

A close-up photograph of a microarray chip, showing a grid of small spots on a glass surface. The chip is mounted on a yellow carrier. The background is dark and slightly blurred.


FACULTAD DE MEDICINA
UNIVERSIDAD DE CHILE


INSTITUTO
DE CIENCIAS
BIOMEDICAS

Análisis de expresión diferencial con
MICROARREGLOS

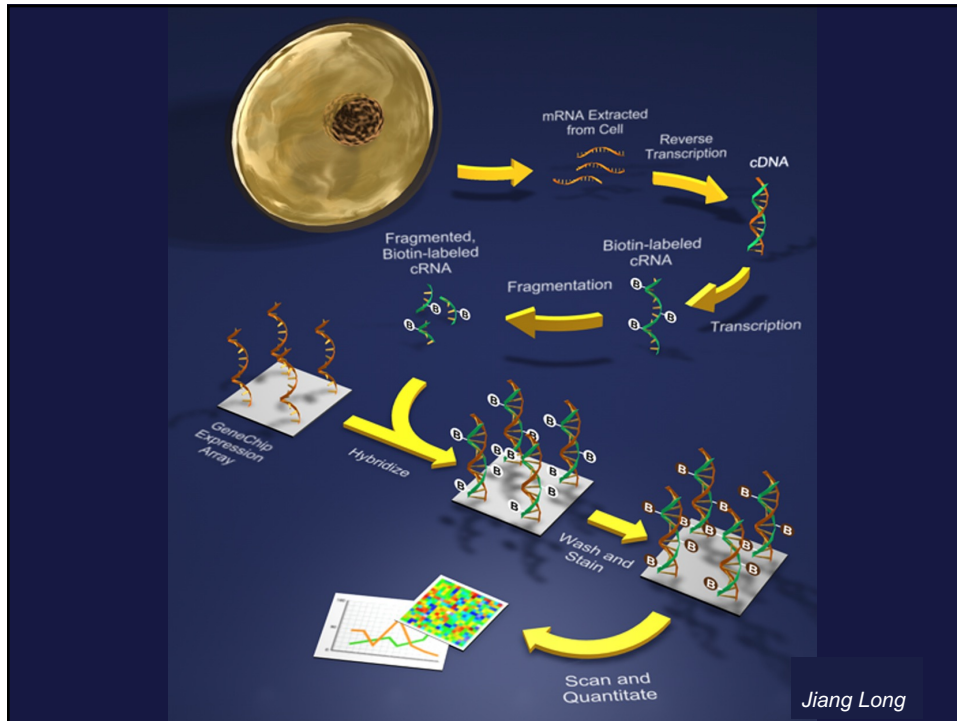
Ricardo A. Verdugo, Ph.D
Programa de Genética Humana, ICBM
Departamento de Oncología Básico Clínica
Facultad de Medicina, U. de Chile

1

Overview

- I. Technology background
- II. Experimental design: Hybridization
- III. Diferential expression analysis
- IV. Experimental design: Power and sample size

