ANÁLISIS SOBRE MÉTODOS DE PRUEBAS DE HIPÓTESIS MÚLTIPLE EN LA IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

Por

Luz Marina Muñiz Rivera

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRO EN CIENCIAS

en

MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

Julio, 2009

Aprobada por:

_____	_____
Edgar Acuña Fernández, PhD Presidente, Comité Graduado	Fecha
_____	_____
Dámaris Santana, PhD Miembro, Comité Graduado	Fecha
_____	_____
Edgardo Lorenzo, PhD Miembro, Comité Graduado	Fecha
_____	_____
Juan Carlos Martínez Cruzado, PhD Representante de Estudios Graduados	Fecha
_____	_____
Silvestre Colón, Profesor Director de Departamento	Fecha

Abstract of Thesis Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master in Mathematics (Statistics)

ANALYSIS OF MULTIPLE HYPOTHESIS TESTING PROCEDURES ON THE IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

By

Luz Marina Muñiz Rivera

July 2009

Chair: Edgar Acuña
Major Department: Mathematical Sciences

The Human Genome Project is the most important reason for the surge of new technologies in the microarray area. These technologies facilitate the experimentation with a large number of genes simultaneously. These experiments allow, for example, to compare the genetic expressions between sick and healthy cells. One of the main goals in analysis of gene expression data from microarray experiments is the identification of genes differentially expressed. To accomplish this goal various methods are used. One of them is to use multiple testing establishing a null hypothesis of not association between the measure of gene expression and the response is tested for each gene. Due to the fact that in microarray experiments thousand of genes are considered, the same amount of tests must be considered. In this situation, the probability of error type I increases with the number of hypothesis. Thus, a p-value near zero can occur randomly due to the large number of hypothesis that are tested simultaneously and not because the gene is differentially expressed.

In this thesis, we study methods for multiple testing on data coming from microarray experiments. We compare theoretically and experimentally ten multiple hypothesis testing methods on six gene expression data sets, which are freely available in the internet.

Resumen de Tesis Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestro en Ciencias

ANÁLISIS SOBRE MÉTODOS DE PRUEBAS DE HIPÓTESIS MÚLTIPLE EN LA IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS

Por
Luz Marina Muñiz Rivera

Julio, 2009

Consejero: Edgar Acuña
Departamento: Ciencias Matemáticas

Con el Proyecto del Genoma Humano surgieron nuevas tecnologías, tales como los microarreglos que facilitan la ejecución de experimentos con un gran número de genes simultáneamente. Estos experimentos permiten, por ejemplo, comparar la expresión genética entre una célula sana y una enferma. Una de las principales metas en el análisis de datos de expresión de genes en experimentos de microarreglos es la identificación de genes diferencialmente expresados. Para alcanzar esta meta se han propuesto varios métodos, entre ellos, los métodos de prueba de hipótesis múltiple, los cuales establecen simultáneamente para cada gen una hipótesis nula de no asociación entre la medida de expresión de los genes y la respuesta. Pero debido a que los experimentos de microarreglos consideran miles de genes, la misma cantidad de pruebas ha de ser considerada. Por tanto la probabilidad de cometer un Error Tipo I aumenta con el número de hipótesis; pues, un *p-value* cercano a cero puede ser producto del azar debido al gran número de hipótesis que se prueban simultáneamente y no porque el gen sea diferencialmente expresado.

En esta tesis, estudiamos métodos de prueba de hipótesis múltiple para datos de expresión genética provenientes de experimentos de microarreglos. Para ello examinamos y

comparamos, teóricamente y experimentalmente diez métodos de prueba de hipótesis múltiple empleando seis conjuntos de datos de expresión genética, disponibles gratuitamente en la internet.

En el transcurso de nuestras vidas existen momentos, experiencias, sueños, metas y personas especiales. De estas últimas podemos decir que algunas de ellas solo van de paso dejando una leve huella, otras pertenecen a ella desde nuestro origen y otras llegan no solo para quedarse sino para formar parte importante de ella.

Cuando llegué a Puerto Rico con el anhelo de materializar unas de mis metas, no imaginé que además de ello, aquí llegaría a apreciar aún más el valor incalculable que para mí tiene mi familia, aquella en la que Dios me puso en este mundo. Por mi mente nunca pasó que además de un título académico, encontraría de entre todos al ser más especial, aquel que siempre me ha brindado no solo su apoyo incondicional, sino su amor, aquel hombre que llenaría mi vida de felicidad, con el que estoy dispuesta a compartir momentos felices y tristes, con el que he llegado a formar un hogar, conociendo la dicha inmensa de ser madre de una hermosa niña, regalo de Dios y fruto de nuestro amor.

Por estas razones dedico este trabajo a mi familia, la cual siempre me dio su apoyo desde el preciso instante de mi partida a Puerto Rico, aún sabiendo la distancia física que nos separaría cuando partí de mi querida Cartagena con una maleta llena de sueños. A mi madre, quien siempre se esmeró en darle a sus hijos lo necesario y suficiente, con lo poco que tenía, Gracias mami. A mi hermanita Edenis, de la cual me separé en un momento crucial de su vida, solo tú sabes lo especial que eres para mí.

A mi esposo William, quien siempre me animó a seguir adelante y de quien he aprendido muchas cosas buenas. A mi hija Luz Verónica, muñequita y princesa mágica de mamá, por animarme con su sola presencia. A ambos, gracias por ser reales y por existir en mi vida, simplemente los amo.

Y por último pero no menos importante, a Dios, por estar siempre a mi lado y porque desde esta tierra a la que he aprendido a querer, me enseñó que los planes que tiene trazados para nosotros son más grandes y sublimes que los nuestros.

AGRADECIMIENTOS

Principalmente a Dios, por la fortaleza, sabiduría y constancia, que su Santo Espíritu me dio para realizar este trabajo.

A mi consejero el doctor Edgar Acuña Fernández por compartirme sus conocimientos y por darme la oportunidad de investigar bajo su dirección y supervisión.

Al apoyo económico proporcionado por la oficina de Investigación Naval (ONR), recibido a través del Grant N0014-03-0359. Y al apoyo económico parcial recibido del Departamento de Defensa, a través del Grant N0014-06-1-0555.

A la profesora Dámaris Santana por compartir con William y conmigo mucho más que academia.

A mi esposo, hija, padres y hermanos, por obvias razones.

A César Herazo por incentivarme siempre a realizar estudios graduados, por la confianza que ha depositado en mí, no solo como estudiante o compañera de trabajo, sino también como amiga, compartiendo su experiencia de vida conmigo, brindándome consejos sinceros y haciendo suyos todos mis éxitos.

Al departamento de Matemáticas en general, por brindarme la oportunidad de realizar mis estudios graduados.

Y a todas aquellas personas que de una u otra forma contribuyeron en este proceso.

TABLA DE CONTENIDO

Página

ABSTRACT . . . . . **I**

RESUMEN . . . . . **IV**

AGRADECIMIENTOS . . . . . **VII**

LISTA DE TABLAS . . . . . **X**

LISTA DE FIGURAS . . . . . **XI**

LISTA DE ABREVIATURAS . . . . . **XII**

1 INTRODUCCIÓN **1**

1.1 MOTIVACIÓN 1

1.2 REVISIÓN LITERARIA 2

1.3 OBJETIVO DE LA TESIS 5

1.4 RESUMEN DE LOS SIGUIENTES CAPÍTULOS 5

2 NOCIONES SOBRE BILOGÍA MOLECULAR **7**

2.1 CÉLULA 8

2.1.1 CÉLULA PROCARIÓTICA 8

2.1.2 CÉLULA EUCARIÓTICA 8

2.2 ESTRUCTURA DE UNA CÉCULA EUCARIÓTICA ANIMAL TÍPICA 9

2.2.1 NÚCLEO. 9

2.2.2 CROMOSOMAS 9

2.3 FUENTE Y TRANSMISIÓN DE INFORMACIÓN GENÉTICA 10

2.3.1 ÁCIDO DESOXIRRIBONUCLEICO "*ADN*" 10

2.3.2 ÁCIDO RIBONUCLEICO "*ARN*" 11

2.4 GENOMA HUMANO 12

2.5 GEN 12

2.5.1 EXPRESIÓN GENÉTICA 13

2.6 DOGMA CENTRAL DE LA BIOLOGÍA MOLECULAR 13

2.7 MICROARREGLOS 14

2.7.1 CONCEPTOS PRELIMINARES RELACIONADOS CON MICROARREGLOS 15

2.7.2 TIPOS DE MICROARREGLOS DE *ADN* 15

2.7.3 IMPORTANCIA DE LOS MICROARREGLOS 17
2.7.4 EXPERIMENTOS CON MICROARREGLOS 17
2.8 DATOS DE EXPRESIÓN GENÉTICA 19
2.9 IDENTIFICACIÓN DE GENES DIFERENCIALMENTE EXPRESADOS 20

3 PRUEBAS MÚLTIPLES EN EXPERIMENTOS DE MICROARREGLOS 21

3.1 PRUEBA DE HIPÓTESIS SIMPLE 22
3.2 PRUEBA DE HIPÓTESIS MÚLTIPLE EN EXPERIMENTOS DE MICROARREGLOS 23
3.3 ERROR TIPO I Y TIPO II 24
3.3.1 TASA DE ERROR TIPO I 25
3.3.2 CONTROL DE LA TASA DE ERROR TIPO I 26
3.3.3 CONTROL DE LA TASA DE ERROR TIPO II 27
3.4 ESTADÍSTICOS DE PRUEBA 28
3.4.1 PRUEBA T PARA COMPARAR DOS MEDIAS POBLACIONALES USANDO MUESTRAS INDEPENDIENTES CUANDO LAS VARIANZAS POBLACIONALES SON DESCONOCIDAS E IGALES 29
3.4.2 PRUEBA T PARA COMPARAR DOS MEDIAS POBLACIONALES USANDO MUESTRAS INDEPENDIENTES CUANDO LAS VARIANZAS POBLACIONALES SON DESCONOCIDAS Y DIFERENTES 30
3.4.3 PRUEBA *MANN-WITHNEY* PARA DOS MUESTRAS INDEPENDIENTES 30
3.4.4 PRUEBA *F* PARA COMPARAR VARIAS MEDIAS POBLACIONALES 31
3.4.5 PRUEBA *KRUSKALL-WALLIS* PARA COMPARAR VARIAS MUESTRAS 33
3.5 *P-VALUES* NO AJUSTADOS (*UNADJUSTED*) Y AJUSTADOS (*ADJUSTED*) . . 34

4 MÉTODOS DE PRUEBA DE HIPÓTESIS MÚLTIPLE 38

4.1 MÉTODO DE PRUEBA MÚLTIPLE (*MTP*) 38
4.2 MÉTODOS DE UN SOLO PASO QUE CONTROLAN EL *FWER* 39
4.2.1 MÉTODO DE BONFERRONI 40
4.2.2 MÉTODO DE ŜIDÁK (*SS*) 41
4.2.3 MÉTODOS DE *MINP* Y *MAXT* 42
4.3 MÉTODOS DE PASOS HACIA ABAJO QUE CONTROLAN EL *FWER* 43
4.3.1 MÉTODO DE HOLM 46
4.3.2 MÉTODO DE ŜIDÁK (*SD*) 47
4.3.3 MÉTODOS DE *MINP* Y *MAXT* 47
4.4 MÉTODOS DE PASOS HACIA ARRIBA QUE CONTROLAN EL *FWER* . . . 48
4.4.1 DESIGUALDAD DE SIMES 48
4.4.2 MÉTODO DE HOCHBERG 49
4.5 MÉTODOS QUE CONTROLAN EL *FDR* 50

4.5.1 MÉTODOS DE PASOS HACIA ARRIBA DE BENJAMINI Y HOCHBERG 51
4.5.2 MÉTODOS DE PASOS HACIA ARRIBA DE BENJAMINI Y YEKUTIELI 51
4.5.3 ANÁLISIS DE SIGNIFICANCIA DE MICROARREGLOS (*SAM*) . . . 52
4.6 MÉTODOS DE *P-VALUES* AJUSTADOS POR PERMUTACIÓN 54

5 METODOLOGÍA Y RESULTADOS 57

5.1 BASE DE DATOS 57
5.1.1 APO AI 58
5.1.2 LEUKEMIA 58
5.1.3 PROSTATE 58
5.1.4 COLON 59
5.1.5 BREAST CÁNCER (BREASTCC) 59
5.1.6 SRBCT 59
5.2 EXISTENCIA DE GENES DIFERENCIALMENTE EXPRESADOS 59
5.3 NÚMERO DE HIPÓTESIS NULAS RECHAZADAS 63
5.3.1 RESULTADOS PARA APO AI 65
5.3.2 RESULTADOS PARA LEUKEMIA 67
5.3.3 RESULTADOS PARA PROSTATE 68
5.3.4 RESULTADOS PARA COLON 69
5.3.5 RESULTADOS PARA BREASTCC 70
5.3.6 RESULTADOS PARA SRBCT 71
5.4 NÚMERO DE GENES DIFERENCIALMNETE EXPRESADOS OBTENIDOS POR SIMULACIÓN 72
5.5 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS 74
5.6 CORRELACIÓN 77
5.6.1 CORRELACIÓN PARA APO AI 78
5.6.2 CORRELACIÓN PARA LEUKEMIA 79
5.6.3 CORRELACIÓN PARA PROSTATE 81
5.6.4 CORRELACIÓN PARA COLON 82
5.6.5 CORRELACIÓN PARA BREASTCC 84
5.6.6 CORRELACIÓN PARA SRBCT 85

6 **CONCLUSIONES 87**

6.1 CONCLUSIONES 87
6.2 TRABAJOS FUTUROS 89

APÉNDICE A CÓDIGOS DE FUNCIONES EN R USADOS EN ESTA TESIS 90

APÉNDICE B MÉTODO DE BENJAMINI Y HOCHBERG (*BH*) 94

APÉNDICE C MÉTODO DE BENJAMINI Y YEKUTIELI (*BY*) 97

REFERENCIAS BIBLIOGRÁFICAS 101

LISTA DE TABLAS

Tablas Página

TABLA 3.3 ERROR TIPO I Y TIPO II 24

TABLA 3.4.4 ANÁLISIS DE VARIANZA 32

TABLA 5.1 CARACTERÍSTICAS DE LAS BASES DE DATOS. 57

TABLA 5.3 NÚMERO DE HIPÓTESIS NULAS RECHAZADAS A UN NIVEL DE SIGNIFICANCIA DEL 5% 64

TABLA 5.3.1 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS PARA APO AI 65

TABLA 5.3.2 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS PARA LEUKEMIA 67

TABLA 5.3.3 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS PARA PROSTATE 68

TABLA 5.3.4 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS PARA COLON 69

TABLA 5.3.5 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS PARA BREASTCC 70

TABLA 5.3.6 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS PARA SRBCT 71

TABLA 5.4 NÚMERO DE GENES DIFERENCIALMENTE EXPRESADOS OBTENIDOS POR SIMULACIÓN, PARA UNA MUESTRA ALEATORIA DE 500 GENES CON DISTRIBUCION NORMAL MULTIVARIADA, A UN NIVEL DE SIGNIFICANCIA DEL 5% 74

TABLA 5.5.1 TOP GENES DIFERENCIALMENTE EXPRESADOS PARA APO AI . . 75

TABLA 5.5.2 TOP GENES DIFERENCIALMENTE EXPRESADOS PARA LEUKEMIA . 75

TABLA 5.5.3 TOP GENES DIFERENCIALMENTE EXPRESADOS PARA PROSTATE . 75

TABLA 5.5.4 TOP GENES DIFERENCIALMENTE EXPRESADOS PARA COLON . . 76

TABLA 5.5.5 TOP GENES DIFERENCIALMENTE EXPRESADOS PARA BREASTCC . 76

TABLA 5.5.6 TOP GENES DIFERENCIALMENTE EXPRESADOS PARA SRBCT . . 76

LISTA DE FIGURAS

Figuras Página

FIGURA 2.3.1 REPRESENTACIÓN DE CADENA DE *ADN* DOBLE 11

FIGURA 2.6 REPRESENTACIÓN GRÁFICA DEL DOGMA DE LA BIOLOGÍA MOLECULAR 14

FIGURA 2.7.4 EXPERIMENTO DE MICROARREGLO 19

FIGURA 4.6.1 ALGORITMO DE PERMUTACIÓN PARA P-VALUES NO AJUSTADOS . . 55

FIGURA 4.6.2 ALGORITMO DE PERMUTACIÓN PARA P-VALUES AJUSTADOS, BASADO EN EL ALGORITMO DE WESTFALL Y YOUNG 56

FIGURA 5.2.1 HISTOGRAMA Y Q-Q PLOT DE NORMALIDAD PARA APO AI . . . 60

FIGURA 5.2.2 HISTOGRAMA Y Q-Q PLOT DE NORMALIDAD PARA LEUKEMIA . . 61

FIGURA 5.2.3 HISTOGRAMA Y Q-Q PLOT DE NORMALIDAD PARA PROSTATE . . 61

FIGURA 5.2.4 HISTOGRAMA Y Q-Q PLOT DE NORMALIDAD PARA COLON . . . 62

FIGURA 5.2.5 HISTOGRAMA Y Q-Q PLOT DE NORMALIDAD PARA BREASTCC . . 62

FIGURA 5.2.6 HISTOGRAMA Y Q-Q PLOT DE NORMALIDAD PARA SRBCT . . . 63

FIGURA 5.3.1(a) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA APO AI 66

FIGURA 5.3.1(b) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA 500 HIPÓTESIS NULAS. APO AI 66

FIGURA 5.3.2(a) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA LEUKEMIA 67

FIGURA 5.3.2(b) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA 500 HIPÓTESIS NULAS. LEUKEMIA 67

FIGURA 5.3.3(a) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA PROSTATE 68

FIGURA 5.3.3(b) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA 500 HIPÓTESIS NULAS. PROSTATE 68

FIGURA 5.3.4(a) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA COLON 69

FIGURA 5.3.4(b) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA 500 HIPÓTESIS NULAS. COLON 69

FIGURA 5.3.5(a) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA BREASTCC 70

FIGURA 5.3.5(b) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA 500 HIPÓTESIS NULAS. BREASTCC 70

FIGURA 5.3.6(a) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA SRBCT 71

FIGURA 5.3.6(b) PLOT DE NÚMERO DE HIPÓTESIS RECHAZADAS VERSUS TASA DE ERROR TIPO I PARA 500 HIPÓTESIS NULAS. SRBCT 71

FIGURA 5.4.1 COMPORTAMIENTO DE MÉTODOS DE PRUEBA MÚLTIPLE VS NÚMERO DE HIPÓTESIS, PARA APO AI 72

FIGURA 5.4.2 COMPORTAMIENTO DE MÉTODOS DE PRUEBA MÚLTIPLE VS NÚMERO DE HIPÓTESIS, PARA LEUKEMIA 72

FIGURA 5.4.3 COMPORTAMIENTO DE MÉTODOS DE PRUEBA MÚLTIPLE VS NÚMERO DE HIPÓTESIS, PARA PROSTATE 73

FIGURA 5.4.4 COMPORTAMIENTO DE MÉTODOS DE PRUEBA MÚLTIPLE VS NÚMERO DE HIPÓTESIS, PARA COLON 73

FIGURA 5.4.5 COMPORTAMIENTO DE MÉTODOS DE PRUEBA MÚLTIPLE VS NÚMERO DE HIPÓTESIS, PARA BREASTCC 73

FIGURA 5.4.6 COMPORTAMIENTO DE MÉTODOS DE PRUEBA MÚLTIPLE VS NÚMERO DE HIPÓTESIS, PARA SRBCT 73

FIGURA 5.6.1(a) CORRELACIÓN PARA MUESTRAS DE *ARN* EN APO AI 78

FIGURA 5.6.1(b) CORRELACIÓN PARA MUESTRA ALEATORIA DE 500 GENES. APO AI 78

FIGURA 5.6.1(c) CORRELACIÓN PARA MUESTRA ALEATORIA NORMAL MULTIVARIADA DE 500 GENES. APO AI 78

FIGURA 5.6.2(a) CORRELACIÓN PARA MUESTRAS DE *ARN* EN LEUKEMIA 79

FIGURA 5.6.2(b) CORRELACIÓN PARA MUESTRA ALEATORIA DE 500 GENES. LEUKEMIA 80

FIGURA 5.6.2(c) CORRELACIÓN PARA MUESTRA ALEATORIA NORMAL MULTIVARIADA DE 500 GENES. LEUKEMIA 80

FIGURA 5.6.3(a) CORRELACIÓN PARA MUESTRAS DE *ARN* EN PROSTATE 81

FIGURA 5.6.3(b) CORRELACIÓN PARA MUESTRA ALEATORIA DE 500 GENES. PROSTATE

. 81

FIGURA 5.6.3(c) CORRELACIÓN PARA MUESTRA ALEATORIA NORMAL MULTIVARIADA DE 500 GENES. PROSTATE 81

FIGURA 5.6.4(a) CORRELACIÓN PARA MUESTRAS DE *ARN* EN COLON 82

FIGURA 5.6.4(b) CORRELACIÓN PARA MUESTRA ALEATORIA DE 500 GENES. COLON 83

FIGURA 5.6.4(c) CORRELACIÓN PARA MUESTRA ALEATORIA NORMAL MULTIVARIADA DE 500 GENES. COLON 83

FIGURA 5.6.5(a) CORRELACIÓN PARA MUESTRAS DE *ARN* EN BREASTCC 84

FIGURA 5.6.5(b) CORRELACIÓN PARA MUESTRA ALEATORIA DE 500 GENES. BREASTCC 84

FIGURA 5.6.5(c) CORRELACIÓN PARA MUESTRA ALEATORIA NORMAL MULTIVARIADA DE 500 GENES. BREASTCC 84

FIGURA 5.6.6(a) CORRELACIÓN PARA MUESTRAS DE *ARN* EN SRBCT 85

FIGURA 5.6.6(b) CORRELACIÓN PARA MUESTRA ALEATORIA DE 500 GENES. SRBCT 86

FIGURA 5.6.6(c) CORRELACIÓN PARA MUESTRA ALEATORIA NORMAL MULTIVARIADA DE 500 GENES. SRBCT 86

Capítulo 1
INTRODUCCIÓN

El descubrimiento de la secuencia del genoma humano ha motivado que se creen tecnologías que facilitan la ejecución de experimentos con miles de genes simultáneamente. Los experimentos de microarreglos son de este tipo. Además, se ha descubierto que la secuencia genética es idéntica en un 99.9% en todos los seres humanos. Sin embargo, a pesar que todas las células poseen la misma información genética, se diferencian unas de otras en que diferentes genes se expresan en diferentes niveles; determinando así, las diferencias en las funciones específicas de cada célula del cuerpo y entre los estados de salud y enfermedad.

Una de las metas de mayor interés en el análisis de datos de expresión genética, es la identificación de genes diferencialmente expresados para múltiples arreglos. Es decir, la identificación de genes cuyos niveles de expresión se asocian con una variable de respuesta o covariable de interés; entendiéndose por variable de respuesta algún resultado clínico o patológico y por covariable un estado (tratamiento/control), tipo de célula o droga, dosis de una droga o un tiempo particular. Para alcanzar esta meta se han sugerido varios métodos. Entre ellos, los métodos de prueba de hipótesis múltiple, los cuales establecen simultáneamente para cada gen una hipótesis nula de no asociación entre la medida de expresión de los genes y la respuesta.

1.1 Motivación

Estudios científicos recientes han mostrado cómo la consideración de los datos moleculares aportados mediante el análisis sistemático de la expresión genética utilizando microarreglos de *ADN* permite diferenciar nuevas clases de tumores dentro de grupos ya conocidos, así como predecir la evolución clínica y la respuesta al tratamiento de

determinados casos en los que los métodos tradicionales no alcanzaban a aportar tal información.

Los métodos de pruebas de hipótesis múltiples para experimentos de microarreglos de múltiples arreglos asumen una prueba estadística para la expresión diferencial de cada gen y debido a que estos tipos de experimentos consideran miles de genes simultáneamente debe entonces considerarse la misma cantidad de pruebas. Sin embargo, según el marco estadístico clásico, en cualquier situación de prueba pueden ser cometidos dos tipos de errores, Error Tipo I y Error Tipo II, llamados en este contexto falsos positivos y falsos negativos, respectivamente; los cuales para cada prueba tienen una probabilidad de ocurrencia específica. Por tanto, para los experimentos de microarreglos, la probabilidad de cometer un Error Tipo I aumenta con el número de hipótesis; pues, un *p-value* cercano a cero puede ser producto del azar debido al gran número de hipótesis que se prueban simultáneamente y no necesariamente debido a que el gen sea diferencialmente expresado. Sobre estos métodos existe una amplia literatura, aunque en lo que se refiere a conjuntos de datos con dos clases.

En esta tesis se describen y estudian los métodos de prueba de hipótesis múltiple dentro del marco estadístico estándar para datos de expresión genética provenientes de experimentos de microarreglos, con dos, tres y cuatro clases. Para ello se examinan y comparan, teóricamente y experimentalmente diez métodos de prueba de hipótesis múltiple y se hallan los genes diferencialmente expresados de seis conjuntos de datos de expresión genética, disponibles gratuitamente en sitios de internet.

1. 2 Revisión Literaria

Conceptualmente, la técnica de los microarreglos de ácido desoxirribonucleico (*ADN*) no es nueva, ya en los años 80 existían técnicas como el *differential screening* que permitían detectar genes cuyo nivel de expresión era diferente en distintas situaciones biológicas, por ejemplo entre células tumorales y células normales, o entre regiones diferentes del cerebro.

A partir de la segunda mitad del siglo XX comenzaron a desarrollarse técnicas de biología molecular que permitían analizar y manipular el *ADN*. El Proyecto Genoma Humano

fundado en 1990 y cuyo interés es determinar las posiciones relativas de todos los nucleótidos (o pares de bases que constituyen el *ADN* cromosómico) e identificar los 20,000 a 25,000 genes presentes en él. Gracias a la gran colaboración internacional, a los avances en el campo de la genómica (especialmente, en el análisis de secuenciación), así como los avances en la tecnología informática, se presentó un borrador inicial del genoma en el año 2003. Esta secuenciación del genoma humano ha sentado las bases para el desarrollo de cambios revolucionarios en la medicina molecular, la cuál junto al desarrollo de las técnicas de microarreglos permite analizar simultáneamente miles de genes para los cuales, su secuencia es conocida.

El proceso de expresión de genes en muestras de ácido ribonucleico (*ARN*) se conoce desde 1963, pero solo hasta 1977 se desarrolló un procedimiento práctico de medida, que sólo permitía un gen. En 1989 se desarrollan técnicas, basados en la tecnología de construcción de microchips, que permiten fijar millones de cadenas de *ADN* sobre una placa (los microarreglos).

En 1994, la compañía de biotecnología *Affymetrix*, empleando técnicas de fotolitográficas, desarrolló arreglos de *ADN* de alta densidad. Arreglos de menor tamaño se idearon después constituyendo los microarreglos de *ADN* o *CHIPS*. Los microarreglos modernos pueden contener aproximadamente 20,000 a 30,000 genes. En los últimos años los microarreglos de *ADN* se han convertido en la herramienta universalmente más utilizada para el análisis de expresión genética.

A pesar de su reciente introducción, esta tecnología ha experimentado un gran auge al proveernos de una herramienta que permite realizar avances rápidos en el conocimiento de diversas patologías desde el punto de vista de la biología molecular. En el año 2000, *Alizadeh et al.* [1] muestran cómo la utilización de un microarreglo especializado, el *linfochip*, permite distinguir, dentro de los tumores tradicionalmente clasificados como linfomas *no-Hodgkin*, dos distintos tipos de patología con pronóstico diferente. En 1999, *Golub et al.* [15] mencionan cómo los datos moleculares basados en el comportamiento de un sólo gen no permiten distinguir entre dos tipos distintos de leucemia aguda, pero sin embargo pueden distinguirse de manera sólida cuando la predicción se basa en los niveles de

expresión de 50 genes (seleccionados de entre los más de 6,000 representados en los arreglos utilizados). En el 2002, el cáncer de mama presenta manifestaciones clínicas muy diversas; *Van´t Veer et al*. [46], utilizando chips de *ADN*, muestran cómo la aparición de un patrón específico de expresión genética en algunos casos de cáncer de mama se correlaciona con manifestaciones agresivas de esta enfermedad.

Los datos de los experimentos de microarreglos consisten de dos medidas de intensidad de fluorescencia, (*R, G*), las cuales representan respectivamente, los niveles de expresión de los genes en las muestras de *ARN* mensajero etiquetadas como rojo (*R*) y verde (*G*), respectivamente. A partir del siglo XXI se han realizado muchos estudios con este tipo de experimentos para identificar genes diferencialmente expresados; entre estos estudios se encuentran los realizados por *Callow et al*. [6], *Friddle et al*. [13], *Galitski et al.* [14], *Golub et al.* [15], and *Spellman et al.* [42].

Para los experimentos de microarreglos que consisten de un solo arreglo se distinguen dos tipos de métodos principales: aquellos que solo consideran el valor de la razón de las fluorescencias *R/G* y los que además consideran la abundancia de transcripción del producto *RG*. Los primeros análisis de microarreglos (*DeRisi et al.* [9], *Schena et al.* [36, 37]) identifican los genes diferencialmente expresados estableciendo cotas superiores e inferiores para el logaritmo de la razón *R/G*. Por otra parte *Sapir* y *Churchill* [35] lo hacen usando sólo la probabilidad posterior del cambio del logaritmo de tal razón bajo un modelo de mezclas, con la limitación de que ignoran la información que proporciona el producto *RG*. Sin embargo, *Newton et al.* [31] notan esta limitación e identifican los genes diferencialmente expresados basados en la probabilidad posterior del cambio ya mencionado bajo un modelo jerárquico para (*R, G*), tales probabilidades son funciones de *R + G* y *RG*, considerando de esta forma un método que considera la abundancia total de la transcripción. *Roberts et al.* [32] establecen un método también con esta misma característica, basado en que *R* y *G* son aproximadamente independientes y distribuidos normalmente, con varianza dependiente de la media. A fin de cuentas, cada uno de estos métodos produce un modelo dependiente de las cantidades *R* y *G*. Aunque los datos de expresión de genes pueden ser muy variables

(ruidosos) para identificar genes diferencialmente en experimentos sin réplica, las reglas antes mencionadas pueden considerarse buenas en general.

Para la identificación de genes diferencialmente expresados en experimentos de múltiples arreglos, se han realizado también varios estudios, algunos de ellos son: el análisis exploratorio de conglomerados considerando modelos jerárquicos (*Alizadeh et al.* [1], *Ross et al.* [34], *Tamayo et al.* [43]); la inspección visual de los conglomerados obtenidos (*Eisen et al.* [12]); el cálculo para cada gen de la correlación de su perfil de expresión (*Galitski et al.* [14] and *Golub et al.* [15]); la aplicación de técnicas de análisis de variación (*Kerr et al.* [25]). En el año 1993, *Westfall* y *Young*, proponen métodos de prueba múltiple que controlan la tasa de error *family wise* y considera la estructura de dependencia entre los estadísticos de prueba, basándose en el remuestreo de los *p-values* ajustados. *Benjamini* y *Hochberg* en 1995 proponen métodos que controlan la tasa de descubrimiento falso. Más tarde en el año 2002 Storey, motivado por la teoría Bayesiana, desarrolla un nuevo concepto llamado tasa de descubrimiento falso positivo.

1.3 Objetivo de la Tesis

Esta investigación se centra principalmente en evaluar la efectividad de algunos métodos de pruebas de hipótesis múltiple utilizados para identificar genes diferencialmente expresados, usando datos de expresión genética obtenidos mediante experimentos de microarreglos, con muestras pertenecientes a dos, tres y cuatro clases de condición o tejido. Para ello se discute el tipo de control que ofrece cada método para la tasa de Error Tipo I y se evalúa experimentalmente la efectividad de diez métodos de prueba de hipótesis múltiple, identificando los correspondientes genes diferencialmente expresados en seis conjuntos de datos de expresión genética, evaluando la correlación entre las muestras y la correlación entre los genes, y comparando estos resultados con los obtenidos para conjuntos de datos simulados. En los experimentos, se usa el software Bioconductor, basado en el lenguaje de programación R. Este software proporciona las listas de genes diferencialmente expresados para cada método.

