Universidad de Valladolid
Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Estadística

Métodos estadísticos para evaluar la expresión diferencial en experimentos de microarrays de ADN

Autora:
D.ª Rocío González Silos

Tutores:
D. Agustín Mayo Íscar
D.ª Itziar Fernández Martínez

Índice

1. INTRODUCCIÓN

El ADN contiene toda la información de la estructura y funcionamiento de un organismo. Pequeñas diferencias en la secuencia de ADN pueden tener un efecto importante sobre la salud y la enfermedad. Aunque todas las células poseen el código genético completo, todos los genes no se expresan en todas las células. La expresión de un gen puede medirse al observar la cantidad de ARN mensajero o de la proteína elaborada con el ARN mensajero que produce ese gen. Los microarrays son una tecnología que permite a los investigadores medir niveles de expresión de miles de genes en un solo experimento.

Uno de los objetivos del análisis de datos de microarrays es la identificación de genes con expresión diferencial, es decir, de genes que varían su nivel de expresión en función de diferentes condiciones biológicas. En algunas situaciones estas condiciones biológicas aparecen recogidas en la variable principal o respuesta en un estudio, siendo las medidas de expresión genética variables candidatas a explicar ese comportamiento diferencial. Uno de los casos más simples corresponde por ejemplo a identificar qué genes presentan diferentes niveles de expresión en individuos enfermos que en individuos sanos, lo que permitirá asociar un grupo de genes a una determinada enfermedad. Estos genes podrían ser candidatos a ser incluidos en pruebas diagnósticas para la enfermedad si adicionalmente las distribuciones de sus niveles de expresión en enfermos y sanos están muy separadas.

Si una investigación reclutara solamente individuos sanos para ser sometidos a un seguimiento durante un periodo de tiempo en el que recoger la aparición o no de una enfermedad y la expresión genética se obtuviese en el momento inicial los genes que fueran identificados con expresión diferencial podrían ser candidatos a ser utilizados en reglas de ayuda en el pronóstico de dicha enfermedad.

Otra situación, a la que corresponden muchas de las investigaciones que se llevan a cabo en la actualidad, aparece cuando se está interesado en la detección de genes ligados a la respuesta o ausencia de respuesta a un tratamiento. En otros casos, las condiciones biológicas pueden constituir variables candidatas para explicar la diferente expresión observada, de esta forma se puede estar interesado en genes con diferente comportamiento en individuos de diferente sexo o en individuos expuestos o no a determinado factor de riesgo.

Todas estas situaciones diferentes corresponderían a la identificación de genes con expresión diferencial en los dos tipos de individuos definidos en cada caso y, en un estudio gen a gen, podrían ser respondidas desde el punto de vista estadístico de forma similar utilizando un contraste de hipótesis en el que la hipótesis nula fuese la igualdad de comportamiento entre los dos tipos de individuo estudiados, en algún sentido pre-determinado, en las distribuciones de expresión de cada gen correspondientes a las dos tipologías estudiadas (enfermos y sanos, individuos que desarrollan enfermedad o no, individuos con respuesta o no al tratamiento, expuestos o no a un factor de riesgo e individuos de diferente sexo).

En el caso de asumir normalidad, el contraste a utilizar en cada gen, podría ser el de igualdad de medias de la t de Student para muestras independientes y en el caso de no asumirla el contraste podría ser el de Wilcoxon-Mann-Whitney basado en rangos para la hipótesis de igualdad de localización de las distribuciones. En todos estos casos, habría que tener en cuenta que *se* realizan tantos contrastes como genes haya y que por tanto, la seguridad que se puede tener en el resultado de uno de ellos no se puede comparar con la que se tendría si solo se realiza un contraste para un gen pre-especificado.

Desde el punto de vista estadístico, se puede responder con el ajuste de modelos más o menos complicados diversas situaciones con múltiples variables explicativas correspondientes a las variables, cuantitativas o cualitativas, y con la tipología demográfica y la situación clínica del individuo, además de la variable correspondiente a la expresión de cada gen que podría jugar o no el papel de variable respuesta.

En todos estos casos, la aplicación de un contraste de hipótesis permitiría evaluar la relevancia o no del modelo para explicar la expresión de cada gen o la relevancia o no de la expresión de cada gen para explicar la variable respuesta, según cual fuera el rol de la variable que contiene la expresión del gen. En estos casos se puede etiquetar también al gen como diferencialmente expresado si el modelo aporta explicación a la variación genética o si el gen la aporta a la correspondiente variable respuesta, según corresponda. En cualquier caso, sea cual sea la situación entre las planteadas, se debe tener en cuenta que se está realizando una multiplicidad de contrastes.

Como hemos señalado, la tecnología de los microarrays nos permite medir miles de genes simultáneamente. Derivándose de esta posibilidad, el problema de contrastar múltiples hipótesis, pertenecientes a alguna de las tipologías anteriormente mencionadas que aparecerán al estudiar problemas biológicos, merecerá una metodología estadística que tenga en cuenta esta multiplicidad. Ésta será muy superior a la prevista en la estadística clásica para las hipótesis que aparecían, por ejemplo, en las comparaciones a pares derivadas del análisis de la varianza o de las que pudieran aparecer en el estudio del comportamiento de la media de una distribución multivariante.

En la realización de dichos contrastes aparecerán los dos posibles errores, el de rechazar hipótesis correspondientes a genes no diferencialmente expresados (error de tipo I) y el de no descubrir genes diferencialmente expresados (error de tipo II). El enfrentamiento habitual en la realización de un solo contraste correspondía a, fijado un nivel de tipo I, tratar de minimizar la probabilidad de error de tipo II. En este caso, la probabilidad de aparición de algún error de tipo I crecerá con el número de contrastes realizados. Además, al realizar muchos contrastes fijando el nivel de significación más convencional, 0.05, se sabe que si los genes implicados tuvieran un comportamiento independiente, se encontraría en media, un 5% de genes con falsa expresión diferencial debida al azar. Serán necesarias por tanto definiciones adecuadas del error de tipo I y de la potencia o de su asociado error de tipo II, que tengan en cuenta la multiplicidad, para poder conocer las propiedades de los contrastes que se

aplican. Esto permitirá en el problema en el que se contrastan miles de hipótesis simultáneamente (una por cada gen), replicar la estrategia que se aplica cuando se quiere contrastar una sola hipótesis de una vez, controlar el error de tipo I en este caso el correspondiente al problema múltiple, maximizando la potencia, también ésta correspondiente al problema definido por las múltiples hipótesis.

En los últimos años han aparecido publicados muchos procedimientos que estudian de manera global el problema del contraste de múltiples hipótesis en las revistas de Estadística, pero muchos más han aparecido publicados en revistas del área de Bioquímica, sobre todo, ligados a la aparición de la tecnología de los microarrays. De estos últimos, muchos corresponden solo a soluciones heurísticas cuyas propiedades estadísticas no están demostradas. El trabajo de Dudoit y otros (2003) constituye un intento de poner orden en toda la confusión creada por la aparición de toda esa metodología. Previamente la monografía de Westfall y Young (1993), que supone uno de los puntos de partida del trabajo anterior, analizó las posibilidades que los procedimientos de remuestreo ofrecían en su aplicación a problemas de contraste de hipótesis múltiples.

Este trabajo de fin de grado quiere facilitar una aproximación a la lectura del trabajo de Dudoit y otros (2003), explicando los ejemplos ahí tratados y añadiendo un análisis de un conjunto de datos provenientes de microarrays y un ejemplo resultante de la aplicación a datos obtenidos por método de Monte Carlo.

2. PROBLEMA DE LOS CONTRASTES MÚLTIPLES

El punto de partida de un problema de contrastes múltiples, a partir de datos provenientes de microarrays, será una matriz de datos de expresión $X= (x_{ji})$. Esta matriz se obtiene por un preprocesado de los datos brutos obtenidos de los microarrays. Es una matriz de tamaño m x n, donde cada fila contiene los datos de un gen y cada columna contiene los datos de un individuo. Esta representación no siempre es así de exacta y, en ocasiones, debido a ambigüedades en la definición de los genes, varias filas podrían estar conteniendo datos de expresión (obtenidos de forma diferente, y por tanto, diferentes) correspondientes al mismo gen. Esta matriz de expresión es distinta a las matrices de datos que se utilizan habitualmente en estadística, por dos motivos, por la localización de los individuos y las variables, ya que las filas suelen representar variables y las columnas individuos, y porque el número de individuos es muy reducido en comparación con el número de variables. En muchos casos el número de individuos es inferior a la decena, siendo experimentos con varias decenas o varios centenares de individuos, más infrecuentes. *El* número de variables puede rondar desde los miles a varias decenas de miles, situación esta última que incluye la correspondiente a analizar todos los genes en el genoma completo. Adicionalmente habrá una matriz de datos conteniendo el resto de información demográfica y clínica del individuo.

El problema de contrastes múltiples aparece por la multiplicidad de las hipótesis nulas a contrastar, una para cada gen, correspondiente a la no existencia de expresión diferencial (en el sentido mencionado en la introducción, correspondiente al modelo de relación escogido entre los valores de expresión de cada gen y las variables clínicas). De esta forma para cada gen j, se tiene la hipótesis nula H_j. Para cada hipótesis se supone definido un estadístico de contraste T_j, basado en la información proveniente de las variables clínicas y de la expresión del gen j de forma que los valores más extremos que uno dado para la hipótesis nula sean los más grandes (en valor absoluto) que ese valor.

Dado que se van a aplicar los múltiples contrastes T_j a las hipótesis H_j simultáneamente, se necesita utilizar una metodología que permita a partir de los valores de los estadísticos decidir que hipótesis rechazar y cuáles no. Muchas de las metodologías disponibles para estudiar esta multiplicidad, entre ellas las que se presentan en esta memoria, pueden aplicarse conjuntamente con cualquier elección de test estadístico para los contrastes individuales.

A continuación se van a introducir una serie de conceptos básicos que se utilizarán al analizar los diferentes procedimientos de contrastes múltiples para el análisis de datos provenientes de microarrays.

2.1 CONTRASTE DE HIPÓTESIS

Un contraste de hipótesis es un método estadístico que permite elegir una hipótesis entre dos posibles. El contraste comienza con la formulación de dos hipótesis incompatibles (si una es cierta, la otra necesariamente ha de ser falsa), que son etiquetadas como hipótesis nula, H_0, e hipótesis alternativa, H_1. El punto de partida es suponer cierta la hipótesis nula, H_0, y la aplicación del contraste trata de medir hasta qué punto la información recogida de la muestra es compatible con H_0. Sólo en el caso de que la información muestral ofrezca fuertes evidencias en contra de que la hipótesis nula H_0 sea cierta, esta será rechazada.

En el siguiente esquema aparecen representadas las cuatro combinaciones posibles (en función de la decisión tomada y de la certeza o no de la hipótesis nula) para todo contraste de hipótesis:

	No se rechaza la hipótesis nula	**Se rechaza la hipótesis nula**
No diferencialmente expresado (H0 cierta)	Acierto	Error tipo I
diferencialmente expresado (H0 falsa)	Error tipo II	Acierto

Tabla 1.

Se define error de tipo I al que se produce al rechazar la hipótesis nula siendo ésta cierta, y error de tipo II al correspondiente a no rechazar la hipótesis nula cuando ésta es falsa. Estos errores aparecen porque al rechazar o no la hipótesis nula, a partir de la

información muestral, no es posible garantizar que la decisión tomada sea la correcta, y por tanto, no se sabrá si se está cometiendo o no error. Cuando se realiza un contraste se mide la probabilidad de cometer estos errores. En los contrastes, se aprovecha el diferente papel jugado por las dos hipótesis para situar como hipótesis alternativa la que corresponde a la decisión que se quiere tomar, y que, por tanto, solo será tomada si los datos aportan mucha evidencia en contra de la nula. Por ello la estrategia al contrastar hipótesis es controlar el riesgo de error al rechazar la H_0, el error de tipo I. Situar este error en niveles bajos permite equivocarse menos veces al rechazar la hipótesis nula, pero producirá un aumento en el error de tipo II, porque al dificultar el rechazo de la hipótesis nula, también será más difícil que cuando esta sea falsa se tome la decisión correcta.

La potencia del contraste corresponderá a la probabilidad de rechazar la hipótesis nula siendo ésta falsa. No se obtiene un único valor de potencia, tendremos uno por cada hipótesis alternativa. Para un mismo umbral de control del error de tipo I, cuanto mayor sea la potencia tendremos un mejor contraste, porque dispondremos de un menor error de rechazo de la hipótesis alternativa cuando es cierta. Habitualmente se nombran con α y con β a las probabilidades de cometer un error de tipo I y de tipo II, respectivamente y, por tanto, a la potencia se nombra con $1 - \beta$.

La elección de un error de tipo I, α, define una región crítica de ese tamaño (para el tipo de test mencionado anteriormente con valores del estadístico (en valor absoluto) más extremos bajo la hipótesis nula tendrá una representación del tipo, $|T|>c_\alpha$, para algún umbral c_α), de forma que observar el valor del estadístico en esa región llevará a rechazar la hipótesis nula. Con la identificación de esta región se puede tomar la decisión de rechazar o no comprobando si el valor observado del estadístico está dentro de ella o no. Al utilizar este tipo de estrategia, pierde la posibilidad de saber si el valor de este estadístico aparece como muy extremo o no para la hipótesis nula. Como alternativa a la toma de decisiones, basada en la determinación de una región critica, está el p-valor.

El p-valor es la probabilidad bajo H_0 de que el estadístico tome valores tan extremos o más que el observado. Su valor corresponde al mínimo nivel α que, de haber sido escogido previamente al experimento, conducirá al rechazo de la hipótesis nula. El p-valor da una medida de lo extremo que es el valor observado del estadístico, suponiendo que la hipótesis nula es cierta. Por ello valores bajos del p-valor, al no dudar de los resultados del experimento y por tanto de la realización obtenida del estadístico, conducirán a dudar de H_0 y por tanto a rechazarla. El p-valor permitirá su aplicación en la toma de decisiones, incluso sin haber fijado previamente el nivel del test o simular la decisión tomada para varias posibles elecciones de este nivel.

2.2 PROBABILIDAD DE ERROR TIPO I

En los experimentos con microarrays, como se ha mencionado anteriormente, no se tiene una sola hipótesis, sino hay que contrastar simultáneamente m hipótesis nulas

H_j, j=1,…, m. La tabla siguiente muestra la situación que aparece tras la realización de los contrastes correspondientes a cada uno de los m genes:

Número de	**Hipótesis no rechazadas**	**Hipótesis Rechazadas**	
Genes no diferencialmente expresados (Verdaderas H0)	**U**	**V**	$\mathbf{m_0}$
Genes diferencialmente expresados (Falsas H0)	**T**	**S**	$\mathbf{m_1}$
	m – R	**R**	**m**

Tabla 2. Reproducida de Dudoit y otros (2003).

En ella aparecen dos parámetros desconocidos, m_0 y m_1 (m_1=m–m_0), que corresponden al número de hipótesis nulas ciertas o genes no diferencialmente expresados y falsas o genes diferencialmente expresados. Estos valores son desconocidos, como es obvio, antes de hacer la investigación, pero también después de llevarla a cabo. R es una variable aleatoria observable, que corresponde al número de hipótesis rechazadas. Las variables aleatorias S, T, U y V, que aparecen al combinar los rechazos con el estado real de la naturaleza en el que se encuentra cada gen (en el sentido de ser o no diferencialmente expresado) son variables que no se pueden observar al no conocer dicho estado.

En la matriz anterior los errores aparecen en las celdas no diagonales, V y T, que corresponden al número de falsos positivos y de falsos negativos respectivamente. Por tanto, lo ideal sería utilizar métodos que minimicen V y T. Pero la estrategia, debido a la ya mencionada asimetría existente entre los dos tipos de hipótesis, debería ser para un cierto control del error tipo I (que habrá que definir, pero que estará relacionado con V que es el número de errores de tipo I) minimizar el error de tipo II (que estará relacionado con T, el número de errores de tipo II). Como cabe esperar, cuando aumenta el número de contrastes, la probabilidad de que aparezca algún gen como diferencialmente expresado sin serlo, aumenta. De hecho, el número esperado de estos falsos positivos crece proporcionalmente al número de contrastes realizados.

A continuación se ofrecen diferentes definiciones de error de Tipo I que aparecen asociadas al problema de los contrastes múltiples

- La definición más clásica para el error del tipo I en esta situación de multiplicidad de contrastes es la FWER (Family-Wise Error Rate) que corresponde a la Pr(V≥1), o lo que es lo mismo a la probabilidad de que aparezcan falsos positivos. Por tanto, se controla que no aparezca ninguno. Ha sido la más utilizada clásicamente ligada a contrastar un número reducido de hipótesis.
- Otra posible definición viene dada por la *PCER* (Per-Comparison Error Rate) que aparece definida como el valor esperado del número de errores tipo I dividido por el número de hipótesis a contrastar: PCER=E(V)/m. Controlar la PCER es

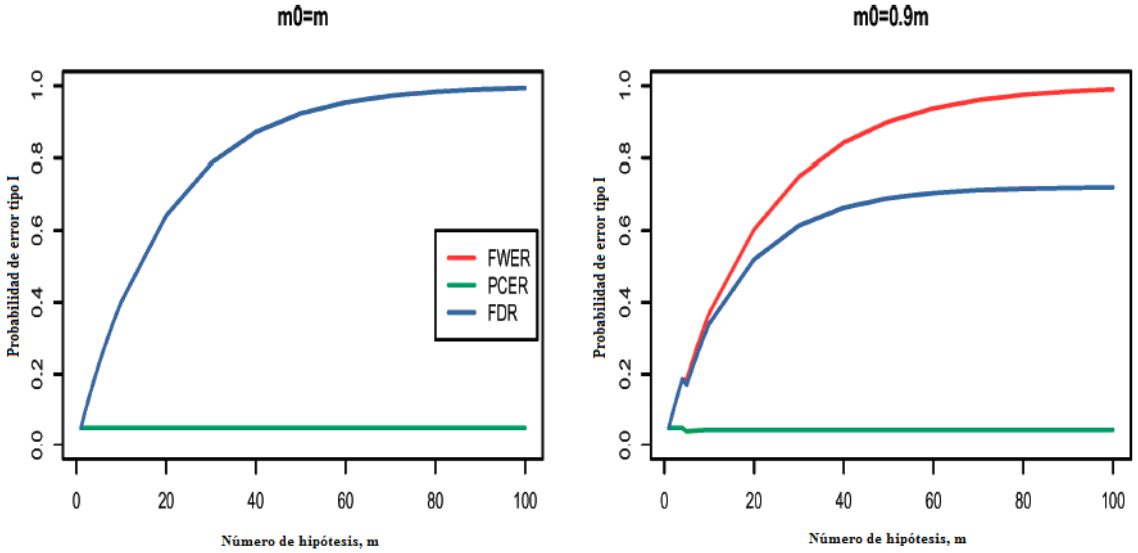
equivalente a controlar la *PFER* (Per-Family Error Rate), que es un múltiplo de ella que aparece al multiplicar por el número de hipótesis.
PFER=E (V)

- Otra definición alternativa corresponde a la *FDR* (False Discovery Rate), definida como la proporción esperada de error tipo I, calculada ésta respecto al número de hipótesis rechazadas.

FDR = E (Q) , con Q= V/R si R>0 y Q=0 si R=0

Estas definiciones verifican las siguientes desigualdades: PCER ≤ FDR ≤ FWER. Estas desigualdades definen un orden para clasificar los procedimientos de contrastes múltiples basados en su control de menos a más conservadores. En este sentido, procedimientos basados en el PCER serán menos conservadores que los basados en la FDR y estos a su vez menos que los basados en el FWER. Los procedimientos más conservadores rechazaran menos veces la hipótesis de no expresión diferencial, incluidas situaciones en las que los genes están diferencialmente expresados.

En la estadística clásica, cuando la multiplicidad venía dada por un número reducido de contrastes, ha sido muy frecuente controlar la FWER, tratando de evitar hasta la aparición de un solo falso positivo. Sin embargo, en los experimentos de microarrays, en los que se plantean miles de contrastes simultáneamente, los usuarios prefieren detectar un número mayor de genes diferencialmente expresados, aún a riesgo de tener algún falso positivo. Por tanto, los procedimientos que controlan la FDR o la PCER son más populares que los que controlan la FWER.