2



3

I. Technology background

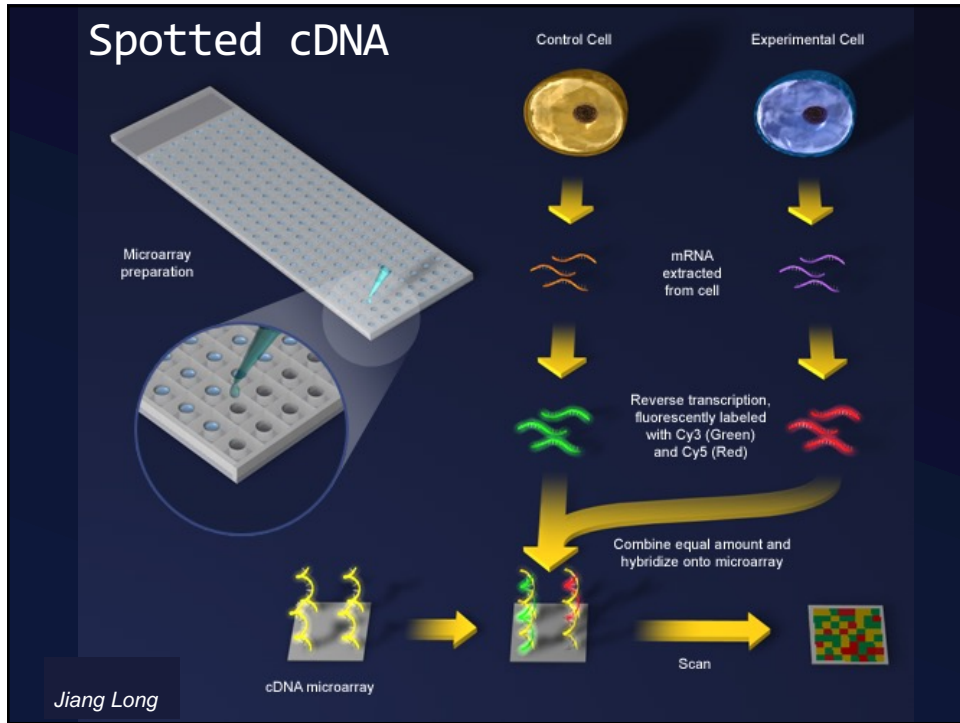
Two-color

- Spotted cDNA
- Agilent spotted probes

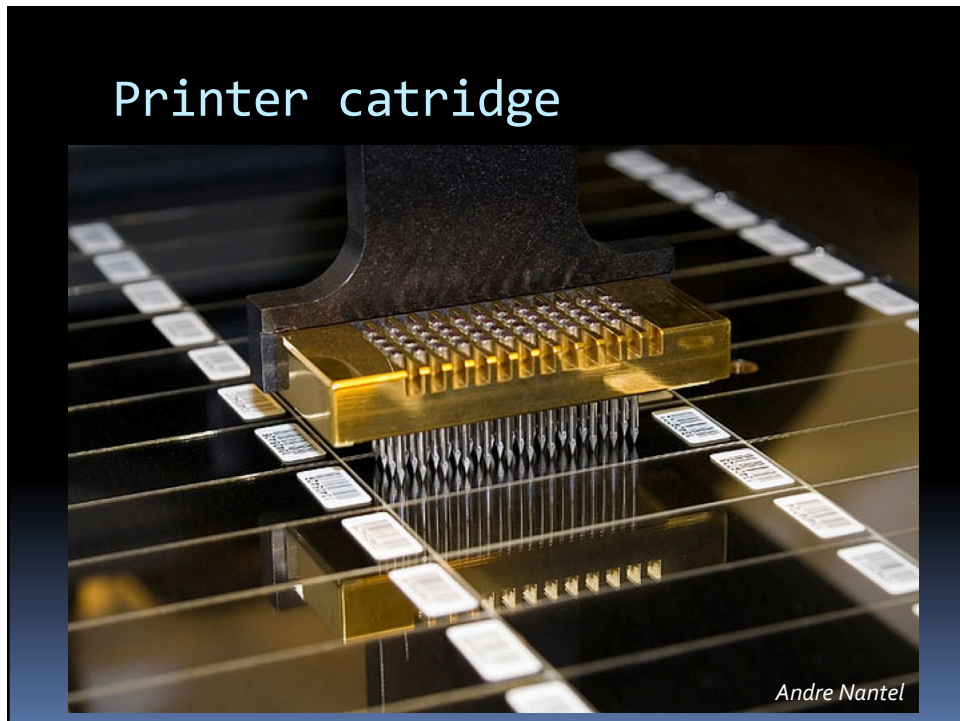
One-color

- Affymetrix GeneChip
- NimbleGene
- Illumina BeadChip

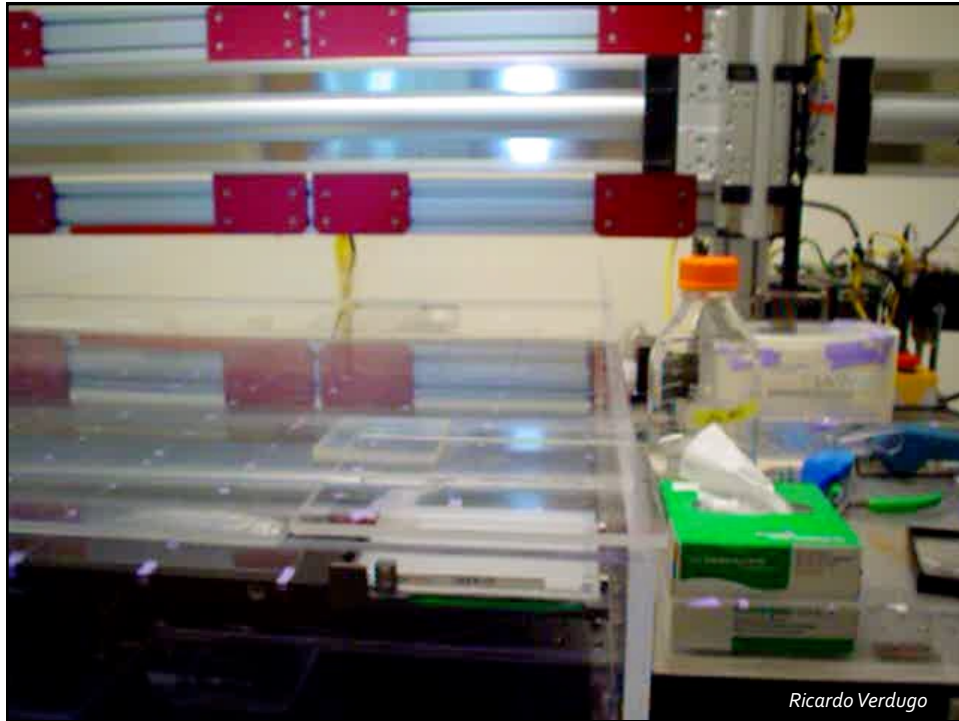
4



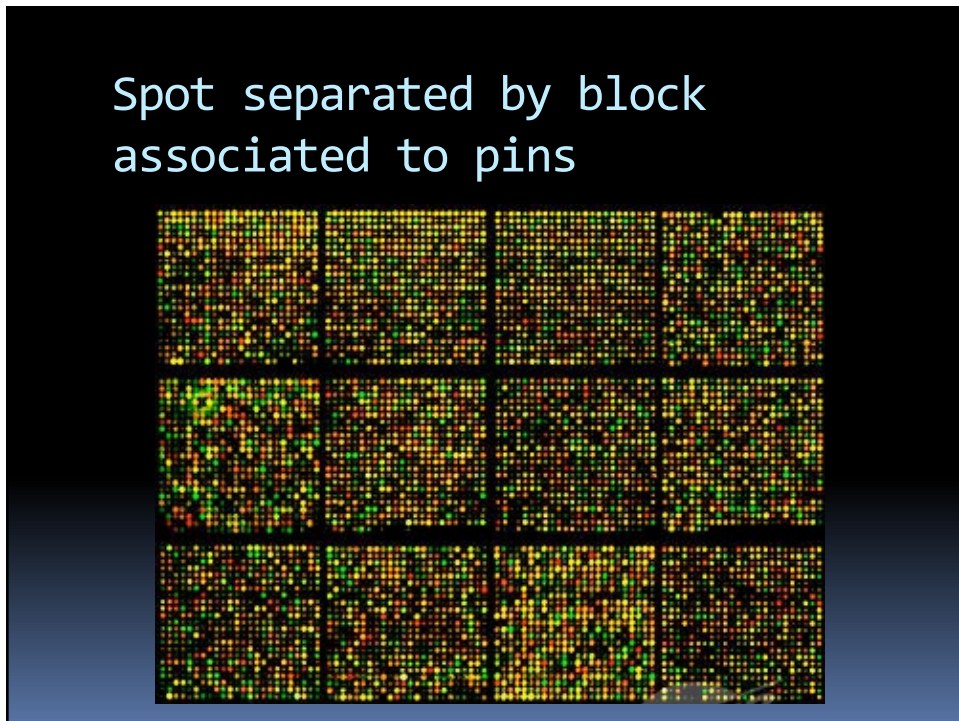
5



6

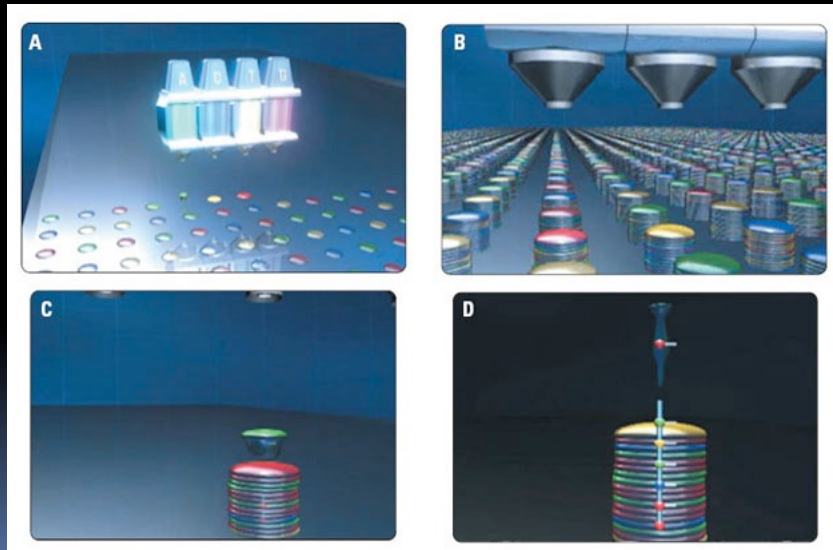


7



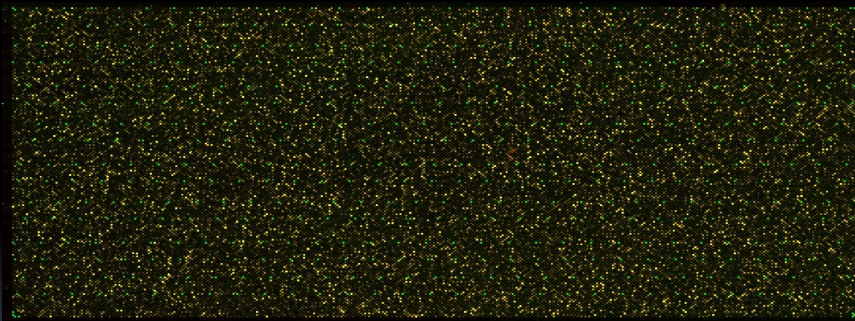
8

Agilent printed oligo arrays



