Análisis de datos de microarrays

Alex Sánchez-Pla y M. Carme Ruíz de Villa
Departament d'Estadística. Universitat de Barcelona.
Facultat de Biologia. Avda. Diagonal 643. 08028 Barcelona. Spain.
asanchez@ub.edu;mruiz_de_villa@ub.edu

xx
PID_00191035

Módulo

Índice

Part I

Preliminares

1. Selección de genes diferencialmente expresados

1.1 Introducción

El motivo más habitual para el que se suelen utilizar microarrays es la búsqueda de genes cuya expresión cambia entre dos o más condiciones experimentales, por ejemplo a consecuencia de un tratamiento, una enfermedad u otras causas (distintos tiempos, distintas lineas celulares, ...).

El problema consiste en identificar estos genes y suele denominarse *selección de genes diferencialmente expresados ("DEG")* o bien comparación de clases.

El problema de seleccionar genes diferencialmente expresados se traduce de manera casi inmediata al problema estadístico de comparar variables y, en años recientes, se han desarrollado un gran número de métodos estadísticos para resolverlo. La mayoría son extensiones de los métodos estadísticos clásicos –pruebas-t o análisis de la varianza– adaptados en uno u otro sentido para tener en cuenta las peculiaridades de los microarrays.

Aunque el problema de la selección de genes diferencialmente expresados puede relacionarse directamente con la realización de pruebas estadísticas, en el caso de los microarrays, el hecho de que haya dos tecnologías que miden la expresión de dos formas distintas hace que se deba diferenciar la metodología a emplear en cada caso. Los arrays de dos colores combinan dos muestras en un chip y generan una medida de expresión relativa. Esto hace que para comparar dos muestras de un mismo individuo sean la opción naturalmente más apropiada

En el caso de querer comparar muestras independientes de diferentes individuos los arrays de un color son la mejor opción. Evidentemente, lo m'as común será disponer de una sola técnica y tener que adaptar los análiaisis estadísticos a la misma.

Vamos a plantear un posible esquema de trabajo, para situar la mejor opción en cada caso:

- Situación 1: experimento con 5 individuos diabéticos y 5 no diabéticos, independientes entre si (muestras independientes)

  – Caso 1: Arrays de cDNA (2 colores): Utilizaríamos 5 arrays Diabético/Referencia y 5 arrays No diabético/Referencia

– Caso 2: Arays de Affymetrix (1 color): Utilizaríamos 5 arrays de Diabético y 5 arrays de No diabético

- Situación 2: experimento con 6 individuos de los que se ha tomado una muestra de tejido sano y otra de tejido tumoral (muestras apareadas o dependientes)

– Caso 3: Arrays de cDNA (2 colores): Utilizaríamos 6 arrays, uno por individuo, y en cada uno se realizaría la comparaci'on Tejido Tumoral/Tejido sano.

– Caso 4: Arays de Affymetrix (1 color): Utilizaríamos 12 arrays, 2 por individuo, 6 con muestras de Tejido Tumoral y 6 con muestras de Tejido sano.

1.1.1 Medidas *naturales* para comparar dos muestras

Recordemos que una vez se han hecho los experimentos con microarrays y NA valor detectado por el escaner.Esto hace que algunas operaciones que se realicen tengan en cuenta esta característica. Segun si la comparación a realizar se llevará a cabo con datos *independientes* (2 muestras, casos 1 y 2) o con datos *dependientes* (muestras apareadas, casos 3, 4) algunas medidas *naturales* o razonables para la comparación de expresiones son las siguientes:

- Para comparaciones directas, con expresiones relativas entre muestras apareadas o bien diferencias apareadas de expresiones absolutas:

– log ratio promedio : $\overline{R} = \frac{1}{n}\sum_{i=1} R_i$

– t–test de una muestra $\frac{\overline{R}}{SE}$, donde SE estima el error estándar del *log ratio* promedio

– t–test robusto: Substituir en el anterior medidas robustas del error estándar

- Para comparaciones indirectas entre muestras independientes de expresiones relativas o absolutas:

– Diferencia media $\overline{R}_1 - \overline{R}_2 = \frac{1}{n}_1 \sum_{i=1}^{n_1} R_i - \frac{1}{n}_2 \sum_{j=1}^{n_2} R_j$

– t-test (clásico) de dos muestras $\frac{\overline{R}_1 - \overline{R}_2}{SE_{12}\sqrt{\frac{1}{n}_1 + \frac{1}{n}_2}}$

– t-test robusto de dos muestras: Substituir en el anterior medidas robustas del error estándar

Al "ratio" o razón de expresiones se suele denominar también "Fold Change (FC)" porque en arrays de dos colores representaba cuantas veces más expresado está el gen en una (R) que en otra condición (G). Al logRatio también se le llama logFC y por extensión a la diferencia de medias en escala logarítmica tambien se la denomina logFC dado que se realiza de forma implícita la aproximación siguiente: $\overline{X}_1 - \overline{X}_2 = \overline{\log Y}_1 - \overline{\log Y}_2 \simeq \log(\overline{Y}_1) - \log(\overline{Y}_2) = \log(\frac{\overline{Y}_1}{\overline{Y}_2})$

Un primer ejemplo

Consideremos la tabla siguiente que representa una matriz de expresión simplificada que contiene las expresiones relativas (por ejemplo entre tejido tumoral y sano del mismo individuo) de 5 genes en 6 muestras.

