



UNIVERSIDAD  
DE CHILE

# Introducción a Microarreglos de DNA

Luis Valenzuela Villa,  
luis.valenz.v@gmail.com

Laboratorio de Genética, Facultad de Ciencias,  
Laboratorio de BioMatemática y Ómica Integrativa, Facultad de Medicina,  
Universidad de Chile.

6 de Septiembre, 2017



UNIVERSIDAD DE CHILE

# Introducción a Microarreglos de DNA

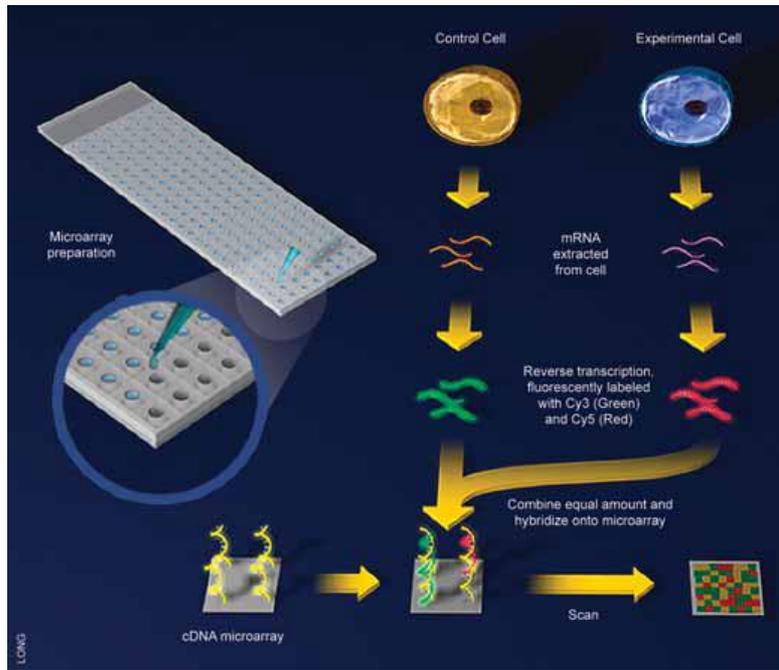
## Algunas Tecnología de microarreglos



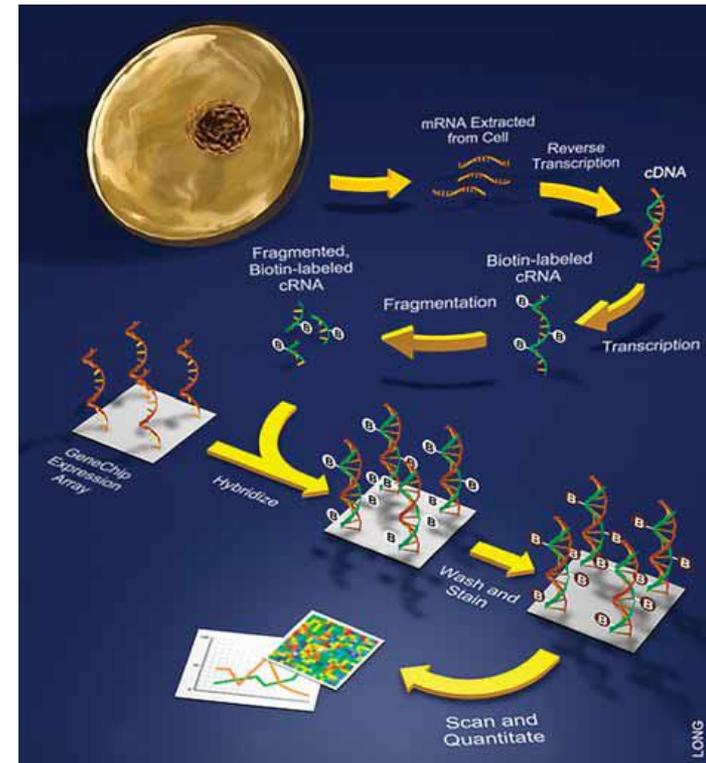
- Un color
- Spotted cDNA
- Agilent spotted probes