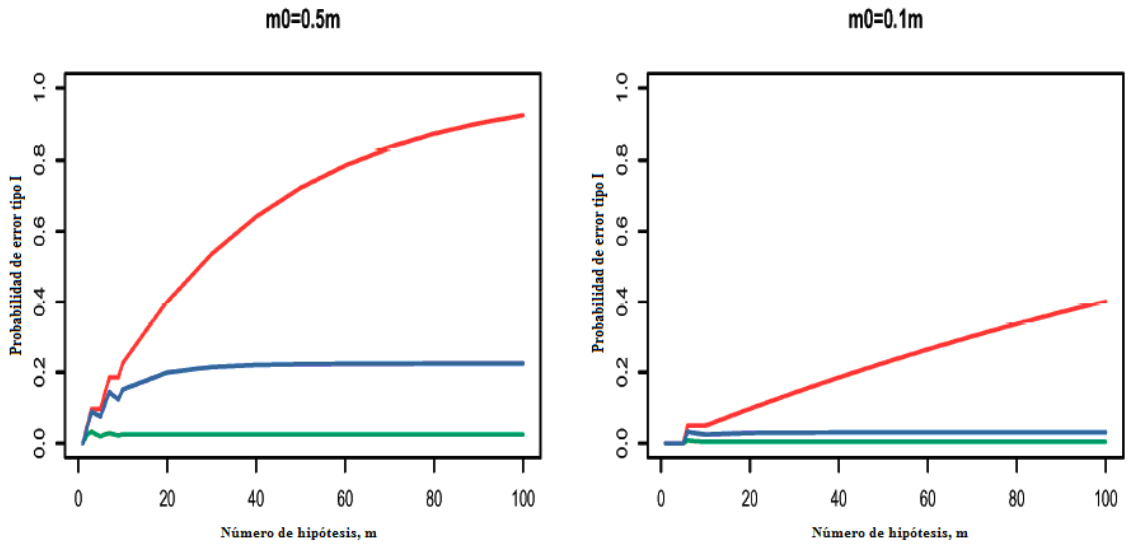


Figura 1. Gráfico de la probabilidad de error tipo I frente al número de hipótesis m para diferentes proporciones de verdaderas hipótesis nulas m_0/m = 1, 0.9, 0.5, 0.1; α=0.05; FWER curva roja, FDR curva azul, PCER curva verde. Figura tomada de Dudoit y otros (2003)

En Dudoit y otros (2003) ofrecen un ejemplo, quizá el más simple posible, para ilustrar el crecimiento de los errores de tipo I (globales para el problema de contrastes múltiples) FWER, PCER y FDR en función del nivel α escogido para controlar los errores de tipo I individuales correspondientes a cada gen, bajo el supuesto de haber elegido el mismo tamaño de error para todos ellos. El ejemplo corresponde a contrastar la hipótesis de que la media es el vector de 0 para una distribución normal m dimensional con marginales independientes de varianza conocida igual a 1. En la generación de los datos, para las primeras m_0 componentes se sitúa la media en 0 y para las restantes en $\frac{d}{\sqrt{n}}$. Para contrastar la hipótesis correspondiente a la media de cada componente el estadístico de contraste que utilizan, que es el más potente bajo normalidad y varianza conocida, sería la media muestral. Suponiendo que de las m hipótesis, las m_0 primeras son ciertas, se tendría que la PCER es igual a $m_0\alpha/m$ y que la FWER es igual a $1-(1-\alpha)^{m_0}$ (debido a que el suceso correspondiente a tener al menos un falso positivo es el complementario del correspondiente a no tener ninguno, que por la independencia corresponde al producto de las probabilidades de no fallo debido a falso positivo individuales, 1-α). La FDR se puede obtener aplicando directamente su definición como esperanza de la proporción de falsos positivos, obtenida ésta respecto del número total de hipótesis rechazadas:

$$E\left(\frac{V}{R}\right)=\sum_{s=0}^{m}\sum_{v=0}^{m_0}\frac{v}{v+s}\,p(V=v)p(S=s)=\sum_{s=0}^{m}\sum_{v=0}^{m_0}\frac{v}{v+s}\,p(V=v)p(S=s)$$

$$=\sum_{s=0}^{m}\sum_{v=0}^{m_0}\frac{v}{v+s}\binom{m_0}{v}\alpha^{v}(1-\alpha)^{m_0-v}\binom{m_1}{s}\beta^{s}(1-\beta)^{m_1-s}$$

Para ello se utiliza la definición de la FDR (por el convenio de resolver la indeterminación 0/0 como 1, incluido en su definición) y que las distribuciones de V y S son binomiales con parámetros correspondientes al número de repeticiones de Bernoulli realizadas, dadas por m_0 y m_1 y con probabilidades de éxito dadas por α y β. La distribución binomial aparece gracias a que los ensayos de Bernoulli son

independientes debido al diseño de la matriz de datos. Las distribuciones binomiales son independientes por el mismo motivo. Las probabilidades de éxito α y β corresponden a la probabilidad de rechazar una hipótesis cuando es cierta, que por construcción es α y a la probabilidad de rechazar una de las m_1 hipótesis falsas por estar la media situada en $d/\sqrt{n}$ (que es función explícita de n, d y α) respectivamente.

La figura 1 representa cuatro gráficos con la probabilidad de error tipo I frente al número de hipótesis m, trabajando con un tamaño del test individual α=0.05 y una media de las diferencialmente expresadas correspondiente a una elección de d=1. Los distintos gráficos corresponden a cuatro situaciones diferentes según el número de hipótesis nulas verdaderas, m_0/m = 1, 0.9, 0.5, 0.1.

Se puede observar como la FWER aumenta fuertemente según crece el número de hipótesis, también se puede apreciar como la PCER se mantiene constante y la FDR se va aproximando cada vez más a ella, la FDR se vuelve más estable cuanto más pequeña se hace la proporción de m_0/m.

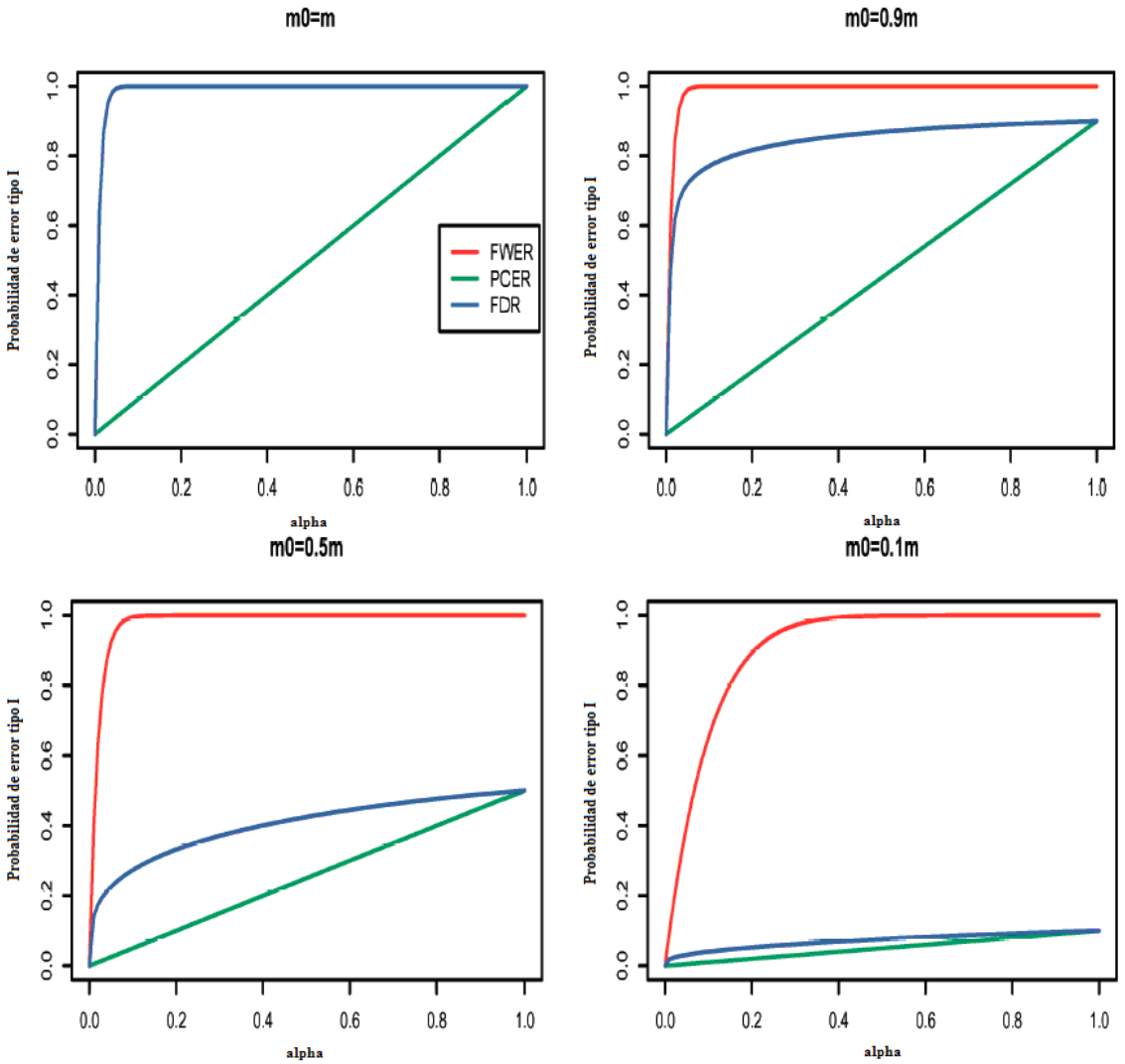


Figura 2. Gráfico de la probabilidad de error tipo I frente a α para diferentes proporciones de verdaderas hipótesis nulas m_0/m = 1, 0.9, 0.5, 0.1; número de hipótesis m=1000; FWER curva roja, FDR curva azul, PCER curva verde. Figura tomada de Dudoit y otros (2003)

En la figura 2 se representan gráficos de la probabilidad de error tipo I frente al test individual de tamaño α; en todos los gráficos se tiene un número de hipótesis m=1000. Los distintos gráficos son según cuatro situaciones correspondientes a un número de hipótesis nulas verdaderas diferentes, $m_0/m = 1, 0.9, 0.5, 0.1$.

Se puede apreciar como la FWER toma valores más grandes que la PCER, siendo la diferencia entre ellos máxima cuando todas las hipótesis nulas son ciertas, también se observa que según va disminuyendo la proporción de m_0/m, la FDR se va aproximando más a la PCER. La FWER va aumentando con el α en forma, en algún sentido similar, a como lo hacía con el número de hipótesis.

2.3 CONTROL FUERTE FRENTE A CONTROL DÉBIL

La FWER, la PCER y la FDR, antes definidas, controlan la probabilidad de error de tipo I independientemente de cual sea el conjunto de hipótesis nulas ciertas, o lo que es equivalente, bajo cualquier combinación de genes no diferencialmente expresados. A esta forma de control sobre el error de tipo I se le conoce como control fuerte.

Frente a este control fuerte algunas metodologías fueron definidas para conseguir un control del error de tipo I (para las probabilidades de error correspondientes a las diferentes definiciones anteriores) solo cuando todas las hipótesis nulas son ciertas. A esta situación se le denomina hipótesis nula completa, $H_0^C = \bigcap_{j=1}^{m} H_j$, y el tipo de control del error de tipo I que aparece ligado a ella se le conoce como control débil.

Es más fácil el cálculo de las probabilidades ligadas al control débil que al control fuerte, pero es necesario tener en cuenta qué genes están diferencialmente expresados, lo cual es desconocido.

El "milagro" de obtener el control fuerte del error de tipo I tras haber asegurado el control débil lo va a permitir la pivotalidad en subconjuntos (subset pivotality). Esta propiedad que corresponde a que la distribución conjunta de cualquier vector de estadísticos/p-valores (T/P) correspondientes a las hipótesis en un subconjunto del microarray, bajo la hipótesis de ningún gen diferencialmente expresado en ese subconjunto, sea la misma que la distribución de ese subconjunto de (T/P) bajo la hipótesis de que ningún gen este diferencialmente expresado en el microarray (hipótesis nula completa). Esta propiedad también será crucial para asegurar el control fuerte cuando se aplican procedimientos de remuestreo para ajustar los p-valores. El remuestreo, como veremos, nos generará estimaciones de la distribución conjunta de los estadísticos bajo la hipótesis nula completa lo que permitirá asegurar el control débil del error de Tipo I. En las situaciones en las que se pueda aplicar esta propiedad, ese control débil será suficiente para asegurar el control fuerte.

Trabajando con datos procedentes de microarrays es, biológicamente, muy poco posible que ningún gen este diferencialmente expresado, por lo que el control débil, que controla errores de tipo I bajo esta hipótesis, no es nada interesante. Controlar el

error solamente en una circunstancia que seguramente no va a suceder, significa no controlar nada, y por tanto, es fundamental tener un control fuerte de la probabilidad de error tipo I.

2.4 POTENCIA

Es fundamental disponer de definiciones de la potencia adaptadas al contexto de los contrastes múltiples. Esto permitirá, cuando se dispone de varios procedimientos, utilizar el mismo tipo de estrategia que se emplea en el caso de un solo contraste: dado un control de la probabilidad de error de tipo I, elegir el procedimiento que maximiza la potencia o que, por tanto, minimiza la probabilidad de error tipo II.

Hay varias definiciones candidatas a generalizar el concepto de potencia en los contrastes múltiples, igual que ocurre con el error de tipo I. Todas están basadas en las variables aleatorias T y S que informan del acierto y del fallo en los genes diferencialmente expresados.

- La probabilidad de rechazar al menos una hipótesis correspondiente a un gen diferencialmente expresado,
  $Pr(S \geq 1) = Pr(T \leq m_1 - 1)$.
- El número esperado de hipótesis que son rechazadas correspondientes a genes diferencialmente expresados
  $E(S)/m_1$.
- La probabilidad de rechazar todas las hipótesis nulas correspondientes a genes no diferencialmente expresados,
  $Pr(S=m_1) = Pr(T=0)$
- Una definición que iría apareada a la de la FDR vendría dada por la esperanza de la proporción de genes diferencialmente expresados con hipótesis rechazadas respecto al total de hipótesis rechazadas
  $E(S/R|R>0)\ P(R>0) = P(R>0) - FDR$

En todas las definiciones anteriores, la potencia depende de la hipótesis alternativa elegida, aunque en el caso de los microarrays las posibilidades de escoger dichas hipótesis son más numerosas que en el contexto univariante.

2.5 P-VALORES NO AJUSTADOS Y AJUSTADOS

El p-valor en un contraste, como ya fue señalado, es una medida del grado de evidencia que ofrecen los datos en contra de la hipótesis nula, que supone una alternativa a la determinación de regiones críticas a la hora de tomar la decisión de rechazar o no en la realización de un contraste de hipótesis. El p-valor, no solo permite tomar la decisión, además da una medida de la proximidad o lejanía a la que se encuentra del umbral de rechazo.

El p-valor no ajustado para el gen i está diseñado para trabajar con un solo contraste ya que no está ajustado por el número de hipótesis contrastadas, este vendría dado por $p_i = P(|T_i| \geq |t_i|\ |H_i)$, donde t_i es el valor observado del estadístico para los datos

provenientes del i-ésimo gen. Si solo interesase contrastar la hipótesis correspondiente a un solo gen, dado p_i, rechazaremos la hipótesis H_i a nivel α si $p_i \leq \alpha$ y no rechazando en otro caso.

Cuando se realizan simultáneamente todos los contrastes, sería interesante disponer de p-valores que estuvieran ajustados por la multiplicidad, $\tilde{p}_i$, y que permitieran con una simple comparación, con el nivel α pre-especificado, del tipo $\tilde{p}_i < \alpha$, tomar la decisión de rechazar o no. De manera análoga a la de los p-valores no ajustados, se definen los p-valores ajustados como el mínimo nivel al que se debería situar α (utilizando el control múltiple elegido) para rechazar el contraste, teniendo en cuenta que se están realizando contrastes simultáneos. De esta forma, por ejemplo, el p-valor ajustado utilizando un control del tipo FWER, viene dado por $\tilde{p}_i = \min\{\alpha / Hj\ es\ rechazada\ a\ nivel\ FWER\ \alpha\}$.

2.6 REMUESTREO

El remuestreo ha sido utilizado en la realización de contrastes de hipótesis como alternativa a los procedimientos paramétricos, evitando suponer hipótesis distribucionales para las observaciones. Incluso en el caso de conocer las distribuciones correspondientes a las variables aleatorias que constituyen las fuentes de aleatoriedad en un experimento, puede ser interesante el uso de estas técnicas de remuestreo para aproximar las distribuciones de estadísticos que aparecen como funciones complejas de aquellas variables. Recientemente se ha incrementado el interés por los procedimientos de remuestreo, que incluyen la permutación de índices, la extracción de muestras de la propia muestra dada por el bootstrap,.. debido a la disponibilidad de modernos ordenadores con grandes capacidades de computo.

En Dudoit y otros (2003) se muestra un ejemplo de las posibilidades del remuestreo en la realización de contrastes múltiples correspondiente a la situación en la que se quiere contrastar simultáneamente la relación entre la expresión en cada uno de los genes y una variable Y. Las permutaciones de las columnas de la matriz de expresión permiten obtener una estimación de la distribución conjunta del vector con los estadísticos de contraste correspondientes a todos los genes bajo la hipótesis de que ningún gen está diferencialmente expresado, en este caso correspondiente a la no ausencia de relación entre la expresión de cada gen y la variable Y. En muchas ocasiones el número total de permutaciones es muy alto y no va a ser posible generar todas. En estos casos se genera un número grande de permutaciones aleatoriamente, B, para basar en ellas la estimación de la distribución del vector de estadísticos. El hecho de permutar columnas completas permite conservar la estructura de correlación en la matriz de expresión. Un detalle importante es que la distribución estimada que se genera con las simulaciones es la correspondiente a que todas las hipótesis nulas son ciertas, por lo que en principio, se basa en esta distribución estimada, se tiene solo un control débil del error de tipo I. Si en lugar de permutar columnas completas, se obtienen permutaciones aleatorias independientes en cada fila se eliminaría la estructura de correlaciones, mantenida en el caso anterior, sustituyéndola por una en la que la expresión de los genes fuera independiente.

No cualquier remuestreo es útil en cualquier situación. Un remuestreo basado en permutaciones, como el anteriormente planteado, no sería conveniente en una situación en la que los datos de la matriz de expresión provinieran de dos tipos de individuos que, en los genes diferencialmente expresados presentan distribuciones con diferente media y que, en los genes no diferencialmente expresados tienen distribuciones con la misma media pero con posiblemente desviación típica diferente. En el caso de aplicarlo, las distribuciones de los genes obtenidas en el remuestreo se igualarían no solo en las medias, también en las desviaciones típicas, que es más restrictivo que lo que permite la hipótesis nula. Este tipo de error en la generación de la distribución estimada bajo la hipótesis nula producirá que la metodología no funcione.

El algoritmo planteado en Dudoit y otros (2003) para obtener los p-valores ajustados en la situación anteriormente planteada correspondiente a contrastar la relación de la expresión genética con una variable es el siguiente:

Para obtener la b-ésima permutación, b=1,…, B:

1. Permutar las n columnas de la matriz de datos X.
2. Calcular los test estadísticos $t_{1,b}$,…, $t_{m,b}$ para cada hipótesis (para cada gen)

La distribución permutada de los test estadísticos T_j para la hipótesis de que todas las hipótesis nulas (H_j, j=1,…, m) son ciertas, viene dada por la distribución empírica de $t_{j,1}$,…, $t_{j,B}$. Para las hipótesis de dos lados, el p-valor permutado para la hipótesis H_j es:

$$p_j^* = \frac{\sum_{b=1}^{B} I\left(|t_{j,b}| \geq |t_j|\right)}{B}$$

donde I(.) es la función indicador, que es igual a 1 si la condición dada en los paréntesis es cierta, sino, vale 0 y donde t_j son los valores observados del estadístico T_j

3. MÉTODOS

Una primera clasificación de las diferentes metodologías disponibles para abordar el problema de contrastar hipótesis múltiples de forma simultánea está relacionada con la forma de ajustar los p-valores, si ésta se realiza en un solo paso o secuencialmente.
Según esta clasificación tenemos:

- *Procedimientos un-paso (Single-step)*: Los ajustes de multiplicidad se realizan en un solo paso de la misma forma para todas las hipótesis, y de forma independiente a los resultados observados en los otros genes, sin tener en cuenta algún orden de los genes en función de valores de los estadísticos o de los p-valores sin ajustar.

- *Procedimientos secuenciales (Stepwise)*: En este caso el ajuste del p-valor de cada gen se realiza tras ordenar los p-valores sin ajustar (o los estadísticos de los contrastes). El objetivo es obtener ventajas en la potencia del método manteniendo el control de la probabilidad de error tipo I. En ellos, el rechazo de cada hipótesis se basa en la cantidad total de hipótesis y en el resultado de los contrastes para las hipótesis anteriores, ordenadas según sus p-valores. Según este orden aparecen dos sub-tipologías, los procedimientos *secuenciales hacia abajo (Step-down),* que utilizando el orden de los p-valores correspondiente a comenzar por el más significativo, van ajustando p-valores hasta llegar a uno no estadísticamente significativo, momento en que no se rechazan más hipótesis. En los procedimientos *secuenciales hacia arriba (Step-up),* se ordenan los p-valores comenzando por el menos significativo, obteniendo el ajuste de los mismos hasta rechazar la primera hipótesis nula, momento a partir del cual, todas son rechazadas.

Otra clasificación, más importante que la anterior se puede hacer atendiendo al tipo de error de tipo I controlado. Según éste aparecerán procedimientos que controlan la FWER, la PCER y la FDR.

Al final de la sección se incluye la tabla 3 que procede de Dudoit y otros (2003) y que muestra las propiedades de procedimientos disponibles para contrastes múltiples. En ella aparece el nombre del procedimiento, la clase de control que ofrece para la probabilidad de error tipo I e información sobre si su aplicación es secuencial o no. Algunos procedimientos incluidos en esta tabla como los debidos a Golub y colaboradores no se van a comentar en este trabajo ya que, al no proporcionar un control fuerte sobre la probabilidad de error de tipo I, su interés es menor.

3.1 MÉTODOS BASADOS EN EL CONTROL DE LA FWER

Los procedimientos que controlan la FWER, como ya se ha señalado, tratan de controlar la probabilidad de que algún gen no diferencialmente expresado aparezca entre los correspondientes a hipótesis rechazadas. Dentro de estos procedimientos, su aplicación en un solo paso o paso a paso, servirá para subclasificar estos procedimientos.

3.1.1 Procedimientos single-step

El primer procedimiento que vamos a comentar es debido a Bonferroni, que quizá sea el más conocido y uno de los más utilizados dentro de los procedimientos de comparaciones múltiples en la Estadística clásica. Esta popularidad no se debe precisamente a su potencia, que al ponerse en la peor de las situaciones posible es generalmente muy baja, sino a su simplicidad. El ajuste de los p-valores corresponde simplemente a multiplicar el p-valor sin ajustar por el número de hipótesis a contrastar.

