



**INTEGRACIÓN DE DIVERSAS
TECNOLOGÍAS DE MICROARRAYS PARA
LA EXPRESIÓN DIFERENCIAL DE GENES
EN PATOLOGÍAS DE CÁNCER**

Memoria presentada por

DANIEL CASTILLO SECILLA

Para optar al máster en
**CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES POR
LA UNIVERSIDAD DE GRANADA**

Fdo. Daniel Castillo Secilla

Septiembre 2016

VISTO BUENO

El Prof. Dr. D. Ignacio Rojas Ruiz y el Prof. Dr. D. Fernando Rojas Ruiz, ambos del Departamento de Departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada,

CERTIFICAN:

Que la memoria titulada:

"Integración de diversas tecnologías de microarrays para la expresión diferencial de genes en patologías de cáncer"

ha sido realizada por D. Daniel Castillo Secilla bajo nuestra dirección en el Departamento de Departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada para optar al máster de Ciencia de Datos e Ingeniería de Computadores por la Universidad de Granada.

En Granada, a 12 de Septiembre 2016.

Los Directores del trabajo fin de grado:

Fdo. Ignacio Rojas Ruiz

Fdo. Fernando Rojas Ruiz

INTEGRACIÓN DE DIFERENTES TECNOLOGÍAS DE MICROARRAYS PARA EL ANÁLISIS DIFERENCIAL DE GENES APLICADO AL DIAGNÓSTICO DE CÁNCER DE MAMA



Palabras clave:

Cáncer de mama - microarray - expresión diferencial de genes - Affymetrix - Illumina - biomarcadores - bioconductor - GEO - RStudio - Matlab - virtualArray

Resumen:

Hoy en día, existe muchos repositorios públicos que contienen una gran base de datos con datasets provenientes de microarrays de análisis diferencial. Esto conlleva también un problema, la mayoría de estos microarrays contienen un número muy reducido de muestras y por ello, no son estadísticamente significativas. Por tanto, una buena opción podría ser integrar diferentes datasets provenientes de diferentes tecnologías y de una misma enfermedad con el propósito de obtener un mejor número de muestras para hacer el estudio mucho mas robusto.

En este proyecto se han unido datasets provenientes de diferentes tecnologías de microarrays(Illumina y Affymetrix). Gracias a esto se ha conseguido una mayor base de datos relacionada con el cáncer de mama para así tener más significación en la expresión de los genes para los diferentes estudios que se han realizado.

INTEGRATION OF MICROARRAYS TECHNOLOGIES FOR THE ANALYSIS DIFFERENTIAL GENE EXPRESSION APPLY TO BREAST CANCER DIAGNOSIS



Keywords:

Breast cancer - microarray - differentially expressed genes - Affymetrix - Illumina - biomarker - bioconductor - GEO - RStudio - Matlab - virtualArray

Abstract:

Nowadays there are many public repositories which contains a big database of microarrays gene expression datasets. But there is a problem, most of this microarrays have a little number of samples and they aren't statistically significant. Therefore, may be a good option integrating different datasets from different technologies but with the same disease with the purpose of obtain the better number of probes for doing the study more robust.

In this work we merge datasets from Illumina and Affymetrix microarrays. Thanks to that we achieve a bigger database of breast cancer studies in order to get more significantly expressed genes in the different analysis that we do.

'La valentía es la única forma de vivir la vida intensamente. De avanzar. De crecer. De ir por delante. Es la principal causa de los grandes descubrimientos de nuestra humanidad y de los avances de todas las disciplinas'

— Juana Erice

Mirada al frente, paso corto y mala leche.

— Bonifacio Castillo Prados

AGRADECIMIENTOS

Aunque el camino ha sido largo, he conseguido llegar a la cima de esta montaña rodeado de personas increíbles que han me han hecho recuperar aliento cuando más exhausto estaba y seguir escalando. Por todo ello Gracias a todos.

Gracias Papa, Gracias Mama, se que nunca habría llegado hasta aquí sin vuestro apoyo y vuestro amor. Me faltarían vidas si hubiese más de una para devolveros todo lo que me habéis dado y aunque debido al agobio a veces suframos roces os quiero y os querré siempre.

Para Juan Carlos simplemente decir que has sido el último en llegar como quién dice pero el claro ejemplo de que se pueden hacer hermanos nuevos, aún sin conocerlos de toda la vida. Gracias Romeo.

A mis hermanos, Jesús y José María, que aunque este año ha sido difícil y agobiante para todos, ahí habéis estado siempre, escuchando, aconsejando e intentado que me evadiese un rato aunque fuese de todo. Espero que la vida os devuelva todo lo que os merecéis de una vez. Seguid persiguiendo vuestros sueños y no les dejéis escapar.

A mi abuelo Bonifacio, desde chico me has llevado de la mano a todas partes y me has hecho crecer como el hombre que soy hoy. Sin duda alguna eres el claro ejemplo de la sabiduría, el trabajo de toda una vida y el amor hacia tu familia. No hay palabras suficientes en el mundo para agradecerte todo. Te quiero.

Al resto de mis abuelos, que por desgracia ya partieron, sabed que os he tenido presentes en cada paso que he dado, a cada momento, cada día de mi vida y así pienso seguir haciéndolo hasta el día que me toque partir a mi.

Por último, agradecer a mis directores Ignacio y Fernando, la oportunidad de descubrir y trabajar en un campo tan interesante y beneficioso para la sociedad como lo es la bioinformática. Espero poder seguir más años trabajando con vosotros codo con codo.

LISTA DE CONTENIDOS

i	INTRODUCCIÓN	1
1	INTRODUCCIÓN	3
1.1	Objetivos del proyecto	4
1.2	Cáncer de mama: incidencia y anatomía	5
1.2.1	Definición de cáncer	5
1.2.2	¿Cómo aparece el cáncer?	5
1.2.3	Cáncer de mama	5
1.3	Introducción a los MicroArrays	9
1.4	Repositorio GEO	12
2	ESTADO DEL ARTE	15
2.1	Secuenciación clásica	16
2.1.1	Sanger Sequencing	16
2.1.2	Maxam-Gilbert Sequencing	17
2.1.3	Shotgun Sequencing	19
2.2	Next Generation Sequencing	20
2.2.1	Roche 454 Sequencing	20
2.2.2	Illumina (Solexa) Sequencing	21
2.2.3	SOLiD Sequencing	21
2.3	Métodos en desarrollo	22
2.3.1	Nanopore DNA Sequencing	22
2.3.2	Sequencing with mass spectrometry	23
2.3.3	Multiplex polony technology	23
ii	DESARROLLO DEL ESTUDIO Y RESULTADOS	25
3	METODOLOGÍA	27
3.1	Metodología de trabajo	28
3.2	Procesamiento de los datos de GEO mediante R	29
3.3	Análisis de calidad y eliminación de los Outliers	30
3.4	Preprocesamiento de las series	31
3.5	Análisis de expresión diferencial de los genes	32
3.6	Integración de las series	33
3.7	Efecto Batch	34
3.7.1	Métodos de eliminación del efecto batch basados en escalado	36
3.7.2	Métodos de eliminación del efecto batch basados en discretización	38
3.8	Evaluación de las técnicas de integración en microarrays	39
4	RESULTADOS Y ESTUDIO	41
4.1	Discusión y resultados de las series de Affymetrix en R	43
4.1.1	Análisis serie GSE52712	43
4.1.2	Análisis serie GSE40987	47
4.1.3	Análisis serie GSE52262	50

4.1.4	Análisis serie GSE12790	54
4.1.5	Análisis integrador de Affymetrix mediante VirtualArray	55
4.2	Discusión y resultados de las series de Affymetrix en Matlab	60
4.2.1	Análisis serie GSE52712	60
4.2.2	Análisis serie GSE40987	63
4.2.3	Análisis serie GSE52262	66
4.2.4	Análisis serie GSE12790	69
4.2.5	Comparación entre análisis en R y Matlab de Affymetrix	72
4.3	Discusión y resultados de las series de Illumina	77
4.3.1	Análisis serie GSE46834	77
4.3.2	Análisis serie GSE68651	80
4.3.3	Análisis integrador de Illumina mediante VirtualArray	83
4.4	Discusión y resultados de las series de Illumina en Matlab	86
4.4.1	Análisis serie GSE46834	86
4.4.2	Análisis serie GSE68651	89
4.4.3	Comparación entre análisis en R y Matlab de Illumina	92
4.5	Discusión del estudio integrador de las series de Affymetrix e Illumina	95
4.5.1	Análisis de las series integradas	96
4.5.2	Comparación de técnicas de unión de las muestras	99
4.5.3	Comparación de técnicas de eliminación del efecto batch	100
4.5.4	Interpretación de los resultados del estudio integrador	102
4.6	Clasificación de los datos	105
4.6.1	Clasificación con KNN y SVM	105
iii	CONCLUSIONES	109
5	CONCLUSIONES Y FUTUROS TRABAJOS	111
5.1	Contribuciones originales	112
5.2	Futuros trabajos	113
iv	BIBLIOGRAFÍA	115
	BIBLIOGRAFÍA	117
v	ANEXOS	127
6	ANEXO A	129

LISTA DE FIGURAS

- Figura 1.1 Mamografía de una mama afectada de cáncer 6
Figura 1.2 Anatomía de la mama. Fuente: <http://www.bonomedico.es/informacion/mamoplastia/aumento-pecho/anatomia-mama> 6
Figura 1.3 Doble hélice de ADN con los enlaces de hidrógeno entre los nucleótidos 10
Figura 1.4 Proceso de creación de un microarray. El ARN mensajero es extraído de las células y convertido en ADN complementario y etiquetado de manera fluorescente. La muestra de referencia es de color rojo y verde la muestra de test. Al mezclarlas se pasa una sonda y se eliminan posteriormente el material que no ha hibridado para escanear el chip mediante un láser confocal. 11
Figura 1.5 Representación de un 'Affymetrix GeneChip' y un 'Illumina BeadArray' 12
Figura 2.1 Fotografía de Frederick Sanger 16
Figura 2.2 Método de secuenciación de Sanger 17
Figura 2.3 Fotografía de Walter Gilbert 17
Figura 2.4 Método de Maxam and Gilbert 18
Figura 2.5 Celera Genomics 19
Figura 2.6 Shotgun Sequencing 19
Figura 2.7 Roche 454 Sequencing 20
Figura 2.8 Illumina Solexa Sequencing 21
Figura 2.9 SOLiD Sequencing 22
Figura 2.10 Nanopore Sequencing 23
Figura 3.1 Metodología Pipeline en expresión de genes 28
Figura 3.2 Posibles causas del efecto Batch 34
Figura 4.1 Test de Kolmogorov-Smirnov aplicado a la serie GSE52712 44
Figura 4.2 Test de Kolmogorov-Smirnov aplicado a la serie GSE52712 normalizada 44
Figura 4.3 Test de Kolmogorov-Smirnov aplicado a la serie GSE40987 47
Figura 4.4 Test de Kolmogorov-Smirnov aplicado a la serie GSE40987 normalizada 48
Figura 4.5 Test de Kolmogorov-Smirnov aplicado a la serie GSE52262 50
Figura 4.6 Test de Kolmogorov-Smirnov aplicado a la serie GSE52262 normalizada 51
Figura 4.7 Test de Kolmogorov-Smirnov aplicado a la serie GSE12790 54
Figura 4.8 Test de Kolmogorov-Smirnov aplicado a la serie GSE12790 normalizada 55

- Figura 4.9 Diagrama de Venn de la integración de affymetrix usando 20 genes 57
- Figura 4.10 Diagrama de Venn de la integración de affymetrix usando 100 genes 58
- Figura 4.11 Valores de expresión sin normalizar de las series de Affymetrix por separado 58
- Figura 4.12 Valores de expresión normalizados de las series de Affymetrix por separado 59
- Figura 4.13 Valores de expresión normalizados de las series de Affymetrix integradas 59
- Figura 4.14 Valores de expresión normalizados de la serie GSE52712 de Affymetrix 60
- Figura 4.15 Valores de expresión normalizados de la serie GSE40987 de Affymetrix 63
- Figura 4.16 Valores de expresión normalizados de la serie GSE52262 de Affymetrix 66
- Figura 4.17 Valores de expresión normalizados de la serie GSE12790 de Affymetrix 69
- Figura 4.18 Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE52712 72
- Figura 4.19 Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE52712 73
- Figura 4.20 Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE40987 73
- Figura 4.21 Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE40987 74
- Figura 4.22 Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE52262 74
- Figura 4.23 Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE52262 75
- Figura 4.24 Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE12790 75
- Figura 4.25 Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE12790 76
- Figura 4.26 Test de Kolmogorov-Smirnov aplicado a la serie GSE46834 78
- Figura 4.27 Test de Kolmogorov-Smirnov aplicado a la serie GSE68651 80
- Figura 4.28 Diagrama de Venn de la integración de Illumina usando 20 genes 83
- Figura 4.29 Diagrama de Venn de la integración de Illumina usando 100 genes 85
- Figura 4.30 Valores de expresión normalizados de las series de Illumina por separado 85
- Figura 4.31 Valores de expresión normalizados de las series de Illumina integradas 86
- Figura 4.32 Valores de expresión normalizados de la serie GSE46834 de Affymetrix 87
- Figura 4.33 Valores de expresión normalizados de la serie GSE68651 de Affymetrix 89
- Figura 4.34 Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE46834 92

Figura 4.35	Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE46834	93
Figura 4.36	Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE68651	93
Figura 4.37	Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE68651	94
Figura 4.38	Flujo de actividad seguido en el proyecto	96
Figura 4.39	Diagrama de Venn de la todas las series integradas dejando 2 fuera	97
Figura 4.40	Valores de expresión no normalizados de las series de integradas por separado	97
Figura 4.41	Valores de expresión no normalizados de las series de integradas por separado	98
Figura 4.42	Valores de expresión normalizados de las series integradas	98
Figura 4.43	Diagrama de venn comparando distintos métodos de unión	99
Figura 4.44	Diagrama de venn comparando distintos métodos de eliminación del efecto batch	100
Figura 4.45	Boxplots de la expresión del gen IL20RB usando diferentes técnicas de supresión del efecto batch	104
Figura 4.46	Ranking de variables y accuracy al clasificar mediante diferentes algoritmos las muestras iniciales de cáncer de mama y de control	106
Figura 4.47	Ranking de variables y accuracy al clasificar mediante diferentes algoritmos muestras de cáncer de mama y de control del conjunto de test	107

LISTA DE TABLAS

Tabla 1.1	Posibles estadios	8
Tabla 3.1	Notación en las técnicas de la eliminación del efecto batch	35
Tabla 4.1	Veinte genes más destacados para la serie GSE52712	45
Tabla 4.2	Pathway de los veinte genes más destacados para la serie GSE52712	46
Tabla 4.3	Veinte genes más destacados para la serie GSE40987	48
Tabla 4.4	Pathway de los veinte genes más destacados para la serie GSE40987	49
Tabla 4.5	Veinte genes más destacados para la serie GSE52262	51
Tabla 4.6	Pathway de los veinte genes más destacados para la serie GSE52262	52
Tabla 4.7	Veinte genes más destacados para la serie GSE12790	55
Tabla 4.8	Función de los veinte genes más destacados para la serie GSE12790	56

Tabla 4.9	Veinte genes más destacados para la serie GSE52712 según Matlab	61
Tabla 4.10	Función de los veinte genes más destacados para la serie GSE52712 según Matlab	62
Tabla 4.11	Veinte genes más destacados para la serie GSE40987 según Matlab	64
Tabla 4.12	Función de los veinte genes más destacados para la serie GSE40987 según Matlab	65
Tabla 4.13	Veinte genes más destacados para la serie GSE52262 según Matlab	67
Tabla 4.14	Función de los veinte genes más destacados para la serie GSE52262 según Matlab	68
Tabla 4.15	Veinte genes más destacados para la serie GSE12790 según Matlab	70
Tabla 4.16	Función de los veinte genes más destacados para la serie GSE12790 según Matlab	71
Tabla 4.17	Veinte genes más destacados para la serie GSE46834	78
Tabla 4.18	Pathway de los veinte genes más destacados para la serie GSE46834	79
Tabla 4.19	Veinte genes más destacados para la serie GSE68651	81
Tabla 4.20	Pathway de los veinte genes más destacados para la serie GSE68651	82
Tabla 4.21	Veinte genes más destacados para la serie GSE46834 según Matlab	87
Tabla 4.22	Función de los veinte genes más destacados para la serie GSE46834 según Matlab	88
Tabla 4.23	Veinte genes más destacados para la serie GSE68651 según Matlab	90
Tabla 4.24	Función de los veinte genes más destacados para la serie GSE68651 según Matlab	91
Tabla 4.25	Series usadas en el estudio integrador final	95
Tabla 4.26	Genes en común comparando series con distintos método de supresión de batch	102

Parte I

INTRODUCCIÓN

INTRODUCCIÓN

ÍNDICE

1.1	Objetivos del proyecto	4
1.2	Cáncer de mama: incidencia y anatomía	5
1.2.1	Definición de cáncer	5
1.2.2	¿Cómo aparece el cáncer?	5
1.2.3	Cáncer de mama	5
1.3	Introducción a los MicroArrays	9
1.4	Repositorio GEO	12

Desde hace décadas, el cáncer es una de las enfermedades más combatida por la humanidad debido al gran número de tipos de cáncer existentes y a la naturaleza diferente de cada uno de estos. A día de hoy los avances en este campo han sido sustanciosos pero aún no se ha obtenido una cura total para esta enfermedad. Una posibilidad para asegurar esa curación es conseguir hacer un diagnóstico precoz de dicha enfermedad de manera que se pueda atacar desde su estado más primario y aumentar de manera elevada la posibilidad de supervivencia a dicha enfermedad. En concreto, el cáncer de mama supone una de las principales causas de muerte de la población femenina mundial y el cáncer más frecuente también entre esta población.

1.1 OBJETIVOS DEL PROYECTO

Como objetivo principal, este proyecto trata de adquirir un conjunto fiable de genes para la detección de cáncer de mama. Este estudio se llevará a cabo haciendo un análisis robusto de genes basado en tecnologías de microarrays [51] diferentes. Para ello se realizará dicho estudio de forma independiente para cada microarray y a continuación mediante un estudio unificado de dichos microarrays a través de la herramienta VirtualArray [18]. Para desarrollar este objetivo y el proyecto en sí se definirán una serie de capítulos en los que se dividirá dicho trabajo, en concreto 6 capítulos:

- El capítulo 1 hará una introducción de objetivos del proyecto, así como una explicación acerca de la patología de cáncer de mama y la repercusión que este tiene en la sociedad actual. También se citarán las tecnologías de microarrays existentes y herramientas a utilizar en el desarrollo del proyecto.
- Una breve descripción del actual estado del arte acerca de la expresión diferencial de genes es lo que se detallará en el capítulo 2, para así meter al lector en contexto sobre las tecnologías y herramientas elegidas para la realización del trabajo.
- En el capítulo 3 se procederá a describir la metodología seguida para la realización de este proyecto, describiendo las herramientas usadas como el lenguaje R mediante RStudio o la herramienta GEOquery entre otras. También abarcará el preprocesamiento de los datos como la normalización o la eliminación de ruido y los análisis de calidad de dichos datos.
- El capítulo 4 será una exposición detallada de los resultados empíricos obtenidos a raíz del estudio realizado en el proyecto. Una vez hallados los resultados, se interpretarán tanto de un modo técnico como de un modo útil para el experto médico
- Para terminar, en el capítulo 5, se expondrán las conclusiones obtenidas en el desarrollo y la evaluación del cumplimiento del objetivo principal y se detallarán futuros posibles trabajos como continuación de este estudio.

1.2 CÁNCER DE MAMA: INCIDENCIA Y ANATOMÍA

1.2.1 *Definición de cáncer*

El cáncer hace referencia a un conjunto de enfermedades que guardan relación en su naturaleza. En todos los tipos de cáncer, se produce una división incontrolada de células o tejidos que acaban invadiendo órganos y tejidos cercanos.

Normalmente, el ciclo de vida de una célula contemplan un crecimiento y división para formación de nuevas células a medida que el cuerpo las va necesitando. Cuando estas células envejecen o son dañadas, mueren y son reemplazadas por células sanas.

En el cáncer ese proceso ordenado pierde ese control y las células viejas o dañadas no mueren como deberían, además se siguen creando células nuevas aun sin ser necesarias y si estás células empiezan a dividirse sin control pueden formar lo que se conoce como tumor.

Esas células cancerosas tienen la capacidad de ignorar las señales que les indican que tienen que dejar de dividirse o que empiece su muerte celular programada, también conocido como apoptosis [11], usada por el cuerpo para eliminar las células que ya no son necesarias o útiles.

1.2.2 *¿Cómo aparece el cáncer?*

El cáncer es una enfermedad genética, esto significa que principalmente es causado por alteraciones en los genes que controlan la forma en que las células crecen y se dividen.

Estos cambios en los genes pueden ser hereditarios, también pueden deberse a errores en la división de las células o por el daño del propio ADN debido a factores causados por ciertas exposiciones del ambiente, como por ejemplo exposición al humo del tabaco o radiaciones.

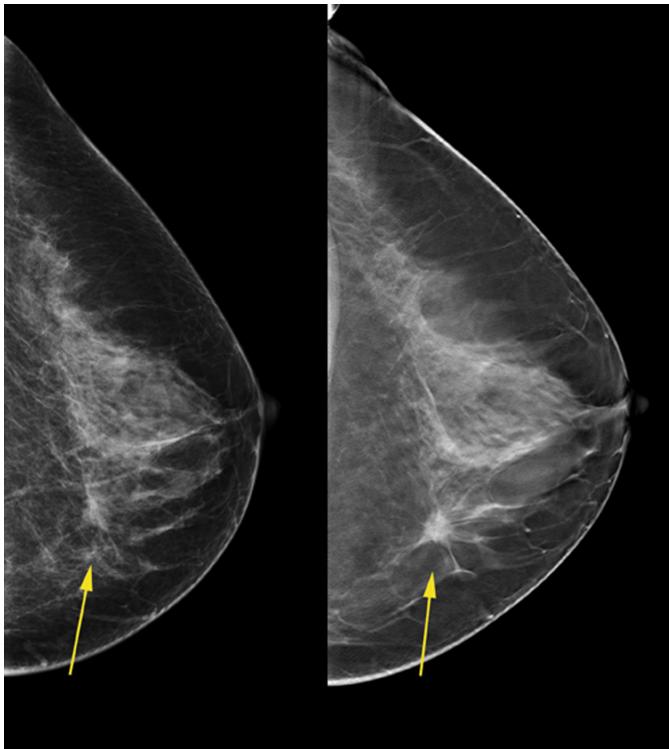
En cada persona el cáncer tendrá una combinación única de cambios genéticos, además ocurrirán cambios adicionales conforme el cáncer se vaya desarrollando.

1.2.3 *Cáncer de mama*

El cáncer de mama es la división descontrolada de las células mamarias, lo que puede dar lugar al desarrollo de un tumor maligno. En la Figura 1.1 se puede observar una masa tumoral recogida a través de una mamografía.

Lo más normal en este tipo de cáncer es que se origine en dos zonas localizadas:

Figura 1.1
Mamografía de una mama afectada de cáncer

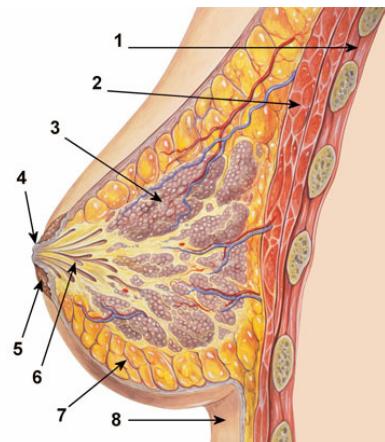


1. En las células de los lobulos, que son las glándulas encargadas de la producción de leche materna. Los lobulos pueden observarse en la Figura 1.2.
2. En los conductos, encargados del transporte de la leche materna desde los lobulillos hasta el pezón. Los conductos también pueden verse reflejados en la Figura 1.2.

En los tejidos estromales también puede originarse este tipo de cáncer, que incluyen tejidos conjuntivos grasos y fibrosos de la mama, aunque en esas zonas se producen con mucha menos frecuencia.

Figura 1.2
Anatomía de la mama. Fuente: <http://www.bonomedico.es/informacion/mamoplastia/aumento-pecho/anatomia-mama>

- 1. Caja torácica
- 2. Músculos pectorales
- 3. Lóbulos
- 4. Superficie del pezón
- 5. Areola
- 6. Conducto lactífero
- 7. Tejido adiposo
- 8. Piel



El cáncer de mama se origina siempre por cambios o anomalías genéticas. De todos los casos, solo alrededor del 5 al 10 % son originados por herencia familiar, el resto están vinculadas al desgaste por la edad.

Las células cancerígenas pueden expandirse e invadir tejidos cercanos sanos así como llegar a los ganglios linfáticos de las axilas.

Si las células consiguen llegar a dichos ganglios, pueden pasar hacia otros tejidos y partes del cuerpo, esto también es llamado metástasis [10]. En la Tabla 1.1 se muestran los posibles estadios del cáncer de mama y lo lejos que pueden llegar a lo largo del cuerpo dependiendo del estado.

Según datos de la Sociedad Española de Oncología Médica (SEOM), el cáncer de mama es el cáncer con mayor incidencia en Mujeres en España, en concreto un 29 % de la población femenina. La mortalidad de este cáncer en España según la fuente antes citada es del 15.5 % situándose en el cáncer con más porcentaje de muerte en la población femenina. También se cita la prevalencia de este cáncer en un tiempo de 5 años con un total del 40.8 %, por tanto cerca de la mitad de la población femenina enferma con este cáncer tendrá esa prevalencia [65].

Según la Organización Mundial de la Salud (OMS), el cáncer de mama representó en 2011 la décima causa de muerte en la población femenina a nivel mundial y la sexta causa en los países con ingresos altos [70]. También se observa como de 8.2 millones de muertes causadas por el cáncer en general en 2012 a nivel mundial, 521.000 de dichas muertes fueron causadas por el cáncer de mama, lo que representa aproximadamente un 6.5 % de las muertes por cáncer a nivel mundial ese año [66].

Según los datos expuestos, se puede observar la magnitud que tiene encontrar una forma eficaz y robusta de diagnosticar el cáncer de mama, que es la principal motivación de este proyecto mediante en análisis diferencial de genes combinando microarrays de diferentes tecnologías.

Tabla 1.1
Posibles estadios

Estadios del Cáncer	Definición
Estadio 0	Las células no han invadido el tejido mamario cercano, permanecen en el conducto mamario.
Estadio IA	El tumor mide hasta 2 cm y el cáncer no se ha extendido más allá de la mama
Estadio IB	No hay tumor en la mama, pero si pequeñas células cancerígenas en los ganglios linfáticos o bien se observa un tumor inferior a 2 cm en la mama y pequeñas células cancerígenas en los ganglios linfáticos.
Estadio IIA	No hay ningún tumor en la mama pero si células cancerígenas en los ganglios linfáticos axilares o bien el tumor mide 2 cm y se ha propagado a los ganglios linfáticos axilares o el tumor mide de 2 a 5 cm y no se propagado fuera de la mama
Estadio IIB	El tumor mide entre 2 y 5 cm y se ha propagado hacia los ganglios linfáticos axilares o el tumor mide mas de 5 cm pero no se ha propagado aún.
Estadio IIIA	No se detecta ningún tumor en la mama y se encuentra en los ganglios linfáticos axilares o en los ganglios linfáticos cercanos al esternón o el tumor es de cualquier tamaño y se encuentra en los ganglios linfáticos axilares o en los ganglios linfáticos cercanos al esternón
Estadio IIIB	El tumor puede ser de cualquier tamaño, además se se habrá propagado a la pared torácica o a la piel de la mama y puede que también a los ganglios linfáticos axilares o cercanos al esternón.
Estadio IIIC	Puede no haber señales de tumor en la mama o puede ser de cualquier tamaño y haberse propagado hacia la pared torácica o piel de la mama y puede que también a los ganglios linfáticos sobre o debajo de la clavícula y puede que se haya propagado a los ganglios linfáticos axilares o a los cercanos al esternón.
Estadio IV	Se ha producido metástasis y se ha propagado por el cuerpo.

1.3 INTRODUCCIÓN A LOS MICROARRAYS

El genoma humano es una secuencia completa de ácido desoxirribonucleico(ADN) del cuerpo humano y contiene las instrucciones genéticas para desarrollar y dirigir las funciones de todo el organismo. Este genoma esta formado por 23 pares de cromosomas y cada cromosoma alberga miles de genes que contienen las instrucciones para crear proteínas. Se estima que el cuerpo humano tiene un total aproximado de 30.000 genes.

El Proyecto del Genoma Humano comenzó en Octubre de 1990 y terminó en Abril de 2003. Gracias a este proyecto se conoce que el genoma contiene 3.000 millones de pares de bases de nucleótidos, también se descubrió que el genoma humano es un 99% igual en todos los seres humanos, aún así se desconoce las funciones que desempeñan el 50% de los genes y a día de hoy continúa investigándose.

En el ADN hay cuatro tipos de nucleótidos:

- Adenina, simbolizado con una A.
- Citosina, simbolizado con una C.
- Timina, simbolizado con una T.
- Guanina, simbolizado con una G.

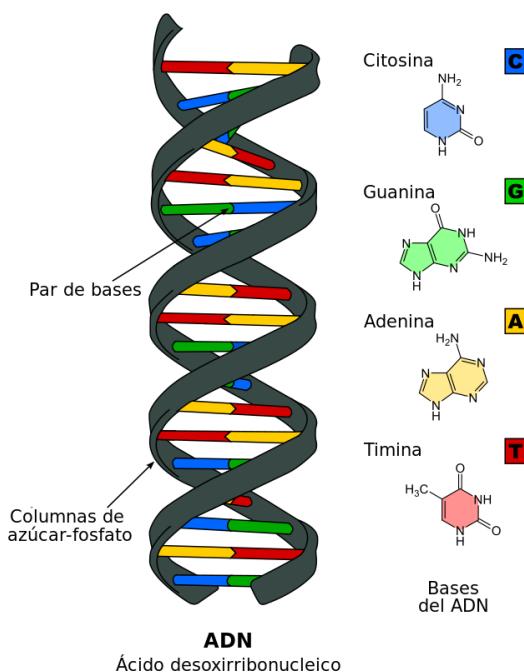
El funcionamiento y fundamento de los micrarray se basa precisamente en el proceso de hibridación del ADN.

El proceso de hibridación del ADN es el proceso mediante el cual dos fragmentos de ADN si hibridan si y sólo si son complementarios entre sí. Para que sean complementarios entre si han de seguir la regla de Watson-Crick según la cual la Adenina (A) se une a la Timina (T) y la Citosina (C) se une a la Guanina (G) tal y como se puede observar en la Figura 1.3.

El proceso que sigue la creación del microarray es el siguiente. Las sondas o 'probes' de oligonucleótidos son adheridas a una superficie de 1 cm² donde se forma un array de ADN consiguiendo que cada sonda diste de las demás en distancias del orden de micras. Cada una de las muestras es sometida a iluminación fluorescente y agregada al array, este proceso puede observarse mejor en la Figura 1.4.