1.4 Resumen de los siguientes Capítulos

El Capítulo 1 de esta tesis, constituido por la Introducción, pretende introducir brevemente al lector en el contexto relacionado con la identificación de genes diferencialmente expresados, explicando la motivación de esta investigación, revisando un poco de literatura relacionada con ella y estableciendo los objetivos de la misma.

Como principio para comprender el marco referencial de esta tesis, en el Capítulo 2 se presentan las nociones básicas sobre biología molecular, las cuales comprenden desde la definición de una célula hasta la explicación del Dogma Central de la Biología Molecular. El Capítulo 3 establece lo concerniente a las técnicas de microarreglos; empezando por explicar en qué consiste un microarreglo y los diferentes tipos existentes y en particular, los que se usan para esta tesis, hasta llegar a establecer las características de los datos de expresión genética y explicar en qué consiste la identificación de genes diferencialmente expresados. En el Capítulo 4 se discuten los métodos de prueba de hipótesis múltiples empleados en la identificación de genes diferencialmente expresados. El Capítulo 5 se describen los conjuntos de datos usados y se reportan y establecen los resultados obtenidos para tales conjuntos y para los conjuntos simulados. Finalmente, en el Capítulo 6 se establecen las conclusiones, recomendaciones e inquietudes que surgieron al término de esta investigación.

Capítulo 2
NOCIONES SOBRE BIOLOGÍA MOLECULAR Y TÉCNICAS DE MICROARREGLOS

La Biología Molecular es el estudio de la vida a nivel molecular. Su desarrollo comienza en el año 1866 cuando *Mendel* publica sus experimentos sobre los principios de segregación y clasificación independiente de los genes. En 1869, el científico suizo *Frederick Miescher* descubrió en el núcleo de las células una sustancia de carácter ácido a la que llamó nucleina. En los años 20, el químico alemán *Robert Feulgen*, utilizando una tinción específica, descubrió que el *ADN* estaba situado en los cromosomas. A partir de este descubrimiento todo sucedió muy rápidamente. En 1944, *Avery, McCleod y McCarty* comprueban que el *ADN* es el portador de la información genética. En 1953, *Watson y Crick* revelan la estructura del *ADN* como una doble hélice complementaria. A partir de entonces y de manera exponencial, suceden los descubrimientos que conducirán a lo que se conoce como tecnología del *ADN* recombinante.

El interés fundamental de la Biología Molecular es comprender todos los procesos celulares que contribuyen a que la información genética se transmita eficientemente de unos seres a otros y se exprese en los nuevos individuos. Con la secuencia del genoma humano, la Biología Molecular avanza significativamente, al establecer que la diferencia entre cada una de las células, a pesar de que poseen la misma información genética, se debe a que diferentes genes se expresan en diferentes niveles. Razón por la cual, es imprescindible medir el nivel de expresión de un gen. Antes, esta medición se hacía empleando técnicas como el análisis por *Northern blot* y el método *RT-PCR*, que suplieron las necesidades durante muchos años pero sufrían la limitación del número de genes a reportar, debido a que su principio fundamental era la hibridización, la cual se basa en la complementariedad de las bases. Actualmente, la meta de los microarreglos es analizar muchos genes a la vez.

2.1 Célula

La célula es la unidad estructural y funcional esencial más pequeña y simple de todo ser vivo, con capacidad para la vida independiente. Toda célula contiene la información hereditaria necesaria para el control de su propio ciclo y del desarrollo y el funcionamiento de un organismo, así como para la transmisión de esa información a la siguiente generación celular.

Algunos organismos microscópicos, como las bacterias, poseen una única célula, mientras que los animales y plantas están formados por millones de células organizadas en tejidos y órganos. La teoría celular distingue dos clases generales de células, las procarióticas y las eucarióticas.

2.1.1 Célula Procariótica

Las células procarióticas son las más primitivas de la tierra. Son pequeñas y de estructura sencilla; carecen de núcleo celular diferenciado, debido a que el material genético (*ADN*) está concentrado en una o más regiones nucleares, que se denominan nucleoides. Estos se encuentran rodeados por citoplasma pero carecen de una membrana que los separe del resto de la célula. Los organismos procariotes son unicelulares y comprenden las bacterias y cianobacterias (antes llamadas algas verdeazuladas).

2.1.2 Célula Eucariótica

Los organismos que poseen este tipo de células son los animales, plantas, hongos y levaduras, protozoos y algas. La célula eucariótica estructuralmente más compleja que la célula procariótica. Se caracteriza principalmente porque su núcleo está rodeado por una membrana nuclear, en el cual reside el *ADN*. Su estructura básica es relativamente estable y está caracterizada por la presencia de diferentes orgánulos intracitoplasmáticos que desarrollan funciones específicas. Para efectos de este trabajo las células de interés son las células animales eucarióticas. A continuación se definen algunas partes de su estructura.

2.2 Estructura de una Célula Eucariótica Animal Típica

La estructura de una célula animal eucariótica está constituida fundamentalmente por Membrana Plasmática y Citoplasma. La Membrana Plasmática es la capa externa de la célula que envuelve, da forma y protege todo el contenido celular. Además, controla y posibilita el transporte e interacción de las clases y cantidades de sustancias que circulan dentro y fuera de la célula.

El Citoplasma es todo el contenido de la célula, formado por los organelos, los cuales son a la célula como los órganos a un cuerpo y constituyen cuerpos individualizados del resto del protoplasma que cumplen funciones específicas.

2.2.1 Núcleo

Dirige las funciones celulares y es generalmente la mayor estructura celular, con forma esférica u oval, envuelto por una membrana doble llamada membrana nuclear, que contiene en su interior moléculas de *ADN* organizadas en cromosomas. Estos contienen a su vez toda la información hereditaria y suelen aparecer dispuestos en pares.

La función principal del núcleo es transmitir de una célula a otra la información genética que posee, sin modificarla ni empobrecerla. También contiene un nucléolo, que es una región especial, que controla la síntesis *ARN* (ácido ribonucleico y ribosomal) de acuerdo con las instrucciones contenidas en el *ADN*. Estas después se dirijen al citoplasma a través de los poros nucleares y forme el andamiaje sobre el cual se ensamblan los ribosomas.

2.2.2 Cromosomas

Son corpúsculos curvos, que forman estructuras en forma de bastoncillos o en forma de *V* o *X* durante la metafase, característicos del núcleo y portadores de los factores hereditarios (genes). Durante la división celular (mitosis) aparecen claramente diferenciados y cada uno consta de una doble hélice de *ADN*, envuelta sobre sí mismas varias veces. Cada especie animal y vegetal tiene un número constante y determinado de cromosomas.

2.3 Fuente y Transmisión de Información Genética

La información genética de todo individuo está contenida en las biomoléculas de ácidos nucleicos. Aunque la función de los ácidos nucleicos es contener la información necesaria para la síntesis de las proteínas celulares, ésta no es la única, pues, hay secuencias regulatorias que controlan la expresión de las diferentes unidades genéticas por sí mismas o a su vez son controladas por otras moléculas. Hay también ácidos nucleicos implicados en la transmisión y procesamiento de la información genética y ácidos nucleicos con funciones catalíticas. Todas las células contienen dos tipos de ácidos nucleicos: ácido desoxirribonucleico o *ADN* y ácido ribonucleico o *ARN*.

2.3.1 Ácido Desoxirribonucleico (*ADN*)

El ácido desoxirribonucleico es una macromolécula, constituida por tres sustancias distintas: un grupo fosfato, un monosacárido aldehídico del tipo pentosa (la desoxirribosa), y una base nitrogenada cíclica que puede ser púrica (adenina "*A*" o guanina "*G*") o pirimidínica (timina "*T*" o citocina "*C*"). La unión de las bases nitrogenadas (*C, A, G* o *T*) con la pentosa (desoxirribosa) forma un nucleósido; el cual al unirse al grupo fosfato, produce un nucleótido. La unión de estos últimos entre sí en enlace fosfo diester origina el polinucleótido, que es en este caso el ácido desoxirribonucleico de hebra sencilla.

En las células eucarióticas, se encuentra organizado dentro del núcleo celular en los cromosomas y una mínima parte en las mitocondrias y en los cloroplastos en caso de tenerlos.

El *ADN* es una doble hélice con forma de escalera en espiral, con barandas de azúcar y fosfato, cuyos peldaños están formados por las bases unidas entre sí siempre por los pares *A-T* y *C-G*, apareamiento que se mantiene debido a la acción de los puentes de hidrogeno entre ambas bases (ver Figura 2.3.1). La información genética está codificada en la secuencia de los peldaños, y su lectura genera productos que representan la base biológica de funciones complejas, tales como respirar y pensar.

El *ADN* contiene toda la información necesaria para construir las características biológicas de una célula, y su papel principal es el de ser portador y transmisor de la información genética

entre generaciones. La información contenida en la secuencia lineal de bases del *ADN* codifica la síntesis de una secuencia lineal de aminoácidos de una proteína, así, cualquier cambio en estas bases causa un cambio en la proteína correspondiente. Las variaciones necesarias para adaptaciones y evoluciones entre generaciones, se obtienen por los procesos de replicación y mutación del *ADN*, que tienen lugar en el núcleo antes de comenzar la división celular.

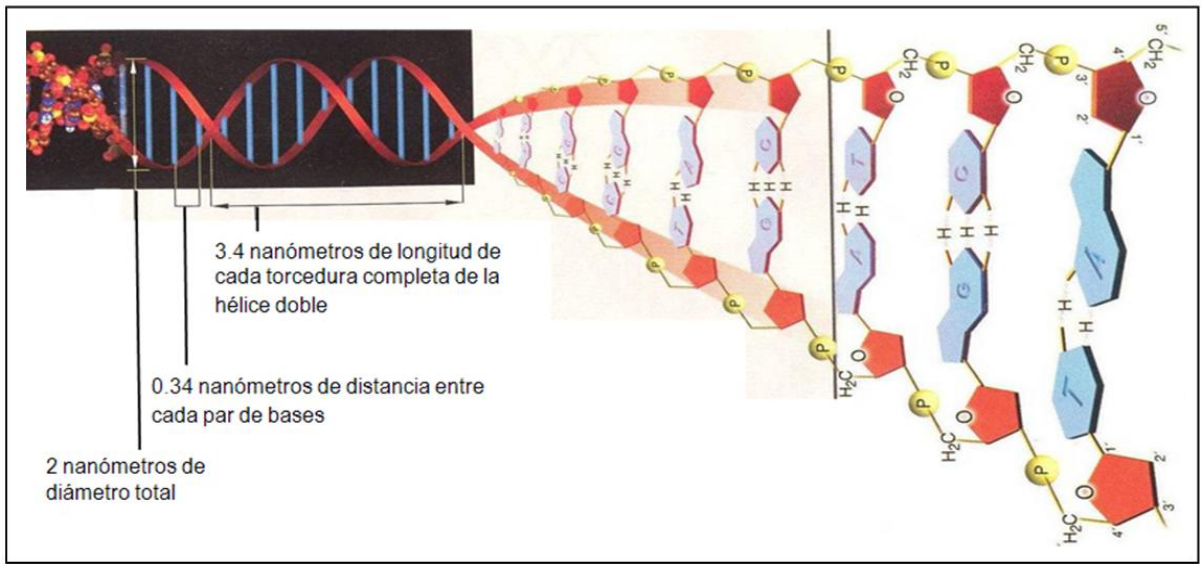


Figura 2.3.1: Representación de la cadena doble de *ADN*. Tomado de texto *Biology The Unity and Diversity of life*, séptima edición, página 211.

2.3.2 Ácido Ribonucleico (*ARN*)

Además de contener información propia, constituye la herramienta de conversión de la información contenida en el *ADN* en proteínas específicas. Se encuentra abundantemente, tanto en el nucleolo como en el citoplasma de las células. Existen varios tipos de *ARN*: *ARN* mensajero (*ARNm*), que porta y transporta la información genética del *ADN* cromosomal a los ribosomas; *ARN* transferente (*ARNt*), que transporta los aminoácidos activados y aptos desde el citosol al lugar de síntesis de proteínas; *ARN* ribosómico (*ARNr*), que integra junto con proteínas la partícula conocida como ribosoma; y *ARN* pequeño nuclear (*ARNsn*), implicado en el proceso de maduración del *ARN* en las células eucarióticas.

El *ARN* difiere del *ADN* en que la pentosa de los nucleótidos constituyentes es ribosa en lugar de desoxirribosa, y en que en lugar de las cuatro bases *A*, *G*, *C*, *T* aparece *A*, *G*, *C*, *U* es decir, uracilo en lugar de timina. Mientras que el *ADN* y el *ARN* contiene la información escrita en un lenguaje de 4 letras (las cuatro bases) llamado código genético, el ribosoma traduce dicha información usando el otro lenguaje, al sintetizar la proteína, pasando del lenguaje de 4 letras a otro con 20 letras (los 20 aminoácidos que conforman las proteínas).

2.4 Genoma Humano

Se denomina genoma al conjunto de todos los cromosomas de una especie. El Proyecto del Genoma Humano que fundado en 1990 y su interés era determinar las posiciones relativas de todos los nucleótidos e identificar los 20,000 a 25,000 genes presentes en el genoma humano. En el año 2003 se presentó un borrador inicial del genoma humano. El Genoma Humano está constituido por 23 pares de cromosomas distintos, de los cuáles 22 pares son autosomas procedentes de cada progenitor y un par son cromosomas sexuales. Tiene un tamaño total aproximado de 3200 millones de pares de bases de *ADN* que contienen unos 20,000-25,000 genes.

La finalidad del genoma humano no representa sólo una lista de todos los genes, sino también, la búsqueda de respuestas a interrogantes sobre la función, regulación e interacción de esos genes en procesos celulares; los cambios en la expresión de esos genes debido a enfermedad o tratamiento, el funcionamiento de las células o su intervención en la formación de organismos.

2.5 Gen

Un gen es una porción de *ADN* que contiene la información necesaria para la síntesis de una macromolécula con función celular específica, y que determina la herencia de una o varias características determinadas. Los genes se disponen a lo largo de cada cromosoma, ocupando una posición específica en ellos llamada locus.

2.5.1 Expresión Genética

Un gen se expresa a través de la síntesis de su correspondiente *ARN* y éste, a su vez, se expresa a través de su correspondiente proteína. Es decir, la expresión de un gen, es el proceso por el cual el gen se activa en una célula para elaborar *ARN* y proteínas. El nivel de expresión de un gen es entonces, la cantidad de copias de *ARNm* transcriptos presentes en la célula en un determinado momento y puede medirse al observar el *ARN* o la proteína elaborada con el *ARN*, o lo que la proteína hace en una célula.

Las diferencias morfológicas y fenotípicas, de salud y enfermedad, en un individuo se deben a que sólo una parte de los genes de cada célula están expresados. Esto hace necesario medir y analizar el nivel de expresión genética. En los últimos años la técnica de microarreglos de *ADN* se ha convertido en la herramienta universalmente más utilizada para este tipo de análisis.

Para más detalles de los conceptos de biología citados, ver *Gonick* y *Wheelis* [16] y *Griffiths* [17].

2.6 Dogma Central de la Biología Molecular

El Dogma Central de la Biología Molecular, establece que una porción del *ADN* del cromosoma se copia (transcripción) a una cadena simple de *ARNm*, el cual sale del núcleo hacia el citoplasma llevando consigo la información necesaria para que en el ribosoma se sintetice y/o codifique (traducción) determinada proteína. Al proceso de copiado de la información genética contenida en el *ADN* cromosomal durante la síntesis del *ARNm* se le llama transcripción, en él se sintetiza en forma complementaria una secuencia de *ARN* mensajero usando como molde el *DNA* original. Así, la secuencia de bases de *ARN* es igual a la hebra de *ADN* estabilizadora pero cambiando las *T's* por las *U's*. Al proceso de lectura de la información transportada por *ARNm*, en el ribosoma, durante la síntesis de proteína, se le conoce como traducción.

El Dogma de la Biología Molecular, llamado así por Francis Harry Compton Crick en 1970, establece que la información fluye de manera unidireccional, una vez en las proteínas no

pueden regresar al *ADN*. Es decir, ya en las proteínas, éstas no pueden ser cambiadas o, lo que es lo mismo, las proteínas no pueden influir los genes. Sin embargo, esto no es del todo cierto, pues, en el mismo año Temin y Baltimore demostraron que una enzima que se encuentra en los retrovirus, la transcriptasa reversa, permite que el *ARN* se copie al *ADN*.

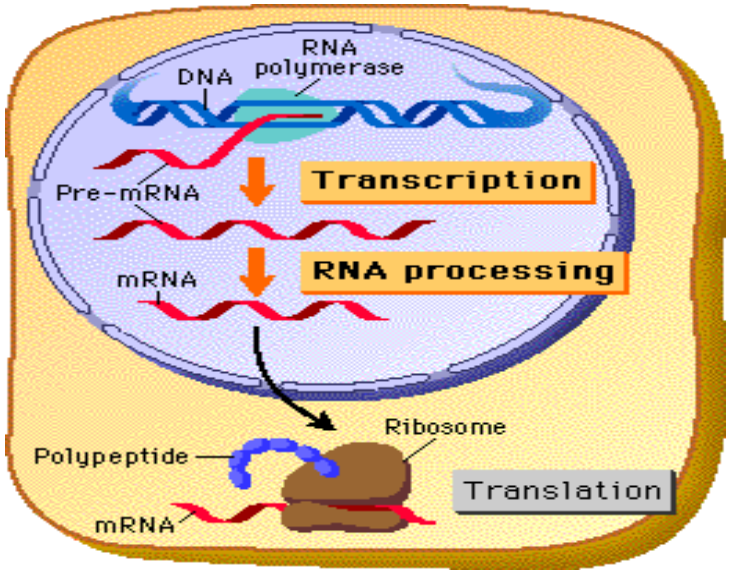


Figura 2.6: Representación del Dogma Central de la Biología Molecular. Tomada de http://biplot.usal.es

2.7 Microarreglos

Un Microarreglo es un soporte sólido en el cual se encuentran inmovilizadas una serie ordenada de muestras (*ADN*, *ARN*, proteína o tejido fino). El soporte sólido puede ser una laminilla de vidrio para microscopio, laminilla de silicón o membrana de nilón. Y el tipo de microarreglo depende del material puesto sobre la laminilla: microarreglo de ácido nucleico, si es un ácido nucleico; microarreglos de proteína, si es proteína; microarreglos de tejido (*TMA*), si es tejido.

El apareamiento *A-T* y *C-G* que se mantiene por los puentes hidrógeno de las cadenas opuestas del *ADN* y que representa la propiedad de complementariedad de las bases que

componen el *ADN*, es la base molecular de los microarreglos; donde hebras de *ADN* de diferente origen van a hibridizarse por complementariedad de las bases que las componen para identificar y medir las secuencias de *ARN* que están presentes en una muestra.

Un microarreglo proporciona una visión instantánea de la actividad de genomas completos o de un grupo selecto de genes, debido a que permite el depósito de miles de puntos conteniendo genes o parte de genes sobre un portaobjetos para su estudio en paralelo. Cada secuencia en un microarreglo se asocia con un único gen y cada punto contiene millones de secuencias clonadas "idénticas" de una secuencia específica conocida, asociada a un gen.

2.7.1 Conceptos Preliminares Relacionados con Microarreglos

Sondas o probes: Muestra de material genético inmovilizado, compuesto de ser secuencias de ácidos nucleicos de hebras sencillas que pueden ser complementarias a genes.

Target, blanco u objetivo: Secuencia determinada *ARNm* que interesa estudiar en un tejido; están marcadas con material fluorescente (biotina o cianina: Cy3, Cy5).

Hibridación: Proceso por el cual se forman cadenas dobles entre las sondas del microarreglo y los blancos.

Desnaturalización: Proceso inverso a de hibridizar.

Spot: Es cada uno de los puntos de la sonda inmovilizada en el microarreglo.

Oligonucleótidos: Son secuencias cortas de nucleótidos de *ADN* o *ARN*, que tienen 20 o menos bases o pares de bases. Si las secuencias son de 50-70 nucleótidos se habla de oligonucleótidos largos.

2.7.2 Tipos de Microarreglos de *ADN*

Los diferentes tipos de microarreglos que existen dependen de las clases de sondas utilizadas para su diseño. Dos de ellas son genómicas, una detecta la pérdida o ganancia de genes (usa grandes cantidades de *ADN* genómico) y la otra detecta mutaciones en el *ADN*

(emplea un único gen); las otras son transcriptómicas, en ellas el *ADN* inmovilizado por hibridación es *ADNc* producto de *ARNm* de genes conocidos, y miden los niveles de *ARNm*.

Los distintos tipos de microarreglos que hay son: los microarreglos de ácidos nucleicos (de *ADN* y Oligonucleótidos), de proteínas, de tejidos (*TMA*) y *Comparative Genomic Hybridization* (*CGH*). A continuación sólo describimos los microarreglos de ácidos nucleicos, puesto que son los de nuestro interés.

Microarreglos de *ADN*

Son los comúnmente más usados porque permiten depositar genes o fragmentos de genes amplificados de cualquier especie y así diseñar y generar de manera sencilla y menos costosa el grupo de sondas. En ellos, las sondas impresas sobre la laminilla pueden ser oligonucleótidos largos (100-3000 nucleótidos) químicamente sintetizados o productos enzimáticos generados por *PCR* (amplificación posterior de la cadena por la reacción de la polimerasa).

Este tipo de microarreglo también es llamado microarreglos punteados (*Spotted microarrays*) y son empleados para detectar cambios en la expresión genética durante un período de tiempo. Por esta razón constituyen una importante herramienta en la investigación en cáncer, ya que se podrían identificar por este métodos nuevos marcadores cancerígenos con propósitos diagnósticos, aunque cabe señalar que no permite hacer una comparación lo suficientemente exacta entre experimentos diferentes.

Microarreglos de Oligonucleótidos

Se conocen también como Arreglos de un solo Canal. Las sondas que emplean son porciones de *ADN* sintético de cadena simple que pueden ser cortas (15-25 nucleótidos) o largas (50-120 nucleótidos). Estos fragmentos pueden ser presintetizados y depositados en portaobjetos por robots o sintetizados *in situ* y depositados por *ink jet* o fotolitografía (*ADNchips*). Esta última técnica, fue ideada por D. *Lockhard* y *Lipshutz* ([28], [29]) y desarrollada por la compañía de biotecnología californiana *Affymetrix*; y emplea 11 cadenas cortas pertenecientes al mismo gen para asegurarse de que se está midiendo el gen correcto.

Este tipo de microarreglos proporciona estimaciones sobre el nivel de expresión, pero permite observar solo una condición por arreglo. Se producen comercialmente por las compañías *Affymetrix* y *Aiglent* Technologies, los más populares son los construidos por *Affymetrix* y se conocen como *GeneChips*. Actualmente se usan oligonucleótidos de longitud corta, ya que su alta precisión, densidad y reproductibilidad permite hacer comparaciones precisas de las señales generadas usando muestras hibridizadas en diferentes arreglos.

2.7.3 Importancia de los Microarreglos

La importancia de los microarreglos radica en que, sin establecer ninguna hipótesis previa, permiten medir la expresión de uno o varios genes durante la enfermedad, los cambios en los tejidos tumorales, la expresión de marcadores relacionados con factores pronósticos de enfermedad o toxicidad, y los polimorfismos de genes, entre otras aplicaciones. Es decir, detectan en cualquier momento o cualquier condición establecida, la cantidad de *ARNm* que fabrica cada gen, si es que lo fabrica. Esto lo hace interrumpiendo el proceso natural establecido por el Dogma Central de la Biología Molecular, al extraer la porción de *ADN* cromosomal que ha sido transcrito a una cadena simple de *ARNm* maduro en uno o más tejidos, para hibridizarlo con el material previamente depositado en el microarreglo.

2.7.4 Experimento con Microarreglos

Se puede decir que un experimento de microarreglos se desarrolla en tres etapas fundamentales, ver Figura 2.7.4. Dada un interrogante biológico, la primera etapa consiste en seleccionar el tipo y la cantidad de sondas que se fijaran en la laminilla, comprobando que las cadenas de *ADN* que las conforman no estén presentes en ningún otro gen en cuestión y no otro, de modo, que cuando una molécula de *ARN* se una al blanco los científicos sepan que el gen se ha expresado. El número de sondas a inmovilizar depende del método de fabricación utilizado.

La segunda etapa consiste en la extracción y marcado del *ARNm* de cada una de las muestras provenientes de dos o más condiciones a comparar, por ejemplo, las de un tejido sano (control) y las de un tejido enfermo (tumor). Los procesos que se siguen en esta etapa son los de extracción y purificación del material a analizar; en caso de tratarse de material genético sería un proceso de amplificación. Después se realiza la transcripción inversa para obtener el *ADNc* y se etiqueta la muestra para poder detectarla en el proceso de revelado; los marcadores más usados son los fluorescentes pero también puede utilizarse marcaje radiactivo o quimioluminiscencia. Convencionalmente se marca con dos colorantes derivados de cianinas, Cy3 (verde) para el control y Cy5 (rojo) para el tumor. Por último, se hibridizan los *ADNc'*s marcados en cantidades iguales en el mismo microarreglo, con sus complementarias inmovilizadas en la superficie del microarreglo.

La tercera y última etapa, consiste en el análisis de imágenes y extracción de datos. Debido a que los datos proporcionados por un experimento de microarreglos son intensidades de fluorescencia, es decir, imágenes; para su lectura, se emplean isótopos radiactivos, escáneres láser o cámaras *CCD* que detectan marcadores fluorescentes y que proporcionan datos constituidos por archivos *tiff* de las imágenes digitales escaneadas, una para cada color (rojo y verde), que usualmente son superpuestas y comprimidas en una sola imagen. Los puntos que se ven rojos o en tonalidades naranja, son genes que aumentaron su expresión; los puntos verdes o de tonalidades entre verde y amarillo, son genes que disminuyeron su expresión y los puntos amarillos son genes que ambas condiciones se expresan igual.

Una vez obtenidos los archivos de imágenes, éstos se transforman a cantidades numéricas, considerando las dimensiones y la forma, la localización y los parámetros estadísticos que pueden asociarse a cada punto analizado de la imagen; para ello, se hace uso de software, algunos de ellos gratuitos, que permiten analizar este tipo de datos. De esta manera se obtiene una matriz de expresión donde las filas son las muestras y las columnas genes. Finalmente, se normalizan y transforman los datos para disminuir las variaciones y facilitar los cálculos posteriores.

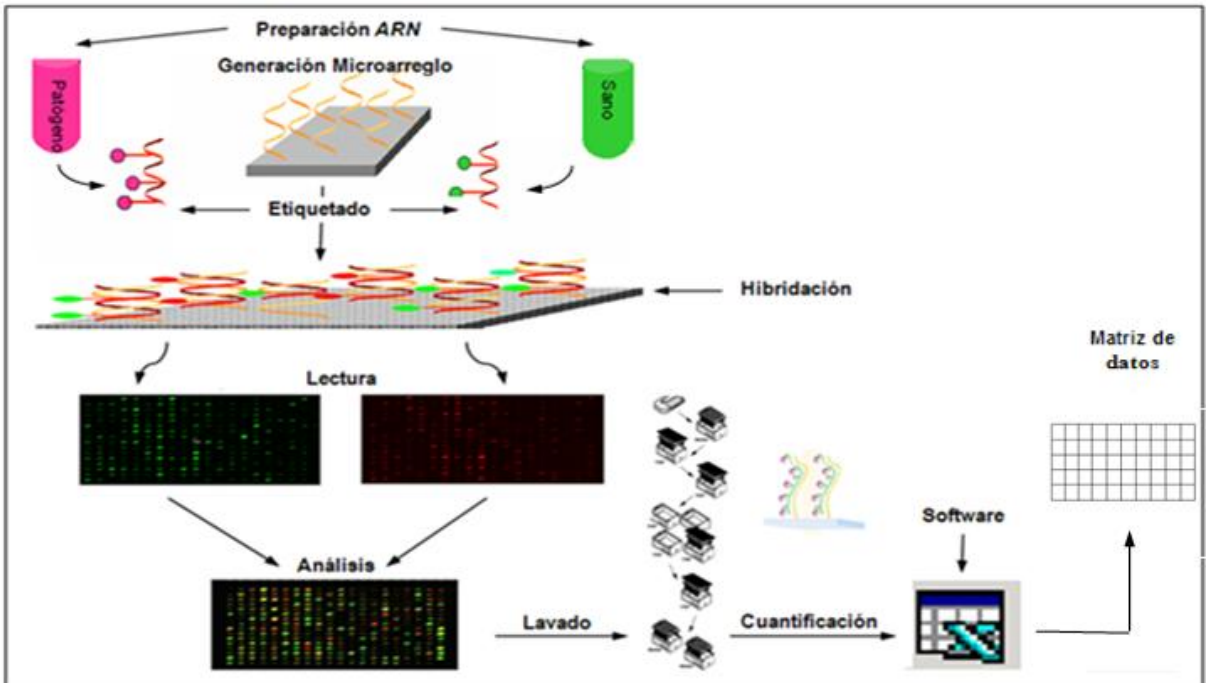


Figura 2.7.4: Experimento de Microarreglos.

2.8 Datos de Expresión Genética

Recordemos que un Microarreglo es un arreglo ordenado de forma rectangular, que contiene puntos que representan las intensidades de fluorescencia de cada gen; para el cual, el análisis de Microarreglos proporciona una matriz de datos. Esta matriz, es una matriz de alta dimensionalidad llamada matriz de expresión genética, donde un gran número de columnas representan los genes (variables o características) y pocas filas representan los arreglos (muestras). Convencionalmente, no es a esta matriz a la que se aplican los métodos estadísticos, antes hay que normalizarla, y transformarla si es necesario, para evitar variaciones que pueden llevar a errores considerables. Además, generalmente se trabaja con la transpuesta de la matriz de expresión.

Nosotros consideraremos específicamente una matriz de expresión denotada por $X = (x_{ji})$, cuyas filas corresponden a los genes y sus columnas a las muestras, es decir, una matriz de dimensión $m \times n$ que almacena datos de expresión para m genes y n muestras de *ARNm*. Supondremos además, que una variable de respuesta o covariable de interés es registrada

para cada muestra, de modo que los datos para una muestra i consisten de una variable de respuesta o covariable y_i y un perfil de la expresión del gen $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{mi})$, donde x_{ji} es generalmente una variable continua que denota la medida de expresión del gen j en la muestra i, para lo cual $i = 1,2,\ldots,n$, y $j = 1,2,\ldots,m$ y entendiéndose, por variable de respuesta algún resultado clínico o patológico (para nuestro caso: el gen es diferencialmente expresado) que puede asumir incluso más de dos valores, y por covariable un estado (tratamiento/control), tipo de célula o droga, dosis de una droga o un tiempo particular.

Sea X_j y Y que denotan, respectivamente, la variable aleatoria que corresponde a la medida de expresión del gen j, $j = 1,2,\ldots,m$, y la variable de respuesta; consideraremos los pares $\{(\boldsymbol{x}_i\, y_i)\}_{i=1,\ldots,n}$ formados por los perfiles de la expresión de los genes $\boldsymbol{x}_i$ y la variable de respuesta y_i, como una muestra aleatoria para la variable aleatoria (X_j , Y). Nuestro objetivo es usar esta muestra aleatoria para hacer inferencia sobre la asociación entre la medida de expresión de los genes X_j y la respuesta Y , es decir, identificar genes diferencialmente expresados.

2.9 Identificación de Genes Diferencialmente Expresados

Como ya hemos mencionado, una de las metas de mayor interés en el análisis de datos de expresión genética en experimentos de microarreglos es la identificación de genes diferencialmente expresados (activos) para múltiples arreglos, es decir, la identificación de genes cuyos niveles de expresión se asocian con una variable de respuesta o covariable de interés. El objetivo científico es identificar genes diferencialmente expresados en tejidos enfermos y compararlos con los genes correspondientes en un tejido del mismo tipo sano. Inicialmente, deben identificarse dentro de los resultados del experimento los genes que presentan cambios significativos, lo cual puede hacerse en dos etapas, la elección del estadístico y la determinación de un punto de corte. La primera etapa, consiste en elegir un estadístico y permite ordenar la evidencia de expresión diferenciada, desde la mayor a la menor evidencia. La segunda etapa consiste en elegir un valor crítico para el ordenamiento anterior por encima del cual cualquier valor resulta significativo, definiendo así la cantidad de genes que se consideran diferencialmente expresados.