El p-valor ajustado viene, por tanto, dado por $\tilde{p}_j = \min\left(mp_j, 1\right)$. Es fácil obtener una demostración basada en la desigualdad de Boole (desigualdad que dice que en un conjunto finito de sucesos - aunque es más general permitiendo incluso conjuntos infinitos numerables - la probabilidad de que al menos uno de esos sucesos ocurra es menor o igual a la suma de las probabilidades de los sucesos individuales). En esta demostración, que aparece en Dudoit y otros (2003), supone que los genes no diferencialmente expresados son los m primeros y obtienen:

$$FWER = \Pr(V \geq 1/\Lambda) = \Pr\left(\bigcup_{j=1}^{m_0}\left(\tilde{P}_j < \alpha\right)\Big/\Lambda\right)$$

$$= \sum_{j=1}^{m_0}\Pr\left(\left(\tilde{P}_j < \alpha\right)/\Lambda\right) = \sum_{j=1}^{m_0}\Pr\left(\left(P_j < \frac{\alpha}{m}\right)\Big/\Lambda\right) \leq \frac{m_0}{m}\alpha \leq \alpha$$

donde Λ representa la intersección de las m_0 primeras hipótesis nulas, que se han supuesto ciertas, y donde la penúltima desigualdad viene de una propiedad que define los p-valores, como es $\Pr\left(\left(P_j < x\right)/\Lambda\right) \leq x$, para cualquier $x \in [0,1]$, dándose la igualdad para distribuciones continuas de los estadísticos subyacentes (en este caso de continuidad los p-valores P_j, estudiados como variable aleatoria, tienen distribución uniforme).

Otro procedimiento interesante, que como el anterior permite que de forma sencilla se puedan obtener sus propiedades, es el debido a Šidák. Este procedimiento hace la suposición de que las distribuciones de los genes no diferencialmente expresados son independientes. Bajo esta hipótesis, controla de forma fuerte la FWER, y, exactamente a nivel α en distribuciones continuas cuando ningún gen aparece como diferencialmente expresado. El p-valor ajustado viene dado por $\tilde{p}_j = 1-\left(1-p_j\right)^m$. De forma fácil se puede comprobar que controla, como se decía, la FWER:

$$FWER = \Pr(V \geq 1/\Lambda) = \Pr\left(\bigcup_{j=1}^{m_0}\left(\tilde{P}_j < \alpha\right)\Big/\Lambda\right)$$

$$= 1 - \Pr\left(\bigcap_{j=1}^{m_0}\left(\tilde{P}_j \geq \alpha\right)\Big/\Lambda\right) = 1 - \prod_{j=1}^{m_0}\Pr\left(\tilde{P}_j \geq \alpha/\Lambda\right) = 1 - \prod_{j=1}^{m_0}\Pr\left(1-\left(1-P_j\right) \geq \alpha^{\frac{1}{m}}\Big/\Lambda\right)$$

$$= 1 - \prod_{j=1}^{m_0}\Pr\left(P_j \geq \alpha^{\frac{1}{m}}\right) = 1 - \left(1-\alpha^{\frac{m_0}{m}}\right) = \alpha^{\frac{m_0}{m}} \leq \alpha$$

El problema que tiene este procedimiento es que en los experimentos de microarrays, en muchas ocasiones, ni el test estadístico ni el p-valor ajustado son independientes, ya que existen en el genoma grupos de genes que tienden a tener niveles de expresión fuertemente correlados. Debido a haber utilizado la hipótesis de independencia, el procedimiento de Šidák no garantiza un control de la FWER para cualquier distribución conjunta del vector de estadísticos utilizados en el contraste. Este control estaría garantizado para estadísticos cuyas distribuciones verifiquen la desigualdad de Šidák:

$$\Pr\left(|T_1| \leq c_1, \ldots, |T_m| \leq c_m\right) \geq \prod_{j=1}^{m}\Pr\left(|T_j| \leq c_j\right)$$

Esta desigualdad la verifican, entre otras, la familia de distribuciones Normal multivariante centrada en el 0 y la distribución t multivariante.

El efecto de las correcciones de los p-valores producidas por el ajuste de tipo Šidák o Bonferroni aumenta con el número de genes, como se ha visto. Cuando el número de genes implicados en los contrastes múltiples sea muy grande, como el que podemos encontrar en un microarray, lo que ocurre es que estos ajustes dificultarán fuertemente la identificación de genes como diferencialmente expresados. El uso de procedimientos que tengan en cuenta la estructura de correlación incluida en la matriz de datos podría ayudar a disminuir este efecto en casos en los que haya una correlación relevante.
Estrategias que tienen en cuenta esta estructura de correlación son el máx T (basado en máximo valor de los estadísticos utilizados en los contrastes) y el mín P (basado en el mínimo p-valor). Estas dos estrategias son equivalentes cuando la distribución de los estadísticos es la misma. En situaciones que no son estándar en el contexto de los microarrays, donde se hubieran utilizado diferentes estadísticos para cada gen o que los tamaños muestrales no fueran iguales, la distribución de los estadísticos bajo la hipótesis nula sería diferente y debería utilizarse el mínimo P-valor que, de alguna forma estaría estandarizado por estas diferencias.
Los p-valores ajustados correspondientes al mín P y al máx T son respectivamente $\tilde{p}_j = \Pr\left(\min_{1\le l\le m} P_l \le p_j \mid H_0^c\right)$ y $\tilde{p}_j = \Pr\left(\max_{1\le l\le m} |T_l| \ge |t_j| \mid H_0^C\right)$, (expresados en función de H_0^c que representa la hipótesis nula completa, la correspondiente a que todos los genes están diferencialmente expresados). En situaciones en las que se verifique la pivotalidad en subconjuntos se podrá extender el control débil del error de tipo I que asegura esta definición de p-valores ajustados a control fuerte. La utilización de técnicas de remuestreo será fundamental para la obtención de estimaciones de los p-valores correspondientes a estos procedimientos.

Estos procedimientos son más potentes que el de Bonferroni y el de Šidák, en todas las situaciones, siendo igualados por el de Šidák bajo independencia. El interés de utilizar la estructura de correlación está motivado por aumentar la potencia de los procedimientos. En un caso extremo, que tampoco se espera que ocurra en el contexto de los microarrays, correspondiente a que la expresión en todos los genes estuviera perfectamente correlada, el estadístico del mínimo P-valor coincidiría, trivialmente, con cualquier p-valor sin ajustar al ser todos iguales. La ventaja que ofrecería esta situación no sería aprovechada al aplicar los ajustes de Bonferroni o Sidak que producirían p-valores ajustados muchísimo más grandes que los no ajustados, sobre todo para grandes valores de m que son los que aparecen en este contexto biológico.

3.1.2 Procedimientos stepwise:

Los procedimientos secuenciales con toma de decisiones basados en el orden de los p-valores pueden mejorar la potencia de los métodos basados en un solo paso.
Dentro de ellos estarían los procedimientos *Step down*, secuenciales hacia abajo, en los que se ordenan los p-valores ajustados de menor a mayor. Secuencialmente, siguiendo este orden, se toma la decisión de rechazar o no cada hipótesis. Los procedimientos basados en un solo paso se pueden modificar para ser aplicados en

forma secuencial. La idea intuitiva que mueve el interés por estos métodos secuenciales (concretamente los secuenciales Step down) es la siguiente: en los métodos de un solo paso todos los p-valores son ajustados utilizando el mismo criterio, en cambio, al trabajar secuencialmente se ajustaría el primer p-valor teniendo en cuenta el número total de hipótesis a contrastar m, si resulta significativo se podría pensar que ese gen está diferencialmente expresado y, por tanto, no debería ser tenido en cuenta a la hora de controlar el error de tipo I (que está basado en los no diferencialmente expresados), por lo que en el ajuste del siguiente gen se podría trabajar como si solo hubiera m-1 hipótesis a contrastar. Esta forma de trabajar se puede extender iterativamente y tras rechazar la hipótesis correspondiente al k-ésimo gen se puede ajustar el siguiente p-valor teniendo en cuenta una multiplicidad dada por m-k genes. De esta forma se consiguen contrastes menos conservadores.
Entre los procedimientos estudiados el step-down Holm sería la versión secuencial hacia abajo del de Bonferroni y sus p-valores ajustados vienen dados por

$$\tilde{p}_{rj} = \max_{k=1,\ldots,j} \left\{ \min((m-k+1)\, p_{r_k}, 1) \right\}$$

Este procedimiento, como ha sido señalado, se espera que sea menos conservador que el procedimiento en el que se basa, el de Bonferroni, que para obtener los p-valores ajustados multiplica los no ajustados por m. La definición utilizada para los p-valores ajustados quiere asegurar que estos conserven el orden observado en los p-valores sin ajustar.
La réplica *step down* de los procedimiento de Šidák, min P y max T tendrá p-valores ajustados dados por

Step down Šidák $$\tilde{p}_{rj} = \max_{k=1,\ldots,j} \left\{ 1-(1-p_{r_k})^{(m-k+1)} \right\}$$

Step down mín P $$\tilde{p}_{rj} = \max_{k=1,\ldots,j} \left\{ \Pr\left(\min_{l \in \{r_k,\ldots,r_m\}} P_l \le p_{r_k} \mid H_0^C \right) \right\}$$

Step down máx T $$\tilde{p}_{rj} = \max_{k=1,\ldots,j} \left\{ \Pr\left(\max_{l \in \{r_k,\ldots,r_m\}} |T_l| \ge |t_{r_k}| \mid H_0^C \right) \right\}$$

También existen estrategias secuenciales step up, o secuenciales hacia arriba, que partiendo del p-valor menos significativo van tomando la decisión de rechazar o no, hasta encontrarse con el primer test a rechazar, a partir del cual todos son significativos. Muchos de estos procedimientos están basados en la desigualdad de Simes: para test estadísticos independientes, el p-valor no ajustado ordenado $P_{(1)} \le \ldots \le P_{(m)}$ satisface

$$\Pr\left(P_{(j)} > \frac{\alpha j}{m}, \forall j = 1,\ldots,m \mid H_0^C \right) \ge 1-\alpha$$ (en caso continuo se dará la igualdad).

Entre ellos el procedimiento *step up* de Hochberg [dado $j^* = \max\left\{ j : p_{r_j} \le \frac{\alpha}{(m-j+1)} \right\}$, rechazar las hipótesis H_{rj}, para j=1…j*. Si no existe j*, no se rechaza ninguna hipótesis] utiliza la misma penalización que el de Holm para construir p-valores ajustados.

3.2 MÉTODOS BASADOS EN EL CONTOL DE LA FDR

El control de la FWER es muy conservador por obsesionarse en evitar la aparición de un solo falso positivo en la búsqueda de expresión diferencial. Los métodos de control del error de tipo I basados en la FDR controlan la proporción de falsos positivos entre los genes declarados como diferencialmente expresados correspondientes a los contrastes de hipótesis rechazados. De esta forma el investigador obtendrá una lista de genes diferencialmente expresados, entre los que espera una proporción de falsos positivos como la que haya prefijado.

La FDR se define como FDR = E(Q), donde Q = V/R si R> 0 y 0 si R = 0, es decir, $FDR = E(\frac{V}{R} \mid R > 0)\Pr(R > 0)$. La diferente definición dependiendo de que R valga 0 o no, se utiliza para evitar la indeterminación 0/0 resolviéndola a 0.

Bajo la hipótesis nula completa, la FDR es igual a la FWER; por tanto, procedimientos que controlan la FDR controlan, en el sentido débil, la FWER.

Entre los métodos desarrollados para controlar la FDR están el step-up de Benjamini y Hochberg [dados los p-valores ordenados de menor a mayor, el mayor j de tal manera que $P_{r_j} \leq \frac{j}{m}\alpha$, y entonces rechazar todas las H_{r_k} para k=1,…, j. Si no existiese j, no se rechazaría ninguna hipótesis], con p-valores ajustados dados por $\tilde{p}_{r_j} = \min_{k=j,\ldots,m}\left\{\min(\frac{m}{k}\, p_{r_k}, 1)\right\}$ y que consigue un control fuerte de la FDR si la distribución de los estadísticos es independiente. Otra versión de este procedimiento debida a Benjamini y Yekutieli modifica la constante utilizada en la construcción de los p-valores ajustados anteriores, por una superior ($P_{r_j} \leq \frac{k}{m \cdot \sum_{j=1}^{m} \frac{1}{j}}\alpha$) para asegurar el control de la FDR independientemente de la estructura de correlación encontrada en los datos.

3.3 MÉTODOS BASADOS EN EL CONTROL DE LA PCER

Los procedimientos de contrastes múltiples basados en el control de la PCER (la proporción de falsos descubrimientos en relación al número total de hipótesis planteadas) son otra alternativa a los métodos que controlan la FWER. Al igual que los procedimientos que controlan la FDR, estos admiten una proporción prefijada de falsos positivos (controlan su valor esperado). Los métodos que controlan la PCER, lo hacen situando en el denominador el número de decisiones tomadas, a diferencia de los basados en el control de la FDR, que sitúan en el denominador el número de rechazos realizados.

La utilización de p-valores no ajustados para tomar decisiones permite controlar la PCER bajo la suposición de independencia de los contrastes. Considerar como diferencialmente expresados genes cuyos p-valores sin ajustar son inferiores a un nivel α prefijado controla la PCER a este nivel.

Incluido dentro de este marco de control de la PCER está, también, el procedimiento más popular en la aplicación de contrastes múltiples para encontrar expresión diferencial en microarrays, el SAM.

3.3.1 Significance Analysis of Microarrays (SAM)

Bajo esta denominación han aparecido dos procedimientos el de Tusher, Tibshirani y Chu (2001) y otro debido a Efron (y otros colaboradores) que no ha suscitado el mismo interés por controlar la PCER solo en el sentido débil, es decir, bajo la hipótesis de ningún gen diferencialmente expresado (Dudoit y otros, 2003).

A continuación describimos el algoritmo asociado a la obtención del SAM. Se calculan los valores del estadístico t_j para cada gen j y su respectivo estadístico ordenado $t_{(j)}$ tal que $t_{(1)} \geq \ldots \geq t_{(m)}$. Seguidamente, se realizan B permutaciones de las respuestas $y_1,\ldots, y_n$, y para cada una de estas permutaciones se calcularán los test estadísticos $t_{j,b}$ y su correspondiente estadístico ordenado $t_{(1),b} \geq \ldots \geq t_{(m),b}$; a partir de estas B permutaciones se obtiene el estadístico esperado para el gen que ocupa la posición j (bajo la hipótesis nula completa) como el promedio de los B valores del estadístico que ocupan dicha posición en cada una de las permutaciones, $\bar{t}_j = \frac{1}{B}\sum_b t_{(j),b}$.

Después, se usa un QQ-plot en el que se representa el estadístico esperado frente al observado, y así, fijando un umbral Δ, para la diferencia entre estos dos estadísticos, se obtienen dos puntos de corte globales, k_1 y k_2, que darán lugar a una región de rechazo.

Para un determinado Δ, el número esperado de falsos positivos, PCER, se estima calculando para cada una de las B permutaciones el número de genes con $t_{j,b}$ superiores al punto de corte superior o inferiores al punto de corte inferior y promediando este número por las permutaciones.

Hay que escoger un valor Δ para poder controlar el número esperado de falsos positivos, bajo la hipótesis nula completa, a un nivel nominal aceptable.

El SAM de Tusher, Tibshirani y Chu (2001) rechaza H_j siempre que $t_j \geq k_2$ o $t_j \leq k_1$ donde k_2 y k_1 son los puntos de corte escogidos a partir del QQ-plot de tal manera que se tiene un control fuerte de la PCER a un nivel dado, estas regiones de rechazo son interesantes ya que pueden ser asimétricas, es decir, puede que los contrastes sean más potentes cuando hay más genes sobre-expresados que infra-expresados o al revés.

La estimación de la FDR que propone el SAM es diferente que la que hemos presentado anteriormente y corresponde a E (V | H_0^C) / R en lugar de E (V/R | H_0^C).

Procedimiento	**Probabilidad de error tipo I**	**Control fuerte o débil**	**Estructura stepwise**
Bonferroni	FWER	Fuerte	Single
Šidák	FWER	Fuerte	Single

Min P	FWER	Fuerte	Single
Máx T	FWER	Fuerte	Single
Holm	FWER	Fuerte	Down
Step-down Šidák	FWER	Fuerte	Down
Step-down min P	FWER	Fuerte	Down
Step-down max T	FWER	Fuerte	Down
Hochberg	FWER	Fuerte	Up
Troendle	FWER	Fuerte	Up
Benjamini y Hochberg	FDR	Fuerte	Up
Benjamini y Yekutieli	FDR	Fuerte	Up
Yekutieli y Benjamini	FDR	Fuerte	Up
P-valores no ajustados	PCER	Fuerte	Single
SAM, Tusher, Tibshirani y Chu (2001)	PFER(PCER)	Fuerte	Single
SAM, Efron et al.	PFER(PCER)	Débil	Single
Golub et al. step-down	Pr*(R ≥ r* | H*C0)* (FWER)	Débil	Down
Golub et al. step-up	Pr*(R ≥ r* | H*C0)*	Débil	Up

Tabla 3. Propiedades de los procedimientos de contrastes múltiples. Reproducida de Dudoit y otros (2003).

4. Ejemplos de Dudoit y otros (2003)

4.1 DATOS DE MICROARRAYS

En los ejemplos de experimentos de microarrays utilizados en Dudoit y otros (2003), se sigue un mismo esquema, primero se cuenta en qué consiste el estudio, después se aplica a cada uno de ellos los procedimientos de contrastes múltiples descritos en el punto 3 usando permutaciones para estimar los p-valores no ajustados y ajustados y por último se explican mediante gráficos que ayudarán a comparar los diferentes métodos.

4.1.1 Experimento de la apolipoproteína AI de Callow et al.

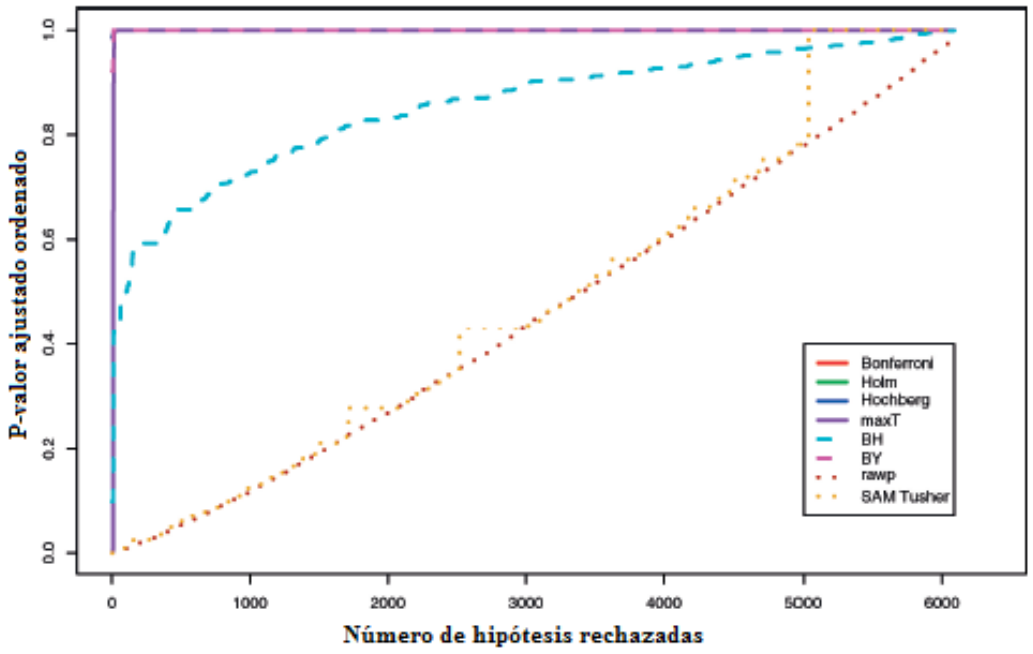
El gen de la apolipoproteína AI (apo AI) tiene un papel importante en el metabolismo del HDL, por lo que se realizó un experimento que consistía en comparar 8 ratones control con 8 ratones tratamiento a los cuales se les había eliminado este gen.
El objetivo era identificar los genes que variaron su nivel de expresión, partiendo de una matriz de datos X de la expresión diferencial.
Se calculó el estadístico t-Welch para dos muestras independientes en cada unos de los genes de manera independiente para identificar los genes diferencialmente expresados.

Los p-valores no ajustados y ajustados se estimaron basándose en todas las posibles permutaciones de tratamiento/control $\binom{16}{8} = 12870$, salvo el SAM en el cual se usaron B_{SAM}= 1000 permutaciones.

Se consideran tanto el p-valor no ajustado (rawp) como los p-valores ajustados utilizando los siguientes procedimientos de contrastes múltiples,

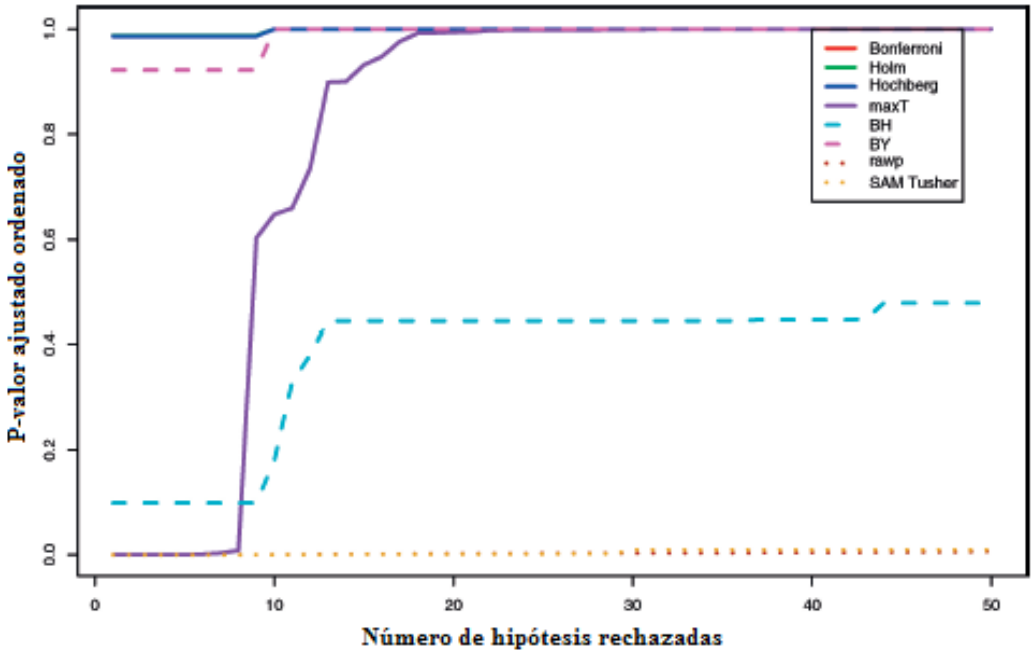
- Control de la FWER: Bonferroni, Holm, Hochberg y maxT
- Control de la FDR: Benjamini & Hochberg y Benjamini & Yekutieli
- Control de la PCER: SAM y rawp

Los dos siguientes gráficos, 1.(a) y 2.(b) son gráficos que representan en el eje x el número de hipótesis rechazadas y en el eje y, el p-valor ajustado ordenado, son los dos el mismo gráfico con distinta escala en el número de hipótesis rechazas.
Se puede apreciar como los p-valores ajustados son más pequeños para los procedimientos que controlan la PCER, el SAM Tusher y el procedimiento de los p-valores no ajustados rawp, de tal manera que rechazan alrededor de 400 hipótesis nulas (declaran que hay unos 400 genes expresados diferencialmente), mientras que los p-valores más grandes son para la mayor parte de los procedimientos que controlan la FWER, Bonferroni, Holm, Hochberg, por su parte el máx T detecta unos 8 genes diferencialmente expresados. El resto de los procedimientos representados no rechazan ninguna hipótesis nula.



