10. Selección de genes diferencialmente expresados

Uno de los principales objetivos del análisis de datos de microarreglos consiste en identificar los genes que muestran buena evidencia de estar diferencialmente expresados (DE). Este objetivo se divide en dos etapas:

- Elección del estadístico
- Determinación de un punto de corte.

La primera, que consiste en elegir un estadístico, permite ordenar la evidencia de expresión diferenciada, desde la mayor a la menor evidencia. La segunda, que es elegir un valor crítico para el ordenamiento anterior por encima del cual cualquier valor resulta significativo, define la cantidad de genes que se considerarán DE.

10.1 Fold change

En la literatura de microarreglos es habitual referirse a cambios en intensidades de fluorescencia en términos del "fold change".
¿Qué es el fold change? Un fold change correspondería a un cambio del 100% de la intensidad. ¿Que es un cambio porcentual? Es 100*(Valor final - Valor inicial) / Valor inicial. Esto corresponde a duplicar el valor inicial, ¿es 1- fold change o 2-fold change?

Recordemos que, para cada probe,

- $M = \log_2(\frac{R}{G})$ es el logaritmo en base 2 del cociente entre las intensidades del canal rojo y el canal verde,

mientras que

- $A = 0.5\log_2(R*G)$.

De acuerdo con Smyth et al (2003), es conveniente utilizar logaritmos en base 2 para M y A de manera que estos estén expresados en *unidades* de un aumento de *2-fold* en luminosidad. En esta escala,

- M=0 representa igualdad de expresión (R= G $\Leftrightarrow$ R/G = 1),
- M=1 representa un cambio en un 2-fold ($R/G = 2^1$),
- M =2 ($R/G = 4 = 2^2$) representa un 4-fold change.

Para Smyth et al. *un fold change es no cambio* y *2 fold change* corresponde a *duplicar* el valor inicial. Pero no hay una criterio establecido, en otras publicaciones aparece que "one fold change = 2 times". En este sentido la cantidad de fold-changes indica la cantidad de veces que el valor se duplica y coincide con el valor de M.

En cada situación es necesario identificar cuál de las dos interpretaciones de fold change se está utilizando.

¿Por qué logaritmo en base 2?

La cuantificación por fold-changes no es simétrica debido a que los cocientes no son simétricos respecto de 1:

2 fold change significa 2 / 1
-2 fold change, significa 1 / 2

Esto hace que sea problemático operar con cocientes.
Pero con logaritmos. en particular

$\log(2x) = \log(x) + \log(2)$, doble $\Leftrightarrow$ *sumar* log(2)
$\log(x/2) = \log(x) - \log(2)$, mitad $\Leftrightarrow$ *restar* log(2

Las diferencias en escala logarítmica (es decir las diferencias de logaritmos) pueden ser interpretadas como "fold change" en la escala original de los datos. Aumentos y reducciones correspondientes al mismo "fold change" tienen el mismo tratamiento en la escala log.

10.2 Selección de genes diferencialmente expresados: Elección del estadístico. Métodos basados en un único microarreglo de dos canales

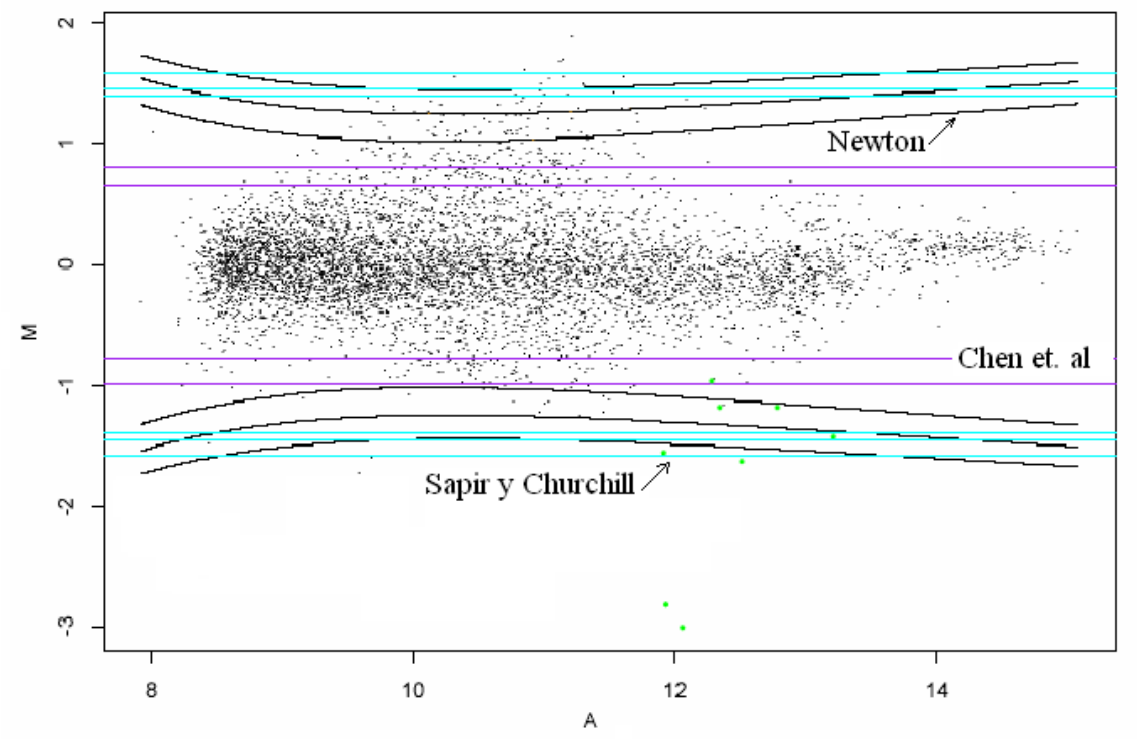
Los trabajos tempranos de datos de microarreglos (DeRisi et al., 1996; Schena et al., 1995, 1996) identificaban genes DE en base a un único microarreglo de 2 canales (o 2 de un canal. Utilizaban puntos de corte para el aumento o disminución de la intensidad (fold increase/decrease cutoffs) entre los canales rojo y verde para identificar genes DE. Por ejemplo Schena et al. 1995 en su estudio de niveles de expresión en la planta modelo Arabidopsis thaliana, utilizaron controles spike-in para normalizar las señales de dos tintes fluorescentes (fluoresceina y lisamina) y declararon que un gen estaba expresado diferencialmente si sus niveles de expresión diferían en más de un factor de 5 en las dos muestras de mRNA. DeRisi et al. 1996 identificaron genes diferencialmente expresados utilizando un punto de corte para los log-ratios de las intensidades de fluorescencia en ±3 desvíos estándar standard, con respecto a la media y desvío estándar de un panel de 90 genes "housekeeping" (i.e., genes que se supone que no están diferencialmente expresados entre los dos tipos de células de interés).

