Introducción a Microarreglos de DNA

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

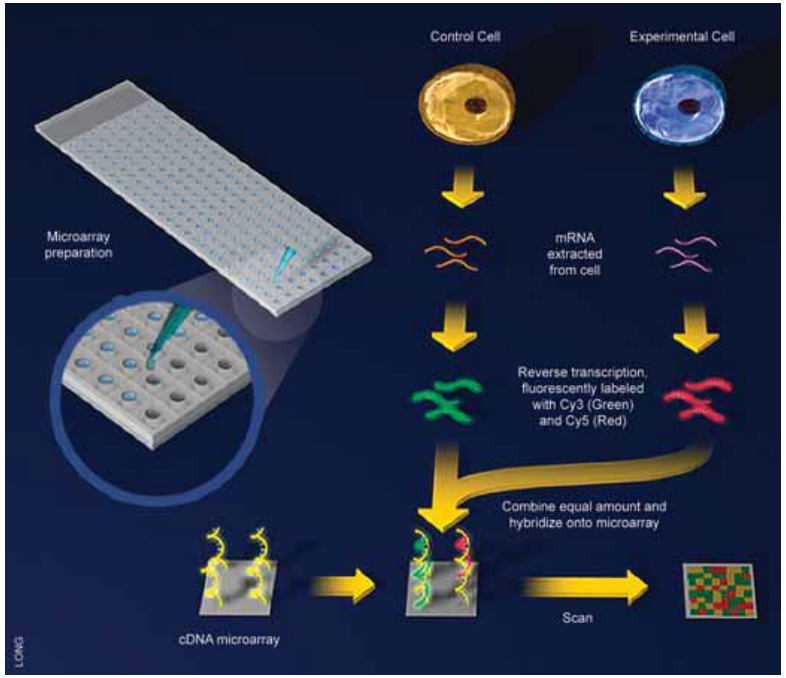
Laboratorio de Genética, Facultad de Ciencias,
Laboratorio de BioMatemática y Ómica Integrativa, Facultad de Medicina,
Universidad de Chile.
6 de Septiembre, 2017

Introducción a Microarreglos de DNA
Algunas Tecnología de microarreglos

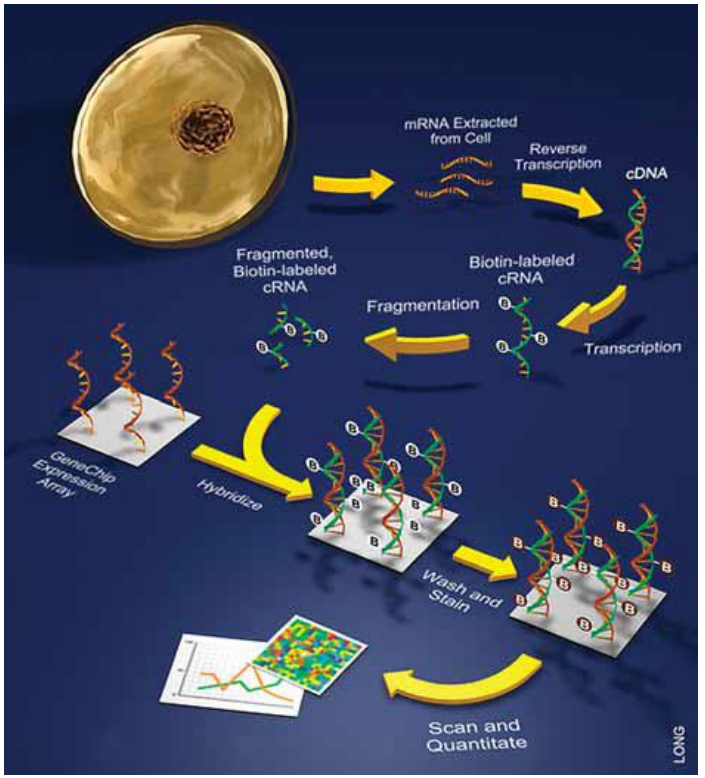
- Un color
  Spotted cDNA
  Agilent spotted probes