Una vez hecho esto, se procede a eliminar aquel material que no se ha hibridado en cada sonda y al finalizar esto, se pasa un láser al material hibridado de forma que la luz reflejada es detectada por un escáner que va escaneando la superficie del microarray. Posteriormente, mediante

Figura 1.3
Doble hélice de ADN
con los enlaces de
hidrógeno entre los
nucleótidos



un análisis de la imagen del microarray, su puede conocer y cuantificar la proporción de muestra que se ha hibridado en total.

El material genética, ya sea en forma de ADN complementario o de oligonucleótidos, es replicado mediante un proceso muy barato y sencillo e injectado mediante agujas finas controladas por un brazo robótico debido a su precisión dentro del microarray.

Como se puede observar en la Figura 1.4, cada hueco o pocillo del microarray contiene una cadena de nucleótidos asociada a cada gen diferente. Para diferenciar las muestras de los individuos se marcan con sondas fluorescentes. Para terminar se mide la intensidad lumínica en cada uno de los pocillos para saber su cuantía y todos ellos se almacenan en un fichero .CEL, una vez obtenido este fichero, se puede abrir con entornos que permitan analizarlo como RStudio o MatLab.

Existen dos tecnologías principales en la creación de Microarrays, una de ellas llamada Ilumina y la otra conocido como Affymetrix. Según estudios realizados, se ha demostrado que existe una buena correspondencia en los datos proporcionados por Affymetrix y por Ilumina conseguido gracias a la realización de réplicas técnicas en ambas tecnologías.

No obstante, debido a las diferencias en el proceso de creación existentes entre ambas tecnologías, no es trivial saber si los datos de ambas pueden integrarse conjuntamente.

Una peculiaridad de Ilumina es que proporcionan los datos de expresión tanto a nivel de sonda o 'probe' como de genes, esto lo consiguen

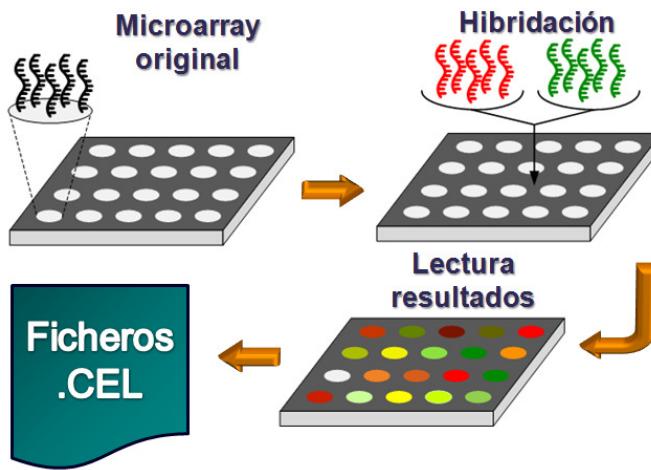


Figura 1.4
Proceso de creación de un microarray. El ARN mensajero es extraído de las células y convertido en ADN complementario y etiquetado de manera fluorescente. La muestra de referencia es de color rojo y verde la muestra de test. Al mezclarlas se pasa una sonda y se eliminan posteriormente el material que no ha hibridado para escanear el chip mediante un láser confocal.

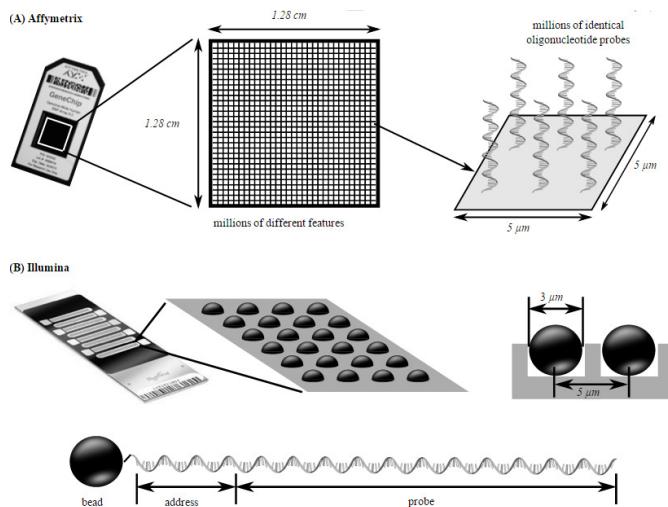
implementando diversas réplicas técnicas en el array, en torno a unas 30 réplicas para cada nucleótido en concreto. Esto quiere decir que el mismo gen puede estar contenido en diferentes sondas. Ilumina también puede proporcionar los datos de expresión por separado de cada una de las muestras o bien de forma agrupada.

Por otra parte Affymetrix usa múltiples 'probes' como controles internos para verificar el correcto funcionamiento y no solo con fines para hibridación.

Affymetrix cuenta con "spotted cDNA", que son tipos de "GeneChips" basados en arrays como el que se muestra en la Figura 1.5. La principal diferencia con los de Illumina es que cada sonda del mallado va en una posición específica y conocida. El proceso en Illumina es diferente debido a que en dicha tecnología existe un paso de decodificación de la posición de cada sonda en el array en función de su dirección molecular tal y como puede apreciarse en la Figura 1.5.

Por último, la hibridación en Illumina se produce en paralelo debido a que se colocan varios arrays sobre el mismo sustrato mientras que en Affymetrix se procesa cada array en un sustrato separado.

Figura 1.5
Representación de un 'Affymetrix GeneChip' y un 'Illumina BeadArray'



1.4 REPOSITORIO GEO

La plataforma GEO (Gene Expression Omnibus) fue creada en un principio para almacenar y proporcionar datos de microarrays referentes a expresión genéticas para que los usuarios pudieran trabajar libremente con ellos. Se puede acceder a el mediante el enlace adjunto a continuación:

<http://www.ncbi.nlm.nih.gov/geo/>

Actualmente, no solo ofrece datos relacionados con microarrays, sino que también ofrece datos extraídos de tecnologías más avanzadas con la secuenciación masiva o NGS(Next Generation Sequencing).

GEO es mantenido por el Centro de Información Biotecnológica de los Estados Unidos o NCBI. GEO forma parte de una red de bases de datos más amplia recogidas mediante el sistema de anotación de genes Entrez, que es una de las maneras por la cual se identifican las características de los genes.

Lo normal es encontrar para cada entrada en el repositorio los datos crudos recogidos en el microarray, los datos de expresión y los datos de anotación. Además, se pueden distinguir cuatro tipos de entidades ordenadas jerárquicamente: Plataformas, Muestras y Series que proceden de contribuciones de los usuarios y, por último, Datasets, generados por el equipo de investigadores bioinformáticos del GEO combinando datos de las Plataformas, Muestras y Series. Dicha información es guardada en un fichero .SOFT(Simple Omnibus Format in Text).

Se procederá a explicar con detalle que son cada una de las entidades anteriormente nombradas:

- Plataformas: Cada identificador de plataforma va unido o a la tecnología usada en el microarray o al secuenciador utilizado a la hora de extraer los valores de expresión. También definen características técnicas, como si es un microarray de ARN mensajero, ARN normal o proteínas, así como la especie en la que se ha empleado(Humano, Raton, etc..). Cada plataforma tiene el siguiente formato: GPLxxx. Donde xxx es un identificador único. Se debe tener en cuenta que una misma plataforma puede estar referenciada o incluida en distintas muestras procedentes de otros usuarios.

Toda la información de cualquier plataforma puede ser encontrada en formato .SOFT o en forma tabulada online, además incluyendo la información de anotación para cada una de las sondas del microarray.

- Muestras: Toda entidad incluida así ha de describir las condiciones en que se registró cada muestra individual así como sus características, la forma en que se manipuló el microarray, la cantidad de muestra existente en ese momento y los datos fenotípicos relevantes(Género, estadio de la enfermedad, etc...). Cada muestra o sample tiene también un identificador único con el formato GSMxxxxx.

Una muestra debe estar referenciada por una única plataforma pero puede estar incluida en mas de una serie. En este nivel se deben incluir los valores de expresión obtenidos para cada una de las sondas de la muestra, de esta forma se puede saber si los valores están en crudo o han sido preprocesados y están, por tanto, normalizados. Cabe la posibilidad de incluir ficheros de imágenes adicional en formato TIFF.

- Series: Una serie puede definirse como un conjunto de muestras que pueden pertenecer a distintas plataformas relacionadas. Las series son identificadas mediante el formato GSExxxxx con un identificador único por tanto. La serie puede obtenerse como documentación en la web o mediante un fichero .SOFT e incluye lo que se conoce como 'Overall Design' o descripción técnica con el número de muestras, etiquetas o réplicas realizadas. GEO también deja disponible los ficheros de expresión del experimento completo indicando si son datos crudos o previamente normalizados. Para el desarrollo de este proyecto se trabajará a nivel de series crudas que ser preprocesadas a posteriori, con el fin de obtener una serie de biomarcadores relevantes y robustos para la patología seleccionada.
- Datasets: Son colecciones de muestras creadas por el equipo del NCBI. Están compuestos por muestras biológicas y estadísticamente comparables. Las muestras de un datasets son extraídas siempre de la misma plataforma y el preprocesamiento realizado es de forma consistente en todas ellas. El formato para identificar un Dataset es el siguiente: GDSxxx.

En este proyecto se usará el término dataset para referenciar las series usadas en este estudio y no para hacer referencia al uso de un dataset de GEO.

2

ESTADO DEL ARTE

ÍNDICE

2.1	Secuenciación clásica	16
2.1.1	Sanger Sequencing	16
2.1.2	Maxam-Gilbert Sequencing	17
2.1.3	Shotgun Sequencing	19
2.2	Next Generation Sequencing	20
2.2.1	Roche 454 Sequencing	20
2.2.2	Illumina (Solexa) Sequencing	21
2.2.3	SOLiD Sequencing	21
2.3	Métodos en desarrollo	22
2.3.1	Nanopore DNA Sequencing	22
2.3.2	Sequencing with mass spectrometry	23
2.3.3	Multiplex polony technology	23

Frederick Sanger fue la primera persona que desarrolló un método para secuenciar ADN, este método es conocido como "Método Sanger"[\[47\]](#). En este capítulo se hará una revisión del estado del arte partiendo de 1975, fecha en la que Sanger planteó su método hasta las más avanzadas técnicas de secuenciación usadas hoy día, como las técnicas de Next Generation Sequencing o NGS [\[13\]](#).

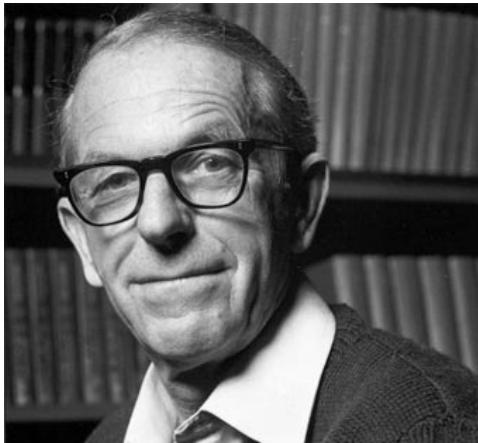
2.1 SECUENCIACIÓN CLÁSICA

En este apartado se expondrán algunas de las principales técnicas de secuenciación genómica que se han usado desde el auge de esta disciplina y las cuales algunas siguen usándose hoy día.

2.1.1 *Sanger Sequencing*

Frederick Sanger [83] ha sido una de las pocas personas del mundo en recibir dos premios Nobel, en concreto en la categoría de Química, puede verse una fotografía de Sanger en la Figura 2.1. El segundo de estos premios Nobel fue gracias a su método de secuenciación de ADN desarrollado en 1975, aunque no sería hasta dos años mas tarde cuando lo usaría para secuenciar el primer ADN de un ser vivo secuenciado en su totalidad, el ADN del bacteriófago Φ -X174 [94].

Figura 2.1
Fotografía de Frederick Sanger



El método de Sanger se basa en la polimerización del ADN y en el uso de dideoxinucleótidos que cumplen la función de finalizadores de la reacción.

El primer paso a seguir en el método de Sanger es calentar la cadena de ADN para separar las dos hebras. Después de separar las dos hebras, se introduce lo que se conoce como un primer o cebador, es una pequeña secuencia de ácido nucleico que inicia todo el proceso de secuenciación al unirse con la ADN polimerasa [55].

Dicho primer es complementario al inicio de la hebra con respecto del gen que se quiere replicar. La DNA polimerasa seguirá replicando hasta que se encuentre con un nucleotido de parada, más conocido como dideoxinucleótido.

El proceso se repite con dideoxinucleótidos de parada de cada uno de los 4 nucleótidos que conforman el ADN: A, C, T y G.

Una vez obtenidos los distintos fragmentos secuenciados del ADN, estos son introducidos en un tubo de electroforesis capilar [22]. En dicho tubo los fragmentos de ADN, unidos a marcas fluorescentes, son pasados por un láser de argón que excita a diferentes longitudes de onda los fragmentos permitiendo así la medición paralela de estos. Una vez hecho esto, Sanger obtuvo satisfactoriamente la secuenciación del ADN. Se puede ver un resumen de este método en la Figura 2.2.

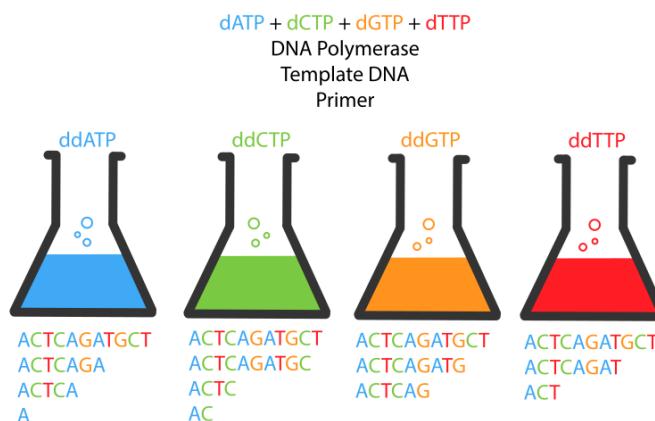


Figura 2.2
Método de secuenciación de Sanger

2.1.2 Maxam-Gilbert Sequencing

Maxam [88] y Gilbert [95] inventaron este método [7] en el año 1976. Era un método muy efectivo aunque limitado por la necesidad de usar secuenciación química. Esto significa que es necesario el uso de procesos químicos para interrumpir las cadenas de ADN. Estos fragmentos que se crean después son pasados a través de un gel para resolver la secuencia del ADN. En la Figura 2.3 puede verse una fotografía de Walter Gilbert.



Figura 2.3
Fotografía de Walter Gilbert

El primer paso para conseguir la secuenciación es el de desnaturalizar la doble hélice de ADN mediante la aplicación de calor. Cuando ambas hebras se han separado, se une radioactivamente el fósforo Gamma- ^{32}p [93] con el extremo 5' del fragmento de ADN mediante una reacción con un polinucleótido cinasa [89].

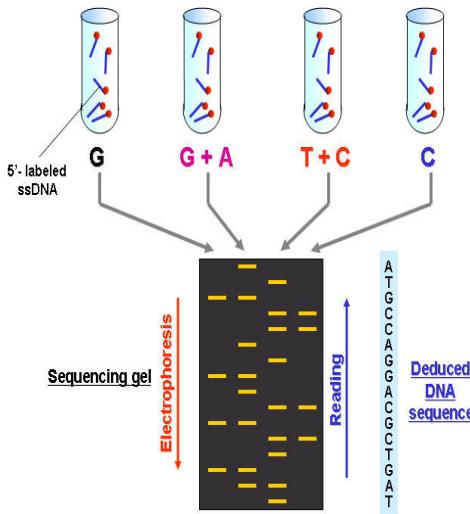
Después se procede a romper la molécula de ADN marcada en uno de los cuatro nucleótidos, de forma que se produzca más o menos una rotura por cada una de las reacciones (G, A+G, C, C+T). Estas reacciones se hacen mediante los siguientes agentes químicos:

- DMS (G)
- Ácido fórmico (A+G)
- Hidrazina (C+T)
- Hidrazinas más sales(C)

Se genera así una serie de fragmentos marcados radiactivamente desde el final hasta el lugar donde se cortó cada molécula.

Una vez que se tienen los fragmentos, se separan por tamaños usando electroforesis en gel [56], separándose así en 4 vías, una al lado de la otra, dependiendo de cada reacción. Para visualizar dichos fragmentos de cada reacción, se hace una autoradiografía [9] de cada fragmento, esto genera unas bandas oscuras que corresponden a cada uno de los fragmentos y que, por inferencia se puede deducir la secuencialidad de la molécula de ADN. Puede verse un resumen de este método en la Figura 2.4.

Figura 2.4
Método de Maxam y Gilbert



2.1.3 *Shotgun Sequencing*

La empresa Celera Genomics [46], representada en la Figura 2.5, fue creada en 1998 con la misión de secuenciar el ADN humano y poder ofrecer a sus clientes un pronto acceso a sus resultados. Celera también fue pionera en la utilización de la técnica de Shotgun Sequencing [58], aunque fue el equipo de Venter [91] en el instituto de investigación genómica(TIGR) en el año 1993 los que crearan esta metodología.



Figura 2.5
Celera Genomics

El método consiste en romper la cadena de ADN en fragmentos al azar que pueden ser leídos por máquinas de secuenciación automática. Una vez hecho esto, se buscan coincidencias mediante software en las secuencias de ADN que se tienen y se van reorganizando hasta volver a reconstruir la cadena original. Debido a la inmensa cantidad de datos generados, los software usados para la reconstrucción del genoma son muy complejos y abarcan campos como la medicina, la estadística, la teoría de grafos o la ingeniería. Puede verse el funcionamiento de este método en la Figura 2.6.

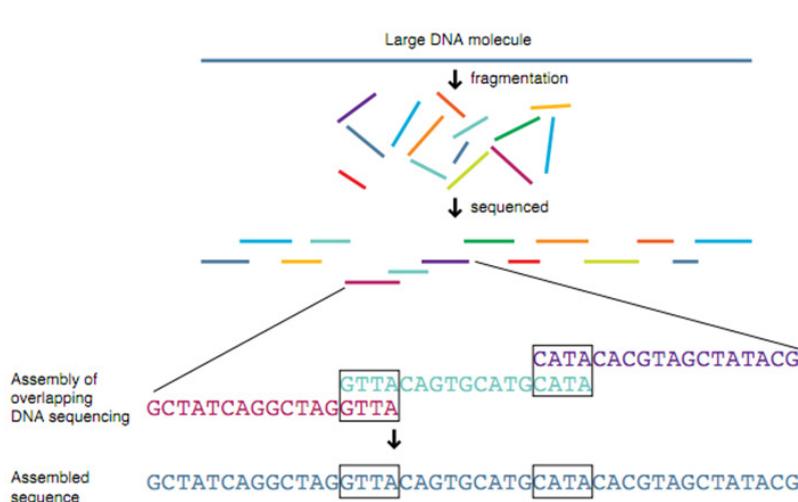


Figura 2.6
Shotgun Sequencing

2.2 NEXT GENERATION SEQUENCING

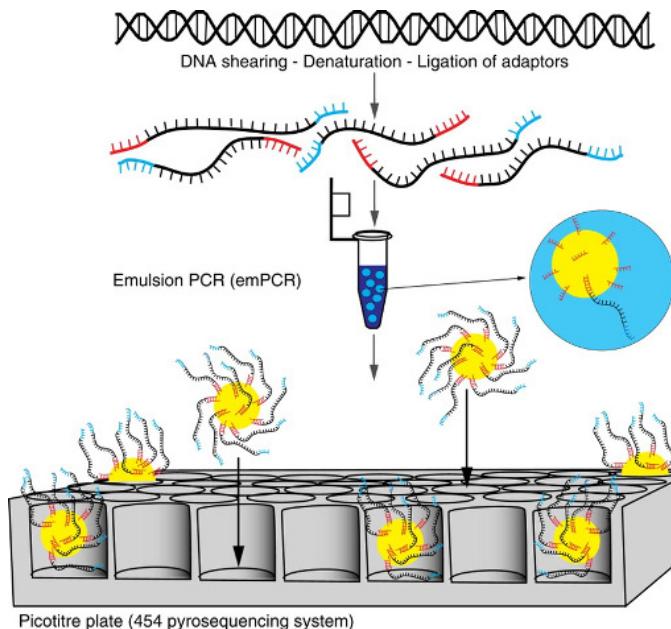
Next Generation Sequencing hace referencia también a secuenciación de alto rendimiento. Las tecnologías bajo este tipo de secuenciación permiten secuenciar ADN y ARN de una forma mucho más rápida y barata que las técnicas de secuenciación más clásicas y gracias a ello, han revolucionado el estudio de la genómica y la biología molecular.

2.2.1 Roche 454 Sequencing

Es considerado el primer método de secuenciación de nueva generación disponible en el mercado [6]. La secuenciación puede realizarse partiendo de diferentes tipos de ácidos nucleicos, como ADN, productos de PCR, BACs o cDNA. Lo primero que se hace es reducir las muestras a fragmentos de 300 a 800 par de bases(pb) de longitud. Una vez realizado esto, los fragmentos se unen a nanoesferas y la secuencia comienza mediante la síntesis de la cadena complementaria mediante el ADN polimerasa.

El proceso de amplificación o lo que es lo mismo, realizar clones del mismo fragmento, se realiza en la propia nanoesfera como si se tratase de una PCR. Mediante una pirosecuenciación [57] adaptada para este tipo de análisis se realiza la identificación de las secuencias. Puede verse el funcionamiento de este método en la Figura 2.7.

Figura 2.7
Roche 454 Sequencing



2.2.2 Illumina (Solexa) Sequencing

Esta técnica consta principalmente de dos partes [40]. En primer lugar el genoma es fraccionado o dividido, una vez esta dividido, se procede a una amplificación clonal de los fragmentos del ADN sobre la superficie sólida de una celda de flujo. Así se consiguen racimos de fragmentos clonados.

En segundo lugar, se procede a una secuenciación mediante síntesis. En este sistema se emplean los cuatro nucleótidos (A, C, G y T) marcados mediante fluorescencia con diferentes colores para secuenciar los millones de racimos presentes en la celda de flujo. Es característico de estos nucleótidos modificados el contener terminaciones reversibles, esto permite que cada ciclo del proceso ocurra con los cuatro nucleótidos presentes al mismo tiempo. En la Figura 2.8 se puede observar como funciona esta técnica.

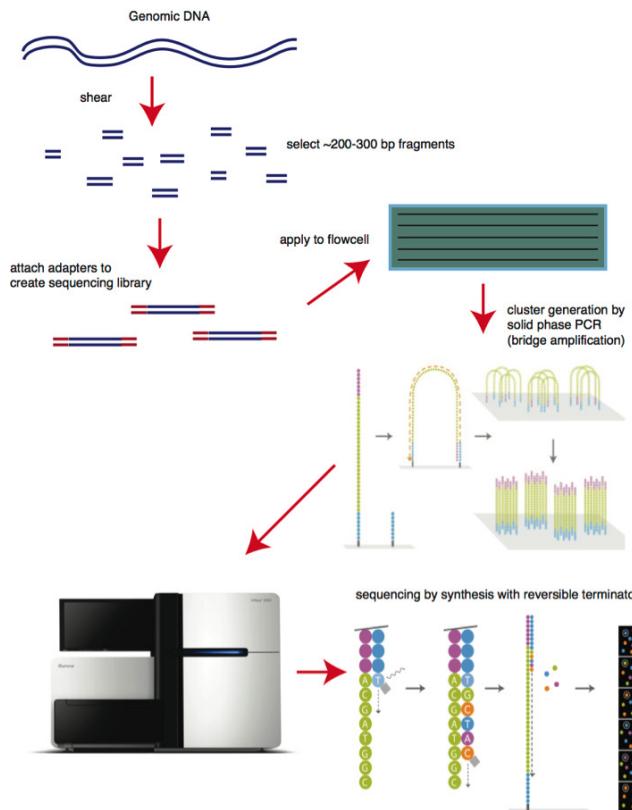


Figura 2.8
Illumina Solexa Sequencing

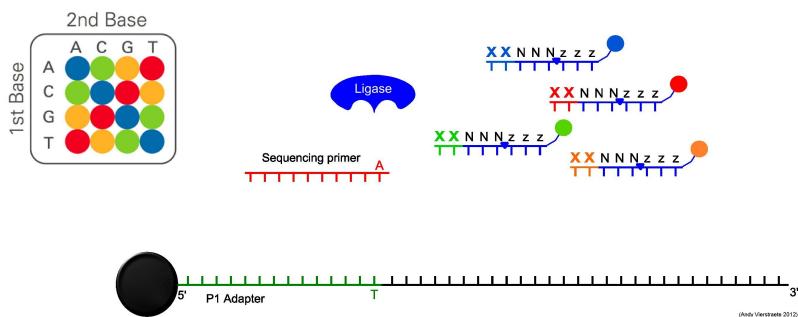
2.2.3 SOLiD Sequencing

Este método [19] fue el último que salió al mercado en comparación con los demás mencionados anteriormente, en concreto en 2008. Este método permite usar distintos tipos de muestras ya sean fragmentos

sueltos o agrupaciones de fragmentos unidos mediante un cebador. El proceso de amplificación es prácticamente igual al utilizado en el método 454, lo realmente innovador es el método de secuenciación usado.

El proceso de secuenciación cuenta con varias rondas de hibridación ligadas a 16 nucleótidos marcados con 4 colores distintos. Empleando un código de colores, se evalúa cada posición dos veces y gracias a esto se aumenta drásticamente la discriminación entre errores de secuencia y polimorfismos de SNPs, estos polimorfismos son variaciones en la secuencia de ADN que afectan a una sola base(A, C, G o T) de una secuencia del genoma. Los SNP constituyen el 90 % de los cambios genómicos humanos. En la Figura 2.9 se puede observar como funciona esta técnica.

Figura 2.9
SOLiD Sequencing



2.3 MÉTODOS EN DESARROLLO

Para terminar este capítulo de estado del arte, también se hablará de las futuras técnicas que prometen revolucionar una vez más, como lo han hecho las técnicas de nueva generación, el campo de la genómica y la secuenciación. Son técnicas que aun están en desarrollo en diferentes lugares del planeta, pero que ya prometen unos resultados brillantes.

2.3.1 Nanopore DNA Sequencing

Las principales ventajas de esta técnica son la capacidad de leer cadenas muy largas de bases, un alto rendimiento y poca cantidad de material necesario para llevarla a cabo.

Esta tecnología [79] se originó a partir de un contador de Coulter [90] y canales de iones [84]. Al aplicarle un voltaje externo, partículas con tamaños ligeramente más pequeñas que el tamaño del poro pasan a través de él. Los poros, de tamaño nanométrico, se incrustan en una membrana biológica o formando una película sólida y así se separan los depósitos que contienen los electrolitos conductores.

Los electrolitos están inmersos en cada uno de los depósitos como se muestra en la figura. Bajo una tensión polarizada negativa, los iones se mueven a través del poro por electroforesis, generando una señal de corriente iónica.

Cuando el poro se bloquea, por ejemplo por el paso de una molécula de ADN, la corriente que fluye a través del nanoporo se bloquearía también y por tanto, se interrumpe la señal de corriente. Las propiedades tanto físicas como químicas de la moléculas pueden ser calculadas mediante el análisis estadístico de la amplitud y duración de los bloqueos generados en la corriente a través del poro, de esta forma se puede leer la translocación de la molécula de ADN. En la Figura 2.10 se puede observar como funciona esta técnica.

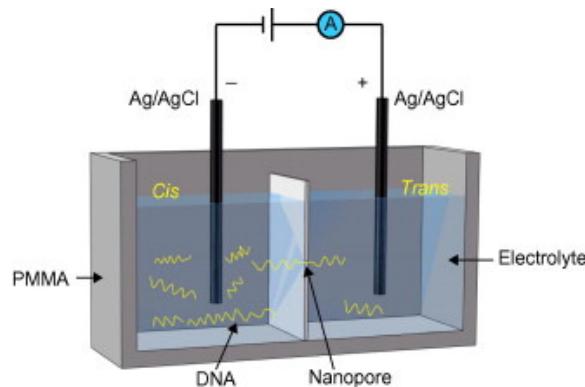


Figura 2.10
Nanopore Sequencing

2.3.2 Sequencing with mass spectrometry

Esta técnica [28] se puede usar para determinar la secuencia de las moléculas de ADN, aunque no solo para eso. El ADN se fragmenta y se hace que vayan pasando los fragmentos por un espectrómetro de masa [2] pueden medirse las masas de cada nucleótido y al ser diferentes, ir averiguando la secuencia de ADN.

2.3.3 Multiplex polony technology

Esta técnica [86] ha sido diseñada por el grupo de investigación del Prof. G Church. En este método, varios cientos de plantillas de secuenciación se depositan en capas delgadas de gel agarosa y las secuencias se calculan en paralelo.

Este método presenta un aumento de varios órdenes de magnitud en el número de muestras que pueden analizarse simultáneamente. Tiene la ventaja, de una gran reducción de los volúmenes de reacción, lo que requiere menores cantidades de reactivos y, por tanto, un costo mucho menor para aplicar la técnica.

Parte II

DESARROLLO DEL ESTUDIO Y RESULTADOS

3

METODOLOGÍA

ÍNDICE

3.1	Metodología de trabajo	28
3.2	Procesamiento de los datos de GEO mediante R	29
3.3	Análisis de calidad y eliminación de los Outliers	30
3.4	Preprocesamiento de las series	31
3.5	Análisis de expresión diferencial de los genes	32
3.6	Integración de las series	33
3.7	Efecto Batch	34
3.7.1	Métodos de eliminación del efecto batch basados en escalamado	36
3.7.2	Métodos de eliminación del efecto batch basados en discretización	38
3.8	Evaluación de las técnicas de integración en microarrays	39

Analizar microarrays tiene como objetivo principal encontrar un conjunto de probes que contengan genes con niveles de expresión anómalos, como pudieran ser por sobreexpresión o por inhibición con respecto a una persona que no padezca la enfermedad a estudiar, el cáncer de mama en el caso de este estudio. Los genes resultantes se conocen como genes diferencialmente expresados, también conocidos como DEG(Differentially Expressed Genes). Estos genes proporcionan información muy valiosa sobre multitud de enfermedades, en nuestro caso, las diferencias entre estos genes, pueden asociarse a presencias o no de tumores mamarios.

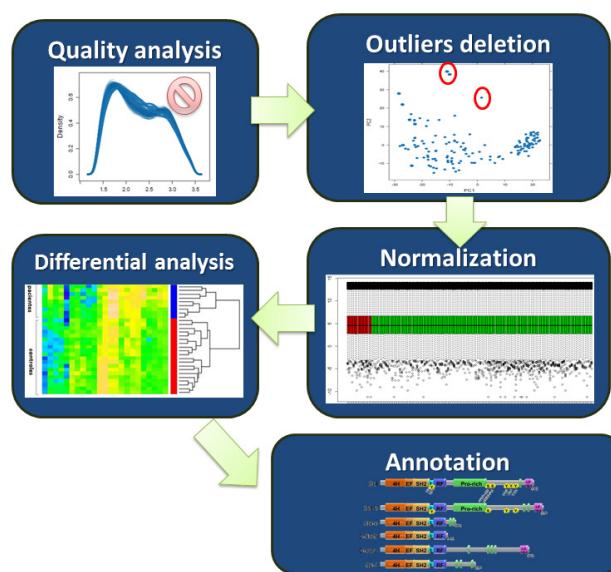
En este capítulo se procederá a explicar la metodología seguida durante este trabajo para conseguir identificar estos genes diferencialmente expresados.