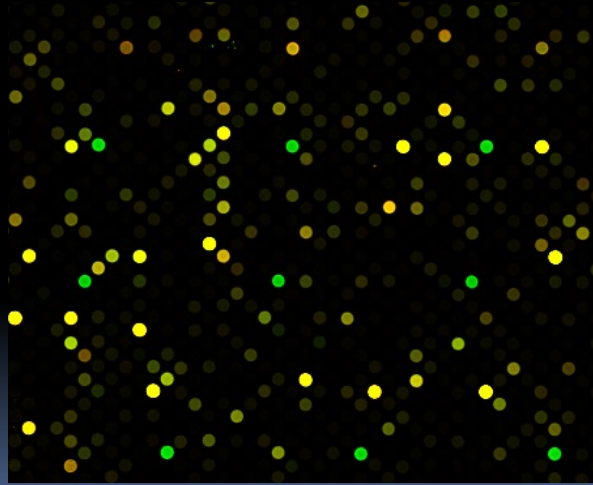
9

Agilent Mouse Whole-Genome Oligo Array



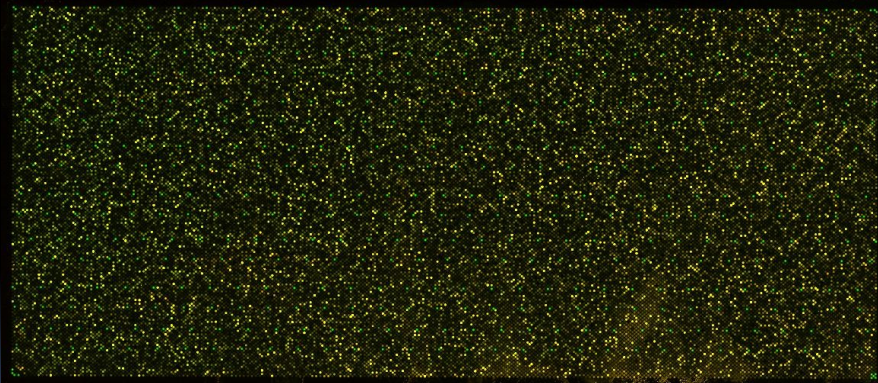
10

Feature Quality

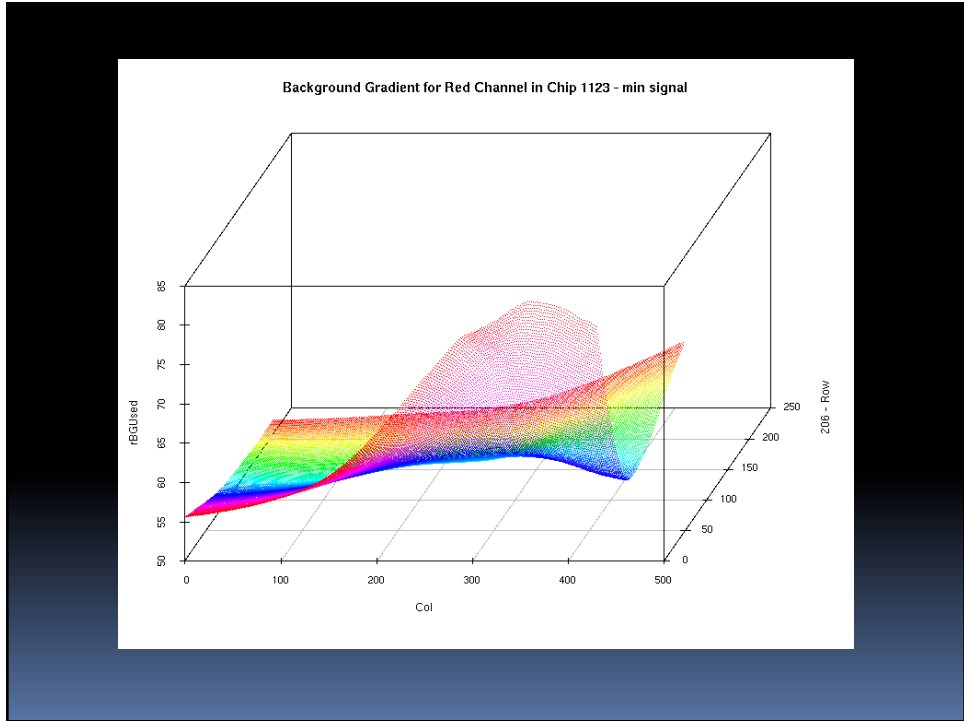


11

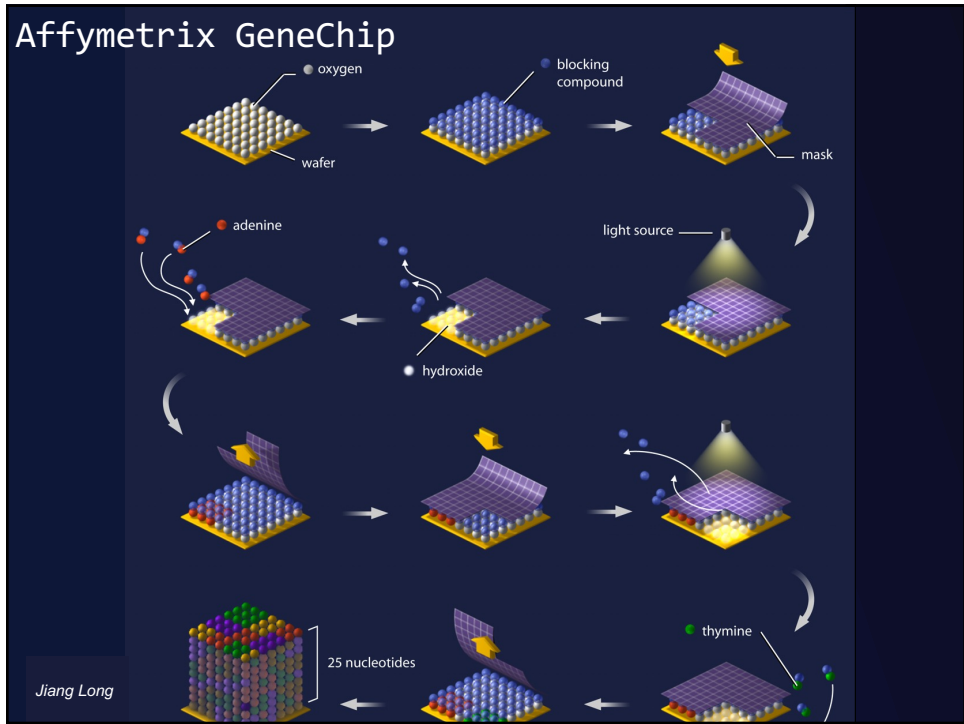
Agilent Mouse Whole Oligo Array



12

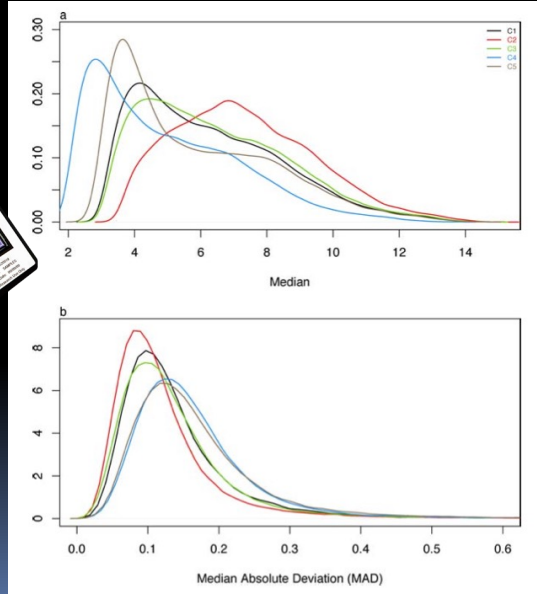


13



14

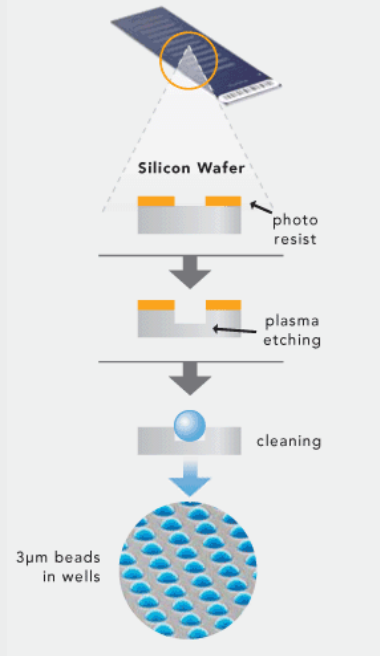
Laboratory effects in Affy data



Yang 2009,
PlosOne
3:e3724

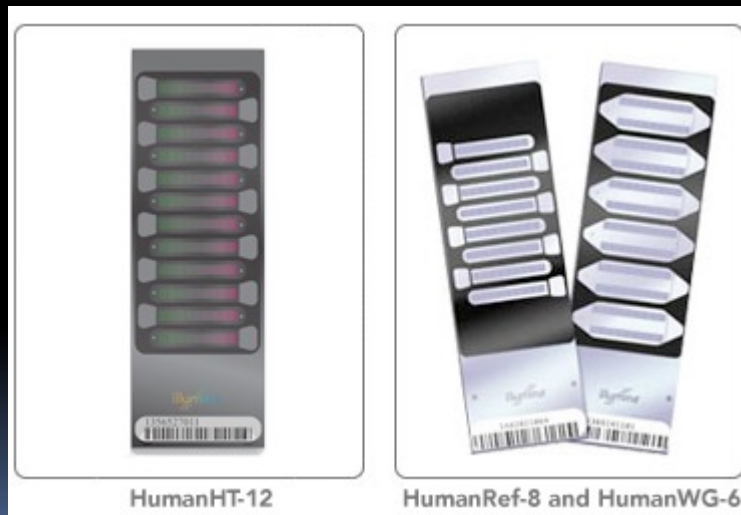