- Dos colores
- AffymetrixGeneChip
- Nimblegene
- Illumina BeadChip

Long J. 2006



Spotted cDNA

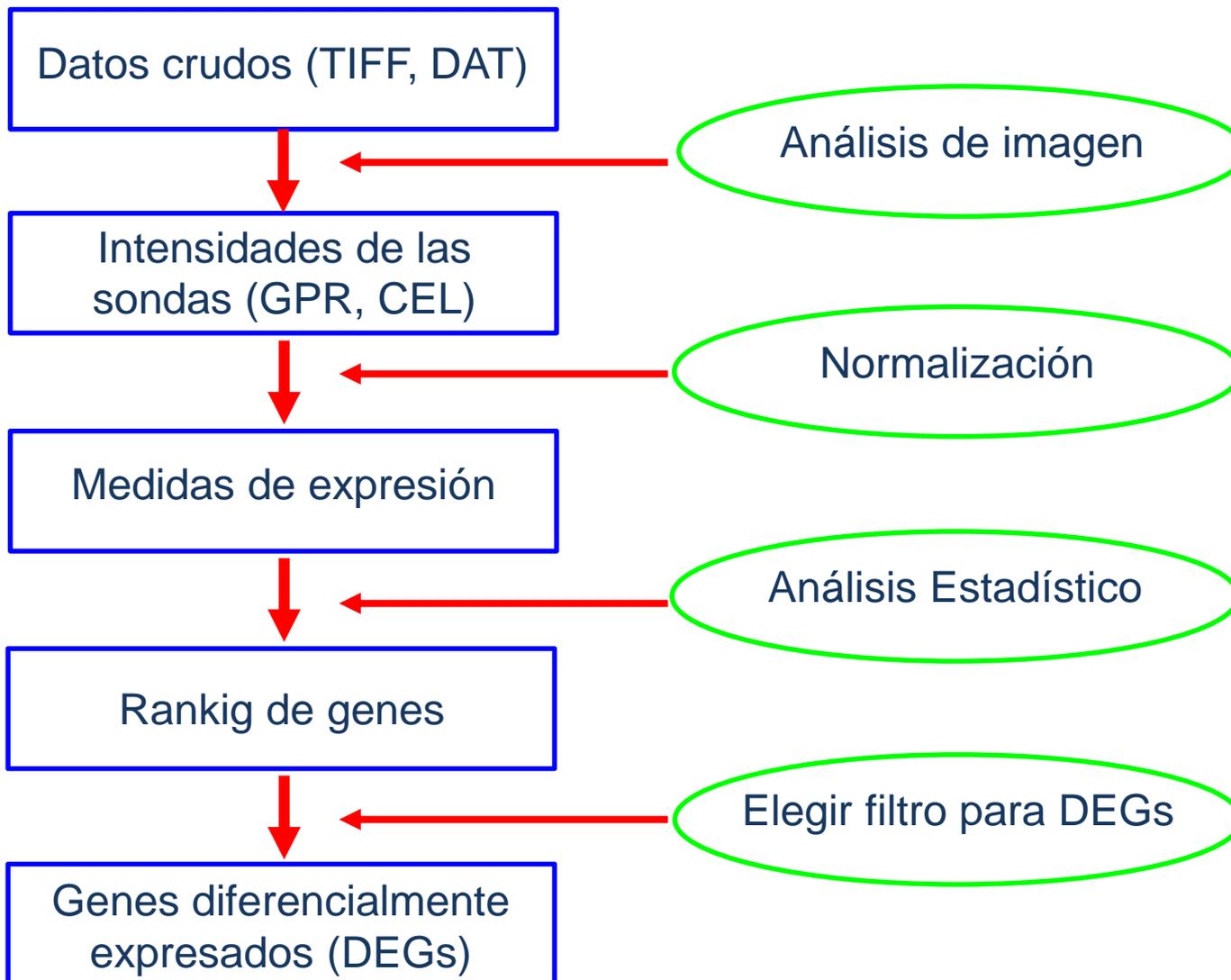


AffymetrixGeneChip



# Introducción a Microarreglos de DNA

## Flujo de trabajo con microarreglos





UNIVERSIDAD  
DE CHILE

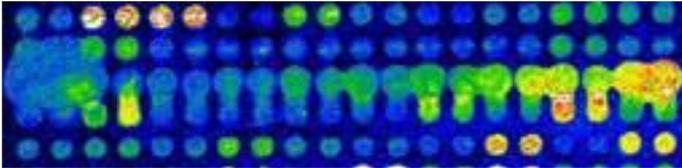
# Introducción a Microarreglos de DNA

## Análisis de imagen

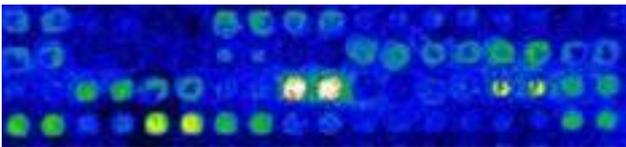


- Tipos de atributos observados en los spots, índices entre 0 y 1 (Wang et al., 2001):

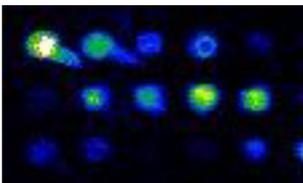
- Tamaño del spot.



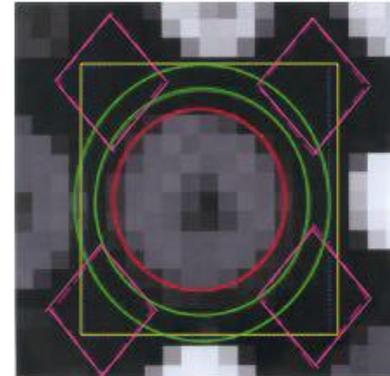
- Saturación.