3.1 METODOLOGÍA DE TRABAJO

Este proyecto sigue la metodología o filosofía de trabajo conocido como "pipeline", usada normalmente cuando se tiene un problema de análisis de expresión diferencial en microarrays. En la Figura 3.1 se pueden observar las distintas fases que se llevaran a cabo en el análisis diferencial mediante el lenguaje de programación R [71].

El primer paso que se tiene que dar es hacer un análisis de calidad, de esta forma se pueden descubrir posibles muestras que añadan algo de ruido a los resultados del estudio y que deberían ser eliminadas al ser de baja calidad. Una vez hecho dicho análisis, se ha de proceder a eliminar dichas muestras, también conocidas como outliers y volver a repetir el análisis por si surgieran más muestras de baja calidad y con valores muy distintos del resto de muestras. Una vez hecho esto, se debe proceder a normalizar todas las muestras para mantener todos los genes en unos valores de rango dinámico similares, eliminando así las variabilidades no relacionadas con las diferencias propiamente biológicas. Por último, se realiza el análisis de expresión de los genes y los resultantes han de ser anotados para poder localizar el símbolo de cada gen de forma correcta cotejando con las librerías genéticas de la plataforma y especie usadas en el estudio. Una vez conseguida esa lista de genes, un experto médico puede determinar como de relevantes son y si está realmente relacionado con la enfermedad que se está estudiando. Cada una de estas fases será explicada con mas detalle en los siguientes puntos de este capítulo.

Figura 3.1
Metodología Pipeline
en expresión de genes



3.2 PROCESAMIENTO DE LOS DATOS DE GEO MEDIANTE R

El lenguaje R ha sido el elegido para el desarrollo de este proyecto gracias a su manejo estadístico y a las librerías de Bioconductor disponibles para él. Estos datasets serán descargados a través de la plataforma GEO explicada en el capítulo 1. Estos datos pueden ser descargados directamente desde la plataforma web de GEO o también existe la opción de descargarlos mediante la librería GeoQuery [15] desde R sin tener que acceder a la web. Normalmente, los datasets descargados vienen ya previamente normalizados y con los outliers eliminados pero, existe también la opción de descargar esos datasets en crudo(RAW) para hacer manualmente todo el proceso de eliminación de outliers y normalización. En este proyecto se trabajará siempre con datos en crudo para asegurarse que todas las normalizaciones se hayan hecho mediante las mismas técnicas en todos los datasets.

Otras librerías fundamentales serán las necesarias para leer los datos de los microarrays de Affymetrix e Illumina desde R. Estas librerías son la librería “affy” [14] y la librería “lumi” [17]. Para los análisis de calidad de los datasets se ha utilizado la librería “arrayQualityMetrics”, que proporciona un análisis completo en formato HTML [85] con representaciones gráficas y métricas estadísticas de calidad para las muestras del datasets. El informe resultante también te da posibles outliers que ha detectado para ser eliminados al ser de baja calidad con respecto al resto.

Por otro lado, para realizar los análisis diferenciales de los genes tanto en datasets de Affymetrix como de Illumina se ha usado la librería “limma” [16]. En esta librería se incluyen métodos como “lmFit”, usado para crear un modelo lineal de la expresión de los genes. Otro método incluido es la función “eBayes”, la cual hace una regulación del error estándar para cada gen.

Para la integración de las diferentes tecnologías, tanto de Affymetrix como de Illumina, se usará la librería “virtualArray”. Para usar esta librería primero se han de preprocesar los datos de manera independiente en cada una de las series que se tengan para unir. VirtualArray también permite modificar una gran cantidad de parámetros, como el método de unión de probes, el umbral de significación P-value, fold-change(FC) mínimo para considerar relevante las diferencias estadísticas (entendiendo fold-change como el cociente entre los promedios de expresión de un determinado gen ante dos condiciones analizadas, en el caso de este proyecto, expresión en pacientes con cáncer de mama y muestras de personas sanas o controles). El enfoque que se seguirá será, por tanto, similar al propuesto por Taminau et al. [41] que consiste en un análisis a dos niveles: a nivel de meta-análisis, que es un análisis independiente de cada serie y de cada tecnología y luego otro análisis integrador, en el cual se analizan las series integradas de una misma enfermedad y en condiciones comparables. Por último, se ha

de usar la librería “annotate” de R, encargada de establecer la relación entre los identificadores (“Entrez”) de los genes expresados con sus nombres reales mediante los cuales los médicos puedan identificarlos. Gracias a estos identificadores se puede extraer información acerca de la funcionalidad de los genes, estructura, secuencia y otras características importantes de los mismos.

3.3 ANÁLISIS DE CALIDAD Y ELIMINACIÓN DE LOS OUTLIERS

Lo primero que se ha de hacer a la hora de afrontar un análisis de calidad es la eliminación de las muestras de baja calidad, también conocidas como outliers, y de las muestras que no interesan debido a su incompatibilidad con el estudio que se esté realizando en ese momento. Es muy importante documentarse acerca de cada serie usada en el estudio, las muestras han de ser de la misma variable biológica de interés, así como haber sido registradas en las mismas condiciones y bajo una tecnología lo suficientemente fiable.

En este proyecto, se ha evaluado la calidad de las muestras mediante el análisis creado a partir de la herramienta virtualArray, observando así las distribuciones de intensidad de los array y la prueba de bondad de ajuste de Kolmogorov-Smirnov [34] entre la distribución de cada array o muestra y la distribución del conjunto total de datos de la serie. Al usar este test para evaluar la calidad obtenida en el preprocesamiento de los datos, se asume la hipótesis de que si es tratamiento de estos datos se ha hecho correctamente, la distribución de los datos apenas ha de haber sufrido modificación, esto quiere decir, que los valores del test de Kolmogorov-Smirnov han de ser pequeños si son dos distribuciones muy parecidas.

Mediante Kolmogorov-Smirnov pueden detectarse outliers en el conjunto de datos, una vez detectados y en función de la cantidad de muestras que tengamos y de lo extremos que sean esos outliers se decidirá si habrá que eliminarlos o no. Lo ideal es después de eliminar estos outliers, volver a realizar el análisis de forma iterativa debido a posibles outliers enmascarados por los outliers anteriormente detectados hasta que se hayan eliminado todos ellos o hasta que se decida dejar de eliminarlo por alguna de las cuestiones anteriormente explicadas.

Han de analizarse las distribuciones de intensidad de todos los arrays del estudio antes y después de la eliminación de los outliers para así poder verificar que las intensidades se han vuelto más homogéneas que antes de la eliminación de dichos outliers.

3.4 PREPROCESAMIENTO DE LAS SERIES

Esta fase es imprescindible para poder minimizar el número de falsos positivos en los genes destacados, de esta forma se pueden evitar o minimizar las posibles variaciones técnicas que existen al hacer el propio experimento. Un punto a tener en cuenta es que esta fase está relacionada directamente con la tecnología usada para tomar los valores de las muestras y, como tal, las herramientas que se usarán dependerán de la tecnología usada.

En el caso de trabajar con Affymetrix, existen tres fases consecutivas:

1. La primera fase es la corrección del efecto de fondo, esto consiste en una corrección de los valores de intensidad para evitar así ruido debido a la iluminación. Se basa en ajustar los valores de intensidad lumínica para que no haya ruido en la fluorescencia y poder incrementar la sensibilidad de las medidas realizadas. Cuando se aglomeran diferentes conjuntos de datos, también conocidos como “batches”, se puede producir el conocido como efecto batch [23] que consiste en una serie de errores inevitables y que no están relacionados con variaciones biológicas.
2. “Between-array normalization”, también conocido como normalización intra-array entre los diferentes conjuntos de datos para poder corregir así la variabilidad de intensidad lumínica de origen técnico ya que los valores de la expresión de los genes han de estar en un rango que se pueda comparar.
3. El último paso es normalizar los valores de expresión de cada gen para todas las muestras que estén contenidas en el estudio, este paso también es conocido como “Reporter summarization”.

Aplicando las fases anteriores se consigue reducir las diferencias de origen técnico entre los diferentes conjuntos de datos de Affymetrix. Estas tres fases de normalización están incluidas en la función RMA (“Robust Multi-Array Average”) [81] en el paquete “affy” de R.

Si por el contrario, se trabaja con Illumina, la finalidad del proceso es la misma y también se hace en tres fases pero esta vez desde el paquete “lumi”:

1. Lo primero es hacer una corrección del efecto de fondo de manera muy similar al llevado a cabo en la tecnología Affymetrix.
2. Lo segundo es hacer una estabilización de la varianza, más conocida como “Variance Stabilizing Transform” [87], y es fundamental para la correcta detección de los genes expresados. El proceso consiste en la estabilización de las varianzas de la expresión de los genes mediante el uso de una serie de réplicas técnicas, también conocidas como beads, disponibles en los microarrays de Illumina.

3. Por último, se hace una normalización de los datos mediante el algoritmo “Robust Spline Normalization”, el cual combina métodos de normalización de cuartiles y de regresión local, conocidos con las siglas LOESS [67].

Para llevar a cabo esta tres fases, existe la función “lumiExpresso” que es la equivalente a RMA en Affymetrix. Antes de hacer este análisis es conveniente, en Illumina, usar la función incluida en la librería “lumi” conocida como “*detectionCall*”, la cual reduce el número de genes a estudiar. Esta función elimina todos aquellos genes con una variabilidad en las réplicas muy grande debido a que esto puede ocultar la variabilidad biológica subyacente y no deben ser considerados en el análisis.

3.5 ANÁLISIS DE EXPRESIÓN DIFERENCIAL DE LOS GENES

Una vez están los datos preparados, se procede a emplear la librería “limma” para encontrar aquellos genes que están estadísticamente expresados diferente del resto. Esta librería cuenta con una función llamada “*topTable*” que devuelve una tabla con dichos genes expresados.

Para detectar estos genes, suele usarse un fold-change como mínimo equivalente a 2, esto quiere decir que para que un gen esté diferencialmente expresado, la expresión de un grupo debe ser como mínimo el doble en media que en otro grupo. En los análisis que se realizarán en los capítulos siguientes, el fold-change usado será más restrictivo y será próximo a 3. Otro estadístico que se usará es el estadístico p-value y devuelve información acerca de la relevancia estadística de las diferencias de expresión entre grupos, o lo que es lo mismo, da información de la probabilidad de un cambio fruto del azar en la expresión de un gen. El valor asignado normalmente a este estadístico equivale a 0.05 mas en este estudio se usará un valor más restrictivo, siendo este equivalente a 0.001. Debido a la gran cantidad de muestras que se tienen, se puede ser más restrictivo y exigente con la calidad de estas. Se debe de tener en cuenta, aparte del p-value, la versión ajustada que nos ofrece la librería “limma” que además, está corregido para el caso de comparaciones múltiples y para minimizar el número de falsos positivos. Por ello, se ha de tener en cuenta al usar el test estadístico “T-student” el cual se aplica a todos los genes para reducir el número de falsos positivos. También se suele emplear un método de corrección conocido como “Benjamini and Hochberg False Discovery Rate” o FDR [92].

Para terminar a modo de resumen, una vez finalizada la expresión diferencial de genes se obtendrá para cada uno los siguientes parámetros estadísticos:

- Fold-change en escala logarítmica en base 2.
- Estadístico T-Student.

- p-value, tanto ajustado como sin ajustar.
 - Coeficiente de Bondad (B). Este estadístico guarda relación con la posibilidad de que el gen este diferencialmente expresado. Si la bondad, por ejemplo, vale 0, existe una posibilidad del 50 % de probabilidad de que el gen este diferencialmente expresado. Cuanto mayor sea la bondad, mayor será la probabilidad de que el gen este realmente expresado o inhibido.
-

3.6 INTEGRACIÓN DE LAS SERIES

Típicamente las series contienen un bajo número de muestras a estudiar y si se pretende hacer un estudio lo suficientemente amplio y con significación estadística se necesita un mayor número de muestras. Para ello pueden integrarse varias series de diversos estudios siempre y cuando estos pertenezcan a una misma especie biológica, estén orientados a una misma enfermedad y hayan sido tomados de una manera comparable. Una vez integradas las series, el estudio es llevado a cabo con un proceso similar al explicado a lo largo de este capítulo. Para realizar esta integración se usará la librería “VirtualArray” y habrá que tener en cuenta tres aspectos antes de realizar dicha integración:

1. Se tiene que mantener la misma transformación logarítmica en la expresión de los diferentes arrays. Depende de si los datos de GEO son crudos o no, se les ha podido pasar una transformación logarítmica previa y se han de comprobar antes de la integración que todas las muestras están en un rango de valores comparables.
2. Mantener la profundidad de precisión de los datos o el número de bits que se utiliza para la representación de dichos datos. Las tecnologías más importantes usan una representación entre 16 y 20 bits, no obstante, en GEO aun se pueden encontrar tecnologías obsoletas de 12 bits. Si las muestras no tienen el mismo rango dinámico de representación, han de ser igualadas antes de su integración.
3. Debe tener cada gen su correcta y consistente anotación para así poder integrarlos correctamente.

Se pueden configurar varios parámetros de la integración mediante la misma librería “VirtualArray”:

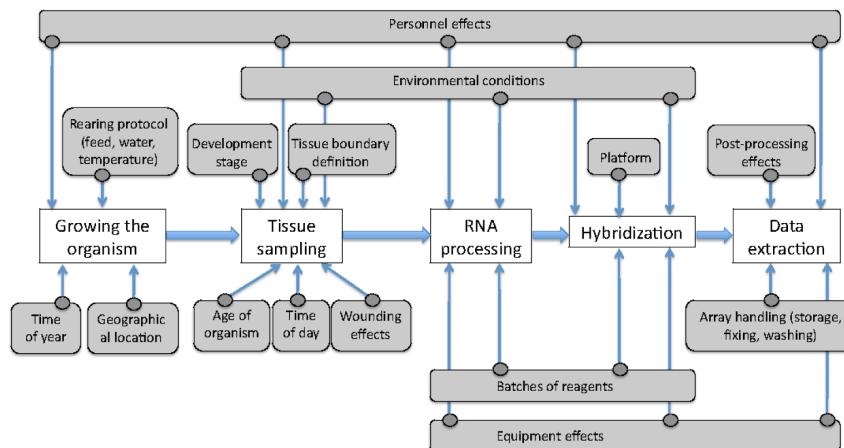
- Combinación de sondas redundantes o réplicas del mismo gen. R, por defecto, emplea la mediana para promediar sus valores, aunque también se puede indicar que use la media u otra función.
- Eliminación del efecto “Batch”, estudiado más profundamente en el siguiente apartado del capítulo.

3.7 EFECTO BATCH

Cuando se trabaja con datasets heterogéneos aparece lo que se conoce como efecto "batch". El efecto "batch" se produce al usar tecnologías diferentes en las series de la integración y no se puede eliminar al quitar el efecto de fondo simplemente al normalizar los datos. Además, pese a ser tomadas las medidas con un mismo instrumental y proceso, pueden llegar a ser un conjunto muy importante de medidas con un comportamiento muy diferente al resto. El origen de este efecto "batch", bias o desviación puede ser muy diversas. Lo cierto es que para tomar las muestras para los análisis diferenciales, estas han pasado antes por varias etapas que son susceptibles a provocar cambios en las medidas involuntariamente(Crecimiento del tejido, selección de muestra, procesamiento del ARN, hibridación y extracción de la información genética).

Todas estas posibles variaciones que pueden dar lugar al efecto "batch" se pueden ver representadas en la Figura 3.2.

Figura 3.2
Posibles causas del efecto Batch



Antes de estudiar las técnicas existentes para reducir el efecto batch se empleará la notación usada en la Tabla 3.1 [49]. Estas técnicas se han de aplicar con los datos ya normalizados, por ejemplo, usando RMA. Se asumirá también, que el efecto batch puede aparecer como un término aditivo y/o multiplicativo, aunque normalmente será aditivo al trabajar con los datos en escala logarítmica.

Parámetro	Definición
$X^{m*n}, \gamma^{m'*n'}$	Datos para estudio MMAGE con m/m' genes y n/n' muestras de tejido
$\hat{X}^{m*n}, \hat{\gamma}^{m'*n'}$	Datos para estudio MMAGE una vez aplicada corrección de efecto batch con m/m' genes y n/n' muestras
X_{ij}, γ_{ij}	Expresión del gen i en la muestra j -ésima del conjunto de datos
$\hat{X}_{ij}, \hat{\gamma}_{ij}$	Expresión del gen i en la muestra j -ésima del conjunto de datos tras suprimir efecto batch
$\bar{X}_i, \bar{\gamma}_i$	Valor medio de la expresión del gen i -ésimo en el conjunto de datos
$\sigma_{x_i}, \sigma_{\gamma_i}$	Desviación estándar del gen i -ésimo en el conjunto de datos
b_{ij}^x, b_{ij}^γ	Bias del gen i -ésimo en la muestra j -ésima en el conjunto de datos
$\epsilon_{ij}^x, \epsilon_{ij}^\gamma$	Ruido en la medición del gen i -ésimo en la muestra j -ésima en el conjunto de datos
$\gamma_i^\gamma, \gamma_i^y$	Bias aditivo del gen i -ésimo en el conjunto de datos
$\delta_i^x, \delta_i^\gamma$	Bias multiplicativo del gen i -ésimo en el conjunto de datos

Tabla 3.1
Notación en las técnicas de la eliminación del efecto batch

La medición del valor de expresión del gen i -ésimo en la muestra j -ésima del dataset X puede calcularse a través de la ecuación 3.1:

$$x_{ij} = x'_{ij} + b_{ij}^x + \epsilon_{ij}^x \quad (3.1)$$

En la ecuación anterior, el término x'_{ij} representa el valor real de la expresión del gen i -ésimo en la muestra j -ésima, por otra parte, el término b_{ij}^x es el efecto batch que se pretende corregir, por último, el término ϵ_{ij}^x representa el ruido inherente que existe en cualquier medición. Los métodos empleados para la eliminación del efecto batch se basan en descomponer el término b_{ij}^x de diferentes formas. Una forma posible es conociéndose alguna de covariante(variable que pueda predecir el resultado probablemente) tratar de aislar su influencia particular. Para que estas técnicas tengan un correcto funcionamiento, es necesario que en los estudios MMAGE usados contengan todos las mismas distribuciones de muestras para cada variable de interés.

La librería VirtualArray incluye algunas técnicas de corrección del efecto batch, las cuales serán explicadas a lo largo de los siguiente apartados y asumiendo que se tienen dos conjuntos de datos que deben de ser integrados y que se denotarán como X e γ respectivamente. En las ecuaciones 3.1 y 3.2 se pueden observar como se pueden hallar estos datos de expresión de los genes para ambos conjuntos de datos.

$$\gamma_{ij} = \gamma'_{ij} + b_{ij}^{\gamma} + \epsilon_{ij}^{\gamma} \quad (3.2)$$

El objetivo es conseguir que al combinar ambos conjuntos, las muestras sean comparables entre sí para poder hacer el análisis posterior de forma correcta. Se puede enfocar de distintas maneras, una de ellas es intentar eliminar la influencia del término bias. Otra puede ser aceptar la presencia de dicho término y con ello ajustar los rangos dinámicos de X_{ij} y γ_{ij} para que dichos datasets sean comparables entre sí. Las técnicas que se emplearán pueden dividirse en dos grupos principalmente:

- Métodos basados en escalado. En estos métodos se asume un modelo para el valor medio y/o la varianza de los datos con el efecto batch ya incorporado y mediante una serie de procedimientos se puede ajustar dicho batch para que así concuerde con el modelo de hipótesis planteado.
- Métodos basados en discretización. En este tipo de métodos se marcan con códigos binarios los valores de cada gen para un cierto nivel de expresión, 1 significará que un gen esta expresado y 0 que no lo está.

3.7.1 *Métodos de eliminación del efecto batch basados en escalado*

La principal idea en la que estos métodos se basan es transformar cada uno de los conjuntos de datos para que estos tengan un valor similar en cada uno de los genes, ya sea en media como en varianza. Por ello, estas transformaciones no eliminan información biológica de interés. Dentro de estos métodos se pueden distinguir los siguientes:

- Mean Centering (MC). Este método se basa en restarle a cada dato la media de cada gen en todas las muestras del mismo dataset, con ello se consigue que la media de cada gen pase a ser cero. Para este método, el segundo término de la ecuación 3.1, b_{ij}^X , representa el efecto batch multiplicativo en vez de aditivo de cada gen. De esta forma, la expresión final de los genes al eliminar el efecto batch en ambos conjuntos quedaría de la siguiente manera:

$$\hat{X}_{ij} = X_{ij} - \bar{X}_i \quad (3.3)$$

$$\hat{\gamma}_{ij} = \gamma_{ij} - \bar{\gamma}_i \quad (3.4)$$

Una modificación en el método anterior da lugar al conocido como estandarización de genes, donde además de dejar la media de los genes a cero, se consigue que todos los genes tengan también desviación estándar uno, mediante una estandarización Z-score [76]. En este caso la expresión de los genes quedaría de la siguiente forma:

$$\hat{X}_{ij} = \frac{X_{ij} - \bar{X}_i}{\sigma_{xi}} \quad (3.5)$$

$$\hat{\gamma}_{ij} = \frac{\gamma_{ij} - \bar{\gamma}_i}{\sigma_{\gamma_i}} \quad (3.6)$$

- Método empírico de Bayes (EB). En dicho método se usan las medias y la varianza de cada uno de los genes. Para hacer el cálculo de los parámetros se une la información procedente de multitud de genes con características parecidas de expresión en cada conjunto de datos. De esta forma, el valor de expresión del gen i-ésimo en la muestra j-ésima del efecto batch se puede calcular como:

$$X_{ij} = \alpha_i + C\beta_i + \gamma_i^X + \delta_i^X \epsilon_{ij}^X \quad (3.7)$$

La ecuación 3.7 es una particularización de la ecuación 3.1, en la cual α_i es aquella parte del gen i-ésimo que no se puede calcular a partir de una covariable conocida. C es una matriz de diseño para las muestras que se asocian a covariables conocidas, por otra parte, β_i es un vector de los coeficientes de regresión asociados a la matriz C . El efecto batch viene representado mediante un término multiplicativo(δ_i^X) y otro aditivo(γ_i^X). Se asume que el ruido ϵ_{ij}^X sigue una distribución normal de media cero y varianza σ_i^2 .

El primer paso de este método consiste en estandarizar los datos empleando las estimaciones de los parámetros de la ecuación anterior, los cuales se denotarán de la siguiente manera: $\tilde{\alpha}_i$, $\tilde{\beta}_i$, $\tilde{\delta}_i^X$ y $\tilde{\sigma}_i^2$. De esta forma, la expresión estandarizada del gen z_{ij} se puede asumir como normalmente distribuida siguiendo una nor-

mal $N(\gamma_i^X, (\delta_i^X)^2)$ y que se define a partir de la siguiente ecuación 3.8:

$$z_{ij} = \frac{X_{ij} - \tilde{\alpha}_i - C\tilde{\beta}_i}{\tilde{\sigma}_i} \quad (3.8)$$

De esta forma, las ecuaciones correspondientes a la corrección del efecto batch en cada conjunto se expresan de la siguiente forma:

$$\hat{X}_{ij} = \tilde{\alpha}_i + C\tilde{\beta}_i + \frac{\tilde{\sigma}_i(z_{ij} - \tilde{\gamma}_i^{X*})}{\tilde{\sigma}_i^{X*}} \quad (3.9)$$

$$\hat{\gamma}_{ij} = \tilde{\alpha}_i + C\tilde{\beta}_i + \frac{\tilde{\sigma}_i(z_{ij} - \tilde{\gamma}_i^{X*})}{\tilde{\sigma}_i^{X*}} \quad (3.10)$$

En las ecuaciones 3.9 y 3.10, los términos $\tilde{\gamma}_i^{X*}$ y $\tilde{\sigma}_i^{X*}$ representan las estimaciones del efecto batch de la ecuación 3.7 empleando a priori método empíricos o no que se muestran con mas detalle en [78].

- Normalización basada en discretización normal (NORDI). Este método trata de mejorar la discretización de la expresión de los genes y su agrupación en conjuntos para poder generar reglas de asociación. Aunque también se emplea para la supresión del efecto batch. Los pasos a seguir en este método son los siguientes:
 1. Primero se procede a la eliminación de los outliers detectados por el método de Grubbs [35] y aplicando el test de normalidad de Jarque-Bera [21] para mejorar de esa manera la normalidad de la distribución.
 2. Luego se verificará que la distribución obtenida coincide con una normal comparando las distribuciones inicial y final mediante una gráfica cuartil-cuartil y el test de normalidad de Lilliefors.
 3. Por último se aplica una metodología Z-score para calcular los umbrales de sobreexpresión e inhibición de los genes.

3.7.2 *Métodos de eliminación del efecto batch basados en discretización*

Estos métodos tratan de transformar la expresión de los genes en un conjunto de datos de categorías limitado. Por ello, habrá perdida de in-

formación, pero dependiendo del análisis mencionado anteriormente, la precisión conseguida puede ser similar o incluso en algunos casos, mayor. A continuación se nombrarán algunas de las técnicas más conocidas:

- Discretización por cuartiles(QD). Implementada por la librería VirtualArray y esta basada en intervalos de igual frecuencia de aparición. Con esto se consigue que todos los valores de expresión de los arrays sean discretizados en un número finito de intervalos o bins. Después se usa una equiprobabilidad empleando los cuartiles para seleccionar por donde cortar dichos bins. Una vez hecho esto, los dos bins centrales, considerando como punto de corte la mediana, son unidos en un único intervalo central. Posteriormente, cada valor de expresión es reemplazado por un entero que corresponde con el bin al que seria asignado.

El criterio seguido es el explicado a continuación. El valor cero se corresponde con el intervalo central, el resto de bins se van numerando secuencialmente conforme más se alejan del bin central. Si están por encima de la mediana, se numeran de forma positiva y, por el contrario, si están por debajo de la mediana, se numeran de forma negativa.

- Median Rank Score(MRS). Se asemeja a la normalización por cuantiles empleada en el algoritmo RMA. Se basa en considerar uno de los datasets como referencia, de forma que se ordenan todos sus genes siguiendo un ranking en función de su mediana del valor de expresión. Por último, se repite este proceso con el resto de datasets sustituyendo sus valores por la correspondiente mediana de referencia.
- Gene Quantiles(GQ). Es una modificación del Median Rank Score. En ella se fuerza una transformación adicional de los valores de expresión de los genes, para igualar los valores de la mediana de todos los genes en los diferentes datasets que se quieren integrar.

3.8 EVALUACIÓN DE LAS TÉCNICAS DE INTEGRACIÓN EN MICROARRAYS

La unión de datos biológicos se facilita gracias a la eliminación del efecto batch, siempre y cuando se hayan tomado dichos datos en condiciones comparables aunque sean de diferentes tecnologías. No obstante, si la técnica se ha empleado incorrectamente, puede ayudar a la aparición de falsos positivos. Es por tanto, muy importante hacer una validación de los datos una vez eliminado el efecto batch para comprobar que efectivamente son comparables. Para validar dichos resultados se pueden emplear herramientas como los boxplots, las curvas de densidad a nivel de gen y el análisis de componentes principales. Es re-

comendable usar todas las herramientas debido a que no proporcionan la misma información entre ellas.

Los boxplot y las curvas de densidad proporcionan información acerca de como afecta el efecto batch a nivel de gen. Por tanto, cabe esperar que si no existen el efecto batch en dichas muestras, los niveles de expresión de los genes de ambos datasets sean muy similares, tengan una distribución muy parecida, bajo la hipótesis de que ambos datasets tengan la misma distribución en lo a la variable biológica a estudiar se refiere. Es importante que los datasets estén equilibrados en cuanto al número de muestras con la patología a estudiar y el número de pacientes sanos, de esa forma el estudio será lo mas balanceado posible. Por último, el análisis de componentes principales brinda una visión más global de la presencia del efecto batch a nivel de estudio o muestra.

4

RESULTADOS Y ESTUDIO

ÍNDICE

4.1	Discusión y resultados de las series de Affymetrix en R	43
4.1.1	Análisis serie GSE52712	43
4.1.2	Análisis serie GSE40987	47
4.1.3	Análisis serie GSE52262	50
4.1.4	Análisis serie GSE12790	54
4.1.5	Análisis integrador de Affymetrix mediante VirtualArray	
	ray	55
4.2	Discusión y resultados de las series de Affymetrix en Matlab	60
4.2.1	Análisis serie GSE52712	60
4.2.2	Análisis serie GSE40987	63
4.2.3	Análisis serie GSE52262	66
4.2.4	Análisis serie GSE12790	69
4.2.5	Comparación entre análisis en R y Matlab de Affymetrix	72
4.3	Discusión y resultados de las series de Illumina	77
4.3.1	Análisis serie GSE46834	77
4.3.2	Análisis serie GSE68651	80
4.3.3	Análisis integrador de Illumina mediante VirtualArray	83
4.4	Discusión y resultados de las series de Illumina en Matlab	86
4.4.1	Análisis serie GSE46834	86
4.4.2	Análisis serie GSE68651	89
4.4.3	Comparación entre análisis en R y Matlab de Illumina	92
4.5	Discusión del estudio integrador de las series de Affymetrix e Illumina	95
4.5.1	Análisis de las series integradas	96
4.5.2	Comparación de técnicas de unión de las muestras	99
4.5.3	Comparación de técnicas de eliminación del efecto batch	100
4.5.4	Interpretación de los resultados del estudio integrador	102
4.6	Clasificación de los datos	105
4.6.1	Clasificación con KNN y SVM	105

En este capítulo se expondrán los resultados obtenidos en la realización de la expresión diferencial de genes a las diferentes series de datos pertenecientes a pacientes de cáncer de mama, estas series provienen de dos tecnologías diferentes de microarrays, como son Affymetrix e Illumina.

Las series usadas han sido obtenidas a través de la plataforma web de NCBI GEO y se ha seguido la metodología explicada en el capítulo 3 para llevar la cabo la realización de los experimentos. Como

se mencionó anteriormente, todos los datos analizados han sido analizados en crudo para asegurar que todos ellos han seguido el mismo preprocesamiento.

Se procederá a mostrar los resultados obtenidos para los estudios realizados con la herramienta bioconductor de R y por otro lado, con la bioinformatic toolbox de Matlab. De esta forma se le dará más robustez al estudio, pudiendo comparar los resultados de las mismas series en distintas herramientas.

A lo largo de estos resultados se seguirá la siguiente notación para el fold-change(FC):

- Si un gen contiene en el FC el signo positivo, significará que este está sobreexpresado en el grupo de paciente con respecto al grupo de las muestras de control.
- Por contra, si un gen contiene el signo negativo en el FC, estará inhibido y, por tanto, menos expresado en el grupo de pacientes que el grupo de control.

4.1 DISCUSIÓN Y RESULTADOS DE LAS SERIES DE AFFYMETRIX EN R

A lo largo de este apartado se estudiarán en profundidad cuatro series de la plataforma GEO, las cuales contienen muestras de pacientes con cáncer de mama. Todas estas series pertenecen a la tecnología Affymetrix.

El proceso para ello será el análisis de cada una de las series individualmente mediante la herramienta de bioconductor del lenguaje R y partiendo de sus datos en crudo. Posteriormente, se realizará un análisis integrador de las series de Affymetrix con el fin de encontrar DEGs destacados entre las series elegidas pertenecientes a esta tecnología.