- Dos colores
  AffymetrixGeneChip
  Nimblegene
  Illumina BeadChip

Long J. 2006

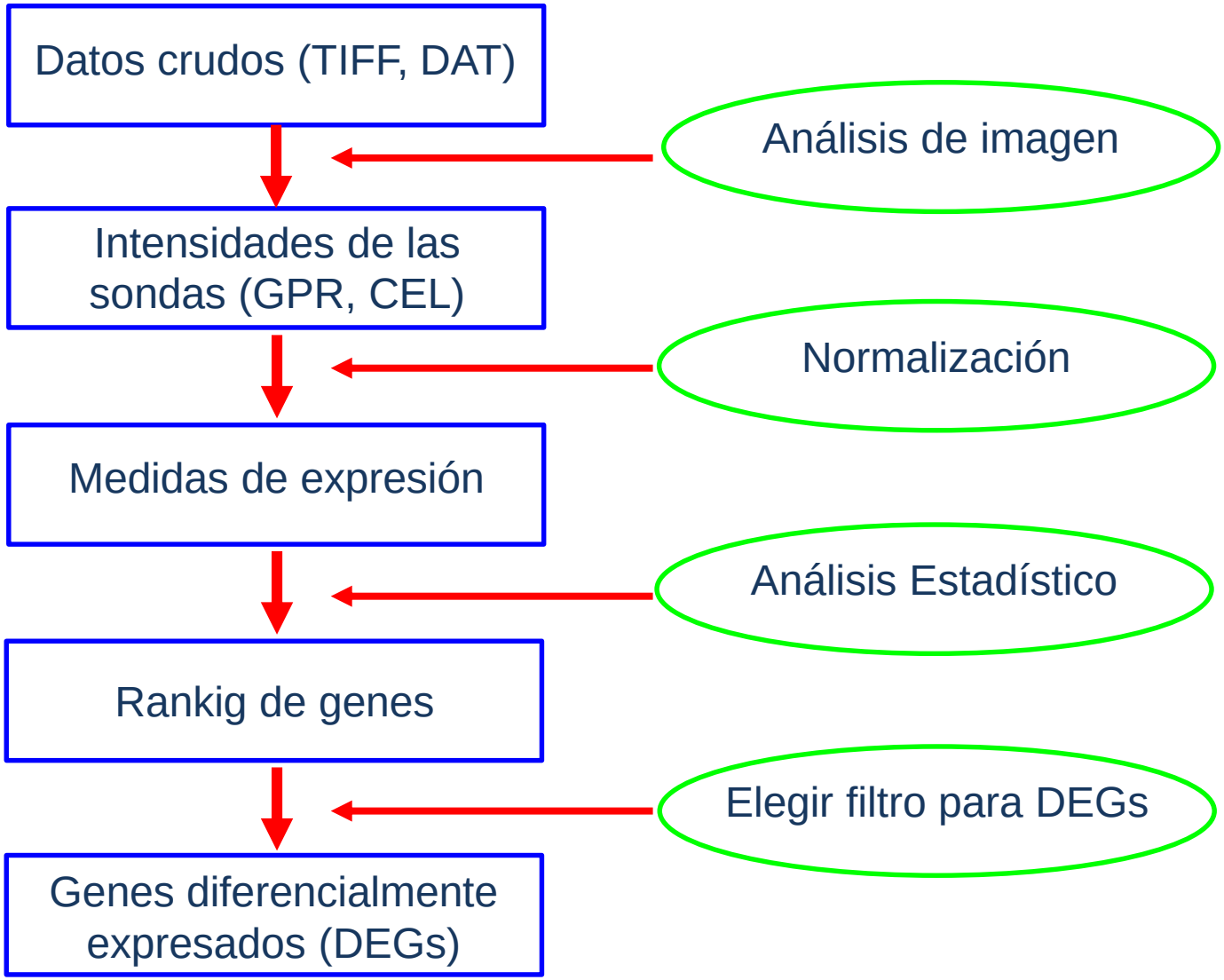


Spotted cDNA



AffymetrixGeneChip

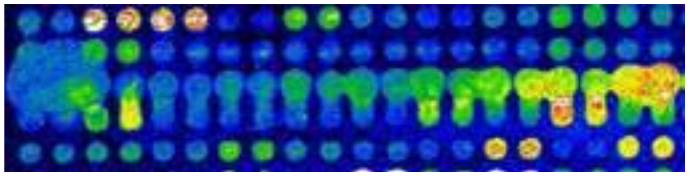
Introducción a Microarreglos de DNA
Flujo de trabajo con microarreglos



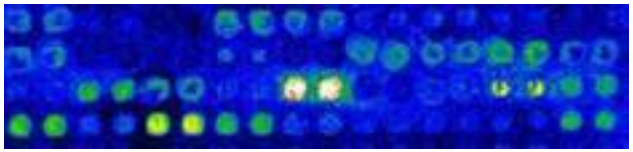
Introducción a Microarreglos de DNA
Análisis de imagen

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

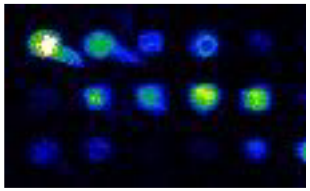
- Tamaño del spot.