Capítulo 3
PRUEBAS MÚLTIPLES EN EXPERIMENTOS DE MICROARREGLOS

Como se discutió en el capítulo anterior, el análisis de los microarreglos permite realizar avances rápidos en el conocimiento de diversas patologías desde el punto de vista de la biología molecular. Uno de tales avances, es la identificación de genes diferencialmente expresados para múltiples arreglos; para lo cual han sido sugeridos varios métodos, tanto para experimentos de microarreglos de un solo arreglo como para experimentos de microarreglos repetidos. Para los experimentos de microarreglos que consisten de un solo arreglo se distinguen dos tipos de métodos principales: aquellos que solo consideran el valor de la razón de la expresión y los que además toman en cuenta la abundancia de transcripción total.

La identificación de genes diferencialmente expresados en experimentos de microarreglos repetidos o de múltiples arreglos, se puede plantear como un problema de prueba de múltiples hipótesis, donde para cada gen se prueba simultáneamente las hipótesis nulas de no asociación entre los niveles de expresión y las variables de respuesta. Estas pruebas pueden hacerse de dos formas: una, considerando una estructura de no dependencia entre los genes la cual es una situación poco real pero útil si lo que se desea es hacer afirmaciones sobre una población particular de células y en casos en los cuales se quieren hacer afirmaciones que se cumplen en presencia de variabilidad biológica; otra, teniendo presente la estructura de dependencia existente entre los genes y los arreglos e incluso la interacción entre la condición de interés y el arreglo.

3.1 Prueba de Hipótesis Simple

Una Hipótesis Estadística es una afirmación que se hace acerca de un parámetro poblacional. Tal afirmación suele ser llamada Hipótesis Nula y se representa por H_o, cuando se refiere a una afirmación que está establecida y que se espera sea rechazada después de aplicar una prueba estadística. La Hipótesis Alterna, la cual se representa por H_a y se refiere a la afirmación que se espera sea aceptada después de aplicar la prueba estadística.

Una prueba de hipótesis es una regla de decisión basada en la distribución del estimador del parámetro que aparece en la hipótesis; su objetivo es decidir cuál de las dos hipótesis complementarias, H_o y H_a, es verdadera, basada en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de la población. Al subconjunto del espacio muestral para el cual H_o es rechazada se le llama Región de rechazo o región crítica, a la región complemento se le llama región de aceptación.

Al realizar una prueba de hipótesis pueden cometerse dos tipos de errores; el Error Tipo I, que consiste en rechazar H_o cuando es verdadera, y el Error Tipo II que se comete al no rechazar H_o cuando es falsa, llamados respectivamente en este contexto, falso positivo y falso negativo; los cuales para cada prueba tienen una probabilidad de ocurrencia específica. La probabilidad de cometer el Error Tipo I, también interpretada como el tamaño de la región crítica, es llamada nivel de significancia y se representa por α. La probabilidad de cometer el Error Tipo II es representada por β y al valor 1-β, que representa la probabilidad de rechazar H_o cuando es falsa, se le llama Potencia de la prueba. Una buena prueba estadística es aquella que tiene una alta potencia de prueba. Lo que generalmente se hace al considerar una sola prueba de hipótesis es pre-establecer un nivel de significancia α aceptable para la tasa del Error Tipo I y buscar pruebas que minimicen la tasa de Error Tipo II , es decir, maximizar la potencia dentro de la clase de pruebas con tasa de Error Tipo I de a lo más α.

La decisión es tomada usando el método clásico o el moderno. El primero evalúa la prueba estadística y define de antemano el nivel de significancia α, para predeterminar una región de rechazo y una de aceptación; si el valor obtenido de la prueba estadística cae en la región de rechazo, entonces se concluye que la muestra tomada proporciona suficiente evidencia estadística para rechazar la hipótesis nula, de lo contrario H_o no es rechazada. El método

moderno, consiste en usar el *p-value*, llamado también nivel de significancia observado, el cual es el valor de α al cual se rechazaría la hipótesis nula si se usa el valor calculado de la prueba estadística. En la práctica un *p-value* cercano a cero indica un rechazo de la hipótesis nula. El cálculo del *p-value* depende de cómo este establecida la hipótesis alterna. Por ejemplo, sea μ la media poblacional, si H_o: $\mu = \mu_o$ y,

i) H_a: $\mu > \mu_o$, entonces *p-value* = Pr ($\xi >$ *valor del estadístico de prueba*).

ii) H_a: $\mu < \mu_o$, entonces *p-value* = Pr ($\xi <$ *valor del estadístico de prueba*).

iii) H_a: $\mu \neq \mu_o$, entonces *p-value* = 2Pr ($\xi >$ | *valor del estadístico de prueba* |).

3.2 Prueba de Hipótesis Múltiple en Experimentos de Microarreglos

Al hablar de prueba de Hipótesis Múltiple, se asume que se quiere probar simultáneamente algún conjunto de hipótesis nulas de interés. Tal conjunto, suele estar constituido por un gran número de hipótesis que en ocasiones puede ser infinito, aunque generalmente en la práctica es finito. Esta situación de considerar varias pruebas de hipótesis se presenta, por ejemplo, al comparar varios tratamientos, procesos, variedades o localizaciones diferentes, cuando se desea probar si estas diferencias tienen algún efecto en el resultado.

Dentro del marco de microarreglos, consideraremos específicamente una matriz de expresión $X = (x_{ji})$, de dimensión $m \times n$, cuyas filas corresponden a los genes y sus columnas a las muestras. Para $X_j = (x_{j1}, \ldots, x_{jn})$ y Y que denotan, las variables aleatorias corresponden a la medida de expresión del gen j y la variable de respuesta, respectivamente; los pares $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,n}$ son considerados como una muestra aleatoria del vector aleatorio (X_j, Y).

La pregunta biológica sobre los genes diferencialmente expresados puede ser expuesta como un problema de prueba de hipótesis múltiple que establece simultáneamente para cada gen j la hipótesis nula, "H_j: no existe asociación entre X_j y Y ", j =1,2,…,m. Estas pruebas de hipótesis múltiples permitirán establecer aquellos genes con expresión diferenciada; pues, el que H_j sea rechazada significa que el gen j es diferencialmente expresado. Aquí, el interés es probar las m hipótesis H_j's simultáneamente; puesto, que si cada H_j se prueba

separadamente, tal situación constituye un problema de hipótesis simple y recordemos que en ese contexto lo que se hace es pre-establecer un nivel aceptable α para la tasa del Error Tipo I y buscar pruebas que minimicen la tasa de Error Tipo II. En cuanto al caso múltiple lo que se busca es controlar la tasa de Error Tipo I, eligiendo y calculando a partir de los datos el estadístico de prueba para cada gen *j*, y luego, se aplica algún método de prueba múltiple para determinar cuales hipótesis nulas se rechazarán. Cabe señalar que el estadístico de prueba apropiado para cada gen depende de H_j, del tipo de control sobre la tasa de Error Tipo I, del diseño de experimento y del tipo de respuesta.

3.3 Error Tipo I y Tipo II

Sabemos que en cualquier situación de prueba de hipótesis pueden cometerse dos tipos de error, Error Tipo I y Error Tipo II, llamados en este contexto Falso Positivo y Falso Negativo, respectivamente. Consideremos el problema de probar simultáneamente *m* hipótesis nulas H_j, denotando por *R* el número de hipótesis rechazadas, y por m_0 y $m_1 = m - m_0$ el número de hipótesis verdaderas y falsas, respectivamente. Esta situación es resumida en la Tabla 3.3.

	Hipótesis no rechazadas	**Hipótesis rechazadas**	
Hipótesis nulas verdaderas	*U*	*V:# Falsos Positivos* Error Tipo I	m_0
Hipótesis nulas falsas	*T: # Falsos Negativos* Error Tipo II	*S*	m_1
	$m - R$	*R*	*M*

Tabla 3.3: Error tipo I y Error Tipo II.

Aquí, *R* es una variable aleatoria observable; m_0 y m_1 son parámetros desconocidos y *S*, *T*, *U* y *V* son variables aleatorias no observables. Lo ideal es minimizar el número *V* de falsos positivos o Error Tipo I y el número *T* de falsos negativos o Error Tipo II.

3.3.1 Tasas de Error Tipo I

Cuando se prueban múltiples hipótesis, existen muchas posibles definiciones para la tasa de Error Tipo I y la potencia de un método de prueba; a continuación establecemos las definiciones más estándar [39].

The per-comparison error rate *(PCER):* Es el valor esperado del número V de falsos positivos dividido por el número de hipótesis, esto es, $PCER = E(V) / m$.

The per-family error rate *(PFER):* Es el número esperado de falsos positivos, esto es, $PFER = E(V)$.

The family-wise error rate *(FWER*): Es la probabilidad de cometer al menos un Error Tipo I, esto es, $FWER = \Pr(V > 0)$.

The false discovery rate *(FDR):* Introducido por Benjamini y Hochberg (1995), es la proporción esperada de falsos positivos entre las hipótesis rechazadas, esto es,

$$FDR = \begin{cases} E\left(\dfrac{V}{R}\right), \ si \quad R > 0 \\ 0, \ si \quad R = 0 \end{cases} \quad \text{ó} \quad FDR = E\left(\frac{V}{R} I_{\{R>0\}}\right),$$

donde $I_{\{R>0\}}$ es la función indicadora, la cual es 1 cuando $R > 0$ y cero en cualquier otro caso.

Ahora bien, por definición de R y V se tiene que $0 \le V \le R \le m$, y que $R = 0$ implica $V = 0$, entonces

$$\frac{V}{m} \le \frac{V}{R} I_{\{R>0\}} \le I_{\{V>0\}} \le V ,$$

luego al aplicar valor esperado tenemos

$$E\left(\frac{V}{m}\right) \le E\left(\frac{V}{R} I_{\{R>0\}}\right) \le E(I_{\{V>0\}}) \le E(V) ,$$

y como

$$E\left(\frac{V}{m}\right) = \frac{E(V)}{m} \qquad \text{y} \qquad E(I_{\{V>0\}}) = \int_{\{V>0\}} f_V(v)\, dv = \Pr(V > 0) ,$$

por definición de las distintas tasas de error, se tiene entonces

$$PCER \leq FDR \leq FWER \leq PFER.$$

Si $m = m_0$, entonces el número total de hipótesis coincide con el número de hipótesis verdaderas y por tanto no se rechazaría ni aceptaría ninguna hipótesis falsa. En este caso $S = T = 0$ y $R = V$. En consecuencia,

$$FDR = E\left(\frac{V}{R} I_{\{R>0\}}\right) = E(I_{\{V>0\}}) = \Pr(V > 0) = FWER. \qquad (3.3.1)$$

Por lo que se establece, que si $m = m_0$ entonces $FDR = FWER$.

3.3.2 Control de la Tasa de Error Tipo I

Debido a que en este caso se están probando miles de hipótesis nulas simultáneamente, la probabilidad de cometer un Error Tipo I aumenta con el número de hipótesis y un *p-value* cercano a cero no garantiza que el gen sea diferencialmente expresado, ya que tal resultado puede ser producto del azar debido al gran número de hipótesis que se prueban. Lo que se pretende entonces, es controlar la tasa de Error Tipo I a un nivel α.

Se dice que un método de prueba múltiple establece el control de una tasa particular de Error Tipo I a un nivel α, si la tasa de error es menor o igual a α al aplicar un método dado para producir una lista de R hipótesis rechazadas. En particular, cada una de las tasas de Error Tipo I definidas, son controladas a un nivel α, para un método de prueba múltiple específico si: $PCER \leq \alpha$, $FDR \leq \alpha$, $FWER \leq \alpha$ y $PFER \leq \alpha$.

Algunos de los métodos que controlan la tasa de Error Tipo I lo hacen solo bajo algunas condiciones, hecho al cual deben su nombre, entre ellos están, el control exacto, el débil y el fuerte.

Control Exacto: Se refiere al control de la tasa Error Tipo I, cuando únicamente las m_0 hipótesis nulas son verdaderas.

Control Débil: Se refiere al control de la tasa de Error Tipo I solamente, cuando todas las m hipótesis nulas son verdaderas, es decir, cuando $m = m_0$, o bajo la que se conoce también como la hipótesis nula completa y que se representa por $H_0^C = \bigcap_{j=1}^{m} \{H_j : \text{es verdadera}\}$. Pero esta hipótesis nula completa no es realista y el control obtenido bajo ella no es satisfactorio, puesto que no habría genes diferencialmente expresados.

Control Fuerte: Se refiere al control de la tasa de Error Tipo I bajo cualquier combinación de hipótesis nulas verdaderas y falsas, es decir, para cualquier subconjunto $\kappa \subseteq \{1,\ldots,m\}$ de hipótesis nulas verdaderas; este tipo de control establece algunos genes como diferencialmente expresados. En la práctica, algunas κ hipótesis nulas son verdaderas y otras falsas, pero κ es desconocido.

El control fuerte implica control exacto y control débil, pero de ningún modo el control débil y control exacto implican control fuerte.

En adelante a menos que se especifique lo contario, estamos interesados en el control fuerte de la tasa de Error Tipo I.

3.3.3 Control de la Tasa de Error Tipo II

Al igual que para la tasa de Error Tipo I en pruebas de hipótesis múltiples, y con el fin de maximizar la potencia, es decir, minimizar convenientemente la tasa de Error Tipo II, el concepto de Potencia puede ser generalizado de varias formas:

Potencia:

- Probabilidad de rechazar al menos una hipótesis nula falsa, esto es,

$$\Pr(S \geq 1) = \Pr(T \leq m_1 - 1).$$

- Probabilidad media de rechazar la hipótesis nula falsa, también llamada Potencia Media, esto es,

$$E(S)/\, m_1.$$

- Probabilidad de rechazar todas las hipótesis nulas falsas, esto es,

$$\Pr(S = m_1) = \Pr(T = 0)$$

- Proporción esperada del número de hipótesis falsas rechazadas entre las hipótesis rechazadas, esto es,

$$\begin{cases} E\left(\dfrac{S}{R}\right), \ si \quad R > 0 \\ \quad 0, \ si \quad R = 0 \end{cases} \quad \text{ó} \quad E\left(\frac{S}{R} I_{\{R>0\}}\right).$$

3.4 Estadísticos de Prueba

Para probar cualquier hipótesis o conjunto de hipótesis estadísticas, debe emplearse un estadístico de prueba, el cual depende de la distribución del estimador del parámetro que aparece en la hipótesis y se basa en una muestra aleatoria de la población. En nuestro caso, el estadístico de prueba apropiado para cada gen depende del diseño de experimento y del tipo de respuesta, y por supuesto del tipo de control sobre la tasa de Error Tipo I. Para aquellas situaciones representadas por variables de respuesta binarias, como son los casos donde se comparan las expresiones de los genes en dos grupos, control y tratamiento, puede considerarse el estadístico *t* o el estadístico *Mann-Whitney*; para variables de respuestas que asumen más de dos valores, puede emplearse el estadístico *F* o la prueba de *Kruskal-Wallis* y; para datos de sobrevivencia, puede usarse el valor del estadístico para el modelo de riesgo proporcional *Cox*.

La elección del estadístico de prueba permite ordenar la evidencia de expresión diferenciada, desde la mayor a la menor evidencia, por ello el interés en que sea el adecuado.

3.4.1 Prueba *t* para comparar dos medias poblacionales usando muestras independientes cuando las varianzas poblacionales son desconocidas e iguales

Supóngase que se tienen dos poblaciones distribuidas normalmente con medias desconocidas μ_1 y μ_2, respectivamente. Basándonos en dos muestras independientes tomadas de ellas, una de tamaño n_1, con media $\bar{x}_1$ y varianza s_1^2 y otra muestra de tamaño n_2, con media $\bar{x}_2$ y varianza s_2^2. Estableciendo las hipótesis complementarias:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_a &: \mu_1 \neq \mu_2 \end{aligned} \tag{3.4.1}$$

Si las varianzas poblacionales son desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), entonces el estadístico de prueba que se emplea es:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.4.2}$$

El cual tiene distribución *t* con $n_1 + n_2 - 2$ grados de libertad. Donde

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

es un estimador de la varianza poblacional común σ^2.

Una vez establecido el nivel de significancia α, se rechaza la hipótesis nula si $| t | > t_{\alpha/2}$; *t* está dado por (3.4.2) y $t_{\alpha/2}$ representa el valor de la distribución *t* de *Student* tal que el área a la derecha de dicho valor es $\alpha/2$. O también, se rechaza H_0 cuando el *p-value* de la prueba es menor que α (*p-value* = 2 Pr($| t |$ > *valor del estadístico de prueba*) < α).

3.4.2 Prueba *t* para comparar dos medias poblacionales usando muestras independientes cuando las varianzas poblacionales son desconocidas y diferentes

En este caso se asume también normalidad de las poblaciones, y las hipótesis establecidas en (3.4.1) se prueban usando la prueba aproximada de *t*, dada por

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}} \tag{3.4.3}$$

La cual tiene distribución *t* con grados de libertad *gl* estimados por la siguiente aproximación:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}.$$

En este caso la regla de decisión para rechazar H_0 es similar a la del caso anterior.

3.4.3 Prueba *Mann-Whitney* para comparar dos muestras independientes

También conocida como prueba de suma de rangos de *Wilcoxon*; es una prueba no paramétrica alterna a la prueba de *t*, que no presupone normalidad de las poblaciones. Para esta prueba como en todas las pruebas no paramétricas, las hipótesis complementarias dependen de la mediana y no de la media; H_0 establece que la mediana de las dos poblaciones son iguales y H_a establece que la mediana de una población es distinta (mayor o menor) de la mediana de la otra población.

Cuando n_1 y n_2 son mayores que 10 y si no hay empates, se puede probar que el estadístico *W* de *Wilcoxon*, que consiste en la suma de todos los rangos de las diferencias positivas entre los valores observados y la mediana; se distribuye aproximadamente como una normal con media $n_1 \cdot (n_1+n_2+1)/2$ y varianza $n_1 \cdot n_2 \cdot (n_1+n_2+1)/12$. Por lo tanto,

$$Z = \frac{W - \frac{n_1 \cdot (n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} \sim N(0,1) \qquad (3.4.4)$$

Cuando hay empates la varianza es modificada y se obtiene:

$$Z = \frac{W - \frac{n_1 \cdot (n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12} - \sum_{i=1}^{g} \frac{t_i^3 - t_i}{(n_1 + n_2)(n_1 + n_2 - 1)}}} \sim N(0,1), \qquad (3.4.5)$$

donde g es el número de grupos empatados y t_i es el tamaño del i-ésimo grupo empatado.

Una vez establecido α, la prueba *Mann-Whitney* rechaza la hipótesis nula cuando $| Z | > Z_{\alpha/2}$ o cuando el *p-value* = 2 Pr($| Z | >$ *valor del estadístico de prueba*) $< \alpha$; Z es el valor dado por (3.4.4) o (3.4.5).

3.4.4 Prueba *F* para comparar varias medias poblacionales

Cuando se dispone de más de dos grupos, la prueba estadística a emplear es la prueba *F* de *Fisher-Snedecor* obtenida por un análisis de la varianza (*ANOVA*). Esta situación surge, por ejemplo, al comparar diferentes tratamientos, procesos o variedades, cuando el interés se centra en probar si estas diferencias tienen algún efecto sobre los promedios de los resultados.

A menos que se especifique lo contrario, en adelante consideraremos un análisis de varianza para un solo factor, para el cual se consideran n grupos clasificados de acuerdo a los niveles 1, 2,…, k del factor (clase) y en cada nivel se tienen n_1, n_2,..., n_k observaciones independientes y obtenidas de forma aleatoria, por lo que el total de los datos está dado por $n = n_1 + n_2 + ... + n_k$. Si se designa de forma general cada observación por x_{ji}, donde el subíndice i indica el grupo (muestra) al que pertenece y j es el número de la observación dentro de ese grupo (la medida de expresión del gen j en la muestra i), con $i = 1,2,\ldots,n$, y $j = 1,2,\ldots,m$. Para probar las hipótesis complementarias de homogeneidad

H_o: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ (La media poblacional de la medida de expresión de los

genes a través de las muestras es la misma para las distintas clases),

H_a: Al menos para una de las clases la media poblacional de la medida de expresión de los genes es distinta,

deben cumplirse las siguientes suposiciones: las observaciones deben ser independientes, la variable de respuesta debe distribuirse normalmente y la variabilidad en cada grupo debe ser igual. Una vez satisfechas estas condiciones, la prueba estadística usada para tomar la decisión es la prueba *F* obtenida de completar la tabla del análisis de varianza, como sigue:

Fuentes de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrados Medios**	**F**
Entre Grupos	*k*-1	*BSS*	*BMS* = *BSS*/(*k*-1)	*BMS*/*MSE*
Dentro de Grupos	*n*-*k*	*SSE*	*MSE* = *SSE*/(*n*-*k*)	
Total	*n*-1	*SST*		

Tabla 3.4.4: Análisis de Varianza

Donde *SST*, *SSE* y *BSS* representan la Suma de Cuadrados del Total, la Suma de Cuadrados Dentro de los grupos y la Suma de Cuadrados Entre los grupos respectivamente.

Sean $n = \sum_{i=1}^{k} n_i$; $X_i. = \sum_{i=1}^{n_i} X_{ji} / n_i$, el estimador por mínimos cuadrados de la media μ_i en el grupo *i*; y $X.. = \sum_{i=1}^{k} n_i X_i. / n$, el estimador por mínimos cuadrados de la media μ de todas las observaciones. Se tiene que

$$\sum\sum (X_{ji} - \mu_i)^2 = \sum\sum (X_{ji} - X_i.)^2 + \sum\sum n_i (X_i. - \mu_i)^2 \quad \text{y}$$

$$\sum\sum (X_{ji} - \mu)^2 = \sum\sum (X_{ji} - X..)^2 + \sum\sum n_i (X.. - \mu_i)^2 .$$

Y la hipótesis nula H_o es rechazada si

$$W = \frac{\sum n_i (X_i. - X..)^2 /(k-1)}{\sum (X_{ji} - X_i.)^2 /)(n-k)} > C. \qquad (3.4.4)$$

Puesto que la expresión (3.4.4) es el cociente de dos variables aleatorias χ^2 independientes, cada una dividida por sus respectivos grados de libertad, *W* tiene distribución *F* con $k-1$ y $n-k$ grados de libertad, y por tanto el valor crítico *C* es determinado por

$$\int_C^{\infty} F_{(k-1),(n-k)}(x)dx = \alpha .$$

La suma de cuadrados en el numerador y denominador en (3.4.4) puede interpretarse como las dos componentes en la descomposición de la variación total

$$SST = \sum\sum (X_{ji} - X..)^2 = \sum\sum (X_{ji} - X_i.)^2 + \sum\sum n_i (X_i. - X..)^2 ,$$

en la cual el primer y segundo término de la derecha representan respectivamente, a *SSE* y *BSS*.

3.4.5 Prueba *Kruskal-Wallis* para comparar varias muestras

Es una ampliación de la prueba *Mann-Whitney* para tres o más grupos y, es una alternativa no paramétrica a la prueba *F* del análisis de varianza para un solo factor, cuando no se cumplen algunas de las condiciones necesarias para aplicar dicho método. Exige solamente la aleatoriedad en la extracción de las muestras. Y al igual que todas las pruebas no paramétricas, asigna rangos a las observaciones y compara varios grupos pero usando la mediana de cada uno de ellos, en lugar de las medias. Establece las hipótesis complementarias de la siguiente forma:

H_o: Las *k* medianas son todas iguales,

H_a: Al menos una de las medianas es diferente.

Para calcular el estadístico de contraste de *Kruskal-Wallis* denotado por *H*, deben hallarse primero los rangos de cada una de los *k* grupos, combinado todos los grupos en uno sólo y en caso de haber datos empatados se asigna un rango promedio a cada dato del grupo empatado. Así, *H* esta dado por

$$H = \frac{12}{n \cdot (n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1), \qquad (3.4.6)$$

donde R_i es la suma de los rangos del grupo i y n es el total de datos.

Cuando hay empates en los datos, se aplica la siguiente corrección a H

$$H' = \frac{H}{1 - \frac{\sum_{i=1}^{g} t_i^3 - t_i}{n^3 - n}},$$

donde g es el número de grupos de diferentes rangos empatados, t_i es el número de observaciones repetidas dentro del grupo i.

Cuando los tamaños de cada grupo son mayores que 5, se puede mostrar que H se distribuye como una *Ji-Cuadrado* con k-1 grados de libertad. Por tanto, la hipótesis nula se rechaza si $H > \chi^2_{k-1,1-\alpha}$, o si su respectivo *p-value* es menor que α.

3.5 *P-values* No Ajustados (*Unadjusted*) y Ajustados (*Adjusted*)

Como ya se ha mencionado, la discusión para probar una sola hipótesis no se centra únicamente en un valor crítico y su región de rechazo; sino también en el *p-value*, el cual proporciona mayor información que simplemente aceptar o rechazar la hipótesis. La extensión de este concepto al contexto de pruebas múltiples aunque no es directa, está dado por el *p-value* ajustado.

***P-values* No ajustados:** Consideremos primero una sola hipótesis nula H_j (igual 0 si cierta y 1 si no lo es) y entonces, consideremos también una familia de hipótesis H_j*'s*, cada una con nivel de significancia α_j y región de rechazo S_{α_j} tal que:

1. $\Pr(T_j \in S_{\alpha_j} \mid H_j = 0) = \alpha_j$, realizable bajo la distribución del estadístico de prueba T_j y para todo $\alpha_j \in [0,1]$.
2. $S_{\alpha_j'} = \bigcap_{\alpha_j \geq \alpha_j'} S_{\alpha_j}$, para todo α_j y α'_j para los cuales estas regiones se definen en (1).

Entonces el *p-value* para el valor observado de $T_j = t_j$, puede definirse como $p_j = \inf\{\alpha_j : t_j \in S_{\alpha_j}\}$, [27]. De este modo, el *p-value* es el menor nivel de significancia de la prueba al cual se rechaza la hipótesis nula H_j, dado t_j. Equivalentemente, si se está considerando una prueba H_j de dos lados, el *p-value* puede definirse en términos del estadístico de prueba como

$$p_j = 2\Pr(|T_j| \geq |t_j| \,|H_j). \tag{3.4.9}$$

El cual manifiesta que H_j es rechazada para valores grandes de $|T_j|$ y por tanto, un p_j pequeño proporciona evidencia fuerte en contra de la hipótesis nula H_j. Al rechazar H_j cuando $p_j \leq \alpha_j$, se obtiene un control de la tasa de Error Tipo I al nivel α_j especificado con anticipación.

Lemma 3.5: Sea θ el parámetro que clasifica la distribución de los datos, dado por

$$\theta = \begin{cases} 0, & si \;\; H_j = 0 \\ 1, & si \;\; H_j = 1 \end{cases},$$

tal θ se considera conocido solamente si pertenece a un cierto conjunto $\Omega = \Omega_0 \cup \Omega_1$, donde Ω_0 y Ω_1 representan, respectivamente, los conjuntos al cual pertenece θ cuando H_j está en la clase de hipótesis nulas verdaderas o en la clase de hipótesis nulas falsas. Denotemos por P_θ la distribución de t_j. Si la hipótesis nula H_j es cierta y asumiendo que se cumple lo establecido por (1) y (2), se tiene que:

i. Si

$$\sup_{\theta \in \Omega_0} P_\theta\{t_j \in S_\alpha\} \leq \alpha, \text{ para todo } 0 < \alpha < 1, \tag{3.4.10}$$

entonces la distribución de p_j bajo $\theta \in \Omega_0$ satisface

$$P_\theta\{p_j \leq u\} \leq u, \text{ para todo } 0 \leq u \leq 1. \tag{3.5.11}$$

ii. Si, para $\theta \in \Omega_0$,

$$P_\theta\{t_j \in S_\alpha\} = \alpha, \text{ para todo } 0 < \alpha < 1, \tag{3.5.12}$$

entonces

$$P_\theta\{p_j \le u\} = u \text{ , para todo } 0 \le u \le 1. \qquad (3.5.13)$$

Es decir, p_j se distribuye uniformemente sobre el intervalo (0,1).

Prueba:

i. Si $\theta \in \Omega_0$, entonces el evento $\{p_j \le u\}$ implica $\{t_j \in S_v\}$ para todo $u < v$. Pero por hipótesis se sabe que

$$\sup_{\theta \in \Omega_0} P_\theta\{t_j \in S_v\} \le v \text{, para todo } 0 < v < 1.$$

Luego si v tiende a u se tiene (3.5.11).

ii. Sea $\theta \in \Omega_0$, como el evento $\{t_j \in S_u\}$ implica $\{p_j \le u\}$, se tiene que

$$P_\theta\{p_j \le u\} \ge P_\theta\{t_j \in S_u\}.$$

Por tanto, si se cumple (3.5.12), entonces $P_\theta\{p_j \le u\} \ge u$ y por (*i*), se cumple entonces (3.5.13). ■

P-values **Ajustados:** Considerando específicamente un método de prueba múltiple, el *p-value* ajustado correspondiente a la prueba de una sola hipótesis H_j, puede definirse como el nivel nominal del método de prueba al cual H_j debe ser rechazada, una vez conocidos los valores de todos los estadísticos de prueba involucrados. Estos *p-values* ajustados dependen de la tasa de Error Tipo I que se desee controlar, así por ejemplo, si el interés es controlar el *FWER* entonces el *p-value* ajustado para la hipótesis H_j está dado por:

$$\tilde{p}_j = \{\alpha \in [0,1]: H_j \text{ es rechazada al nivel } FWER = \alpha\}, \qquad (3.4.10)$$

donde el *FWER* nominal es el nivel α al cual el método considerado es realizado. Así, la hipótesis nula H_j es rechazada al *FWER* nominal α si $\tilde{p}_j \le \alpha$. Los *p-values* ajustados para métodos que controlan las otras tasas de Error Tipo I, se definen de forma similar como lo establece (3.4.10).

Más específicamente, establecido y definido el método de prueba para cualquier nivel α, $0 \leq \alpha \leq 1$, un *p-value* ajustado cumple las siguiente propiedades:

- Monoticidad: Si H_j es rechazada para α, entonces H_j es rechazada para todo $\alpha' > \alpha$.
- Continuidad a Izquierda: Si $\alpha_1 > \alpha_2 > \alpha_3 \cdot \cdot \cdot$ es una secuencia decreciente, $\lim_{n\to\infty} \alpha_n = \alpha$ y H_j es rechazada para todo α_n, con $n = 1, 2, 3, \ldots$, entonces H_j es rechazada para α.
- Para cada $\alpha = 1$, toda hipótesis H_j es rechazada.

Los *p-values* ajustados describen convenientemente algunos métodos de prueba múltiple y con su uso no es necesario determinar el nivel de la prueba por adelantado. En nuestro caso el interés es establecer pruebas que permitan identificar genes que se diferencien de los demás, pero los genes aquí considerados se suponen correlacionados por la forma como se construye el microarreglo, lo que constituye una razón más para usar estos *p-values* ajustados.

De acuerdo al método de prueba múltiple empleado, los *p-values* ajustados se calculan de diferentes formas, en el próximo capítulo se establecen algunas de ellas.

Capítulo 4
MÉTODOS DE PRUEBA MÚLTIPLE

Cuando se prueban simultáneamente muchas hipótesis lo que más interesa es el control de la tasa de Error Tipo I, puesto que la probabilidad de cometer este tipo de error aumenta con el número de hipótesis y muy rara vez se está interesado solo en saber si todas las hipótesis nulas son conjuntamente verdaderas o falsas. En la mayoría de las aplicaciones se infiere sobre las distintas hipótesis nulas, consciente de que algunas de ellas son ciertas y otras no; los métodos de prueba múltiple se interesan en satisfacer tal condición.

4.1 Método de Prueba Múltiple (*MTP*)

Un método de prueba múltiple (*MTP* por sus siglas en ingles) es aquel que proporciona una región de rechazo S_α, constituida por los valores de cada estadístico de prueba T_j que conduce al rechazo de la hipótesis nula H_j. En otras palabras, un *MTP* es un subconjunto aleatorio R de hipótesis rechazadas que estima el conjunto de verdaderos positivos, dado por

$$R = \{j: H_j \text{ es rechazada}\} = \{ j: T_j > c_j \} \subseteq \{1,\ldots,m\},$$

donde, c_j es el valor crítico que determina la región rechazo para la hipótesis H_j y, R es una función que depende de:

- Las observaciones a través del m-vector de los estadísticos de prueba T_j correspondientes a la hipótesis nula H_j.
- Una distribución nula de los estadísticos de prueba T_j*'s*, estimada determinando los puntos de corte (*cut-offs*) para cada T_j y los *p-values* ajustados que resultan.
- El nivel nominal α del *MTP*, es decir, el límite superior deseado para una tasa de Error Tipo I convenientemente definida.