Se mostrará para cada una de las series una tabla con los 20 genes más relevantes ordenados con respecto a su FC. Se impondrá como restricción un FC igual o mayor a 2 independientemente del signo que este tenga y un p-value de como mínimo 0.001, de forma que los resultados tengan relevancia.

Por último, añadir que todas las series seleccionadas de Affymetrix pertenecen a la tecnología "Affymetrix Human Genome U133 Plus 2.0''(GPL570) [61].

4.1.1 Análisis serie GSE52712

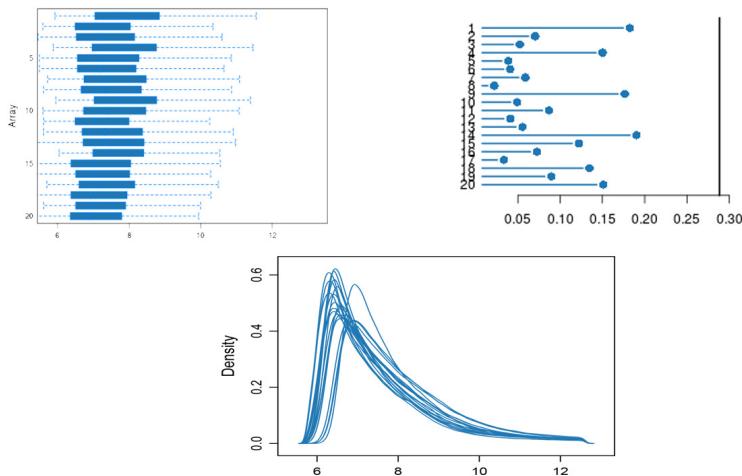
Según la información dada acerca de esta serie por la plataforma web GEO consta de 20 muestras entre ellas se pueden encontrar tanto muestras de célula de cáncer de mama como muestras de control. En concreto, las muestras se dividen en 10 muestras de control pertenecientes a células epiteliales conocidas como MCF10A [68] y 10 muestras de células cancerígenas llamadas MCF7 [5]. Este estudio proviene del Cancer Research UK Manchester Institute en Reino Unido.

De todas las muestras analizadas, se puede observar gracias a todos los tests proporcionados por la librería ArrayQualityMetrics, que no existe ningún outlier entre ellas.

Esto se puede observar en el boxplot de la distribución de intensidades del array de datos y en el estadístico de Kolmogorov-Smirnov aplicado sobre estas intensidades. Para estas intensidades se ha determinado un umbral de K igual a 0.288 y cualquier array que sobrepase este valor, podría ser considerado como un outlier. Como se puede ver en la Figura 4.1, ningún array sobrepasa dicho valor.

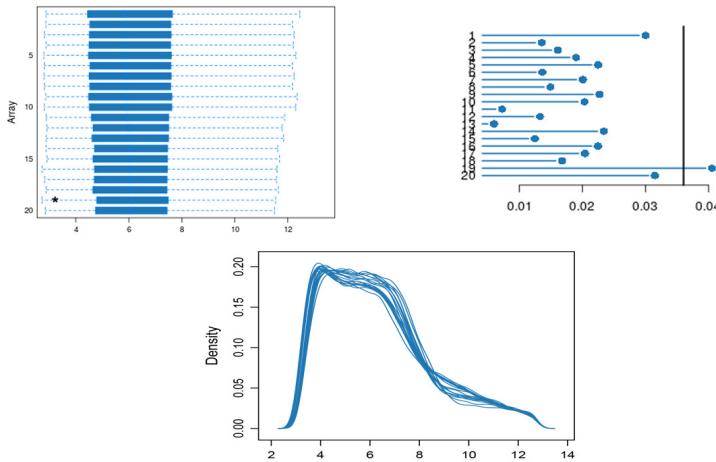
Una vez analizados los posibles outliers, se procede a normalizar los datos mediante el método RMA y a volver a realizar el análisis de

Figura 4.1
Test de Kolmogorov-Smirnov aplicado a la serie GSE52712



calidad. Se puede observar en la Figura 4.2 como los datos están correctamente normalizados y como la muestra número 19 perteneciente a una muestra de control, es considerada outlier por este test. Por lo que se procede a su eliminación y posterior repetición del análisis de calidad.

Figura 4.2
Test de Kolmogorov-Smirnov aplicado a la serie GSE52712 normalizada



A continuación se procederá a mostrar los 20 genes más destacados ordenador por el FC decreciente. Es importante añadir, que estos no han sido todos los genes detectados con esos valores de restricción y que la lista completa de estos se usará en el análisis integrador de todas las series posteriormente.

Una vez expuesta la Tabla 4.1 con los valores estadísticos de los veinte genes más relevantes para la series GSE52712, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET [25].

Como se puede observar en la Tabla 4.2, de los 20 genes destacados, 10 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 10 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido según la plataforma DisGeNET.

Símbolo	logFC	T	Pval	Adj. Pval	B
ARMCX6	8.0083	34.4239	1.77e-18	4.86e-15	30.9429
NNMT	-7.8731	-78.5239	3.55e-25	1.94e-20	39.6209
UCP2	7.7665	36.3217	6.56e-19	2.11e-15	31.6911
BMP7	-7.7069	30.3913	1.78e-14	3.49e-14	29.1277
DSCAM-AS1	7.6869	40.1116	1.03e-19	4.03e-16	33.0157
GPX1	-7.5686	-66.5348	7.97e-24	2.17e-19	38.3892
AGR2	7.4135	19.5134	5.87e-14	4.59e-11	22.0240
FXYD5	-7.4033	-30.2572	1.94e-17	3.66e-14	29.0614
CALD1	-7.3243	-33.3545	3.19e-18	8.31e-15	30.4931
NUP210L	7.1340	35.7399	8.85e-19	2.69e-15	31.4682
LAMB1	-6.9977	-26.7773	1.84e-16	2.34e-13	27.1855
SFRP1	-6.9468	-26.3650	2.45e-16	2.91e-13	26.9416
KRT6A	-6.9439	-26.6083	2.07e-16	2.57e-13	27.0861
TFF1	6.8867	36.3958	6.31e-19	2.11e-15	31.7191
ESR1	6.8757	40.8489	7.36e-20	3.35e-16	33.2500
KRT7	-6.8571	-31.9093	7.69e-18	1.83e-14	29.8036
C3orf14	6.8244	43.6038	2.17e-20	1.70e-16	34.0660
LY6K	-6.7507	-41.7111	4.98e-20	2.72e-16	33.5153
ANXA9	6.6987	20.1872	3.18e-14	2.72e-11	22.5926
IGFBP7	-6.6934	-31.2189	1.08e-17	2.28e-14	29.5278

Tabla 4.1
Veinte genes más destacados para la serie GSE52712

Tabla 4.2
Pathway de los veinte genes más destacados para la serie GSE52712

Símbolo	Pathway	Asociado a cáncer de mama
ARMCX6	-	-
NNMT	Relación con metabolismo	No
UCP2	Relación con metabolismo	No
BMP7	Organización matriz extracelular	No
DSCAM-AS1	-	-
GPX1	Relación con metabolismo	Si
AGR2	-	Si
FXYD5	-	Si
CALD1	Contracción muscular	No
NUP210L	-	-
LAMB1	Organización matriz extracelular	No
SFRP1	Transducción de señales	Si
KRT6A	-	Si
TFF1	-	Si
ESR1	Transducción de señales	Si
KRT7	-	Si
C3orf14	-	No
LY6K	-	Si
ANXA9	-	No
IGFBP7	Respuesta celular al estrés	Si

4.1.2 Análisis serie GSE40987

Como podemos observar en la plataforma GEO, esta serie contiene 10 muestras de las cuales 6 son muestras de control y 4 muestras de pacientes con cáncer de mama. La serie procede del Harvard Medical School en Boston, USA.

Como en la anterior serie, las muestras de control son tomadas de las conocidas como células epiteliales MCF10A mientras que las muestras cancerígenas provienen de las células MCF7.

De todas las muestras analizadas, se puede observar gracias a todos los tests proporcionados por la librería ArrayQualityMetrics, que no existe ningún outlier entre ellas.

Esto se puede observar en el boxplot de la distribución de intensidades del array de datos y en el estadístico de Kolmogorov-Smirnov aplicado sobre estas intensidades. Para estas intensidades se ha determinado un umbral de K igual a 0.476 y cualquier array que sobrepase este valor, podría ser considerado como un outlier. Como se puede ver en la Figura 4.3, ninguno array sobrepasa dicho valor.

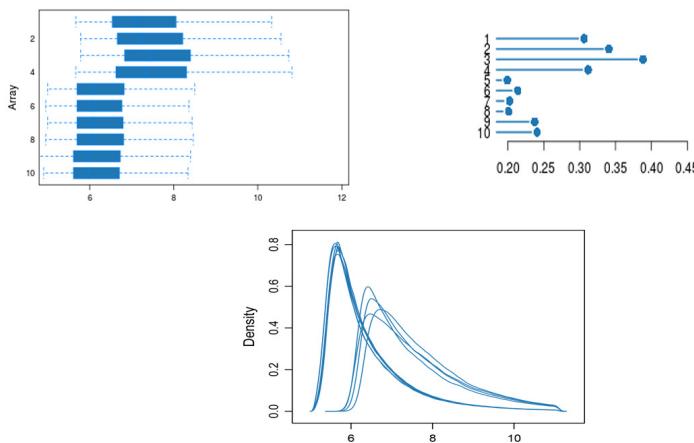


Figura 4.3
Test de Kolmogorov-Smirnov aplicado a la serie GSE40987

Al normalizar mediante RMA la serie y volver a practicar el análisis de calidad se observa como nuevamente no se detecta ningún outlier tal y como se aprecia en la Figura 4.4.

Tal y como se procedió en la anterior serie estudiada, se mostrará una tabla con los veinte primeros genes que han cumplido las restricciones impuestas al principio del capítulo.

Figura 4.4
Test de Kolmogorov-Smirnov aplicado a la serie GSE40987 normalizada

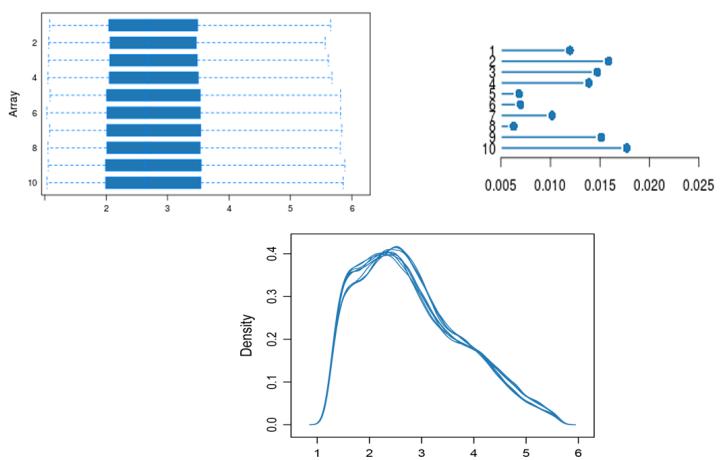


Tabla 4.3
Veinte genes más destacados para la serie GSE40987

Símbolo	logFC	T	Pval	Adj. Pval	B
DSCAM-AS1	9.2979	61.8736	2.60e-17	8.26e-14	29.6786
NNMT	-9.1323	-97.1862	7.90e-20	1.44e-15	33.9400
KRT15	-8.8632	-59.6618	4.15e-17	1.02e-13	29.2865
KRT19	8.5838	39.3792	8.48e-15	3.91e-12	24.4482
GPX1	-8.4906	-79.9397	9.73e-19	1.06e-14	32.2448
SEPP1	-8.2597	-80.0445	9.56e-19	1.06e-14	32.2570
IFI16	-8.2357	-99.4696	5.86e-20	1.4415e-15	34.1249
C3orf14	8.1650	42.6550	3.05e-15	1.87e-12	25.4219
LDBH	-7.9851	-73.6553	2.78e-18	2.17e-14	31.4640
EMP1	-7.8085	-98.5213	6.63e-20	1.44e-15	34.0490
EHF	-7.4904	-57.4707	6.71e-17	1.31e-13	28.8772
LAMB1	-7.4267	-33.1498	7.62e-14	1.75e-11	22.3008
S100A2	-7.4002	-23.1843	7.05e-12	5.05e-10	17.7054
KRT5	-7.3261	-43.6667	2.26e-15	1.66e-12	25.7043
CLCA2	-7.2450	-48.0561	6.64e-16	7.57e-13	26.8409
CALD1	-7.0823	-26.2391	1.48e-12	1.59e-10	19.3098
SLC6A14	7.0259	34.5996	4.41e-14	1.22e-11	22.8401
ESR1	7.0093	27.6312	7.69e-13	9.89e-11	19.9766
ANPEP	-6.9743	-42.8973	2.84e-15	1.78e-12	25.4903
GSTM3	6.9318	61.5156	2.80e-17	8.26e-14	29.6165

Después de mostrarse la Tabla 4.3 en la cual aparecen los valores relevantes de los primeros veinte genes expresados, se procederá a mostrar como en la serie anterior, la función y la relación o no con la enfermedad a estudiar en base a la información dada por las plataformas DisGeNET.

Símbolo	Pathway	Asociado a cáncer de mama
DSCAM-AS1	-	-
NNMT	Relación con metabolismo	No
KRT15	-	Si
KRT19	-	Si
GPX1	Respuesta celular al estrés	Si
SEPP1	-	Si
IFI16	Sistema inmunológico	No
C3orf14	-	No
LDBH	-	-
EMP1	-	Si
EHF	-	Si
LAMB1	Organización matriz extracelular	No
S100A2	-	No
KRT5	Comunicación intercelular	Si
CLCA2	Transporte de pequeñas moléculas	Si
CALD1	Contracción muscular	No
SLC6A14	Transporte de pequeñas moléculas	No
ESR1	Transducción de señales	Si
ANPEP	Metabolismo de proteínas	No
GSTM3	Relación con metabolismo	No

Tabla 4.4
Pathway de los veinte genes más destacados para la serie GSE40987

Como se puede observar en la Tabla 4.4, de los 20 genes destacados, 9 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 11 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso.

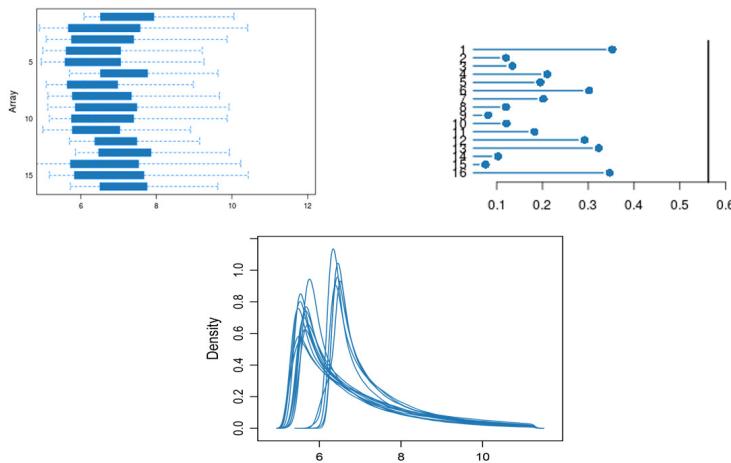
4.1.3 Análisis serie GSE52262

Como podemos observar en la plataforma GEO, esta serie contiene 27 muestras de las cuales 4 son muestras de control y 23 muestras de pacientes con cáncer de mama. Esta serie pertenece al Houston Methodist Research Institution ubicado en Houston, USA.

De todas ellas, en el estudio se han utilizado las 4 pertenecientes a muestras de control tomadas de células epiteliales MCF10A y 16 muestras de células cancerígenas, entre las que se incluyen células como MCF7 entre otras.

Como en las otras series, se ha realizado un análisis de calidad mediante la librería ArrayQualityMetrics. En el análisis se puede ver como ningún test ha detectado outliers. En concreto, se puede observar en el boxplot de intensidades de los arrays y en el gráfico creado a raíz del test de Kolmogorov-Smirnov como ningún array supera el valor K estimado en el test, en este caso, equivalente a 0.562. Esto puede verse en la Figura 4.5.

Figura 4.5
Test de Kolmogorov-Smirnov aplicado a la serie GSE52262



Una vez realizado el análisis, se ha procedido a normalizar mediante el método RMA y a realizar una segunda vez el análisis de calidad. Tal y como puede verse en la Figura 4.6, al igual que en el análisis de los datos sin preprocesar, no se ha detectado ningún outlier.

Una vez realizados los análisis, se han obtenido como en las series anteriores, los veinte genes más relevantes en orden decreciente por su FC.

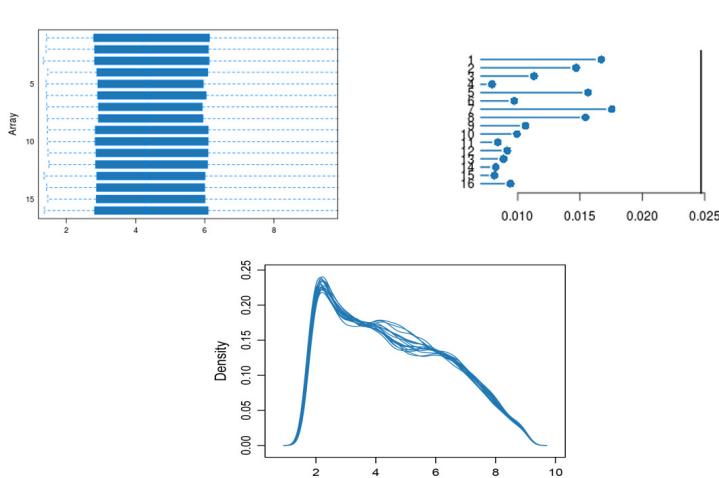


Figura 4.6
Test de Kolmogorov-Smirnov aplicado a la serie GSE52262 normalizada

Símbolo	logFC	T	Pval	Adj. Pval	B
DSG3	-7.6072	-46.372655	1.44e-18	3.94e-14	29.1556
SFRP1	-7.2472	-9.3355	6.73e-08	1.01e-05	8.4941
LTF	-7.0526	-40.1315	1.45e-17	1.58e-13	27.8161
ARMCX6	7.0150	16.5109	1.63e-11	1.07e-08	16.6391
GSTM3	6.9864	11.5546	3.32e-09	9.66e-07	11.4985
ALOX15B	-6.8593	-20.4194	6.19e-13	7.21e-10	19.6347
CHI3L1	-6.3715	-11.2830	4.68e-09	1.26e-06	11.1586
SLC6A15	-6.2592	-17.3260	7.82e-12	6.10e-09	17.3272
CPNE8	-6.2019	-32.1433	4.95e-16	1.93e-12	25.4414
PNLIPRP3	-6.0575	-13.5754	3.12e-10	1.39e-07	13.8187
ZNF655	-6.0557	-30.3888	1.20e-15	4.39e-12	24.7851
BBOX1	-6.0375	-8.4255	2.69e-07	2.94e-05	7.0975
UCP2	6.0223	10.5187	1.27e-08	2.67e-06	10.1621
SCNN1G	-5.8371	-34.8641	1.36e-16	8.27e-13	26.3530
LOX	-5.8269	-7.8662	6.62e-07	5.88e-05	6.1859
ANPEP	-5.6893	-19.1639	1.65e-12	1.64e-09	18.7517
MT1E	-5.5606	-17.0946	9.60e-12	7.00e-09	17.1356
TDRP	-5.5559	-19.8677	9.48e-13	1.05e-09	19.2551
SNCA	-5.5259	-43.7907	3.60e-18	6.56e-14	28.6452
FBN2	-5.4889	-16.6522	1.43e-11	9.67e-09	16.7611

Tabla 4.5
Veinte genes más destacados para la serie GSE52262

Una vez expuesta la Tabla 4.5 con los valores estadísticos de los veinte genes más relevantes para la serie GSE52262, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Tabla 4.6
Pathway de los veinte genes más destacados para la serie GSE52262

Símbolo	Pathway	Asociado a cáncer de mama
DSG3	Muerte celular programada	No
SFRP1	Transducción de señales	Si
LTF	Sistema inmunológico	No
ARMCX6	-	-
GSTM3	Relación con metabolismo	No
ALOX15B	Relación con metabolismo	Si
CHI3L1	-	No
SLC6A15	Transporte de pequeñas moléculas	No
CPNE8	-	No
PNLIPRP3	-	-
ZNF655	-	-
BBOX1	Relación con metabolismo	No
UCP2	Relación con metabolismo	No
SCNN1G	Transporte de pequeñas moléculas	No
LOX	Organización de matriz extracelular	No
ANPEP	Metabolismo de proteínas	No
MT1E	Relación con metabolismo	Si
TDRP	-	No
SNCA	Metabolismo de proteínas	No
FBN2	Organización de matriz extracelular	No

Como se puede observar en la Tabla 4.6, de los 20 genes destacados, 3 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 17 genes restantes, en principio están relacionados

con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso.

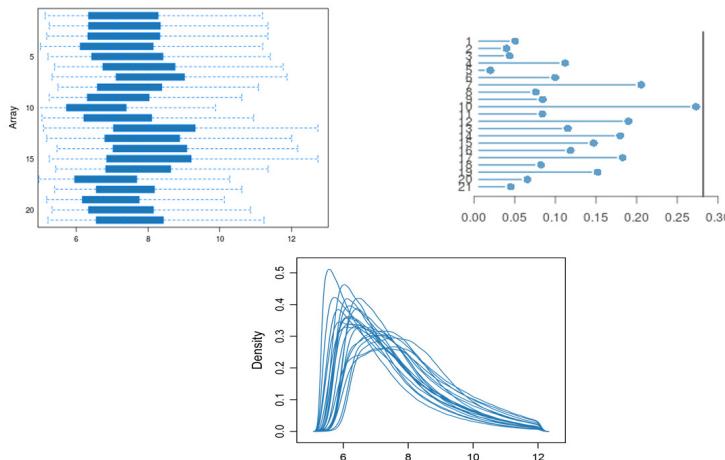
4.1.4 Análisis serie GSE12790

La última serie analizada perteneciente a la tecnología Affymetrix ha sido la serie GSE12790 con un total de 98 muestras. De todas esas muestras se han usado un total de 21 de ellas para este estudio. 11 muestras de control pertenecientes a células epiteliales MCF10A y 10 muestras de células cancerígenas.

Las muestras han sido tomadas en el Genentech de South San Francisco en USA, según muestra la ficha de la serie en la plataforma GEO.

Se ha realizado el análisis de calidad mediante ArrayQualityMetrics y este no detectó ningún outlier como puede verse en la Figura. En la Figura 4.7 se puede observar que ningún array supera el valor umbral del estadístico K calculado por el test de Kolmogorov-Smirnov cuyo valor es igual a 0.282.

Figura 4.7
Test de Kolmogorov-Smirnov aplicado a la serie GSE12790



Se ha procedido a la normalización de las muestras restantes y ha realizar de nuevo el análisis de calidad, el cual ha detectado como outlier la muestra número 7, como se puede ver en la Figura 4.8.

A continuación, se mostrará una tabla con los veinte genes más relevantes ordenados en orden decreciente en base a su FC.

Una vez expuesta la Tabla 4.7 con los valores estadísticos de los veinte genes más relevantes para la series GSE12790, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Como se puede observar en la Tabla 4.8, de los 20 genes destacados, 6 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 14 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso.

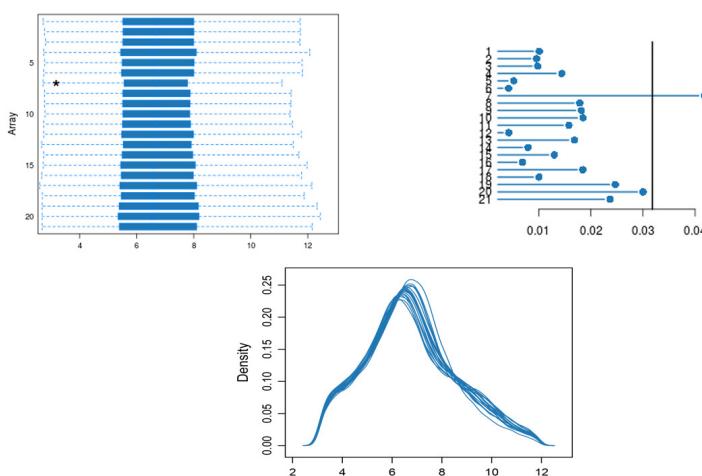


Figura 4.8
Test de Kolmogorov-Smirnov aplicado a la serie GSE12790 normalizada

Símbolo	logFC	T	Pval	Adj. Pval	B
KRT19	7.4922	12.0098	1.07e-10	7.46e-08	14.6543
KRT14	-6.8408	-6.3704	2.98e-06	1.38e-04	4.6293
KRT6A	-6.5911	-9.5574	5.73e-09	1.56e-06	10.8090
IGFBP5	6.5330	7.9629	1.11e-07	1.32e-05	7.8946
CSTA	-6.3877	-6.8740	1.01e-06	6.45e-05	5.6997
DST	-6.1134	-12.8517	3.16e-11	3.53e-08	15.8185
DSG3	-5.9689	-8.7291	2.56e-08	4.75e-06	9.3388
KCTD12	-5.9581	-8.8435	2.07e-08	4.18e-06	9.5473
VGLL3	-5.8690	-8.5254	3.76e-08	6.38e-06	8.9628
S100A2	-5.7588	-8.8289	2.13e-08	4.24e-06	9.5200
IL18	-5.7223	-9.4717	6.67e-09	1.77e-06	10.6612
BNC1	-5.6878	-14.0258	6.35e-12	1.01e-08	17.3264
KRT5	-5.6754	-7.8199	1.47e-07	1.63e-05	7.6158
SGPP2	-5.6433	-14.8062	2.32e-12	4.70e-09	18.2602
ATP8B1	5.5815	17.7803	7.24e-14	2.82e-10	21.3890
C3orf14	5.5324	11.6133	1.96e-10	1.16e-07	14.0801
CLCA2	-5.4724	-11.3474	2.96e-10	1.50e-07	13.6853
NNMT	-5.4646	-6.6582	1.60e-06	8.98e-05	5.2451
GSTM3	5.3609	7.9532	1.13e-07	1.33e-05	7.8758
PLD5	-5.3560	-22.5051	7.49e-16	5.12e-12	25.2606

Tabla 4.7
Veinte genes más destacados para la serie GSE12790

4.1.5 Análisis integrador de Affymetrix mediante VirtualArray

Para poder representar los resultados anteriores de manera conjunta se han usado diagramas de Venn, de esta forma se pueden representar de forma simultanea las intersecciones de los genes en común que tienen

Tabla 4.8
Función de los veinte genes más destacados para la serie GSE12790

Símbolo	Pathway	Asociado a cáncer de mama
KRT19	-	Si
KRT14	Comunicación intercelular	No
KRT6A	-	Si
IGFBP5	Metabolismo de proteínas	Si
CSTA	-	Si
DST	Comunicación intercelular	Si
DSG3	Muerte celular programada	No
KCTD12	-	No
VGLL3	-	No
S100A2	-	No
IL18	Sistema inmunológico	No
BNC1	-	No
KRT5	Comunicación intercelular	No
SGPP2	Relación con metabolismo	No
ATP8B1	Transporte de pequeñas moléculas	No
C3orf14	-	No
CLCA2	Transporte de pequeñas moléculas	Si
NNMT	Relación con metabolismo	No
GSTM3	Relación con metabolismo	No
PLD5	-	No

las cuatro series de affymetrix, añadiendo además, la serie creada integrando estas cuatro series en una usando la herramienta virtualArray.

Con estos diagramas se pretende encontrar genes mas robustos y significativos al estar expresados en las distintas series y muestras, como se observa en las Figuras 4.9 y 4.10 no son muchos los genes comunes encontrados, por lo que los genes encontrados son poco robustos. Este hecho justifica el enfoque dado a este estudio, donde se han integrado diferentes series y muestras de diferentes tecnologías y lugares con el propósito de que los genes que se encuentren sean mucho más robustos y potenciales biomarcadores del cáncer de mama.

Se han realizado dos diagramas de Venn, uno usando los veinte genes relevantes detallados para cada serie en el apartado anterior. Como se

muestra en la Figura 4.9, con este número de genes no se han encontrado genes relevantes comunes al estudio integrador y a las cuatro series a la vez. Por otro lado, si se puede ver como se han detectado un gen común a todas las series(incluido el análisis integrador) menos a la serie GSE52262. El gen común a todas menos a la mencionada en concreto es el gen "NNMT", el cual en principio no estaría relacionado con el cáncer de mama según la plataforma DisGeNET.

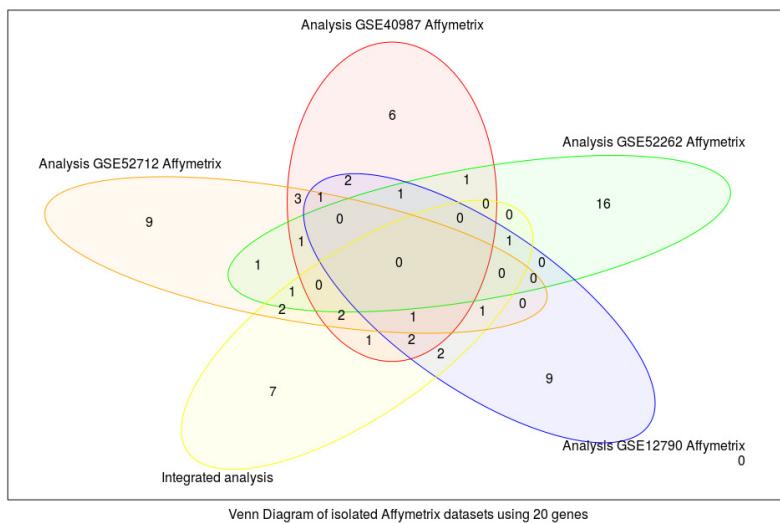


Figura 4.9
Diagrama de Venn de la integración de affymetrix usando 20 genes

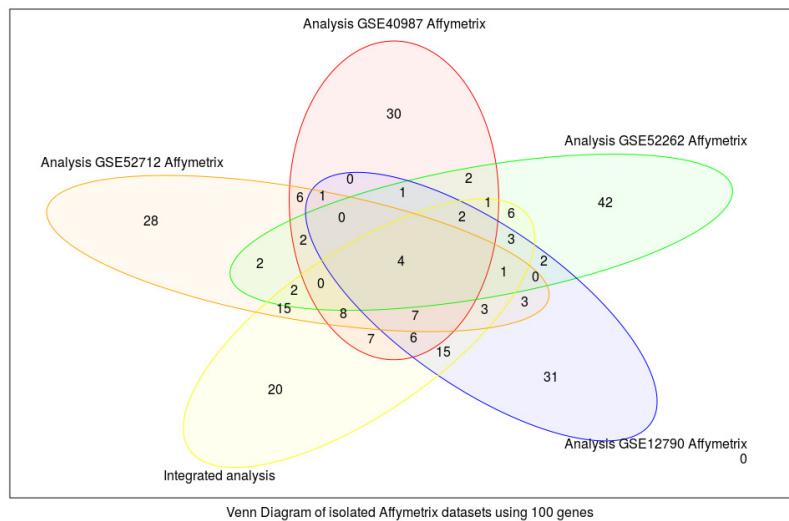
El diagrama de Venn restante, incluye un total de 100 genes destacados por cada serie sin cambiar las restricciones de estas. Tal y como se puede ver en la Figura 4.10, en este caso se han detectado 4 genes robustos en común a estas series. Estos genes detectados han sido los siguientes:

- SFRP1: Según DisGeNET, está relacionado con el cáncer de mama.
- GSTM3: Según DisGeNET, no está relacionado con el cáncer de mama.
- CLMP: Según DisGeNET, no está relacionado con el cáncer de mama.
- SULT1E1: Según DisGeNET, está relacionado con el cáncer de mama.

