11.5 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 experimentos biológicos típicos para su confirmación y estudios posteriores.

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 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 inherente a la necesidad de realizar un test para cada gen.

11.5.1 Tipos de errores en tests de hipótesis, para un único test.

Consideremos, para un gen en particular, las siguientes hipótesis:

H_0: el gen no está DE vs. **H_1:** el gen sí está DE

	Decisión	
Realidad	**H_0**	**H_1**
H_0	bien!	error de tipo II
H_1	error de tipo I	bien!

Cuando se testea una única hipótesis como **H_0** se pueden cometer dos tipos de errores:

Rechazar **H_0** cuando **H_0** es verdadera: error de tipo I ó falso positivo
No rechazar **H_0** cuando **H_0** es falsa: error de tipo II ó falso negativo

Habitualmente se controla la probabilidad de error de tipo I, es decir se fija *un nivel* α de manera que

$$P(\text{error de tipo I}) \leq \alpha$$

y dentro de una familia de posibles tests con nivel α se elige el que tiene menor probabilidad de error de tipo II (mayor potencia). Fijado un test y el nivel es necesario aumentar el tamaño de la muestra para controlar la probabilidad de cometer errores de tipo II (falsos negativos), de manera de asegurarse suficiente potencia para detectar verdaderos DE.

Estadístico del test: es el estadístico en base al cual se toma la decisión, lo llamamos T.

Región de rechazo a nivel α, son los valores del estadístico que resultan en rechazo:

$$|T| \geq c(\alpha),$$

con $P(|T| \geq c(\alpha)$ / cuando $\mathbf{H_0}$ es verdadera) = α.

> Este tipo de regiones de rechazo son las utilizadas para detectar genes estadísticamente DE. Son llamadas a *dos colas*, porque tanto valores positivos como negativos son evidencia a favor de la hipótesis alternativa $\mathbf{H_1}$**.** Un valor negativo del estadístico corresponderá a un gen expresado diferencialmente hacia abajo (down-regulated) y un valor positivo a uno expresado diferencialmente hacia arriba (up-regulated).

p-valor, es el menor nivel para el cual el test resultaría en rechazo para los datos observados.

¿Cómo se calcula el p-valor?

Llamemos $t_{observado}$ al valor que resulta de reemplazar los datos en la expresión del estadístico *T*, entonces

$$\text{p-valor } (t_{observado}) = P(|T| \geq |t_{observado}|).$$

El p-valor es la probabilidad de obtener un valor tan ó más extremo que el valor observado del estadístico del test, cuando la $\mathbf{H_0}$ es verdadera.

Decisión utilizando p-valores El test resulta en rechazo si p-valor $(t_{observado}) \leq \alpha$

11.5.2 Tipos de errores en múltiples tests de hipótesis.

Consideremos, para cada gen i en particular (i: 1...N), tenemos las siguientes hipótesis

$\mathbf{H_0^{(i)}}$**:** el gen no está DE vs. $\mathbf{H_1^{(i)}}$**:** el gen sí está DE

En la práctica N puede ser muy grande. Por ejemplo, si vamos a realizar N=6000 tests la cantidad de falsos positivos puede ser muy grande. En particular, si ninguno de ellos está DE, esperamos tener 300 falsos positivos. Es necesario controlar la tasa de falsos positivos.

Posibles resultados de N tests de Hipótesis

Realidad	Decisión		Total
	$\mathbf{H_0}$	$\mathbf{H_1}$	
$\mathbf{H_0}$	N_{00}	N_{01}	N_0
$\mathbf{H_1}$	N_{10}	N_{11}	N_1
Total	$\mathbf{N - N_R}$	$\mathbf{N_R}$	**N**

Estamos utilizando la siguiente notación

- N: la cantidad de veces que ser realizan tests, es conocida de antemano.
- N_0: cantidad de hipótesis nulas verdaderas, cantidad de genes realmente DE. Parámetro desconocido.
- N_1: cantidad de hipótesis nulas falsas, cantidad de genes no DE. Parámetro desconocido.
- N_R: cantidad de hipótesis nulas rechazadas , es observable
- N_{01}: cantidad de falsos positivos, variable aleatoria no observable (v.a.n.o.)
- N_{10}: cantidad de falsos negativos,v.a.n.o..
- N_{11}: cantidad de verdaderos positivos,v.a.n.o..
- N_{00}: cantidad de verdaderos negativos, v.a.n.o..