Gen	R1	R2	R3	R4	R5	R6
A	2.50	2.70	2.50	2.80	3.20	2.00
B	0.01	0.05	-0.05	0.01	0.00	0.00
C	2.50	2.70	2.50	1.80	20.00	1.00
D	0.50	0.00	0.20	0.10	-0.30	0.30
E	0.10	0.11	0.10	0.10	0.11	0.09

Podemos calcular las medidas descritas para el caso de una muestra para decidir si un gen está expresado o no lo está. Se discutirá más adelante como precisar esto pero de momento nos quedaremos con la idea de que si la medida escogida es (cercana a) cero el gen no está diferencialmente expresado y si es mayor o menor que cero si que lo está. Nos referimos a "cero" porque estamos hablando de logaritmos de razones: si la expresión es la misma en ambas condiciones el cociente es uno y su logaritmo es cero.

Vale la pena insistir en el concepto de expresión diferencial: no nos preocupa cual es la expresión del gen en una u otra muestra sinó si son distintas.

Gen	Promedio	Err. Std	T-test
A	2.617	0.397	14.735
B	0.003	0.032	0.233
C	5.083	7.335	1.550
D	0.133	0.273	1.091
E	0.102	0.008	30.200

La tabla anterior sugiere que podría considerarse el gen A está diferencialmente expresado (promedio y t–test altos) mientras que el gen B o el D no lo están (promedio y test-t próximos a cero). Los genes C y D pueden llevar a conclusiones contradictoria según nos basemos en el promedio o el test t.

Si se observan los valores del test t del gen C se concluye que el gen no aparenta estar diferencialmente expresado. Si en cambio se observa su promedio parece que si que lo esté.

En el gen E pasa exctamente lo opuesto. La explicación de estas aparentes contradicciones se halla en el error estándar. En el gen C es muy elevado, debido a que el valor (20) es probablemente un "outlier". En el gen D el error estándar es muy bajo por lo que, al encontrarse en el denominador del t–test aumenta artificialmente su valor.

1.1.2 Selección de genes diferencialmente expresados

Vamos a plantear la forma como se aborda este problema en el estudio de datos de microarrays. Dadas las características propias de este tipo de datos se consideran otras formas de estimar el error estándar que no sean tan sensibles a valores extremos o muy bajos.

Consideremos el gen g. Si llamamos:

- R_g log-ratio medio observado.
- SE_g error estándar de R_g estimado a partir de los datos en el gen g .
- SE error estándar de R_g estimado a partir de los datos con la información de todos los genes.

Podemos considerar dos variantes para el test–t:

- *Test-t global*: se calcula en base a un único estimador de SE para todos los genes:

$$t = R_g/SE,$$

- *Test-t específico*: Utiliza un estimador distinto del error estandar para cada gen:

$$t = R_g/SE_g.$$

Cada aproximación tiene sus pros y sus contras como muestra la tabla siguiente:

Test	**Pros**	**Contras**
Test-t Global	Estimador estable de σ	Asume homocedasticidad
Test-t específico	Robusto a heterocedasticidad	Estimador de σ no estable

En la práctica muchos métodos de selección de genes diferencialmente expresados han acabado buscando un compromiso entre ambas aproximaciones para lo que proponen o derivan fórmulas que de alguna forma ponderan o combinan dos estimaciones del error estándar: una basada en todos los genes y otra específica de cada gen. La tabla siguiente ilustra como algunos de los métodos más utilizados en la bibliografía incorporan esta idea.

Método (Referencia)	Fórmula
SAM (Tibshirani et al 2001)	$S = \frac{R_g}{c+SE_g}$
Cyber-T (Baldi et al, 2001)	$t = \frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1)SE_g^2}{v_0+n-2}}}$
T-moderado (Smyth, 2003)	$t = \frac{R_g}{\sqrt{\frac{d_0 \cdot SE_0^2 + d \cdot SE_g^2}{d_0+d}}}$

Al hecho de incluir un coeficiente que tenga en cuenta la variabilidad de todos los genes en el array para estimar el error estándar de cada gen se le denomina moderación de la varianza ("variance shrinkage") y es una de las aproximaciones en que existe cierto consenso ([1]) acerca de que sirven para mejorar la selección de genes diferencialmente expresados.

Ejemplo de utilización del test–t

Como ejemplo utilizaremos el conjunto de datos `celltypes` y supondremos que disponemos ya de los datos normalizados y filtrados almacenados en un objetoo `expressionSet`.

Si consideramos que el campo *treat* del objeto contiene la información de los grupos a comparar (ratones estimulados con LPS frente NA `rowttests` del paquete `genefilter` que realiza un test *t* sobre cada una de las filas (genes) de una matriz de expresión.

```
> stopifnot(require(Biobase))
> load ("celltypes-normalized.rma.Rda")
> my.eset <- eset_rma_filtered
> grupo_1 <- as.factor(pData(my.eset)$treat)
> stopifnot(require(genefilter))
> teststat <-rowttests(my.eset, "treat")
> print(teststat[1:5,])
```

```
             statistic         dm      p.value
1415694_at    3.503426  0.9813765 5.693821e-03
1415698_at   -4.253446 -0.9200460 1.680334e-03
1415743_at  -15.642231 -1.0156246 2.335398e-08
1415760_s_at -5.686252 -0.8663532 2.020937e-04
1415772_at   11.394671  1.1306309 4.745989e-07
```

Cuanto mayor sea el valor absoluto del estadístico *t* mayor es la probabilidad de que el gen esté diferencialmente expresado.

Genes diferencialmente expresados estadísticamente significativos

Como hemos visto en la sección anterior dos medidas naturales para la selección de genes son el promedio de "log-ratios" –o la diferencia de promedios en el caso de muestras independientes– o el valor del estadístico de test (t-test) de una o dos muestras según si se trata de muestras apareadas o independientes respectivamente. Los primeros estudios de microarrays eran muy costosos y se hacían con pocas o incluso ninguna réplica por condición experimental. En estas situaciones la única forma fiable de detectar una diferencia de expresión era a traves del "log-ratio" o su diferencias.