Como se puede observar en la lista anterior, de los cuatro genes robustos encontrados en las series de Affymetrix, dos de ellos tienen relación con el cáncer de mama según la plataforma DisGeNET, mientras que los otros dos en principio estarían asociados a otras enfermedades.

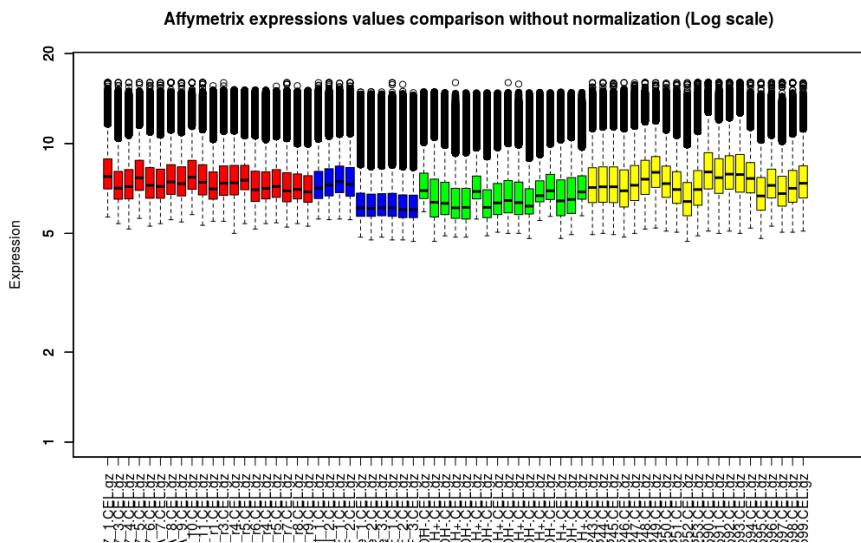
En el boxplot en escala logarítmica representado en la Figura 4.11 se puede observar como la variabilidad de los rangos dinámicos de la ex-

Figura 4.10
Diagrama de Venn
de la integración de
affymetrix usando
100 genes



presión es elevada, en algunos casos no solo entre serie y serie, sino dentro de una misma serie también. Con el fin de darle una solución a esa variabilidad se ha procedido a normalizar cada una de las series por separado usando el método RMA y el resultado muestra una gran reducción de la variabilidad de las muestras tanto entre serie como internamente a cada una de ellas. Esto queda representado en la Figura 4.11

Figura 4.11
Valores de expresión
sin normalizar de las
series de Affymetrix
por separado



Por último, se han representado las mismas series pero esta vez unidas y normalizadas mediante la herramienta virtualArray en una sola serie. En el boxplot que muestra la Figura 4.13 se puede ver como prácticamente la variabilidad ha quedado eliminada para poder así hacer un

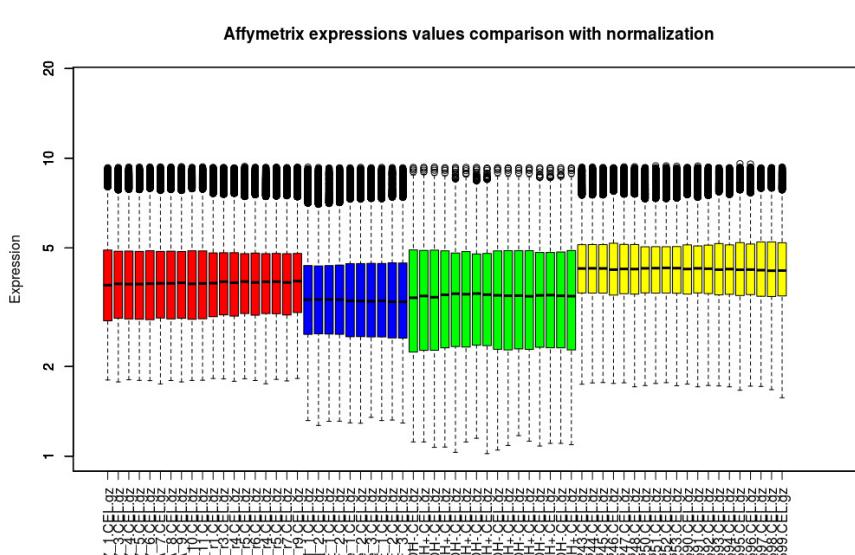


Figura 4.12
Valores de expresión
normalizados de las
series de Affymetrix
por separado

estudio en iguales condiciones entre todas las muestras integradas, lo-
grando así más robustez y fiabilidad en los resultados.

En los boxplots representados, el eje vertical representa el número de bits empleados, también conocido como profundidad de bits el cual es equivalente a 16 para todas las series usadas en este estudio. Si alguna serie tuviese una profundidad de bits distinta al resto, debería igualarse dicha profundidad, puesto que el ruido de cuantización podría afectar al análisis diferencial.

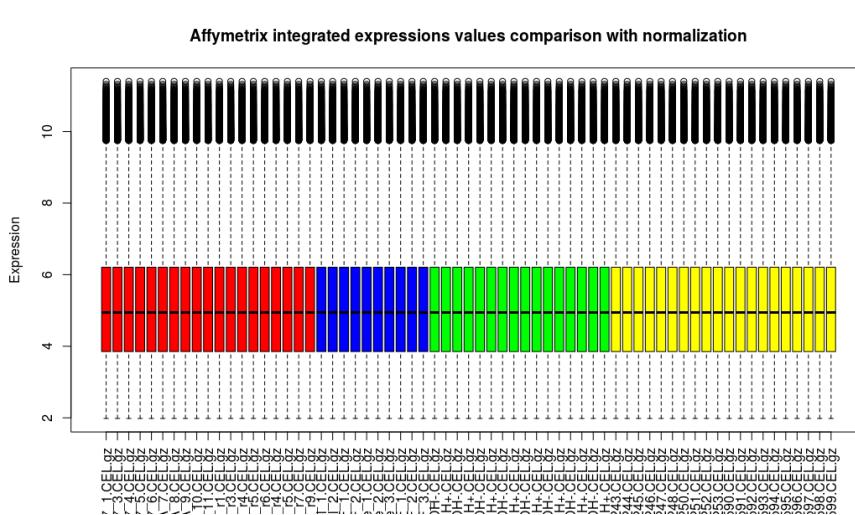


Figura 4.13
Valores de expresión
normalizados de las
series de Affymetrix
integradas

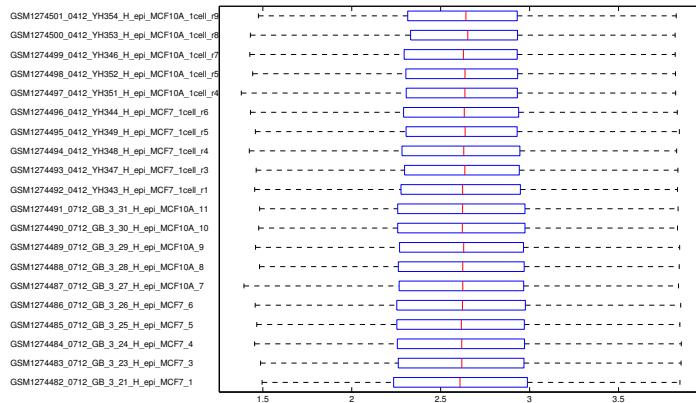
4.2 DISCUSIÓN Y RESULTADOS DE LAS SERIES DE AFFYMETRIX EN MATLAB

Después de realizarse el estudio de las series de Affymetrix usando las librerías de bioconductor para R, se ha procedido a realizar un estudio paralelo esta vez usando la Bioinformatic Toolbox [54] de Matlab. Con dicho estudio se pretende comprobar como varían los resultados en función de la herramienta o la librería utilizada e intentar encontrar genes en ambos estudios en común.

4.2.1 Análisis serie GSE52712

Tal como se hizo en el análisis en R, se ha construido una tabla de los 20 genes más relevantes y sus valores estadísticos según la toolbox de Matlab. También se puede ver en la Figura 4.14, los valores ya normalizados para esta serie a través del proceso RMA de la toolbox.

Figura 4.14
Valores de expresión
normalizados de la
serie GSE52712 de
Affymetrix



A continuación, se mostrará una Tabla 4.9 con los veinte genes más relevantes ordenados en orden decreciente en base a su p-value.

Una vez expuesta la Tabla 4.9 con los valores estadísticos de los veinte genes más relevantes para la serie GSE12790, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Como se puede observar en la Tabla 4.10, de los 20 genes destacados, 9 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 11 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido según la plataforma DisGeNET.

Símbolo	t-scores	PVal	FDR
TFF3	-43.011	1.1145e-18	4.0748e-14
LY6K	40.632	2.8782e-18	5.2615e-14
DSCAM-AS1	-48.24	3.5534e-18	4.3306e-14
TFF1	-39.002	2.6482e-17	2.4206e-13
C3orf14	-46.216	5.297e-17	3.8734e-13
BMP7	-32.248	1.5727e-16	9.5834e-13
UCP2	-36.649	2.25e-16	1.1752e-12
TSPYL5	-30.788	2.4023e-16	1.0979e-12
NNMT	52.573	2.408e-16	9.7823e-13
MT1E	34.083	2.5049e-16	9.1584e-13
HENMT1	-32.917	4.8526e-16	1.6129e-12
CPVL	29.128	6.6552e-16	2.0277e-12
GSTT2	29.647	6.7913e-16	1.91e-12
ESR1	-39.3	7.3635e-16	1.923e-12
IGFBP7	28.608	8.1332e-16	1.9824e-12
PRKCDBP	43.266	1.8588e-15	3.9978e-12
RAC2	26.943	2.5501e-15	4.9071e-12
KRT19	-26.679	3.7533e-15	6.5347e-12
LAMB1	27.288	3.8492e-15	6.3969e-12
GATA3	-26.784	3.8984e-15	6.1971e-12

Tabla 4.9
Veinte genes más destacados para la serie GSE52712 según Matlab

Tabla 4.10
Función de los veinte genes más destacados para la serie GSE52712 según Matlab

Símbolo	Pathway	Asociado a cáncer de mama
TFF3	-	No
LY6K	-	Si
DSCAM-AS1	-	-
TFF1	-	Si
C3orf14	-	No
BMP7	Organización de matriz extracelular	No
UCP2	Relación con metabolismo	No
TSPYL5	-	Si
NNMT	Relación con metabolismo	No
MT1E	Relación con metabolismo	Si
HENMT1	Expresión genética	No
CPVL	-	No
GSTT2	Relación con metabolismo	No
ESR1	Transducción de señales	Si
IGFBP7	Respuesta celular al estrés	Si
PRKCDBP	-	Si
RAC2	Transducción de señales	No
KRT19	-	Si
LAMB1	Organización de matriz extracelular	No
GATA3	Hemostasis	Si

4.2.2 Análisis serie GSE40987

Tal como se hizo en el análisis en R, se ha construido una tabla de los 20 genes más relevantes y sus valores estadísticos según la toolbox de Matlab. También se puede ver en la Figura 4.15, los valores ya normalizados para esta serie a través del proceso RMA de la toolbox.

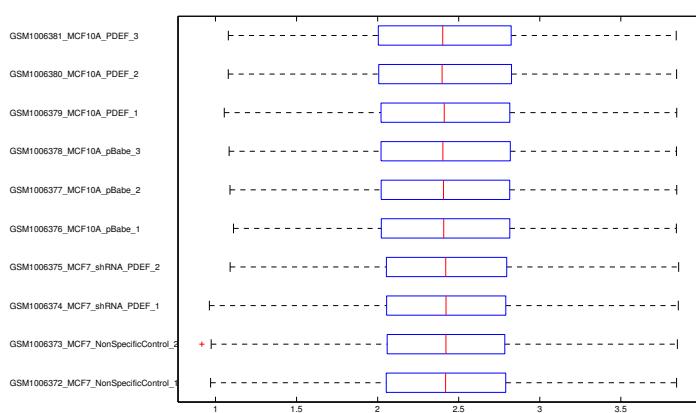


Figura 4.15
Valores de expresión normalizados de la serie GSE40987 de Affymetrix

A continuación, se mostrará una Tabla 4.11 con los veinte genes más relevantes ordenados en orden decreciente en base a su p-value.

Una vez expuesta la Tabla 4.11 con los valores estadísticos de los veinte genes más relevantes para la serie GSE40987, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Como se puede observar en la Tabla 4.12, de los 20 genes destacados, 8 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 12 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido según la plataforma DisGeNET.

Tabla 4.11
Veinte genes más destacados para la serie GSE40987 según Matlab

Símbolo	t-scores	PVal	FDR
NNMT	123.43	2.3277e-14	2.0179e-10
PLAT	102.35	9.3232e-14	4.0412e-10
HADHB	93.607	1.9767e-13	5.7122e-10
WBP5	85.52	6.4263e-13	1.3928e-09
HENMT1	-80.657	8.2827e-13	1.4361e-09
IFI16	79.371	1.9824e-12	2.8643e-09
DMWD	-71.952	2.5625e-12	3.1735e-09
APPBP2	-74.907	6.799e-12	7.3677e-09
AGPS	58.499	8.1001e-12	7.8023e-09
OXCT1	59.976	9.6473e-12	8.3634e-09
HS6ST2	57.385	9.8217e-12	7.7405e-09
DSG2	65.24	1.2784e-11	9.2358e-09
MPZL1	54.407	1.4708e-11	9.1073e-09
BCL11A	57.671	1.7962e-11	1.0381e-08
TM4SF1	65.34	1.9106e-11	1.0352e-08
BCAS3	-79.485	1.9125e-11	9.7527e-09
SFRP1	59.702	2.0336e-11	9.7943e-09
TMEM30B	-54.994	2.0349e-11	9.2844e-09
SREK1IP1	74.241	2.0385e-11	8.8362e-09
PTPN14	124	2.0974e-11	8.6582e-09

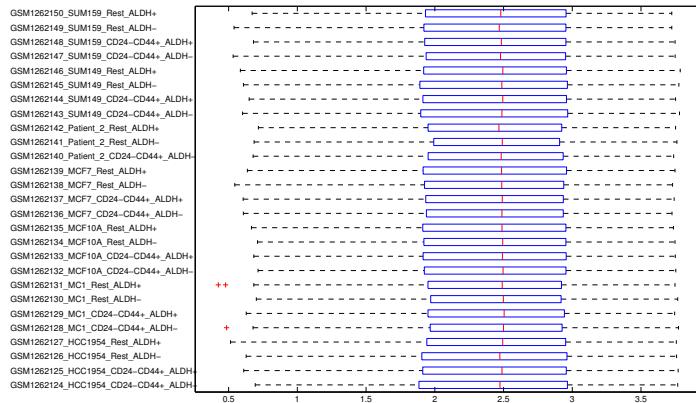
Símbolo	Pathway	Asociado a cáncer de mama
NNMT	Relación con metabolismo	No
PLAT	Hemostasis	No
HADHB	Relación con metabolismo	Si
WBP5	-	No
HENMT1	Expresión de genes	No
IFI16	Sistema inmunológico	No
DMWD	-	No
APPBP2	-	Si
AGPS	Relación con metabolismo	No
OXCT1	Relación con metabolismo	No
HS6ST2	Relación con metabolismo	Si
DSG2	Muerte celular programada	No
MPZL1	-	Si
BCL11A	-	No
TM4SF1	-	Si
BCAS3	-	Si
SFRP1	Transducción de señales	Si
TMEM30B	-	No
TMEM30B	-	No
PTPN14	-	Si

Tabla 4.12
Función de los veinte genes más destacados para la serie GSE40987 según Matlab

4.2.3 Análisis serie GSE52262

Tal como se hizo en el análisis en R, se ha construido una tabla de los 20 genes más relevantes y sus valores estadísticos según la toolbox de Matlab. También se puede ver en la Figura 4.16, los valores ya normalizados para esta serie a través del proceso RMA de la toolbox.

Figura 4.16
Valores de expresión
normalizados de la
serie GSE52262 de
Affymetrix



A continuación, se mostrará una Tabla 4.13 con los veinte genes más relevantes ordenados en orden decreciente en base a su p-value.

Una vez expuesta la Tabla 4.13 con los valores estadísticos de los veinte genes más relevantes para la series GSE52262, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Como se puede observar en la Tabla 4.14, de los 20 genes destacados, 7 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 13 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido según la plataforma DisGeNET.

Símbolo	t-scores	PVal	FDR
DSG3	82.065	4.7408e-20	8.9227e-16
TMEM246	46.133	1.3455e-16	8.4413e-13
TDRP	46.593	2.4485e-16	1.1521e-12
HOXB7	-36.435	3.4623e-15	1.3033e-11
OSBPL5	31.324	3.0546e-14	8.2127e-11
PDE4D	26.592	2.3039e-13	4.3362e-10
CASP1	25.253	4.4663e-13	7.6417e-10
SFRP1	25.08	4.9724e-13	7.7988e-10
CPNE8	31.587	7.0439e-13	1.0198e-09
FKBP5	26.744	1.5239e-12	2.0486e-09
NOTCH1	22.276	3.1388e-12	3.6921e-09
COL8A1	20.492	1.1549e-11	1.144e-08
RBMS3	19.936	1.2076e-11	1.1364e-08
SLC2A10	-22.004	2.8903e-11	2.5904e-08
ZYG11A	-18.916	3.9704e-11	3.3966e-08
NEDD4L	20.809	4.2758e-11	3.4988e-08
OPTN	19.274	4.4621e-11	3.4991e-08
ARMCX6	-18.98	6.1065e-11	4.4204e-08
PER3	-17.718	7.0299e-11	4.9003e-08
QPRT	-17.114	8.8097e-11	5.9216e-08

Tabla 4.13
Veinte genes más destacados para la serie GSE52262 según Matlab

Tabla 4.14
Función de los veinte genes más destacados para la serie GSE52262 según Matlab

Símbolo	Pathway	Asociado a cáncer de mama
DSG3	Muerte celular programada	No
TMEM246	-	-
TDRP	-	No
HOXB7	-	Si
OSBPL5	-	No
PDE4D	Transducción de señales	Si
CASP1	Sistema inmunológico	No
SFRP1	Transducción de señales	Si
CPNE8	-	No
FKBP5	-	No
NOTCH1	Expresión genética	Si
COL8A1	Organización de matriz extracelular	No
RBMS3	-	Si
SLC2A10	Transporte de pequeñas moléculas	Si
ZYG11A	-	-
NEDD4L	Expresión genética	No
OPTN	Ciclo celular	No
ARMCX6	-	-
PER3	-	Si
QPRT	Relación con metabolismo	No

4.2.4 Análisis serie GSE12790

Tal como se hizo en el análisis en R, se ha construido una tabla de los 20 genes más relevantes y sus valores estadísticos según la toolbox de Matlab. También se puede ver en la Figura 4.17, los valores ya normalizados para esta serie a través del proceso RMA de la toolbox.

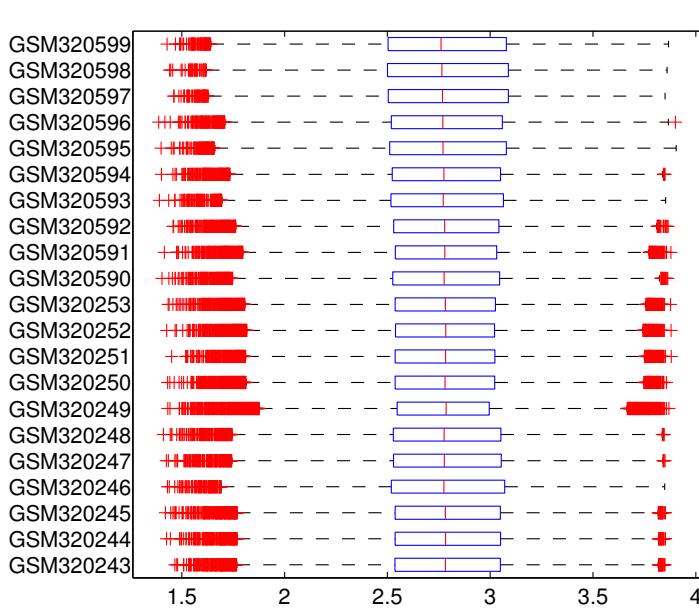


Figura 4.17
Valores de expresión normalizados de la serie GSE12790 de Affymetrix

A continuación, se mostrará una Tabla 4.15 con los veinte genes más relevantes ordenados en orden decreciente en base a su p-value.

Una vez expuesta la Tabla 4.15 con los valores estadísticos de los veinte genes más relevantes para la series GSE12790, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Como se puede observar en la Tabla 4.16, de los 20 genes destacados, 2 de ellos ya están relacionados con el cáncer de mama segúin la plataforma DisGeNET. Los 18 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido segúin la plataforma DisGeNET.

Tabla 4.15
Veinte genes más destacados para la serie GSE12790 según Matlab

Símbolo	t-scores	PVal	FDR
COL4A6	9.1626	3.4213e-08	0.00036891
ROS1	9.5353	7.3436e-08	0.0005279
ABCA8	8.5285	1.0045e-07	0.00043327
MIS18A	-8.3852	1.2694e-07	0.00039109
INPP5D	8.2689	1.6131e-07	0.00043483
PLD5	9.2666	1.7301e-07	0.00041455
PNLIPRP3	8.5136	1.8422e-07	0.00039727
PCDH9	8.1241	2.0446e-07	0.00040084
ZBTB16	8.33	2.0491e-07	0.00036825
SLC9A9	8.2996	2.4102e-07	0.00039983
SULT1E1	8.0427	2.4162e-07	0.00037219
FLI1	8.0043	2.4381e-07	0.00035053
MMP28	8.0469	2.5132e-07	0.00033874
C5orf46	8.1489	2.655e-07	0.0003368
NT5DC1	8.7315	2.7118e-07	0.00032489
SDPR	8.5688	3.2082e-07	0.00036414
FAM65B	8.3815	3.3869e-07	0.00034781
TRNP1	7.8792	3.5912e-07	0.00035203
PHACTR3	9.0661	3.7729e-07	0.00033902
TPT1P8	8.9228	3.9334e-07	0.0003393

Símbolo	Pathway	Asociado a cáncer de mama
COL4A6	Organización de matriz extracelular	No
ROS1	-	No
ABCA8	Transporte de pequeñas moléculas	No
MIS18A	Ciclo celular	No
INPP5D	Sistema inmunológico	No
PLD5	-	No
PNLIPRP3	-	-
PCDH9	-	No
ZBTB16	Sistema inmunológico	No
SLC9A9	Transporte de pequeñas moléculas	No
SULT1E1	Relación con metabolismo	Si
FLI1	-	No
MMP28	-	No
C5orf46	-	-
NT5DC1	-	No
SDPR	-	Si
FAM65B	-	No
TRNP1	-	No
PHACTR3	-	No
TPT1P8	-	-

Tabla 4.16
Función de los veinte genes más destacados para la serie GSE12790 según Matlab

4.2.5 Comparación entre análisis en R y Matlab de Affymetrix

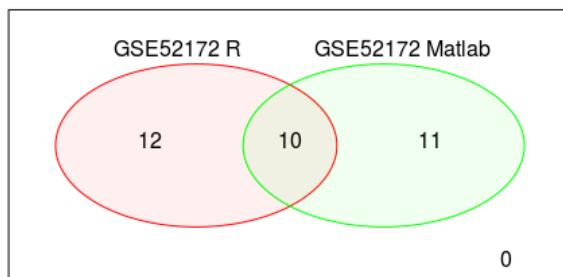
Una vez expuestos los resultados realizados tanto con R como con Matlab, se han realizado dos comparaciones para cada series. Una de ellas con un diagrama de venn con los veinte genes más relevantes según R y Matlab y otro con los cien genes más relevantes.

Primero se procede ha exponer las posibles coincidencias entre los veinte genes más relevantes para R frente a los veinte genes más relevantes para Matlab en la serie GSE52712.

Como se puede observar en la Figura 4.18, de los veinte genes más destacados para esta serie según R y Matlab, diez de ellos son comunes a ambos resultados. Esto revela una buena robustez en dichos genes que siguen expresándose a pesar de variar la herramienta usada para su estudio.

Figura 4.18
Comparación de los
20 genes más rel-
evantes entre R y
Matlab de la serie
GSE52712

Venn diagram of R & Matlab genes comparison of GSE52712



En la Figura 4.19, de los cien genes más destacados para esta serie según R y Matlab, se puede observar como comparten 43 genes. Esto, como ya paso en el diagrama anterior, revela una buena robustez en dichos genes.

Después de analizar los diagramas para la serie GSE52712, se analizará los diagramas comparativos de la serie GSE40987. Se mostrará a continuación el diagrama de venn de los veinte primeros genes expresados, esto puede verse en la Figura 4.20. De estos veinte genes, solo 3 de ellos son comunes a ambos estudios.

El siguiente diagrama mostrado en la Figura 4.21, solo dieciocho de los cien genes aparecen como comunes.

La tercera serie que se expondrá será la GSE52262. En la Figura 4.22 se puede ver que de los veinte genes expresados por cada una de las dos

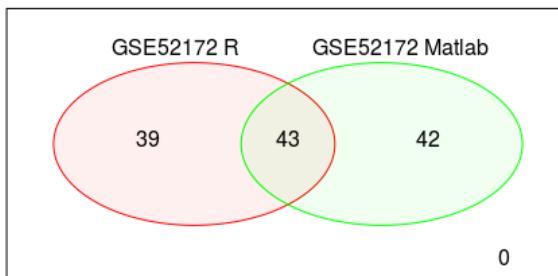
Venn diagram of R & Matlab genes comparison of GSE52712

Figura 4.19
Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE52712

herramientas para esta serie tan solo cinco genes están en común entre ambos estudios.

Analizando los cien genes en común en la Figura 4.23 se puede observar como tienen 27 genes en común.

La serie GSE12790 es la última de Affymetrix, se puede ver en la Figura 4.24 que de los veinte posibles genes en común, solo uno aparece como tal. Esto puede llevar a pensar que los genes encontrados no son muy robustos, al menos en esta serie, al tener resultados muy diferentes dependiendo de la herramienta con la que se ha analizado.

Por último, tal y como se puede ver en la Figura 4.25, de los cien genes, se encuentran dieciséis en común a ambos estudios.

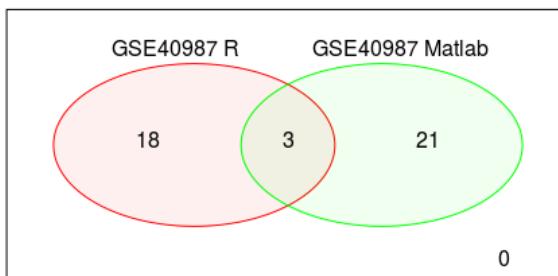
Venn diagram of R & Matlab genes comparison of GSE40987

Figura 4.20
Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE40987

Se ha comprobado en este apartado que dependiendo de la serie analizada se encuentra más o menos genes robustos común a ambas herramientas. La serie GSE52712 ha tenido un gran número de genes en común en ambos casos, mientras que las demás series en algunos casos obtienen un buen número de genes en común y en otros apenas se obtienen.

Figura 4.21
Comparación de los
100 genes más rel-
evantes entre R y
Matlab de la serie
GSE40987

Venn diagram of R & Matlab genes comparison of GSE40987

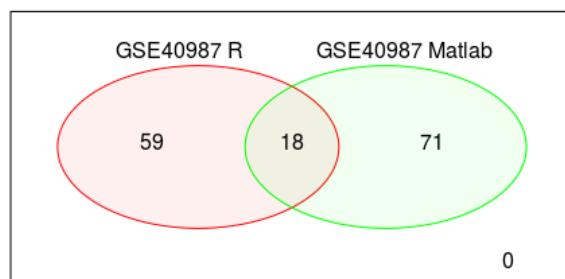
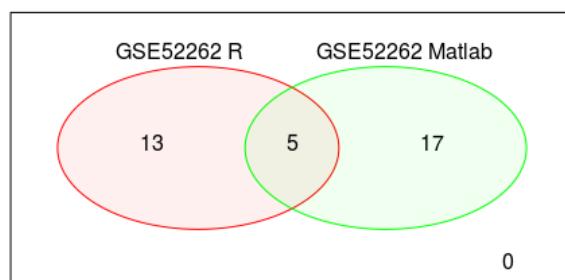


Figura 4.22
Comparación de los
20 genes más rel-
evantes entre R y
Matlab de la serie
GSE52262

Venn diagram of R & Matlab genes comparison of GSE52262



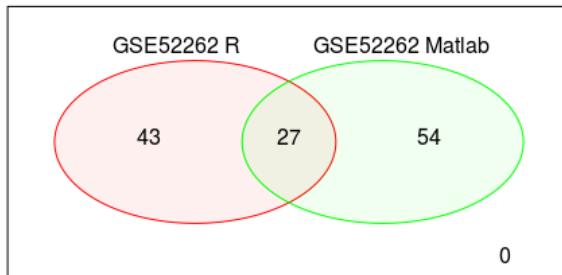
Venn diagram of R & Matlab genes comparison of GSE52262

Figura 4.23
Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE52262

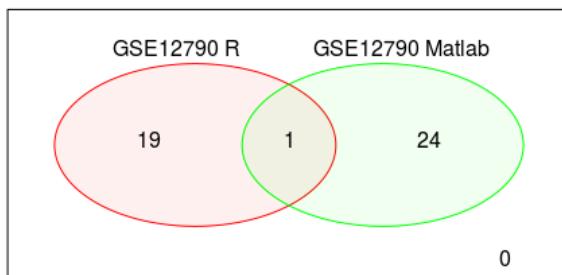
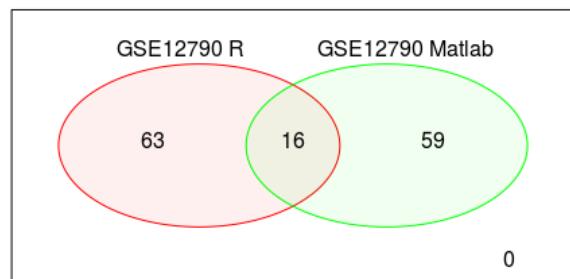
Venn diagram of R & Matlab genes comparison of GSE12790

Figura 4.24
Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE12790

Figura 4.25
Comparación de los
100 genes más rel-
evantes entre R y
Matlab de la serie
GSE12790

Venn diagram of R & Matlab genes comparison of GSE12790



4.3 DISCUSIÓN Y RESULTADOS DE LAS SERIES DE ILLUMINA

Una vez explicados los resultados obtenidos para las series de la tecnología perteneciente a Affymetrix, se pasará a exponer los resultados de las dos series elegidas para el estudio usando la tecnología de Illumina.

Ambas series han sido elegidas a través de la plataforma GEO y a diferencia de las series de Affymetrix, no se han usado los datos en crudo, sino a través de los ficheros matriz que proporciona GEO al ser la mejor manera de trabajar con los microarray de Illumina. Una vez cargados los datos de estas series, se han comprobado si estaban o no normalizados, en caso de no estarlo se ha procedido a su normalización mediante el método lumiExpresso de la librería lumi.