Usualmente se distinguen tres tipos de métodos de prueba múltiple: los de un solo paso (*single-step*), los de pasos hacia abajo (*step-down*) y los de pasos hacia arriba (*step-up*).

Métodos de un solo Paso: Realizan los ajustes de multiplicidad para todas las hipótesis, basándose en los *p-values* ajustados sin importar el orden de los estadísticos de prueba o de los *p-values* no ajustados. Es decir, cada hipótesis se evalúa usando un valor crítico que es independiente de los resultados de las demás pruebas de hipótesis.

Los métodos de prueba múltiple *step-down* y *step-up*, conocidos también como métodos por etapas (*stepwise*) tratan de mejorar la potencia, mientras preservan el control de la tasa de Error Tipo I; considerando que el rechazo de una hipótesis particular no depende solo del número total de hipótesis, sino también de los resultados de las pruebas de las otras hipótesis.

Métodos de Pasos hacia Abajo: Asumen que las hipótesis son enumeradas de acuerdo a los *p-values* no ajustados ordenados y comparan sucesivamente las hipótesis que corresponden al estadístico de prueba más significativo (*p-value* no ajustado menor) con los de las demás pruebas que dependen de los resultados de las pruebas anteriores. Una vez que no se puede rechazar una hipótesis nula no rechazan las demás hipótesis.

Estos métodos se basan en los *p-values* ajustados, los cuales son funciones de sus *p-values* no ajustados ordenados y, mejoran la potencia, mientras preservan el control fuerte del *FWER*.

Métodos de Pasos hacia Arriba: Únicamente se diferencian de los métodos *step-down*, en que consideran el estadístico de prueba menos significativo y, tan pronto como rechazan la primera hipótesis nula rechazan las demás hipótesis restantes.

En adelante denotaremos por P_j y $\tilde{P}_j$, las variables aleatorias correspondientes a los *p-values* no ajustados y ajustados, respectivamente.

4.2 Métodos de un Solo Paso que controlan el *FWER*

Una de las principales preocupaciones en problemas de hipótesis múltiple es controlar la probabilidad de rechazar erróneamente al menos una de las hipótesis nulas verdaderas, es

decir, controlar el *FWER*. Dicho de otra forma, se desea que las hipótesis rechazadas sean las que en verdad deben serlo. En el campo de la medicina y la biología generalmente se espera que pocos genes sean diferencialmente expresados.

$(*)$ Para efectos de este trabajo asumiremos que $\Pr\{p_j \le u\} \le u$ para cualquier $u \in (0,1)$, ver lemma 3.5. Además, por complemento $\Pr\{p_j \ge u\} = 1 - \Pr\{p_j \le u\} \ge 1 - u$.

4.2.1 Método de Bonferroni

Este método se basa en la desigualdad de primer orden de Bonferroni, la cual establece que dado cualquier conjunto de eventos $A_1, A_2, \ldots, A_m$, la probabilidad de su unión es menor o igual que la suma de sus probabilidades, esto es,

$$\Pr\left(\bigcup_{j=1}^{m} A_j\right) \le \sum_{j=1}^{m} \Pr(A_j).$$

Los *p-values* ajustados para este método están dados por

$$\tilde{p}_j = \min(m \cdot p_j, 1) = \begin{cases} m \cdot p_j, & si \quad m \cdot p_j < 1 \\ 1, & , si \quad m \cdot p_j > 1 \end{cases} \tag{4.2.1}$$

El método de Bonferroni simple rechaza la hipótesis nula H_j si $p_j \le \alpha_j$, con los α_j*'s* elegidos de tal forma que $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_m$; generalmente, los α_j*'s* son elegidos a ser iguales al nivel corregido α/m $(**)$. Asumiendo sin pérdida de generalidad que el número de hipótesis nulas H_j verdaderas es m_0 (véase Tabla 3.3), se tiene que

$$FWER = \Pr(V > 0) = \Pr\left(\bigcup_{j=1}^{m_0} \{\tilde{P}_j \le \alpha\}\right) \le \sum_{j=1}^{m_0} \Pr(\tilde{P}_j \le \alpha) \le \sum_{j=1}^{m_0} \left(\frac{\alpha}{m}\right) = \left(\frac{m_0 \cdot \alpha}{m}\right) \le \alpha.$$

La segunda desigualdad se tiene por el lemma 3.5 y por $(**)$; y la última desigualdad se debe a que $m_0 \le m$.

Por tanto, para obtener un control fuerte del *FWER* al nivel α, el método de Bonferroni rechaza cualquier hipótesis nula H_j con *p-value* menor o igual que el nivel corregido α/m,

pero constituye una prueba demasiado conservativa, pues el *FWER* es generalmente menor que α, debido a que cuando se realizan muchas pruebas los niveles corregidos son demasiado bajos, lo que implica que los *p-values* ajustados sean demasiado altos; y esto significa que se seleccionarán pocos genes como diferencialmente expresados.

4.2.2 Método de Šidák (*SS*)

Está altamente relacionado con el método de Bonferroni y al igual que dicho método es altamente conservativo. Sus estadísticos de prueba y *p-values* ajustados son independientes, estos últimos se distribuyen uniformemente sobre el intervalo [0, 1], y están dados por

$$\tilde{p}_j = 1-(1-p_J)^m . \tag{4.2.2}$$

Este método proporciona un control fuerte del *FWER* a un nivel α. En efecto, notar que

$$\Pr(V=0) = \Pr\left(\bigcap_{j=1}^{m_0}\{\tilde{P}_j \geq \alpha\}\right) = \prod_{j=1}^{m_0}(\tilde{P}_j \geq \alpha) = \prod_{j=1}^{m_0}\Pr(P_j \geq 1-(1-\alpha)^{1/m}) = \{(1-\alpha)^{1/m}\}^{m_0} .$$

Donde la segunda igualdad se debe a la independencia de los *p-values*; la tercera igualdad se obtiene remplazando $\tilde{p}_j$ por α y despejando p_j en (4.2.2); y la última igualdad se debe a la distribución de los *p-values*.

Luego,

$$FWER = \Pr(V>0) = 1-\Pr(V=0) = 1-(1-\alpha)^{m_0/m} \leq \alpha .$$

El método de Šidák es menos conservativo que el de Bonferroni; es decir, rechaza más hipótesis nulas y por tanto declara más genes como diferenciados. En efecto, usando los primeros términos de la expansión de las series de Taylor se tiene que

$$1-(1-p_j)^m = m\cdot p_j - \frac{1}{2}m\cdot(m-1)\cdot(1-\xi)^{m-2}\cdot p_j^2$$

$$\tilde{p}_S = \tilde{p}_B - \frac{1}{2}m\cdot(m-1)\cdot(1-\xi)^{m-2}\cdot p_j^2 ,$$

donde $\tilde{p}_S = 1-(1-p_J)^m$ y $\tilde{p}_B = m \cdot p_j$ denotan respectivamente los *p-values* ajustados por el método de Šidák y Bonferroni, y $0 \le \xi \le p_j$.

Esta expresión pone de manifiesto; primero, que los *p-values* ajustados por Šidák son menores que los de Bonferroni, segundo, que la diferencia entre los dos es menor con un p_j menor y tercero, que la diferencia entre los dos es mayor a medida que el número m de hipótesis por probar aumenta.

4.2.3 Métodos *minP* y *maxT*

En la práctica existen muchas situaciones en las que los estadísticos de prueba no son independientes y por tanto ellos, como sus *p-values* están correlacionados. Tal es el caso de los experimentos de microarreglos, donde grupos de genes tienen niveles de expresión altamente correlacionados. Los métodos *minP* y *maxT*, de Westfall y Young, toman en cuenta la estructura de dependencia entre los estadísticos de prueba bajo la hipótesis nula completa $H_0^C = \bigcap_{j=1}^{m} \{H_j : \text{es verdadera}\}$, y establecen *p-values* ajustados para métodos de prueba múltiple menos conservativos, definidos en términos de la variable aleatoria que representa los *p-values* no ajustados P_l y los estadísticos de prueba T_l, de la l-ésima hipótesis nula; dados respectivamente por

$$\tilde{p}_j = \Pr(\min_{1 \le l \le m} P_l \le p_j \,|\, H_0^C), \tag{4.2.3}$$

$$\tilde{p}_j = \Pr(\max_{1 \le l \le m} |T_l| \ge |t_j| \,|\, H_0^C). \tag{4.2.4}$$

Estos métodos proporcionan un control débil del *FWER* bajo cualquier condición. El control fuerte se obtiene bajo la suposición de subconjunto pivotal, [47]; el cual establece, que los *p-values* no ajustados $P_1, P_2, \ldots, P_m$ tienen la propiedad de subconjunto pivotal si para todos los subconjuntos κ de $\{1,\ldots, m_0\}$ de hipótesis nulas verdaderas, las distribuciones conjuntas del vector aleatorio $\{P_j: j \in \kappa\}$ son idénticas bajo la hipótesis nula completa H_0^C y

la hipótesis nula parcial $H_\kappa = \bigcap_{j\in\kappa}\{H_j$: es verdadera$\}$; refiriéndose ésta última a la distribución conjunta de los estadísticos de prueba T_j o de los *p-values* P_j, para probar las hipótesis H_j, con $j \in \kappa$. En efecto, sea $\kappa = \{j_1,..., j_i\}$ la colección de hipótesis H_j que son verdaderas, *S* el espacio muestral de los *p-values* mínimos y c_α el valor observable mayor de la variable aleatoria $\min_{1\le j\le m} P_j$ para la cual la probabilidad acumulativa es a lo más α, esto es,

$$c_\alpha = \max\{\, p_j \in S : \Pr(\min_{1\le l\le m} P_l \le p_j \,|\, H_0^C) \le \alpha\},$$

Entonces

$$\begin{aligned} FWER &= \Pr(\text{al menos una } H_j \text{ verdadera sea rechazada para } j \in \kappa \,|\, H_\kappa) \\ &= \Pr(\min_{l\in\kappa} \tilde{P}_l \le \alpha \,|\, H_\kappa) \\ &= \Pr(\min_{l\in\kappa} P_l \le c_\alpha \,|\, H_\kappa) \\ &= \Pr(\min_{l\in\kappa} P_l \le c_\alpha \,|\, H_0^C) \\ &\le \Pr(\min_{1\le l\le m} P_l \le c_\alpha \,|\, H_0^C) \\ &\le \alpha. \end{aligned}$$

La cuarta igualdad se cumple por la propiedad del subconjunto pivotal.

Cuando los estadísticos de prueba son idénticamente distribuidos, el *minP* y *maxT* coinciden. De esta forma los métodos *minP* y *maxT* proporciona un control fuerte del *FWER*. ■

Debido a que la propiedad de subconjunto pivotal es necesaria para asegurar que los *MTP* basados en los *p-values* calculados bajo H_0^C proporcionen un control fuerte del *FWER*; a menos que se indique lo contrario, en este trabajo consideraremos que tal propiedad es satisfecha. En efecto, como cada hipótesis nula H_j se refiere a un solo gen *j* y cada estadístico de prueba T_j es solamente función de la variable de respuesta *Y* y la medida de expresión X_j, entonces la hipótesis nula parcial H_κ se refiere a la distribución conjunta de las variables *Y* y el vector aleatorio de medidas de expresión $\{X_j: j \in \kappa\}$, el cual es independiente de *Y*. Por lo tanto la distribución conjunta de $\{X_j: j \in \kappa\}$ es idéntica para todos los niveles de *Y*, lo que implica que los estadísticos de prueba T_j (y por tanto sus correspondientes *p-values* no

ajustados) tienen la misma distribución en el subconjunto κ sin importar el valor de verdad de las hipótesis en el complemento de κ [47].

Sobre los cuatro métodos vistos hasta el momento, debe observarse lo siguiente:

1. Si los *p-values* $P_1, P_2, \ldots, P_m$ son independientes y cada P_j se distribuye uniformemente sobre el intervalo [0, 1] bajo H_j, entonces los *p-values* ajustados del *minP* coinciden con los *p-values* de Šidák, puesto que

$$\begin{aligned}\tilde{p}_j &= \Pr(\min_{1\le l\le m} P_l \le p_j \,|\, H_0^C)\\ &= 1-\Pr(P_l > p_j \ \forall j \,|\, H_0^C)\\ &= 1-(1-p_j)^m\\ &= \tilde{p}_S. \ \blacksquare\end{aligned}$$

2. Para distribuciones arbitrarias de los estadísticos de prueba, el método de Šidák no garantiza el control del *FWER*; sin embargo, controla el *FWER* para estadísticos de prueba que satisfacen la desigualdad de Šidák, la cual establece:

$$\Pr(|T_1| \le c_1, \ldots, |T_m| \le c_m) \ge \prod_{j=1}^{m} \Pr(|T_j| \le c_j).$$

Esta desigualdad fue inicialmente derivada por Dunn (1958), para $T_1, T_2, \ldots, T_m$ con distribución normal multivariada, con media cero y ciertos tipos de matriz de covarianza. En 1967, este resultado es extendido por Šidák [40], para matrices de covarianza arbitrarias. Y en 1977, Jogdeo [22] demuestra que tal desigualdad se cumple para una clase más grande de distribuciones, que incluyen las distribuciones multivariadas *t* y *F*.

Bajo la desigualdad de Šidák, los *p-values* ajustados del *minP* son menores que los *p-values* ajustados de Šidák. En efecto,

$$\begin{aligned}\tilde{p}_j &= \Pr(\min_{1\le l\le m} P_l \le p_j \,|\, H_0^C)\\ &= 1-\Pr(P_l > p_j \ \forall j \,|\, H_0^C)\\ &= 1-\Pr(|T_l| \le c_j \ \forall j \,|\, H_0^C)\\ &\le 1-\prod_{j=1}^{m} \Pr(|T_j| \le c_j)\\ &= \tilde{p}_S.\end{aligned}$$

El último reglón se cumple si los *p-values* se distribuyeran uniformemente sobre el intervalo [0,1]; es decir, si se cumplen la suposición bajo la cual se define el método de Šidák.

3. Teniendo en cuenta la relación que existe entre los *p-values* ajustados de Bonferroni y Šidák dados anteriormente, por la parte (2) se puede ver que los métodos basados en los *p-values* ajustados del *minP* dados por la ecuación (4.2.3) son menos conservativos que los métodos de Bonferroni o Šidák. En el caso que los estadísticos de prueba sean independientes, sucede lo establecido en la parte (1).

4. Los *p-values* para el *maxT* son más fáciles de calcular y requieren menos computaciones que los del *minP*; aunque, cuando los estadísticos de prueba se distribuyen idénticamente ambos son iguales. Generalmente estos dos métodos proporcionan *p-values* ajustados diferentes, y la selección de una u otra aproximación depende de las consideraciones de balance, potencia y viabilidad computacional. En el caso en que los estadísticos de prueba T_j no tienen los mismos grados de libertad, es decir, no se distribuyen idénticamente, no todas las pruebas contribuyen igualmente en el cálculo de los *p-values* ajustados para el *maxT*, y esto conduce a establecer juicios desequilibrados [5,47]. Cuando los *p-values* se estiman por permutación y un gran número de hipótesis son probadas, los métodos basados en los *p-values* para el *minP* tienden a ser más sensitivos al número de permutaciones y más conservativos que los métodos basados en el *maxT*, razón por la cual en este estudio no consideraremos el método *minP*.

4.3 Métodos de Pasos hacia Abajo que controlan el *FWER*

En comparación con los métodos de un solo paso, los cuales tienden a ser conservativos para el control del *FWER*; los métodos de paso hacia abajo, mejoran la potencia mientras preservan el control fuerte del *FWER*. A continuación se describen los *p-values* ajustados de los cuatro métodos análogos a los descritos en la sección 4.2, en términos de sus *p-values* no ajustados ordenados $p_{r_1} \leq p_{r_2} \leq ... \leq p_{r_m}$, para los cuales, $H_{r_1} \leq H_{r_2} \leq ... \leq H_{r_m}$ denotan sus correspondientes hipótesis nulas.

4.3.1 Método de Holm

Establecido un nivel α para el control de la tasa de Error tipo I, este método ajusta los *p-values* como sigue:

1. Ordena los *p-values* no ajustados de cada uno de los *m* genes, denotándolos por $p_{r_1}, p_{r_2}, ..., p_{r_m}$, para $r = 1, \ldots, m$.
2. Evalúa las siguientes desigualdades

   $p_{r_1} > \alpha/m;\ p_{r_2} > \alpha/(m-1);\ \ldots\ ;\ p_{r_j} > \alpha/(m-j+1);\ \ldots\ ;\ p_{r_m} > \alpha/1.$

3. Define $j^* = \min\{ j : p_{r_j} > \alpha/(m-j+1)\}$ como el menor de los *j*'s si y solo si $(m-j+1)\cdot p_{r_j} > \alpha$.
4. Si tal j^* existe, se rechaza la hipótesis nula H_{r_j}, para $j = 1, \ldots, j^*-1$. Si j^* no existe, se rechazan todas las hipótesis nulas.

Lo anterior es equivalente a decir que los *p-values* ajustados para el método de Holm están dados por:

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \{\min((m-k+1)\cdot p_{r_k}, 1)\}. \qquad (4.3.1)$$

Cabe resaltar que al considerar el máximo de $\min\{((m-k+1)\cdot p_{r_k}, 1)\}$, el método de Holm garantiza la monoticidad de los *p-values* ajustados, es decir, garantiza que $\tilde{p}_{r_1} \le \tilde{p}_{r_2} \le \ldots \le \tilde{p}_{r_m}$ y que solamente se rechaza una hipótesis nula cuando todas las hipótesis con *p-value* no ajustado menor son rechazadas previamente.

Este método brinda un control fuerte del *FWER* a un nivel α. En efecto, suponiendo que j^* existe y considerando cada una de las etapas establecidas por las desigualdades descritas en (2); para cada etapa *j* = 1, …, *m* se tiene

$$FWER = \Pr(V > 0) = \Pr(\tilde{P}_{r_j} \le \alpha) \le \Pr\left(P_{r_j} \le \frac{\alpha}{m-k+1}\right) \le \frac{\alpha}{m-k+1} \le \alpha.$$

La última desigualdad en la expresión anterior se cumple, puesto que si el número de hipótesis verdaderas es *k*, como el número de hipótesis nulas H_j verdaderas es en realidad $m_0 \leq m$; si $k = m$, entonces la prueba en esta etapa dará lugar a un Error Tipo I con probabilidad menor o igual que α. Si $k = m-1$, aunque pudo ocurrir un Error Tipo I en la etapa anterior, ciertamente tal error ocurrirá con una probabilidad menor o igual que α, si hay un rechazo en esta etapa, debido a que ninguna de las $m-1$ hipótesis verdaderas puede ser rechazada a menos que al menos una de ellas tenga un *p-value* asociado mayor que $\alpha/(m-1)$. Análogamente, para cualquier otro valor de *k*, puede ocurrir en cualquiera de las otras etapas $k-1$un Error Tipo I también con probabilidad menor o igual que α, pero ciertamente este error ocurrirá si hay un rechazo en la etapa $m-k+1$. De esta manera es claro que el $FWER \leq \alpha$ para cualquier posible combinación de hipótesis verdaderas y falsas. ■

El método de Holm es menos conservativo que el método de Bonferroni, cuyos *p-values* son multiplicados por *m* en cada paso. Sin embargo, ninguno de estos dos métodos, consideran la estructura de dependencia entre las variables.

4.3.2 Método de Šidák (*SD*)

Así como para los métodos de un solo paso se definen los *p-values* ajustados de Šidák, éstos también se definen para los métodos de pasos hacia abajo y están dados como sigue:

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \{1-(1-p_{r_k})^{(m-k+1)}\} \tag{4.3.2}$$

4.3.3 Métodos *minP* y *maxT*

Los *p-values* ajustados del *minP* y el *maxT* para los métodos de pasos hacia abajo, se definen respectivamente como sigue:

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \{\Pr(\min_{l\in\{r_k,\ldots,r_m\}} P_l \leq p_{r_k} \left| H_0^C \right.)\}, \tag{4.3.3}$$

$$\tilde{p}_{r_j} = \max_{k=1,\ldots,j} \{\Pr(\max_{l\in\{r_k,\ldots,r_m\}} |T_l| \geq \left| t_{r_k} \right| \left| H_0^C \right.)\}. \tag{4.3.4}$$

Aquí $\left|t_{r_1}\right| \geq \left|t_{r_2}\right| \geq \cdots \geq \left|t_{r_m}\right|$, denotan los estadísticos de prueba observados ordenados.

4.4 Métodos de Pasos hacia Arriba que controlan el *FWER*

Como se mencionó anteriormente, estos métodos consideran el estadístico de prueba menos significante y tan pronto como una hipótesis nula es rechazada, rechaza todas las demás. Por tanto, conducen a rechazar mayor número de hipótesis nulas que los métodos de paso hacia abajo.

Estos métodos controlan tasa del Error Tipo I bajo la suposición de que los estadísticos de prueba satisfacen la desigualdad de Simes.

4.4.1 Desigualdad de Simes

Sea $X_{(1)},...,X_{(m)}$ un conjunto ordenado de variables aleatorias continuas, con función de distribución conjunta F y *p-values* no ajustados ordenados $U_{(1)},...,U_{(m)}$ dados por $U_{(j)} = F(X_{(j)})$ y que se distribuyen uniformemente sobre el intervalo [0,1], para $j = 1,...,m$. La desigualdad de Simes establece que

$$P\left(U_{(j)} \geq \frac{\alpha \cdot j}{m},\ \forall j = 1,...,m\right) \geq 1-\alpha. \qquad (4.4.1)$$

Simes (1986) probó esta desigualdad bajo la hipótesis nula completa H_0^C y para estadísticos de prueba independientes, y la igualdad para el caso continuo. Hasta hace poco los estudios de simulación realizados por Simes han probado empíricamente la desigualdad (4.4.1) y sugieren que la misma es conservativa para una variedad de distribuciones Gaussiana y Gamma multivariadas de los estadísticos de prueba. Además, aunque no proporciona un método formal, establece que este resultado puede ser usado en situaciones de prueba múltiple.

De acuerdo a la notación aquí establecida, según la cual $P_{r_1},...,P_{r_m}$ representan los *p-values* no ajustados ordenados, $\tilde{P}_{r_1},...,\tilde{P}_{r_m}$ los *p-values* ajustados ordenados y $H_{r_1},...,H_{r_m}$ sus hipótesis nulas correspondientes; la desigualdad de Simes estaría dada por

$$\begin{aligned} &\Pr\left(P_{r_j} \geq \frac{\alpha \cdot j}{m},\ \forall j = 1,...,m \,\middle|\, H_0^C \right) \\ &= \Pr\left(\bigcap_{j=1}^{m} \left\{ P_{r_j} \geq \frac{\alpha \cdot j}{m} \right\} \right) \\ &\geq 1-\alpha. \end{aligned} \tag{4.4.2}$$

o equivalentemente, por complemento,

$$\begin{aligned} &\Pr\left(P_{r_j} \leq \frac{\alpha \cdot j}{m},\ \exists j = 1,...,m \,\middle|\, H_0^C \right) \\ &= \Pr\left(\bigcup_{j=1}^{m} \left\{ P_{r_j} \leq \frac{\alpha \cdot j}{m} \right\} \right) \\ &\leq \alpha \end{aligned} \tag{4.4.3}$$

4.4.2 Método de Hochberg

Este método usa la desigualdad de Simes para derivar el control del *FWER* a un nivel α, ajustando los *p-values* como sigue

$$\tilde{p}_{r_j} = \min_{k=j,...,m} \{\min((m-k+1)\cdot p_{r_k}, 1)\} \tag{4.4.4}$$

Define $j^* = \max\{ j : p_{r_j} \leq \alpha/(m-j+1)\}$ y rechaza la hipótesis H_{r_j} para $j = 1,..., j^*$, si tal j^* no existe, no se rechazan hipótesis.

Este método proporciona un control fuerte del *FWER* a un nivel α. En efecto; sea Π_0 el conjunto de todas las hipótesis nulas verdaderas y $h = \min\left\{ j : \left|\{r_1,...,r_j\} \cap \Pi_0\right| = j' \right\}$, donde los $r_k's$ representan los índices de los *p-values* ajustados ordenados para $k = 1,..., j$. La forma como se definió $j' = \left|\{r_1,...,r_j\} \cap \Pi_0\right| = 1,..., m_0$, establece que j^* es el rango de todas

las m hipótesis verdaderas con h-iésimo *p-value* no ajustado menor de entre todas las m_0 hipótesis nulas verdaderas. Entonces $j' \leq h \leq m - m_0 + j'$, por lo que se tiene

$$p_{r_h} \leq \frac{\alpha}{m_0 - j' + 1},$$

Luego,

$$\begin{aligned} FWER = \Pr(V > 0) &\leq \Pr\left(\exists\, j^* = 1,\ldots, m_0 \;/\; p_{r_h} \leq \frac{1}{m_0 - j' + 1} \cdot \alpha \right) \\ &= \Pr\left(\bigcup_{j=1}^{m_0} \left\{ p_{r_h} \leq \frac{1}{m_0 - j' + 1} \cdot \alpha \right\} \right) \\ &\leq \Pr\left(\bigcup_{j=1}^{m_0} \left\{ p_{r_h} \leq \frac{j'}{m_0} \cdot \alpha \right\} \right) \\ &\leq \alpha. \end{aligned}$$

El método de Hochberg puede ser descrito como una modificación del método de paso hacia arriba de Holm, debido a que los *p-values* no ajustados ordenados son comparados con el mismo valor crítico $\alpha/(m - j + 1)$.

4.5 Métodos que controlan el *FDR*

Para problemas de alta dimensionalidad, generalmente es preferible trabajar con las tasas de error basadas en la proporción de falsos positivos (*FDR*), que con las tasas de error basadas en el número de falsos positivos (*FWER*, *PFER*); las primeras se mantienen estables así el número de hipótesis por probar aumente, en cambio, por ejemplo, bajo la hipótesis nula completa ($m = m_0$), el *FWER* crece exponencialmente a medida que el número de hipótesis por probar aumenta. En efecto, supóngase que cada hipótesis es probada individualmente al nivel α, si t_j es el valor observado de los estadísticos de prueba T_j correspondientes a cada hipótesis nula H_j y si c_j es un valor crítico tal que $\Pr(|t_j| > c_j) = \alpha$, $j = 1,\ldots,m$, entonces

$$
\begin{aligned}
FWER = \Pr(V > 0) &= 1 - \Pr(V = 0) \\
&= 1 - \Pr\left(\bigcap_{j=1}^{m_0} \left\{ \left| t_j \right| \le c \right\} \right) \\
&= 1 - \prod_{j=1}^{m_0} \Pr\left(\left| t_j \right| \le c \right) \\
&= 1 - (1-\alpha)^{m_0}. \ \blacksquare
\end{aligned}
$$

Por tanto, así se esté dispuesto a tolerar un número pequeño de errores Tipo I en comparación con el número de hipótesis rechazadas, el control del *FWER* puede proporcionar métodos excesivamente conservativos, por tanto es necesario un enfoque menos conservador, el del *FDR*; puesto que bajo la hipótesis nula el *FDR* es igual al *FWER*, ver (3.3.1). Así, en este sentido, los métodos que controlan el *FDR* controlan también el *FWER*.

4.5.1 Método de pasos hacia Arriba de Benjamini y Hochberg (*BH*)

Al igual que el método anterior este también se basa en la desigualdad de Simes.

Sean $p_{r_1},...,p_{r_m}$ los *p-values* no ajustados ordenados y $H_{r_1},...,H_{r_m}$ sus correspondientes hipótesis nulas; este método proporciona un control fuerte del *FDR* para estadísticos de prueba independientes al nivel α (ver Apéndice B), definiendo $j^* = \max\{ j : p_{r_j} \le (j/m) \cdot \alpha \}$ y rechazando H_{r_j} para $j = 1,..., j^*$, si tal j^* no existe, no se rechazan hipótesis. Este método ajusta los *p-values* como sigue

$$
\tilde{p}_{r_j} = \min_{k=j,...,m} \left\{ \min\left(\left(\frac{m}{k} \right) \cdot p_{r_k}, 1 \right) \right\}. \tag{4.5.1}
$$

4.5.2 Método de pasos hacia Arriba de Benjamini y Yekutieli (*BY*)

Es una modificación simple y conservativa del método Benjamini y Hochberg. Proporciona un control fuerte del *FDR* (ver Apéndice C) para estadísticos de prueba con distribuciones conjuntas arbitrarias (dependencia de regresión positiva) a un nivel α,

definiendo $j^* = \max\left\{j : p_{r_j} \le \left(\left\lfloor j \Big/ \left(m \cdot \sum_{j=1}^{m}(1/j)\right)\right\rfloor \cdot \alpha\right)\right\}$ y rechazando H_{r_j} para $j = 1,..., j^*$, si tal j^* no existe, no se rechazan hipótesis. Este método establece los *p-values* ajustados como sigue

$$\tilde{p}_{r_j} = \min_{k=j,...,m}\left\{\min\left(\left(\frac{m \cdot \sum_{j=1}^{m}(1/j)}{k}\right) \cdot p_{r_k}, 1\right)\right\}. \qquad (4.5.2)$$

Los dos métodos anteriores se diferencian en el factor de penalidad, es decir, el multiplicador aplicado a los *p-values* no ajustados. Para el método estándar de Benjamini y Hochberg, la penalidad es m/k, mientras para el método conservativo de Benjamini y Yekuteili es mayor, $m \cdot \sum_{j=1}^{m}(1/j) \Big/ k$. Luego para un gran número *m* de hipótesis, las penalidades difieren por un factor aproximado de log(*m*) para el método *BY*, (ver Apéndice C) lo cual representa un precio muy alto a pagar por mantener la estructura de dependencia.

4.5.3 Análisis de Significancia de Microarreglos (*SAM*)

El Análisis de Significancia de Microarreglos llamado *SAM* (Tusher et al., 2001 [45]) es uno de los métodos no paramétricos más populares para detectar genes significantes y controlar el *FDR*. Asume que la estructura de dependencia es la misma bajo la hipótesis nula completa, y es una modificación de la prueba *t* establecida en la ecuación 3.4.3 y se define como sigue

$$d_{(j)} = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{s(j) + s_0}, \qquad (4.5.3)$$

donde $\bar{x}_{1j}$ y $\bar{x}_{2j}$ representan la media de los niveles de expresión para el gen *j* en la clase o condición 1 y 2, respectivamente. La dispersión específica de los genes *s*(*j*), se define por

$$s(j) = \sqrt{a\left\{\sum_{i=1}^{n_1}(x_{ji} - \bar{x}_{1j})^2 + \sum_{i=1}^{n_2}(x_{ji} - \bar{x}_{2j})^2\right\}},$$

para el cual $a = (1/n_1 + 1/n_2)/(n_1 + n_2 - 2)$. Por otra parte, la constante s_0 se elige de modo que minimice el coeficiente de variación $d_{(j)}$ dentro de las clases de los genes con varianza aproximadamente igual.

El algoritmo *SAM* propuesto por Tusher, Tibshirani y Chu (2001), puede establecerse como sigue:

1. Calcular los estadísticos de prueba $d_{(j)}$ para cada gen j ($j = 1,...m$) y ordenarlos de acuerdo a su magnitud como: $d_{(1)} \geq d_{(2)} \geq \cdots \geq d_{(m)}$.

2. Realizar B permutaciones para las variables de respuestas $y_1, y_2,..., y_n$. Calcular el estadístico de prueba $d^b_{(j)}$ para cada permutación $b = 1,...B$ y sus correspondientes estadísticos ordenados $d^b_{(1)} \geq d^b_{(2)} \geq \cdots \geq d^b_{(m)}$. Cabe aclarar que el estadístico $d^b_{(j)}$ no corresponde necesariamente al mismo gen al que corresponde el estadístico observado $d_{(j)}$.

3. Para las B permutaciones y bajo la hipótesis nula completa, estimar el valor esperado de los estadísticos de orden, dado por

$$\bar{d}_{(j)} = \frac{1}{B} \cdot \sum_{b=1}^{B} d^b_{(j)} .$$

4. Realizar una gráfica llamada *plot SAM*, de los estadísticos de prueba ordenados $d_{(1)}, d_{(2)},\ldots, d_{(m)}$ versus los estadísticos esperados $\bar{d}_{(1)}, \bar{d}_{(2)},..., \bar{d}_{(m)}$.