Rápidamente se puso en evidencia que para poder obtener los genes que estaban realmente diferencialmente expresados era preciso disponer de un soporte estadístico que permitiera tener en cuenta la variación aleatoria existente entre muestras.

En la práctica esto se reduce a afirmar que si, además de la diferencia de expresión entre las condiciones experimentales, se lleva a cabo un test estadístico dispondremos de una medida objetiva, el p–valor que nos servirá para decidir qué genes se declaran diferencialmente expresados, a saber, aquellos en los que el p–valor del test sea inferior a un cierto umbral como 0.05 o 0.01.

Tal como se ha indicado en el capítulo **??**, un test estadístico procede decidiendo rechazar la hipótesis nula si el p–valor es más pequeño que el nivel de significación del test.

Siguiendo con el ejemplo anterior podemos ordenar los resultados de los tests en base a los p–valores:

```
> ranked <-teststat[order(teststat$p.value),]
> print(ranked[1:5,])
```

```
             statistic        dm      p.value
1449383_at   -48.61014 -2.044928 3.276616e-13
1430127_a_at  46.90904  1.508843 4.672074e-13
1451421_a_at -40.72173 -1.445182 1.909283e-12
1450826_a_at  37.62239  5.147992 4.193941e-12
1416122_at    34.66314  1.482676 9.460684e-12
```

Ahora podríamos seleccionar, por ejemplo los genes cuyo p–valor fuera inferior a 0.01

```
> selectedTeststat <- ranked[ranked$p.value < 0.01,]
```

Esto deja un total de 1594 con un p–valor inferior a 0.01

"Volcano plots"

Si se opta por computar los valores de significación (p–valores) de los genes, resulta interesante comparar el tamaño del cambio del nivel de significación estadístico. El "volcano plot" es una representación gráfica que permite ordenar los genes a lo largo de dos dimensiones, la biológica, representada por el "fold change" y la estadística representada por el logaritmo negativo del p–valor.

En la escala horizontal se representa el cambio entre los dos grupos (en escala logarítmica, de manera que la regulación positiva o negativa se representa de forma simétrica). En la escala vertical se representa el p–valor del test en una escala logarítmica negativa, de forma que los p–valores más pequeños aparecen mayores.

Así pues puede considerarse que el primer eje indica impacto biológico del cambio (a más efecto biológico mayor "fold-change") y el segundo muestra la evidencia estadística, o la fiabilidad del cambio.

La figura 1 muestra un "volcano–plot" para ejemplo de los "celltypes".

Figura 1. Volcano plot

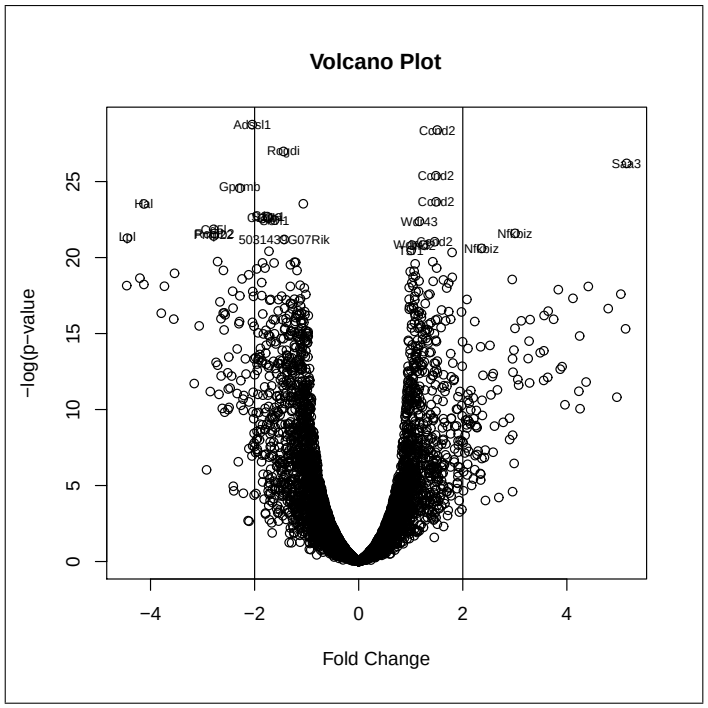


Figura 1

Ejemplo de Volcano plot que muestra los genes candidatos a considerarse como diferencialmente expresados en la comparación "LPS" frente a "Medium"

1.1.3 Potencia y tamaño muestral

Tal como se ha indicado más arriba, para realizar un *buen test* suele controlarse la probabilidad de error de tipo I (de falsos positivos) y buscar, de entre los tests candidatos, aquellos que tengan una menor probabilidad de error de tipo II, o equivalentemente una mayor potencia. A partir de este planteamiento existe, en el contexto estadístico estándar, multitud de formas de determinar cual debe ser el tamaño muestral necesario para obtener una potencia dada fijados el tamño de efecto ("fold-change") y la probabilidad de error de tipo I.

En el caso de los microarrays se han desarrollado diversas fórmulas para realizar cálculos de este tipo, pero la compleja estructura de los datos microarrays hace que sean relativamente *discutibles*.

Simon ([6]) sugiere la fórmula siguiente que es una generalización de las fórmulas clásicas para problemas de dos muestras:

El tamaño total requerido para detectar genes diferencialmente expresados en al menos una diferencia δ con una probabilidad de error de tipo I (FP), α y una probabilidad de error de tipo II (FN) $1-\beta$ se calcula:

$$n = \frac{4(z_{\alpha/2} + z_{\beta})^2}{(\delta/\sigma)^2},$$

donde z_α y z_β son los percentiles $100\alpha/2$ y 100β de la distribución Normal $N(0,1)$ y σ es la desviación estandar de un gen dentro de una clase (de un grupo). Obviamente σ *es siempre desconocida* por lo que, sin una muestra piloto con que estimarla el calculo e más imaginativo que realista.