Tal y como se detalló en el apartado de Affymetrix, se mostrará para cada una de las series una tabla con los 20 genes más relevantes ordenados con respecto a su FC. Se impondrá como restricción un FC igual o mayor a 2 independientemente del signo que este tenga y un p-value de como mínimo 0.001, de forma que los resultados tengan relevancia.

Por último, añadir que una de las series de Illumina pertenece en concreto a la tecnología "Illumina HumanRef-8 v3.0"(GPL6883) [62] mientras que la serie restante pertenece a la tecnología "Illumina HumanHT-12 V4.0"(GPL10558) [60].

4.3.1 Análisis serie GSE46834

Como podemos observar en la plataforma GEO, esta serie contiene 8 muestras de las cuales 3 son muestras de control perteneciente a células epiteliales llamadas en concreto, HMLE y 5 muestras pertenecientes a diferentes células cancerígenas implicadas en el cáncer de mama. Esta serie pertenece al Mount Sinai Medical Center ubicado en New York, USA.

Como en las otras series, se ha realizado un análisis de calidad mediante la librería ArrayQualityMetrics. En el análisis se puede ver como ningún test ha detectado outliers. En concreto, se puede observar en el boxplot de intensidades de los arrays y en el gráfico creado a raíz del test de Kolmogorov-Smirnov como ningún array supera el valor K estimado en el test, en este caso, equivalente a 0.00361. Esto se puede ver más detallado en la Figura 4.26.

Una vez realizados los análisis, se han obtenido como en las series anteriores, los veinte genes más relevantes en orden decreciente por su FC.

Figura 4.26
Test de Kolmogorov-Smirnov aplicado a la serie GSE46834

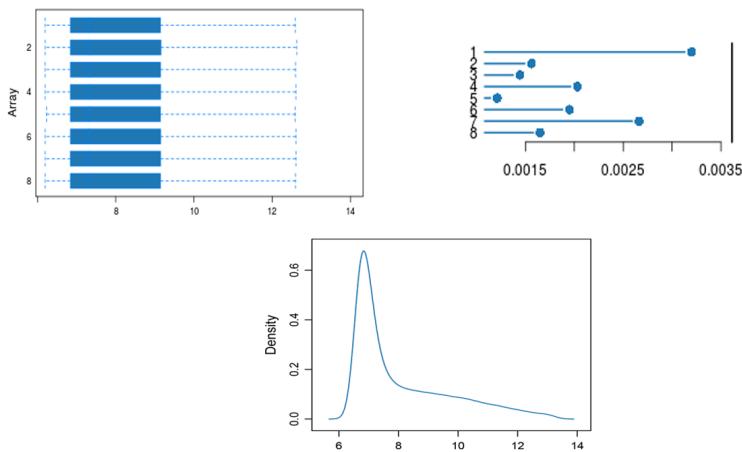


Tabla 4.17
Veinte genes más destacados para la serie GSE46834

Símbolo	logFC	T	Pval	Adj. Pval	B
KRT6C	-7.9270	-22.319	2.80e-08	6.88e-05	8.9815
KRT6B	-7.7794	-13.3937	1.30e-06	7.53e-04	6.0262
KRT6A	-7.5369	-14.0425	9.19e-07	6.27e-04	6.3325
KRT14	-7.1333	-29.4105	3.42e-09	1.68e-05	10.1750
KRT5	-6.8643	-32.9253	1.44e-09	1.25e-05	10.5726
FGFBP1	-6.7424	-31.4555	2.05e-09	1.25e-05	10.4180
KLK5	-6.3648	-13.8358	1.02e-06	6.29e-04	6.2370
ADIRF	-5.9636	-15.6670	4.06e-07	4.21e-04	7.0198
EEF1A2	5.6310	13.5437	1.20e-06	7.20e-04	6.0988
SOX15	-5.1140	-14.0968	8.93e-07	6.27e-04	6.3572
GJB2	-5.0539	-15.5508	4.29e-07	4.21e-04	6.9741
FAM83A	-4.9847	-19.6421	7.39e-08	1.51e-04	8.3248
S100A2	-4.5044	-14.5062	7.22e-07	5.53e-04	6.5397
PKP1	-4.3109	-18.7828	1.03e-07	1.81e-04	8.0804
MTAP	-3.9039	-34.2278	1.07e-09	1.25e-05	10.6974
RBP1	-3.7583	-27.2730	6.10e-09	2.13e-05	9.8799
FBN2	-3.7511	-15.0320	5.53e-07	4.53e-04	6.7638
SERINC2	-3.6314	-32.2441	1.69e-09	1.25e-05	10.5029
SFRP1	-3.5777	-15.5689	4.25e-07	4.21e-04	6.9812
DMKN	-3.4313	-16.4867	2.77e-07	3.39e-04	7.3285

Una vez expuesta la Tabla 4.17 con los valores estadísticos de los veinte genes más relevantes para la serie GSE46834, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Símbolo	Pathway	Asociado a cáncer de mama
KRT6C	-	No
KRT6B	-	No
KRT6A	-	Si
KRT14	Comunicación intercelular	No
KRT5	Comunicación intercelular	No
FGFBP1	Transducción de señales	Si
KLK5	-	Si
ADIRF	-	No
EEF1A2	Metabolismo de proteínas	No
SOX15	-	No
GJB2	Transporte vesícula mediada	No
FAM83A	-	Si
S100A2	-	No
PKP1	Muerte celular programada	No
MTAP	Relación con metabolismo	No
RBP1	Relación con metabolismo	Si
FBN2	Organización de matriz extracelular	No
SERINC2	-	No
SFRP1	Transducción de señales	Si
DMKN	-	No

Tabla 4.18
Pathway de los veinte genes más destacados para la serie GSE46834

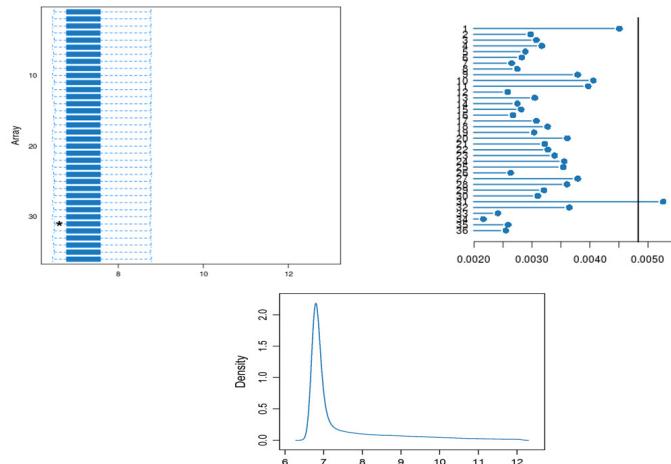
Como se puede observar en la Tabla 4.18, de los 20 genes destacados, 6 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 14 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso.

4.3.2 Análisis serie GSE68651

Por último, se ha analizado la serie GSE68651 de Illumina, esta serie contiene 36 muestras de las cuales 12 son muestras de control tomadas de células epiteliales MCF10A y 24 muestras de células cancerígenas de cáncer de mama de dos tipos, MCF7 y Hs578T. Esta serie pertenece al University of Southampton ubicado en Southampton, UK.

En el análisis de calidad se ha detectado un outlier como se puede ver en la Figura 4.27. En concreto, el outlier pertenece a la muestra número 31 que representa una muestra cancerígena. Se puede observar en el boxplot de intensidades de los arrays y en el gráfico creado a raíz del test de Kolmogorov-Smirnov como ningún array supera el valor K estimado en el test, en este caso, equivalente a 0.00483. Se ha procedido a la eliminación del outlier y ha repetir el análisis de calidad, que esta vez no encontró ningún outlier más.

Figura 4.27
Test de Kolmogorov-Smirnov aplicado a la serie GSE68651



Una vez realizados los análisis, se han obtenido como en las series anteriores, los veinte genes más relevantes en orden decreciente por su FC.

Símbolo	logFC	T	Pval	Adj. Pval	B
KRT6A	-6.7108	-56.8362	7.81e-36	5.51e-33	71.8302
S100A2	-5.5548	-139.0589	2.99e-49	2.35e-45	99.4022
KRT5	-5.2385	-160.9981	1.87e-51	8.83e-47	103.1536
SERPINB5	-5.0473	-31.6207	3.60e-27	6.96e-25	51.9871
KRT6C	-5.0132	-153.8747	8.97e-51	2.11e-46	102.0278
COL17A1	-4.9435	-45.7245	1.34e-32	5.38e-30	64.5069
IGFBP5	4.9373	39.5205	1.92e-30	5.83e-28	59.5598
LAMC2	-4.7839	-105.9063	3.71e-45	1.75e-41	91.7014
MT1E	-4.7102	-8.0258	2.03e-09	3.11e-08	10.4128
MT1IP	-4.6863	-19.7182	1.97e-20	1.48e-18	36.2336
C3orf14	4.6708	13.4288	2.63e-15	9.56e-14	24.2009
KRT7	-4.5639	-61.9488	4.04e-37	3.41e-34	74.6968
KRT81	-4.4945	-6.4499	2.06e-07	2.24e-06	5.7512
MT2A	-4.4363	-20.0345	1.18e-20	9.22e-19	36.7520
ALDH3A1	-4.3237	-28.6786	9.44e-26	1.45e-23	48.6839
LONC00857	-4.2902	-62.5664	2.87e-37	2.56e-34	75.0253
ANGPTL4	-4.2774	-22.4333	3.09e-22	3.08e-20	40.4675
LAMB3	-4.1584	-89.1733	1.41e-42	3.35e-39	86.4557
MT1G	-4.1425	-48.2339	2.16e-33	1.04e-30	66.3138
LAMA3	-4.1104	-36.1913	3.80e-29	9.93e-27	56.5695

Tabla 4.19
Veinte genes más destacados para la serie GSE68651

Una vez expuesta la Tabla 4.19 con los valores estadísticos de los veinte genes más relevantes para la serie GSE68651, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Tabla 4.20
Pathway de los veinte genes más destacados para la serie GSE68651

Símbolo	Pathway	Asociado a cáncer de mama
KRT6A	-	Si
S100A2	-	No
KRT5	Comunicación intercelular	No
SERPINB5	-	Si
KRT6C	-	No
COL17A1	Comunicación intercelular	No
IGFBP5	Metabolismo de proteínas	Si
LAMC2	Comunicación intercelular	No
MT1E	Relación con metabolismo	Si
MT1IP	-	Si
C3orf14	-	No
KRT7	-	Si
KRT81	-	No
MT2A	Sistema inmunológico	No
ALDH3A1	Relación con metabolismo	Si
LONCo0857	-	-
ANGPTL4	Relación con metabolismo	Si
LAMB3	Comunicación intercelular	No
MT1G	Relación con metabolismo	No
LAMA3	Organización de matriz extracelular	Si

Como se puede observar en la Tabla 4.20, de los 20 genes destacados, 9 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 11 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso.

4.3.3 Análisis integrador de Illumina mediante VirtualArray

Tal y como se llevó a cabo con las series pertenecientes a Affymetrix, se procede a hacer el análisis integrador de las dos series estudiadas de Illumina con el fin de encontrar mediante la herramienta VirtualArray y los diagramas de Venn, genes más robustos que tengan en común ambas series más la serie creada a partir de la integración de estas.

Se han realizado dos diagramas de Venn tal y como se hizo para las series de Affymetrix, uno usando los veinte genes relevantes detallados para cada serie en el apartado anterior. Como se muestra en la Figura 4.28, encontrado genes relevantes comunes al estudio integrador, en concreto, cinco genes en común de los cuales tres de ellos guardarían relación con el cáncer de mama. Los genes encontrados han sido los siguientes:

- KRT6C: Según DisGeNET, no está relacionado con el cáncer de mama.
- KRT6A: Según DisGeNET, está relacionado con el cáncer de mama.
- KRT5: Según DisGeNET, está relacionado con el cáncer de mama.
- FGFBP1: Según DisGeNET, está relacionado con el cáncer de mama.
- S100A2: Según DisGeNET, no está relacionado con el cáncer de mama.

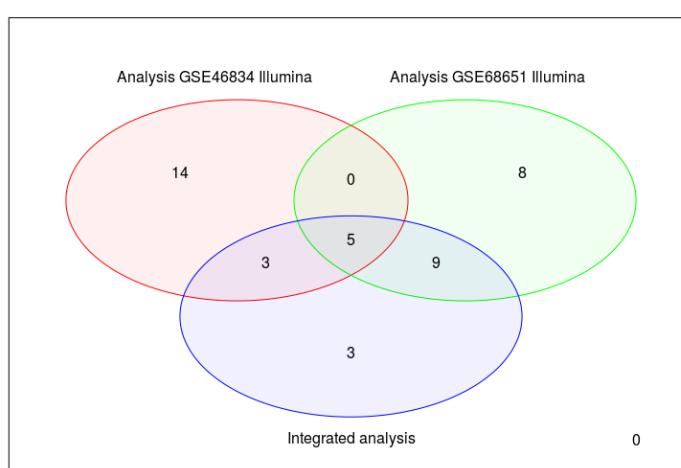
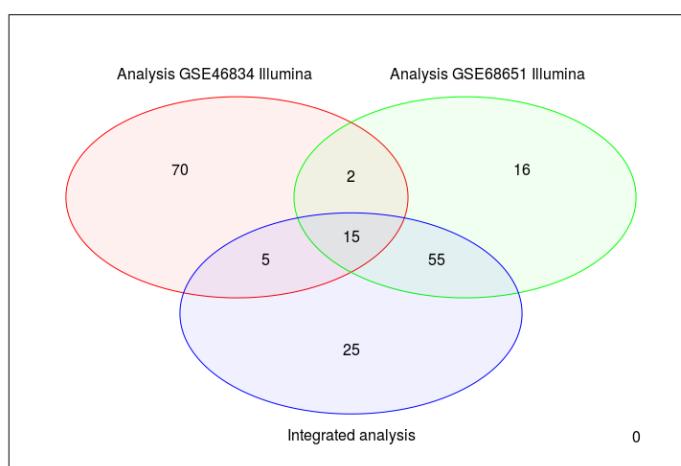


Figura 4.28
Diagrama de Venn de la integración de Illumina usando 20 genes

El diagrama de Venn restante, incluye un total de cien genes destacados por cada serie sin cambiar las restricciones de estas. Tal y como se puede ver en la Figura 4.29, en este caso se han detectado quince genes robustos en común a estas series. De estos quince genes, puede verse como 6 de ellos aparecen relacionados con el cáncer de mama en la plataforma DisGeNET. Estos genes detectados han sido los siguientes:

- KRT6C: Según DisGeNET, no está relacionado con el cáncer de mama.
- KRT6A: Según DisGeNET, está relacionado con el cáncer de mama.
- KRT14: Según DisGeNET, no está relacionado con el cáncer de mama.
- KRT17: Según DisGeNET, está relacionado con el cáncer de mama.
- KRT5: Según DisGeNET, está relacionado con el cáncer de mama.
- FGFBP1: Según DisGeNET, está relacionado con el cáncer de mama.
- ADIRF: Según DisGeNET, no está relacionado con el cáncer de mama.
- SOX15: Según DisGeNET, no está relacionado con el cáncer de mama.
- GJB2: Según DisGeNET, no está relacionado con el cáncer de mama.
- FAM83A: Según DisGeNET, está relacionado con el cáncer de mama.
- KRT16: Según DisGeNET, no está relacionado con el cáncer de mama.
- S100A2: Según DisGeNET, no está relacionado con el cáncer de mama.
- COL17A1: Según DisGeNET, no está relacionado con el cáncer de mama.
- FBN2: Según DisGeNET, no está relacionado con el cáncer de mama.
- SFRP1: Según DisGeNET, está relacionado con el cáncer de mama.



Venn Diagram of isolated Illumina datasets using 100 genes

En los siguientes boxplots se representaran la variabilidad de los rangos dinámicos en diferentes casos. El primero de estos casos, tal como se puede ver en la Figura 4.30, se muestra la variabilidad entre las series de Illumina antes de integrarlas. Se puede ver bastante variabilidad entre ambas series pero no entre las muestras de una misma serie.

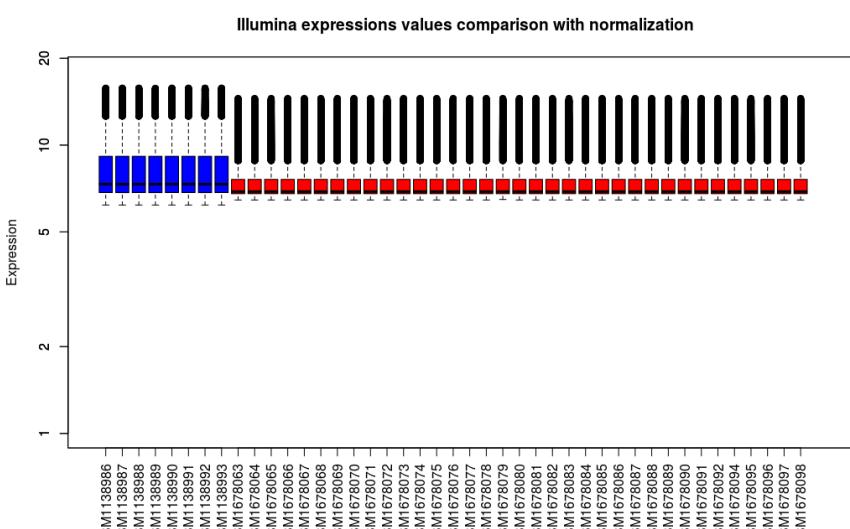


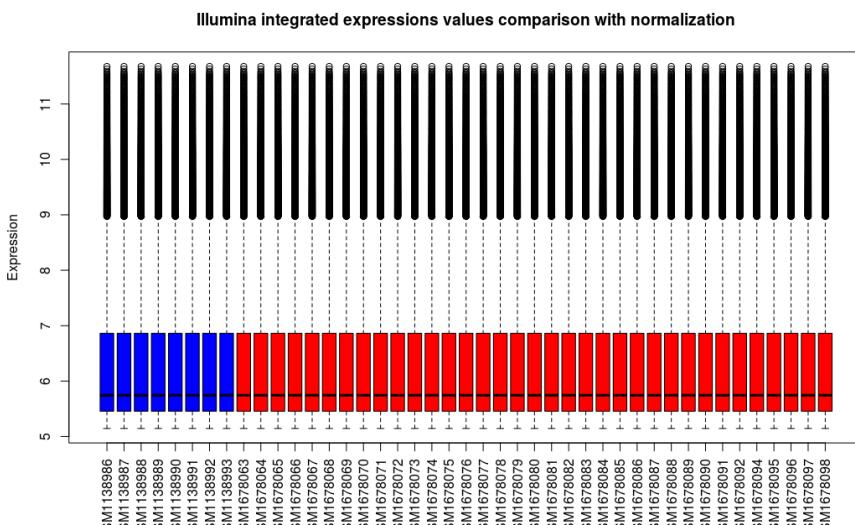
Figura 4.30
Valores de expresión
normalizados de las
series de Illumina por
separado

En el otro caso representado, se han unido las series y normalizado mediante la herramienta `virtualArray` en una sola serie. En el boxplot que muestra la Figura 4.31 se puede ver como prácticamente la variabilidad ha quedado eliminada para poder así hacer un estudio en

iguales condiciones entre todas las muestras integradas, logrando así más robustez y fiabilidad en los resultados.

En los boxplots representados, el eje vertical representa el número de bits empleados, también conocido como profundidad de bits el cual es equivalente a 16 para todas las series usadas en este estudio. Si alguna serie tuviese una profundidad de bits distinta al resto, debería igualarse dicha profundidad, puesto que el ruido de cuantización podría afectar al análisis diferencial.

Figura 4.31
Valores de expresión normalizados de las series de Illumina integradas



4.4 DISCUSIÓN Y RESULTADOS DE LAS SERIES DE ILLUMINA EN MATLAB

Después de realizarse el estudio de las series de Illumina usando las librerías de bioconductor para R, se ha procedido a realizar un estudio paralelo esta vez usando la Bioinformatic Toolbox de Matlab. Con dicho estudio se pretende comprobar como varían los resultados en función de la herramienta o la librería utilizada e intentar encontrar genes en ambos estudios en común.

4.4.1 Análisis serie GSE46834

Tal como se hizo en el análisis en R, se ha construido una tabla de los 20 genes más relevantes y sus valores estadísticos según la toolbox de Matlab. También se puede ver en la Figura 4.32, los valores ya normalizados para esta serie.

A continuación, se mostrará una Tabla 4.21 con los veinte genes más relevantes ordenados en orden decreciente en base a su p-value.

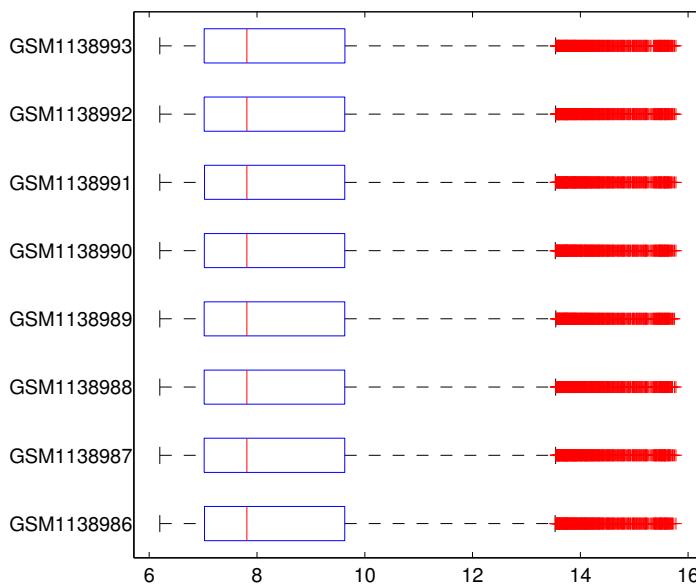


Figura 4.32
Valores de expresión normalizados de la serie GSE46834 de Affymetrix

Símbolo	t-scores	PVal	FDR
TTLL2	-17.992	2.1344e-06	0.025236
FILIP1L	15.518	4.847e-06	0.028654
ATL1	13.883	1.2397e-05	0.048858
NSMF	-12.734	1.7176e-05	0.050769
HAGH	-12.109	1.9267e-05	0.04556
DHCR24	11.776	2.4261e-05	0.047808
FCRLA	12.834	2.5822e-05	0.043615
FGFR3	11.045	4.2205e-05	0.062375
ZEB1	-10.626	4.2447e-05	0.055762
ACTR3	9.8046	7.9366e-05	0.093837
FABP4	9.3118	8.8413e-05	0.09503
CITED4	9.0228	0.00010479	0.10325
ZNF32	-9.0072	0.00010529	0.095759
S100A16	8.8832	0.0001203	0.1016
PKN1	-9.9521	0.00013181	0.1039
MBOAT2	8.8362	0.00013194	0.097494
CAP1	8.5019	0.00015328	0.10661
MID2	9.1341	0.00016332	0.10728
HAP1	8.2772	0.00016883	0.10506
ALOX12P2	9.1515	0.00016925	0.10006

Tabla 4.21
Veinte genes más destacados para la serie GSE46834 según Matlab

Una vez expuesta la Tabla 4.21 con los valores estadísticos de los veinte genes más relevantes para la serie GSE46834, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores según la plataforma DisGeNET.

Tabla 4.22
Función de los veinte genes más destacados para la serie GSE46834 según Matlab

Símbolo	Pathway	Asociado a cáncer de mama
TTL2	-	-
FILIP1L	-	No
ATL1	-	No
NSMF	-	Si
HAGH	Relación con metabolismo	Si
DHCR24	Relación con metabolismo	No
FCRLA	-	No
FGFR3	-	Si
ZEB1	Sistema inmunológico	No
ACTR3	Sistema inmunológico	No
FABP4	Relación con metabolismo	Si
CITED4	-	Si
ZNF32	-	Si
S100A16	-	Si
PKN1	Transducción de señales	Si
MBOAT2	Relación con metabolismo	Si
CAP1	Hemostasis	Si
MID2	-	No
HAP1	-	Si
ALOX12P2	-	No

Como se puede observar en la Tabla 4.22, de los 20 genes destacados, 11 de ellos ya están relacionados con el cáncer de mama según la plataforma DisGeNET. Los 9 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido según la plataforma DisGeNET.

4.4.2 Análisis serie GSE68651

Tal como se hizo en el análisis en R, se ha construido una tabla de los 20 genes más relevantes y sus valores estadísticos según la toolbox de Matlab. También se puede ver en la Figura 4.33, los valores ya normalizados para esta serie.

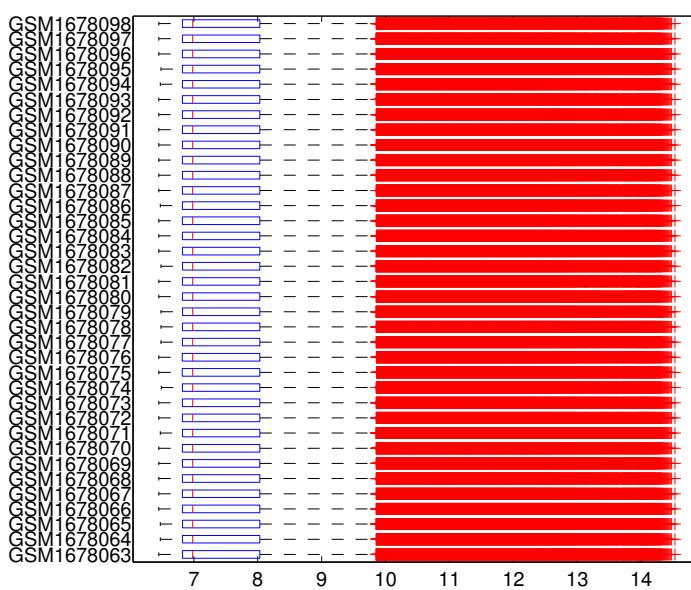


Figura 4.33
Valores de expresión
normalizados de la
serie GSE68651 de
Affymetrix

A continuación, se mostrará una Tabla 4.23 con los veinte genes mas relevantes ordenados en orden decreciente en base a su p-value.

Una vez expuesta la Tabla 4.23 con los valores estadísticos de los veinte genes más relevantes para la series GSE68651, se procederá a mostrar otra tabla con la explicación de la función de cada uno de los genes anteriores segúrn la plataforma DisGeNET.

Como se puede observar en la Tabla 4.24, de los 20 genes destacados, 6 de ellos ya están relacionados con el cáncer de mama segúrn la plataforma DisGeNET. Los 14 genes restantes, en principio están relacionados con otro tipo de enfermedades y, al menos directamente, no estarían relacionados con la enfermedad a estudiar en este caso. Además, se puede observar como muchos de estos genes tienen un pathway desconocido segúrn la plataforma DisGeNET.

Tabla 4.23
Veinte genes más destacados para la serie GSE68651 según Matlab

Símbolo	t-scores	PVal	FDR
IGFBP5	-73.11	1.8649e-37	3.0933e-33
PRSS23	-52.463	4.3542e-30	3.6111e-26
SERPINA1	41.721	5.373e-30	2.9707e-26
ARNTL	39.826	5.2339e-29	1.7363e-25
CBS	-46.496	4.2733e-28	1.1813e-24
PYGB	36.467	4.4791e-28	1.0613e-24
TBL1X	-51.189	5.2209e-28	1.0825e-24
FAM111A	34.694	1.6127e-27	2.9721e-24
KLHDC3	-46.849	1.9077e-27	3.1643e-24
MT1E	38.628	2.0245e-27	3.0527e-24
FAM46A	-40.975	5.3171e-27	6.7842e-24
ANPEP	34.209	6.0155e-27	6.6519e-24
ARV1	33.661	6.1955e-27	6.4227e-24
NUDT11	42.951	1.3097e-26	1.2778e-23
KRT81	35.927	1.7169e-26	1.5821e-23
TEAD2	-40.054	2.35e-26	2.0515e-23
TFPT	-35.006	3.958e-26	3.2825e-23
KLF2	-32.252	4.6622e-26	3.6824e-23
DUT	39.287	6.394e-26	4.8207e-23
RBMS1	-30.476	1.8165e-25	1.31e-22

Símbolo	Pathway	Asociado a cáncer de mama
IGFBP5	Metabolismo de proteínas	Si
PRSS23	-	Si
SERPINA1	Metabolismo de proteínas	Si
ARNTL	Relación con metabolismo	No
CBS	Relación con metabolismo	No
PYGB	Relación con metabolismo	No
TBL1X	Relación con metabolismo	Si
FAM111A	-	No
KLHDC3	-	-
MT1E	Relación con metabolismo	Si
FAM46A	-	Si
ANPEP	Metabolismo de proteínas	No
ARV1	Relación con metabolismo	No
NUDT11	Relación con metabolismo	No
KRT81	-	No
TEAD2	Relación con metabolismo	No
TFPT	Reparación de ADN	No
KLF2	-	No
DUT	Relación con metabolismo	No
RBMS1	-	No

Tabla 4.24
Función de los veinte genes más destacados para la serie GSE68651 según Matlab

4.4.3 Comparación entre análisis en R y Matlab de Illumina

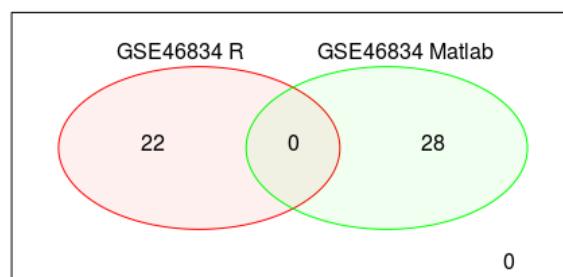
Una vez expuestos los resultados realizados tanto con R como con Matlab como en el caso de las series de Affymetrix, se han realizado dos comparaciones para cada series. Una de ellas con un diagrama de venn con los veinte genes más relevantes según R y Matlab y otro con los cien genes más relevantes.

Primero se procede ha exponer las posibles coincidencias entre los veinte genes más relevantes para R frente a los veinte genes más relevantes para Matlab en la serie GSE46834.

Como se puede observar en la Figura 4.34, de los veinte genes más destacados para esta serie según R y Matlab, no aparece ningún gen común entre ambos estudios y tan solo existen cinco genes en común en caso de mirar la comparación entre lo cien genes más destacados tal y como se ve en la Figura 4.35.

Figura 4.34
Comparación de los
20 genes más rel-
evantes entre R y
Matlab de la serie
GSE46834

Venn diagram of R & Matlab genes comparison of GSE46834



Se analizará por último la serie GSE68651. En la Figura 4.36 se puede observar como de los veinte genes solo tienen tres en común y en el caso de la comparación de los cien genes como se ve en la Figura 4.37 se detecta doce genes en común a ambos estudios.

Después de analizar y comparar los resultados entre R y Matlab de las series elegidas de Illumina, se puede observar el bajo número de genes expresados en común. Esto podría denotar falta de robustez en los genes expresados en dichas series.