11.5.3 Nivel global - FWER Family wise error rate

La probabilidad de cometer un error de tipo I en alguno de los N tests de hipótesis es denominada *nivel global* o *family wise error rate*. Esto es rechazar $\mathbf{H_0^{(i)}}$ cuando $\mathbf{H_0^{(i)}}$ es verdadera para una o más hipótesis:

$$FWER = P(N_{01} \geq 1)$$

11.5.4 Métodos para controlar el FWER

El método más conocido para controlar el FWER es el de *Bonferroni*. Por este método, si se quiere tener un nivel global de a lo sumo α, se debe tomar para el test de cada gen un nivel corregido $\alpha^{(i)} = \alpha /N$.

Demostración: Tomemos $\alpha^{(i)} = \alpha /N$

$$\text{Sean } N_{01}{}^{(i)} = \begin{cases} 1 \text{ el test resulta en rechazo cuando } \mathbf{H_0^{(i)}} \text{ es verdadera} = R_{01}{}^{(i)} \\ 0 \text{ el test resulta en no rechazo cuando } \mathbf{H_0^{(i)}} \text{ es verdadera} = R_{00}{}^{(i)} \end{cases}$$

por lo tanto $N_{01} = \sum_{i=1}^{N} N_{01}{}^{(i)}$

$$\text{Pero } FWER = P(N_{01} \geq 1) = P(\sum_{i=1}^{N} N_{01}{}^{(i)} \geq 1) = P(\bigcup_{i=1}^{N} R_{01}{}^{(i)}) \leq \sum_{i=1}^{N} P(N_{01}{}^{(i)} = 1)$$

$$= \sum_{i=1}^{N} \alpha^{(i)} = N * (\alpha /N)$$

Luego $FWER \leq \alpha$ cqd.

Observaciones, para Bonferroni.

Se obtiene un nivel global de a lo sumo α,

- Si para un gen (i) el p-valor ($t^{(i)}_{observado}$) = $p^{(i)}$, entonces

p- valor ajustado para el gen i $p_A^{(i)} = \text{mín}(N^* p^{(i)},1)$

- Si todos los p-valores se multiplican por el mismo número N, para ajustarse.

- Si se eligen como estadísticamente DE aquellos genes para los que $p^{(i)} \leq \alpha / N$.

Esta propuesta, da una cota para el nivel global. Puede ser *demasiado conservativa* y cuando la cantidad de tests es grande los niveles corregidos resultan demasiado bajos, equivalentemente los p-valores demasiado altos. Esto significa que se *seleccionarán pocos genes* como candidatos a estar DE.

Método de Sidak

Igual que antes llamemos al nivel global α y $\alpha^{(i)}$ el nivel de cada uno de los tests individuales, que tomamos iguales. Sidak propone tomar $\alpha^{(i)} = 1-\sqrt[N]{(1-\alpha)}$ para obtener un nivel global α. Esta propuesta está apoyada en el supuesto de independencia entre los tests.

Demostración:

$$\alpha = \text{FWER} = P\left(\bigcup_{i=1}^{N} R_{01}^{(i)}\right) = 1 - P\left(\bigcap_{i=1}^{N} R_{00}^{(i)}\right) \underset{\text{independencia}}{=} 1 - \prod_{i=1}^{N} P(R_{00}^{(i)}) = 1 - \prod_{i=1}^{N} (1-\alpha^{(i)})$$

Luego $\alpha = 1-(1-\alpha^{(1)})^N \Rightarrow \alpha^{(i)} = 1-\sqrt[N]{(1-\alpha)}$

p-valor ajustado por Sidac, para el gen i $p_A^{(i)} = 1-(1-p^{(i)})^N$

Equivalentemente, con un nivel global de sumo α, se eligen como estadísticamente DE aquellos genes para los que $p^{(i)} \leq 1-\sqrt[N]{(1-\alpha)}$

p-valores ajustados- Westfall y Young, 1993

Dado un procedimiento que utiliza N tests simultáneos, el p-valor ajustado para una única hipótesis $\mathbf{H_0^{(i)}}$ se define como el mínimo nivel del procedimiento completo al cual $\mathbf{H_0^{(i)}}$ sería rechazada dados los valores observados de todos los estadísticos.