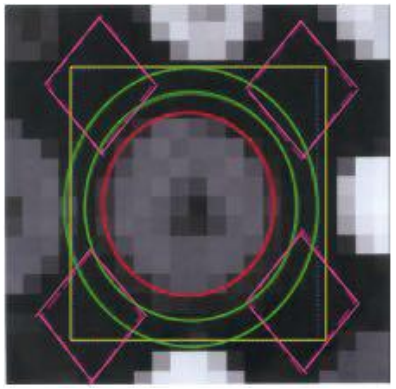
- Saturación.



- Tamaño del background local.



- Radio señal ruido.



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*

Introducción a Microarreglos de DNA
Normalización: Robust Multi-Array Average (RMA)

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

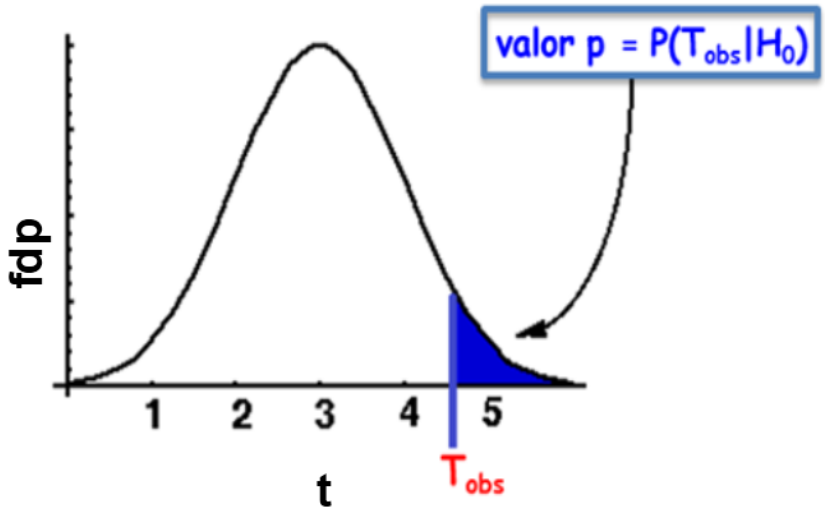
$$Y_{ij} = m_i + a_j + e_{ij}$$

Irizarry et al. (2003).

mi: expresión log del conjunto de sondas en el chip
aj: afinidad de la sonda
eij: error