5. Para cada posible cota fija Δ, un gen es llamado **significativo** si

$$\left| d_{(j)} - \bar{d}_{(j)} \right| > \Delta .$$

   Entonces el número total de genes significativos está dado por

$$tgs = \#\left\{ j : \left| d_{(j)} - \bar{d}_{(j)} \right| > \Delta \right\}.$$

   Sea

$$j_0 = \max\{\, j : \bar{d}_{(j)} \geq 0\},$$
$$j_1 = \max\{\, j \leq j_0 : d_{(j)} - \bar{d}_{(j)} \geq \Delta\},$$

$$j_2 = \min\{\ j > j_0 : d_{(j)} - \bar{d}_{(j)} \le -\Delta\},$$

todos los genes con $j \le j_1$ son llamados **significativos positivos** y todos los genes con $j \ge j_2$ son llamados **significativos negativos**.

6. Para un Δ dado, se define el punto de corte superior $\delta_U = \min\{\ d_{(j)} : j \le j_1\}$ y el punto de corte inferior $\delta_L = \max\{\ d_{(j)} : j \ge j_2\}$. Si tales δ_U y δ_L no existen, entonces $\delta_U = \infty$ y $\delta_L = -\infty$.

   El número de falsos positivos, *PFER*, se estima calculando para cada una de las *B* permutaciones los números de genes con $d^b_{(j)}$ superiores a δ_U o inferiores a δ_L, y promediando este número sobre las permutaciones. Es decir, si denotamos el conjunto de genes significativos declarados en (5) por *gS*, entonces

$$TgS = \frac{1}{B} \cdot \sum_{b=1}^{B} \#\left\{ j \in gS : d^b_{(j)} > \delta_U \ \vee\ d^b_{(j)} < \delta_L \right\}.$$

   Así, el *FDR* es estimado por *FDR=TgS/gS*.

7. Δ se elige para controlar el número de falsos positivos bajo la hipótesis nula completa y a un nivel nominal aceptable.

4.6 Método de *p-values* ajustados por Permutación

Los *p-values* no ajustados y ajustados pueden ser estimados usando métodos de remuestreo como *bootstrapping* y permutación; con el fin de evitar las suposiciones sobre la distribución conjunta de los estadísticos de prueba, la cual generalmente en la práctica es desconocida.

En el contexto de los microarreglos, se considera la hipótesis nula H_j de no asociación entre la variable X_j y la respuesta o covariable Y, para $j = 1,...,m$; y se estima la distribución conjunta de los estadísticos de prueba $(T_1,...,T_m)$, bajo la hipótesis nula completa, permutando las columnas de la matriz X de medidas de expresión de los genes (ver Figura 4.6.1). Al permutar las columnas enteras de la matriz, se tiene que la variable de respuesta Y es

independiente de las medidas de expresión y se preserva la estructura de correlación y las características distribucionales de las medidas de expresión. Generalmente, para un tamaño de muestra *n* grande, se considera un conjunto *B* de permutaciones, en lugar de considerar todas las posibles permutaciones.

Algoritmo de permutación para *p-values* no ajustados (*rawp*)

Para la *b*-iésima permutación, $b = 1,...,B$:

1. Se permutan las *n* columnas de la matriz de datos *X*.
2. Se calculan los estadísticos de prueba $t_{1,b},...,t_{m,b}$ para cada hipótesis , es decir, cada gen.

Después de hechas las *B* permutaciones, para una hipótesis alterna de dos lados, los *p-values* por permutación para la hipótesis H_j es

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

donde

$$I(\cdot) = \begin{cases} 1, & si \;\; |t_{j,b}| \geq |t_j| \\ 0, & en \;\; otro \;\; caso. \end{cases}$$

Figura 4.6.1: Algoritmo de permutación para *p-values* no ajustados. Sandrine Dudoit (2003).

Los *p-values* ajustados por permutación para los métodos de Bonferroni, Ŝidák, Holm y Hochberg, que controlan el *FWER*; y los métodos de Benjamini y Hochberg, y Benjamini y Yekutieli, que controlan el *FDR*, pueden obtenerse remplazando p_j por p_j^* en (4.2.1), (4.2.2), (4.3.1), (4.3.2), (4.4.4), (4.5.1) y (4.5.2) en el algoritmo dado por la Figura 4.6.1. El cálculo de los *p-values* ajustados por permutación para el método de paso hacia abajo *maxT*, resulta computacionalmente inefectivo, por ello Westfall y Young sugieren el uso de teoría asintótica para aproximarlos. Esta versión del método *maxT* es la usada para el análisis realizado en esta tesis.

Algoritmo de permutación para *p-values* ajustados del método de pasos hacia debajo de *maxT* basado en el algoritmo de Westfall y Young (1993, Algoritmo 4.1, p. 116 - 117)

Para los datos originales, se ordenan los estadísticos de prueba tal que $|t_{s_1}| \geq \cdots \geq |t_{s_m}|$. Para la b-iésima permutación, $b = 1,...,B$:

1. Se permutan las n columnas de la matriz de datos X.
2. Se calculan los estadísticos de prueba $t_{1,b},...,t_{m,b}$ para cada hipótesis.
3. Después, se calcula $u_{i,b} = \max_{l=1,...m} |t_{s_l,b}|$, el máximo sucesivo de los estadísticos de prueba esta dado por

$$u_{m,b} = |t_{r_m,b}|$$
$$u_{j,b} = \max(u_{j+1,b}, |t_{r_j,b}|),\ j = m-1,...,1$$

Los pasos anteriores se repiten B veces y lo *p-values* ajustados son estimados por

$$\tilde{p}^*_{r_j} = \frac{\sum_{b=1}^{B} I(u_{j,b} \geq |t_{r_j}|)}{B},$$

con los contrates de monotocidad asegurados por

$$p_{r_1} \leftarrow p_{r_1}, \quad p_{r_j} \leftarrow \max(p_{r_j}, p_{r_{j-1}}),\ j = 2,...,m.$$

Figura 4.6.2: Algoritmo de permutación para *p-values* ajustados, basado en el algoritmo de Westfall y Young (1993).

Capítulo 5
METODOLOGÍA Y RESULTADOS

Con el fin de encontrar el método de prueba múltiple más eficaz en la labor de identificar genes diferencialmente expresados; en este trabajo, se han considerado diez métodos de prueba múltiple de los descritos en el Capítulo 4, para identificar tales genes, se usaron seis conjuntos de datos de expresión genética provenientes de experimentos de microarreglos y disponibles gratuitamente en sitios de internet. Para ello se establecen simultáneamente para los *m* genes la hipótesis nula "H_j" $(j = 1,..., m)$ de no asociación entre la medida de expresión de los genes X_j y la respuesta Y. Se calculan los respectivos *p-values* ajustados para cada uno de los métodos de prueba de hipótesis múltiple, como se indicó en el capítulo anterior, usando el software Bioconductor, basado en el lenguaje de programación R. El paquete multtest de R, proporciona las listas de genes diferencialmente expresados para cada método.

5.1 Base de Datos

A continuación se da una breve descripción de las bases de datos que se utilizaron en esta tesis, y en Tabla 5.1, se da un resumen de sus características.

Base de Datos	***Apo AI***	***Leukemia***	***Prostate***	***Colon***	***Breastcc***	***SRBCT***
Genes	6384	3051	6033	2000	3226	2308
Muestras	16	38	102	62	22	63
Clases	2	2	2	2	3	4

Tabla 5.1: Características de las bases de datos.

5.1.1 *Apo AI*

Es un conjunto de expresión de 6384 genes, estudiados en 16 muestras. El gen de la *Apolipoproteína AI* (*Apo AI*) es conocido por desempeñar un papel central en la lipoproteína de alta densidad (HDL) del metabolismo. Ratones con el gen *Apo AI* inutilizado tienen muy bajos niveles de colesterol HDL. El objetivo del experimento fue el de identificar los genes con expresión alterada de la Apo AI en el hígado. El tratamiento previo que se realiza a estos datos se describe en Dudoit et. al 2002.

Este conjunto de datos está disponible en el paquete *DEDS* de R, pero debido a que posee valores faltantes, tales valores se imputan por el método *Gene Average* (*Troyanskaya et, al.* [44]), el cual calcula el valor faltante para un gen promediando los valores conocidos del gen. Existen otros métodos más elaborados y rigurosos de imputación, pero para el objetivo de ésta tesis, el método *Gene Average* es suficiente.

5.1.2 *Leukemia*

Es un conjunto de datos de expresión genética introducido por Golub et al. (1999) en su estudio de microarreglos para identificar y distinguir entre dos tipos de cáncer, leucemia linfoblástica aguda (*acute lymphocytic leukemia*, *ALL*) y leucemia mielógena aguda (*acute myelogenous leukemia*, *AML*). Consiste de los niveles de expresión de 7129 genes para 72 pacientes, de los cuales 47 son casos *ALL* y los otros 25 corresponden a casos *AML*. Después del preprocesamiento, hecho como se describe en Dudoit et al. (2002), solo se siguen usando 3051 genes. Las muestras fueron ensayadas utilizando los chips de Affimetrix Hgu6880, las cuales incluyen solo 27 casos *ALL* y 11 casos *AML*. Estos datos se encuentran en el paquete multtest de R.

5.1.3 *Prostate*

Este es un conjunto de datos ampliamente estudiado por Singh et al. (2002). Consiste de la expresión de 12600 genes para 50 tejidos normales de próstata y 52 tejidos de cáncer de próstata. Después de filtrar los datos, el conjunto se reduce a 6033 genes.

5.1.4 *Colon*

Es un conjunto de datos de expresión genética proveniente de experimentos de microarreglos de Affimetrix, ampliamente descrito en Alon et al. (1999) y disponible al público. Contiene los niveles de expresión de 6500 genes para 62 pacientes de dos clases, 22 pacientes son pacientes sanos y los otros tienen cáncer de colon; después de preprocesarlo los genes se reducen a 2000.

5.1.5 *Breast Cáncer* (*Breastcc*)

Este conjunto de datos contiene la expresión genética de 3226 genes para 22 muestras. Consta de tres distintas clases *sporadic*, *BRCA1* y *BRCA 2*. Fue publicada por Hendenfalk et al. (2001).

5.1.6 *SRBCT*

Este conjunto de datos de expresión genética obtenido por microarreglos punteados o *ADNc* se presenta en Kahn et al. (2001). Contiene los niveles de expresión de 6767 genes para 63 pequeñas células tumorales redondas y azules "*Small Round Blue Cells Tumor*" (*SRBCT*) de los pacientes pertenecientes a una de las 4 clases de tumor: *Ewing family of tumors* (*EWS*), *non-Hodgkin lymphoma* (*BL*), *neuroblastoma* (*NB*) y *habdomyosarcoma* (*RMS*). Después de preprocesarlos el conjunto se reduce a 2308 genes.

Los datos anteriores, a excepción de *Apo AI* y *Leukemia*, se obtuvieron a partir de la página web de Dettling (http://stat.ethz.ch/ dettling/bagboost.html).

5.2 Existencia de Genes Diferencialmente Expresados

Para evaluar la existencia de genes diferencialmente expresados para cada conjunto de datos con dos clases, se evaluó la normalidad de los mismos mediante un histograma y un q-q (*quantil-quantil*) *plot* de normalidad [11], mediante la función *plot.qqline* del paquete *sma*. En este último gráfico, los puntos corresponden a los genes con estadísticos de prueba menor o mayor que el punto crítico correspondiente al nivel de significancia establecido. A

partir de estos dos gráficos se asume que habrá pocos genes diferencialmente expresados e incluso podrá no haberlos, si en el *plot* los puntos son cercanos a la línea y si el histograma tiende a ser o es simétrico; de lo contrario, se puede concluir que existen genes con expresión diferenciada. Para los conjuntos con más de dos clases la normalidad se evalúa empleando el *volcanoplot*, a través de la función *volcanoplot* del paquete *limma*; este *plot* indica normalidad si existe simetría, es decir, si el *volcanoplot* es simétrico, entonces no existen genes diferenciados. A continuación se evalúa visualmente la existencia de genes diferenciados para todos los conjuntos de datos.

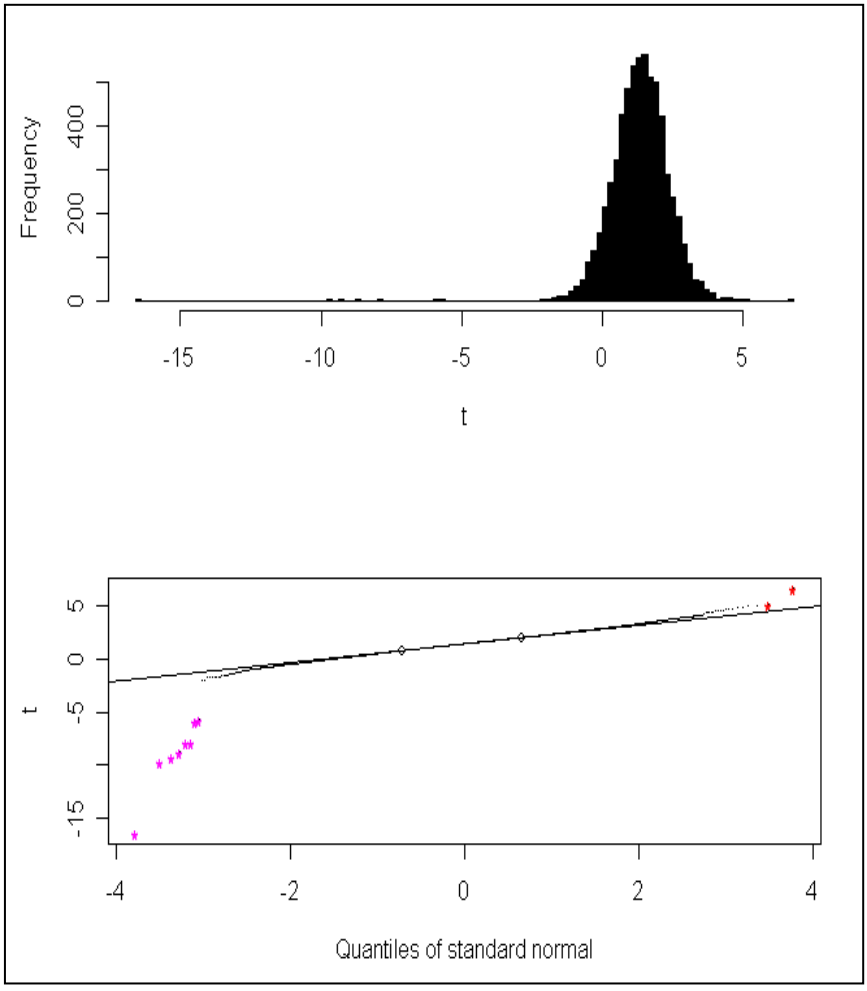


Figura 5.2.1: Histograma y *q-q plot* de normalidad para *Apo AI*.

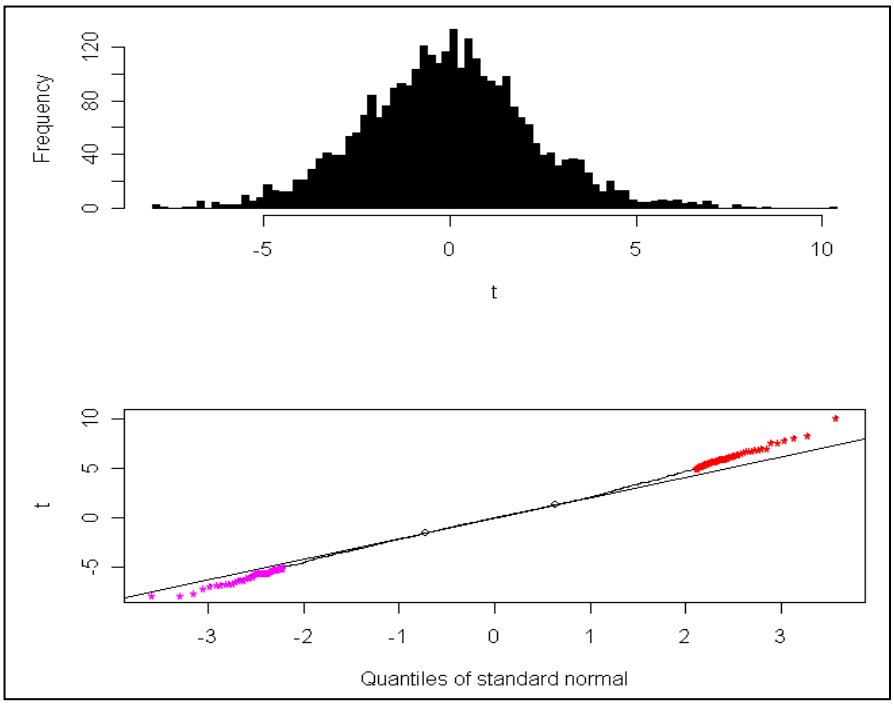


Figura 5.2.2: Histograma y *q-q plot* de normalidad para *Leukemia*.

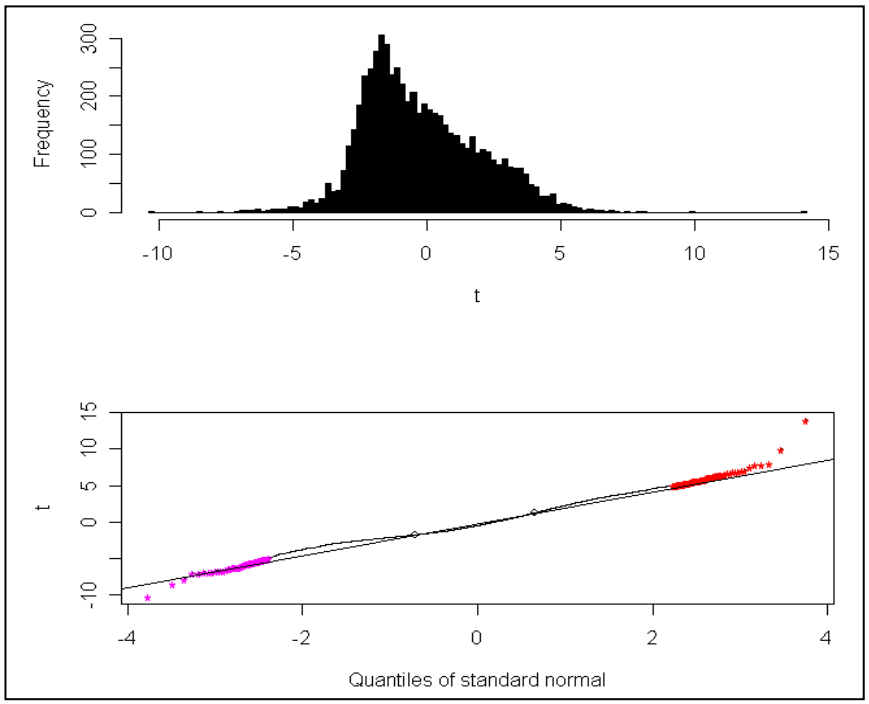


Figura 5.2.3: Histograma y *q-q plot* de normalidad para *Prostate*.

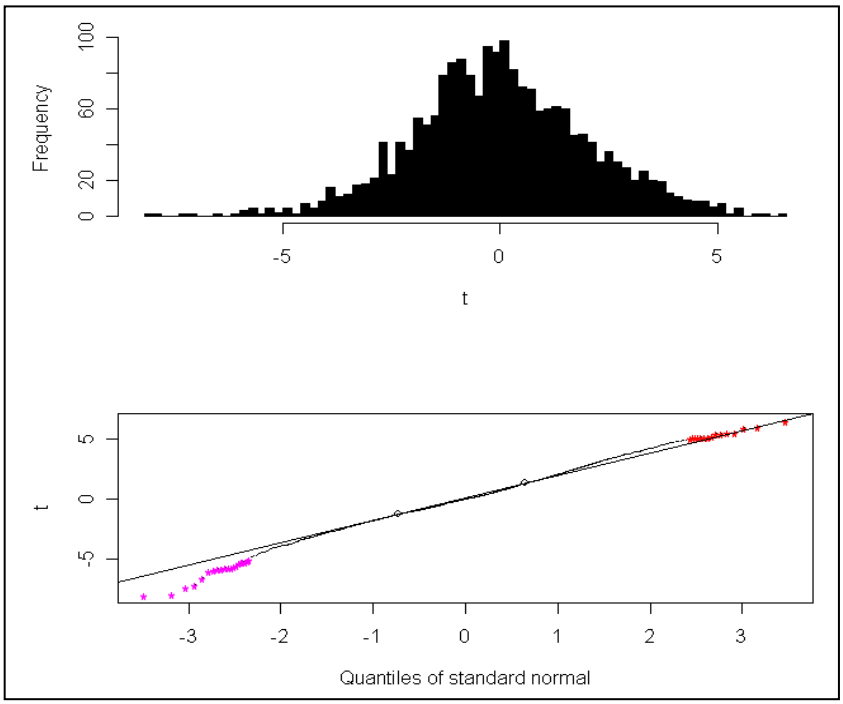


Figura 5.2.4: Histograma y *q-q plot* de normalidad para *Colon*.

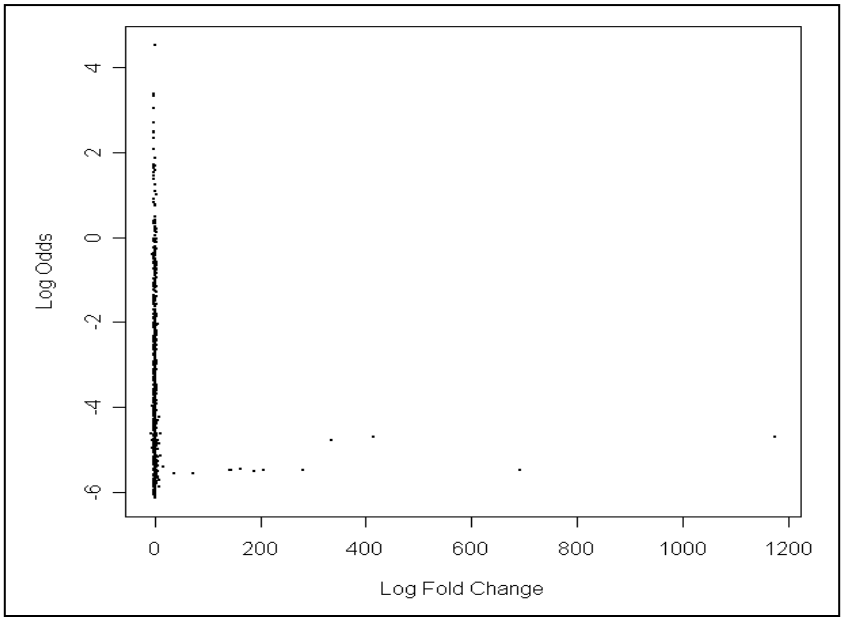


Figura 5.2.5: *Volcanoplot* para *Breastcc*.

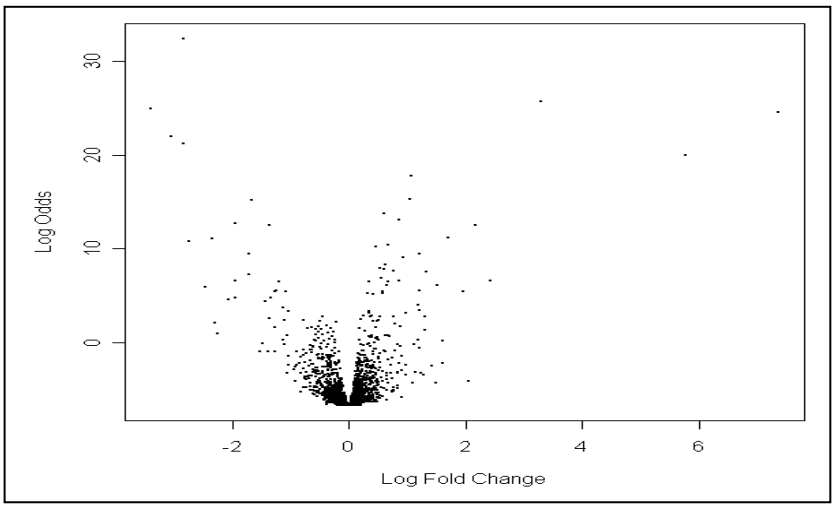


Figura 5.2.6: *Volcanoplot* para *SRBCT*.

El análisis de normalidad basado en las graficas anteriores indica que para todos los conjuntos de datos, existen genes con expresión diferenciada; en particular, por las características del histograma se aprecia que Prostate será el conjunto con mayor número de genes diferencialmente expresados y según el *volcanoplot*, *Breast* y *SRBCT* también.

5.3 Número de Hipótesis Nulas Rechazadas

Recordemos que en este trabajo el interés es analizar los métodos de prueba de hipótesis múltiple en identificación de genes diferencialmente expresados. Para tal fin y dentro de este contexto, se prueban simultáneamente las m hipótesis nulas H_j's de no asociación entre la medida de expresión de los genes X_j y la respuesta Y, las cuales permiten establecer aquellos genes con expresión diferenciada. Para ello se identifican tales genes para los conjuntos de datos antes descritos, empleando diez de los métodos descritos en el capítulo anterior. Cabe anotar que si H_j es rechazada, esto significa que el gen j es diferencialmente expresado. La decisión de rechazar o no un cierto número de las hipótesis nulas se toma, primero, calculando el estadístico de prueba para cada hipótesis j a partir de la distribución generada por los datos y, segundo estableciendo un nivel de significancia del 5% y comparándolo con los *p-values* calculados por algún método de prueba múltiple que controle el *FWER* y *FDR*. Debe recordarse que los *p-values* para el método *maxT* fueron calculados considerando un

número bastante grande de permutaciones, en este caso 100,000; para *SAM* se consideraron 1000 permutaciones.

A continuación se muestra el número de hipótesis rechazadas (o número de genes diferencialmente expresados) para cada uno de los seis conjuntos: Apo AI, Leukemia, Prostate, Colon, *SRBCT* y Breastcc; a un nivel de significancia del 5% (Ver Tabla 5.3).

	Bonferroni	**Holm**	**Hochberg**	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***t ó f***
Apo AI	6	6	6	6	6	9	6	7	8	1190
Leukemia	98	98	98	98	98	681	269	93	125	1045
Prostate	168	169	169	168	170	1349	520	239	298	2345
Colon	53	55	55	55	57	354	143	57	81	598
Breastcc	0	0	0	0	0	0	0	1	1	0
SRBCT	4	4	4	4	4	21	0	230	214	346

Tabla 5.3: Número de hipótesis nulas rechazadas a un nivel de significancia del 5%.

Por el análisis de normalidad expuesto en la sección anterior se sabe que para todos los conjuntos de datos existen genes diferencialmente expresados. Por tanto una vez comprobada la existencia de este tipo de genes, el interés se centra entonces en conocer la cantidad de tales genes. La Tabla 5.3 registra esta cantidad para cada uno de los seis conjuntos de datos aquí considerados, en ella se puede corroborar la afirmación hecha para Prostate a partir del análisis de normalidad. Además, establece que en general, entre los métodos que se basan en estadísticos independientes, los métodos de Bonferroni, Holm, Hochberg, *SD* (Šidák de un solo paso) y *SS* (Šidák de pasos hacia abajo), son los más conservativos, es decir, los que menos hipótesis nulas rechazan y por tanto los que declaran menos genes como diferenciados; a diferencia de los métodos basados t ó f (estadístico de prueba t para los conjuntos de datos con dos clases y estadístico de prueba f para aquellos con más de dos clases), los cuales son los menos conservativos. Los métodos por etapas *BY* (Benjamín y Yekutieli) y *maxT*, los cuales consideran una estructura de dependencia entre los estadísticos de prueba, son más conservativos que el método *BH* (Benjamín y Hochberg) que no considera tal estructura. El método por etapas, no paramétrico de *SAM*, proporciona resultados similares a todos los

otros métodos, a excepción de los métodos *BH* y *t* ó *f*. En particular, puesto que *SAM* al igual que *maxT* es un método calculado por permutación, sus resultados son bastante similares.

Los valores dados por la Tabla 5.3 son ciertos solo para un nivel de significancia del 5%; por ello se consideró entonces buscar el número de genes diferencialmente expresados para distintos niveles de significancia. A continuación se muestran los resultados obtenidos.

5.3.1 Resultados para *Apo AI*

α	**Bonferroni**	**Holm**	**Hochberg**	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***T***
0.05	6	6	6	6	6	9	6	7	8	1190
0.1	7	7	7	7	7	10	7	9	11	2127
0.15	7	7	7	7	7	14	7	9	12	2760
0.2	7	7	7	7	7	28	7	10	14	3278
0.25	8	8	8	8	8	949	7	12	18	3681
0.3	8	8	8	9	9	2126	8	14	19	4028
0.35	8	8	8	9	9	2785	8	17	21	4315
0.4	9	9	9	9	9	3379	9	20	24	4584
0.45	9	9	9	9	9	3842	9	23	27	4815
0.5	9	9	9	9	9	4194	9	28	31	5011
0.55	9	9	9	9	9	4530	9	28	36	5205
0.6	9	9	9	9	9	4843	9	31	46	5357
0.65	9	9	9	10	10	5065	9	36	50	5519
0.7	9	9	9	10	10	5299	9	47	57	5644
0.75	9	9	9	12	12	5516	9	56	72	5766
0.8	9	9	9	12	12	5675	9	70	92	5898
0.85	9	9	9	13	13	5833	9	90	120	6023
0.9	9	9	9	14	14	6022	10	119	157	6143
0.95	10	10	10	17	17	6198	10	153	222	6270
1	6384	6384	6384	6384	6384	6384	6384	6384	6384	6384

Tabla 5.3.1: Número de genes diferencialmente expresados para *Apo AI*.

A continuación se muestra gráficamente el comportamiento de los distintos métodos de prueba múltiple considerados en este trabajo, a un nivel de significancia del 5%.

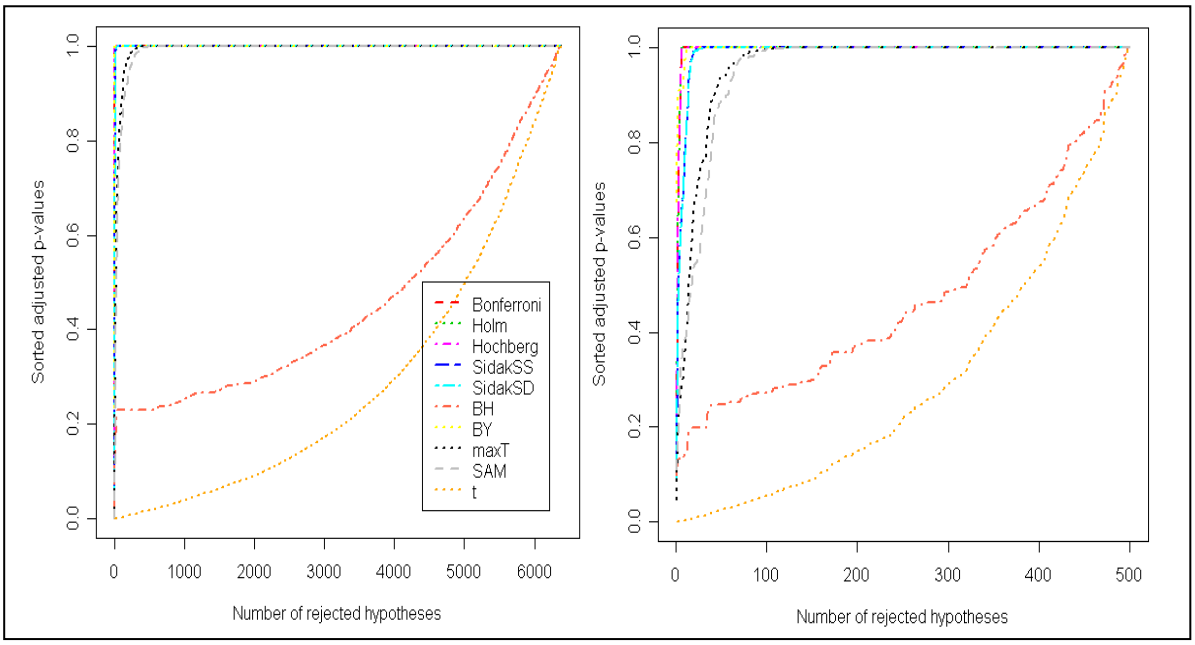


Figura 5.3.1(a): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I para *Apo AI*.

Figura 5.3.1(b): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I, para 500 hipótesis nulas.