15

Illumina BeadChip



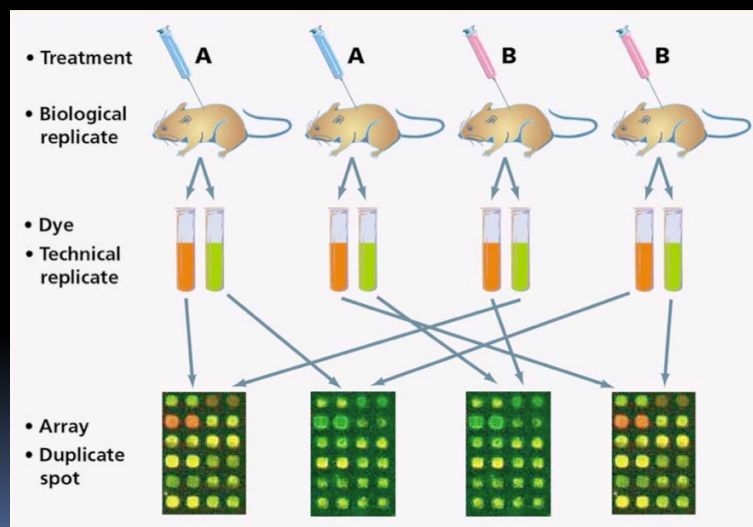
16

Multiple arrays per slide



17

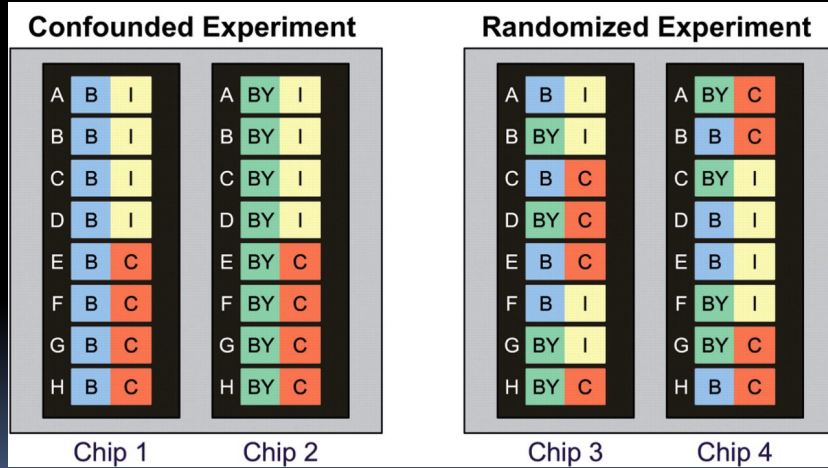
II. Experimental design: Hybridization



Churchill 2002, Nat Genet 32:490

18

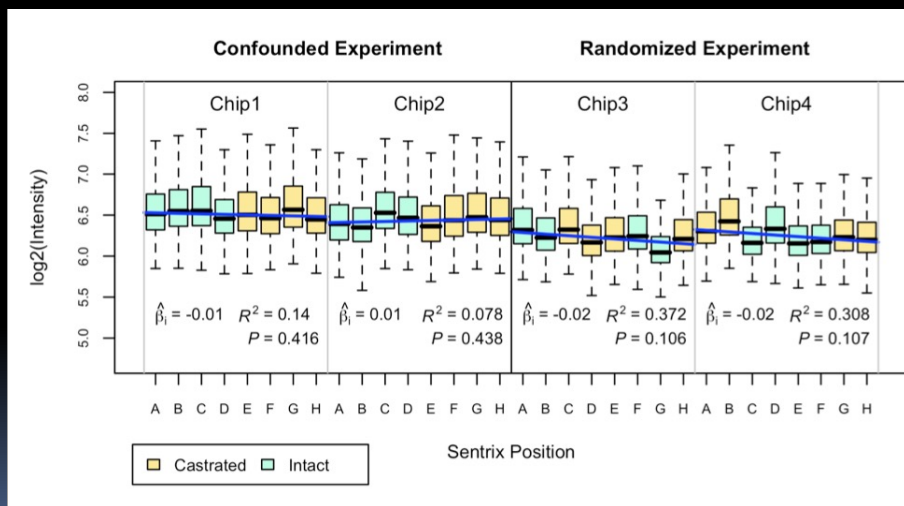
Alternative hybridization layouts



Verdugo 2009, Nucl. Acids Res. 37:5610

19

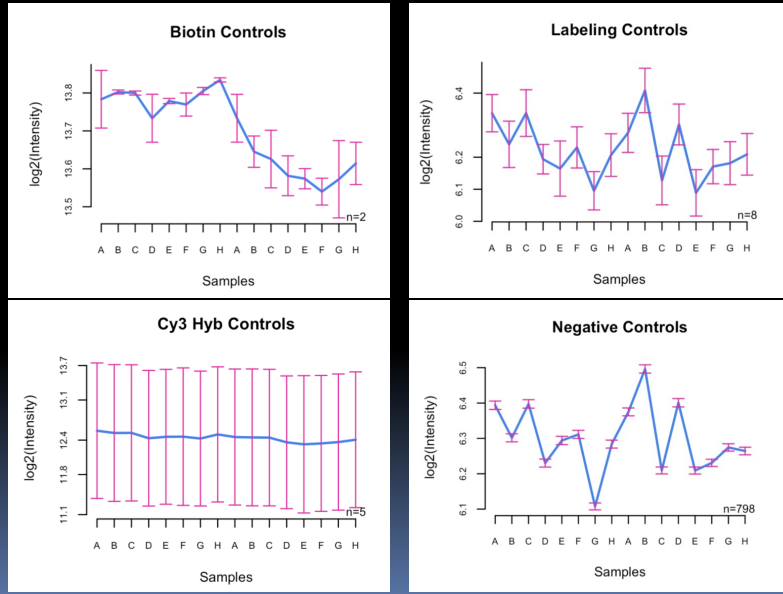
Effect of array position in Illumina BeadChips