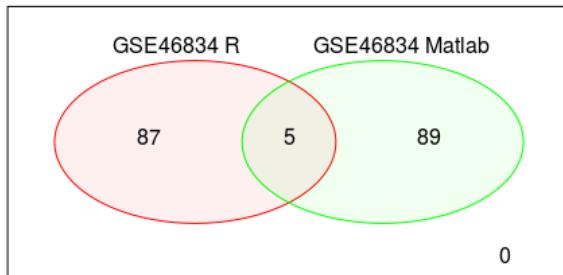
Venn diagram of R & Matlab genes comparison of GSE46834

Figura 4.35
Comparación de los 100 genes más relevantes entre R y Matlab de la serie GSE46834

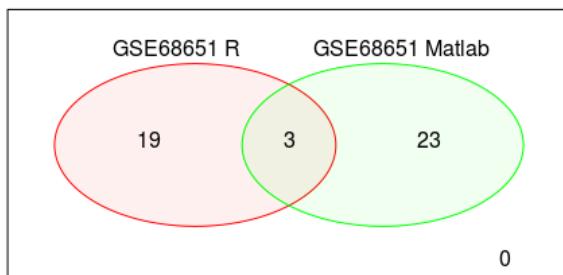
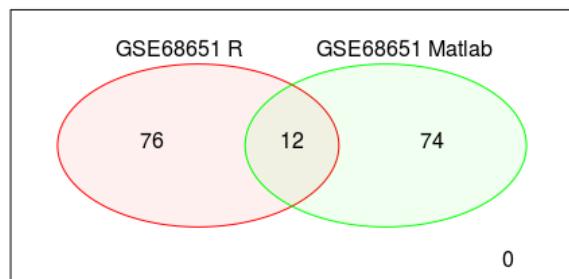
Venn diagram of R & Matlab genes comparison of GSE68651

Figura 4.36
Comparación de los 20 genes más relevantes entre R y Matlab de la serie GSE68651

Figura 4.37
Comparación de los
100 genes más rel-
evantes entre R y
Matlab de la serie
GSE68651

Venn diagram of R & Matlab genes comparison of GSE68651



4.5 DISCUSIÓN DEL ESTUDIO INTEGRADOR DE LAS SERIES DE AFFYMETRIX E ILLUMINA

Una vez analizadas las series de cada tecnología por separado, se procede al análisis conjunto de las series de ambas tecnologías. Como bien se ha explicado durante la redacción de este proyecto, todas las series elegidas pertenecen a muestras de células cancerígenas de cáncer de mama y a muestras de control sanas de células epiteliales de personas sin la enfermedad a estudiar.

El objetivo final de esta integración es doble. Por un lado, se pretende aumentar el número de muestras a estudiar para dotar al análisis de mayor robustez y significancia. Por otro lado, se intenta gracias a este estudio, intentar independizar los resultados de la tecnología usada para tomar las muestras y el posible ruido y efecto que esto pueda tener en los resultados finales. En la Tabla 4.25 se puede ver un resumen del número de muestras y outliers que conforman este estudio.

Serie	Tecnología	Nº muestras de buena calidad	Nº outliers excluidos
GSE52712	Affymetrix	19	1
GSE40987	Affymetrix	10	0
GSE52262	Affymetrix	16	0
GSE12790	Affymetrix	20	1
GSE46834	Illumina	8	0
GSE68651	Illumina	35	1
TOTAL	Integrado	108	3

Tabla 4.25
Series usadas en el estudio integrador final

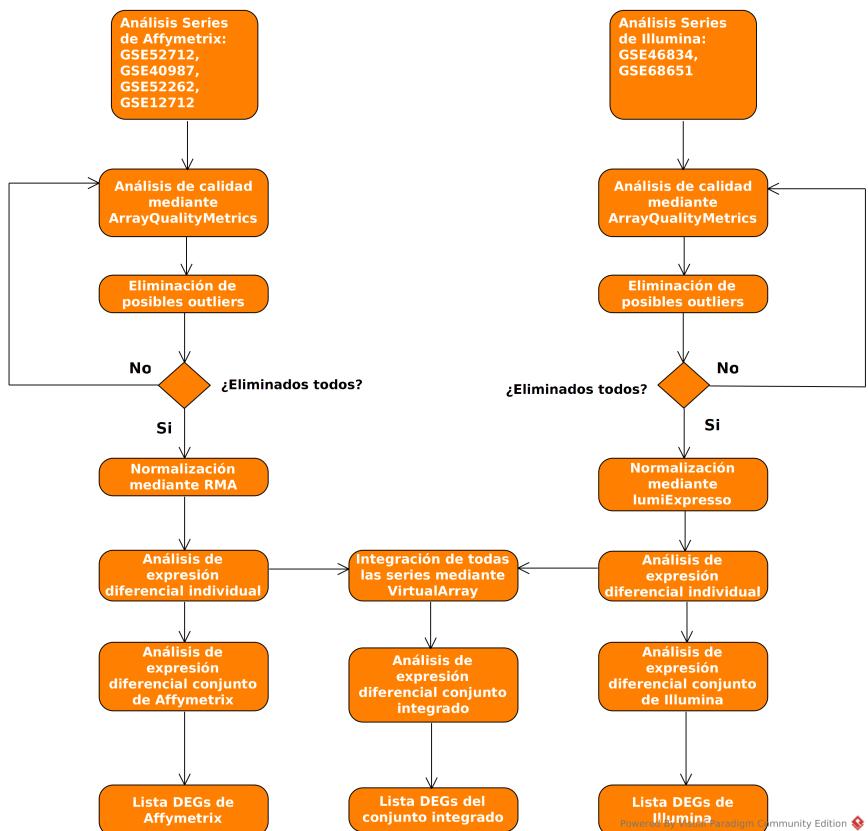
En la Figura 4.38 puede verse el flujo seguido durante el estudio hasta llegar al estudio integrador que se describe en este apartado.

Se procederá a describir y analizar el resultado del estudio integrador en tres subsecciones. Primero se expondrán los resultados del estudio integrador aplicando el método de unión de las muestras por defecto, que es mediante la mediana, y sin aplicar ninguna técnica de supresión del efecto batch.

Después, se explicarán los resultados obtenidos al unir las muestras con las diferentes opciones que virtualArray contiene, en este caso con la mediana y con la media, y ha comparar dichos resultados entre sí. Con esto se intentará ver si hay variación en los genes expresados en función de la técnica usada al crear el conjunto integrado.

Por último, se realizará una comparación pero esta vez usando diferentes técnicas de supresión del efecto batch y con el mismo fin que al variar las técnicas de unión de las muestras, encontrar si se producen variación en los genes expresados o si, en cambio, los resultados son iguales.

Figura 4.38
Flujo de actividad seguido en el proyecto



4.5.1 Análisis de las series integradas

Una vez expuestos los análisis integradores de Affymetrix e Illumina por separado, se ha procedido a unir las series de ambas tecnologías en una sola serie integradora.

En el diagrama de Venn que muestra la Figura 4.39, se incluye un total de cien genes destacados por cada serie sin cambiar las restricciones de estas. En este caso se han detectado cuatro genes robustos en común a estas series. Para crear este diagrama se han dejado dos series fuera, en concreto las GSE52262 y GSE12790, ante la limitación de la herramienta a representar como máximo 5 conjuntos a la vez. No obstante, se han creado los respectivos diagramas de Venn introduciendo dichas series y cambiándolas por otras para poder ver así como varían los genes expuestos en cada caso y buscar genes comunes a todas las combinaciones. Se ha comprobado que todas las combinaciones tienen en común un gen llamado "SFRP1" y que al menos otros dos genes están en todas las combinaciones menos en una de ellas. Se expondrán estos tres genes a continuación:

- SFRP1 (Común a todas): Según DisGeNET, está relacionado con el cáncer de mama.

- KRT14: Según DisGeNET, no está relacionado con el cáncer de mama.
- S100A2: Según DisGeNET, no está relacionado con el cáncer de mama.

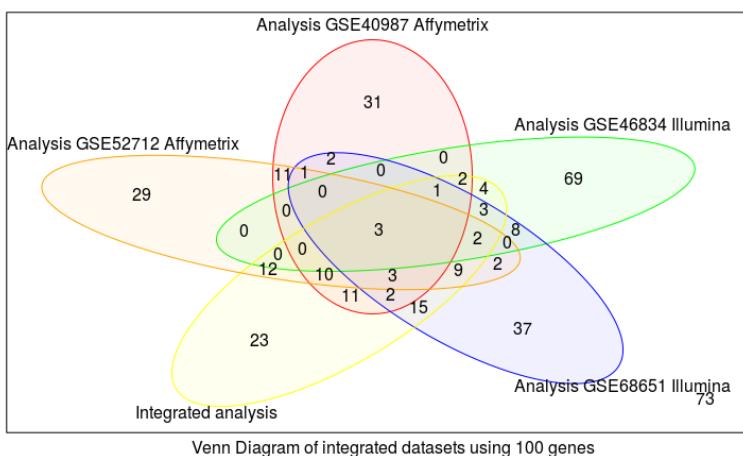


Figura 4.39
Diagrama de Venn de la todas las series integradas dejando 2 fuera

En el boxplot en escala logarítmica representado en la Figura 4.40 se puede observar como la variabilidad de los rangos dinámicos de la expresión es elevada, en algunos casos no solo entre serie y serie, sino dentro de una misma serie también.

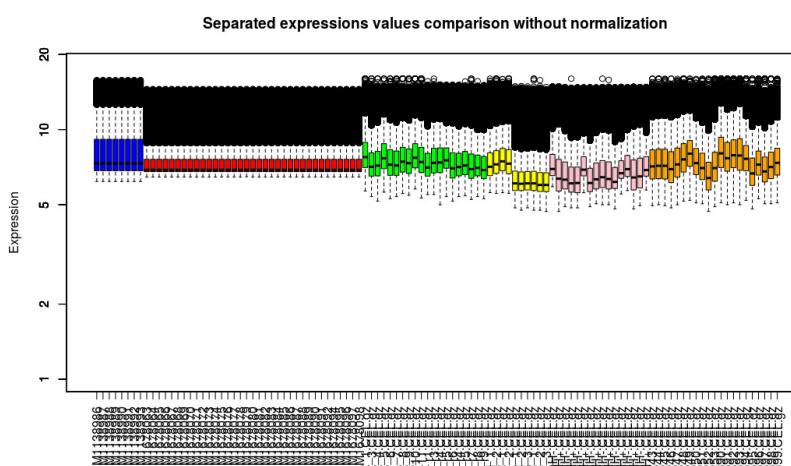


Figura 4.40
Valores de expresión no normalizados de las series de integradas por separado

Tal como se puede ver en la Figura 4.41, se muestra la variabilidad entre las series antes de integrarlas una vez normalizadas por separado. Se puede ver bastante variabilidad entre ambas series pero no entre las muestras de una misma serie.

En el último boxplot representado, se han unido las series y normalizado mediante la herramienta virtualArray en una sola serie. En el

Figura 4.41
Valores de expresión no normalizados de las series de integradas por separado

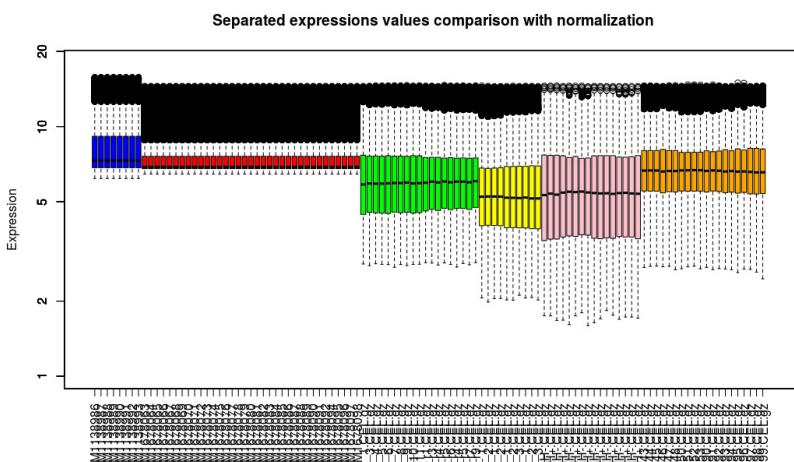
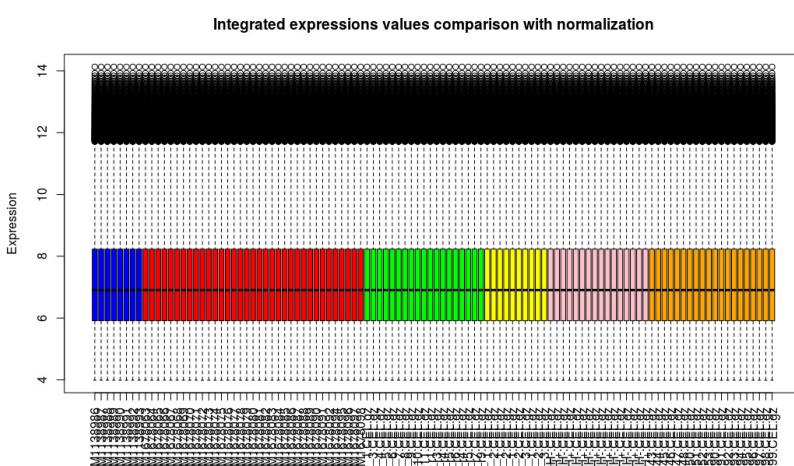


Figura 4.42
Valores de expresión normalizados de las series integradas

boxplot que muestra la Figura 4.42 se puede ver como prácticamente la variabilidad ha quedado eliminada totalmente.



4.5.2 Comparación de técnicas de unión de las muestras

En este apartado se han creado dos series a partir de las muestras de Affymetrix e Illumina. Una de las series usando la media como método de unión de las muestras y la otra usando la mediana. Para la comparación, a ambas series se les ha impuesto un FC mínimo de 2 y un p-value menor de 0.001.

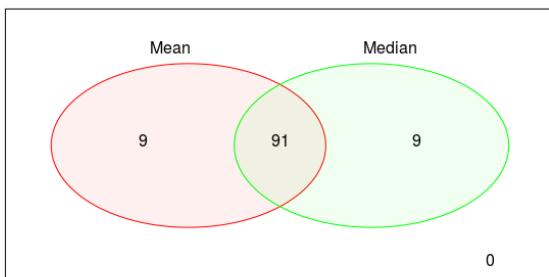
Tal y como se observa en la Figura 4.43, de los 100 genes expresados en ambas series, comparten 91, teniendo 9 genes diferentes expresados debido al método de unión usado para cada una de ellas. Los genes expresados de manera diferente para cada una de ellas han sido los siguientes:

- Genes expresados en caso de unión por media: IGFBP5, OSR2, EMP1, COL5A1, GJB2, S100A3, SLC24A3, LAMB3, TRIM22.
- Genes expresados en caso de unión por mediana: CASP1, RHOBTB3, BIRC3, IGFBP7, KRT7, TBL1X, SLC26A2, C1orf116, FAM174B.

Se puede comprobar como entre los genes diferenciados entre ambas series no están los tres genes destacados en el apartado anterior, de esta manera se puede concluir que son genes bastante robustos al no afectarle el modo de unión de las muestras.

Venn Diagram of median and mean union comparison

Figura 4.43
Diagrama de venn
comparando distintos
métodos de unión



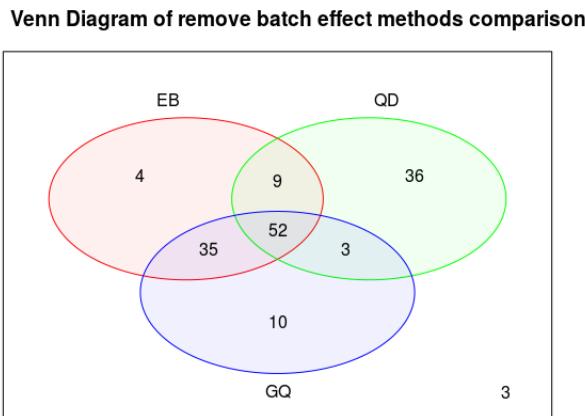
4.5.3 Comparación de técnicas de eliminación del efecto batch

Por último, se ha procedido la comparación de los genes destacados entre distintas series creadas con los distintos métodos de supresión del efecto batch que la herramienta virtualArray provee. Estas técnicas fueron explicadas en el apartado 3.

De las seis técnicas de supresión del efecto batch sólo se usarán tres de ellas en la comparación y esto se debe a que las técnicas EB, NORDI, MRS y MC presentan los mismo resultados entre ellas y destacan los mismo genes, al menos significativos con las restricciones impuestas. Se usarán, por tanto, las técnicas EB, GQ y QD en la comparación.

Se puede ver el resultado de dicha comparación en el diagrama de Venn que muestra la Figura 4.44.

Figura 4.44
Diagrama de venn
comparando distintos
métodos de elimi-
nación del efecto
batch



Como se puede ver en el diagrama anterior, existen 52 genes en común que indican la robustez de dichos genes debido a que a pesar de haber usado diferentes técnicas de supresión de efecto batch, dichos genes siguen expresándose. También se puede observar como tanto la técnica EB como GQ aportan resultados muy similares, en cambio, la técnica QD tiene resultados muy diferentes con 36 genes expresados de forma diferente con respecto a las otras técnicas. Se han recogido los 52 genes expresados en la comparación en la Tabla 4.26 para comprobar que todos cumplen con las restricciones impuestas anteriormente.

El listado de genes mostrado en la tabla tiene gran relevancia puesto que contiene un número de genes diana moderado, algunos de los cuales podrían ser relevantes y aún no reportados en el cáncer de mama. Esta información podría ser proporcionada al personal médico con el fin de poder validar dichos genes en laboratorio.

Símbolo	logFC	T	Pval	Adj. Pval	B
KRT6A	-5.6244	-19.0885	1.24e-36	1.44e-33	72.8355
S100A2	-5.0766	-18.3875	2.84e-35	2.31e-32	69.7432
KRT5	-4.9778	-17.5340	1.40e-33	8.14e-31	65.8917
KRT14	-4.5092	-12.7368	2.37e-23	2.98e-21	42.5626
IL20RB	-4.1427	-18.1454	8.52e-35	6.14e-32	68.6606
FGFBP1	-4.0367	-19.2290	6.68e-37	9.03e-34	73.4475
NNMT	-3.8638	-10.5582	2.13e-18	1.16e-16	31.2539
KRT19	3.5779	8.9265	1.15e-14	3.02e-13	22.7429
SFRP1	-3.5532	-21.7776	1.32e-41	1.07e-37	84.1124
SERPINB5	-3.5242	-20.0800	1.64e-38	2.96e-35	77.1006
ADRB2	-3.4898	-18.9143	2.68e-36	2.91e-33	72.0733
DSG3	-3.4566	-19.1403	9.87e-37	1.23e-33	73.0617
CLCA2	-3.4201	-16.9787	1.86e-32	8.89e-30	63.3348
SLPI	-3.4063	-13.5676	3.35e-25	5.91e-23	46.7873
C3	-3.3636	-15.2953	5.98e-29	1.73e-26	55.3427
HENMT1	3.2119	14.1969	1.39e-26	2.89e-24	49.9421
CXCL1	-3.1760	-13.4876	5.04e-25	8.80e-23	46.3831
COL17A1	-3.1649	-17.6471	8.33e-34	5.01e-31	66.4075
PRKCDBP	-3.1516	-11.8589	2.28e-21	2.06e-19	38.0361
UCP2	3.1333	10.4029	4.83e-18	2.46e-16	30.4411
EFHD1	3.1331	15.9505	2.47e-30	8.74e-28	58.4961
RGS2	-3.0960	-14.6618	1.36e-27	3.35e-25	52.2448
IFI16	-3.0862	-9.9880	4.32e-17	1.89e-15	28.2704
ZBTB16	-3.0713	-17.3165	3.84e-33	1.95e-30	64.8950
DNER	-3.0676	-13.4137	7.35e-25	1.21e-22	46.0095
GNA15	-3.0133	-20.2966	6.49e-39	1.50e-35	78.0153
PNLIPRP3	-2.9116	-17.5079	1.58e-33	8.57e-31	65.7728
KRT6B	-2.9108	-14.6821	1.23e-27	3.07e-25	52.3451
BNC1	-2.8126	-21.4404	5.29e-41	2.86e-37	82.7484
FBP1	2.7716	7.8603	2.86e-12	4.73e-11	17.2920
RAB38	-2.7691	-15.9002	3.15e-30	1.06e-27	58.2560
TSPYL5	2.7624	8.4461	1.40e-13	2.92e-12	20.2689
NMU	-2.6986	-13.3432	1.05e-24	1.69e-22	45.6526
MAOA	-2.6737	-9.9403	5.56e-17	2.35e-15	28.0213
EVA1C	-2.6693	-16.5302	1.54e-31	6.49e-29	61.2410
GPR87	-2.6683	-16.4441	2.33e-31	9.22e-29	60.8357
CPVL	-2.6535	-12.1749	4.38e-22	4.44e-20	39.6718
CBS	2.6525	16.5332	1.52e-31	6.49e-29	61.2548
CASP1	-2.6211	-12.0410	8.81e-22	8.56e-20	38.9794

Tabla 4.26
Genes en común
comparando series
con distintos método
de supresión de batch

FAM83A	-2.5881	-13.2689	1.54e-24	2.35e-22	45.2757
SDPR	-2.5646	-14.4011	5.00e-27	1.14e-24	50.9565
MSLN	-2.5447	-13.8136	9.62e-26	1.86e-23	48.0251
WBP5	-2.5391	-9.1515	3.54e-15	1.03e-13	23.9089
DFNA5	-2.5289	-11.3492	3.30e-20	2.42e-18	35.3860
TNNI2	-2.4727	-17.7094	6.25e-34	3.90e-31	66.6909
IRX4	-2.4620	-13.0144	5.68e-24	8.16e-22	43.9810
TFCP2L1	-2.4567	-19.8150	5.15e-38	8.37e-35	75.9732
BEX2	2.4219	10.6875	1.07e-18	6.22e-17	31.9302
BIRC3	-2.4074	-15.6758	9.35e-30	3.03e-27	57.1804
INHBB	2.3721	13.7278	1.48e-25	2.76e-23	47.5945
SLC26A2	-2.3175	-17.3073	4.01e-33	1.97e-30	64.8526
C3orf14	2.2990	12.0198	9.83e-22	9.44e-20	38.8698
ACOT4	2.2538	10.9259	3.06e-19	1.91e-17	33.1766

También se puede ver en la tabla como los genes destacados en anteriores apartados(SFRP1, S100A2 y KRT14), están reflejados en ella. Tal y como se esperaba, el FC y el p-value de los 52 genes en común supera el impuesto en el estudio, además, se observan que los coeficientes de bondad también son elevados, lo cual significa que son genes muy robustos. Todo ello podría garantizar la idoneidad de estos genes como biomarcadores en el cáncer de mama.

4.5.4 Interpretación de los resultados del estudio integrador

En este apartado se expondrán los resultados desde un punto más biológico con apoyo de bases de datos que permiten ver más en detalle la funcionalidad de los genes, dichas bases de datos, serán GeneCards [3], WikiGenes [8] y CancerIndex [1].

Los 53 genes hallados independientes a la tecnología y a los métodos de unión y supresión del efecto batch podrían ser genes potencialmente robustos para el cáncer de pulmón. Es por esta razón que se procederá a verificar la función biológica de dicho genes y si han sido registrados en la literatura como genes asociados a esta enfermedad o no, o si por el contrario, aún no se han hecho estudios de ello.

De los 53 genes, se mostrará en este apartado la función de los 10 primeros al tener el FC más relevante. No obstante, el resto de genes serán explicados en el Anexo 6. El conjunto de genes pueden ser entregados a un equipo médico que pueda validarlos de manera clínica.

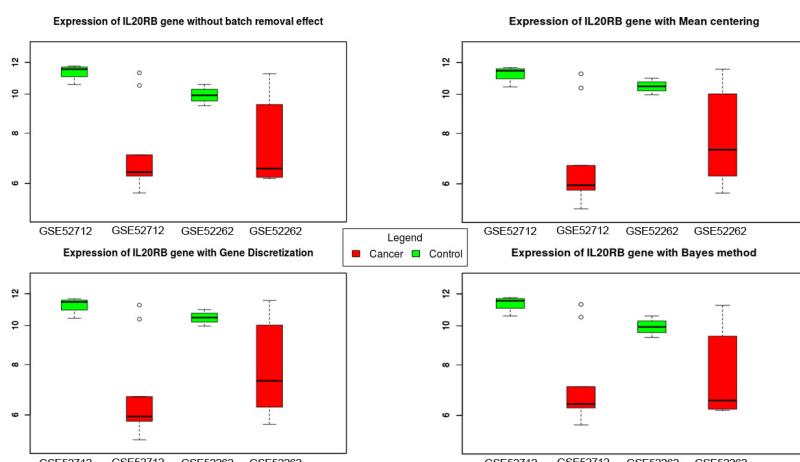
- KRT6A: Según la base de datos de WikiGenes, este gen se muestra sobreexpresado en el cáncer gástrico. La proteína codificada por este gen es un miembro de la familia de genes de queratina.

Las mutaciones en estos genes se han asociado con paquioniquia congénita. Estudios han detectado expresión en este gen en tumores del tipo TNBC(Triple Negative Breast Cancer) [20].

- S100A2: La proteína codificada por este gen es un miembro de la familia de las proteínas S100. Las enfermedades asociadas con S100A2 incluyen fibroma condromixoide y cáncer de pulmón. Se ha detectado entre la inhibición de este gen durante el avance del cáncer de mama y la sobreexpresión en muestras de control por ejemplo, tal y como se puede leer aquí [69].
- KRT5: Este gen codifica un miembro de la familia de la queratina. Una enfermedad asociada con KRT5 es la epidermolisis bullosa. Estudios han detectado expresión en este gen en tumores del tipo TNBC(Triple Negative Breast Cancer) al igual que el gen explicado anteriormente(KRT6) [20].
- KRT14: Este gen codifica un miembro de la familia de la queratina, el grupo más diverso de los filamentos intermedios. Estudios han detectado expresión en este gen en tumores del tipo TNBC(Triple Negative Breast Cancer) al igual que los genes explicados anteriormente(KRT5 y KRT6) [20].
- IL20RB: Entre su pathway está PEDF señalización inducida y la vía de señalización Jak-STAT. Algunas de las enfermedades que se le asocian son psoriasis, enfermedades autoinmunes, cáncer de pulmón y melanoma. En la literatura no se han encontrado evidencias claras de que este gen haya sido detectado como un gen diana en el cáncer de mama, por ello, sería interesante que fuera validado por expertos médicos en PCR. Se ha demostrado que tiene un fold-change muy elevado, lo que significa una diferencia importante de expresión entre el grupo de cáncer y de control. La expresión de este gen puede verse en la Figura 4.45. En dicha figura, se pueden ver diferentes boxplot, cada uno de ellos usando una técnica diferente de eliminación del efecto batch y tal y como puede observarse, estas técnicas no influyen en el nivel de expresión de este gen.
- FGFBP1: Este gen codifica una proteína transportadora del factor de crecimiento de fibroblastos secretada. La proteína codificada juega un papel crítico en la proliferación celular, la diferenciación y la migración mediante la unión a factores de crecimiento de fibroblastos y la potenciación de sus efectos biológicos sobre las células diana. En la literatura ya se ha reportado la expresión de este gen en el cáncer de mama [45].
- NNMT: Las enfermedades asociadas con la NNMT incluyen la enfermedad de parkinson, de inicio tardío. Entre sus rutas relativas son Metabolismo y NAD metabolismo. Se ha comprobado que este gen podría guardar relación con el cáncer de mama también [52].

- KRT19: Este gen codifica un miembro de la familia de la queratina. Las enfermedades asociadas con KRT19 incluyen cáncer de tiroides y cáncer de pulmón. Estudios han detectado expresión en este gen en tumores del tipo TNBC(Triple Negative Breast Cancer) al igual que los genes explicados anteriormente(KRT5, KRT6 y KRT14) [20].
- SFRP1: Este gen puede estar involucrado en la determinación de la polaridad de las células fotorreceptoras en la retina. Las enfermedades asociadas con SFRP1 incluyen el síndrome de Seckel 1 y telangiectasia cutánea. Los estudios determinan sobreexpresión en este gen para la enfermedad estudiada y para otro tipo de cánceres [42].
- SERPINB5: Las enfermedades asociadas con SERPINB5 incluyen cáncer mandibular y carcinoma laringeo. En su pathway se encuentra microARNs en el cáncer. Según algunos estudios [50], se ha encontrado relación entre la metástasis en el cáncer de mama y la sobreexpresión de este gen.

Figura 4.45
Boxplots de la expresión del gen IL20RB usando diferentes técnicas de supresión del efecto batch



De los diez genes expuestos, nueve de ellos han sido reportados alguna vez como posible relación o relación directa con el cáncer de mama. Solo para uno de ellos no se ha encontrado aparente relación en los estudios o no se ha reportado aún. En el Anexo 6 se analizará el resto de genes destacados con el fin de encontrar más genes no reportados.

4.6 CLASIFICACIÓN DE LOS DATOS

En esta última parte de los resultados se expondrán una serie de pruebas de clasificación realizadas con los 53 genes hallados en común en el apartado anterior. Con ello se tratará de comprobar mediante técnicas de machine learning si esos mediante esos genes podría diagnosticarse o no cáncer de mama en caso de tener una muestra nueva sin etiqueta y con qué precisión lo haría dicho sistema.

Esta clasificación se ha llevado a cabo mediante una serie de algoritmos que se explicarán brevemente en el transcurso de este apartado. Los datos han sido las series estudiadas durante este proyecto. Para ello, se han exportado esos datos desde R al formato requerido para su lectura en Matlab y su posterior clasificación.

4.6.1 Clasificación con KNN y SVM

Como bien se ha explicado en el comienzo de este apartado, se creará un clasificador con las series estudiadas en este proyecto y con los genes expresados finalmente como conjunto de train y se sacarán resultados de la clasificación. Una vez hecho esto, también se usarán series de GEO no usadas hasta ahora para crear un conjunto de test y comprobar la eficacia de dicho clasificador ante la presencia de muestra no incluidas en la creación del modelo de clasificación. Se ha hecho uso de Matlab y sus librerías y algoritmos de machine learning. En concreto, se ha creado un clasificador con las muestras que ya se tenían usando los algoritmos KNN y SVM, además de realizar para cada uno de estos algoritmos varias versiones usando validación cruzada con LOO y K-FOLD. Primero se explicarán en qué consisten dichos algoritmos para después pasar a la interpretación de sus resultados.

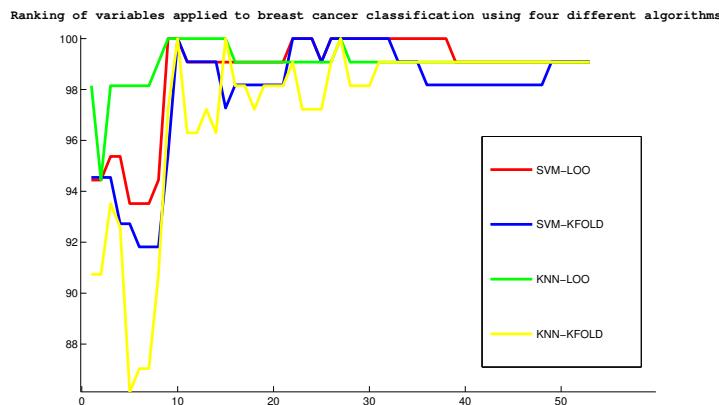
- KNN(K Nearest Neighbors): Este algoritmo pertenece a la clase de aprendizaje supervisado. Calcula la función de densidad de probabilidad posteriormente de que un elemento pertenezca a una clase, es decir, este método no hace ninguna suposición acerca de la distribución de las variables predictorias. El método se basa en la clasificación por proximidad de ejemplos cercanos y parecidos que el elemento a clasificar tenga [48].
- SVM(Support Vector Machine): Este tipo de algoritmo de clasificación crea un modelo para representar nuestro conjunto de muestras en el espacio, para posteriormente tratar de separar mediante un hiperplano las posibles clases que existan a la hora de clasificar, este hiperplano es también conocido como vector de soporte. Dependiendo de la dimensionalidad del problema será un hiperplano o un conjunto de estos de forma que cuanto mejor sea la separación realizada, mejor será la clasificación [99].