Además de esto, el número de arrays usualmente recomendado queda lejos de la cantidad asequible para la mayor parte de los experimentos ([5, 9]). Lo que muchos usuarios hacen a la práctica, es buscar un equilibrio entre los costes y la reproducibilidad y, de hecho, tienden a usar una cantidad fija de arrays tal como 3 o 5 sin consideraciones adicionales.

Por ejemplo si ponemos $\alpha = 0.001$, $1-\beta = 0.95$, $\delta = 1$ y estimamos σ entre todas las muestras el número de réplicas biológicas que necesitaremos será de 35.8.

1.1.4 El problema de la múltiplicidad de tests ("multiple testing"

El análisis de microarrays se realiza en base gen a gen e involucra múltiples tests, miles probablemente. Esto significa que, a medida que crece el número de genes, la probabilidad de declarar erroneamente al menos un gen diferencialmente expresado va en aumento, y si no se realiza algún tipo de ajuste el número de falsos positivos será tanto más alto cuantos más genes estemos analizando.

Hay muchas formas de intentar controlar estas probabilidades de error y puede verse un excelente revisión en Dudoit [3]).

De forma simplificada consideramos las dos aproximaciones más utilizadas.

Una posibilidad es mirar de controlar la probabilidad de obtener *al menos un falso positivo* o "Family-wise-error-rate (FWER)". El más popular de estos métodos de control es la corrección de Bonferroni, consistente en multiplicar el p–valor por el número de tests realizados. La misma Dudoit y muchos otros autores han desarrollado variantes de los métodos clásicos de ajuste FWER usando por ejemplo tests de permutaciones.

El criterio FWER es quizás demasiado restrictivo dado que el control de los falsos positivos implica un considerable incremento de falsos negativos. En la práctica, sin embargo, muchos biólogos parecen estar dispuestos a aceptar que se produzcan algunos errores, siempre y cuando esto permita realizar descubrimientos. Por ejemplo un investigador debe considerar aceptable cierta pequeña proporción de errores (digamos del 10 al 20%) entre sus descubrimientos. En este caso, el investigador está expresando interés en controlar el porcentaje de falsos descubrimientos (FDR), es decir lo que es la proporción de falsos positivos sobre el total de genes inicialmente identificados como expresados diferencialmente. A diferencia del nivel de significación que queda determinado antes de examinar los datos, la FDR es una medida de confianza a posteriori ya que emplea información disponible en los datos para estimar las proporciones de falsos positivos que han tenido lugar. Si se obtiene una lista de los genes expresados diferencialmente en los que el FDR se controla hasta, digamos, el 20%, cabe esperar que el 20% de estos genes representen falsos positivos. Lo cual supone un enfoque menos restrictivo que controlar el FWER.

La decisión de controlar el FDR o el FWER depende de los objetivos del experimento. Si, por ejemplo, el objetivo es la *captura de genes* es razonable permitir cierta cantidad de falsos positivos y es preferible seleccionar FDR. Si por el contrario se trabaja con una lista de un tamaño menor al deseado para verificar la expresión de ciertos genes específicos, entonces el FWER es el criterio apropiado.

El ejemplo siguiente muestra como se realiza el ajuste de p–valores usando el paquete `multtest` de Bioconductor para ajustar por los métodos de Bonferroni (FWER), Benjamini & Hochberg (FDR) o Benjamini & Yekutieli (BY, FDR).

```
> stopifnot(require(multtest))
> procs <- c("Bonferroni","BH", "BY")
> adjPvalues <- mt.rawp2adjp(teststat$p.value, procs)
> names(adjPvalues)
```

```
[1] "adjp"    "index"   "h0.ABH"  "h0.TSBH"
```

```
> ranked.adjusted<-cbind(ranked, adjPvalues$adjp)
> head(ranked.adjusted)
```

```
             statistic        dm      p.value         rawp   Bonferroni
1449383_at   -48.61014 -2.044928 3.276616e-13 3.276616e-13 1.478082e-09
1430127_a_at  46.90904  1.508843 4.672074e-13 4.672074e-13 2.107573e-09
1451421_a_at -40.72173 -1.445182 1.909283e-12 1.909283e-12 8.612774e-09
1450826_a_at  37.62239  5.147992 4.193941e-12 4.193941e-12 1.891887e-08
1416122_at    34.66314  1.482676 9.460684e-12 9.460684e-12 4.267714e-08
1448303_at   -31.92365 -2.283765 2.140612e-11 2.140612e-11 9.656299e-08
                       BH           BY
1449383_at   1.053786e-09 9.475226e-09
1430127_a_at 1.053786e-09 9.475226e-09
1451421_a_at 2.870925e-09 2.581421e-08
1450826_a_at 4.729717e-09 4.252773e-08
1416122_at   8.535429e-09 7.674717e-08
1448303_at   1.609383e-08 1.447093e-07
```

Si seleccionamos los genes en base a su p–valor ajustado por ejemplo por le método de Banjamini y Yekutieli se obtienen los siguientes genes

```
> selectedAdjusted<-ranked.adjusted[ranked.adjusted$BY<0.001,]
> nrow(selectedAdjusted)
```

```
[1] 449
```

1.2 Modelos lineales para la selección de genes: `limma`

En la sección anterior se ha discutido el uso del test t y sus extensiones para la selección de genes diferencialmente expresados en situaciones relativamente sencillas, es decir cuando sólo hay dos grupos.

En muchos estudios el número de grupos a considerar es más de dos y las fuentes de variabilidad pueden ser más de una, por ejemplo una puede ser el tratamiento pero otra puede ser la edad de los individuos o cualquier otra condición fijada por el investigador o derivadad de la heterogeneidad de las muestras.