Verdugo 2009, Nucl. Acids Res. 37:5610

20

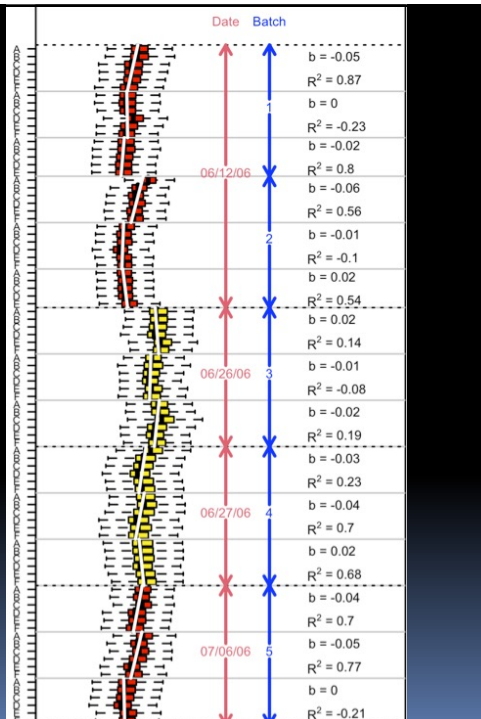
Effects control probes



21

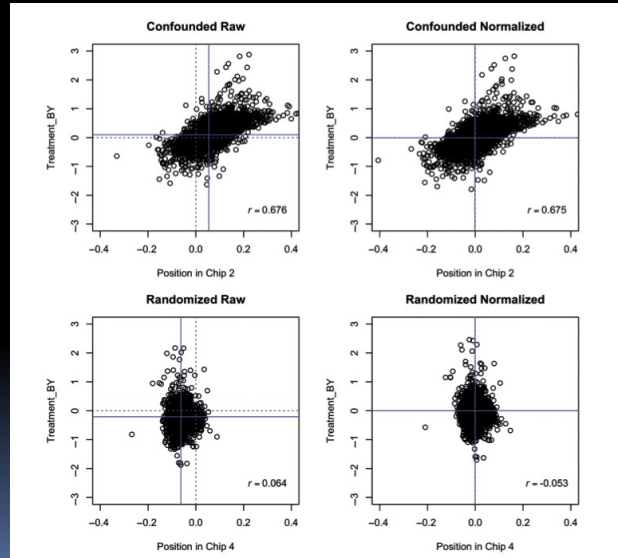
It may not
be position,
but time

Verdugo 2009, Nucl. Acids Res. 37:5610



22

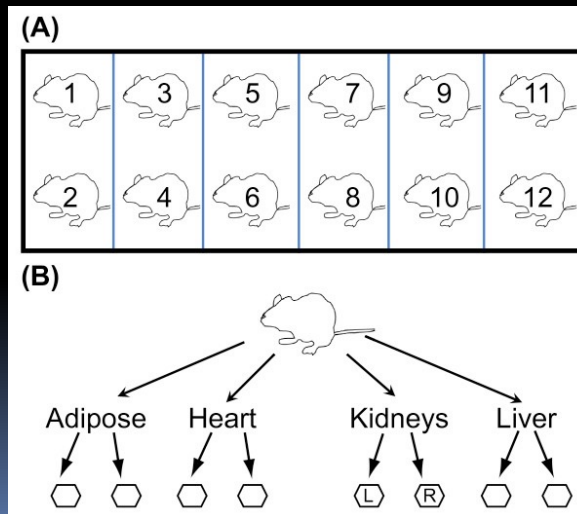
Randomization is critical



Verdugo 2009, Nucl. Acids Res. 37:5610

23

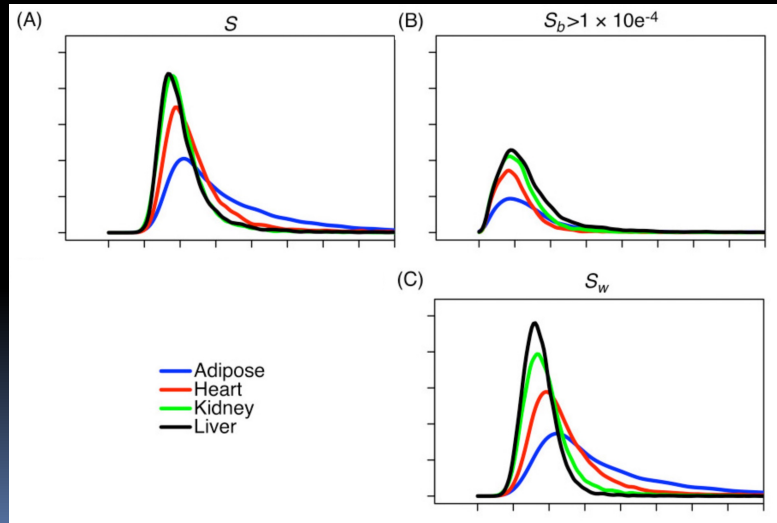
Technical variation may not come from microarray effects



Vedell
2011, BMC
Genomics
12: 167.

24

Within Mouse Variance can be larger than between mice



25

Experimental design also important for: RNA sequencing



8 lanes in one flow cell!

26

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.

27

III. Differential expression analysis

28

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 ($i=1 \rightarrow 5$)
ε_{ij}	residual effect

29

Sums of squares in a One-way ANOVA

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

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

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

$$SS(Total) = SS(A) + SSE$$

30

Statistical model in a Factorial design

- Gene by gene ANOVA

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

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 ($i=1 \rightarrow 2$)

B_j effect of the j^{th} treatment ($i=1 \rightarrow 2$)

AxB_{ij} effect of the ij^{th} treatment

ε_{ijk} residual effect

31

Gene expression for one gene

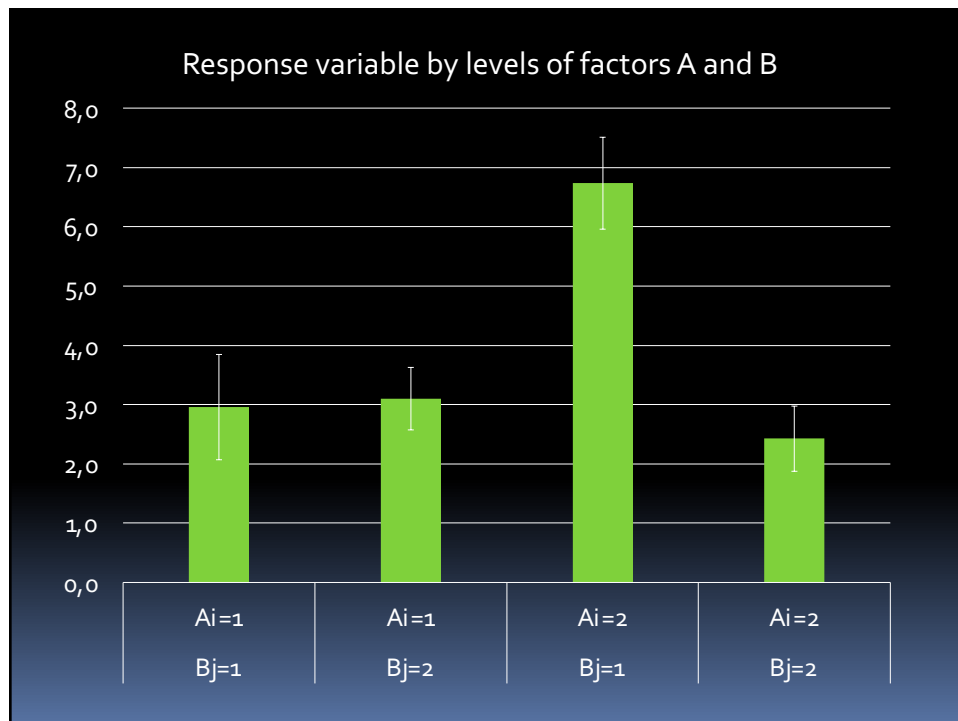
		Factor A	
		i=1	i=2
Factor B	j=1	3,0	6,7
		5,4	8,3
		2,0	5,2
		1,4	9,1
	j=2	1,5	2,3
		2,0	1,5
		3,2	4,0
		4,1	1,9

