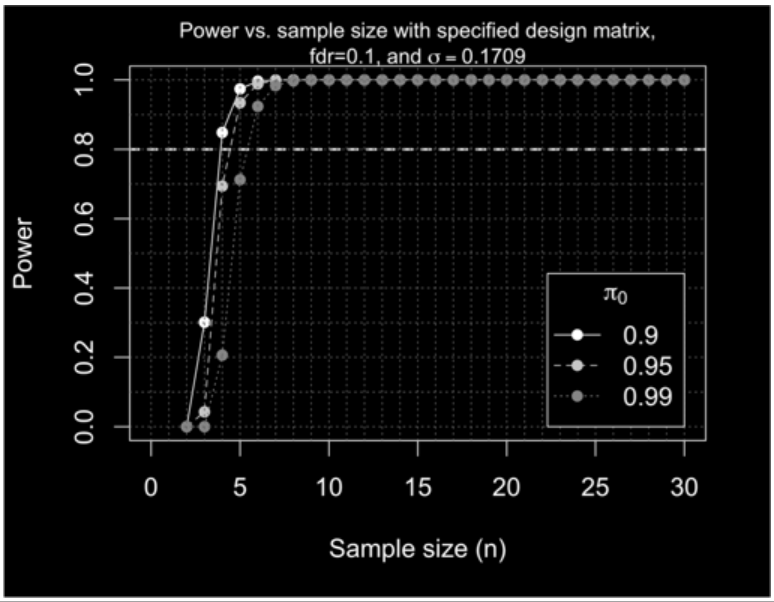
```
> des <- matrix(c(1,0,0,0,1,0,0,0,1),
  ncol=k,byrow=TRUE)
> des
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
> b <- c(-log(FC)/2, 0, log(FC)/2)
> df.f <- function(n) 3*(n-1)
>  liu_ssize <- ssize.F(X=des, beta=b, dn=df.f,
  sigma=sqrt(0.026), fdr=0.1, power=0.8,
  pi0=c(.9,.95, .99), maxN=30)
```

ssize.fdr output

```
> liu_ssize
$ssize
      pi0 ssize     power
[1,] 0.90     4 0.8485756
[2,] 0.95     5 0.9341481
[3,] 0.99     6 0.9238113

$power
       n       0.9      0.95      0.99
 [1,]  2 0.0000000 0.0000000 0.0000000
 [2,]  3 0.3013482 0.0427180 0.0000000
 [3,]  4 0.8485756 0.6939236 0.2062821
 [4,]  5 0.9741036 0.9341481 0.7123410
 [5,]  6 0.9963947 0.9885850 0.9238113
 [6,]  7 0.9995848 0.9983811 0.9837162
```

Sample size in mouse Illumina experiments



Conclusions

- Power to detect DE depends on both technical and biological factors
- Technical sources of variation can offset biological factors
- Variation is affected by technology, site, and protocols, and tissue (dissection)
- Pilot projects are essential to estimate variance, power, and replication
- 5 animals per condition is a good starting point