En estas situaciones una aproximación razonable en problemas con *una variable respuesta* es el análisis de la varianza, discutido en el capítulo **??**. En esta sección se presenta una aproximación equivalente que de forma general se denomina *el modelo lineal*. Este método -que engloba el análisis d la varianza y la regresión- es uno de los más usados en estadística y ha sido popularizado en el campo de microarrays gracias a los trabajos de Gordon Smyth quien ha creado el paquete

`limma` que se ha convertido en la herramiente más utilizada para el análisis de microarrays.

1.2.1 El modelo lineal general

El modelo lineal (ver por ejemplo Faraway, [4]) es un marco general para la modelización y el análisis de datos estadística.

EL método consiste en asumir una relación lineal entre los valores observados de una variable *respuesta* y las condiciones experimentales. A partir de aquí se obtienen estimadores para los parámetros del modelo y de sus errores estándar, y (con algunas condiciones extra) es posible hacer inferencia acerca del experimento.

La aplicación de modelos lineales puede ser visto como un proceso secuencial, con los siguientes pasos:

1) Especificar el diseño del experimento: qué muestras se asignan a qué condiciones.

2) (Re-)Escribir un modelo lineal para este diseño en forma de $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, donde $\mathbf{X}$ se denomina *la matriz de diseño.*

3) Una vez que el modelo se ha especificado aplicar la teoría general de estimar los parámetros y los contrastes (comparaciones entre los valores de los parámetros).

4) Si se cumplen ciertas condiciones de validez para el modelo es posible realizar inferencia sobre los parámetros del modelo, es decir se pueden contrastar hipótesis sobre dichos parámetros.

El esquema anterior se puede aplicar a casi cualquier tipo de situación experimental. En la sección siguiente se presentan un par de ellas.

1.2.2 Ejemplos de situaciones *modelizables* linealmente

Ejemplo 1: Experimento “Swirl–Zebrafish“

Swirl es una mutación puntual que provoca defectos en la organización del embrión en desarrollo a lo largo de su eje dorsal-ventral. Como resultado, algunos tipos celulares se reducen y otros se expanden. Un objetivo de este trabajo fue identificar los genes con expresión alterada en el mutante Swirl en comparación con ”wild zebrafish“.

1) El diseño experimental para este estudio fue el siguiente:

Slide	Cy3	Cy5
1	W	M
2	M	W
3	W	M
4	M	W

- Cada microarray contenía 8848 sondas de cDNA (genes o sequencias EST).
- Cuatro réplicas por array (slide): 2 juegos de pares de intercambio de color
- El cDNA del mutante swirl (S) se marca con Cy5 o Cy3 y el cDNA del "wild type" se marca con el otro

2) El modelo lineal derivado del diseño anterior fue:

- El parámetro de interés es: $\alpha = \mathbf{E}\left(log\frac{S}{W}\right)$.
- Las muestras 1 y 3 estan marcadas con : S (Verde="Green") y W (Rojo="Red"), y las muestras 2 y 4 son "dye-swapped".
- El modelo, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, es:

$$\begin{array}{rcrcl} y_1 &=& \alpha &+& \varepsilon_1 \\ y_2 &=& -\alpha &+& \varepsilon_2 \\ y_3 &=& \alpha &+& \varepsilon_3 \\ y_4 &=& -\alpha &+& \varepsilon_4 \end{array} \Longrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}}_{\text{Matriz de Diseño,}\mathbf{X}} \alpha + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix} \quad (1)$$

1.2.3 Ejemplo 2: Comparación de tres grupos

Los plásmidos IncHI codifican genes de resistencia múltiple a los antibióticos en *S. enterica*.

El plásmido R27 de la cepa salvaje es termosensible al transferirse.

Algunos fenotipos mutantes relacionados con *hha* y *hns* cromosómicos participan en diferentes procesos metabólicos de interés en la conjugación termoregulada.

El objetivo del experimento es encontrar qué genes se expresean diferencialmente en tres tipos de mutantes diferentes, M_1, M_2 y M_3.

Posibles estrategias de diseño Este experimento debe ser planteado de forma diferente según el tipo de arrays (uno o dos colores) y qué comparaciones son las de mayor interés.

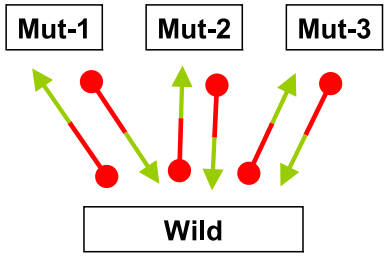
- Array de dos colores

– A *diseño de referencia*: Hibridar cada Mutante (M_i) vs. Salvaje ("Wild type") (W).

– A *diseño loop*: Hibridar cada mutante el uno al otro en un doble "loop" que incluye ("dye-swapping").

- Array de un color: hibridar mutantes y "wild types" separadamente.

Representación del diseño de referencia

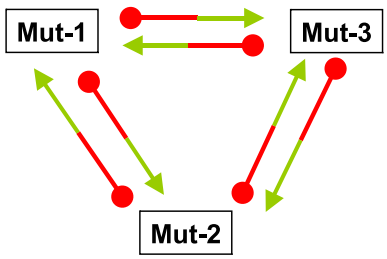


- Permite la comparación directa de Mutantes vs "Wild".

- Número de parámetros a estimar es 3, relación intuitiva entre el número de parámetros y mutantes.

- Las comparaciones de Mutante vs Mutante son menos eficientes.