32

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}

33



34

Sums of squares in a factorial design

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

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

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

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

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

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

35

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)	

36

Hypothesis testing

- H_0 : Effect of A = 0
 - H_1 : Effect of A \neq 0
- $$F(A) = \frac{MS(A)}{MSE}$$
- H_0 : Effect of B = 0
 - H_1 : Effect of B \neq 0
- $$F(B) = \frac{MS(B)}{MSE}$$
- H_0 : Effect of AxB = 0
 - H_1 : Effect of AxB \neq 0
- $$F(AxB) = \frac{MS(AxB)}{MSE}$$
- Reject H_0 if $P < \alpha$ Reject H_0 if $fdr < FDR$

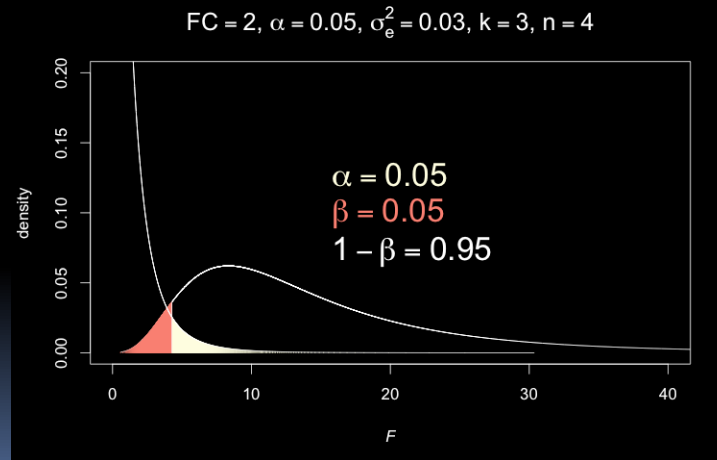
37

IV. Experimental design: Power and sample size

- What is appropriate n to detect DE?
- Find the power of a test at a given n
- $$F = \frac{MS_{genotype}}{MS_{error}} = \frac{SS_{genotype} / df_1}{SS_{error} / df_2}$$
- $$\beta = Pr(F_{H_1} | df_1, df_2, \lambda)$$
- $$\lambda = \frac{(k-1) n \sigma^2_{genotype}}{\sigma^2_{error}}$$

38

Power from an F test



39

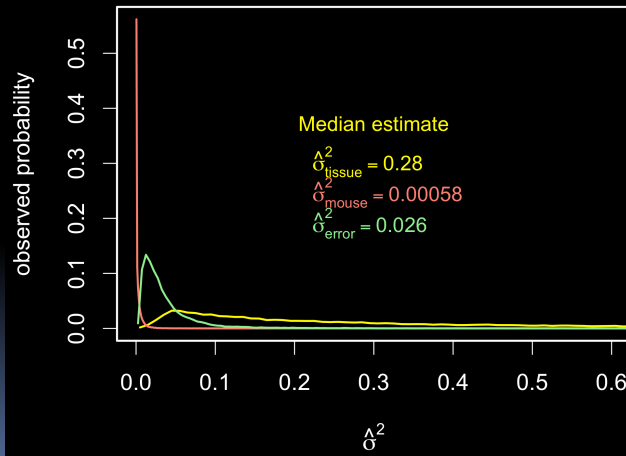
Mouse Illumina Tissue Survey

- 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

40

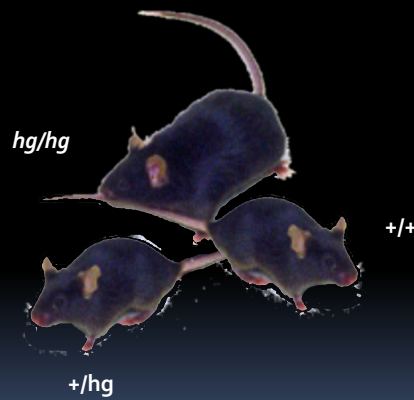
Variance of gene expression in Mice

Variance components in JAX Tissue Survey



41

Differential expression between genotypes



42

Theoretical Genetic Variance (treatment effect)

FC	τ_1 (a)	τ_2 (d)	τ_3 (-a)	$\sigma^2_{\text{genotype}}$
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

$$\tau = \log(\text{FC})/2$$

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

43

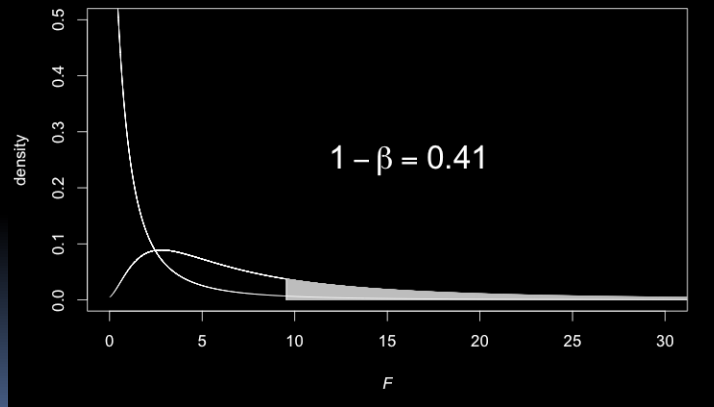
Power calculation for ANOVA

- $k = 3, n = 4, \alpha = 0.05$
- $\sigma^2_{\text{genotype(FC}_2)} = 0.08$
- $\sigma^2_{\text{error}} = \text{median}(\hat{\sigma}_e) + 4 * \text{median}(\hat{\sigma}_m) = 0.029$
- $\lambda = \frac{(k-1) n \sigma^2_{\text{genotype}}}{\sigma^2_{\text{error}}} = \frac{2 * 4 * 0.08}{0.029} = 22.07$
- $F_o = qf(1-0.05, 3-1, 3(4-1)) = 4.3$
- $\beta = Pr(F_{H_1} | df_1, df_2, \lambda)$
- $= pf(F_o, 3-1, 3(4-1), 22.07) = 0.05$
- $\text{Power} = 1 - \beta = 0.95$

44

Power from F test by n

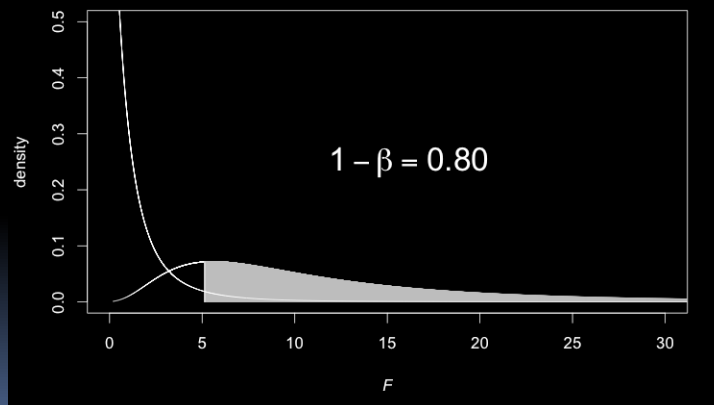
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 2$