(a)

Gráfico 1. (a) Gráfico de los p-valores ajustados permutados ordenados $\tilde{p}_j^*$ frente a j. P-valores ajustados de los procedimientos que controlan la FWER, la FDR y la PCER se representan con una línea sólida, discontinua y a puntos respectivamente. Gráfico reproducido de Dudoit y otros (2003)



(b)

Gráfico 2. (b) Gráfico de los p-valores ajustados permutados ordenados $\tilde{p}_j^*$ frente a j. P-valores ajustados de los procedimientos que controlan la FWER, la FDR y la PCER se representan con una línea sólida, discontinua y a puntos respectivamente. Gráfico reproducido de Dudoit y otros (2003)

El siguiente gráfico 3.(c) se llama volcano plot, es un tipo de gráfico de dispersión que se utiliza para identificar rápidamente cambios en un gran conjunto de datos, en el eje

de las x se representa el estadístico test y en el de la y, el logaritmo negativo en base 10 del p-valor ajustado.
De esta manera, los puntos que se encuentren más arriba y más separados del 0, bien a la derecha o bien a la izquierda representan a genes con valor del estadístico grande (en valor absoluto) y muy significativos.

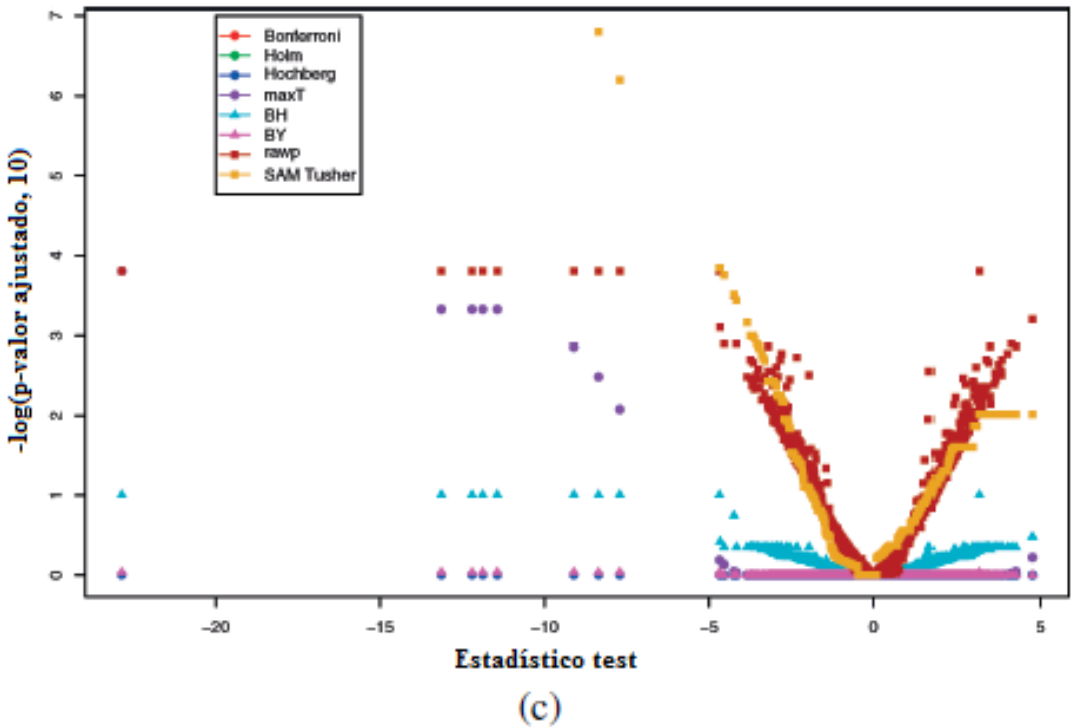


Gráfico 3. (c) Gráfico de p-valores ajustados $-\log_{10} \tilde{p}_j^*$ frente al test estadístico t_j. Gráfico reproducido de Dudoit y otros (2003)

Se puede apreciar como todos los genes en todos los procedimientos tienen valores del test estadístico entre -5 y 5, salvo unos pocos que tienen valores más negativos, estos, a su vez son los que tienen valores más elevados en el eje y, con respecto a cada procedimiento; estos nos da la idea de que estos genes están infra-expresados.
Se puede apreciar fijándose globalmente en los tres gráficos anteriores que tienen comportamientos muy similares los procedimientos de Bonferroni, Holm y Hochberg por un lado, y por otro el SAM Tusher y el de los p-valores no ajustados (rawp).
Por último, en el gráfico cuantil-cuantil 4.(d) se representa en el eje de las x los cuantiles de las permutaciones y en el eje y los cuantiles observados. Si los genes estuviesen igualmente expresados, este gráfico tendría que tener todos los puntos en la diagonal, sin embargo, vemos como en los valores más pequeños de los dos ejes, hay unos puntos que no se encuentran próximos a esa línea diagonal, esos son los que se podrían declarar diferencialmente expresados.

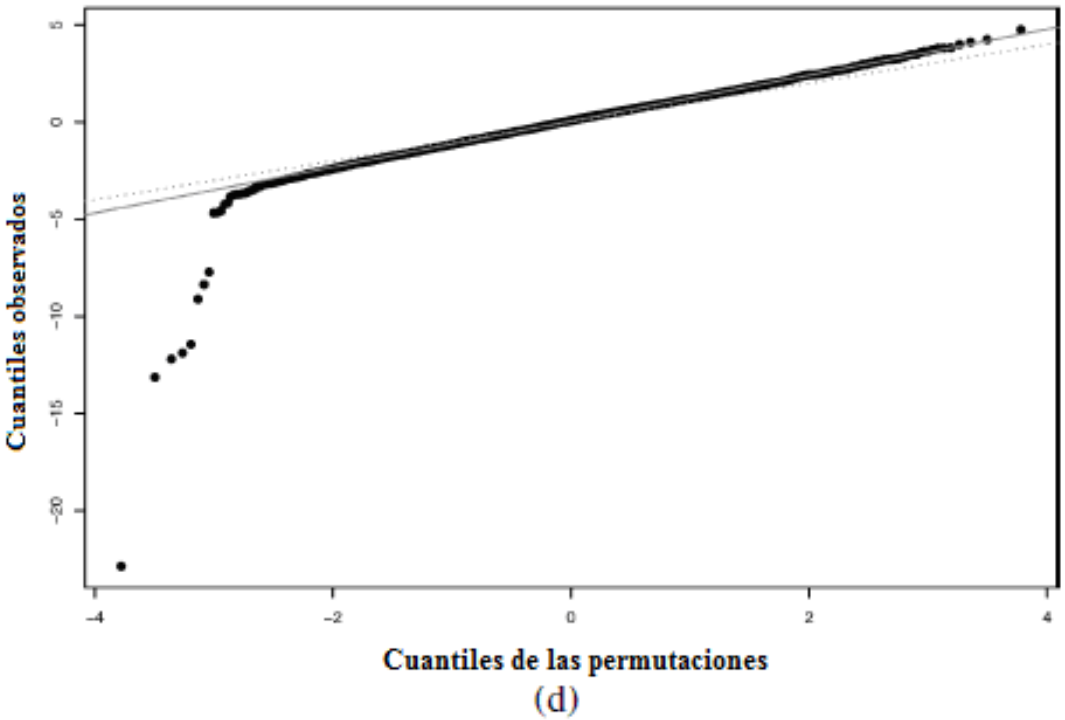


Gráfico 4. (d) Gráfico cuantil-cuantil del estadístico t; con una línea continua se representa la línea identidad y con una discontinua, la línea que pasa por el primer y el tercer cuantil. Gráfico reproducido de Dudoit y otros (2003)

En este experimento, se trabajó con el procedimiento máx T, el cual, con p-valores ajustados menores que 0.05 diferenciaba 8 secuencias de ADN correspondientes solo a 4 genes, el gen eliminado que tenía tres copias, y otros tres con dos, dos y una copia asociados directamente con el apo AI.

4.1.2 Estudio de la leucemia de Golub et al.(1999)

Golub et al. (1999) estaban interesados en la identificación de los genes que se expresan diferencialmente en los pacientes con dos tipos de leucemia, la leucemia linfoblástica aguda (ALL, clase 1) y la leucemia mieloide aguda (AML, clase 2).
Se hizo el estudio para un conjunto de 38 muestras, de las cuales 27 eran de la clase ALL y 11 de la clase AML.

Los genes expresados diferencialmente en los pacientes con ALL y AML se identificaron calculando el estadístico Welch de dos muestras independientes en cada uno de los genes de manera independiente.

Se consideran tanto el p-valor no ajustado (rawp) como los p-valores ajustados utilizando los siguientes procedimientos de contrastes múltiples,

- Control de la FWER: Bonferroni, Holm, Hochberg y maxT
- Control de la FDR: Benjamini & Hochberg y Benjamini & Yekutieli
- Control de la PCER: SAM y rawp

Los p-valores no ajustados y ajustados se estimaron basándose en 500000 permutaciones de ALL/AML, salvo el SAM en el cual se usaron B_{SAM}= 1000 permutaciones.
En los dos gráficos que vienen a continuación del número de hipótesis rechazadas frente al p-valor obtenido ordenado se puede observar el mismo comportamiento que en los gráficos del ejemplo de la apo AI, obteniéndose p-valores más pequeños para procedimientos que controlan la PCER (SAM Tusher y rawp) y p-valores más grandes para procedimientos que controlan la FWER (Bonferroni, Holm, Hochberg).

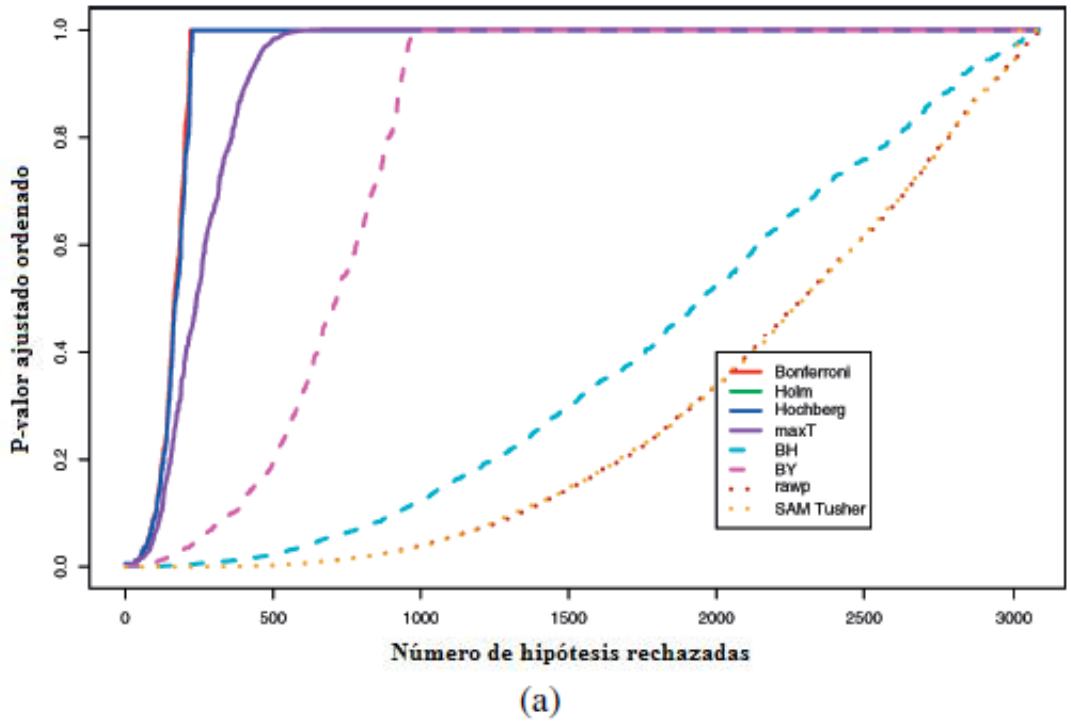
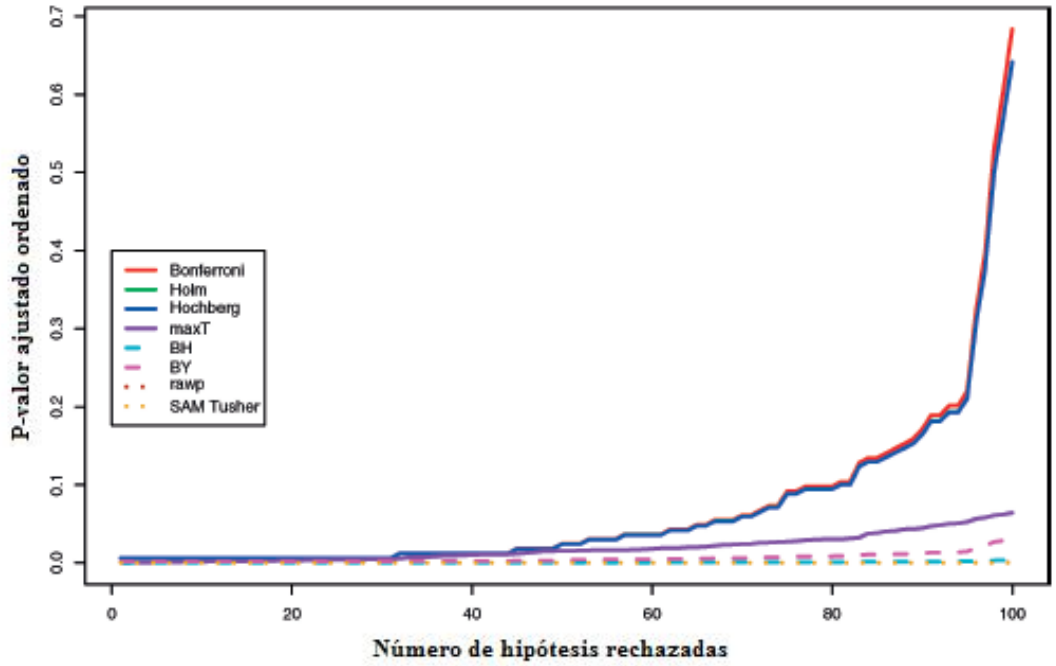


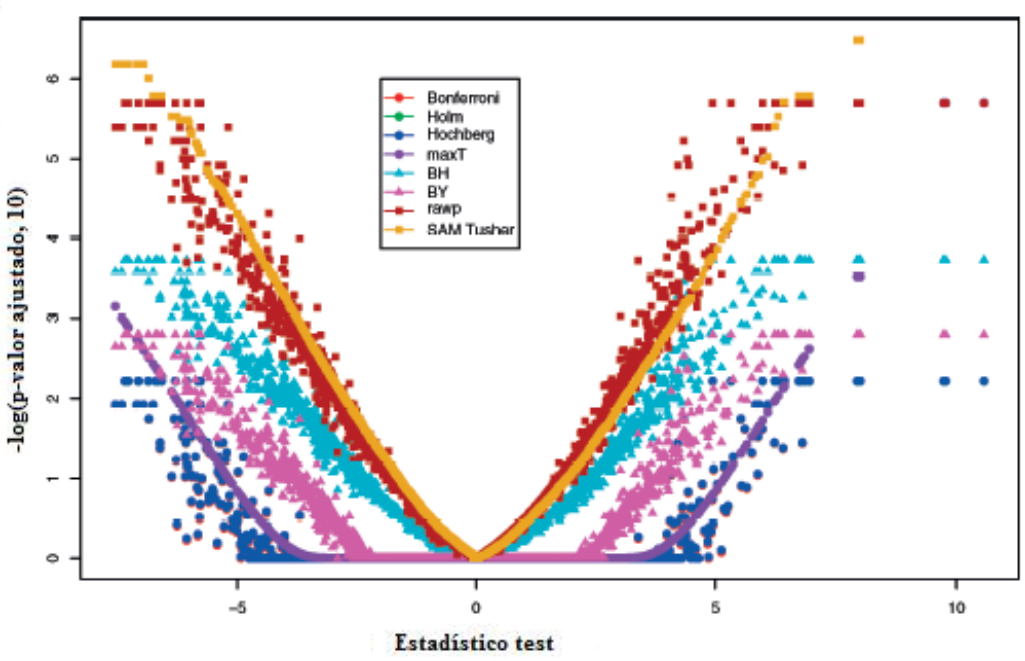
Gráfico 5. (a) Gráfico de los p-valores ajustados permutados ordenados $\tilde{p}_j^*$ frente a j. P-valores ajustados de los procedimientos que controlan la FWER, la FDR y la PCER se representan con una línea sólida, discontinua y a puntos respectivamente. Gráfico reproducido de Dudoit y otros (2003)



(b)

Gráfico 6. (b) Gráfico de los p-valores ajustados permutados ordenados $\tilde{p}_j^*$ frente a j. P-valores ajustados de los procedimientos que controlan la FWER, la FDR y la PCER se representan con una línea sólida, discontinua y a puntos respectivamente. Gráfico reproducido de Dudoit y otros (2003)

En el siguiente gráfico 7.(c), el volcano plot, se puede ver como hay varios genes que podrían estar diferencialmente expresados ya que hay varios con p-valores ajustados muy pequeños, que se encuentran con valores de la coordenada más grandes.



(c)

Gráfico 7. (c) Gráfico de p-valores ajustados $-\log_{10} \tilde{p}_j^*$ frente al test estadístico t_j. Gráfico reproducido de Dudoit y otros (2003)

Si analizamos los tres gráficos anteriores se puede ver como los procedimientos de Bonferroni, Holm y Hochberg obtienen comportamientos muy parecidos, igual que los del p-valor no ajustado (rawp) y SAM Tusher.

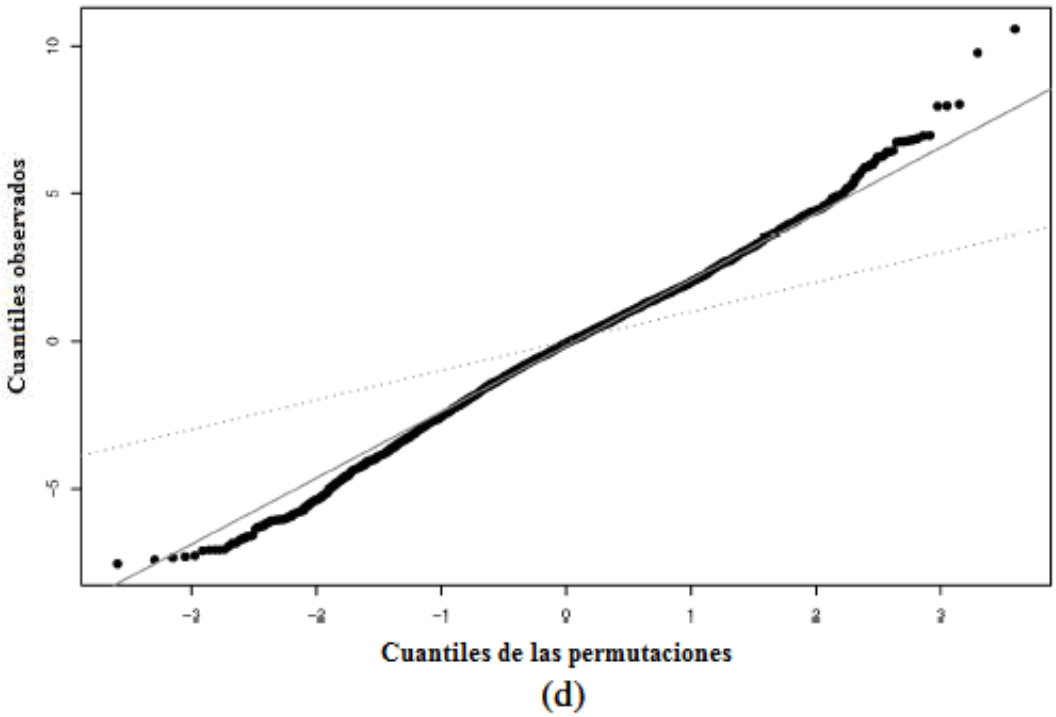


Gráfico 8. (d) Gráfico cuantil-cuantil del estadístico t; con una línea continua se representa la línea identidad y con una discontinua, la línea que pasa por el primer y el tercer cuantil. Gráfico reproducido de Dudoit y otros (2003)

En el plot cuantil-cuantil, gráfico 8.(d), se observa que los puntos no aparecen representados sobre la diagonal y que no es posible identificar grupos de genes con comportamientos claramente diferentes, como ocurría en el ejemplo anterior.