- Tamaño del background local.



- Radio señal ruido.





UNIVERSIDAD  
DE CHILE

# Introducción a Microarreglos de DNA

## Normalización



Por qué?

- Efecto de trabajar con fluorescencias distintas entre canales.
- Error experimental: cambios no controlados.
  
- Tipos de normalización
  - Normalización interna.
  - Normalización entre arrays.
  
- Efecto de la normalización
  - Lleva datos a la misma escala.



# Introducción a Microarreglos de DNA

## Normalización: Robust Multi-Array Average (RMA)



- Corrige por background en función de las intensidades de las sondas con Perfect Match (PM) en cada chip.
- Los valores corregidos son transformados con log2.
- Se normaliza usando el método de los cuantiles.

Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
1	1	10.33333	10.33333	14
1	2	6.66667	5	8.66667
1	3	5	6.66667	6.66667
2	1	8.66667	8.66667	5
2	2	12	14	12
2	3	14	12	10.33333



- Se establece un modelo lineal para la intensidad de cada una de las sondas

$$Y_{ij} = m_i + a_j + e_{ij}$$

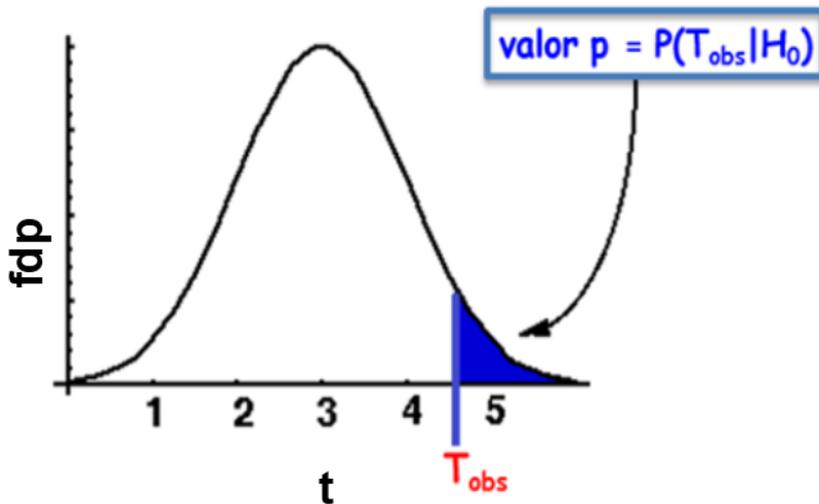
Irizarry et al. (2003).

$m_i$ : expresión log del conjunto de sondas en el chip

$a_j$ : afinidad de la sonda

$e_{ij}$ : error

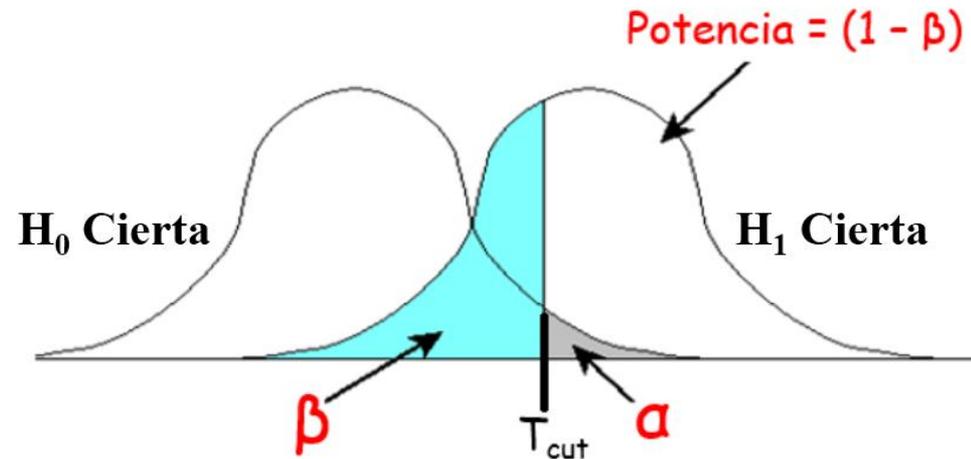
Fisher (1925)



Define valor p:  
Probabilidad de observar los datos,  
dado que la hipótesis nula es  
verdadera.

(¡Es una variable aleatoria que  
depende de n!)