Representación del diseño de "loop" (bucle)

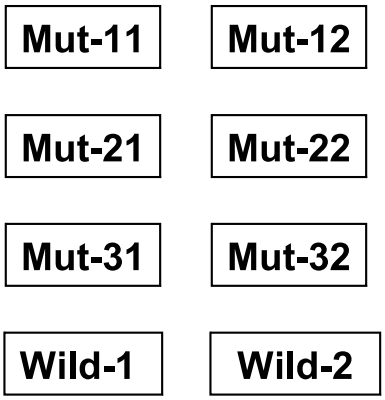


- Permite la comparación directa de Mutantes vs Mutantes.

- El número de parámetros a estimar es 2. Menos intuitivo.

- Las comparaciones Mutante vs Mutante son más eficientes.

Representación del diseño del array de un color



- Permite la comparación directa de

– Mutante vs "Wild"

– Mutante vs Mutante

- El número de parámetros a estimar es 4.

- Todas las comparaciones se pueden hacer de forma eficiente.

Modelo lineal para el diseño de referencia Modelo, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, y contrastes $\mathbf{C}'\beta$

- Parámetros del modelo:

$$\alpha_1 = \mathbf{E}\left(log\frac{M_1}{W}\right),\ \alpha_2 = \mathbf{E}\left(log\frac{M_2}{W}\right),\ \alpha_3 = \mathbf{E}\left(log\frac{M_3}{W}\right).$$

- *Contrastes*: Comparaciones interesantes.

$$\begin{aligned}\beta_1 &= \alpha_1 - \alpha_2,\\ \beta_2 &= \alpha_1 - \alpha_3,\\ \beta_3 &= \alpha_2 - \alpha_3.\end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}}_{\text{Matriz de Diseño,}\mathbf{X}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \tag{2}$$

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}}_{\text{Matriz de Contraste,}\mathbf{C}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}. \tag{3}$$

Modelo lineal para el diseño de "loop" Modelo, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, y contrastes $\mathbf{C}'\beta$

- Parámetros del modelo:

$$\alpha_1 = \mathbf{E}\left(log\frac{M_1}{M_2}\right),\ \alpha_2 = \mathbf{E}\left(log\frac{M_2}{M_3}\right).$$

  α_3 no es necesaria: $\log\left(\frac{M_1}{M_3}\right) = \log\left(\frac{M_1}{M_2}\right) - \log\left(\frac{M_2}{M_3}\right)$.

- *Contrastes*: Algunas de las comparaciones deseadas son precisamente los parámetros.

$$\begin{aligned} \beta_1 &= \alpha_1, \\ \beta_2 &= \alpha_2, \\ \beta_3 &= \alpha_1 + \alpha_2. \end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \\ -1 & 0 \\ 0 & -1 \\ -1 & 1 \end{pmatrix}}_{\text{Matriz de Diseño,}\mathbf{X}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \tag{4}$$

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & +1 \end{pmatrix}}_{\text{Matriz de Contraste,}\mathbf{C}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}. \tag{5}$$

Modelo lineal para el diseño del array de un color Modelo, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, y contrastes $\mathbf{C^{1'}}\beta$, $\mathbf{C^{2'}}\beta$

- Parámetros del modelo:

$$\alpha_1 = \mathbf{E}(logM_1),\ \alpha_2 = \mathbf{E}(logM_2),\ \alpha_3 = \mathbf{E}(logM_3), \alpha_4 = \mathbf{E}(logW).$$

- *Contrastes*: Dos posibles conjuntos de comparaciones interesantes.

  1) Comparación entre tipo de mutantes ($\mathbf{C^{1'}}\beta$)

$$\begin{aligned} \beta_1^1 &= \alpha_1 - \alpha_2, \\ \beta_2^1 &= \alpha_3 - \alpha_2, \\ \beta_3^1 &= \alpha_2 - \alpha_3. \end{aligned}$$

2) Comparación entre cada mutantes y el wild type ($\mathbf{C^{2'}}\beta$)

$$\begin{aligned}\beta_1^2 &= \alpha_4 - \alpha_1,\\ \beta_2^2 &= \alpha_3 - \alpha_1,\\ \beta_3^2 &= \alpha_2 - \alpha_1.\end{aligned}$$

$$\begin{pmatrix} y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\text{Matriz de Diseño,}\mathbf{X}} \begin{pmatrix} \alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\\ \varepsilon_7\\ \varepsilon_8 \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} \beta_1^1\\ \beta_2^1\\ \beta_3^1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0\\ 1 & 0 & -1 & 0\\ 0 & 1 & -1 & 0 \end{pmatrix}}_{\text{Matriz de Contraste,}\mathbf{C^1}} \begin{pmatrix} \alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4 \end{pmatrix}. \tag{7}$$

$$\begin{pmatrix} \beta_1^2\\ \beta_2^2\\ \beta_3^2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & -1\\ 0 & 1 & 0 & -1\\ 0 & 0 & 1 & -1 \end{pmatrix}}_{\text{Matriz de Contraste,}\mathbf{C^2}} \begin{pmatrix} \alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4 \end{pmatrix}. \tag{8}$$

Ejemplo 3: Estudio de la influencia de las citoquinas en ratones viejos

El objetivo de este experimento es estudiar el efecto de las citoquinas sobre la estimulación de una sustancia (LPS) y ver como esta relación se ve afectada por

la edad.

Se trata de un modelo de un un factor (tratamiento, asignable por el individuo) y un bloque (edad, no asignable, prefijada en cada ratón). En la práctica el modelo equivale al del análisis de la varianza de dos factores.

1) El diseño experimental para este estudio fue el siguiente:

Array	Tratamiento	Edad
1	LPS	Viejo
2	LPS	Joven
3	Medio	Viejo
4	Medio	Joven
5	LPS	Viejo
6	LPS	Joven
7	Medio	Viejo
8	Medio	Joven
9	LPS	Viejo
10	LPS	Joven
11	Medio	Viejo
12	Medio	Joven

- Se utilizáron microarrays de un color (Affymetrix).
- Cada condición se replicó tres veces.
- Las preguntas específicas a responder:
  - ¿Cual es el efecto del tratamiento en ratones viejos?
  - ¿Cual es el efecto del tratamiento en ratones jovenes?
  - ¿En que genes el efecto es diferente?.