6.2 DATOS SIMULADOS

Se van a analizar los datos simulados realizados por Dudoit y otros (2003) para así poder evaluar las propiedades de las diferentes definiciones de error tipo I y de la potencia para cada uno de los procedimientos descritos en la tabla 4.

Nombre	**Descripción**
Bonf t	Procedimiento de Bonferroni, rechaza H_j si $\tilde{p}_j \leq \alpha$ (ecuación (1)), p_j calculado de la distribución t con n_1+n_2-2 grados de libertad
Bonf perm	Procedimiento de Bonferroni, rechaza H_j si $\tilde{p}_j^* \leq \alpha$ (ecuación (1)), p_j^* calculado por permutación
Holm t	Procedimiento de Holm, rechaza H_j si $\tilde{p}_{r_j} \leq \alpha$ (ecuación (5)), p_j calculado de la distribución t con n_1+n_2-2 grados de libertad
Holm perm	Procedimiento de Holm, rechaza H_j si $\tilde{p}_{r_j}^* \leq \alpha$ (ecuación (5)), p_j^* calculado por permutación

Hoch t	Procedimiento de Hochberg, rechaza H_j si $\tilde{p}_j \leq \alpha$ (ecuación (9)), p_j calculado de la distribución t con n_1+n_2-2 grados de libertad
Hoch perm	Procedimiento de Hochberg, rechaza H_j si $\tilde{p}^*_{r_j} \leq \alpha$ (ecuación (9)), p^*_j calculado por permutación
máx T ss	Procedimiento single-step de max T, rechaza H_j si $\tilde{p}^*_j \leq \alpha$ (ecuación (4))
máx T sd	Procedimiento step-down de max T, rechaza H_j si $\tilde{p}^*_{r_j} \leq \alpha$ (ecuación (4))
FDR BH t	Procedimiento de Benjamini y Hochberg (1995), rechaza H_{rj} si $\tilde{p}_{r_j} \leq \alpha$ (ecuación (10)), p_j calculado de la distribución t con n_1+n_2-2 grados de libertad
FDR BH perm	Procedimiento de Benjamini y Hochberg (1995), rechaza H_{rj} si $\tilde{p}^*_{r_j} \leq \alpha$ (ecuación (10)), p^*_j calculado por permutación
FDR BY t	Procedimiento Benjamini y Yekutieli (2001), rechaza H_{rj} si $\tilde{p}_{r_j} \leq \alpha$ (ecuación (11)), p_j calculado de la distribución t con n_1+n_2-2 grados de libertad
FDR BY perm	Procedimiento Benjamini y Yekutieli (2001), rechaza H_{rj} si $\tilde{p}^*_{r_j} \leq \alpha$ (ecuación (11)), p^*_j calculado por permutación
PCER ss t	Rechaza H_j si $p_j \leq \alpha$, p_j calculado de la distribución t con n_1+n_2-2 grados de libertad
PCER ss perm	Rechaza H_j si $p^*_j \leq \alpha$, p^*_j calculado por permutación
SAM Tusher	Procedimiento SAM de Tusher, Tibshirani y Chu(2001) (punto 5.3.1), rechaza H(j) si $\tilde{p}^*_{(j)} \leq \alpha$, estimado por permutación
Golub sd	Análisis de vecindad de Golub et al(1999), versión step-down (punto 5.3.2) rechaza $H_{(j)}$ si $\tilde{p}^*_{(j)} \leq \alpha$, estimado por permutación
Golub su	Análisis de vecindad de Golub et al(1999), versión step-up (punto 5.3.2) rechaza $H_{(j)}$ si $\tilde{p}^*_{(j)} \leq \alpha$, estimado por permutación

Tabla 4. Procedimientos de contrastes múltiples aplicados al estudio de simulación. Tabla reproducida de Dudoit y otros (2003)

Mediante los siguientes puntos Dudoit y otros (2003) generaron una variable x que es la expresión génica perfil artificial y la variable respuesta binaria y, para un número m de 500 genes.

Cálculos de la probabilidad de error tipo I y de la potencia para datos simulados:

1. Para un tipo de individuo i=1,2, se generan n_i vectores de dimensión m según una distribución normal con media μ_i y matriz de covarianzas sigma. La matriz de expresión génica artificial está formada por los $n_1 + n_2$ m-vectores. m_0 filas de esta matriz se corresponden con genes no expresados diferencialmente y caracterizados por $\mu_1 = \mu_2$. Los m_1=m- m_0 genes restantes estarán sobre o infra-expresados en una cantidad variable. La estructura de correlación de la matriz se simula utilizando muestras aleatorias de un conjunto de datos reales. Los parámetros del modelo se recogen en la tabla 5.

2. Para cada uno de los m genes, se calcula el t-estadístico de dos muestras (con varianzas iguales en las respuestas de los dos grupos) comparando las medidas de la expresión génica en los dos grupos respuesta. Se aplica un procedimiento determinado de contrastes múltiples para determinar que genes son diferencialmente expresados para una preespecífica probabilidad de error de tipo I α. Un resumen de los procedimientos de contrastes múltiples usados vienen dados en la tabla 3.
3. Para cada procedimiento, se calcula el número R_b de genes declarados diferencialmente expresados, el número V_b y T_b de error tipo I y tipo II respectivamente, y la probabilidad de falso descubrimiento Q_b, donde $Q_b=V_b/R_b$ si $R_b>0$ y $Q_b = 0$ si $R_b=0$.

Repitieron los tres pasos B veces y estimaron las diferentes definiciones de error tipo I y la potencia media para cada uno de los procedimientos de la siguiente manera:

$$PCER = \frac{\sum_b \frac{V_b}{m}}{B},$$

$$FWER = \frac{\sum_b I\left(V_b \geq 1\right)}{B},$$

$$FDR = \frac{\sum_b Q_b}{B},$$

$$Potencia\_media = 1 - \frac{\sum_b \frac{T_b}{\left(m - m_0\right)}}{B}.$$

Calcularon cada uno de los 17 procedimientos para cada conjunto de datos simulados.

Los p-valores no ajustados para cada uno de los genes se calcularon de dos formas:
1. por permutación de las $n = n_1 + n_2$ respuestas
2. obtenidos de la distribución t con $n_1 + n_2 - 2$ grados de libertad.

En la tabla 5 recogieron los parámetros usados en las simulaciones:

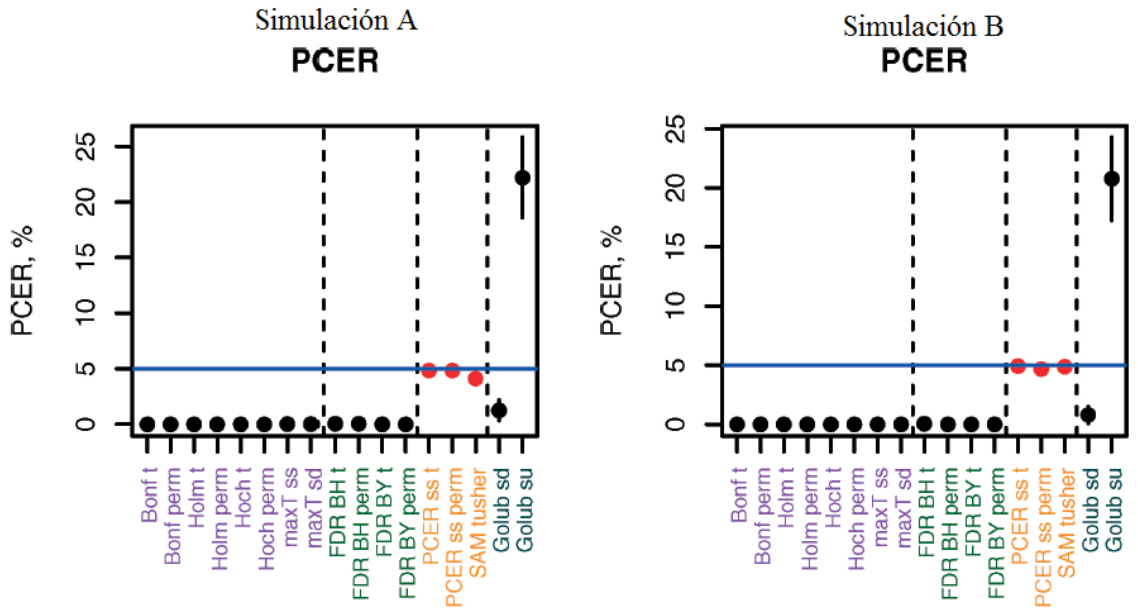
Parámetro	**Simulación A**	**Simulación B**	**Simulación C**	**Simulación D**
Número de genes, m	500	500	500	500
Vectores de las medias				
μ_1	0_m	0_m	0_m	0_m
μ_2	0_m	0_m	$[b_{m.0.1},-b_{m*0.1},0_{m-0.8}]$	$[b_{m.0.1},-b_{m*0.1},0_{m-0.8}]$
Matriz de covarianzas sigma	S_m	S_m	S_m	S_m
Tamaños de las muestras				
n_1	25	5	25	5
n_2	25	5	25	5
Número de simulaciones, B	500	500	500	500

Número de permutaciones para el SAM, B_{sam}	1000	$\binom{n_1+n_2}{n_1}$	1000	$\binom{n_1+n_2}{n_1}$
Número de permutaciones para el análisis de vecindad, B_{nbd}	1000	$\binom{n_1+n_2}{n_1}$	1000	$\binom{n_1+n_2}{n_1}$
Número de permutaciones para los p-valores no ajustados, B_{perm},	25000	$\binom{n_1+n_2}{n_1}$	25000	$\binom{n_1+n_2}{n_1}$
Probabilidad de error tipo I nominal, α (PCER; FWER, FDR)	0.05	0.05	0.05	0.05

Tabla 5. Parámetros de simulación. 0_n es un n-vector igual a 0, b_n es un n-vector 1.5*(1,…,n)/n , S_m es una matriz de covarianzas m x n para un subconjunto aleatorio de m genes. Tabla reproducida de Dudoit y otros (2003)

Para los distintos datos simulados se han realizado gráficos con las probabilidades de error tipo I y la potencia para los distintos procedimientos de contrastes múltiples.
Los p-valores ajustados se han calculado para cada método como se ha descrito en el punto 5, bien sea mediante la distribución t como con las permutaciones.
Se fija un nivel α a partir del cual, si los p-valores ajustados son más pequeños que él, se rechazará la hipótesis nula correspondiente.

En los gráficos, los procedimientos que controlan la FWER están pintados de morado, los que controlan la FDR en verde y los que lo hacen de la PCER en naranja.
Para una definición de probabilidad de error tipo I, hay unos procedimientos que deberían controlarla, estos procedimientos están representados con puntos rojos; para cada definición de probabilidad de error tipo I, los puntos rojos, los procedimientos que la controlan, irán cambiando.



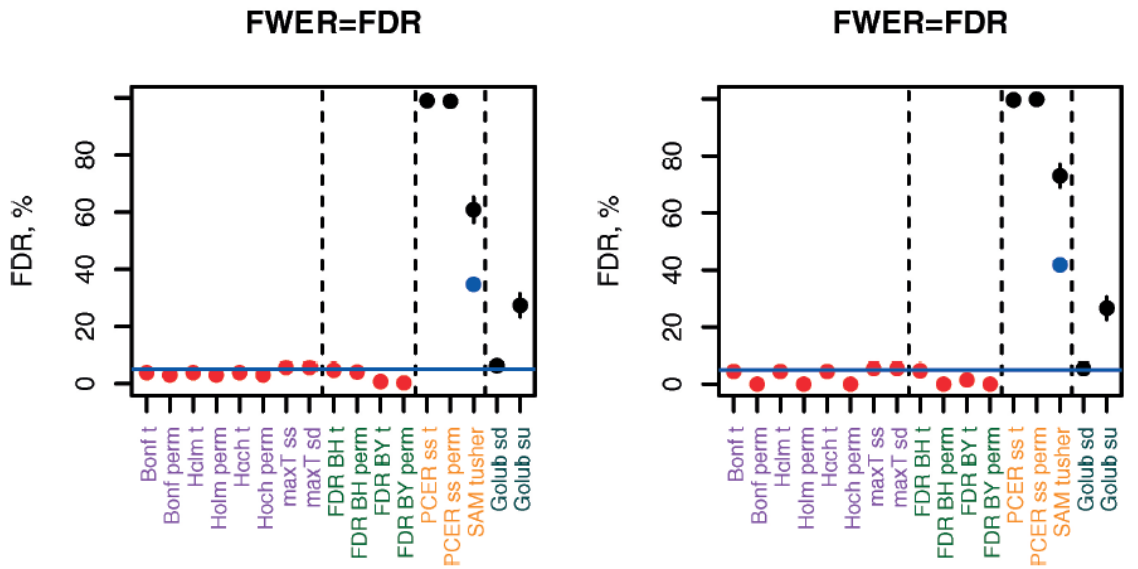
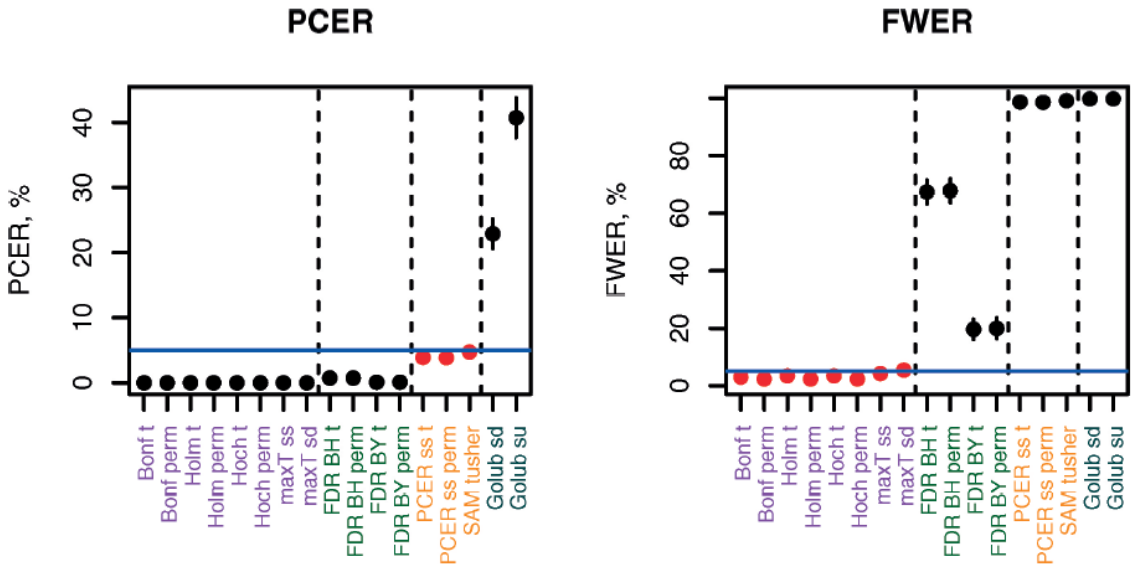


Figura 3. Simulaciones A y B – Completa nula. PCER, FWER y FDR para diferentes procedimientos de contrastes múltiples en simulación A (izquierda) y simulación B (derecha). Los gráficos de arriba representan la PCER = sum(b) Rb/mB y los errores estándar simulados (2SE), los gráficos de abajo representan la FWER=FDR= sum(b) I(Rb≥1)/B y los errores estándar simulados (2SE). Para cada definición de probabilidad de error tipo I, el procedimiento que controla dicha probabilidad está pintado de rojo. La línea azul corresponde a la probabilidad de error tipo I nominal de α=5%. En el gráfico de la FDR, la media simulada del nominal SAM FDR está pintada de azul. Detalles de cada procedimiento de contraste múltiple y de los parámetros de simulación están en las tablas 2 y 3 respectivamente. Figura reproducida de Dudoit y otros (2003)

Bajo la hipótesis nula completa los procedimientos que controlan la FWER, para la simulación A, la controlan bien a un nivel alpha=0.05, mientras que en la simulación B, los que se basan en permutaciones, lo controlan a un nivel más pequeño.
Por su parte los procedimientos que controlan la FDR, que en este caso es igual a la FWER, el de Benjamini y Hochberg controla a un nivel alpha=0.05 la probabilidad de error, mientras que el de Benjamini y Yekutieli la controla a un nivel menor.
Por otra parte los procedimientos que controlan la PCER, tanto para la simulación A como la B, la controlan bien a un nivel alpha=0.05.



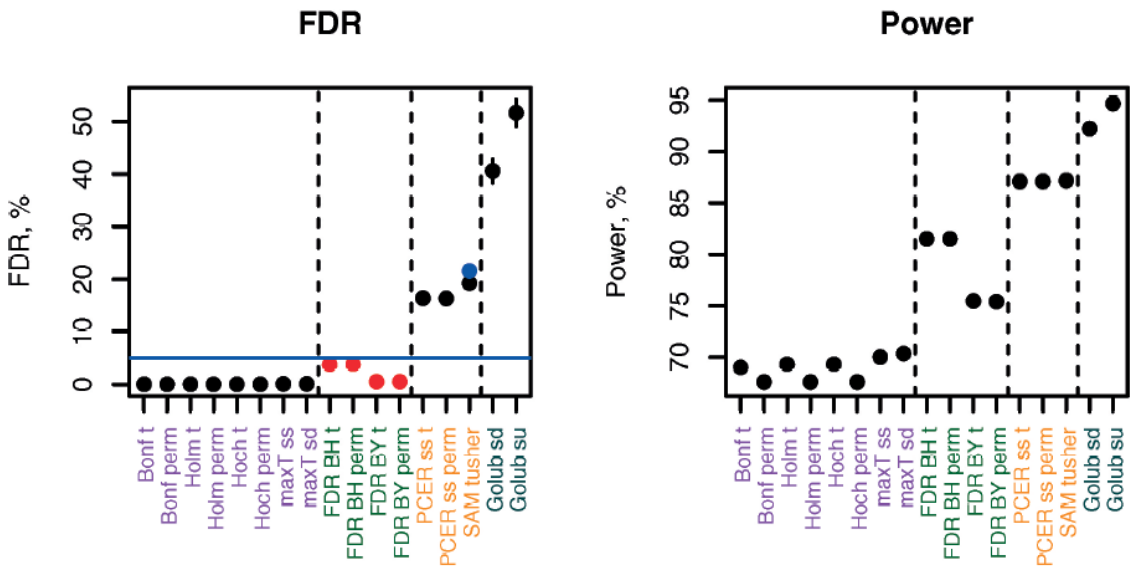
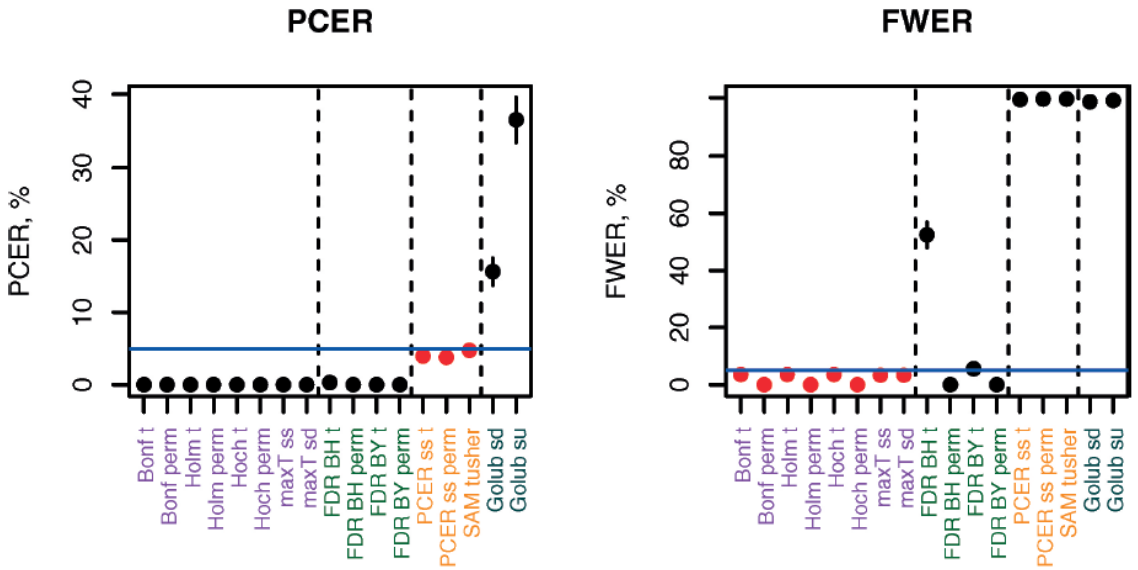


Figura 4. Simulación C – 20% de falsas hipótesis nulas (m0/m=0.8). PCER, FWER y FDR para diferentes procedimientos de contrastes múltiples en simulación A (izquierda) y simulación B (derecha). Los gráficos de arriba representan la PCER = sum(b) Rb/mB y los errores estándar simulados (2SE), y la FWER = sum(b) I(Rb≥1)/B y los errores estándar simulados (2SE). Los gráficos de abajo representan la FDR = SUM(B) Qb / B y los errores estándar simulados (2SE), y la potencia media = 1- sum(b) Tb / (m-m0)B y los errores estándar simulados (2SE). Para cada definición de probabilidad de error tipo I, el procedimiento que controla dicha probabilidad está pintado de rojo. La línea azul corresponde a la probabilidad de error tipo I nominal de α=5%. En el gráfico de la FDR, la media simulada del nominal SAM FDR está pintada de azul. Detalles de cada procedimiento de contraste múltiple y de los parámetros de simulación están en las tablas 2 y 3 respectivamente. Figura reproducida de Dudoit y otros (2003)



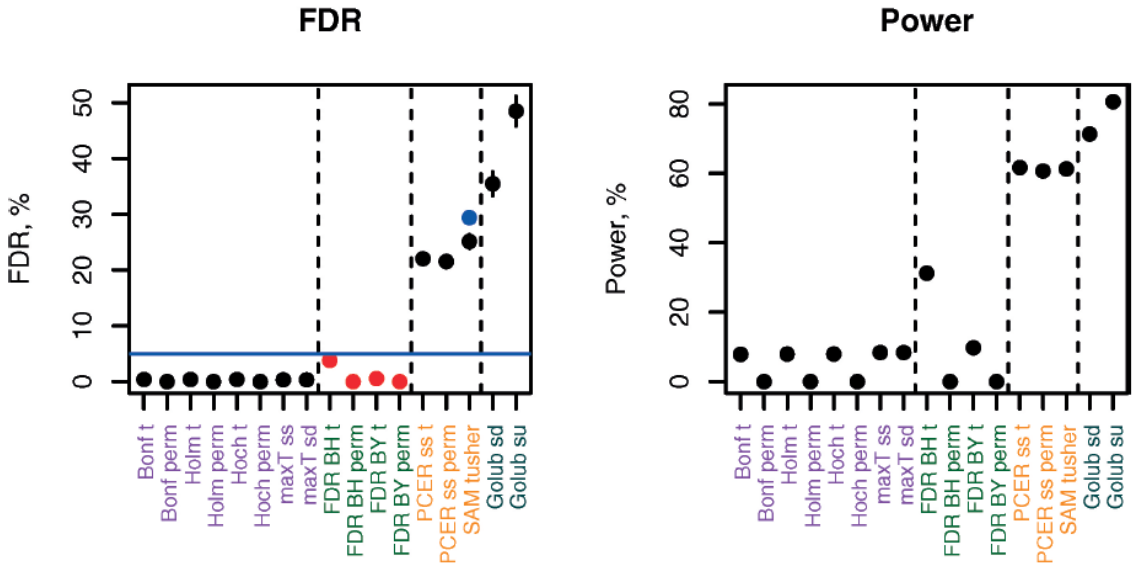


Figura 5. Simulación D – 20% de falsas hipótesis nulas (m0/m=0.8). PCER, FWER y FDR para diferentes procedimientos de contrastes múltiples en simulación A (izquierda) y simulación B (derecha). Los gráficos de arriba representan la PCER = sum(b) Rb/mB y los errores estándar simulados (2SE), y la FWER = sum(b) I(Rb≥1)/B y los errores estándar simulados (2SE). Los gráficos de abajo representan la FDR = SUM(B) Qb / B y los errores estándar simulados (2SE), y la potencia media = 1- sum(b) Tb / (m-m0)B y los errores estándar simulados (2SE). Para cada definición de probabilidad de error tipo I, el procedimiento que controla dicha probabilidad está pintado de rojo. La línea azul corresponde a la probabilidad de error tipo I nominal de α=5%. En el gráfico de la FDR, la media simulada del nominal SAM FDR está pintada de azul. Detalles de cada procedimiento de contraste múltiple y de los parámetros de simulación están en las tablas 2 y 3 respectivamente. Tabla reproducida de Dudoit y otros (2003)

En las simulaciones C y D, en las que se tiene un 20% de falsas hipótesis nulas, es fácil darse cuenta de que los procedimientos que controlan la FWER son más conservadores que los que controlan la FDR, y a la vez, estos son más conservadores que los que controlan la PCER.