Neyman-Pearson



Definen las probabilidades  $\alpha$ ,  $\beta$  y  $(1 - \beta)$



### Decisión

Condición Real

Rechazar  $H_0$

No Rechazar  $H_0$

$H_0$  Verdadera



**Error tipo I**  
(falso positivo)



Acierto  
(verdadero positivo, VP)

$H_0$  Falsa



Acierto  
(verdadero negativo, VN)

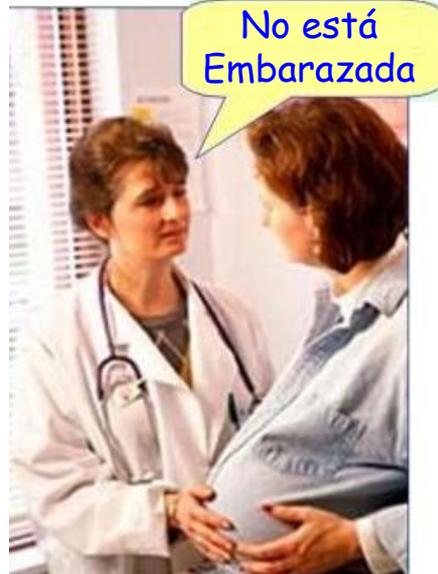


**Error tipo II**  
(falso negativo)

**Error tipo I**  
falso positivo



**Error tipo II**  
falso negativo



$$P(\text{error tipo I}) = P(\text{Rechazar } H_0 | H_0 \text{ es Verdadera}) = \alpha$$

$$P(\text{error tipo II}) = P(\text{No Rechazar } H_0 | H_0 \text{ es Falsa}) = \beta$$

$$(1 - \beta) = \text{Potencia} = \text{Sensibilidad} = \text{Tasa VP}$$

$$(1 - \alpha) = \text{Especificidad} = \text{Tasa VN}$$



Expresión de 1 gen en dos condiciones:

$$P(\text{error tipo I}) = P(\text{Rechazar } H_0 | H_0 \text{ es Verdadera}) = \alpha$$

Fijamos en 0.05, es decir una confianza de 0.95 de no cometer error tipo I (falso positivos)

Ahora si comparo 5 genes en dos condiciones, la probabilidad de no cometer error tipo I en ninguna de las pruebas sería:

$$0.95 * 0.95 * 0.95 * 0.95 * 0.95 = 0.7738$$

Supongamos que tenemos un microarreglo o RNA-seq con 20.000 genes, la probabilidad de no cometer error tipo I:

$$0.95^{20000} \sim 0$$



# Introducción a Microarreglos de DNA

## Múltiples Pruebas: Family Wise Error Rate (FWER)



Condición Real	Rechazar Ho	No Rechazar Ho	Total
Ho Verdadera	V: <b>Error tipo I (<math>\alpha</math>)</b> <b>(falso positivo)</b>	U: Acierto (verdadero positivo, VP)	$m_0$
Ho Falsa	S: Acierto (verdadero negativo, VN)	T: <b>Error tipo II</b> <b>(falso negativo)</b>	$m - m_0$
Total	R	$m - R$	m

m: número total de hipótesis;

$m_0$ , número de hipótesis nulas

V: número de falsos positivos;

T: número de falsos negativos

S,U: número de verdaderos negativos y positivos, respectivamente.

R: número de rechazos.

FWER:  $P(V > 0)$ , es decir, que cometamos uno o más errores de falsos positivos.

$$\text{FWER} = 1 - P(\text{no rechazar } H_0 \mid H_0 \text{ verdadera}) = 1 - (1 - \alpha)^{m_0}$$

Cuando m es grande, FWER tiende a 1.



# Introducción a Microarreglos de DNA

## Múltiples Pruebas: Corrección de Bonferroni



Condición Real	Rechazar $H_0$	No Rechazar $H_0$	Total
$H_0$ Verdadera	V: <b>Error tipo I (<math>\alpha</math>)</b> <b>(falso positivo)</b>	U: Acierto (verdadero positivo, VP)	$m_0$
$H_0$ Falsa	S: Acierto (verdadero negativo, VN)	T: <b>Error tipo II</b> <b>(falso negativo)</b>	$m - m_0$
Total	R	$m - R$	m

m: número total de hipótesis;

$m_0$ , número de hipótesis nulas

V: número de falsos positivos;

T: número de falsos negativos

S,U: número de verdaderos negativos y positivos, respectivamente.

R: número de rechazos.

Cómo elegimos  $\alpha$  si queremos controlar FWER?

$$\alpha = \alpha_{\text{FWER}}/m.$$

Si hago 3 pruebas, mi nuevo  $\alpha$  límite será:

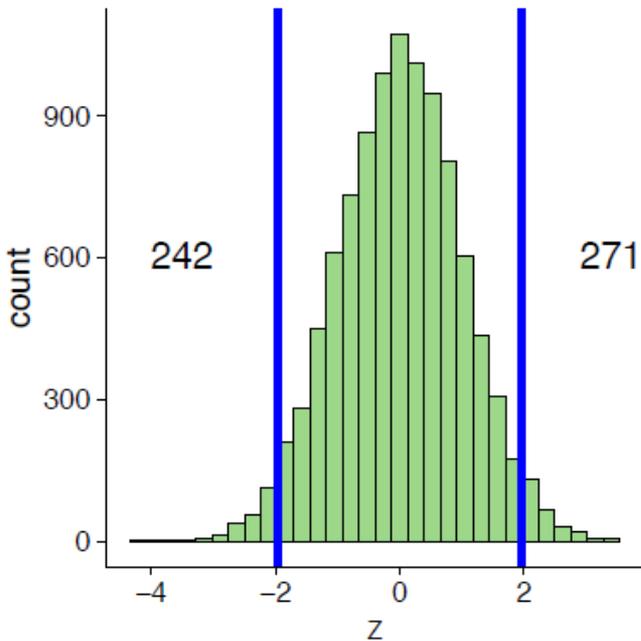
$$\alpha = 0.05/3 = 0.0167$$

Es decir, sólo rechazaría la hipótesis nula cuando el valor p sea menor a 0.0167

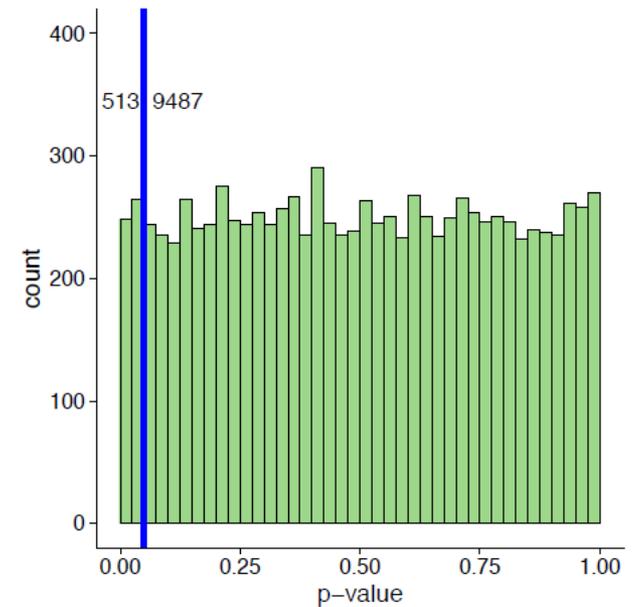


# Introducción a Microarreglos de DNA

## Múltiples Pruebas: False Discovery Rate (FDR)



Bajo hipótesis nula los valores p se distribuyen de modo uniforme.



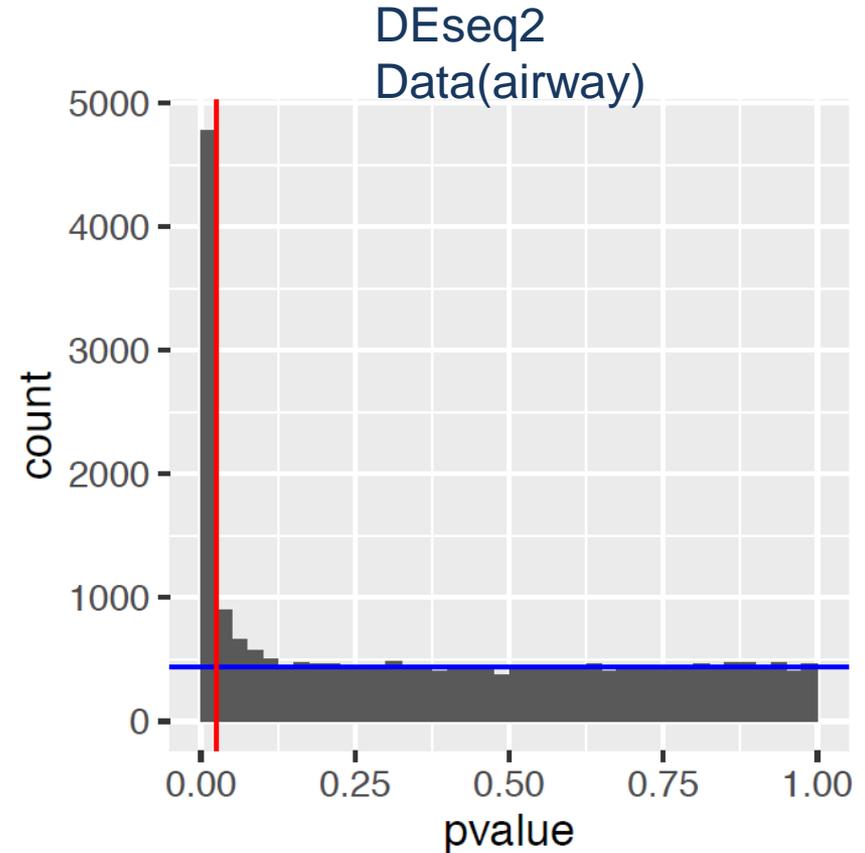


Hay 4783 genes con valores p entre 0 y  $\alpha$  (línea roja), y bajo la línea azul el conteo es 439 genes.