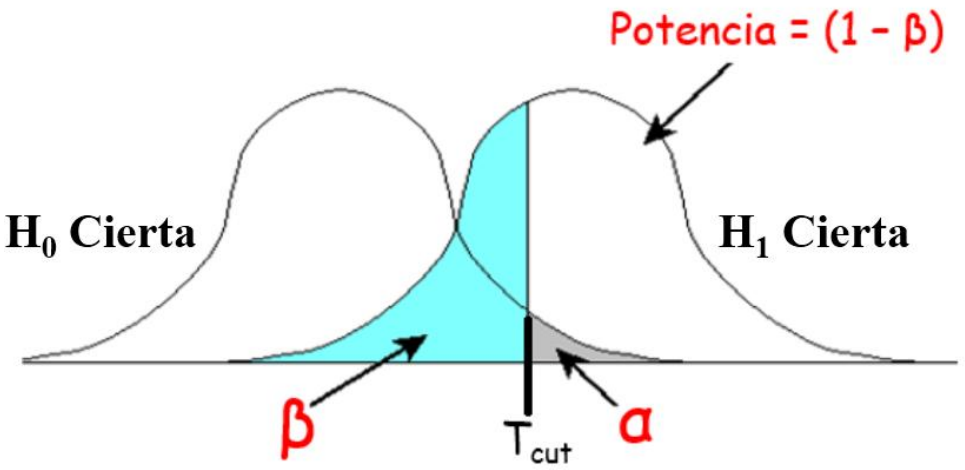
Introducción a Microarreglos de DNA
Recordatorio Estadística Clásica: Pruebas de Hipótesis

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 α, β y (1 - β)

Introducción a Microarreglos de DNA
Recordatorio Estadística Clásica: Pruebas de Hipótesis

Condición Real	Decisión: Rechazar Ho	Decisión: No Rechazar Ho
Ho Verdadera	Error tipo I (falso positivo)	Acierto (verdadero positivo, VP)
Ho Falsa	Acierto (verdadero negativo, VN)	Error tipo II (falso negativo)

Error tipo I
falso positivo

Error tipo II
falso negativo

P(error tipo I) =
P(Rechazar Ho| Ho es Verdadera) = α

P(error tipo II) =
P(No Rechazar Ho| Ho es Falsa) = β

(1- β) = Potencia =
= Sensibilidad = Tasa VP

(1- α) = Especificidad =
= Tasa VN

Expresión de 1 gen en dos condiciones:

P(error tipo I) = P(Rechazar Ho| Ho es Verdadera) = α

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: Error tipo I (α) (falso positivo)	U: Acierto (verdadero positivo, VP)	m_o
Ho Falsa	S: Acierto (verdadero negativo, VN)	T: Error tipo II (falso negativo)	$m-m_o$
Total	R	m-R	m

m: número total de hipótesis; mo, 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.

FWER = 1 – P(no rechazar Ho | Ho verdadera) = $1 - (1 - \alpha)^{m0}$

Cuando m es grande, FWER tiende a 1.

Introducción a Microarreglos de DNA
Múltiples Pruebas: Corrección de Bonferroni

Condición Real	Rechazar Ho	No Rechazar Ho	Total
Ho Verdadera	V: Error tipo I (α) (falso positivo)	U: Acierto (verdadero positivo, VP)	m_o
Ho Falsa	S: Acierto (verdadero negativo, VN)	T: Error tipo II (falso negativo)	$m-m_o$
Total	R	m-R	m

m: número total de hipótesis; mo, 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 α si queremos controlar FWER?

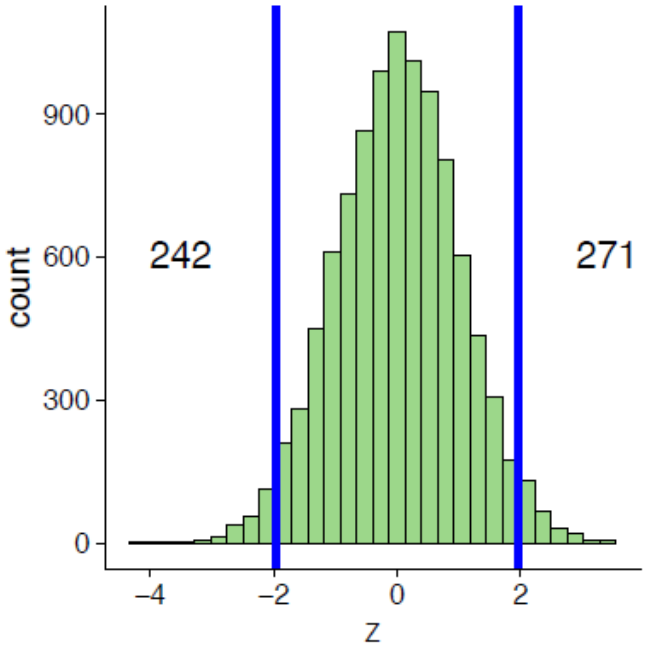
$\alpha = \alpha_{FWER}/m$.