Los métodos que controlan la FWER son muy conservadores, tienden a controlar las respectivas probabilidades de error a un nivel menor que 0.05, lo que repercute en la potencia, de tal manera que ésta es más pequeña.
En la potencia se puede apreciar como se obtienen valores más elevados para la simulación D, es decir, para tamaños de muestra mayores.
Por otra parte se puede ver también como el procedimiento máx T, que tiene en cuenta la estructura de dependencia de los test estadísticos, obtiene cierta mejora en la potencia que el resto de los procedimientos que controlan la FWER.
Por último, se puede observar como los procedimientos en los que los p-valores no ajustados se estimaron mediante permutación tienen valores más bajos en la potencia que los estimados mediante la distribución t, salvo en el máx T.

En los procedimientos que controlan la FDR, se ve como los métodos de Benjamini y Yekutieli son más conservadores que los de Benjamini y Hochberg, los primeros controlan la FDR a un nivel menor que 0.05 pero a cambio de tener valores más bajos en la potencia. Hay que tener en cuenta que el procedimiento de Benjamini y Hochberg, que se basa en la independencia de los test estadísticos, en estos datos simulados no obtiene tan malos resultados.

Este grupo de métodos se ve también afectado si los p-valores no ajustados se estiman mediante permutación o no, siendo con lo primero con lo que se consiguen valores más bajos de la potencia (siendo mayor la diferencia en el procedimiento de Benjamini y Yekutieli).

Por último los procedimientos que controlan la PCER obtienen los valores más elevados en la potencia sin perder el control fuerte de esta probabilidad de error tipo I a un nivel 0.05. A diferencia de la mayor parte de los métodos anteriores, estos no se ven afectados por si los p-valores no ajustados se han estimado por permutación o no. Son los procedimientos que tienen valores de la potencia más altos.

5. APLICACIÓN DE LA METODOLOGÍA

Los análisis presentados en este trabajo se han realizado usando R y las utilidades disponibles en el marco del proyecto Bioconductor (http://www.bioconductor.org). Bioconductor es un proyecto de desarrollo de software abierto para el análisis e interpretación de datos genómicos basado en R. Actualmente, prácticamente la mayoría de los métodos disponibles en análisis de microarrays tiene su propio paquete en este entorno. Concretamente se han utilizado tres paquetes,

- affy: que contiene funciones para la lectura, pre-procesado y descripción de datos de microarrays de Affymetrix.
- multtest: que contiene una colección de funciones para contrastes de hipótesis múltiples utilizado para identificar genes diferencialmente expresados en experimentos de microarrays.
- samr: en el que se implementa el método SAM de Tusher.

El código utilizado se adjunta en el anexo.

5.1 DATOS REALES

En los conjuntos de datos que se analizan en este trabajo se utilizan microarrays de Affymetrix, llamados también GeneChip. Con este sistema cada una de las muestras biológicas se corresponde con un dispositivo o chip independiente, y cada gen estará representado por un conjunto de secuencias de ADN que constituyen una o varias filas en la matriz de expresión, también llamadas probesets.

Se analizan 5 matrices de expresión. Cada una de estas matrices se corresponde con un experimento cuyo objetivo es comparar los niveles de expresión en dos tipos de ratón: ratón *knockout*, al que se le ha eliminado un gen concreto, respecto de un ratón control. Los genes eliminados en cada caso y su correspondencia con filas en las matrices de expresión, se presentan en la Tabla 6. En cada experimento, se dispone de 3 ratones tratamiento y de 3 ratones control. Las muestras biológicas fueron procesadas utilizando el microarray GeneChip ® Mouse Genome 430 2.0, que incluye 45101 probesets. Estos datos han sido cedidos por el laboratorio de Bioinformática y Genómica Funcional del Instituto de investigación del Cáncer de Salamanca.

Gen	APOE	ENG	IRS2	NRAS	SCD1
Probeset	1432466_a_at	1417271_a_at	1443969_at	160925_at 94362_at	94056_at 94057_g_at

Tabla 6: genes suprimidos en cada conjunto de datos y probeset equivalente en los microarrays de Affymetrix.

Los genes diferencialmente expresados se identifican utilizando el estadístico t-Welch para dos muestras independientes en cada uno de los genes de manera independiente. Se consideran tanto el p-valor no ajustado (rawp) como los p-valores ajustados utilizando los siguientes procedimientos de contrastes múltiples,

- Control de la FWER: Bonferroni, Holm, Hochberg y maxT con 20 permutaciones (todas las posibles)

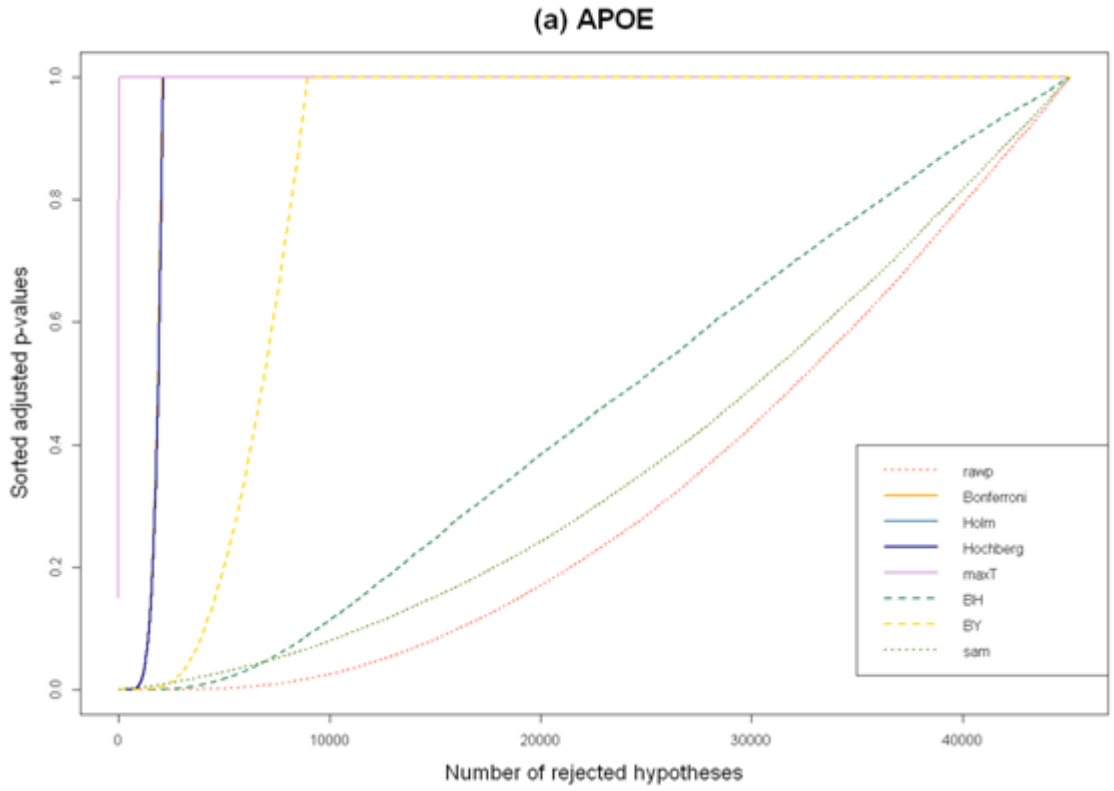
- Control de la FDR: Benjamini & Hochberg y Benjamini & Yekutieli

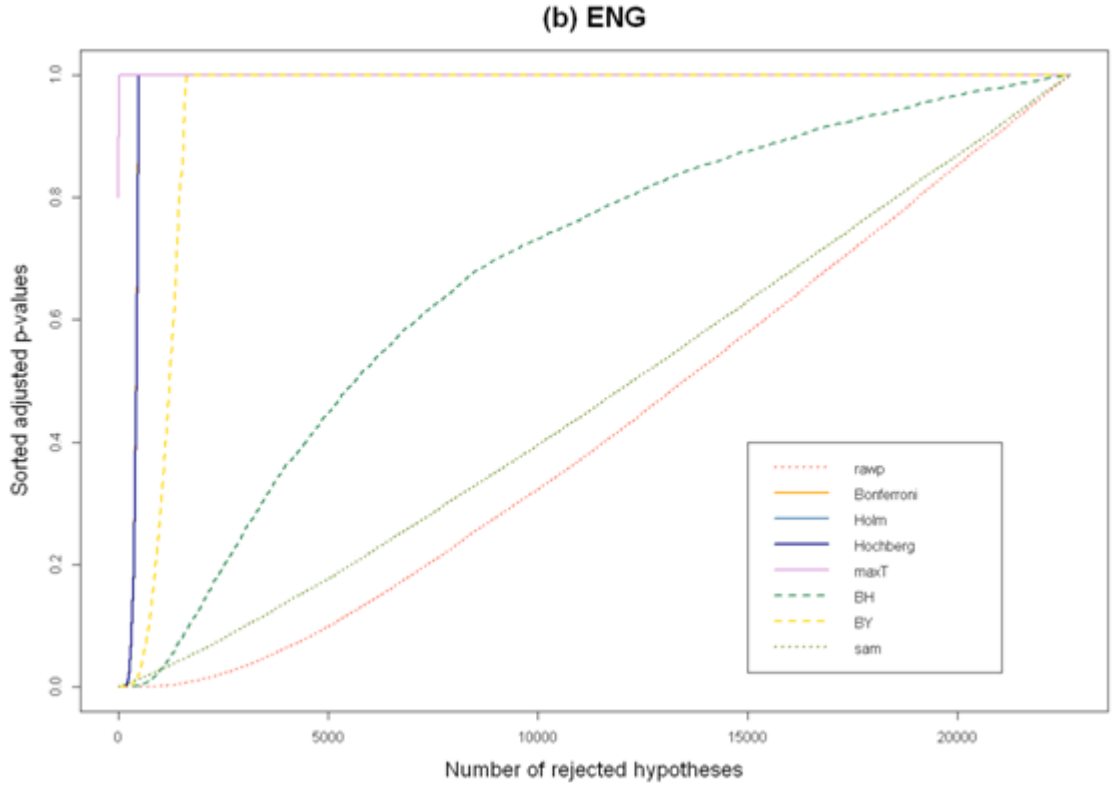
- Control de la PCER: SAM con 100 permutaciones

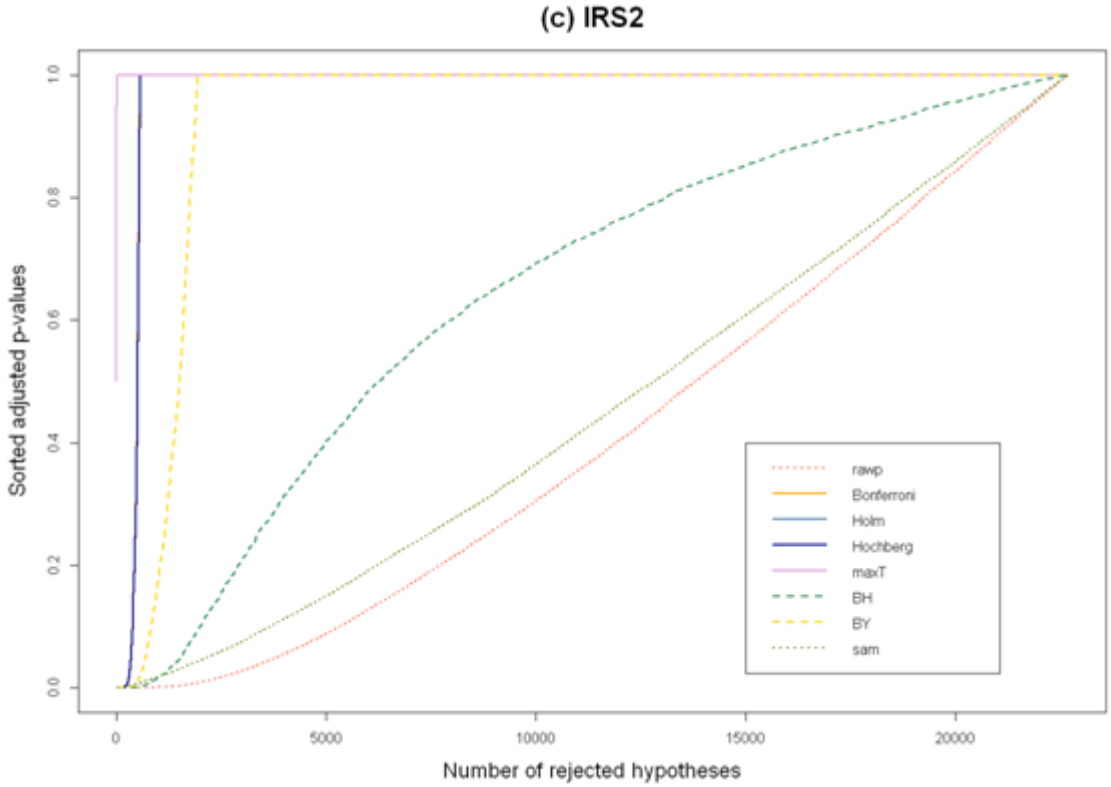
En la Figura 6 se representa el número de hipótesis rechazas frente al p-valor ajustado ordenado para cada uno de los conjuntos de datos analizados. Para todos los conjuntos de datos la curva para los distintos procedimientos de comparaciones múltiples evaluados es similar: los p-valores son más pequeños para los procedimientos que controlan la PCER (SAM y rawp) y más grandes para los p-valores ajustados a partir de procedimientos basados en el control de la FWER (Bonferroni, Holm, Hochberg y maxT). En la Tabla 7 se muestra el número de genes significativos a nivel 0.05 para cada uno de los procedimientos de comparaciones múltiples comparados.

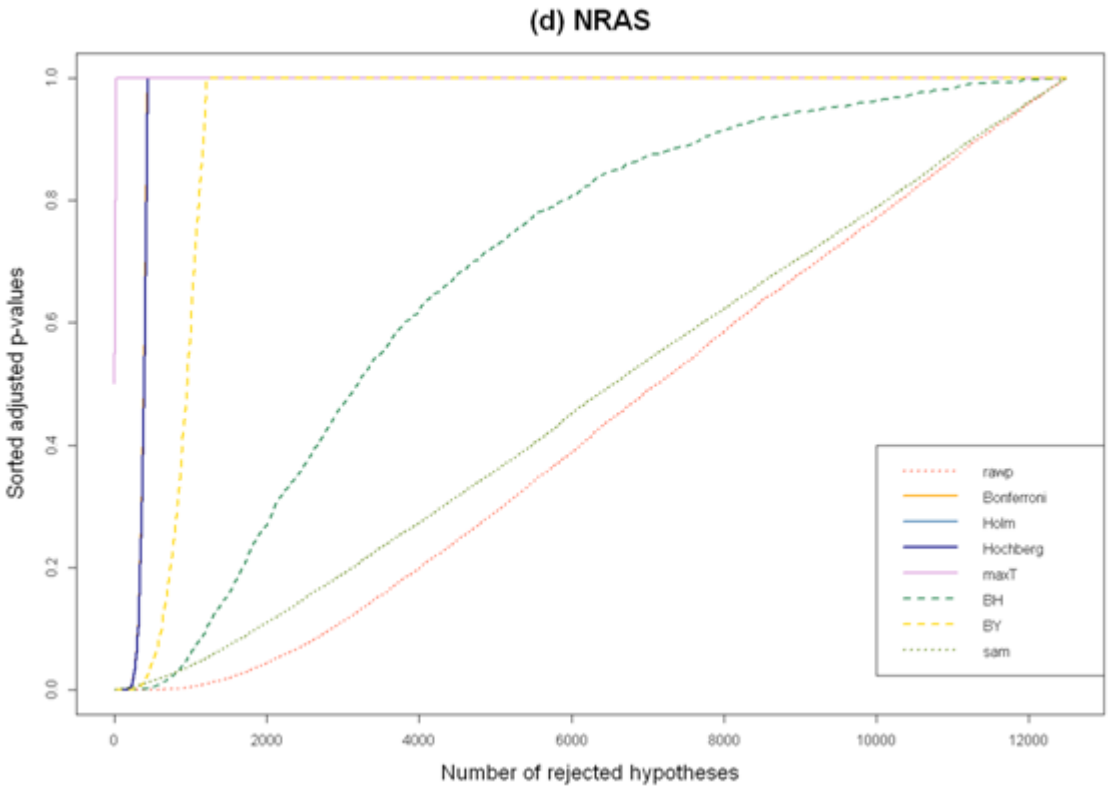
	rawp	Bonferroni	Holm	Hochberg	maxT	BH	BY	SAM
APOE	12538	1302	1303	1303	0	7168	3419	7402
ENG	3580	285	286	286	0	1249	595	1684
IRS2	3846	348	349	349	0	1545	702	2178
NRAS	2101	281	283	283	0	956	517	1188
SCD1	4790	804	812	812	0	3360	1807	2911

Tabla 7. Número de genes significativos a nivel 0.05 para cada uno de los procedimientos de comparaciones múltiples evaluados.


(a) APOE
Sorted adjusted p-values
0.0
0.2
0.4
0.6
0.8
1.0
0
10000
20000
30000
40000
Number of rejected hypotheses
rawp
Bonferroni
Holm
Hochberg
maxT
BH
BY
sam


(b) ENG
Sorted adjusted p-values
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
Number of rejected hypotheses
rawp
Bonferroni
Holm
Hochberg
maxT
BH
BY
sam


(c) IRS2
Sorted adjusted p-values
Number of rejected hypotheses
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
rawp
Bonferroni
Holm
Hochberg
maxT
BH
BY
sam


(d) NRAS
Sorted adjusted p-values
Number of rejected hypotheses
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000
rawp
Bonferroni
Holm
Hochberg
maxT
BH
BY
sam

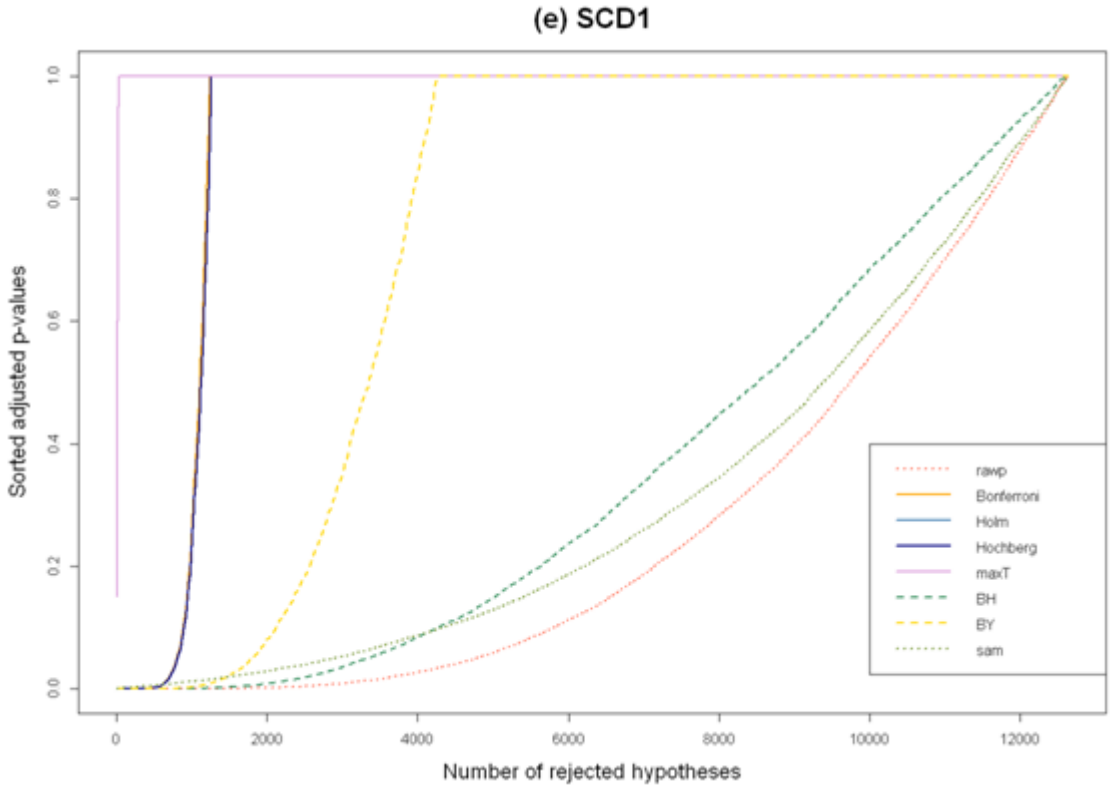


Figura 6. Se representa el número de hipótesis rechazadas frente a p-valor ordenado ajustado para cada uno de los conjuntos de datos utilizados (a-e).