En general

- Se trata de encontrar qué genes se expresan en forma diferencial entre dos muestras
- Se usan reglas para decidir qué par (R,G) corresponde a un gen expresado diferencialmente
- Básicamente, estas reglas permiten trazar dos curvas en el plano (R,G) o el (M,A)) y decidir en base a los puntos que quedan fuera del área delimitada por las curvas que se trata de genes expresados diferencialmente

Existen diferentes propuestas, cada una de ellas basadas en diferentes supuestos.

Chen (1997), Newton (2001), Sapir & Churchill (2000)
Dependiendo de los métodos se obtienen diferentes curvas como ejemplifica la siguiente figura



10.3 Selección de genes diferencialmente expresados: Elección del estadístico. Dos grupos, comparación directa.

Para identificar los genes que muestran buena evidencia de estar diferencialmente expresados (DE) necesitamos elegir un estadístico, permita ordenar la evidencia de expresión diferenciada, desde la mayor a la menor evidencia.

Consideremos el experimento más simple, comparación de dos grupos (material A y material B). Suponemos que tenemos una serie de *n* microarreglos replicados en los cuales se han hibridado las muestras A y B. Se pueden realizar diferentes enfoques para el análisis.

Métodos Clásicos

Para cada réplica se calcula $M_i = \log2(R_i/G_i)$ y se calcula su media M y su varianza muestral s^2.

- Podría ser natural identificar a los genes DE tomando aquellos cuyos valores de $|M|$ excedan algún umbral k, determinado tal vez por la variabilidad observada en hibridaciones self-self en experimentos relacionados.

$$|M| > k$$

- Equivalentemente para los estadísticos calcular el estadístico $t = \sqrt{n}M / s$ y tomar la decisión en base a $|t|$:

$$|t| > k$$

En la primera opción se está asignando en forma implícita igualdad de varianzas de M_i sobre las replicaciones para cada gen. En la segunda se permite explícitamente que esas varianzas cambien entre genes.

Como la variabilidad de M_i sobre las replicaciones no es constante entre genes, los genes con mayor varianza tienen mayor chance de dar valores de M grandes incluso cuando no están DE.

Ninguna de las dos estrategias es completamente satisfactoria. Se pueden obtener valores grandes de *M* debido a la presencia de outliers, en especial debido a que el tamaño de muestra, *n*, es típicamente pequeño (de 2 a 8 Speed(2003)) y la tecnología es bastante ruidosa. Por otro lado dadas las decenas de miles de estadísticos $|t|$, siempre existe la posibilidad de que algunos sean grandes debido a que sus denominadores son muy pequeños tal vez cercanos a cero.

Soluciones al problema de varianzas pequeñas

Varias soluciones aproximadamente equivalentes están disponibles para el problema de las varianzas muy pequeñas son un compromiso entre utilizar únicamente *M* o únicamente *t*.

- Eliminar los genes cuyos errores estándar se encuentran dentro del 1% inferior de su error estándar.

Otros métodos más elaborados consisten en estandarizar *M* por algo intermedio entre una constante y el error estándar específico para cada gen.