Podemos distinguir tres maneras de ajustar los p-valores:

- procedimiento single-step. Un paso.
- procedimiento step-down. Pasos hacia abajo.
- procedimiento step-up. Pasos hacia arriba.

En los procedimientos de un paso se realizan correcciones equivalentes, para todas las hipótesis, que no dependen del orden de los estadísticos individuales ni de los p-valores no ajustados. Este es el tipo de ajuste que hemos presentado en las propuestas de Bonferroni y Sidac.

Es posible obtener un aumento en la potencia, manteniendo el control de la tasa de error de tipo I mediante procedimientos en pasos en los cuales el rechazo de una hipótesis se se corrige no solo por la cantidad total de hipótesis si no en el resultado de los otros tests.

En los procedimientos de step-down, los p-valores no ajustados (ó los valores observados de los estadísticos del test) se ordenan comenzando con el más significativo, mientras que los procedimientos step-up comienzan por los menos significativos.

Los valores ajustados de un paso son simples de calcular pero tienden a ser muy conservativos. Presentamos a continuación un procedimiento step-down de p-valores ajustado.

Método (step-down) *de Holm para ajustar p-valores*

Dado un nivel global α, el procedimiento de ajuste de los p-valores es el siguiente:

1. Se ordenan los p-valores de c/u de los genes i:1,...., N
   $p^{(1)} \leq p^{(2)} \leq \ .... \leq p^{(i)} \ .... \leq p^{(N)}$
2. Se evalúan las siguientes desigualdades
   $p^{(1)} < \alpha/N \quad p^{(2)} < \alpha/(N-1) \quad .... \quad p^{(i)} < \alpha/(N-i+1) \quad .... \quad p^{(N)} < \alpha/1$
3. Sea k el mayor i para el cual $p^{(i)} < \alpha/(N-i+1) \Leftrightarrow (N-i+1)\, p^{(i)} < \alpha$
4. Se reachaza la $\mathbf{H_0^{(i)}}$ de igualdad de expresión para i=1, ..., k

Lo anterior es equivalente a decir que

p- valor ajustado para el gen i
$p_A^{(i)} = \max_{j=1,..i}(\text{mín}((N-j+1)\, p^{(j)},1)$

El procedimiento de Holm es menos conservativo que Bonferroni cuyos p-valores son multiplicados por N en cada paso. Sin embargo, ni el método de Holm ni el de Bonferroni (ni ningún método de un paso) tiene en cuenta la dependencia entre los estadísticos de los tests, que puede ser muy fuerte para genes corregulados.

11.5.5 Tasa de descubrimientos falsos. FDR false discovery rate.

Consideremos el siguiente ejemplo

Realizamos un test para cada gen i (i: 1,...,20000), tenemos las siguientes hipótesis

$\mathbf{H_0^{(i)}}$**:** el gen no está DE vs. $\mathbf{H_1^{(i)}}$**:** el gen sí está DE

Resultados de los 20000 tests de Hipótesis

Realidad	Decisión		Total
	H_0	**H_1**	
H_0	N_{00}=18790	N_{01}=1000	N_0 =19790
H_1	N_{10}=110	N_{11}=100	N_1=210
Total	**N - N_R=18900**	**N_R=1100**	**N=20000**

La tasa de falsos positivos:

$$1000 / 19790 = 0.050 ,$$

es decir la proporción de veces que un gen fue declarado significativo cuando en realidad no lo era, **es adecuada**.
Sin embargo **la tasa de descubrimientos falsos positivos** (FDR): 1000 / 1100 = 0,909 es decir la proporción de veces que un gen fue declarado significativo y no lo era entre el total de genes declarados significativos, **es inaceptable**.

¡Pero, N_{01} es no observable!

La tasa de falsos descubrimiento global es 10000/20000, también es aceptable. Es menos exigente que el nivel global y depende de la cantidad de tests realizados.