Una ventaja de estos conjuntos de datos es que conocemos qué gen debería estar diferencialmente expresado, en este caso infra-expresado. En la Figura 7 se representan los niveles de expresión de cada gen en cada conjunto de datos. Un resultado interesante en este tipo de ejemplo, es la posición que ocupa cada gen en la lista de p-valores ajustados ordenados. Aunque de manera ideal estos genes deberían ocupar la primera posición y ser los únicos que presentan un comportamiento diferencial, en datos reales no ocurre necesariamente así, debido a las interrelaciones entre genes y posiblemente debido al azar por el gran número de genes estudiados. Como se observa en la Tabla 8, dos de los métodos, maxT y SAM, basados ambos en permutaciones, no conservan la posición de p-valor ajustado respecto de los rawp. En general, es SAM el que proporciona posiciones más altas en todos los casos.

Gen	**Probeset**	**Posición en la lista de genes diferencialmente expresados**							
		rawp	***Bonferroni***	***Holm***	***Hochberg***	***maxT***	***BH***	***BY***	***sam***
APOE	1432466_a_at	93.5	93.5	93.5	93.5	22562	93.5	93.5	22
ENG	1417271_a_at	85	85	85	85	11349.5	85	85	2
IRS2	1443969_at	174	174	174	174	11348.5	174	174	19
NRAS	160925_at	110	110	110	110	6254	109.5	109.5	54
	94362_at	21	21	21	21	1.5	21	21	1
SCD1	94056_at	65	65	65	65	6335.5	65	65	2
	94057_g_at	65	65	65	65	9	65	65	1

Tabla 8. Posición de cada gen en la lista ordenada de p-valores ajustados utilizando cada uno de los procedimientos.

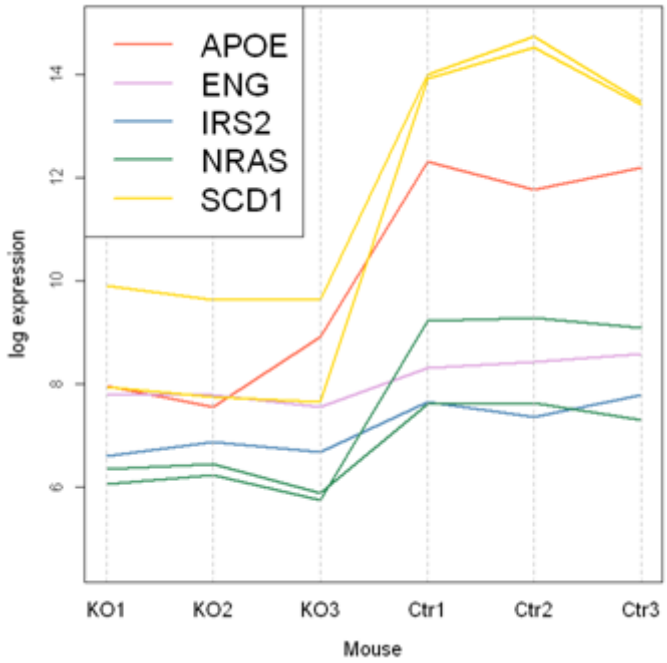


Figura 7. Niveles de expresión de cada gen en cada conjunto de datos. KO: KnockOut; Ctr: Control

5.2 SIMULACIONES

Se considera un experimento con dos grupos de tratamiento, i=1,2. Para cada grupo de tratamiento, se generan 10 vectores independientes de tamaño m con distribución normal de media μ_i y matriz de covarianzas S_m, estimada a partir de un conjunto de m genes seleccionados de manera aleatoria del conjunto de datos perteneciente al experimento en el que se suprime el gen APOE. El 5% de los genes estarán diferencialmente expresados en una cantidad de 2 desviaciones típicas del gen correspondiente.

Los genes diferencialmente expresados se identifican utilizando en cada gen dos estadísticos, el estadístico t-Student y el estadístico de Wilcoxon, ambos para dos muestras independientes. Se consideran tanto el p-valor no ajustado (rawp) como los p-valores ajustados utilizando los procedimientos de contrastes múltiples,

- Control de la FWER: Bonferroni, Holm, Hochberg y maxT con 1000 permutaciones
- Control de la FDR: Benjamini & Hochberg y Benjamini & Yekutieli
- Control de la PCER: SAM con 1000 permutaciones

Se simulan 100 conjuntos de datos en tres escenarios cuyos parámetros se resumen en la Tabla 9. Para cada procedimiento, las tasas de error tipo I y la potencia se estiman utilizando las mismas expresiones que las indicadas en Dudoit y otros.

	Simulación 1	Simulación 2	Simulación 3
Número de genes, m	100	500	1000
Vector de medias			
μ_1	0_m	0_m	0_m
μ_2	$[2_{m1}\cdot\sigma_{m1},0_{m0}]$	$[2_{m1}\cdot\sigma_{m1},0_{m0}]$	$[2_{m1}\cdot\sigma_{m1},0_{m0}]$
Número de genes diferencialmente expresados, m1	5	25	50

Tabla 9. Parámetros de simulación. 0_m es el vector de ceros de tamaño m, σ_{m1} es el m-vector de desviaciones típicas.

En las siguientes figuras se comparan las estimaciones de la PCER, FWER, FDR y potencia para cada simulación y en cada uno de los procedimientos evaluados con los dos estadísticos.

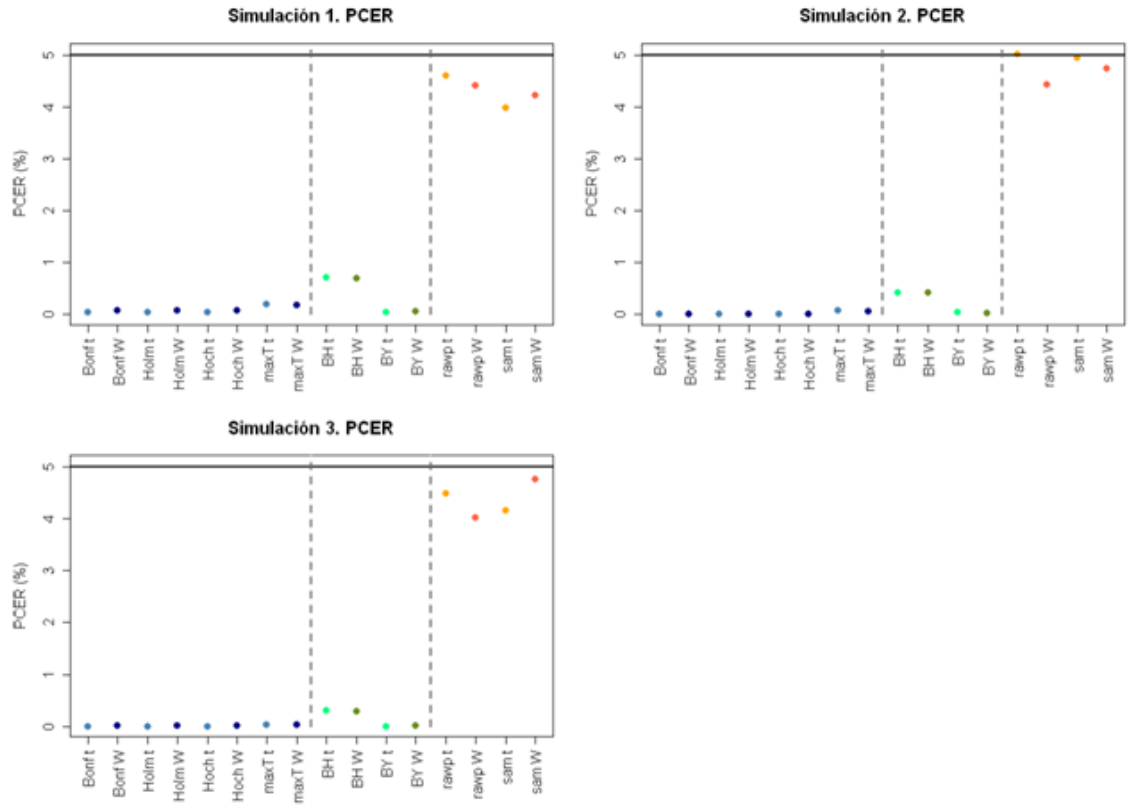


Figura 8.1. PCER (%) para cada uno de los procedimientos y cada simulación. La línea horizontal corresponde a la probabilidad de error tipo I nominal , $\alpha = 5\%$. En azul se representan los procedimientos que tratan de controlar la FWER, más oscuro cuando el estadístico utilizando es el de Wilcoxon. En verde los procedimientos basados en el control de la FDR, un tono más claro cuando el estadístico utilizado es el t-test. En naranja se representan las estimaciones obtenidas con el t-test y utilizando el procedimiento SAM, y en rojo las estimaciones equivalentes utilizando el estadístico de Wilcoxon.

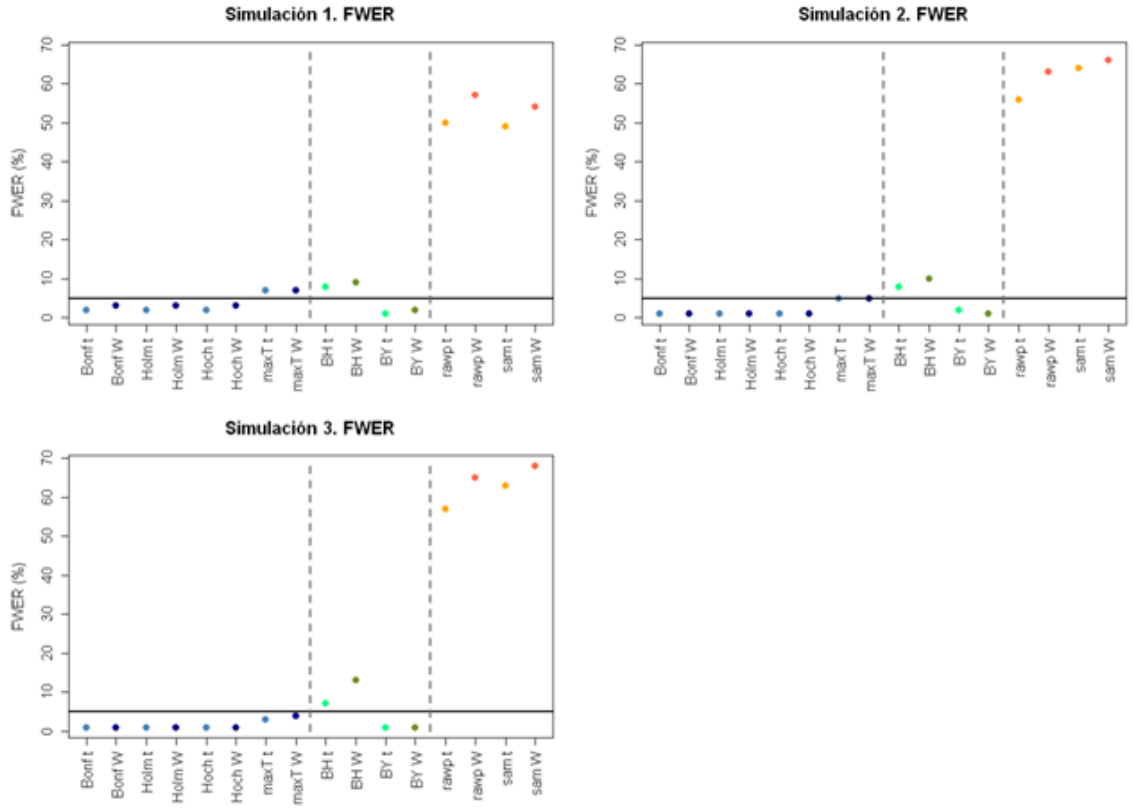


Figura 8.2. FWER (%) para cada uno de los procedimientos y cada simulación. La línea horizontal corresponde a la probabilidad de error tipo I nominal , $\alpha = 5\%$. En azul se representan los procedimientos que tratan de controlar la FWER, más oscuro cuando el estadístico utilizando es el de Wilcoxon. En verde los procedimientos basados en el control de la FDR, un tono más claro cuando el estadístico utilizado es el t-test. En naranja se representan las estimaciones obtenidas con el t-test y utilizando el procedimiento SAM, y en rojo las estimaciones equivalentes utilizando el estadístico de Wilcoxon.

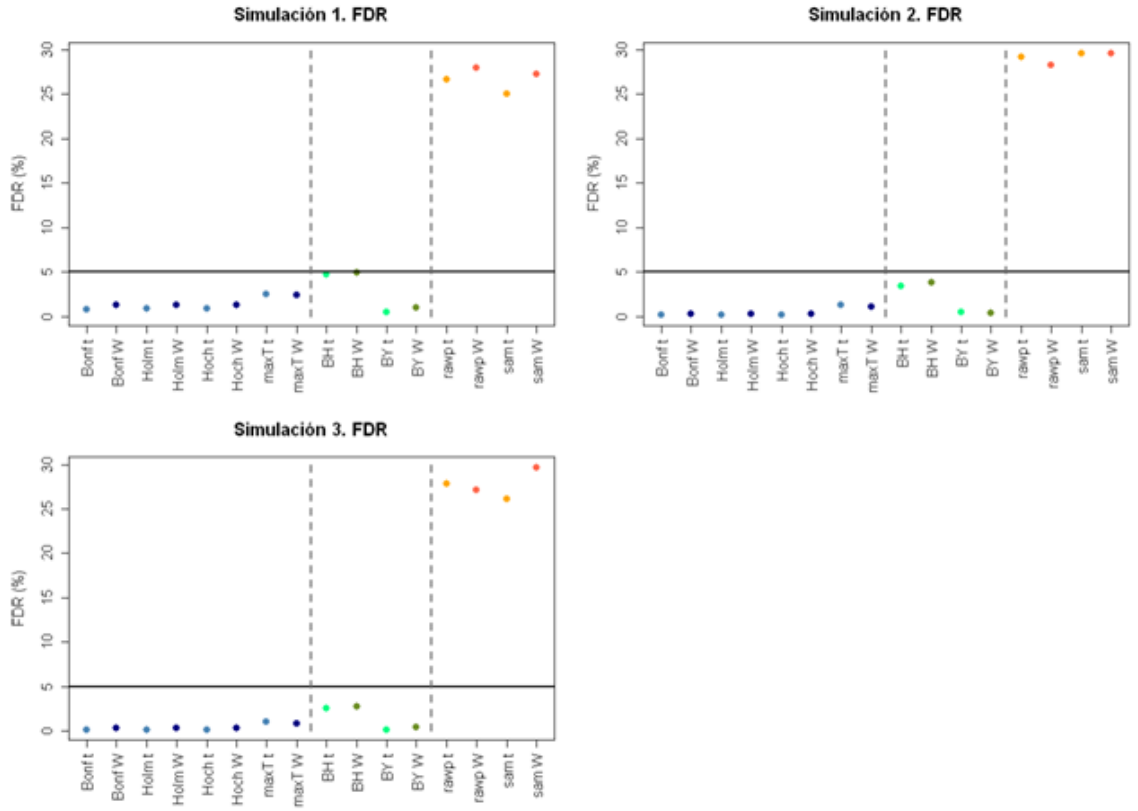


Figura 8.3. PCER (%) para cada uno de los procedimientos y cada simulación. La línea horizontal corresponde a la probabilidad de error tipo I nominal , $\alpha = 5\%$. En azul se representan los procedimientos que tratan de controlar la FWER, más oscuro cuando el estadístico utilizando es el de Wilcoxon. En verde los procedimientos basados en el control de la FDR, un tono más claro cuando el estadístico utilizado es el t-test. En naranja se representan las estimaciones obtenidas con el t-test y utilizando el procedimiento SAM, y en rojo las estimaciones equivalentes utilizando el estadístico de Wilcoxon.

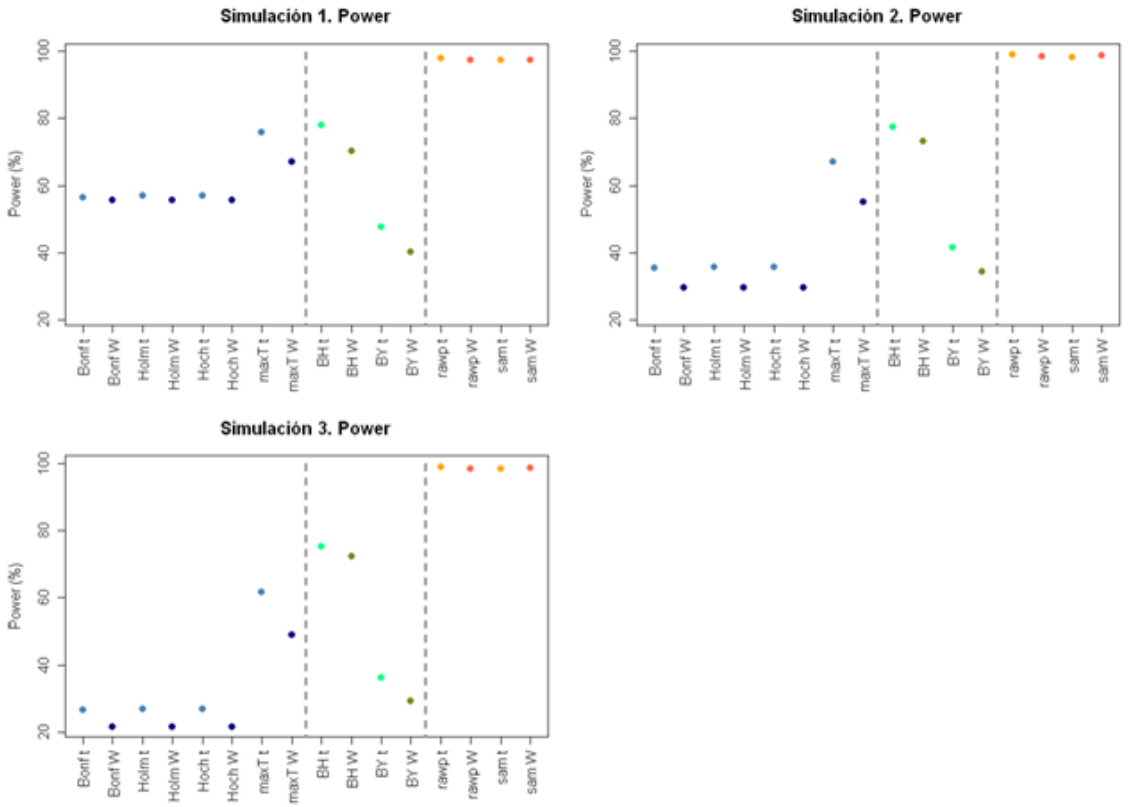


Figura 8.4. Potencia (%) para cada uno de los procedimientos y cada simulación. En azul se representan los procedimientos que tratan de controlar la FWER, más oscuro cuando el estadístico utilizando es el de Wilcoxon. En verde los procedimientos basados en el control de la FDR, un tono más claro cuando el estadístico utilizado es el t-test. En naranja se representan las estimaciones obtenidas con el t-test y utilizando el procedimiento SAM, y en rojo las estimaciones equivalentes utilizando el estadístico de Wilcoxon.

PCER (Figura 8.1), FWER (Figura 8.2) y FDR (Figura 8.3) tienen un comportamiento similar al de las simulaciones del trabajo de Dudoit y otros, proporcionando estimaciones más cercanas al valor nominal los procedimientos basados en la respectiva definición de probabilidad de error tipo I.

En cuanto a la potencia (Figura 8.4), en los procedimientos que controlan la PCER no hay diferencias entre usar el estadístico t-Student y el estadístico de Wilcoxon, si bien ofrecen estimaciones de la potencia cercanas al 100% en todos los casos. Este mismo resultado se observa para los métodos de Bonferroni, Holm y Hochberg, pero sólo en la simulación 1, escenario que se corresponde con la matriz de expresión más pequeña (m=100 genes). En el resto, y tal y como cabía esperar, utilizar el estadístico t-Student proporciona una potencia mayor. En estos tres procedimientos, todos ellos basados en el control de la FWER, se observa un efecto del tamaño del microarray, cuanto mayor número de genes, menor potencia.

7. CONCLUSIONES

Para conseguir un procedimiento de contrastes múltiples para análisis de datos de microarrays de ADN se necesita, antes de nada, conseguir un control adecuado de la probabilidad de error tipo I, para ello habrá procedimientos que controlen la FWER o la FDR o la PCER.