Podemos estimar la tasa de rechazos falsos, como:

$$\frac{439}{4783} = 0.092$$

~Tasa de falsos descubrimientos:  
FDR.





# Introducción a Microarreglos de DNA

## Múltiples Pruebas: False Discovery Rate (FDR)



Condición Real	Rechazar Ho	No Rechazar Ho	Total
Ho Verdadera	V: <b>Error tipo I (<math>\alpha</math>)</b> <b>(falso positivo)</b>	U: Acierto (verdadero positivo, VP)	$m_0$
Ho Falsa	S: Acierto (verdadero negativo, VN)	T: <b>Error tipo II</b> <b>(falso negativo)</b>	$m - m_0$
Total	R	$m - R$	m

$$\text{FDR} = E \left[ \frac{V}{\max(R, 1)} \right]$$

Proporción promedio de los errores tipo I con respecto a los rechazos realizados.



# Introducción a Microarreglos de DNA

## Múltiples Pruebas: FDR con Benjamini y Hochberg



Sea  $k$  el mayor  $i$  para el cual  $p_i \leq i\alpha/m$ :  
Rechazar los primeros  $k$  tests.

Supongamos  $\alpha = 0.05$ , y valores  $p$ :

0.0001,

0.0002,

0.0005,

0.001,

0.002,

0.003,

0.007,

0.01,

0.02,

0.11

$0.0001 \leq 1 * 0.005$ ,

$0.0002 \leq 2 * 0.005 = 0.01$ ,

$0.0005 \leq 3 * 0.005 = 0.015$ ,

$0.001 \leq 4 * 0.005 = 0.02$ ,

$0.002 \leq 5 * 0.005 = 0.025$ ,

$0.003 \leq 6 * 0.005 = 0.03$ ,

$0.007 \leq 7 * 0.005 = 0.035$ ,

$0.01 \leq 8 * 0.005 = 0.04$ ,

$0.02 \leq 9 * 0.005 = 0.045$ ,

$0.11 > 10 * 0.005 = 0.05$

Con Bonferroni:

$\alpha = 0.05/10 = 0.005$



FDR = 0.1

$\pi = m_0/m = 0.9, 0.95, 0.99$

\$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



# Introducción a Microarreglos de DNA

## Recordatorio Estadística Clásica: Pruebas de Hipótesis



Décimas para el promedio:

- 1 muestra:  $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$        $t_{n-1} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$

- 2 muestras independientes (con y sin homocedasticidad):

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_c^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}, \quad s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t_{g.l.} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad g.l. = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left( \frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left( \frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

- 2 muestras pareadas:

$$t_{n-1} = \frac{\bar{d}}{s_d / \sqrt{n}}$$



Inconvenientes de la Prueba de t simple:

- Suposiciones paramétricas difíciles de justificar con pocos arrays.
- La varianza en muestras pequeñas puede ser ruidosa.
- Genes con un fold change pequeño pueden ser significativos desde el punto de vista estadístico, pero desde el punto de vista biológico.



# Introducción a Microarreglos de DNA

## Prueba de t moderada



Prueba de t moderada, aproximación bayesiana:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{\nu_g}}$$

$S_0$ : Estimación de variación total;

$S_g$ : variación por gen;

$\beta_g$ : diferencia de promedios

$d_0$ : Estimación de grados de libertad

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

- En vez de solo estimar variabilidad para cada gen, incluye la información global de todos los genes.
- Elimina la ocurrencia de grandes valores de t accidentales debido a pequeñas varianzas en cada gen.