Puesto que la Figura 5.3.1(a) considera el máximo de hipótesis nulas que pueden ser rechazadas, no brinda una apreciación clara del comportamiento de los distintos métodos, por tal motivo y para tener una mejor apreciación se prueban únicamente 500 hipótesis nulas seleccionando una muestra aleatoria de 500 genes. Los resultados obtenidos se ilustran en la Figura 5.3.1(b). Para la muestra, las permutaciones hechas para el métodos *maxT* fue 10,000.

La información dada por la Tabla 5.3.1 y la Figura 5.3.1(b), para diferentes niveles de significancia, indican que en efecto los métodos más conservativos son los de Bonferroni, Holm, Hochberg, *SS*, *SD*, *maxT*, y los menos conservativos son los métodos *BH*, *BY*, *SAM* y *t*.

Al examinar estos métodos de prueba múltiple para los demás conjuntos de datos, de forma similar como se hizo con Apo AI, puede constatarse que los métodos que declaran menos genes como diferenciados son los de Bonferroni, Holm, Hochberg, *SS*, *SD* y *maxT*, y los que declaran más genes como diferencialmente expresados son los de *BH*, *BY*, *SAM* y *t*.

5.3.2 Resultados para *Leukemia*

α	Bonferroni	Holm	Hochberg	*SS*	*SD*	*BH*	*BY*	*maxT*	*SAM*	*T*
0.05	98	98	98	98	98	681	269	93	125	1045
0.1	126	126	126	126	129	876	392	119	145	1295
0.15	143	143	143	144	146	1069	466	133	166	1488
0.2	153	155	155	159	160	1204	528	150	188	1640
0.25	161	163	163	169	171	1325	564	163	202	1767
0.3	170	171	171	177	179	1459	597	177	224	1868
0.35	176	179	179	186	191	1583	629	187	243	1975
0.4	182	186	186	200	204	1703	663	202	265	2062
0.45	190	194	194	207	210	1812	691	223	282	2148
0.5	196	201	201	213	219	1889	716	237	297	2255
0.55	204	206	206	221	226	1996	737	256	318	2346
0.6	207	210	210	233	243	2077	769	265	331	2431
0.65	210	213	213	245	249	2191	790	286	350	2518
0.7	214	219	219	251	258	2308	811	312	375	2601
0.75	219	222	222	259	268	2420	828	327	394	2677
0.8	221	226	226	269	277	2550	850	357	408	2744
0.85	226	233	233	284	294	2668	874	376	430	2816
0.9	231	242	242	304	315	2778	890	408	473	2891
0.95	239	244	244	330	353	2899	913	448	507	2975
1	3051	3051	3051	3051	3051	3051	3051	3051	3051	3051

Tabla 5.3.2: Número de genes diferencialmente expresados para *Leukemia*.

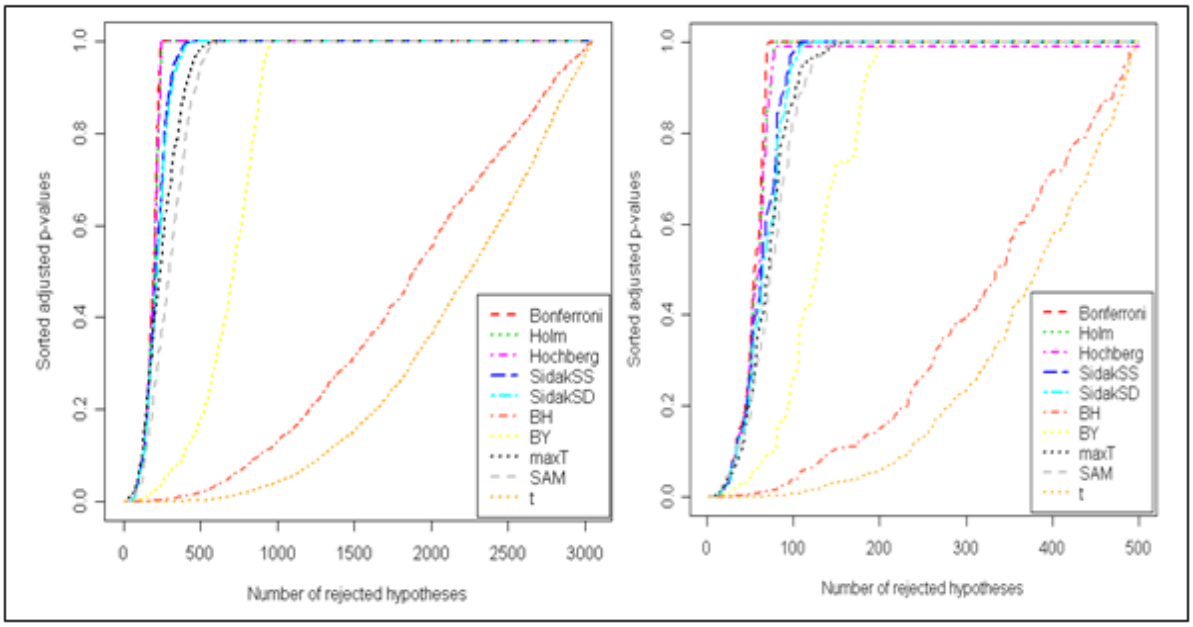


Figura 5.3.2(a): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I para *Leukemia*.

Figura 5.3.2(b): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I, para 500 hipótesis nulas.

5.3.3 Resultados para *Prostate*

α	Bonferroni	Holm	Hochberg	*SS*	*SD*	*BH*	*BY*	*maxT*	*SAM*	*T*
0.05	168	169	169	168	170	1349	520	239	298	2345
0.1	203	204	204	207	208	2017	667	293	360	2996
0.15	230	230	230	234	235	2541	782	334	425	3398
0.2	245	247	247	249	251	2983	885	364	470	3721
0.25	252	254	254	263	271	3309	989	389	512	3932
0.3	270	273	273	284	286	3627	1084	420	541	4173
0.35	283	285	285	300	304	3826	1178	460	576	4366
0.4	290	299	299	310	316	4002	1243	482	603	4524
0.45	303	307	307	327	331	4243	1333	519	633	4684
0.5	309	314	314	339	343	4422	1419	550	683	4826
0.55	318	324	324	351	355	4580	1503	578	704	4954
0.6	327	331	331	364	372	4753	1551	605	733	5094
0.65	333	339	339	380	384	4909	1624	629	756	5231
0.7	340	345	345	389	398	5050	1711	650	798	5342
0.75	346	351	351	404	410	5233	1784	671	839	5464
0.8	352	356	356	420	435	5370	1846	706	882	5572
0.85	356	364	364	450	463	5530	1918	742	914	5682
0.9	363	371	371	482	497	5679	1983	800	961	5800
0.95	371	377	377	527	538	5843	2037	872	1051	5914
1	6033	6033	6033	6033	6033	6033	6033	6033	6033	6033

Tabla 5.3.3: Número de genes diferencialmente expresados para *Prostate*.

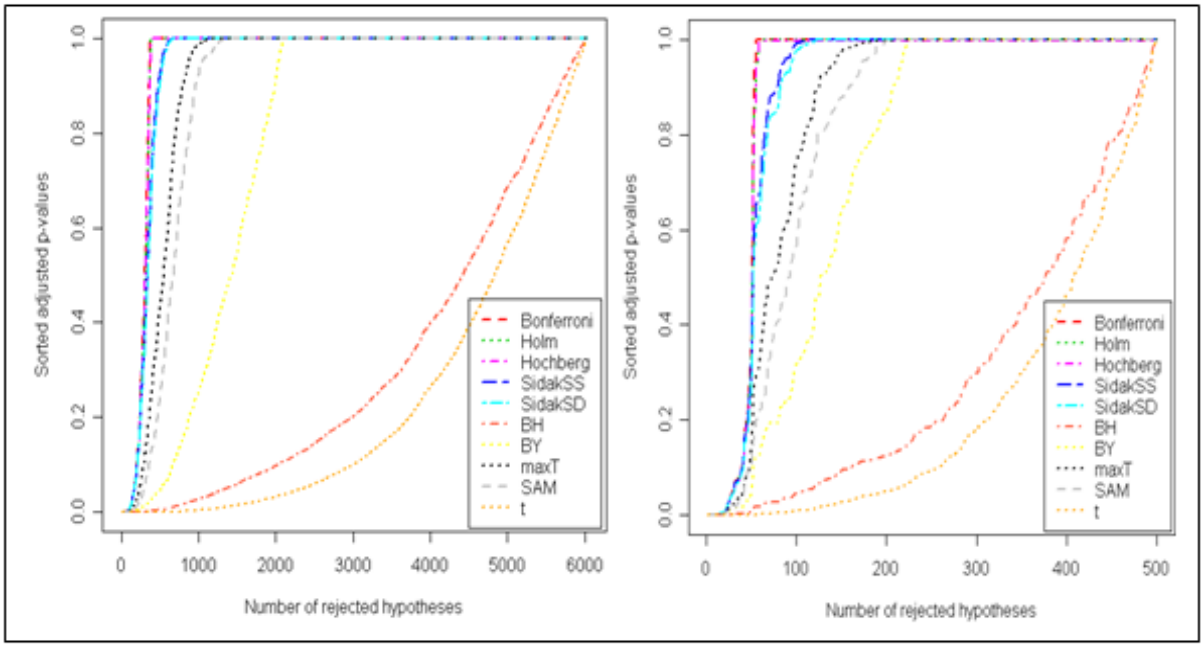


Figura 5.3.3(a): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I para *Prostate*.

Figura 5.3.3(b): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I, para 500 hipótesis nulas.

5.3.4 Resultados para *Colon*

α	Bonferroni	Holm	Hochberg	*SS*	*SD*	*BH*	*BY*	*maxT*	*SAM*	*T*
0.05	53	55	55	55	57	354	143	57	81	598
0.1	70	70	70	70	70	478	188	82	101	761
0.15	74	77	77	78	79	563	230	100	111	884
0.2	84	85	85	88	91	661	253	122	125	988
0.25	91	93	93	94	96	748	277	135	138	1078
0.3	96	96	96	100	101	805	302	149	144	1167
0.35	100	101	101	105	105	899	319	157	154	1236
0.4	103	105	105	117	122	982	348	175	165	1311
0.45	105	110	110	126	126	1060	369	184	179	1377
0.5	116	120	120	129	130	1137	394	200	187	1434
0.55	123	124	124	137	141	1217	404	207	204	1492
0.6	126	126	126	147	148	1299	420	218	216	1553
0.65	126	129	129	151	153	1369	435	227	230	1601
0.7	129	131	131	157	160	1449	444	249	242	1640
0.75	131	137	137	161	166	1534	468	263	258	1691
0.8	137	141	141	169	172	1589	475	282	275	1760
0.85	139	147	147	173	180	1635	490	304	289	1824
0.9	144	148	148	188	193	1734	500	325	314	1874
0.95	147	150	150	201	214	1856	509	362	349	1934
1	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000

Tabla 5.3.4: Número de genes diferencialmente expresados para *Colon*.

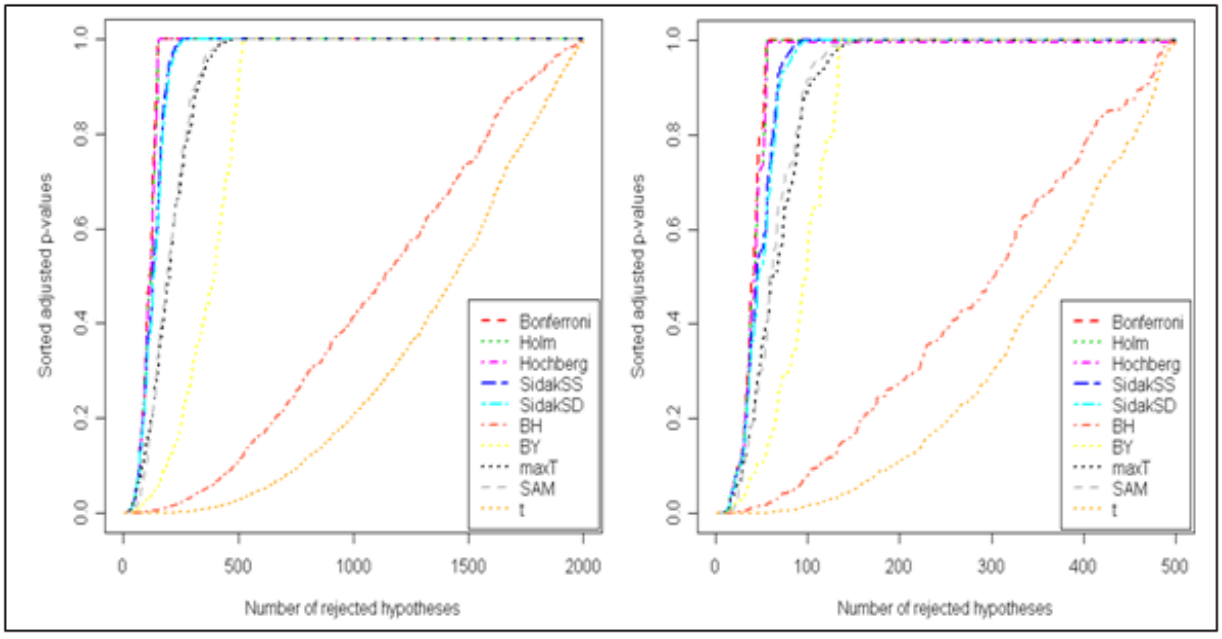


Figura 5.3.4(a): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I para *Colon*.

Figura 5.3.4(b): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I, para 500 hipótesis nulas.

5.3.5 Resultados para *Breastcc*

α	Bonferroni	Holm	Hochberg	*SS*	*SD*	*BH*	*BY*	*maxT*	*SAM*	*T*
0.05	0	0	0	0	0	0	0	1	1	0
0.1	0	0	0	0	0	0	0	1	1	15
0.15	0	0	0	0	0	0	0	2	4	72
0.2	0	0	0	0	0	0	0	5	7	207
0.25	0	0	0	0	0	0	0	8	11	399
0.3	0	0	0	0	0	0	0	13	14	634
0.35	0	0	0	0	0	0	0	14	18	931
0.4	0	0	0	0	0	0	0	17	21	1247
0.45	0	0	0	0	0	0	0	23	25	1563
0.5	0	0	0	0	0	0	0	25	27	1870
0.55	0	0	0	0	0	0	0	27	32	2136
0.6	0	0	0	0	0	0	0	29	34	2379
0.65	0	0	0	0	0	0	0	35	40	2589
0.7	0	0	0	0	0	0	0	42	47	2770
0.75	0	0	0	0	0	0	0	48	52	2911
0.8	0	0	0	0	0	0	0	53	55	2999
0.85	0	0	0	0	0	2975	0	56	56	3085
0.9	0	0	0	0	0	3110	0	65	67	3145
0.95	0	0	0	0	0	3170	0	85	93	3186
1	3226	3226	3226	3226	3226	3226	3226	3226	3226	3226

Tabla 5.3.5: Número de genes diferencialmente expresados para *Breastcc*.

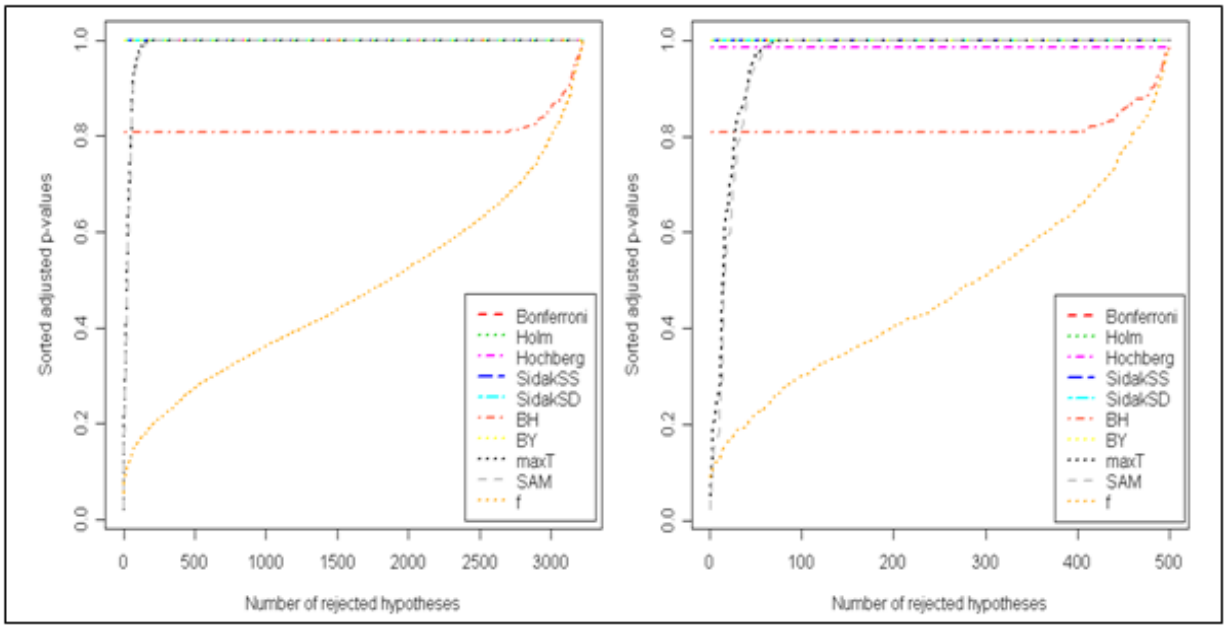


Figura 5.3.5(a): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I para *Breastcc*.

Figura 5.3.5(b): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I, para 500 hipótesis nulas.

5.3.6 Resultados para *SRBCT*

A	Bonferroni	Holm	Hochberg	*SS*	*SD*	*BH*	*BY*	*maxT*	*SAM*	*T*
0.05	4	4	4	4	4	21	0	230	214	346
0.1	6	6	6	6	6	46	4	269	267	627
0.15	7	7	7	8	8	78	6	297	298	922
0.2	8	8	8	9	9	109	13	325	331	1168
0.25	9	9	9	11	11	171	17	349	347	1392
0.3	11	11	11	14	14	252	19	367	357	1591
0.35	13	13	13	14	14	477	21	392	390	1736
0.4	14	14	14	17	17	1216	21	418	412	1875
0.45	14	14	14	18	18	1687	22	441	434	1974
0.5	17	17	17	19	19	1888	23	458	456	2050
0.55	17	17	17	20	20	2038	32	475	473	2110
0.6	18	18	18	21	21	2108	33	491	501	2162
0.65	19	19	19	21	21	2169	34	517	519	2195
0.7	19	19	19	22	22	2226	35	535	540	2234
0.75	20	20	20	23	23	2254	39	552	562	2260
0.8	20	20	20	24	24	2273	45	576	601	2278
0.85	21	21	21	27	27	2287	48	608	631	2290
0.9	21	21	21	32	32	2296	57	656	668	2297
0.95	21	21	21	35	35	2307	59	735	734	2307
1	2308	2308	2308	2308	2308	2308	2308	2308	2308	2308

Tabla 5.3.6: Número de genes diferencialmente expresados para *SRBCT*.

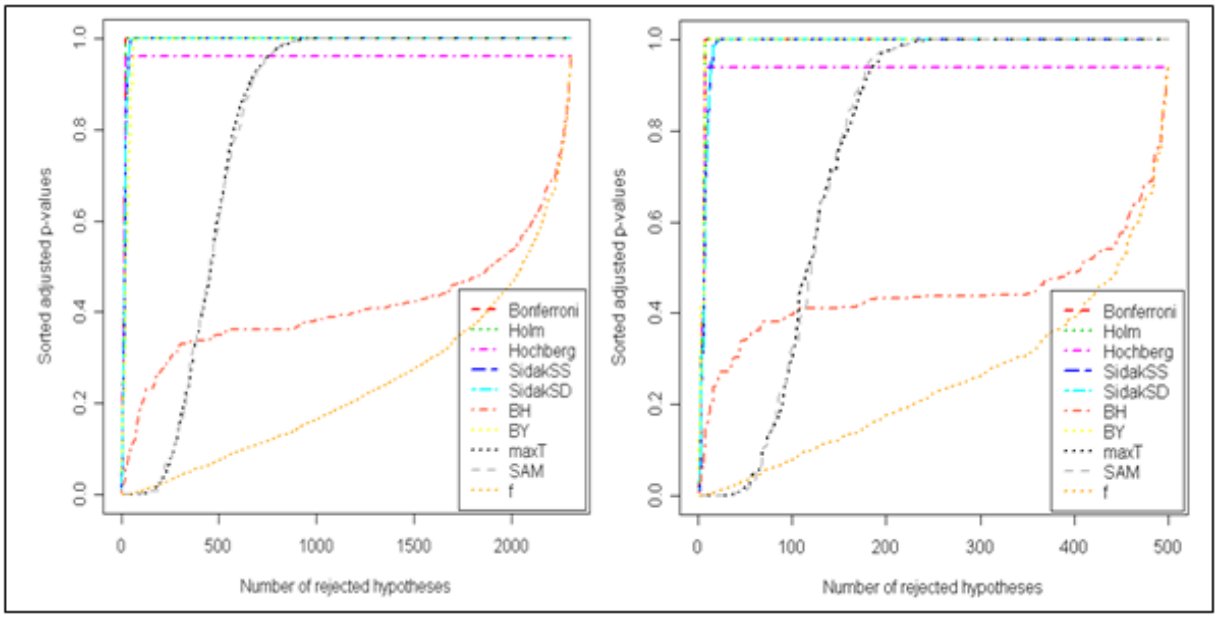


Figura 5.3.6(a): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I para *SRBCT*.

Figura 5.3.6(b): *Plot* de número de hipótesis rechazadas versus tasas de Error Tipo I, para 500 hipótesis nulas.

5.4 Número de Genes Diferencialmente Expresados Obtenidos por Simulación.

Debido a que cada método depende de los supuestos hechos en su definición, hasta el momento en este trabajo sólo se establece experimentalmente qué métodos de prueba múltiple de los examinados son menos o más conservativos. Por esta razón, para tratar de validar los resultados obtenidos y teniendo en cuenta que los métodos empleados para probar las hipótesis nulas múltiples se basaron en los estadísticos de prueba t y F, los cuales asumen normalidad, se simula una muestra de 500 genes con distribución normal multivariada (no Gaussiana) para cada conjunto de datos y se busca reiteradamente (500 veces) el número de genes a ser expresados como diferenciados a un nivel de significancia del 5%. Al final solo se da un promedio de la cantidad de genes diferencialmente expresados obtenidos de cada una de las 500 simulaciones realizadas. La muestra de la que en esta sección se habla tiene por vector de medias y matriz de covarianza los correspondientes a los de una muestra de 500 genes seleccionada aleatoriamente del conjunto completo de datos. También se calcula el número de genes con expresión diferenciada simulando una muestra aleatoria de 500 genes con distribución Gaussiana, pero debido a que una distribución Gaussiana garantiza la no existencia de genes con expresión diferenciada, a continuación solo se muestran los resultados obtenidos para la distribución normal multivariada.

***Plot* del número de hipótesis rechazadas versus tasas de Error Tipo I, para una muestra aleatoria de 500 genes con distribución Normal Multivariada.**

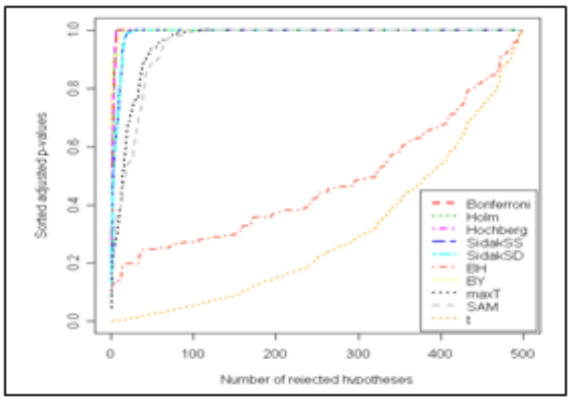


Figura 5.4.1: Comportamiento de métodos de prueba múltiple vs número de hipótesis, para *Apo AI*.

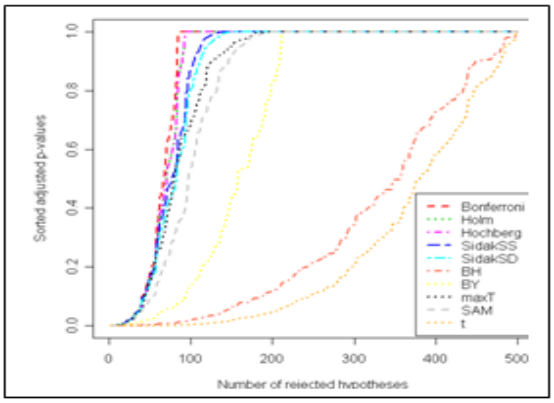


Figura 5.4.2: Comportamiento de métodos de prueba múltiple vs número de hipótesis, para *Leukemia*.

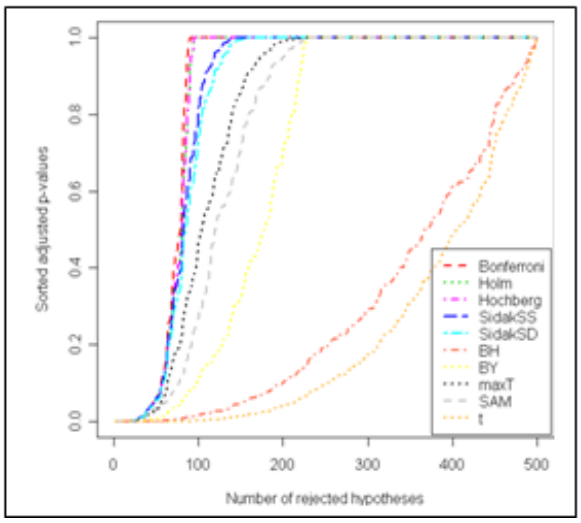


Figura 5.4.3: Comportamiento de métodos de prueba múltiple vs número de hipótesis, para *Prostate*.

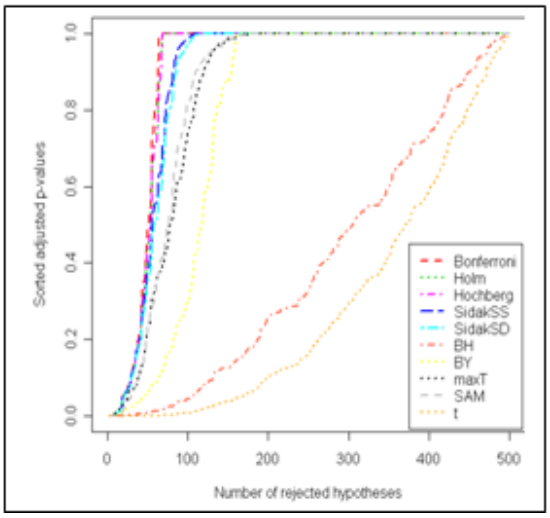


Figura 5.4.4: Comportamiento de métodos de prueba múltiple vs número de hipótesis, para *Colon*.

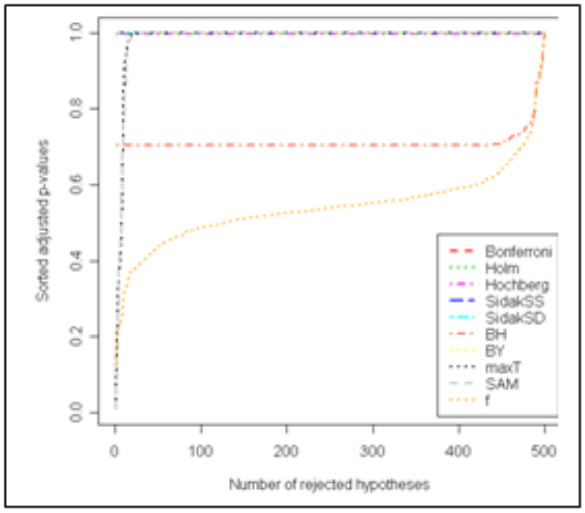


Figura 5.4.5: Comportamiento de métodos de prueba múltiple vs número de hipótesis, para *Breastcc*.

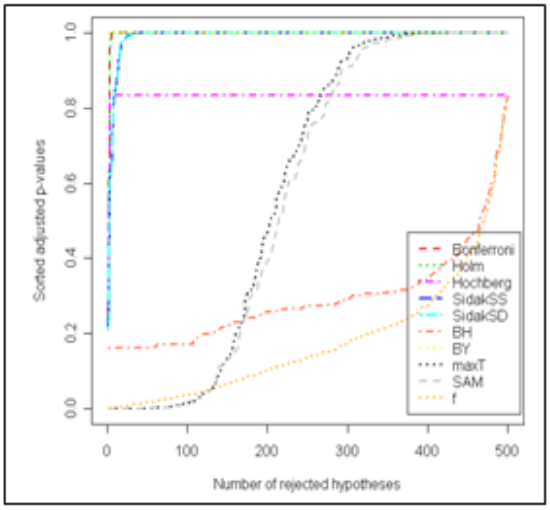


Figura 5.4.6: Comportamiento de métodos de prueba múltiple vs número de hipótesis, para *SRBCT*.

Las Figuras (5.4.1) - (5.4.6), permiten constatar lo antes afirmado sobre aquellos métodos que suelen ser menos o más conservativos.

La Tabla 5.4 contiene para cada conjunto de datos el número de genes diferencialmente expresados calculados reiteradamente para una muestra aleatoria de 500 genes, otra muestra simulada con distribución normal multivariada y por último una muestra con distribución Gaussiana; aunque para ésta última muestra los resultados obtenidos solo se muestran para los métodos *t* ó *F*, pues para los demás métodos se obtiene que no hay genes diferencialmente expresados. Esta tabla pone de manifiesto que los métodos que controlan el *FWER* (Bonferroni, Holm, Hochberg, *SS*, *SD* y *maxT*) son más conservativos que los métodos que controlan el *FDR* (*BH*, *BY* y *SAM*), como era de esperarse según lo establecido por la desigualdad (∗) de la Sección 3.3.1. Por los resultados obtenidos para las tres muestras y en especial aquella con distribución Gaussiana, se puede afirmar que los métodos *t* ó *F* que controlan tanto el *FWER* como el *FDR* suelen ser generalmente los menos conservativos de todos los métodos evaluados.

	Bonferrini		*Holm*		*Hochberg*		*SS*		*SD*		*BH*		*BY*		*maxT*		*SAM*		*t*	*ó*	*f*
Data	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M	S.N.M	M.G
Apo Imp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	102	15	102
Leukemia	33	37	33	38	33	38	33	37	34	38	108	176	58	85	32	34	31	40	162	210	30
Prostate	28	35	28	35	28	35	28	35	28	35	111	142	49	54	34	38	35	41	206	223	22
Colon	16	25	16	26	16	26	16	25	16	26	65	119	28	51	18	26	20	30	144	171	27
Breastcc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1
Srbct	7	3	7	3	7	3	7	3	7	3	30	39	8	0	62	115	44	67	118	203	3

Tabla 5.4 Número de Genes Diferencialmente Expresados obtenidos por Simulación, para una muestra aleatoria de 500 genes con distribución Normal Multivariada, a un nivel de significancia del 5%.

5.5 Genes Diferencialmente Expresados

Ya en las secciones anteriores de este capítulo se ha examinado primero, la existencia de genes diferencialmente expresados y segundo, el número de ellos. Y para estar seguros de que los resultados obtenidos por los métodos considerados son confiables, se realizó un proceso de simulación, como lo explica la Sección 5.4. Ahora el interés se centra entonces en saber cuáles de esos son los genes diferencialmente expresados. A continuación, teniendo en cuenta el número de genes declarados con expresión diferenciada dados por la Tabla 5.3,

la Tabla 5.5 presenta hasta los diez primeros genes expresados como diferenciados por los métodos de prueba múltiple considerados, para cada uno de los conjuntos de datos (completos).

	Bonferroni	***Holm***	***Hochberg***	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***T***
1	2149	2149	2149	2149	2149	2149	2149	**2149**	**2149**	2149
2	1739	1739	1739	1739	1739	1739	1739	**1739**	**540**	1739
3	5356	5356	5356	5356	5356	5356	5356	**5356**	**5356**	5356
4	540	540	540	540	540	540	540	**540**	**1739**	540
5	4139	4139	4139	4139	4139	4139	4139	**4139**	**4139**	4139
6	2537	2537	2537	2537	2537	2537	2537	**2537**	**2537**	2537
7						2		**2**	**2**	2
8						4941			**4941**	4941
9						1496				1496
10										1204

Tabla 5.5.1: Top Genes Diferencialmente Expresados para *Apo AI*.