Procedimientos que controlan la FWER.
Parece que el procedimiento step-down del máx T, es el mejor adaptado a este tipo de problemas, ya que entre los que controlan de forma fuerte la probabilidad de error tipo I es el que tiene valores más elevados en la potencia. Esto es debido a que tiene en cuenta la estructura de dependencia de los test estadísticos a diferencia del resto de estos procedimientos.
Por otra parte a este procedimiento, en los ejemplos presentados, parece que no le afecta demasiado la manera de calcular los p-valores no ajustados se basan en permutaciones o en la distribución t.

Procedimientos que controlan la FDR.
En los procedimientos que controlan la FWER es muy probable que al intentar que no haya ningún falso positivo, no se encuentre ningún gen diferencialmente expresado aunque los haya. Por ello, el control de la FDR es más efectivo ya que permite encontrar el mayor número de genes diferencialmente expresados al permitir un pequeño valor de falsos positivos.
Sin embargo, los procedimientos disponibles de control de la FDR que se han estudiado no tienen en cuenta la distribución conjunta de los test estadísticos. Por lo que nuevos procedimientos basados en la FDR que si lo hagan, podrían mejorar sus resultados.

Procedimiento que controlan la PCER
Estos procedimientos obtienen mejores resultados que todos los anteriores en la potencia sin perder el control fuerte de la PCER.

En general se podría decir que para tener procedimientos adecuados de contrastes múltiples para datos provenientes de experimentos de microarrays, donde se dan grandes problemas de multiplicidad, sería necesario definir una adecuada probabilidad de error tipo I y tener un control fuerte de ella. En este tipo de experimentos es fundamental poder explicar los resultados por medio de los p-valores ajustados, que tienen en cuenta los contrastes que se llevan a cabo simultáneamente; es importante que los procedimientos reflejen las estructuras de dependencia que hay entre las medidas de expresión de diferentes genes, aprovechando las ventajas que ofrecen los procedimientos de remuestro.

8. BIBLIOGRAFÍA

La primera parte del trabajo se basa en el artículo y technical report que aparecen a continuación. Las tablas y gráficos de esta primera parte están tomados de allí:

-S. DUDOIT, J.P. SHAFFER and J.C. BOLDRICK (2003). Multiple hypothesis testing in microarray experiments. Statistical Science: Vol 18, no. 1, 71-103;

-S. DUDOIT, J.P. SHAFFER and J.C. BOLDRICK (2002). Multiple hypothesis testing in microarray experiments. Technical Report 110, Division of Biostatistics, Univ. California, Berkeley. Disponible en http://www.bepress.com/ucbbiostat/ paper110/;

Este artículo parte de la monografía siguiente en la que aparecen algunos de los procedimientos incluidos en el trabajo e introducidas las posibilidades que ofrecen las técnicas de remuestreo en el contexto de los contrastes múltiples:

-WESTFALL, P. H. and YOUNG, S. S. (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York.

Por la importancia de la metodología SAM se ha tenido en cuenta también otro artículo:

- VIRGINIA GOSS TUSHER, ROBERT TIBSHIRANI and GILBERT CHU (2001). Significance analysis of microarrays applied to the ionizing radiation response: Vol. 98, no. 9, 5116-5121.

ANEXO 1

Para obtener los resultados del primer ejemplo de datos reales de microarrays se ha ejecutado el siguiente código en R:

```
########## Cargar librerias
library(affy)
library(multtest)
library(samr)

########## Leer datos
workingDir<- "C:/TFG"
setwd(workingDir)

### APOE
celpath <- file.path(workingDir,"datos/APOE")
fns <- list.celfiles(path=celpath,full.names=TRUE)
datos.APOE<- ReadAffy(filenames=fns)
### ENG
celpath <- file.path(workingDir,"datos/ENG")
fns <- list.celfiles(path=celpath,full.names=TRUE)
datos.ENG<- ReadAffy(filenames=fns)
### IRS2
celpath <- file.path(workingDir,"datos/IRS2")
fns <- list.celfiles(path=celpath,full.names=TRUE)
datos.IRS2<- ReadAffy(filenames=fns)
### NRAS
celpath <- file.path(workingDir,"datos/NRAS")
fns <- list.celfiles(path=celpath,full.names=TRUE)
datos.NRAS<- ReadAffy(filenames=fns)
### SCD1
celpath <- file.path(workingDir,"datos/SCD1")
fns <- list.celfiles(path=celpath,full.names=TRUE)
datos.SCD1<- ReadAffy(filenames=fns)

########## Obtener la matriz de expresión
X.APOE<-rma(datos.APOE)
X.APOE<-exprs(X.APOE)
X.ENG<-rma(datos.ENG)
X.ENG<-exprs(X.ENG)
X.IRS2<-rma(datos.IRS2)
X.IRS2<-exprs(X.IRS2)
X.NRAS<-rma(datos.NRAS)
X.NRAS<-exprs(X.NRAS)
X.SCD1<-rma(datos.SCD1)
X.SCD1<-exprs(X.SCD1)

########## Contrastes
###### FUNCION que calcula los p-valores ajustados
### X: matriz de expresion
### cl: tipo de individuo
### B.maxT: numero de permutaciones para maxT
### B.sam: permutaciones para SAM
fun.TEST<-function(X,cl,B.maxT=1000,B.sam=1000) {
  T.GEN<-mt.teststat(X, cl,test="t")
```

```
  ### 'raw' p-valores
  rawp<-2 * (1 - pnorm(abs(T.GEN)))

  ### Ajuste FWER
  procs.FWER<-c("Bonferroni", "Holm", "Hochberg")
  FWER1.GEN<-mt.rawp2adjp(rawp, procs.FWER)
  pvalores.GEN<-FWER1.GEN$adjp[order(FWER1.GEN$index), ]
  # maxT
  maxT.GEN <- mt.maxT(X, cl, B = B.maxT)
  pvalores.GEN<-cbind(pvalores.GEN, maxT.GEN$adjp[order(maxT.GEN$index)])
  colnames(pvalores.GEN)[5]<-"maxT"

  ### Ajuste FDR
  procs.FDR<-c("BH","BY")
  FDR.GEN<-mt.rawp2adjp(rawp, procs.FDR)
  pvalores.GEN<-cbind(pvalores.GEN, FDR.GEN$adjp[order(FDR.GEN$index),2:3])

  ### SAM
  data<-list(x=X,y=cl+1, geneid=as.character(1:nrow(X)),
genenames=paste("g",as.character(1:nrow(X)),sep=""), logged2=TRUE)
  sam.GEN<-samr(data, resp.type="Two class unpaired", nperms=B.sam)
  psam.GEN<-samr.pvalues.from.perms(sam.GEN$tt, sam.GEN$ttstar)
  pvalores.GEN<-cbind(pvalores.GEN, psam.GEN)
  colnames(pvalores.GEN)[8]<-"sam"

  return(pvalores.GEN)
}   # FIN FUNCION

pvalores.APOE<-fun.TEST(X.APOE,c(rep(1,3),rep(0,3)),B.maxT=20)
pvalores.ENG<-fun.TEST(X.ENG,c(rep(1,3),rep(0,3)),B.maxT=20)
pvalores.IRS2<-fun.TEST(X.IRS2,c(rep(1,3),rep(0,3)),B.maxT=20)
pvalores.NRAS<-fun.TEST(X.NRAS,c(rep(1,3),rep(0,3)),B.maxT=20)
pvalores.SCD1<-fun.TEST(X.SCD1,c(rep(1,3),rep(0,3)),B.maxT=20)

##########Resultados
alpha<-0.05

######### Número de hipótesis rechazas a nivel alpha
WHICH<-rbind(colSums(mt.reject(pvalores.APOE, alpha)$which,na.rm=TRUE),
colSums(mt.reject(pvalores.ENG, alpha)$which,na.rm=TRUE), colSums(mt.reject(pvalores.IRS2,
alpha)$which,na.rm=TRUE), colSums(mt.reject(pvalores.NRAS, alpha)$which,na.rm=TRUE),
colSums(mt.reject(pvalores.SCD1, alpha)$which,na.rm=TRUE))

######### Plot pvalores ajustados ordenados vs hipótesis rechazadas
cols <- c("tomato", "orange", "steelblue", "navyblue", "plum","lightgreen","gold","olivedrab")
ltypes <- c(3,rep(1,4),rep(2,2),3)
mt.plot(pvalores.APOE,plottype = "pvsr",proc = colnames(pvalores.APOE),logscale=TRUE,leg = c(35000,
0.4),col=cols,lty = ltypes,  main="(a) APOE",lwd=2,cex.lab=1.5,cex.main=2)
mt.plot(pvalores.ENG,plottype = "pvsr",proc = colnames(pvalores.ENG),logscale=TRUE,leg = c(15000,
0.4),col=cols,lty = ltypes, main="(b) ENG",lwd=2,cex.lab=1.5,cex.main=2)
mt.plot(pvalores.IRS2,plottype = "pvsr",proc = colnames(pvalores.IRS2),logscale=TRUE,leg = c(15000,
0.4),col=cols,lty = ltypes, main="(c) IRS2",lwd=2,cex.lab=1.5,cex.main=2)
mt.plot(pvalores.NRAS,plottype = "pvsr",proc = colnames(pvalores.NRAS),logscale=TRUE,leg = c(10000,
0.4),col=cols,lty = ltypes,  main="(d) NRAS",lwd=2,cex.lab=1.5,cex.main=2)
mt.plot(pvalores.SCD1[rowSums(is.na(pvalores.SCD1))==0,],plottype = "pvsr",proc =
colnames(pvalores.SCD1),logscale=TRUE,leg = c(10000, 0.4),col=cols,lty = ltypes, main="(e)
SCD1",lwd=2,cex.lab=1.5,cex.main=2)
```

```
######### Orden de los p-valores
## Probeset de cada gen
genes<-vector("list",5)
genes[[1]]<-"1432466_a_at"
genes[[2]]<-"1417271_a_at"
genes[[3]]<-"1443969_at"
genes[[4]]<-c("94362_at","160925_at")
genes[[5]]<-c("94056_at","94057_g_at")

TBL<-
rbind(apply(pvalores.APOE,2,rank,na.last=TRUE,ties.method="average")[is.element(rownames(X.APOE),
genes[[1]]),],
      apply(pvalores.ENG,2,rank,na.last=TRUE,ties.method="average")[is.element(rownames(X.ENG),
genes[[2]]),],
      apply(pvalores.IRS2,2,rank,na.last=TRUE,ties.method="average")[is.element(rownames(X.IRS2),
genes[[3]]),],
      apply(pvalores.NRAS,2,rank,na.last=TRUE,ties.method="average")[is.element(rownames(X.NRAS),
genes[[4]]),],
      apply(pvalores.SCD1,2,rank,na.last=TRUE,ties.method="average")[is.element(rownames(X.SCD1),
genes[[5]]),])

## plot de expresiones
par(mar=c(4, 4, 1, 1))
plot(X.APOE[genes[[1]],],type="l",ylim=range(X.APOE),xaxt="n",xlab="Mouse",ylab="logexpression",mai
n="",col="tomato",lwd=2,cex.lab=1.25)
lines(X.ENG[genes[[2]],],col="plum",lwd=2)
lines(X.IRS2[genes[[3]],],col="steelblue",lwd=2)
lines(X.NRAS[genes[[4]][1],],col="seagreen",lwd=2)
lines(X.NRAS[genes[[4]][2],],col="seagreen",lwd=2)
lines(X.SCD1[genes[[5]][1],],col="gold",lwd=2)
lines(X.SCD1[genes[[5]][2],],col="gold",lwd=2)
axis(1,at=1:6,tck=TRUE,lty=2,col="grey",labels=rep("",6))
axis(1,tck=FALSE,at=1:6,labels=c(paste("KO",1:3,sep=""),paste("Ctr",1:3,sep="")),cex.axis=1.25)
legend("topleft",lwd=2,legend=c("APOE","ENG","IRS2","NRAS","SCD1"),col=c("tomato","plum","steelblu
e","seagreen","gold"),bg="white",cex=2)
```

ANEXO 2

Para obtener los resultados de las simulaciones se ha ejecutado el siguiente código en R:

```
##########Cargar librerias
library(affy)
library(MASS)
library(multtest)
library(samr)

##########Leer datos
workingDir<- "C:/TFG"
setwd(workingDir)

### APOE
celpath <- file.path(workingDir,"datos/APOE")
fns <- list.celfiles(path=celpath,full.names=TRUE)
datos.APOE<- ReadAffy(filenames=fns)
X.APOE<-rma(datos.APOE)
X.APOE<-exprs(X.APOE)


##########Función que calcula PCER, FWER, FDR y potencia
### X: matriz de expresión real para estimar la varianza
### m: numero de genes
### m1: numero de genes DE
### n1, n2: n arrays grupos 1 y 2
### B: n simulaciones
### B.maxT: n permutaciones maxT
### B.sam: n permutaciones SAM
### alpha: nivel significacion
fun.SIM<-function(X,m,m1,n1,n2,B=100,B.maxT=1000,B.sam=100,alpha=0.05) {

 ## Grupos
 Y<-c(rep(0,n1),rep(1,n2))

 ## Expresion grupo control
 mu1<-rep(0,m)

 Tb.Vb<-Tb.Qb<-Tb.Teb<-Wb.Vb<-Wb.Qb<-Wb.Teb<-NULL

 for(b in 1:B) {
   Sb<-cov(t(X[sample(1:nrow(X),size=m),]))
   mu2<-2*c(sqrt(diag(Sb)[1:m1]),rep(0,m-m1))
   Xb<-cbind(t(mvrnorm(n=n1, mu1,Sb)),t(mvrnorm(n=n2, mu2,Sb)))

   ## Estadistico t
   Tb<-sapply(1:nrow(Xb),function(i) t.test(Xb[i,]~Y)$p.value)
   procs<-c("Bonferroni", "Holm", "Hochberg","BH","BY")
   Adj.Tb<-mt.rawp2adjp(Tb, procs)
   pvalores.Tb<-Adj.Tb$adjp[order(Adj.Tb$index), ]
   maxT.Tb <- mt.maxT(Xb, Y, B = B.maxT)
   pvalores.Tb<-cbind(pvalores.Tb, maxT.Tb$adjp[order(maxT.Tb$index)])
```

```
    data<-list(x=Xb,y=Y+1,
geneid=as.character(1:nrow(Xb)),genenames=paste("g",as.character(1:nrow(Xb)),sep=""),logged2=TRUE
)
    sam.Tb<-samr(data, resp.type="Two class unpaired", nperms=B.sam)
    pvalores.Tb<-cbind(pvalores.Tb,samr.pvalues.from.perms(sam.Tb$tt, sam.Tb$ttstar))
    colnames(pvalores.Tb)<-c("rawp",procs,"maxT","sam")

    Tb.Vb<-rbind(Tb.Vb,colSums(pvalores.Tb[(m1+1):m,]<=alpha))
    Tb.Qb<-rbind(Tb.Qb,Tb.Vb[b,]/colSums(pvalores.Tb<=alpha))
    Tb.Qb[b,colSums(pvalores.Tb<=alpha)==0]<-0
    Tb.Teb<-rbind(Tb.Teb,colSums(pvalores.Tb[1:m1,]>alpha))

    ## Estadistico Wilcoxon
    Wb<-sapply(1:nrow(Xb),function(i) wilcox.test(Xb[i,]~Y)$p.value)
    Adj.Wb<-mt.rawp2adjp(Wb, procs)
    pvalores.Wb<-Adj.Wb$adjp[order(Adj.Wb$index), ]
    maxT.Wb <- mt.maxT(Xb, Y, test="wilcoxon", B = B.maxT)
    pvalores.Wb<-cbind(pvalores.Wb, maxT.Wb$adjp[order(maxT.Wb$index)])
    data<-list(x=Xb,y=Y+1,
geneid=as.character(1:nrow(Xb)),genenames=paste("g",as.character(1:nrow(Xb)),sep=""),logged2=TRUE
)
    sam.Wb<-samr(data, resp.type="Two class unpaired", nperms=B.sam,testStatistic="wilcoxon")
    pvalores.Wb<-cbind(pvalores.Wb,samr.pvalues.from.perms(sam.Wb$tt, sam.Wb$ttstar))
    colnames(pvalores.Wb)<-c("rawp",procs,"maxT","sam")

    Wb.Vb<-rbind(Wb.Vb,colSums(pvalores.Wb[(m1+1):m,]<=alpha))
    Wb.Qb<-rbind(Wb.Qb,Wb.Vb[b,]/colSums(pvalores.Wb<=alpha))
    Wb.Qb[b,colSums(pvalores.Wb<=alpha)==0]<-0
    Wb.Teb<-rbind(Wb.Teb,colSums(pvalores.Wb[1:m1,]>alpha))
  }

  PCER<-rbind(colMeans(Tb.Vb)/m,colMeans(Wb.Vb)/m)
  FWER<-rbind(colMeans(Tb.Vb>=1),colMeans(Wb.Vb>=1))
  FDR<-rbind(colMeans(Tb.Qb),colMeans(Wb.Qb))
  Power<-rbind(1-colMeans(Tb.Teb)/m1,1-colMeans(Wb.Teb)/m1)

  return(list(PCER=PCER,FWER=FWER,FDR=FDR,Power=Power))
}


##########Simulación 1
Sim1.m<-100 # n genes
Sim1.m1<-5 # ngenes DE
set.seed(1)
SIM1<-fun.SIM(X=X.APOE,m=Sim1.m,m1=Sim1.m1,n1=10,n2=10)

########## Simulación 2
Sim2.m<-500
Sim2.m1<-25
set.seed(2)
SIM2<-fun.SIM(X=X.APOE,m=Sim2.m,m1=Sim2.m1,n1=10,n2=10)

########## Simulación 3
Sim3.m<-1000
Sim3.m1<-50
set.seed(3)
SIM3<-fun.SIM(X=X.APOE,m=Sim3.m,m1=Sim3.m1,n1=10,n2=10)
```

```
########## Resultados
cols<-
c(rep(c("steelblue","navyblue"),4),rep(c("springgreen1","olivedrab"),2),rep(c("orange","tomato"),2))
PROCs<-c("Bonf t","Bonf W","Holm t","Holm W","Hoch t","Hoch W","maxT t","maxT W","BH t","BH
W","BY t","BY W","rawp t","rawp W","sam t","sam W")

###### PCER
par(mfrow=c(2,2))
PCER.LIM<-range(c(SIM1$PCER,SIM2.PCER,SIM3$PCER))*100

plot(as.vector(SIM1$PCER[,c(2:4,7,5:6,1,8)])*100,ylim=PCER.LIM,pch=20,ylab="PCER (%)", xlab="",
main="Simulación 1. PCER", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,PCER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,PCER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM2$PCER[,c(2:4,7,5:6,1,8)])*100,ylim=PCER.LIM,pch=20,ylab="PCER (%)", xlab="",
main="Simulación 2. PCER", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,PCER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,PCER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM3$PCER[,c(2:4,7,5:6,1,8)])*100,ylim=PCER.LIM,pch=20,ylab="PCER (%)", xlab="",
main="Simulación 3. PCER", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,PCER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,PCER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))


###### FWER
par(mfrow=c(2,2))
FWER.LIM<-range(c(SIM1$FWER, SIM2$PWER, SIM3$PWER))*100

plot(as.vector(SIM1$FWER[,c(2:4,7,5:6,1,8)])*100,ylim=FWER.LIM,pch=20,ylab="FWER (%)", xlab="",
main="Simulación 1. FWER", xaxt="n",cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,FWER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,FWER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM2$FWER[,c(2:4,7,5:6,1,8)])*100,ylim=FWER.LIM,pch=20,ylab="FWER (%)", xlab="",
main="Simulación 2. FWER", xaxt="n",cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,FWER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,FWER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM3$FWER[,c(2:4,7,5:6,1,8)])*100,ylim=FWER.LIM,pch=20,ylab="FWER (%)", xlab="",
main="Simulación 3. FWER", xaxt="n",cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,FWER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,FWER.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
```

```
###### FDR
par(mfrow=c(2,2))
FDR.LIM<-range(c(SIM1$FDR,SIM2$FDR,SIM3$FDR))*100

plot(as.vector(SIM1$FDR[,c(2:4,7,5:6,1,8)])*100,ylim=FDR.LIM,pch=20,ylab="FDR (%)", xlab="",
main="Simulación 1. FDR", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,FDR.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,FDR.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM2$FDR[,c(2:4,7,5:6,1,8)])*100,ylim=FDR.LIM,pch=20,ylab="FDR (%)", xlab="",
main="Simulación 2. FDR", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,FDR.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,FDR.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM3$FDR[,c(2:4,7,5:6,1,8)])*100,ylim=FDR.LIM,pch=20,ylab="FDR (%)", xlab="",
main="Simulación 3. FDR", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(c(0,17),rep(5,2),lwd=2)
lines(rep(8.5,2),c(0,FDR.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,FDR.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))


###### Power
par(mfrow=c(2,2))
Power.LIM<-range(c(SIM1$Power,SIM2$Power,SIM3$Power))*100

plot(as.vector(SIM1$Power[,c(2:4,7,5:6,1,8)])*100,ylim=Power.LIM,pch=20,ylab="Power (%)",xlab="",
main="Simulación 1. Power", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(rep(8.5,2),c(0,Power.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,Power.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM2$Power[,c(2:4,7,5:6,1,8)])*100,ylim=Power.LIM,pch=20,ylab="Power (%)",xlab="",
main="Simulación 2. Power", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(rep(8.5,2),c(0,Power.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,Power.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))

plot(as.vector(SIM3$Power[,c(2:4,7,5:6,1,8)])*100,ylim=Power.LIM,pch=20,ylab="Power (%)",xlab="",
main="Simulación 3. Power", xaxt="n", cex=2,cex.main=1.5,cex.lab=1.25,cex.axis=1.25,col=cols)
axis(1,at=1:16,labels=PROCs,las=2,cex.axis=1.25)
lines(rep(8.5,2),c(0,Power.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
lines(rep(12.5,2),c(0,Power.LIM[2]+1),lty=2,lwd=2,col=grey(0.4))
```