Si hago 3 pruebas, mi nuevo α límite será:

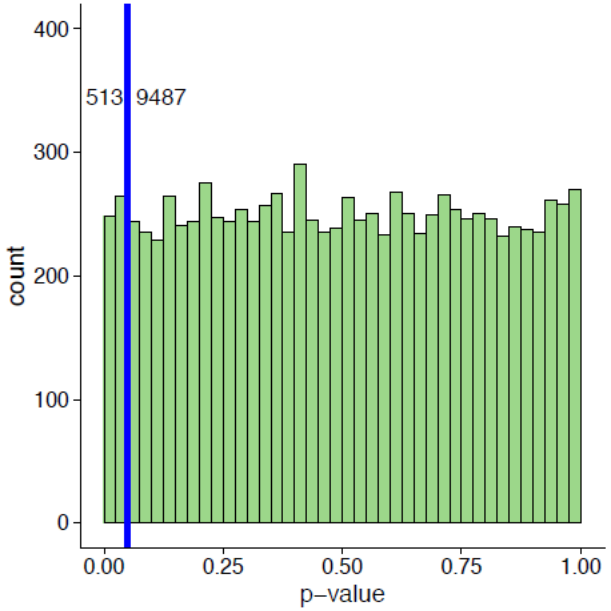
$\alpha = 0.05/3 = 0.0167$

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.