	Bonferroni	***Holm***	***Hochberg***	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***T***
1	829	829	829	829	829	829	829	**2124**	**829**	829
2	378	378	378	378	378	378	378	**829**	**378**	378
3	2124	2124	2124	2124	2124	2124	2124	896	**1009**	2124
4	808	808	808	808	808	808	808	766	**2124**	808
5	2489	2489	2489	2489	2489	2489	2489	2600	**2670**	2489
6	394	394	394	394	394	394	394	2939	**2489**	394
7	2670	2670	2670	2670	2670	2670	2670	**1995**	2663	2670
8	1009	1009	1009	1009	1009	1009	1009	2386	**1995**	1009
9	1995	1995	1995	1995	1995	1995	1995	717	**808**	1995
10	937	937	937	937	937	937	937	**2489**	2664	937

Tabla 5.5.2: Top Genes Diferencialmente Expresados para *Leukemia*.

	Bonferroni	***Holm***	***Hochberg***	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***T***
1	2619	2619	2619	2619	2619	2619	2619	**2619**	**2619**	2619
2	5016	5016	5016	5016	5016	5016	5016	**5016**	**1839**	5016
3	1839	1839	1839	1839	1839	1839	1839	**1839**	**5016**	1839
4	4701	4701	4701	4701	4701	4701	4701	**4701**	**4701**	4701
5	4155	4155	4155	4155	4155	4155	4155	**4155**	**4155**	4155
6	3934	3934	3934	3934	3934	3934	3934	**3934**	**2425**	3934
7	2425	2425	2425	2425	2425	2425	2425	**2425**	**5808**	2425
8	1640	1640	1640	1640	1640	1640	1640	**1640**	5249	1640
9	5808	5808	5808	5808	5808	5808	5808	**5808**	3665	5808
10	4849	4849	4849	4849	4849	4849	4849	**3705**	3705	4849

Tabla 5.5.3: Top Genes Diferencialmente Expresados para *Prostate*.

	Bonferroni	**Holm**	**Hochberg**	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***T***
1	493	493	493	493	493	493	493	**493**	**493**	493
2	377	377	377	377	377	377	377	**1635**	**249**	377
3	249	249	249	249	249	249	249	**377**	**1423**	249
4	1635	1635	1635	1635	1635	1635	1635	**249**	**1635**	1635
5	1423	1423	1423	1423	1423	1423	1423	**1423**	**377**	1423
6	625	625	625	625	625	625	625	1843	**765**	625
7	245	245	245	245	245	245	245	**625**	**897**	245
8	1771	1771	1771	1771	1771	1771	1771	1042	1494	1771
9	765	765	765	765	765	765	765	513	1843	765
10	1772	1772	1772	1772	1772	1772	1772	897	**625**	1772

Tabla 5.5.4: Top Genes Diferencialmente Expresados para *Colon*.

maxT	***SAM***
1443	1443

Tabla 5.5.5: Top Genes Diferencialmente Expresados para *Breastcc*.

	Bonferroni	**Holm**	**Hochberg**	***SS***	***SD***	***BH***	***BY***	***maxT***	***SAM***	***f***
1	1389	1389	1389	1389	1389	1389		**1389**	**1389**	1389
2	742	742	742	742	742	742		**742**	**742**	742
3	123	123	123	123	123	123		**123**	**1158**	123
4	1158	1158	1158	1158	1158	1158		**1158**	**846**	1158
5						846		**846**	**246**	846
6						246		**246**	**545**	246
7						545		**545**	**1955**	545
8						1955		**1955**	**123**	1955
9						1606		**1606**	509	1606
10						335		**335**	**1606**	335

Tabla 5.5.6: Top Genes Diferencialmente Expresados para *SRBCT*.

En las tablas anteriores se puede apreciar que los genes declarados como diferencialmente expresados por los métodos Bonferroni, Holm, Hochberg, Šidák, Benjamini y Hochberg, Benjamini y Yekutieli, y *t* ó *f*, son los mismos. Por lo que concierne al método *maxT* y *SAM*, como los genes que identifican como diferenciados no coinciden totalmente con los identificados por los métodos anteriores para los seis conjuntos de datos, se puede afirmar que estos métodos suelen declarar como genes con expresión diferenciada casi los mismos genes que los métodos anteriores. Sin embargo, podría pensarse que esto es algo que depende

del número de permutaciones realizadas para el cálculo de sus *p-valores* ajustados, o de las características de los datos.

5.1 Correlación

Los resultados obtenidos en las secciones anteriores de este capítulo han permitido asignar y constatar algunas características de los métodos de prueba múltiple considerados; sin embargo, al identificar los genes diferencialmente expresados existen diferencias entre algunos métodos, lo que conduce a revisar entonces la estructura de los datos. Además, cabe recordar que algunos de los métodos examinados consideran una estructura de dependencia, característica de los datos de expresión genética. Por estas razones, a continuación se evalúa la dependencia entre los estadísticos de prueba examinando la correlación entre las muestras de *ARN* (columnas de la matriz de datos) y entre los genes, para los seis conjuntos de datos. Pero debido a que nuestros datos son de alta dimensionalidad, tal examen se hará gráficamente y se examinará primero la correlación entre la muestras de *ARN* y segundo, la correlación de una muestra aleatoria de 500 genes sin y con distribución normal multivariada, estas distribuciones se obtienen como se explicó en la sección 5.4.

A continuación, las diferentes gamas de color indican alta correlación positiva o negativa, o la no existencia de la misma; si el color es rojo, azul o blanco, respectivamente.

5.6.1 Correlación para *Apo AI*

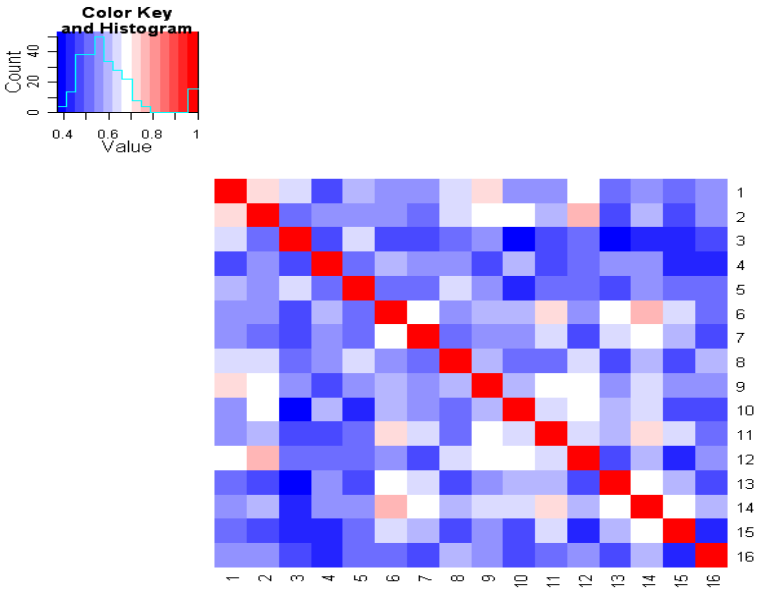


Figura 5.6.1(a): Correlación para muestras de *ARN* en *Apo AI*.

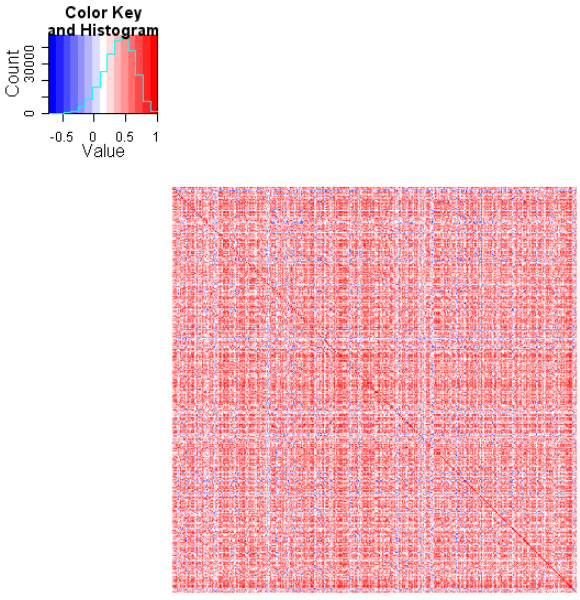


Figura 5.6.1(b): Correlación para muestra aleatoria de 500 genes. *Apo AI*.

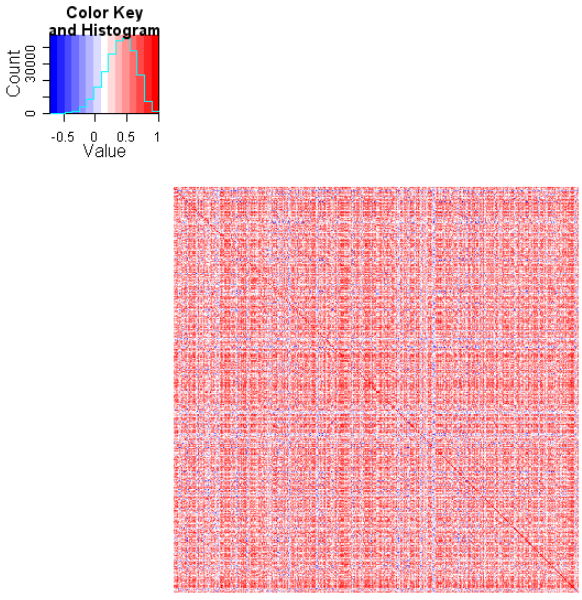


Figura5.6.1(c): Correlación para muestra aleatoria normal multivariada de 500 genes. *Apo AI*.

La Figura 5.6.1(a) indica que la correlación para las muestras de *ARN* en *Apo AI* es altamente negativa, lo que lleva a concluir que para identificar los genes diferencialmente expresados de este conjunto de datos resulta conveniente emplear un método de prueba múltiple que tome en cuenta una estructura de dependencia entre los genes. Esta afirmación es respaldada por los resultados presentes en las Figuras 5.6.1(b) y 5.6.1(c), las cuales indican que la correlación de los 500 genes seleccionados aleatoriamente con y sin distribución normal multivariada tienen dependencia aproximadamente positiva. Luego resulta conveniente aplicar métodos como *BY* y *maxT*, lo cual puede constatarse revisando los resultados de la Tabla 5.5.1, la cual indica que tanto el método *BY* como el método *maxT* declaran como genes diferencialmente expresados los mismos genes que los demás métodos. Aunque si lo que se quiere es un método no paramétrico que proporciones buenos resultados, también se puede optar por usar *SAM*.

5.6.2 Correlación para *Leukemia*

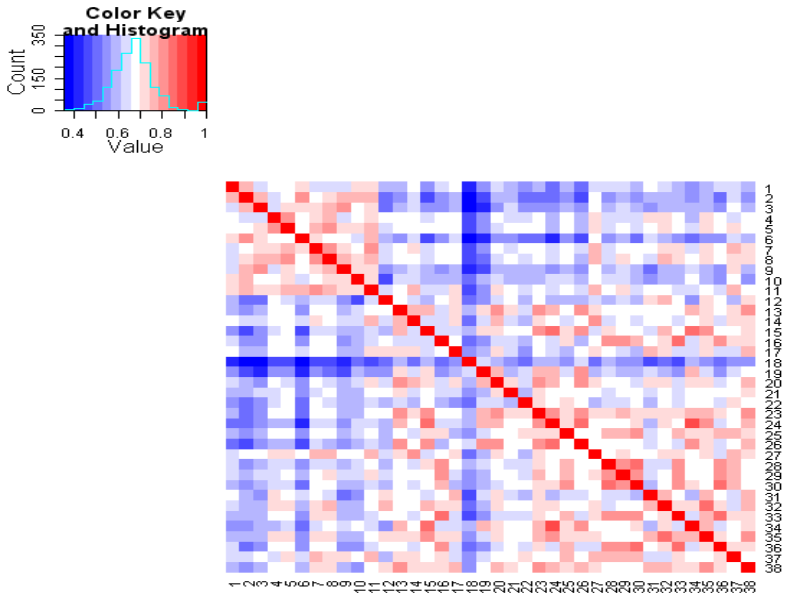


Figura 5.6.2(a): Correlación para muestras de *ARN* en *Leukemia*.

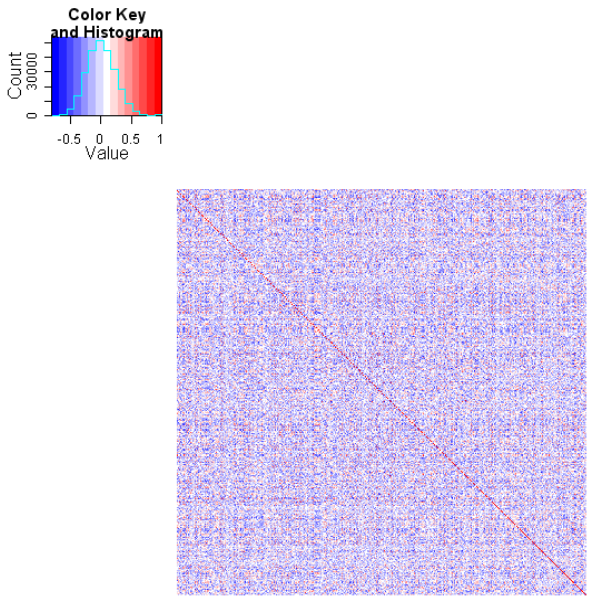


Figura 5.6.2(b): Correlación para muestra aleatoria de 500 genes. *Leukemia*.

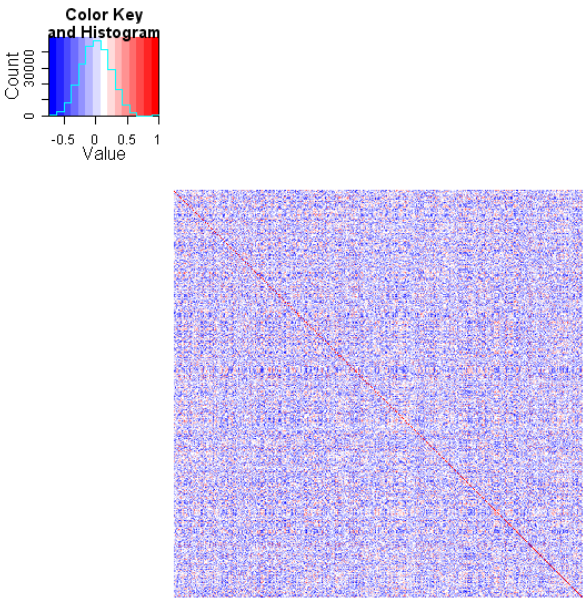


Figura 5.6.2(c): Correlación para muestra aleatoria normal multivariada de 500 genes. *Leukemia*.

De las Figuras 5.6.2(a), 5.6.2(b) y 5.6.2(c) puede concluirse que tanto las muestras de *ARN* como los genes seleccionados para *Leukemia* no son dependientes entre sí, por lo que podría pensarse que para seleccionar los genes con expresión diferenciada en este conjunto de datos deben considerarse métodos basados en estadísticos independientes, como es el caso de los métodos de Bonferroni, Šidák, Holm, Hochberg, *BH* y *t*. El hecho de que no convenga usar métodos como *BY*, *maxT* y *SAM* se nota en los resultados obtenidos en la Tabla 5.5.2, en la cual se observa que los genes declarados como diferenciados por esto métodos no coinciden en su totalidad con los declarados por los métodos que se aconseja usar.

5.6.3 Correlación para *Prostate*

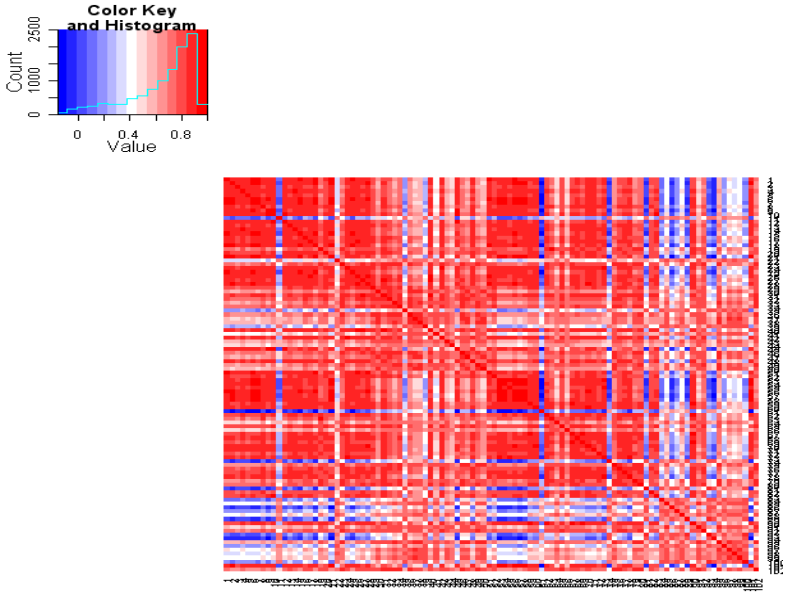


Figura 5.6.3(a): Correlación para muestras de *ARN* en *Prostate*.

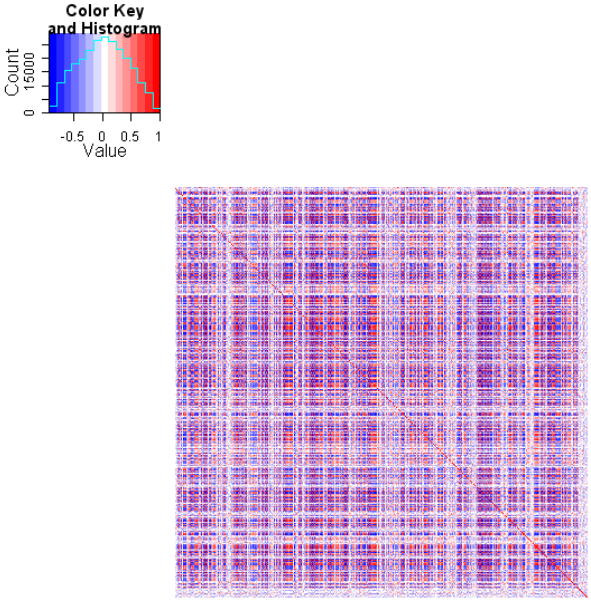


Figura 5.6.3(b): Correlación para muestra aleatoria de 500 genes. *Prostate*.

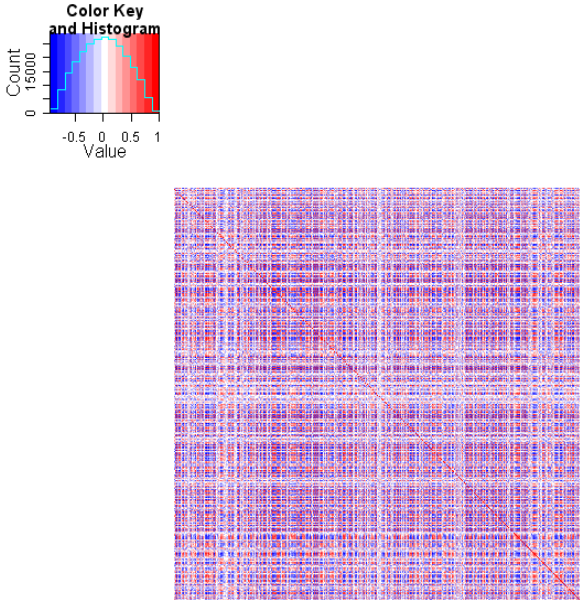


Figura 5.6.3(c): Correlación para muestra aleatoria normal multivariada de 500 genes. *Prostate*.

Por la Figura 5.6.3(a) se sabe que la muestras de *ARN* en *Prostate* tienen correlación altamente positiva, y por las Figuras 5.6.3(b) y 5.6.3(c), que los genes seleccionados tienen tanto correlación positiva como negativa; lo que lleva a pensar, que la cantidad de genes diferencialmente expresados es grande y que para identificarlos deben emplearse métodos como el *BY* y *maxT*; pues al examinar la Tabla 5.5.3 es claro que *BY* declara como genes diferenciados los mismos que los métodos Bonferroni, Šidák, Holm, Hochberg, *BH* y *t*, y que *maxT* declara prácticamente también los mismos (a excepción de un solo gen). En caso que se requiera emplear un método no paramétrico, resulta conveniente usar *SAM*. Es decir, que para datos con estas características, pueden usarse los métodos que controlan el *FDR*.

5.6.4 Correlación para *Colon*

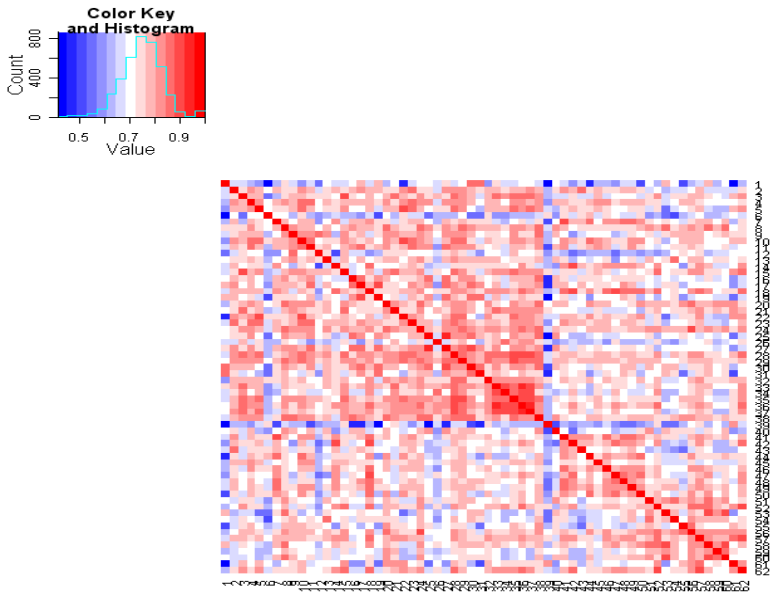


Figura 5.6.4(a): Correlación para muestras de *ARN* en *Colon*.

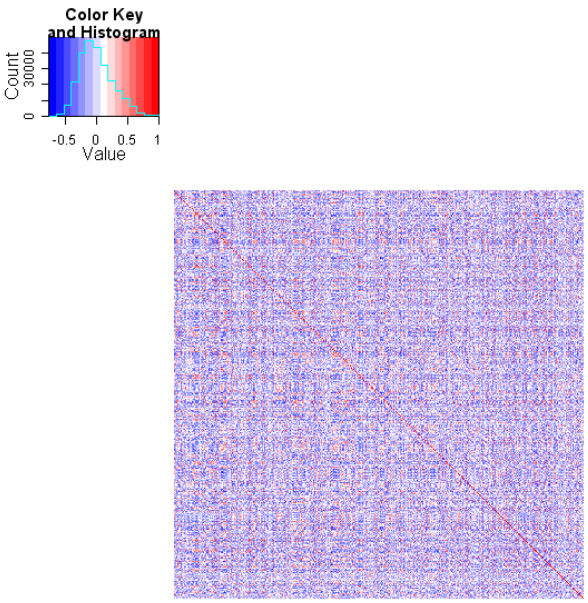


Figura 5.6.4(b): Correlación para muestra aleatoria de 500 genes. *Colon*.

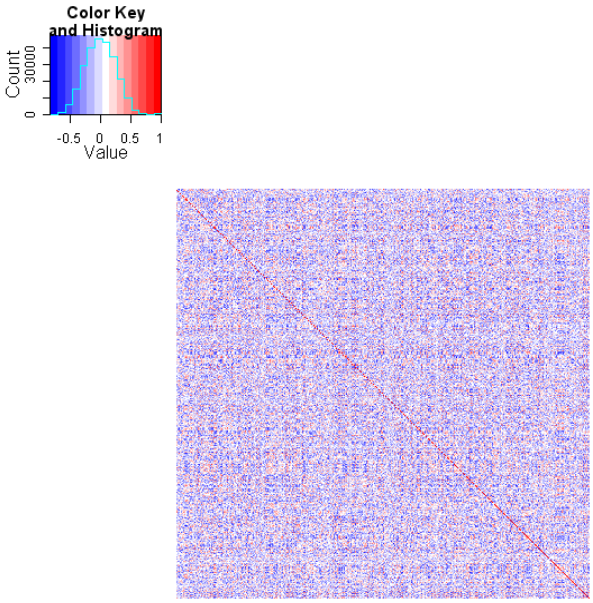


Figura 5.6.4(c): Correlación para muestra aleatoria normal multivariada de 500 genes. *Colon*.

La Figura 5.6.4(a) establece que las muestras de *ARN* en *Colon* son muy poco correlacionadas. Las Figuras 5.6.4(b) y 5.6.4(c), aunque poca, establecen mayor correlación negativa entre los genes seleccionados. Luego, podría pensarse en usar métodos como Bonferroni, Šidák, Holm, Hochberg, *BH* o *t*, para identificar los genes con expresión diferenciada. En la Tabla 5.5.4 se aprecia que los métodos *BY*, *maxT* y *SAM* no identifican totalmente los mismos genes diferencialmente expresados que los métodos anteriores.

5.6.5 Correlación para *Breastcc*

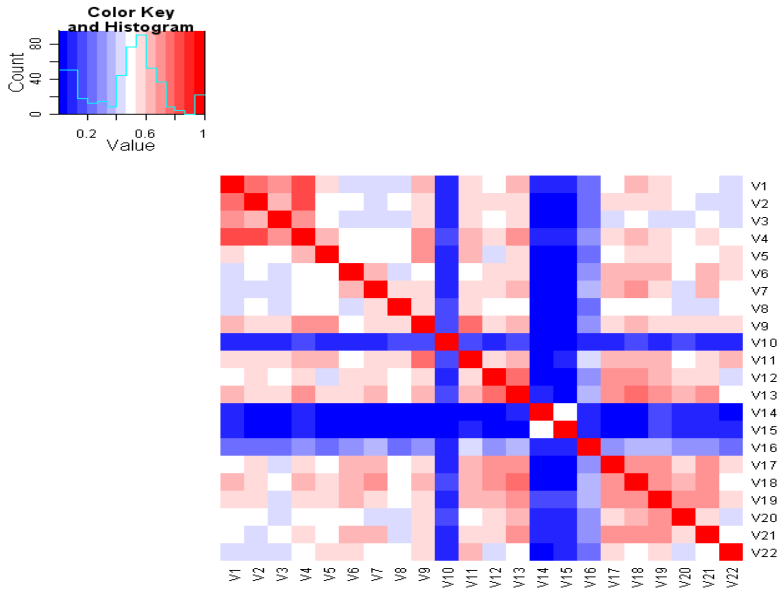
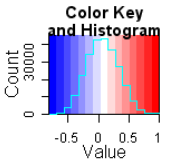


Figura 5.6.5(a): Correlación para muestras de *ARN* en *Breastcc*.



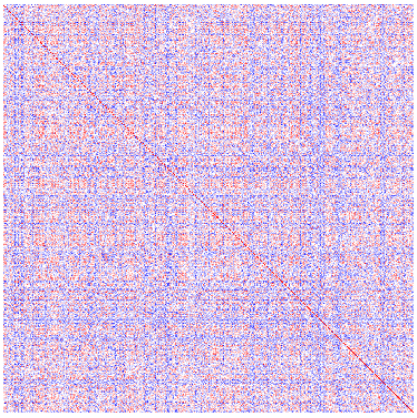
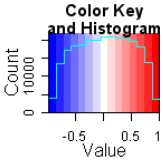


Figura 5.6.5(b): Corrón para muestra aleatoria de 500 genes. *Breastcc*.



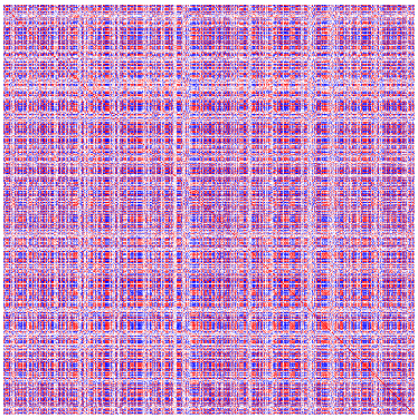


Figura 5.6.5(c): Correlación para muestra aleatoria normal multivariada de 500 genes. *Breastcc*.

Es claro, por la Figura 5.6.5(a), que las muestras de *ARN* en *Breastcc* tiene poca correlación negativa, y por la Figuras 5.6.5(b) y 5.6.5(c), que los genes seleccionados, poseen poca correlación negativa y correlación positiva casi nula, lo que implica declarar muy pocos genes como expresados, como lo establece la Tabla 5.3; usando los métodos que no consideran dependencia entre los datos (Bonferroni, Šidák, Holm, Hochberg, *BH* y *f*), o mejor aún, los métodos que controlan el *FWER*.

5.6.6 Correlación para *SRBCT*

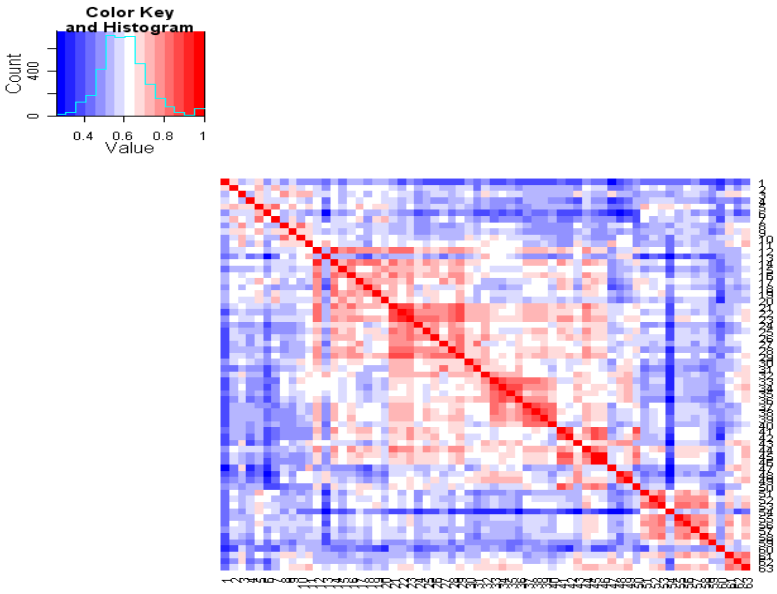


Figura 5.6.6(a): Correlación para muestras de *ARN* en *SRBT*.

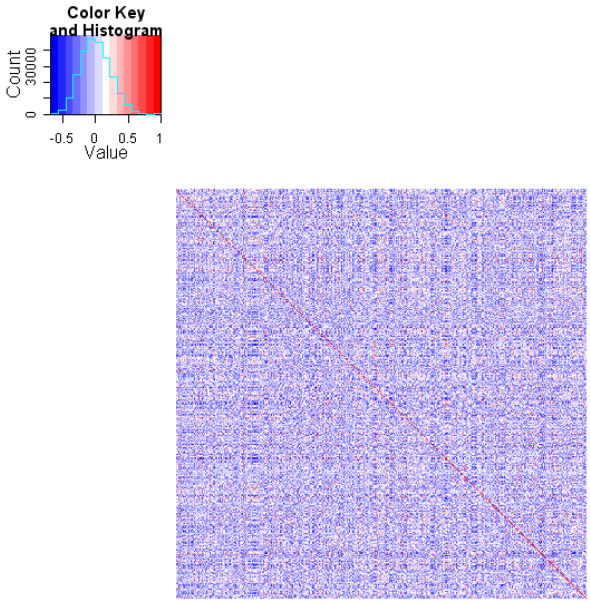


Figura 5.6.5(b): Correlación para muestra aleatoria de 500 genes. *SRBCT*.

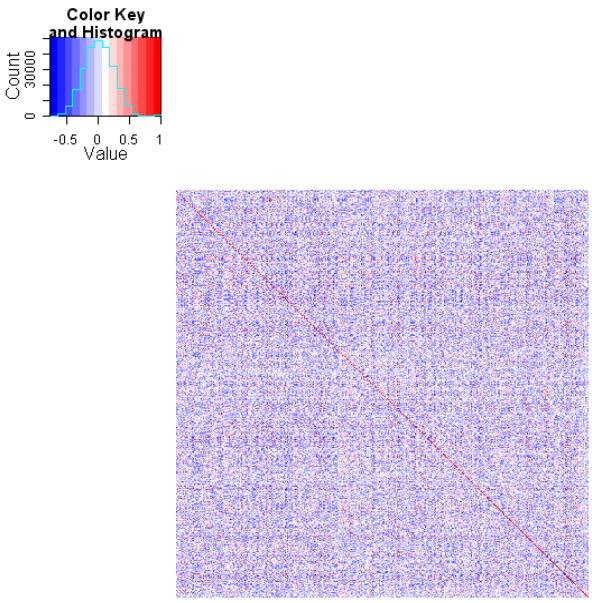


Figura5.6.5(c): Correlación para muestra aleatoria normal multivariada de 500 genes. *SRBCT*.