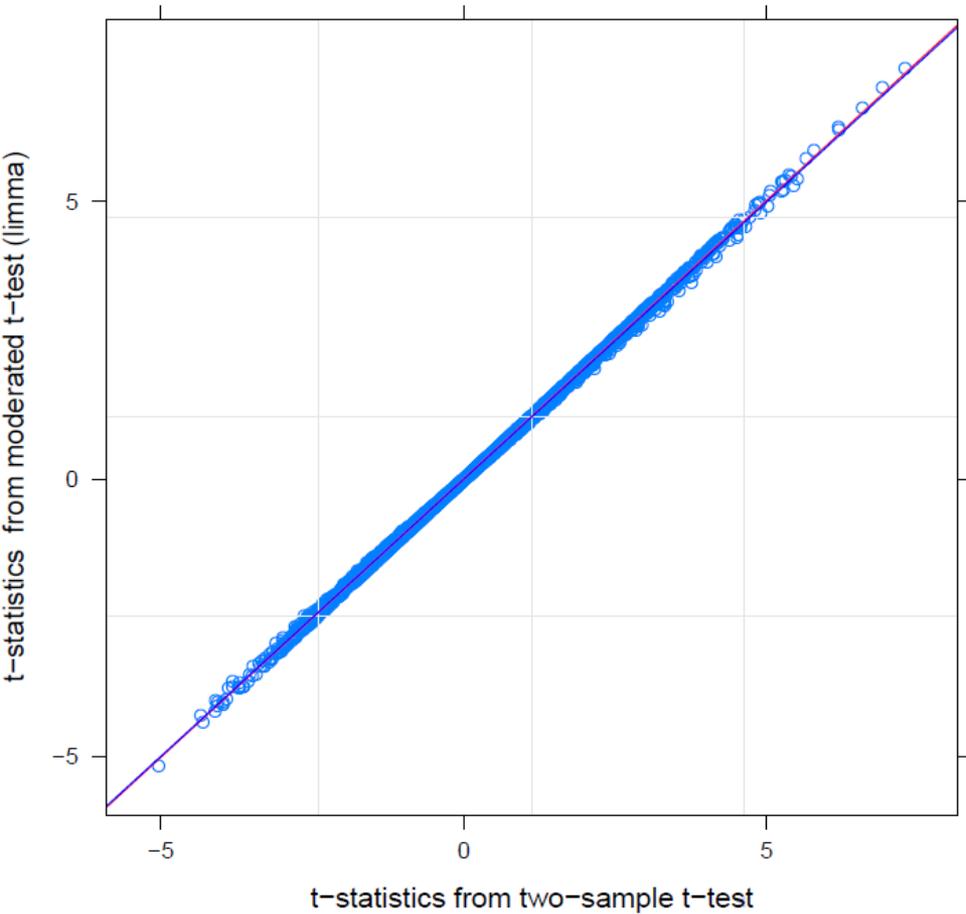
UNIVERSIDAD  
DE CHILE

# Introducción a Microarreglos de DNA

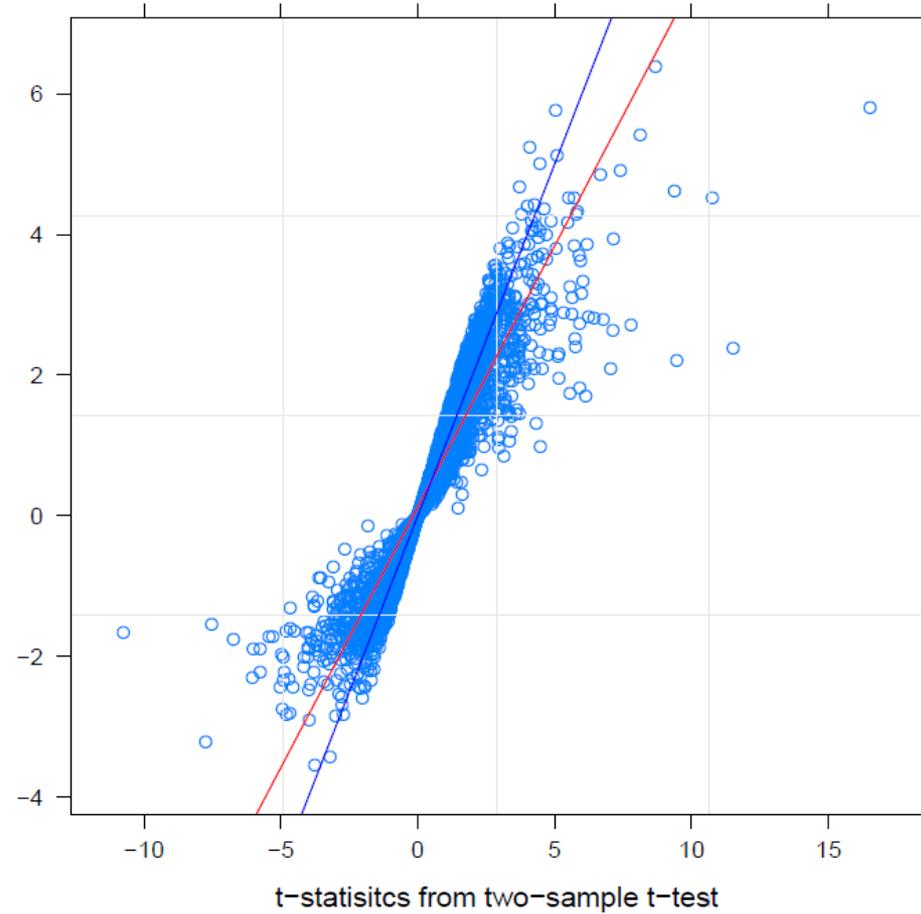
## Prueba de t moderada



70 muestras



6 muestras





Limma: “Linear Models for Microarray and RNA-Seq Data”

$$y_j = \beta_0 + \beta_1 a_{ij} + \epsilon$$

- Definir una matriz de diseño para establecer los parámetros del modelo lineal con `model.matrix()`
- Ajustar un modelo lineal para cada gen con `conlmFit()`
- Calcular las pruebas de t moderadas y el F-moderado con `eBayes()`



UNIVERSIDAD DE CHILE

# Introducción a Microarreglos de DNA NCBI Gene Expression Omnibus (GEO)



Gene Expression Omnibus

GEO Publications | FAQ | MIAMI | Email GEO

Login

NCBI Reso

GEO DataSe

HOME SEARCH

NCBI > GEO > A

GEO help: Mous

Scope: Self

Series GSE4

Status

Title

Organism

Experiment t

Summary

Author

Customize ...