- Efron et al. (2000),

$$t^* = \frac{\sqrt{n}M}{a+s}$$

  a es el percentil 90 de los desvíos estándar ó se elige de manera de minimizar el coeficiente de variación (Efron et al., 2000; Tusher et al., 2001).

- Lönnstedt and Speed 2001 adoptan un enfoque Bayesiano empírico parametrico, obtienen un *estadístico B* que cuando los supuestos paramétricos se cumplen los valores de B mayores que cero se corresponden con chances mayores que 50-50 de que el gen en cuestión esté DE. Con el propósito de ordenar los genes es equivalente a tomar el siguiente estadístico t-penalizado.

$$t^* = \frac{\sqrt{n}M}{\sqrt{a+s^2}}$$

Otras propuestas para elegir están dadas mediante enfoques de "Empirical Bayes": Efron et al. 2001, Long et al., 2001, Baldi and Long, 2001, Efron and Tibshirani, 2002.

10.4 Selección de genes diferencialmente expresados: Elección del estadístico. Dos grupos, comparación indirecta.

Un experimento de comparación indirecta entre un grupo A y un grupo B se realiza utilizando una muestra de referencia R. Tendremos n_A replicaciones en la que se han hibridado A y R simultáneamente dando M_A= media ($\log_2$ (A/R)) sobre las n_A y análogamente n_B muestras dando M_B = media($\log_2$ (B/R)).

Las propuestas para este problema del tipo estadísticos t tienen la siguiente forma general

$$t = \frac{M_A - M_B}{s(M_A - M_B)}$$

difieren en la forma en que calculan la variabilidad de la diferencia $s(M_A - M_B)$.

- Usando $s_p = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$ este cálculo es válido cuando puede suponerse que las varianzas de M_i son iguales en ambos grupos

$$s(M_A - M_B) = s_p \sqrt{1/n_A + 1/n_B}$$

- Welsh t-statistic

$$s(M_A - M_B) = \sqrt{s_A^2/n_A + s_B^2/n_B}$$

Las propuestas anteriores tienen los mismos problemas de desvíos casi nulos como en el caso de la comparación directa

- Estadístico d (Tusher et al.)

$$s(M_A - M_B) = s_p \sqrt{1/n_A + 1/n_B} + a$$

y hay muchas, muchas más Wilcoxon rank sum, Z-scores, likelihood ratio, etc.

10.5 Selección de genes expresados diferencialmente: determinación de un punto de corte.

Hemos visto diferentes propuestas para la elección del estadístico en base al cual se ordenarán los genes de acuerdo a la evidencia de expresión diferencial, desde la más débil a la más fuerte.

La importancia principal de este ordenamiento surge del hecho que solamente una cantidad limitada de genes puede seguirse en estudios posteriores. Además muchas veces se realizan experimentos biológicos típicos para confirmar los hallazgos resultantes de los experimentos de microarreglos.

En la mayoría de las veces será práctico continuar con una cantidad limitada de genes del orden de unos cientos. Por esta razón es importante identificar, por ejemplo, a los 100 candidatos más probables de estar diferencialmente expresados. La lista completa de genes que pueden considerarse DE estadísticamente significativos puede ser menos interesante si ésta es muy grande para su seguimiento (Smyth et al. 2003).

Una vez que se han ordenado los genes en base a un estadístico adecuado, el paso siguiente consiste en hallar un punto de corte por encima del cual los genes serán identificados como significativos.

La cuestión crucial en este punto es el control del nivel global (probabilidad de decidir que algún gen está DE cuando en realidad no lo está) del procedimiento resultante de realizar miles de tests, uno para cada gen. Desarrollaremos este tema en el capítulo 11.

10.5.1 Gráfico Volcán (Volcano plot)

Cualquiera sea el estadístico que se ha elegido para realizar los tests, es interesante evaluar simultáneamente la magnitud del "fold change" (M) con el nivel de significación estadístico (p-valor) de cada gen. Interesa detectar aquellos genes con un M grande y con diferencia estadísticamente significativa.

El 'volcano plot' es un diagrama de dispersión de - $\log_{10}(p)$ (en el eje vertical) en función de M (en el eje horizontal). Los p-valores más bajos, correspondientes a las diferencias estadísticamente más significativas, están en la parte superior del gráfico. Los genes que están regulados hacia arriba y hacia abajo (up and down regulated) aparecen en forma simétrica respecto de la recta vertical (M = 0). Por lo tanto los genes con diferencia estadísticamente significativa y "fold change" tenderán a encontrarse en los extremos derecho e izquierdo superiores del gráfico.

El eje horizontal muestra el impacto biológico del cambio y el vertical el estadístico permitiendo seleccionar visualmente los genes candidatos a estudios posteriores en base a ambos criterios.

En R se obtiene simplemente con

```
> plot(M, -log10(P), pch = '.')
```

Referencias

Smyth GK, Yang YH, Speed T. (2003). "Statistical issues in cDNA microarray data analysis". Methods Mol Biol;**224**:111-36.