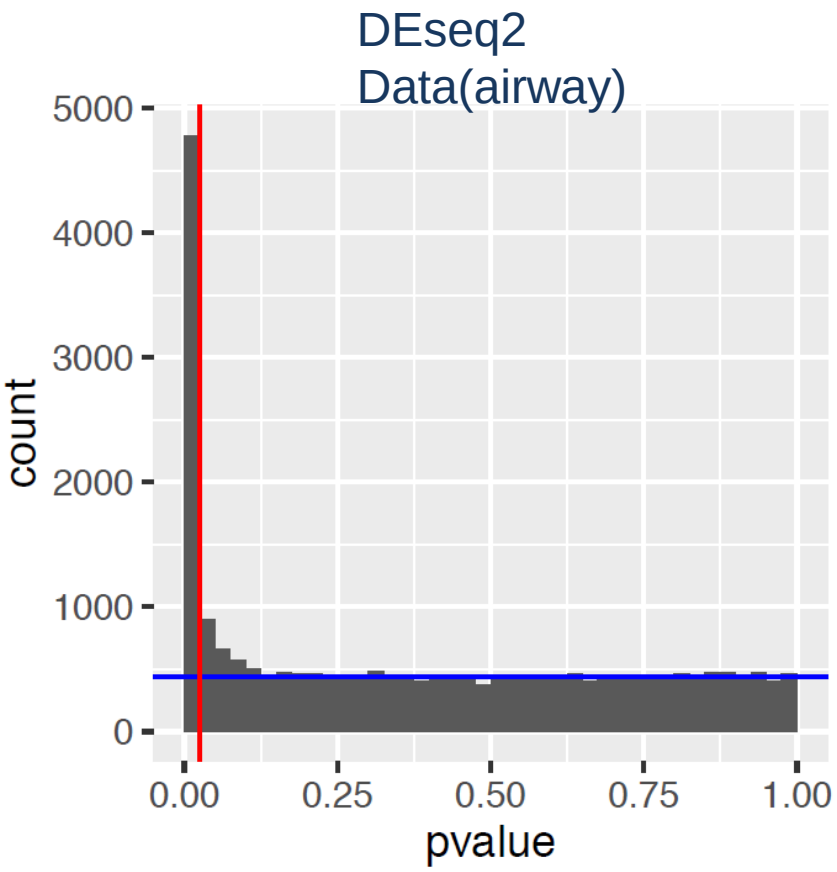


Introducción a Microarreglos de DNA
Múltiples Pruebas: False Discovery Rate (FDR)

Hay 4783 genes con valores p entre 0 y α (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: Error tipo I (α) (falso positivo)	U: Acierto (verdadero positivo, VP)	m_o
Ho Falsa	S: Acierto (verdadero negativo, VN)	T: Error tipo II (falso negativo)	$m-m_o$
Total	R	m-R	m

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

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 α = 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≤1*0.005,
0.0002≤2*0.005=0.01,
0.0005≤3*0.005=0.015,
0.001≤4*0.005=0.02,
0.002≤5*0.005=0.025,
0.003≤6*0.005=0.03,
0.007≤7*0.005=0.035,
0.01≤8*0.005=0.04,
0.02≤9*0.005=0.045,
0.11>10*0.005=0.05

Con Bonferroni:
α =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

Dócimas para el promedio:

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

- 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:

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

Introducción a Microarreglos de DNA

Recordatorio Estadística Clásica: Pruebas de Hipótesis

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.

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;

β_g: diferencia de promedios

d_0: Estimación de grados de libertad

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

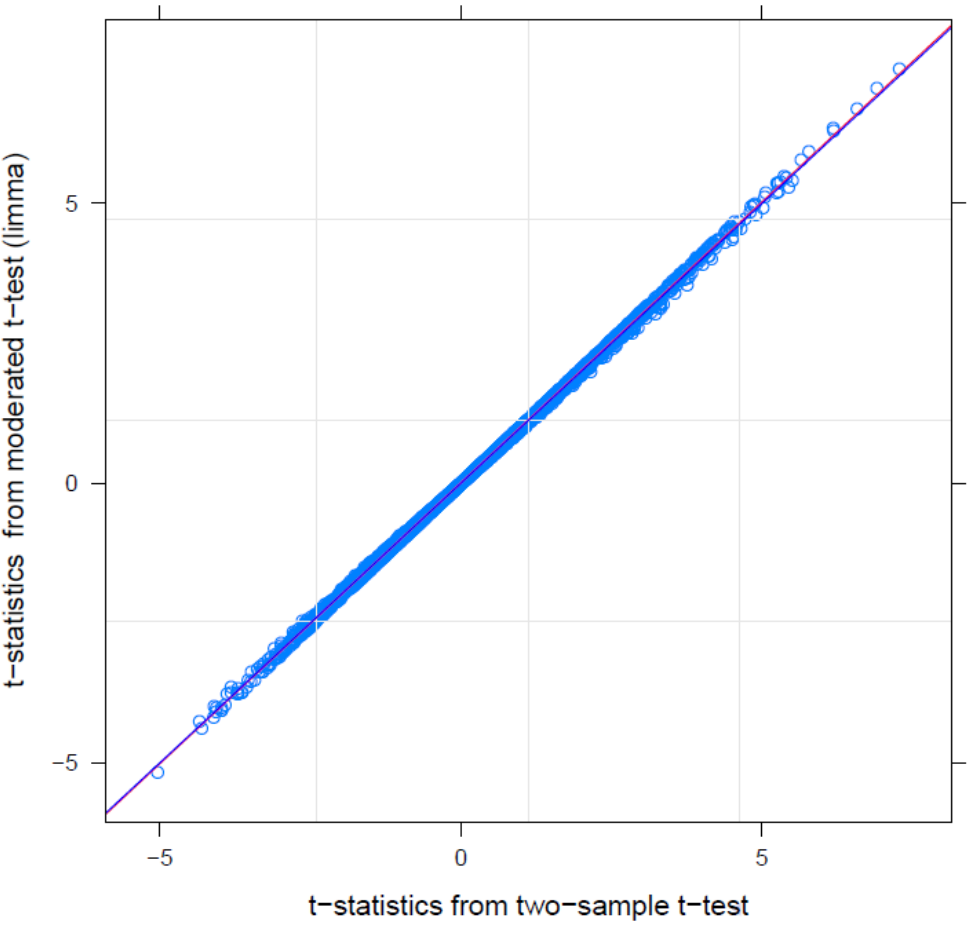
- En vez de solo estimar variabilidad para cada gen, incluye la información global de todos los genes.