45

Power from F test by n

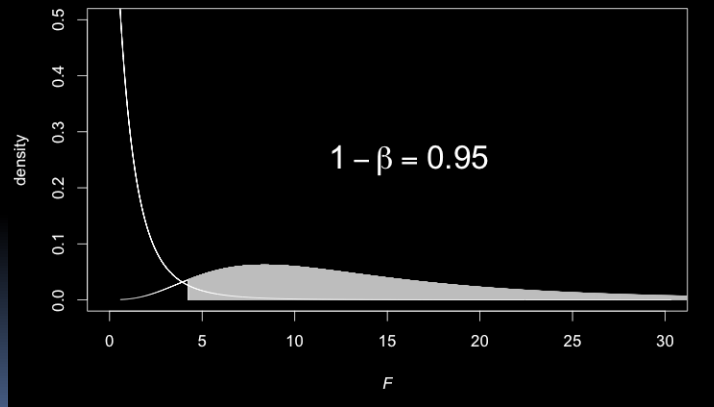
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 3$



46

Power from F test by n

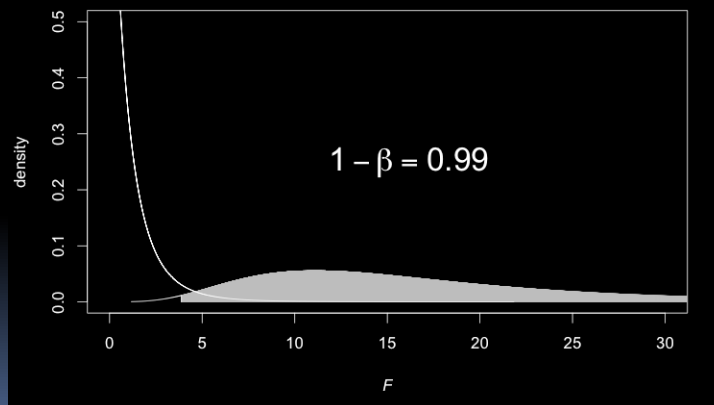
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 4$



47

Power from F test by n

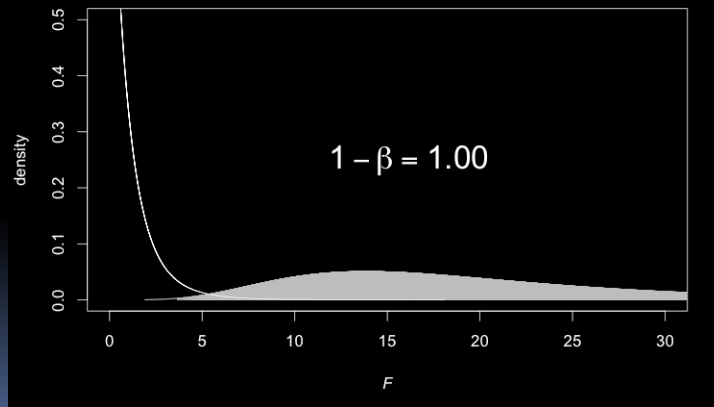
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 5$



48

Power from F test by n

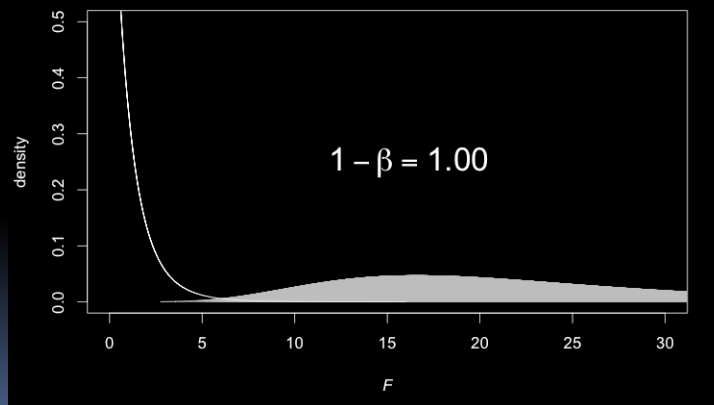
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 6$



49

Power from F test by n

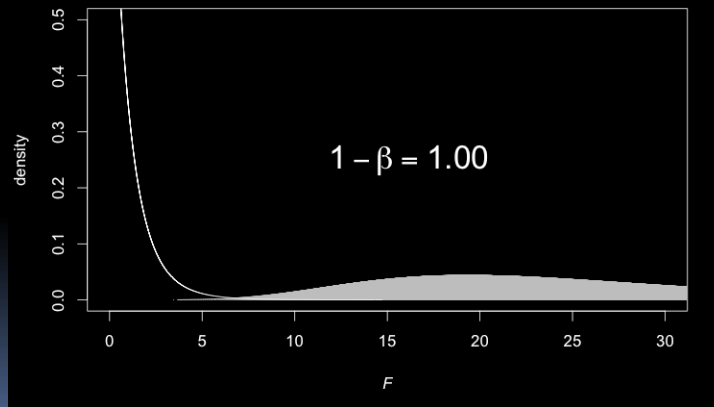
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 7$



50

Power from F test by n

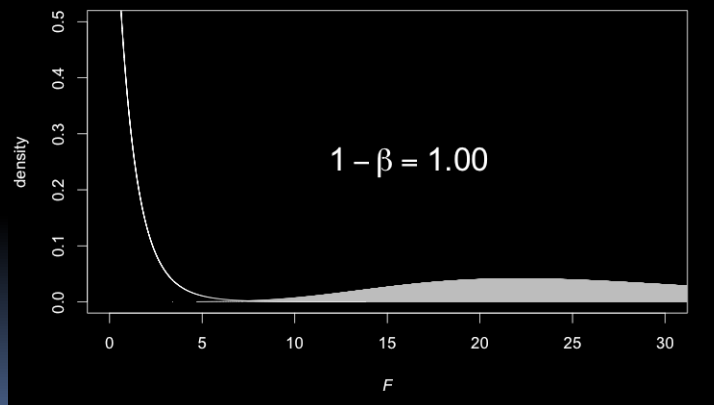
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 8$



51

Power from F test by n

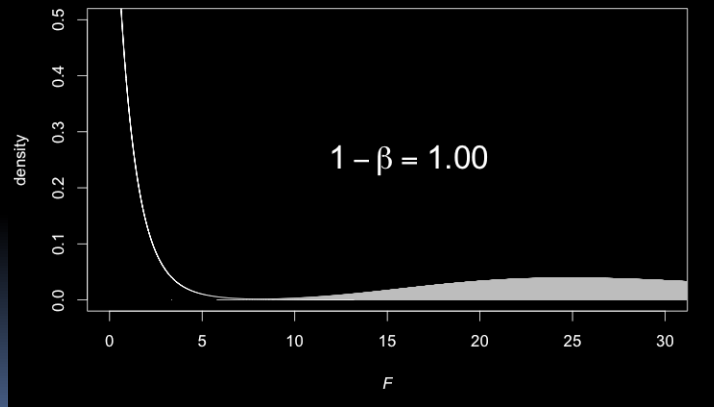
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 9$



52

Power from F test by n

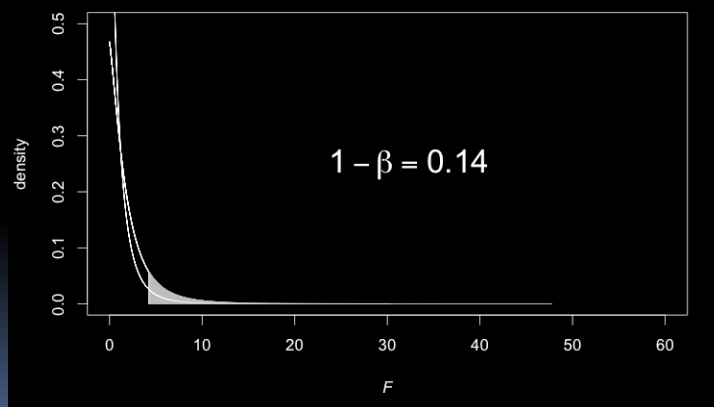
$FC = 2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 10$