El FDR es la medida adecuada porque interesa hallar la mayor cantidad de genes verdaderamente DE con pocos falsos positivos. El FDR da la tasa a la cual verificaciones biológicas futuras resultarán nulas.

La razón por la cual un bajo nivel global no garantiza una baja tasa de descubrimientos falsos es la altísima prevalencia de genes no DE. Es por esto que para poder controlar el positive FDR y el FDR es necesario estimar esa la prevalencia.

Benjamini y Hochberg (1995) proponen un ajuste de los p-valores de manera de controlar el FDR que está basado en el supuesto de independencia de los p-valores. Es un procedimiento similar al de Holm. Sea, nuevamente, $p^{(i)}$ el p-valor en la iésima posición, se rechaza $\mathbf{H_0^{(i)}}$ cuando $p^{(i)} < (i/N)(\alpha/p_0)$. Como p_0, la proporción de hipótesis nulas realmente verdaderas, es desconocida se la toma en forma consevativa igual a 1.

Método (step-down) *de BH para ajustar p-valores*

Dado un nivel global α, el procedimiento de ajuste de los p-valores es el siguiente:

1. Se ordenan los p-valores de c/u de los genes i:1,...., N
   $p^{(1)} \le p^{(2)} \le .... \le p^{(i)} .... \le p^{(N)}$
2. Se comparan los p-valores con puntos de corte que dependen de la posición del gen en la lista de los p-valoes ordenados
   $p^{(1)} < \alpha/N \quad p^{(2)} < 2\alpha/N \quad .... \quad p^{(i)} < i\alpha/N \quad .... \quad p^{(N)} < \alpha$
3. Sea rechaza la $\mathbf{H_0^{(i)}}$ de igualdad de expresión para aquellos genes cuyos p-valores son menores que la cota correspondiente.

Sea k el mayor i para el cual $p^{(i)} < i\alpha/N \Leftrightarrow (N/i)\, p^{(i)} < \alpha$. Se rechaza la hip. nula para i:1,..,k.

Métodos basados en permutaciones

Westfall y Young (1993) proponen un método step-down basado en permutaciones que tiene en cuenta la estructura de dependencias entre genes. Este método no es recomendado por Speed (2003) debido al pequeño tamaño que tienen las muestras en experimentos de microarrays.

False discovery rate - FDR - Definiciones

- Family-Wise Error Rate: probability of including at least one non-differentially expressed gene
- False discovery rate (FDR): expected proportion of Type I errors among the rejected hypotheses
- pFDR: Expected proportion of false discoveries among the genes in your list conditioning on at least one gene is included in the differential list.

El "false discovery rate" (FDR), tasa de descubrimientos falsos, es la proporción esperada de errores de tipo I entre las hipótesis rechazadas

False discovery rate - Benjamini and Hochberg 1995

$$FDR = E(N_{01}/ N_R ; N_R > 0) = E(N_{01}/ N_R \setminus N_R > 0)\, P(N_R > 0)$$

mientras que el "positive false discovery rate" (pFDR), es la tasa de descubrimientos falsos sabiendo que se la decisión fue rechazo

Positive false discovery rate - Storey 2002

$$pFDR = E(N_{01}/ N_R \setminus N_R > 0)$$

Ver el multtest package

Referencias

• Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society B, 57, 289 -300.
• Holm, S. (1979). "A Simple Sequentially Rejective Bonferroni Test Procedure," Scandinavian Journal of Statistics, 6, 65 -70.
• Smyth GK, Yang YH, Speed T. (2003). "Statistical issues in cDNA microarray data analysis". Methods Mol Biol;**224**:111-36.
• Westfall, PH and SS Young (1993) Resampling-based multiple testing: Examples and methods for p-value adjustment, John Wiley & Sons, Inc

• J Storey (2001): 3 papers (some with other authors), www-stat.stanford.edu/~jstorey/ The positive false discovery rate: a Bayesian interpretation and the q-value. A direct approach to false discovery rates
• Estimating false discovery rates under dependence, with applications to microarrays Y Ge et al (2001) Fast algorithm for resampling based p-value adjustment for multiple testing