Attribute name

tissue (0)

strain (28)

Customize ...

Publication dates

30 days

1 year

Custom range...

Clear all

Show additional filt

Contributor(s) Sameith K, Kemm

Citation(s) Sameith K, Amini  
gene expression a  
exposes potential  
23;13:112. PMID

Submission date Nov 27, 2012

Last update date Mar 24, 2016

Contact name Patrick Kemmerer

Organization name UMC Utrecht

Department Department of M

Lab Holstege Lab

Street address Universiteitsweg 3

City Utrecht

State/province Utrecht

ZIP/Postal code 3584 CG

Country Netherlands

Platforms (1) GPL11232 A-UMC

Samples (287) GSM636383 dot6  
More... GSM636384 dot6  
GSM636460 isw2

Relations  
BioProject: PRNA182291

Analyze with GEO2R

Downloa  
SOFT formatted family file(s)  
MINIML formatted family file(s)  
Series Matrix File(s)

Supplementary file

GSE42536\_RAW.tar

GSE42536\_final\_GeneExpressionMatr

GSE42536\_protocols.txt.gz

Processed data included within Sample table

Raw data provided as supplementary file

NCBI

GEO Publications | FAQ | MIAMI | Email GEO

NCBI > GEO > GEO2R > GSE42536

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. Full instructions [View](#)

GEO accession GSE42536 Set Transcription profiling by array of yeast single and double deletion mutants of gene-specific transcription factors

Samples Define groups

Enter a group name: List

Group	Accession	Title	Source name 1	Source name 2	Strain (Ch1)	Genotype (Ch1)	Strain (Ch2)	Genotype (Ch2)
A	GSM636383	dot6	refpool	dot6-del	BY4742	wild type	BY4742	dot6-del
B	GSM636384	dot6	refpool	refpool	BY4742	dot6-del	BY4742	wild type
-	GSM636460	isw2	refpool	refpool	BY4742	isw2-del	BY4742	wild type
-	GSM636461	isw2-del-1-c	refpool	isw2-del	BY4742	wild type	BY4742	isw2-del
-	GSM636501	nhp6a-del-1-b	refpool	nhp6a-del	BY4742	wild type	BY4742	wild type
-	GSM636502	nhp6b-del-3-a	refpool	nhp6b-del	BY4742	wild type	BY4742	nhp6b-del
-	GSM636503	nhp6b-del-3-b	refpool	nhp6b-del	BY4742	wild type	BY4742	wild type
-	GSM636524	rph1-del-2-a	refpool	rph1-del	BY4742	wild type	BY4742	rph1-del
-	GSM636525	rph1-del-2-b	refpool	rph1-del	BY4742	wild type	BY4742	wild type
-	GSM636594	snt1-del-1-a	refpool	snt1-del	BY4742	wild type	BY4742	snt1-del
-	GSM636595	snt1-del-1-b	refpool	snt1-del	BY4742	wild type	BY4742	wild type
-	GSM636609	sum1-del-1-a	refpool	sum1-del	BY4742	wild type	BY4742	sum1-del
-	GSM636610	sum1-del-1-b	refpool	sum1-del	BY4742	wild type	BY4742	wild type
-	GSM636628	tod6-del-1-a	refpool	tod6-del	BY4742	wild type	BY4742	tod6-del
-	GSM636629	tod6-del-1-b	refpool	tod6-del	BY4742	wild type	BY4742	wild type
-	GSM819521	cat8-del-1-b	refpool	cat8-del	BY4742	wild type	BY4742	wild type
-	GSM819531	gis1-del-1-a	refpool	gis1-del	BY4742	wild type	BY4742	gis1-del
-	GSM819532	gis1-del-1-b	refpool	gis1-del	BY4742	wild type	BY4742	wild type
-	GSM819579	msn2-del-2-a	refpool	msn2-del	BY4742	wild type	BY4742	msn2-del

GEO2R Value distribution Options Profile graph R script

```
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
# R scripts generated Mon Sep 4 10:19:11 EDT 2017

#####
# Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(limma)

# load series and platform data from GEO

gset <- getGEO("GSE42536", GSEMatrix = TRUE, AnnotGPL = FALSE)
if (length(gset) > 1) idx <- grep("GPL11232", attr(gset, "sample")) else idx <- 1
```

VITV