2) En este caso se pueden considerar distintos modelo lineales derivados del hecho de que este experimento admite diferente parametrizaciones:

- Factores separados con 2 niveles cada uno para tratamiento (LPS/Med), Edad (Joven/Viejo) y su interacción:

$$Y_{ijk} = \underbrace{\alpha_i}_{\text{Trat}} + \underbrace{\beta_j}_{\text{Edad}} + \underbrace{\gamma_{ij}}_{\text{Interacción}} + \varepsilon_{ijk},\ i = 1, 2,\ j = 1, 2,\ k = 1, 2, 3$$

  Esta primera parametrización parece natural pero es más complicado confiar en ella para responder las preguntas propuestas.

- Un factor combinado con 4 niveles
  (*LPS.Aged, Med.Aged, LPS.Young, Med.Young*)

$$Y_{ij} = \alpha i + \varepsilon_{ij}, \quad i = 1, ..., 4, \; j = 1, 2, 3.$$

  Esta parametrización parece más rígida pero se adapta mejor para responder a las preguntas planteadas.

Aquí se adopta la segunda parametrización.

- Parámetros del modelo:

$$\alpha_1 = \mathbf{E}(logLPS.Aged), \; \alpha_2 = \mathbf{E}(logMed.Aged),$$

$$\alpha_3 = \mathbf{E}(logLPS.Young), \; \alpha_4 = \mathbf{E}(logMed.Young).$$

- *Contrastes*: Preguntas que interesa responder:

$$\beta_1^1 = \alpha_3 - \alpha_1, \quad \text{Efecto del tratamiento en ratones viejos}$$

$$\beta_2^1 = \alpha_4 - \alpha_2, \quad \text{Efecto del tratamiento en ratones jóvenes}$$

$$\beta_3^1 = (\alpha_3 - \alpha_1) - (\alpha_2 - \alpha_4), \quad \text{Interacción: diferencia entre efectos}$$

- Modelo lineal:

  Modelo, $\mathbf{y} = \mathbf{X}\alpha + \varepsilon$, y contrastes $\mathbf{C}'\beta$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\text{Matriz de Diseño,}\mathbf{X}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 1 & 1 & -1 \end{pmatrix}}_{\text{Matriz de Contraste,}\mathbf{C}} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} \tag{10}$$

1.2.4 Estimación e inferencia con el modelo lineal

Una vez se ha expresado el experimento como un modelo lineal:

$$\mathbf{E}(\mathbf{y_g}) = \mathbf{X}\alpha_\mathbf{g}, \quad \text{var}(y_g) = W_g\sigma_g, \tag{11}$$

es posible usar *la teoría estándar del modelo lineal* (ver [4]) para obtener:

- Estimación de los parámetros: $\hat{\alpha}_g (\approx \alpha)$.
- Desviación estándar de las estimaciones: $\hat{\sigma}_g = s_g (\approx \sigma)$.
- Error estándar de las estimaciones:$\widehat{\text{var}\hat{\alpha}_g} = V_g\, s_g^2$.

Estas estimaciones son la base para realizar inferencia sobre α i.e. test $H_0 : \alpha = 0$?, basado en el hecho que:

$$t_{gj} = \frac{\alpha_{gj}}{s_g\sqrt{v_{gj}}} \sim \text{Distribución de Student.} \quad (12)$$

De forma análoga pueden derivarse fórmulas para $\alpha_1 - \alpha_2$, es decir para decidir acerca de las comparaciones.

Los procedimientos de estimación y de inferencia no dependen de qué parametrización se ha adoptado, a pesar de que distintas parametrizaciones piedan dar lugar a distintos valores numéricos..

Fortaleza y debilidades del modelo lineal

El enfoque del modelo lineal es flexible y potente:

- Se puede adaptar a situaciones diferentes y complejas.
- Siempre produce buenas estimaciones ("BLUE").
- Si las suposiciones son ciertas proporciona una base para la inferencia.

Por otro lado hay que tener en cuenta que, si las suposiciones no se cumplen, entonces las conclusiones deben tomarse con precaución.

Y lo que es peor, aún siendo válidas las suposiciones del modelo los resultados pueden verse afectados por el tamaño de la muestra de forma que, en muestras pequeñas donde pueden haber variaciones más grandes es fácil que se obtengan valores t no significativos o, por el contrario, excesivamente significativos (si la variación es muy reducida).

La metodología desarrollada por Smyth ([7]) basada en los resultados de Lönsted & Speed ([8]) está dirigida a abordar cómo hacer frente a estas debilidades.

1.2.5 Modelos lineales para Microarrays

Smyth ([7]) considera el problema de la identificación de genes que se expresan diferencialmente en las condiciones especificadas en el diseño de experimentos de microarrays. Como hemos dicho repetídamente la variabilidad de los valores de expresión difiere entre genes, pero la naturaleza paralela de la inferencia en microarrays sugiere la posibilidad de usar la información *de todos los genes a la vez* para mejorar la estimación de los parámetros, lo que puede llevar a una inferencia más fiable.

Básicamente lo que propone Smyth ([7]) es una solución en tres pasos:

- Se plantea el problema como un model lineal con una componente bayesiana ya que se supone que los mismos parámetros a estimar son variables (no constantes) con distribuciones *prior* que se estimaran a partir de la información de todos los genes.

- A continuación se obtienen las estimaciones de los parámetros del modelo. La aproximación utilizada garantiza que estos estimadores tienen un comportamiento robusto incluso para pequeño número de arrays.