Las Figuras 5.6.5(a), 5.6.5(b) y 5.6.5(c) indican que aunque poca, existe correlación negativa tanto entre las muestras de *ARN* como entre los genes seleccionados para *SRBCT*. Por lo tanto los genes declarados como diferenciados deben ser pocos y deben identificarse usando métodos que controlen el *FWER*.

Capítulo 6
CONCLUSIONES Y TRABAJOS FUTUROS

6.1 CONCLUSIONES

En este trabajo se estudiaron diez métodos de prueba múltiple: los métodos de Bonferroni, Šidák de un solo paso, Šidák de pasos hacia abajo, Holm, Hochberg, Benjamini y Hochberg, Benjamini y Yekutieli, Westfall y Young (*maxT*), *SAM* y *t* ó *f* (estadístico de prueba *t* para los conjuntos de datos con dos clases y estadístico de prueba *f* para aquellos con más de dos clases). El interés consistía en establecer el método más confiable para la identificación de genes diferencialmente expresados en datos de expresión genética obtenidos de experimentos de microarreglos. Primero, se revisó la definición de cada método y se discutió el tipo de control que ofrecen para la tasa de Error Tipo I. Después, fueron aplicados a cada conjunto de datos para identificar los genes con expresión diferenciada. Los resultados obtenidos fueron los siguientes:

i. Entre los métodos paramétricos que consideran estadísticos de prueba independientes, los más conservativos, es decir, los que menos hipótesis nulas rechazan y por tanto los que declararon menos genes como diferenciados fueron los métodos de Bonferroni; Šidák; Holm; Hochberg; y Benjamini y Hochberg. Los menos conservativos fueron los basados en los estadísticos *t* y *f*.

ii. Entre los métodos que consideran dependencia entre los estadísticos de prueba, el más conservativo fue el *maxT*, los menos conservativos fueron el de Benjamini y Yekutieki, y el de *SAM*.

iii. A pesar que el método de Benjamini y Yekutieli considera la dependencia entre los estadísticos de prueba, los genes declarados como expresados por dicho método no difieren de los declarados por los métodos Bonferroni; Šidák; Holm; Hochberg;

Benjamini y Hochberg; *t* y *f*, lo cual sí sucede con los genes declarados por *maxT* y *SAM* para algunos de los conjuntos, en particular los que no son correlacionados. Esto puede deberse a que los *p-values* para los métodos *maxT* y *SAM* son calculados por permutación.

iv. Al comparar el método *SAM* con los otros métodos a excepción del *maxT*, podemos concluir que a pesar de ser un método no paramétrico, la alta aleatoriedad alcanzada con su técnica de remuestreo proporciona buenos resultados; además, suele ser menos conservativo que los métodos que controlan el *FWER*, por lo que constituye una buena herramienta en la identificación de genes diferenciados cuando lo que se desea es evitar asumir suposiciones a priori sobre los datos.

v. Aunque en la comparación teórica de los métodos considerados pueden sugerir diferencias al momento de identificar aquellos genes con expresión diferenciada, por los resultados obtenidos en el estudio experimental se puede afirmar, sin pérdida de generalidad, que se obtuvieron resultados razonablemente satisfactorios para todos los conjuntos de datos, los esperados según la definición de cada método de prueba múltiple aquí considerado, en especial para el conjunto de datos *Prostate* y aquellos conjuntos de datos con más de dos clases, como lo son *Breastcc* y *SRBCT*. Razón por la cual se puede afirmar que es conveniente extender la teoría de métodos de prueba de hipótesis múltiple a la identificación de genes diferencialmente expresados en conjuntos de datos con más de dos clases.

vi. En general, no se puede afirmar que un método es mejor que otro; sin embargo, sí se puede concluir que el método de prueba múltiple más eficiente en la identificación de genes diferencialmente expresados depende de la validez, de las suposiciones bajo las cuales el método es definido, del tipo de control que proporcione (débil, fuerte o exacto), de su efectividad computacional y del tipo de pregunta que se desee contestar.

Cuando se tiene información previa que establezca la cantidad de genes que se espera sean diferencialmente expresados, conviene usar los métodos que controlan el *FWER*, cuando son pocos, y los que controlan el *FDR*, cuando son muchos.

6.2 TRABAJOS FUTUROS

Considerar y evaluar métodos que controlen la tasa de Error *generalized family-wise* (*gFWER*), una variante del *FWER*, la cual consiste en permitir que algún número $k = 0,..., m_0$ de hipótesis sean rechazadas falsamente y se define como la probabilidad de que ocurran al menos *k* Errores Tipo I, esto es

$$gFWER(k) = \Pr(V > k) . \tag{6.2.1}$$

Si $k = 0$ en (6.2.1), entonces *gFWER* es la usual tasa de error *family-wise*, *FWER*.

Otra clase de error Tipo I para la cual deseamos también estudiar métodos que la controlen es la proporción de falsos positivos entre las hipótesis rechazadas, *FDP*; es decir, el valor observado del *FDR*, la cual se define para $\gamma \in (0,1)$, como

$$FDP(\gamma) = \Pr(V / R > \gamma) . \tag{6.2.2}$$

La definición en (6.2.2) resulta interesante, ya que ofrece un método tal que $\Pr(V / R > \gamma) \leq \alpha$ para un nivel de significancia α. En otras palabras, en comparación con el *FDR* constituye un mejor criterio puesto que el *FDR* considera todo el valor esperado y no únicamente lo que sucede con los datos observados. En cambio el *FDP* se relaciona directamente con la información observada.

Es importante notar que el *gFWER* es un parámetro que depende solamente de la distribución marginal del número de Errores Tipo I, *V*, y el *FDP* es un parámetro que depende de la distribución conjunta de (*V, R*).

APÉNDICE A
CÓDIGOS DE FUNCIONES EN R USADOS EN ESTA TESIS

Instalación de software Bioconductor

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

Lectura de paquetes

```
library(DEDS)
library(dprep)
library(sma)
library(splines)
library(survival)
library(multtest)
library(gdata)
library(gtools)
library(gplots)
library(limma)
```

Prueba *t* (por Edgar Acuña) para conjunto de datos con dos clases

```
stat.t2.acu=function (X, cl, x.ratio = FALSE, var.equal = TRUE, ...)
{#adaptado del paquete sma
#X: matriz de expresiones
#cl: vector indicando las columnas de los dos grupos
        if (!x.ratio) {
           n <- ncol(X)/2
           res <- t(apply(X$M, 1, function(z) t2stat.acu(z, cl, var.equal, ...)))
```

```
        A.bar <- apply(X$A, 1, mean.na)
      }
      if (x.ratio) {
        n <- ncol(X)
        res <- t(apply(X, 1, function(z) t2stat.acu(z, cl, var.equal, ...)))
        A.bar <- NULL
      }
      list(t = res[, "t"], Num = res[, "Num"], Den = res[, "Den"],
        n1 = res[, "n1"], n2 = res[, "n2"], pvalues=res[,"pval"],A.bar = A.bar)
    }
#*****************************************************************************
t2stat.acu=function (x, cl, var.equal = TRUE, ...)
    {
      x.ok = x[!(is.na(x) | is.infinite(x))]
      cl.ok = cl[!(is.na(x) | is.infinite(x))]
      x1 = x.ok[cl.ok == 0]
      x2 = x.ok[cl.ok == 1]
      n1 = length(x1)
      n2 = length(x2)
      if ((n1 > 2) & (n2 > 2)) {
        tmp = t.test(x1, x2, var.equal = var.equal, ...)
        tstat = -(tmp$stat)
        num = tmp$est[2] - tmp$est[1]
        den =num/tstat
        pval=tmp$p.value
          res = c(tstat, num, den, n1, n2,pval)
      }
      if ((n1 <= 2) | (n2 <= 2))
            res <- c(NA, NA, NA, n1, n2,0)
        names(res) <- c("t", "Num", "Den", "n1", "n2","pval")
      res
```

}

****

Para esta función se tienen los siguientes valores:

lt=l$t ; el estadístico de prueba *t*,

l$pvalues; los respectivos *p-values* para los estadísticos de prueba *t*,

indl=which(l$pvalues<0.05)

length(indl); el número de genes diferencialmente expresados a un nivel de significancia del 5% según el estadístico *t*,

pt=order(l$pvalues)

pt[1:10]; lista de los top 10 genes diferenciados a un nivel de significancia del 5% según el estadístico *t*.

****

Estadístico de prueba *f*, para conjunto de datos con más de dos clases

f=mt.teststat(data,clase,test="f")

k=dim(table(clase))

n=length(clase)

pvalf=1-pf(k-1,n-k,f); *p-values* correspondientes a los estadísticos de prueba *f*,

indf=which(pvalf<.05)

length(indf); número de genes diferencialmente expresados a un nivel de significancia del 5% según el estadístico *f*,

pf=order(pvalf)

pf[1:10]; lista de los top 10 genes diferenciados a un nivel de significancia del 5% según el estadístico *f*.

Tasa de Error Tipo I para los métodos Bonferroni, Holm, Hochberg, Šidák, Benjamini y Hochberg, Benjamini y Yekutieli.

met1=mt.rawp2adjp(pval, proc)

****

pval corresponde a los *p-values* de los estadísticos *t* ó *f*, dependiendo del número de clases) y proc es el método de interés.

```
**************************************************************************
indbo=which(met1$adjp[,2]<.05)
length(indbo); número de genes diferencialmente expresados por el método de Bonferroni a un nivel de significancia del 5%,
met1$index[1:10]; lista de los top 10 genes diferenciados por el método de Bonferroni.
**************************************************************************
```

Tasa de Error Tipo I para el método de Westfall y Young (*maxT*)

```
resT=mt.maxT(data, clase, B=100000)
indmax=which(resT$adjp<.05)
length(indmax); número de genes diferencialmente expresados a un nivel de significancia del 5%,
resT$index[1:10]; lista de los top 10 genes diferenciados por el método maxT.
```

Tasa de Error Tipo I para *SAM*

```
sam=comp.adjp(data,cl,B=1000,test="sam",tail="abs",extra = NULL)
order1=sam[,1]
adjps=sam[,4]
SAM2=adjps[order1]
indSAM2=which(SAM2<.05)
cat("\nel numero de genes diferenciados por SAM2 es=",length(indSAM2),"\n")
cat("\nlos top 10 genes diferenciados por SAM2 son",order1[1:10],"\n")
```

APÉNDICE B
MÉTODO DE BENJAMINI Y HOCHBERG (*BH*)

Antes de probar que el método de Benjamini y Hochberg proporcionan un control fuerte del *FWER*, se establecen algunas convenciones de nomenclatura y conceptos; también se demuestran ciertos resultados indispensables para la prueba.

- $A_{v,s}$, denota el evento en el que el método *BH* rechaza exactamente *v* hipótesis nulas verdaderas y *s* hipótesis nulas falsas.
- $\boldsymbol{P}^{(j)}$, denota los $m-1$ *p-values* restantes después de seleccionar P_j, para $j = 1,...,m_0$.
- $C_{v,s}^{(j)}$, denota el evento en el cual si P_i es rechazado entonces $v-1$ hipótesis nulas verdaderas y s hipótesis nulas falsas, son rechazadas junto con él. Es decir, $C_{v,s}^{(j)}$ es la proyección de $\{P_j \le \alpha_{v+s}\} \cap A_{v,s}$ en el rango de $\boldsymbol{P}^{(j)}$. Lo que implica, que

$$\{P_j \le \alpha_{v+s}\} \cap A_{v,s} = \{P_j \le \alpha_{v+s}\} \cap C_{v,s}^{(j)}. \tag{B.1.1}$$

- $C_k^{(j)} = \bigcup \{C_{v,s}^{(j)} : v+s=k\}$, $C_k^{(j)}$ es disjunto para cada *j*.

Lemma B.1.1: Pr($A_{v,s}$) puede ser expresado como un promedio, de la siguiente forma,

$$\Pr(A_{v,s}) = \frac{1}{v}\sum_{j=1}^{m_0} \Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}). \tag{B.1.2}$$

Prueba: Para *v* y *s* fijos. Sea κ un subconjunto de $\{1,\ldots, m_0\}$ de tamaño *v*, y sea $A_{v,s}^{\kappa}$ el evento en $A_{v,s}$ que denota, que las v hipótesis nulas verdaderas rechazadas están en κ.

Definamos la función indicadora

$$I_{\{\cdot\}} = \begin{cases} 1, & si \quad j \in \kappa \\ 0, & en \quad otro \quad caso. \end{cases}$$

Como $\Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}^{\kappa}) = \Pr(A_{v,s}^{\kappa})$ si $j \in \kappa$ y cero en otro caso, entonces

$$\begin{aligned} \Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}^{\kappa}) &= \sum_{j=1}^{m_0} \sum_{\kappa} \Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}^{\kappa}) \\ &= \sum_{\kappa} \sum_{j=1}^{m_0} \Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}^{\kappa}) \\ &= \sum_{\kappa} \sum_{j=1}^{m_0} I_{\{j \in \kappa\}} \Pr(A_{v,s}^{\kappa}) \\ &= \sum_{\kappa} v \cdot \Pr(A_{v,s}^{\kappa}) \\ &= v \cdot \Pr(A_{v,s}) . \end{aligned} \qquad \text{(B.1.3)}$$

Asi, (B.1.2) es inmediata de (B.1.3). ■

Según la Tabla 3.3 y la definición del *FDR* y $A_{v,s}$, se puede establecer que

$$FDR = \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \cdot \Pr(A_{v,s}). \qquad \text{(B.1.4)}$$

Así, combinando (B.1.4) con el lemma anterior se tiene entonces que

$$\begin{aligned} FDR &= \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{v}{v+s} \left\{ \sum_{j=0}^{m_0} \frac{1}{v} \cdot \Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}) \right\} \\ &= \sum_{j=0}^{m_0} \left\{ \sum_{s=0}^{m_1} \sum_{v=1}^{m_0} \frac{1}{v+s} \cdot \Pr(\{P_j \le \alpha_{v+s}\} \cap A_{v,s}) \right\} \end{aligned} \qquad \text{(B.1.5)}$$

Luego, por (B.1.1) y por la definición de $C_k^{(j)}$, el *FDR* puede ser expresado como

$$FDR = \sum_{j=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \cdot \Pr(\{P_j \le \alpha_k\} \cap C_k^{(j)}), \qquad \text{(B.1.6)}$$

por lo tanto el *FDR* no depende de *v* ni de *s*.

Teorema B.1.1: Para estadísticos de prueba independientes, el método de Benjamini y Hochberg controla el *FDR* a un nivel menor o igual a $(m_0/m)\cdot\alpha$. Si los estadísticos de pruebas son continuos, el *FDR* es exactamente controlado al nivel $(m_0/m)\cdot\alpha$.

Demostración: Sea $\alpha_k = (k/m)\cdot\alpha$, entonces por (B.1.6) se tiene

$$\begin{aligned} FDR &= \sum_{j=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\Pr\left(\left\{P_j \le \frac{k}{m}\cdot\alpha\right\}\cap C_k^{(j)}\right) \\ &= \sum_{j=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\Pr\left(P_j \le \frac{k}{m}\cdot\alpha\right)\cdot\Pr(C_k^{(j)}) \\ &\le \sum_{j=1}^{m_0}\frac{\alpha}{m}\cdot\sum_{k=1}^{m}\Pr(C_k^{(j)}) \\ &= \frac{m_0}{m}\cdot\alpha \end{aligned} \qquad \text{(B.1.7)}$$

Por otra parte, el lemma 3.5 establece que

$$\Pr\left(P_j \le \frac{k}{m}\cdot\alpha\right) = \frac{k}{m}\cdot\alpha, \text{ para } j = 1,\ldots,m_0.$$

Luego, para estadísticos de prueba continuos distribuidos uniformemente se tiene

$$FDR = \frac{m_0}{m}\cdot\alpha\ .\ \blacksquare \qquad \text{(B.1.8)}$$

Puesto que $m_0 \le m$ implica que las expresiones (B.1.7) y (B.1.8) del teorema (B.1.1) son ambas menores o iguales que α, es claro entonces que el método de Banjamini y Hochberg también proporciona un control fuerte del *FDR* a un nivel α.

APÉNDICE C
MÉTODO DE BENJAMINI YEKUTIELI (*BY*)

Propiedad de Dependencia de Regresión Positiva (***PDRS***)

Sea $\boldsymbol{X} = X_1,...,X_m$ el vector de los estadísticos de prueba. Se define la propiedad *PDRS* en un subconjunto J_0, para cualquier conjunto creciente D y para cada $j \in J_0$, si Pr($\boldsymbol{X} \in D \,|\, X_j = x$) es no creciente en *x*.

Asumimos que un conjunto D es llamado creciente si $x \in D$ y $y \geq x$, implica que $y \in D$ también.

Método de Benjamini y Hochberg

Teorema C.1.1: Si la distribución conjunta de los estadísticos de prueba cumplen la propiedad *PRDS* en el subconjunto de estadísticos de prueba correspondientes a las hipótesis nulas verdaderas, el método de *BH* controla el *FDR* a un nivel menor o igual a $(m_0/m)\cdot\alpha$.

Demostración: Sea $D_k^{(j)} = \bigcup\{C_i^{(j)} : i \leq k\}$ para $k = 1,...,m$. $D_k^{(j)}$ también puede expresarse usando el conjunto de *p-values* ordenados $\{p_{(1)}^{(j)},...,p_{(m-1)}^{(j)}\}$ en el rango de $\boldsymbol{P}^{(j)}$, de la siguiente forma

$$D_k = \{\boldsymbol{p} : \alpha_{k+1} < p_{(k)}^{(j)}, \alpha_{k+2} < p_{(k+1)}^{(j)},..., \alpha_m < p_{(m-1)}^{(j)}\}, \qquad \text{(C.1.1)}$$

para $k = 1,...,m-1$, y $D_m^{(j)}$ es simplemente el espacio entero. Así, $D_k^{(j)}$ es un conjunto creciente para cada *k*.

Por la propiedad *PRDS*, se sabe que para $p \leq p'$

$$\Pr(D \mid P_j = p) \le \Pr(D \mid P_j = p') . \qquad \text{(C.1.2)}$$

Luego para $i \le l$ como $\alpha_i \le \alpha_l$,

$$\Pr(D \mid P_j = \alpha_i) \le \Pr(D \mid P_j = \alpha_l) , \qquad \text{(C.1.3)}$$

para cualquier conjunto creciente *D*, o equivalentemente,

$$\frac{\Pr\left(\{P_j \le \alpha_k\} \bigcap D_k^{(j)}\right)}{\Pr(P_j \le \alpha_k)} \le \frac{\Pr\left(\{P_j \le \alpha_{k+1}\} \bigcap D_k^{(j)}\right)}{\Pr(P_j \le \alpha_{k+1})} . \qquad \text{(C.1.4)}$$

Considerando (C.1.4) junto con el hecho de que el factor $D_{i+1}^{(j)} = D_i^{(j)} \bigcup C_{i+1}^{(j)}$ para todo $k \le m-1$, se tiene

$$\begin{aligned} &\frac{\Pr\left(\{P_j \le \alpha_k\} \bigcap D_k^{(j)}\right)}{\Pr(P_j \le \alpha_k)} + \frac{\Pr\left(\{P_j \le \alpha_{k+1}\} \bigcap C_{k+1}^{(j)}\right)}{\Pr(P_j \le \alpha_{k+1})} \\ &\le \frac{\Pr\left(\{P_j \le \alpha_{k+1}\} \bigcap D_k^{(j)}\right)}{\Pr(P_j \le \alpha_{k+1})} + \frac{\Pr\left(\{P_j \le \alpha_{k+1}\} \bigcap C_{k+1}^{(j)}\right)}{\Pr(P_j \le \alpha_{k+1})} \\ &= \frac{\Pr\left(\{P_j \le \alpha_{k+1}\} \bigcap D_{k+1}^{(j)}\right)}{\Pr(P_j \le \alpha_{k+1})} . \end{aligned} \qquad \text{(C.1.5)}$$

Sea C_1=D_1. Al aplicar repetidamente la desigualdad (C.1.5) para $j = 1,...,m-1$, a partir de la segunda repetición, la suma de la izquierda se transforma en un asola expresión obteniéndose que

$$\sum_{k=1}^{m} \frac{\Pr\left(\{P_j \le \alpha_k\} \bigcap C_k^{(j)}\right)}{\Pr(P_j \le \alpha_k)} \le \frac{\Pr\left(\{P_j \le \alpha_m\} \bigcap D_m^{(j)}\right)}{\Pr(P_j \le \alpha_m)} = 1, \qquad \text{(C.1.6)}$$

donde la última igualdad se cumple porque $D_m^{(j)}$ es el espacio entero.

Combinando las expresiones (B.1.6) y (C.1.6), y considerando el hecho de que bajo la hipótesis nula $\Pr(P_j \le \alpha_k) \le \alpha_k = (k/m) \cdot \alpha$, se tiene

$$FDR = \sum_{j=1}^{m_0}\sum_{k=1}^{m}\frac{1}{k}\cdot \Pr(\{P_j \le \alpha_k\} \cap C_k^{(j)}),$$

$$\le \sum_{j=1}^{m_0}\sum_{k=1}^{m}\frac{\alpha}{m}\cdot\frac{\Pr\left(\{P_j \le \alpha_k\}\bigcap C_k^{(j)}\right)}{\Pr(P_j \le \alpha_k)} \qquad \text{(C.1.7)}$$

Luego, por (C.1.6) y (C.1.7) se obtiene

$$FDR \le \frac{\alpha}{m}\sum_{j=1}^{m_0}\left(\sum_{k=1}^{m}\frac{\Pr\left(\{P_j \le \alpha_k\}\bigcap C_k^{(j)}\right)}{\Pr(P_j \le \alpha_k)}\right) \le \frac{m_0}{m}\cdot\alpha\,. \blacksquare$$

La igualdad se cumple para estadísticos de prueba continuos distribuidos uniformemente.

Método de Benjamini y Yekutieli

Teorema C.2.1: Cuando se emplea el método de Benjamini y Hocberg considerando $\alpha\Big/\left(\sum_{i=1}^{m}(1/i)\right)$ en lugar de α, tal método proporciona siempre un control fuerte del *FDR* a un nivel menor o igual a $(m_0/m)\cdot\alpha$, bajo la propiedad *PDRS*. Es decir, que el método de Benjamini y Yekutieli controla el *FDR* a un nivel $(m_0/m)\cdot\alpha$.

Demostración:

Por simplicidad, usaremos α en lugar de $\alpha\Big/\left(\sum_{i=1}^{m}(1/i)\right)$ y demostraremos que el *FDR* es menor o igual que $\sum_{i=1}^{m}(1/i)$.

Sea

$$p_{ijk} = \Pr\left(\left\{P_j \in \left[\frac{(j-1)}{m}\cdot\alpha, \frac{j}{m}\cdot\alpha\right]\right\}\bigcap C_k^{(j)}\right).$$

Como

$$\sum_{k=1}^{m} p_{ijk} = \Pr\left(\left\{P_j \in \left[\frac{(j-1)}{m}\cdot\alpha, \frac{j}{m}\cdot\alpha\right]\right\}\bigcap\left(\bigcup_{k=1}^{m} C_k^{(j)}\right)\right) = \frac{\alpha}{m}. \qquad \text{(C.2.1)}$$

Por la expresión (B.1.6), el *FDR* puede ser expresado entonces por

$$\begin{aligned} FDR &= \sum_{i=1}^{m_0} \sum_{k=1}^{m} \frac{1}{k} \sum_{j=1}^{k} p_{ijk} \\ &= \sum_{i=1}^{m_0} \sum_{j=1}^{m} \sum_{k=j}^{m} \frac{1}{k} \cdot p_{ijk} \\ &\leq \sum_{i=1}^{m_0} \sum_{j=1}^{m} \sum_{k=j}^{m} \frac{1}{j} \cdot p_{ijk} \\ &\leq \sum_{i=1}^{m_0} \sum_{j=1}^{m} \frac{1}{j} \sum_{k=1}^{m} p_{ijk} \\ &= m_0 \cdot \left(\frac{\alpha}{m} \right) \cdot \sum_{j=1}^{m} \left(\frac{1}{j} \right). \blacksquare \end{aligned}$$

Si el número de hipótesis *m* es muy grande, se sabe que

$$\sum_{j=1}^{m} \frac{1}{j} \approx \log(m) + \frac{1}{2},$$

por lo que en este caso, la penalidades para los métodos de *BH* y *BY* difieren por un factor de aproximadamente log(*m*).

REFERENCIAS BIBLIOGRÁFICAS

[1] Alizadeh, A.A., Eisen, M.B. and Davis, R.E. et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". Nature 403:503-511.

[2] Alon, U., Barkai, N., Notterdam, D., Gish, K., Ybarra, S., Mack , A. , and Levine, D. (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed oligonucleotide arrays". *Proc Natl Acad Sci* 96:6745-6750.

[3] Benjamini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing". J. Roy. Statist. Soc. Ser. B 57**:** 289–300.

[4] Benjamini, Y., and Yekutieli, D. (2001). "The control of the false discovery rate in multiple testing under dependency". Ann. Statist 29**:** 1165–1188.

[5] Beran, R. (1988). "Balanced simultaneous confidence sets". J. Amer. Statist. Assoc. 83**:** 679–686.

[6] Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. "Microarray Expression profiling identifies genes with altered expression in hdl deficient mice". Genome Research, Submitted.

[7] Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). "Ratio - based decisions and the quantitative analysis of cDNA microarray images". Journal of Biomedical Optics 2:364-374.

[8] Crick, F. (1970). "Central Dogma of Molecular Biology". Nature 227, 561-563.

[9] DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., et al. (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer". Nature Genetics 14:457-460.

[10] Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). "Multiple hypothesis testing in Microarray experiments". Statistical Science Vol. 18 No. 1, 71-103.

[11] Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments". Statistica Sinica 12(1): 111-139.

[12] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns". Proc. Natl. Acad. Sci. 95:14863-14868.

[13] Friddle, C. J., Koga, T., Rubin, E. M., and Bristow, J. (2000). "Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. Proc. Natl. Acad. Sci. 97:6745-6750.

[14] Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S., and Fink, G. R. (1999). "Ploidy regulation of gene expression". Science 285:251-254.

[15] Golub, T.R., Slonim, D.K., and Tamayo, P., et al. (1999). "Molecular classification of cancer: Class discovery and class prediction y gene expression monitoring". Science 286: 531–537.

[16] Gonick, L., and Wheelis, M. (1991). "The cartoon guide to genetics". Harper Perennial, updated edition.

[17] Griffths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C., and Gelbart, W. M. (1996). "An Introduction to Genetic Analysis. W. H. Freeman and Company, New York, 6th edition,

[18] Hedenfalk, D., Duggan, I., Radmacher, M., Bittner, M., Simon, P., Meltzer, R.,

Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Loman, S., Pittaluga, N., Gruvberger, S., Johannsson, O., Olsson, H, Wilfond, B., Sauter, G., Kallioniemi, O., Borg, J., and Trent, P. (2001).“Gene-expression proffles in hereditary breast cancer”. *Journal of Medicine* 344:539-548.

[19] Hochberg, Y., and Tamhane, A. C. (1987). “Multiple Comparison Procedures”. Wiley, New York.

[20] Hommel, G. (1988). “A stagewise rejective multiple test procedure based on a modified Bonferroni test”. Biometrika 75**:** 383–386.

[21] Hommel, G., and Bernard, G. (1999). “Bonferroni procedures for logically related hypotheses”. J. Statist. Plann. Inference 82**:** 119–128.

[22] Jogdeo, K. (1977). “Association and probability inequalities”. *Annals of Statistics* 5(3):495–504.

[23] Kaminsky, N., Friedman N. (2002). “Practical Approaches to Analizing Results of Microarrays Experiments”. American Journal of Respiratory Cell and Molecular Biology 27: 125-132.

[24] Kelmansky, D. M. (2006). “Análisis Exploratorio y Confirmatorio de Datos de Experimetos de Microarrays”.

[25] Kerr, M. K., Martin, M., and Churchill, G. A.(2000). “Analysis of variance for gene expression microarray. Technical report, The Jackson Laboratory. http://www.jax.org/research/churchill/pubs/index.html.

[26] Khan, J., Wei, J., Markus, R., Saal, L., Ladany, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, P., and Meltzer, C. (2001).“Classiffation and diagnostic prediction of cancers using gene expression profling and artifcial neural networks. *nature medicine*, 7:673-679.

[27] Lehmann, E. L. (1986). "Testing Statistical Hypotheses". 2nd ed. Wiley, New York.

[28] Lipshutz, R. J., Fodor, S. P., Gingeras D. J., and Lockhart T. R. (1999). "High density synthetic oligonucleotide arrays". Nat Genet 21(1 Suppl):20-24.

[29] Lockhard, D. , Brown, E. , Wong, G., Chee, T., and Gingeras, M. (1996). Expression monitoring by hibridation high density oligonucleotide arrays. United States Patent.

[30] Mills J., Roth, K. and Cols. (2001). "DNA microarrays and beyond: completing the journey from tissue to cell." Nature Cell Biology 3: E175-178.

[31] Newton, M.A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (1999). "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data". Technical Report 139, Department of Biostatistics and edical Informatics, UW Madison. http://www.stat.wisc.edu/ newton/papers/abstracts/btr139a.html.

[32] Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennet, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., M. Tyers, Boone C., and Friend, S. H. (2000). "Singnaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression". Science 287:873-880.

[33] Rom, D. M. (1990)."A sequentially rejective test procedure based on a modified Bonferroni inequality". Biometrika 77**:** 663–665.

[34] Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown P. O. (2000). "Systematic variation in gene expression patterns in human cancer cell lines". Nature Genetics 24:227-234.

[35] Sapir, M. and Churchill, G. A. (2000). "Estimating the posterior probability of differential gene expression from microarray data". Poster. The Jackson Laboratory. http://www.jax.org/research/churchill/.

[36] Schena, M., Shalon, D., Davis, R. W., and Brown,. P. O. (1995). "Quantitative Monitoring of gene expression patterns with a complementary dna microarray". Science 270:467-470.

[37] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis., R. W. (1996). "Parallel human genome analysis: microarray-based expression monotoring of 1000 Genes". Proc. Natl. Acad. Sci., 93:10614-10619,.

[38] Schulze, A. (2001). "Navigating gene expression using microarrays a technology review". Nature Cell Biology 3: E190-195.

[39] Shaffer, J. P. (1995). "Multiple hypothesis testing: A review". Annual Review of Psychology 46: 561–584.

[40] Šidák, Z. (1967). "Rectangular confidence regions for the means of multivariate normal distributions". *Journal of the American Statistical Association* 62:626–633.

[41] Simes, R. J. (1986). "An improved Bonferroni procedure for multiple tests of Significance". Biometrika 73**:** 751–754.

[42] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R.,et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9:3273-3297.

[43] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S, and Golub,. T. R. (1999). "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation". Proc. Natl. Acad. Sci. 96:2907-2912,

[44] Troyanskaya, O.,Cantor, M.l, Sherlock, G., Pat, B., Hastie, T., Tibshirani, R., Botstein, R. B. , and Altman, D. (2001) "Missing value estimation methods for dna microarrays". Bioinformatics, 17:520-525.

[45] Tusher, V. G., Tibshirani, R. and Chu, G. (2001). "Significance analysis of microarrays applied to the ionizing radiation response". *Proc. Natl. Acad. Sci. U.S.A.* 98 5116–5121.

[46] Van´t Veer, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer". Nature 415: 530 - 536

[47] Westfall, P. H. and Young, S. S. (1993). "Resampling - Based Multiple Testing: Examples and Methods for p-Value Adjustment". Wiley, New York.

[48] Wright, S. P. (1992). "Adjusted p-values for simultaneous inference". Biometrics 48**:** 1005–1013.

[49] Yekutieli, D. and Benjamini, Y. (1999). "Resampling - based false discovery rate controlling multiple test procedures for correlated test statistics". Statist. Plann. Inference 82**:** 171– 196.