53

Power from F test by FC

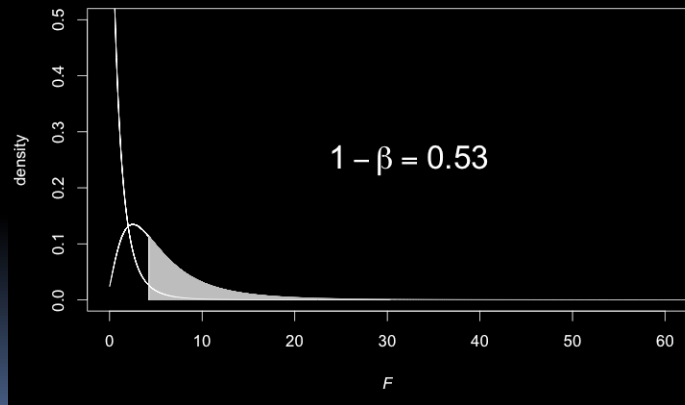
$FC = 1.2, \alpha = 0.05, \sigma_e^2 = 0.03, k = 3, n = 4$



54

Power from F test by FC

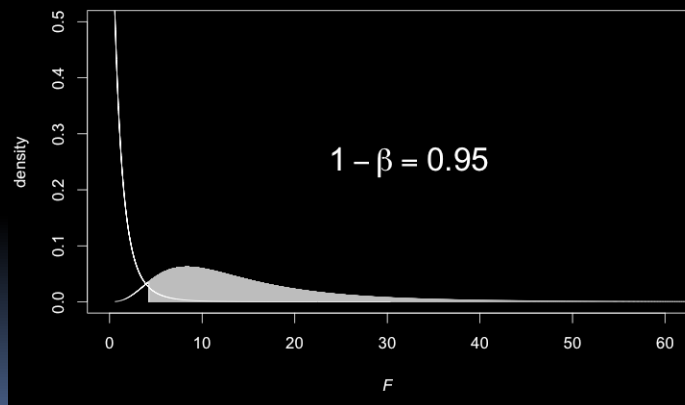
FC = 1.5, $\alpha = 0.05$, $\sigma_e^2 = 0.03$, $k = 3$, $n = 4$



55

Power from F test by FC

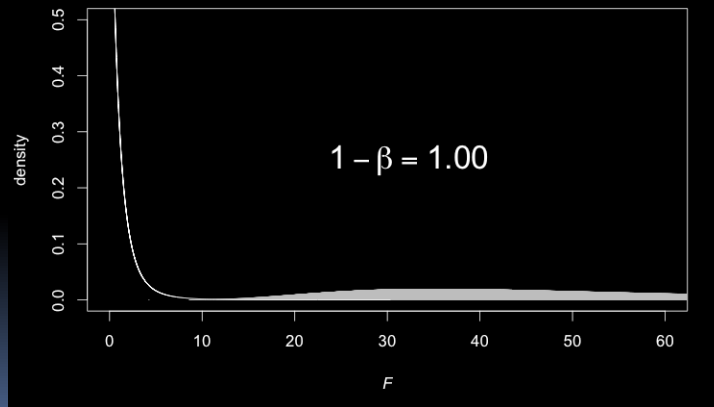
FC = 2, $\alpha = 0.05$, $\sigma_e^2 = 0.03$, $k = 3$, $n = 4$



56

Power from F test by FC

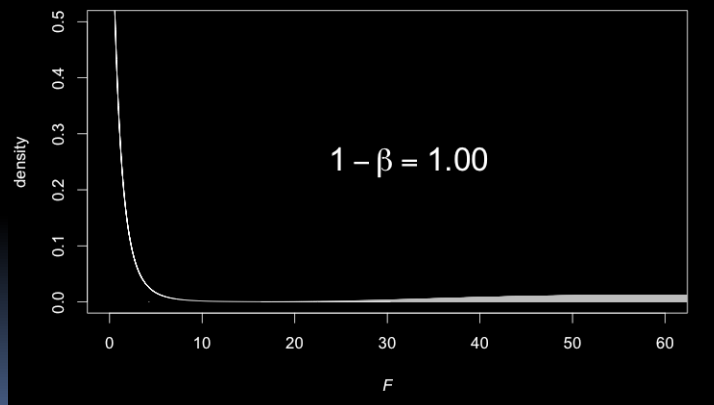
FC = 4, $\alpha = 0.05$, $\sigma_e^2 = 0.03$, $k = 3$, $n = 4$



57

Power from F test by FC

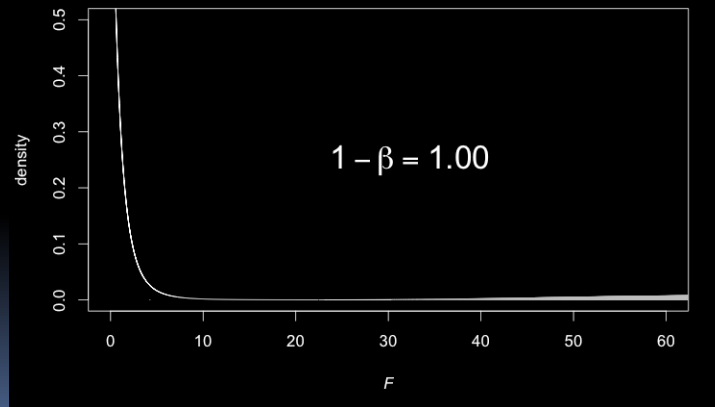
FC = 6, $\alpha = 0.05$, $\sigma_e^2 = 0.03$, $k = 3$, $n = 4$



58

Power from F test by FC

FC = 8, $\alpha = 0.05$, $\sigma_e^2 = 0.03$, $k = 3$, $n = 4$



59

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_0 is the proportion of non-differentially expressed genes

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

Bioinformatics 23: 739

Erratum: *Bioinformatics* 24: 149

60

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

61

Sample size in mouse Illumina experiments

- $m = 28000$ (number of genes)
- $FDR = 0.1$
- $\pi_0 = 0.90, 0.95, 0.99$
- $k = 3$
- $\sigma^2_{error} = 0.026$
- $FC = 2$

62

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)
```

63

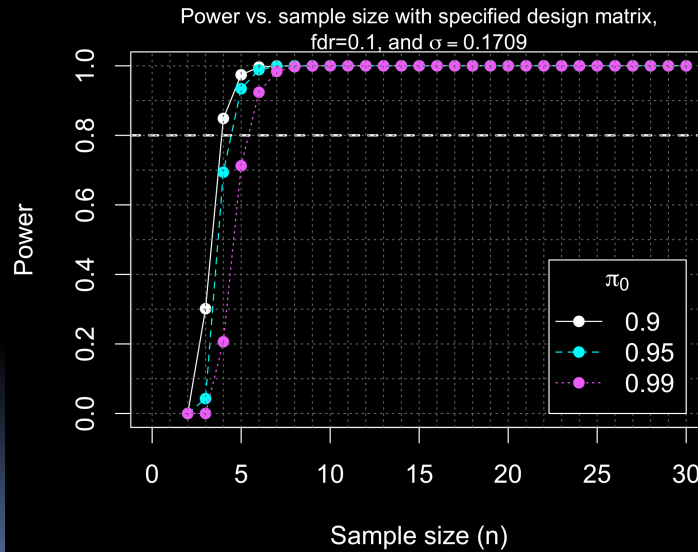
ssize.fdr output

```
> 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
```

64

Sample size in mouse Illumina experiments



65

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

66