- Finalmente se calcula un "odd-ratio" que viene a ser la probabilidad de que un gen esté diferencialmente expresado frente a la de que no lo esté y se asocia este valor denominado estadístico B con un estadístico t moderado y su p–valor.

$$B = \log \frac{P[\text{Afectado}|M_{ij}]}{P[\text{No Afectado}|M_{ij}]},$$

  gen=i ($i = 1...N$), réplica=j ($j = 1, ..., n$).

  El hecho de trabajar con logaritmos permite poner el punto de corte en el cero: A mayor valor positivo más probable es que el gen esté diferencialmente expresado. A mayor valor negativo, más probable es que no lo esté

1.2.6 Implementación y ejemplos

El paquete de *Bioconductor* `limma` al que hemos hecho referencia en los párrafos anteriores implementa el método de Smyth para la regularización de la varianza lo que lo ha convertido en muy popular entre los usuarios de microarrays

El código siguiente muestra como se crea la matriz del diseño y los contrastes para realizar el análisis mediante modelos lineales;

```
                    Aged.LPS Aged.MED Young.LPS Young.MED
Aged_LPS_80L.CEL           1        0         0         0
Aged_LPS_86L.CEL           1        0         0         0
Aged_LPS_88L.CEL           1        0         0         0
Aged_Medium_81m.CEL        0        1         0         0
Aged_Medium_82m.CEL        0        1         0         0
Aged_Medium_84m.CEL        0        1         0         0
Young_LPS_75L.CEL          0        0         1         0
Young_LPS_76L.CEL          0        0         1         0
Young_LPS_77L.CEL          0        0         1         0
Young_Medium_71m.CEL       0        0         0         1
```

```
Young_Medium_72m.CEL        0        0         0         1
Young_Medium_73m.CEL        0        0         0         1
attr(,"assign")
[1] 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$lev
[1] "contr.treatment"
```

```
> require(limma)
> cont.matrix <- makeContrasts (
+       LPS.in.AGED=(Aged.LPS-Aged.MED),
+       LPS.in.YOUNG=(Young.LPS-Young.MED),
+       AGE=(Aged.MED-Young.MED),
+       levels=design)
> cont.matrix
```

```
           Contrasts
Levels      LPS.in.AGED LPS.in.YOUNG AGE
  Aged.LPS            1            0   0
  Aged.MED           -1            0   1
  Young.LPS           0            1   0
  Young.MED           0           -1  -1
```

NA estimar el modelo, estimar los contrastes y realizar las pruebas de NA puede considerarse diferencialmente expresado.

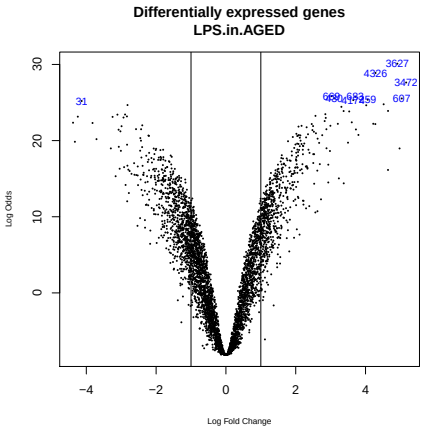
La penúltima instrucción ejecuta el proceso de regularización de NA NA obtener estimaciones de error mejoradas.

NA `Fold-change` t-moderados o p-valores ajustados que se utilizan para ordenar los genes de mas a menos diferencialmente expresados.

A fin de controlar el porcentaje de falsos positivos que puedan resultar del alto numero de contrastes realizados simultaneamente los p–valores se ajustan de forma que tengamos control sobre la tasa de falsos positivos utilizando el metodo de Benjamini y Hochberg ([2]).

La funcion `topTable` genera para cada contraste una lista de genes ordenados de mas a menos diferencialmente expresados.

```
> topTab_LPS.in.AGED <- topTable (fit.main, number=nrow(fit.main), coef="LPS.in.AGED", adjust="fdr")
> topTab_LPS.in.YOUNG <- topTable (fit.main, number=nrow(fit.main), coef="LPS.in.YOUNG", adjust="fdr")
> topTab_AGE  <- topTable (fit.main, number=nrow(fit.main) , coef="AGE", adjust="fdr")
```


Differentially expressed genes
LPS.in.AGED
Log Odds
Log Fold Change
3627
4326
3472
687
31
30
20
10
0
−4
−2
0
2
4

NA NA

Resumen

El análisis de datos de microarrays és una disciplina que combina la bioinformática la estadística y la biología para esclarecer problemas que aparecen en el estudio de la expresión génica con microarrays, que son herramientas que permiten el estudio de la expresión de manera simultánea en todos los genes de un organismo. Con los microarrays se pueden tratar multitud de problemas entre los que podemos destacar la *comparación de clases*, el *descubrimiento de nuevos grupos* o la construcción de predictores.

%beginpreguntes

Bibliography

[1]David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, January 2006.

[2]Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[3]S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

[4]Julian J. Faraway. *Linear Models with R.* Chapman and Hall/CRC, 1 edition, July 2004.

[5]M.L.T. Lee and GA Whitmore. Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21(23):3543–3570, 2002.

[6]Richard M. Simon, Edward L. Korn, Lisa M. McShane, Michael D. Radmacher, George W. Wright, and Yingdong Zhao. *Design and Analysis of DNA Microarray Investigations.* Springer-Verlag, 2003.

[7]Gordon K Smyth, Joëlle Michaud, and Hamish S Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–75, May 2005.

[8]T. Speed. *Statistical Analysis of Gene Expression Data.* Boca Raton, Fla.: Chapman & Hall/CRC, 2003.

[9]R. Tibshirani. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7(1):106, 2006.