- Elimina la ocurrencia de grandes valores de t accidentales debido a pequeñas varianzas en cada gen.

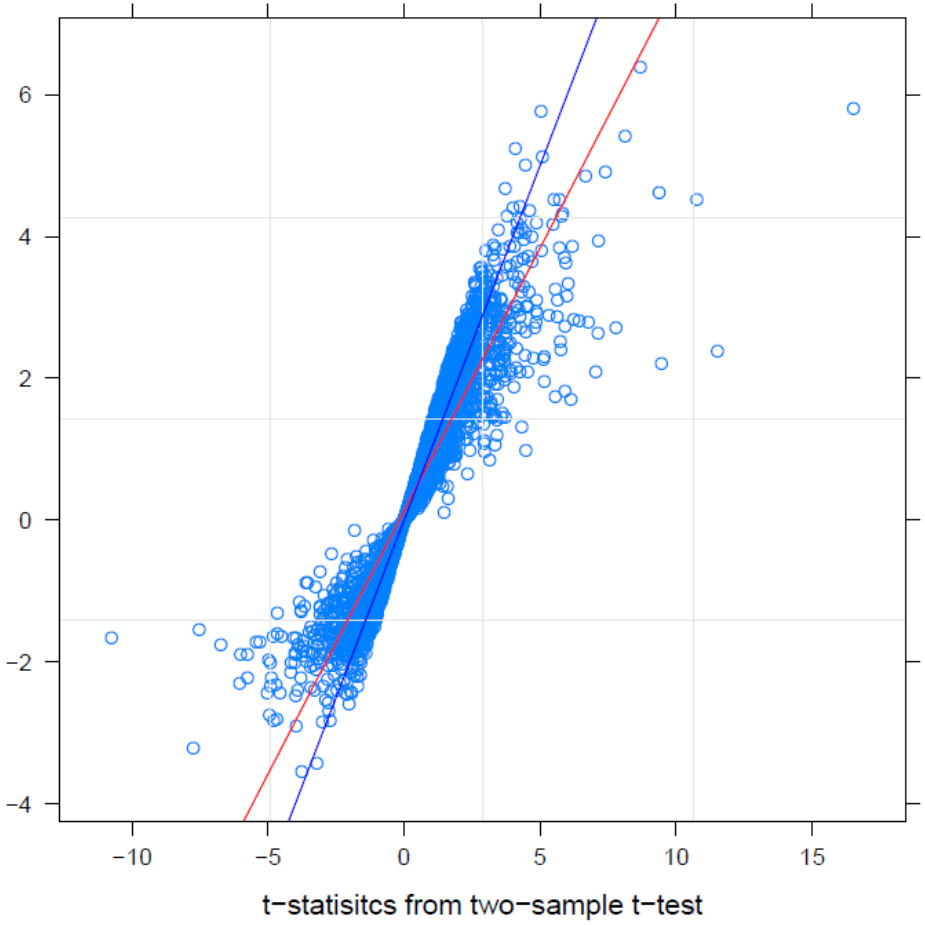
Introducción a Microarreglos de DNA

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 conlmFit()
- Calcular las pruebas de t moderadas y el F-moderado con eBayes()

Introducción a Microarreglos de DNA

NCBI Gene Expression Omnibus (GEO)