- LOO(Leave One Out): Este algoritmo al ser de validación cruzada, pretende particionar los datos para así garantizar la independencia entre los resultados de entrenamiento y de test. Para ello este algoritmo saca en cada iteración a un elemento del conjunto y entrena el conjunto de datos excluyendo dicho elemento. Garantiza un error muy bajo, pero mucho coste de computación [31].
- K-FOLD: En el caso de este algoritmo, la idea es la misma que en el algoritmo LOO pero en vez de excluir esta vez un solo dato, se excluye un conjunto de estos de igual número en cada iteración [4].

Además, también se ha usado un algoritmo de selección de variables, en concreto el algoritmo mRMR, que crea un ranking ordenando las variables en función de la información que estas aporten a la clasificación del conjunto de datos.

Como se puede ver en la Figura 4.46, donde el eje Y representa la precisión obtenida y el eje X el número de genes usados desde 1 a 53 en la clasificación del conjunto train, todos los algoritmos propuestos alcanzan con 9 o 10 de esos genes ya un porcentaje total de clasificación en el conjunto inicial de datos llegando incluso a empeorar en algunas ocasiones al seleccionar más genes para la clasificación. Esto supone un ahorro a la hora de computar el hecho de poder rebajar de 53 genes a 10 para crear el clasificador.

Figura 4.46
Ranking de variables y accuracy al clasificar mediante diferentes algoritmos las muestras iniciales de cáncer de mama y de control



Una vez expuestos los resultados de clasificación del conjunto de train, pasará a mostrarse los resultados de clasificación del conjunto de test, formado por 120 muestras de líneas celulares de cáncer de mama y de muestras de control. Este conjunto se ha formado a partir de las siguientes series procedentes de la plataforma GEO:

- GSE75292
- GSE29327

- GSE59734

Como se puede observar en la Figura 4.47, el algoritmo SVM alcanza con 9 o 10 de los genes un porcentaje muy bueno de clasificación, un 97.5 % en el caso de SMV con LOO y un 94.83 % en el caso de SVM con KFOLD. En el caso del algoritmo KNN, para ambas versiones del algoritmo, lo máximo conseguido es 87.96 % usando 26 genes.

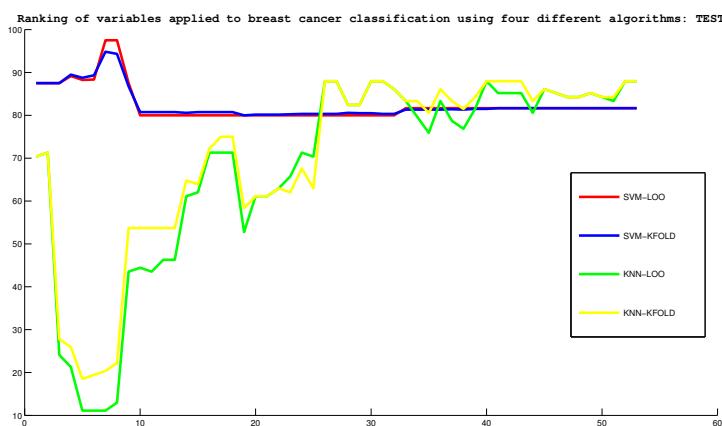


Figura 4.47
Ranking de variables y accuracy al clasificar mediante diferentes algoritmos muestras de cáncer de mama y de control del conjunto de test

Los 10 genes que proporcionan las accuracies más óptimos son los siguientes seg\xfcreo el ranking calculado:

1. KRTA6
2. TNNI2
3. BEX2
4. RGS2
5. CASP1
6. GNA15
7. TFCP2L1
8. CPVL
9. KRT19
10. IRX4

Parte III

CONCLUSIONES

5

CONCLUSIONES Y FUTUROS TRABAJOS

ÍNDICE

- 5.1 Contribuciones originales **112**
 - 5.2 Futuros trabajos **113**
-

Durante la redacción de este proyecto, se han analizado diversas series de diversas tecnologías y con diversos métodos relacionados con el cáncer de mama. En este último capítulo se expondrán las conclusiones finales a las que se han llegado durante la realización del proyecto y posibles futuros trabajos a abordar una vez finalizado este.

5.1 CONTRIBUCIONES ORIGINALES

En el trascurso de este proyecto se han ido haciendo una serie de contribuciones que se expondrán a continuación:

1. Análisis integrados de las tecnologías de Affymetrix e Illumina. Dicho análisis se hizo primeramente por separado para cada una de las tecnologías, algo que ya se había hecho en la literatura pero no se había llegado a hacer este análisis integrando todas ellas por separado y luego integrando ambas tecnologías también en un serie para darle robustez al estudio. La dificultad en este estudio ha sido principalmente la localización de estas series en GEO y que sirvieran al propósito de este estudio. Se han llevado a cabo escrupulosamente los análisis de calidad y la normalización de dichas series para garantizar la idoneidad de los datos usados.
2. Programación en R con las librerías de Bioconductor de todos los análisis, tanto para los análisis en separado como integrados de las series usadas en el proyecto.
3. Programación en Matlab usando la Bioinformatics Toolbox de los análisis por separado de cada serie y comparación de resultados con los respectivos análisis en R.
4. Creación de dos clasificador, uno con SMV y otro con KNN, aplicando también técnicas de validación cruzada y selección de variables para corroborar la robustez y precisión a la hora de clasificar muestras a raíz de los genes destacados y comparar ambos clasificadores.

5.2 FUTUROS TRABAJOS

Durante el desarrollo del presente trabajo han surgido varias posibles líneas de investigación.

1. Los análisis de este proyecto en R, se han realizados usando las librerías específicas de Bioconductor. No obstante, y sobre todo en datasets con gran número de muestras, se ha desbordado la memoria del computador usado(Intel I5 con 8Gb de RAM). Una posible línea sería tratar de paralelizar estos análisis y optimizar así tanto el tiempo como los recursos de ejecución de dichos análisis.
2. Posible búsqueda de un biomarcador robusto de cáncer de mama que se expresase en sangre, de manera que una biopsia solo fuese imprescindible cuando previsiblemente el diagnóstico fuese positivo.
3. Búsqueda de series que tengan etiquetas sobre el tiempo de supervivencia del paciente tras el tratamiento, de forma que estos biomarcadores puedan anticipar la esperanza de vida de estos pacientes. De esa forma, se podrían usar para el diagnóstico y el pronóstico también.
4. Desarrollo de una Toolbox para Matlab que permite la integración de muestras de diferentes tecnologías de microarrays tal y como hace la herramienta virtualArray de Bioconductor.
5. La tecnología de los microarrays está en un entorno de constante cambio, por lo que en unos años los datos tomados pueden quedar obsoletos. No obstante, la obtención de estos datos ha llegado a suponer un gran coste tanto en tiempo como en dinero, por lo cual, sería muy interesante el desarrollo de alguna tecnología que permitiese adaptar mediciones antiguas a las mediciones tomadas con las nuevas tecnologías. Por ejemplo, integrar datos de microarrays con datos tomados mediante secuenciación masiva.
6. Extensión de dicho análisis a otros tipos de cánceres, de forma que se puede ampliar el número de genes diana para dichos cánceres también.

Parte IV

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- [1] *Cancer index.* <http://www.cancerindex.org/>. (Citado en la página 102.)
- [2] *Espetrómetro de masas.* http://www.mncn.csic.es/docs/repositorio/es_ES/investigacion/cromatografia/espectrometria_de_masas.pdf. (Citado en la página 23.)
- [3] *Genecards.* <http://www.genecards.org/>. (Citado en la página 102.)
- [4] *K fold.* <https://www.cs.cmu.edu/~schneide/tut5/node42.html>. (Citado en la página 106.)
- [5] *Mcf7.* <http://www.mcf7.com/>. (Citado en la página 43.)
- [6] *Roche 454 sequencing.* <http://www.yourgenome.org/facts/what-is-the-454-method-of-dna-sequencing>. (Citado en la página 20.)
- [7] *Secuenciación maxam y gilbert.* <http://binf.snipcademy.com/lessons/dna-sequencing-techniques/maxam-gilbert>. (Citado en la página 17.)
- [8] *Wikigenes.* <http://www.wikigenes.org/>. (Citado en la página 102.)
- [9] *Autoradiografía*, June 2016. <http://www.ehu.eus/biomoleculas/isotopos/auto.htm>. (Citado en la página 18.)
- [10] *Que es la metástasis*, March 2016. <http://www.cancer.org/espanol/cancer/metastasisen huesos/guiadetallada/metastasis-en-los-huesos-what-is-bone-mets>. (Citado en la página 7.)
- [11] Isabel Marzo Almudena Porras: *Apoptosis: una forma controlada de muerte celular*. Sociedad Española de Bioquímica y Biología Molecular - SEBBM, 2010. <http://www.sebbm.es/web/es/divulgacion/rincon-profesor-ciencias/articulos-divulgacion-cientifica/289-apoptosis-una-forma-controlada-de-muerte-cellular>. (Citado en la página 5.)
- [12] Peter M. Siegel Paula D. Bos Weiping Shu Dilip D. Giri Agnes Viale Adam B. Olshen William L. Gerald Andy J. Minn, Gaorav P. Gupta and Joan Massague: *Genes that mediate breast cancer metastasis to lung*. 2005. (Citado en la página 130.)
- [13] Sam Behjati and Patrick S Tarpey: *What is the next generation sequencing?*, June 2016. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841808/>. (Citado en la página 15.)

- [14] Bioconductor: *Affy package*. <https://bioconductor.org/packages/release/bioc/html/affy.html>. (Citado en la página 29.)
- [15] Bioconductor: *Geoquery*. <https://www.bioconductor.org/packages/devel/bioc/html/GEOquery.html>. (Citado en la página 29.)
- [16] Bioconductor: *Limma package*. <https://bioconductor.org/packages/release/bioc/html/limma.html>. (Citado en la página 29.)
- [17] Bioconductor: *Lumi package*. <http://bioconductor.org/packages/release/bioc/html/lumi.html>. (Citado en la página 29.)
- [18] Bioconductor: *Virtualarray packaged for differential expression genes*, March 2016. <http://www.bioconductor.org/packages/2.12/bioc/html/virtualArray.html>. (Citado en la página 4.)
- [19] Applied Biosystem: *Solid sequencing*. <http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html>. (Citado en la página 21.)
- [20] Xi Chen Melinda E. Sanders A. Bapsi Chakravarthy Yu Shyr Brian D. Lehmann, Joshua A. Bauer and Jennifer A. Pietenpol: *Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies*. 2011. (Citado en las páginas 103, 104 y 131.)
- [21] Anil K. Bera Carlos M. Jarque: *Efficient tests for normality, homoscedasticity and serial independence of regression residuals*. 2002. (Citado en la página 38.)
- [22] Juan Miguel Castagnino: *Electroforesis capilar*. Acta Biológica Clinica Latinoamericana, 1999. <http://www.aefa.es/wp-content/uploads/2014/04/Electroforesis-capilar.pdf>. (Citado en la página 17.)
- [23] Judith Badner Dandan Zhang Elliot Gershon Li Jin Chao Chen, Kay Grennan and Chunyu Liu: *Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods*, 2011. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3046121/>. (Citado en la página 31.)
- [24] Ali O. Güre Matthew J. Scanlan Julia Karbach Elke Jäger Alexander Knuth Lloyd J. Old Dirk Jäger, Elisabeth Stockert and Yao Tseng Chen: *Identification of a tissue-specific putative transcription factor in breast tissue by serological screening of a breast cancer library*. 2001. (Citado en la página 131.)

- [25] DisGeNET: *Disgenet platform.* <http://www.disgenet.org/web/DisGeNET/menu/search?0>. (Citado en la página 44.)
- [26] F Monville P Finetti J Adélaïde N Cervera S Fekairi L Xerri J Jacquemier D Birnbaum E Charafe-Jauffret, C Ginestier and F Bertucci: *Gene expression profiling of breast cell lines identifies potential new basal markers.* 2005. (Citado en la página 131.)
- [27] Barry Komm Edmund C. Chang, Jonna Frasor and Benita S. Katzenellenbogen: *Impact of estrogen receptor on gene networks regulated by estrogen receptor in breast cancer cells.* 2011. (Citado en la página 132.)
- [28] Ruparel H Edwards JR: *Mass-spectrometry dna sequencing.* <http://www.ncbi.nlm.nih.gov/pubmed/15829234>. (Citado en la página 23.)
- [29] Philippe Dessen Attila Tordai Stefan Michiels Cornelia Liedtke Catherine Richon Kai Yan Bailang Wang Gilles Vassal Suzette Delalage Gabriel N. Hortobagyi W. Fraser Symmans Vladimir Lazar Fabrice Andre, Bastien Job and Lajos Pusztai: *Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array.* 2009. (Citado en la página 131.)
- [30] Jacques Rougemont Emmanuelle Charafe Jauffret Valéry Nasser Béatrice Loriod Jacques Camerlo Rebecca Tagett Carole Tarpin Gilles Houvenaeghel Catherine Nguyen Dominique Marandinchi Jocelyne Jacquemier Rémi Houlgate Daniel Birnbaum François Bertucci, Pascal Finetti and Patrice Viens: *Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy.* 2004. (Citado en las páginas 129 y 130.)
- [31] Nicola L.C. Talbot Gavin C. Cawley: *Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers.* (Citado en la página 106.)
- [32] GeneCards: *Bex2.* <http://www.genecards.org/cgi-bin/carddisp.pl?gene=BEX2&keywords=bex2>. (Citado en la página 132.)
- [33] GeneCards: *Clca2.* <http://www.genecards.org/cgi-bin/carddisp.pl?gene=CLCA2&keywords=clca2>. (Citado en la página 129.)
- [34] Universidad las Palmas de Gran Canarias: *Prueba de bondad de ajuste de kolmogorov-smirnov,* 2016. http://www2.ulpgc.es/hege/almacen/download/5/5015/Complemento_3_Prueba_de_Bondad_de_Ajuste_de_Kolmogorov_Smirnov.pdf. (Citado en la página 30.)
- [35] Frank E. Grubbs: *Procedures for detecting outlying observations in samples.* 2012. <http://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657>. (Citado en la página 38.)

- [36] Nancy Lan Guo: *Gene signature for diagnosis and prognosis of breast cancer and ovarian cancer.* 2007. (Citado en la página 132.)
- [37] Jolanthe Kropidlowski Isabell Witzel Karin Milde Langosch Guido Sauter Manfred Westphal Katrin Lamszus Klaus Pantel Harriet Wikman, Bettina Sielaff-Frimpong: *Clinical relevance of loss of 11p15 in primary and metastatic breast cancer: Association with loss of prkcdbp expression in brain metastases.* 2012. (Citado en la página 130.)
- [38] W Ryan Diver Weining Tang Alpa V Patel Victoria L Stevens Eugenia E Calle Michael J Thun Heather Spencer Feigelson, Lauren R Teras and Mark Bouzyk: *Genetic variation in candidate obesity genes adrb2, adrb3, ghrl, hsd11b1, irs1, irs2, and shc1 and risk for breast cancer in the cancer prevention study ii.* 2008. (Citado en la página 129.)
- [39] Marc M. Van Hulle: *Self-organizing maps.*
- [40] Illumina: *Solexa sequencing.* <http://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/> (Citado en la página 21.)
- [41] S. Meganck J. Taminau, C. Lazar and A. Nowe: *Comparison of margin and meta-analysis as alternative approaches for integrative gene expression analysis.* International Scholarly Research Notes, vol. 2014, 2014. (Citado en la página 29.)
- [42] H An E Klopocki F Wiesmann B Betz O Galm O Camara M Dürst G Kristiansen C Huszka R Knüchel J Veeck, D Niederacher and E Dahl: *Aberrant methylation of the wnt antagonist sfrp1 in breast cancer is associated with unfavourable prognosis.* (Citado en la página 104.)
- [43] Shinya Okada Nobuko Fujiuchi Takao Ohtsuka Jennifer C Kwak Yi Wang Ricky W Johnstone Chuxia Deng Jun Qin Jason A Aglipay, Sam W Lee and Toru Ouchi: *A member of the pyrin family, ifi16, is a novel brca1-associated protein involved in the p53-mediated apoptosis pathway.* 2003. (Citado en la página 130.)
- [44] Madhumita Pradhan Yang Dai Lance D. Miller Chin Yo Lin Jonna Frasor, Aisha Weaver and Adina Stanculescu: *Positive cross-talk between estrogen receptor and nf-kb in breast cancer.* 2009. (Citado en la página 132.)
- [45] Mina J. Bissell Marcia V. Fournier Katherine J. Martin, Denis R. Patrick: *Prognostic breast cancer signature identified from 3d culture model accurately predicts clinical outcome across independent datasets.* (Citado en la página 103.)
- [46] Cold Spring Harbor Laboratory: *Celera genomics*, June 2016. <https://www.dnalc.org/view/15571-Celera-Genomics.html>. (Citado en la página 19.)
- [47] Cold Spring Harbor Laboratory: *Sanger method of dna sequencing.* Cold Spring Harbor Laboratory, June 2016. <https://www.dnalc.org/view/>

- 15479-Sanger-method-of-DNA-sequencing-3D-animation-with-narration.html. (Citado en la página 15.)
- [48] Daniel T. Larose: *k-nearest neighbor algorithm*. 2005. (Citado en la página 105.)
- [49] Meganck S. Taminau J. Steenhoff D. Coletta A. Molter C. Lazar, C.: *Batch effect removal methods for microarray gene Expression data integration: a survey*. *Briefings in bioinformatics*. 2013. (Citado en la página 35.)
- [50] P Nuciforo M A Vigano M Capra M Bianchi D Nicosia F Bianchi V Galimberti G Viale G Palermo A Riccardi R Campanini M G Daidone M A Pierotti S Pece M Vecchi, S Confalonieri and P P Di Fiore: *Breast cancer metastases are molecularly distinct from their primary tumors*. (Citado en la página 104.)
- [51] Miguel Vega Marta López, Paloma Mallorquín: *Microarrays y biochips de adn*, March 2016. <http://www.cecalc.ulb.ve/bioinformatica/BIOTUTOR/Microarrays.pdf>. (Citado en la página 4.)
- [52] Hongxia Sun Jeffrey A Drake Sally Gaddis Keith Baggerly Aysegul Sahin Martin C Abba, Yuhui Hu and C Marcelo Aldaz: *Gene expression signature of estrogen receptor status in breast cancer*. 2005. (Citado en la página 103.)
- [53] Etienne Dardenne Sophie Germann Samaan Samaan Rosette Lidereau Kelouma Driouch Pierre de la Grange Martin Dutertre, Lise Gratadou and Didier Auboeuf: *Estrogen regulation and physiopathologic significance of alternative promoters in breast cancer*. 2010. (Citado en la página 130.)
- [54] Matlab: *Bioinformatics toolbox*. <http://es.mathworks.com/products/bioinfo/>. (Citado en la página 60.)
- [55] MedMol: *Adn polimerasa*, June 2016. <http://medmol.es/glosario/65/>. (Citado en la página 16.)
- [56] MedMol: *Electroforesis en gel*, June 2016. <http://medmol.es/tecnicas/68/>. (Citado en la página 18.)
- [57] MedMol: *Pirosecuenciación*, June 2016. <http://medmol.es/tecnicas/85/>. (Citado en la página 20.)
- [58] Federico Morán: *Secuenciación shotgun*. <http://www.lavozdelaiciencia.com/sin-categoría/la-era-de-la-postgenómica-2.html>. (Citado en la página 19.)
- [59] Jatin K. Nagpal Young Kwang Chae Xiaofei Chang Yiping Huang Tony Chuang Keishi Yamashita Barry Trink Edward A. Ratovitski Joseph A. Califano David Sidransky Myoung Sook Kim, Cinthia Lebron: *Methylation of the dfna5 increases risk of lymph node metastasis in human breast cancer*. 2008. (Citado en la página 132.)

- [60] NCBI: *Gpl10558*. <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL10558>. (Citado en la página 77.)
- [61] NCBI: *Gpl570*. <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gpl570>. (Citado en la página 43.)
- [62] NCBI: *Gpl6883*. <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6883>. (Citado en la página 77.)
- [63] Yoichiro Matsuoka Norihisa Uehara and Airo Tsubura: *Mesothelin promotes anchorage-independent growth and prevents anoikis via extracellular signal-regulated kinase signaling pathway in human breast cancer cells*. 2008. (Citado en la página 132.)
- [64] Dirk Timmerman Yves Moreau Olivier Gevaert, Frank De Smet and Bart De Moor: *Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks*. 2006. (Citado en la página 131.)
- [65] Sociedad Española de Oncología Médica: *Las cifras del cáncer en 2014*, March 2016. http://www.seom.org/seomcms/images/stories/recursos/Las_cifras_del_cancer_2014.pdf. (Citado en la página 7.)
- [66] World Health Organization: *Statistics about cancer in 2012*, March 2016. <http://www.who.int/mediacentre/factsheets/fs297/en/>. (Citado en la página 7.)
- [67] Simon Lin Pan Du: *Robust spline normalization between chips*, 2016. <http://svitsrv25.epfl.ch/R-doc/library/lumi/html/rsn.html>. (Citado en la página 32.)
- [68] Yu Y Yao W Bose S Karlan BY Giuliano AE Cui X Qu Y, Han B: *Evaluation of mcf10a as a reliable model for normal human mammary epithelial cells*. 2015. (Citado en la página 43.)
- [69] Florence A. Scholl Claus W. Heizmann Beat W. Schäfer Roland Wicki, Cornelia Franz: *Repression of the candidate tumor suppressor gene s100a2 in breast cancer is mediated by site-specific hypermethylation*. 2004. (Citado en la página 103.)
- [70] Organización Mundial de la Salud: *Salud de la mujer*, March 2016. <http://www.who.int/mediacentre/factsheets/fs334/es/>. (Citado en la página 7.)
- [71] Julio Sergio Santana Sepulveda: *El arte de programar en r*. https://cran.r-project.org/doc/contrib/Santana_El_arte_de_programar_en_R.pdf. (Citado en la página 28.)
- [72] Gloria Luciani Stacey Champion Zhenhang Meng Lakshmi R. Jakkula Heidi S. Feiler Joe W. Gray Shanaz H. Dairkee, Aejaaz Sayeed and Dan H. Moore: *Immutable functional attributes of histologic grade revealed by context-independent gene expression in primary breast cancer cells*. 2009. (Citado en la página 130.)

- [73] Masao Tanji Ryoiti Kiyama Shunichi Terasaka, Akio Inoue: *Expression profiling of estrogen-responsive genes in breast cancer cells treated with alkylphenols, chlorinated phenols, parabens, or bis- and benzoylphe-*nols for evaluation of estrogenic activity. 2006. (Citado en la página 131.)
- [74] Vincent Tak Kwong Chow Rongxian Jin J. Louise Jones Puay Hoon Tan Anita Jayasurya Boon Huat Bay Siew-Kian Tai, Owen June-Keong Tan: *Differential expression of metallothionein 1 and 2 isoforms in breast cancer lines with different invasive potential : Identification of a novel nonsilent metallothionein-1h mutant variant.* 2003. (Citado en la página 129.)
- [75] Takashi HIRANO Soichiro SAITO, Keiko MORITA: *High frequency of common dna copy number abnormalities detected by bacterial artificial chromosome array comparative genomic hybridization in 24 breast cancer cell lines.* 2009. (Citado en la página 132.)
- [76] Laerd Statistics: *Z score*, 2014. <https://statistics.laerd.com/statistical-guides/standard-score.php>. (Citado en la página 37.)
- [77] Saori Furuta Marc E. Lenburg Paraic A. Kenny Ren Xu Sun-Young Lee, Roland Meier and Mina J. Bissell: *Fam83a confers egfr-ki resistance in breast cancer cells and in mice.* 2012. (Citado en la página 131.)
- [78] Steenhoff D. Coletta A. Meganck S Lazar C. Taminau, J.: *inSilicoDB: an R/Bioconductor package for accessing human Affymetrix expert-curated datadata from GEO.* Bioinformatics. 2011. (Citado en la página 38.)
- [79] Oxford Nanopore Technologies: *Nanopore sequencing.* <https://nanoporetech.com/applications/dna-nanopore-sequencing>. (Citado en la página 22.)
- [80] Akeila Bellahcène Nadia Rucci Soraya Sin Berta Martin Abad Angels Sierra Alain Boudinet Jean Marc Guinebretière Enrico Ricevuto Catherine Noguès Marianne Briffod Ivan Bièche Pascal Cherel Teresa Garcia Vincent Castronovo Anna Teti Rosette Lidereau Thomas Landemaine, Amanda Jackson and Kelouma Driouch: *A six-gene signature predicting breast cancer lung metastasis.* 2008. (Citado en la página 132.)
- [81] Utah State University: *Introduction to preprocessing: Rma (robust multi-array average)*, 2014. http://math.usu.edu/jrstevens/stat5570/1.4.Preprocess_4up.pdf. (Citado en la página 31.)
- [82] Mohamed Mokhtar Desouki Ping Liang Andrei Bakin Kumarasamy Thangaraj Donald J. Buchsbaum Albert F. LoBuglio Keshav K. Singh Vanniarajan Ayyasamy, Kjerstin M. Owens: *Cellular model of warburg effect identifies tumor promoting function of ucp2 in breast cancer and its suppression by genipin.* 2011. (Citado en las páginas 130 y 131.)

- [83] Biografías y Vidas: *Frederick sanger*, June 2016. <http://www.biografiasyvidas.com/biografia/s/sanger.htm>. (Citado en la página 16.)
- [84] Wikipedia: *Canales iónicos*. https://es.wikipedia.org/wiki/Canal_ionico. (Citado en la página 22.)
- [85] Wikipedia: *Html*. <https://es.wikipedia.org/wiki/HTML>. (Citado en la página 29.)
- [86] Wikipedia: *Multiplex polony sequencing*. https://en.wikipedia.org/wiki/Polony_sequencing. (Citado en la página 23.)
- [87] Wikipedia: *Variance-stabilizing transformation*, 2014. https://en.wikipedia.org/wiki/Variance-stabilizing_transformation. (Citado en la página 31.)
- [88] Wikipedia: *Allan maxam*, June 2016. https://en.wikipedia.org/wiki/Allan_Maxam. (Citado en la página 17.)
- [89] Wikipedia: *Cinasas*, June 2016. <https://es.wikipedia.org/wiki/Cinasa>. (Citado en la página 18.)
- [90] Wikipedia: *Contador de coulter*, June 2016. https://es.wikipedia.org/wiki/Contador_Coulter. (Citado en la página 22.)
- [91] Wikipedia: *Craig venter*, June 2016. https://es.wikipedia.org/wiki/Craig_Venter. (Citado en la página 19.)
- [92] Wikipedia: *False discovery rate*, 2016. https://en.wikipedia.org/wiki/False_discovery_rate. (Citado en la página 32.)
- [93] Wikipedia: *Fosforo gamma p32*, June 2016. <https://en.wikipedia.org/wiki/Phosphorus-32>. (Citado en la página 18.)
- [94] Wikipedia: *Phi x 174*, June 2016. https://en.wikipedia.org/wiki/Phi_X_174. (Citado en la página 16.)
- [95] Wikipedia: *Walter gilbert*, June 2016. https://es.wikipedia.org/wiki/Walter_Gilbert. (Citado en la página 17.)
- [96] Elizabeth Richardson Mark Erlander Xiao-Jun Ma, Sonika Dahiya and Dennis C SgroiEmail author: *Gene expression profiling of the tumor microenvironment during breast cancer progression*. 2009. (Citado en la página 130.)
- [97] Yongquan Shen Hitoshi Ichikawa Jonathan Jarvik Robert G. Nagele Gary S. Goldberg Xun Li, Zhenyu Jia: *Coordinate suppression of sdpr and fhl1 expression in tumors of the breast, kidney, and prostate*. 2008. (Citado en la página 132.)
- [98] Prof Jan GM Klijn MD Yi Zhang PhD Anieta M Sieuwerts BSc Maxime P Look MSc Fei Yang MSc Dmitri Talantov MD Mieke Timmermans BSc Marion E Meijer van Gelder MD Jack Yu PhD Tim Jatkoe BSc Els MJJ Berns PhD David Atkins PhD Dr John A Foekens PhD Yixin Wang, PhD: *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*. 2005. (Citado en la página 129.)

- [99] Huiqi Li Zhengyang Chin Yuanqing Li, Cuntai Guan: *A self-training semi-supervised svm algorithm and its application in an eeg-based brain computer interface speller system.* 2008. (Citado en la página 105.)

Parte V

ANEXOS

6

ANEXO A

Como pudo verse en el apartado 4, se destacaron un gran número de genes, en total 52, que eran independientes a la tecnología y métodos de unión y eliminación del efecto batch usados. Por lo tanto en este anexo, se explicarán aquellos genes que no se quedaron explicados y se intentarán relacionar con el cáncer de mama. Como todos los genes aquí destacados han tenido un gran valor de fold-change y un p-value inferior a 0.001 puede ser interesante estudiar o resaltar dichos genes si alguno de ellos no estuviese ya en la literatura.

- ADRB2: Las diferentes formas polimórficas, mutaciones puntuales, y / o regulación a la baja de este gen están asociadas con el asma nocturna, la obesidad y la diabetes tipo 2. Esta asociado según un estudio al riesgo de padecer cáncer de mama debido a la obesidad y en la postmenopausia [38].
- DSG3: La proteína codificada por este gen se ha identificado como el autoantígeno de las ampollas autoinmunes en la enfermedad pénfigo vulgar. Se ha encontrado relación entre este gen y algunos tumores de cáncer de mama [74].
- CLCA2: Las enfermedades asociadas con CXCL2 incluyen fibrosis quística y cáncer de mama. Entre sus vías relacionadas son la secreción pancreática y la secreción de renina [33].
- SLPI: Este gen codifica un inhibidor secretado que protege los tejidos epiteliales de proteasas de serina. Las enfermedades asociadas con SLPI incluyen telangiectasia, hemorragia hereditaria y la bronquitis. Se han encontrado estudios que mantienen una relación de este gen con el cáncer de mama [30].
- C3: Las enfermedades asociadas con C3 incluyen la degeneración macular, relacionada con la edad y deficiencia de C3. Se ha encontrado relación de este con un tipo de cáncer de mama [98].
- HENMT1: Entre sus vías relacionadas son la Expresión Génica y mitótico Profase. En principio, no se han encontrado estudios que relacionen este gen con el cáncer de mama.

- CXCL1: Este gen codifica una proteína, la expresión aberrante de esta proteína se asocia con el crecimiento y la progresión de ciertos tumores. Las enfermedades asociadas con CXCL1 incluyen la artritis reumatoide y la gastritis. Se han encontrado estudios que asocian este gen a la metástasis de este cáncer hacia el pulmón [12].
- COL17A1: Las mutaciones en este gen están asociadas con epidermolisis ampollosa generalizada y con la unión atrófica benigna. Se han encontrado estudios que mantienen una relación de este gen con el cáncer de mama [30].
- PRKCDBP: Se ha comprobado que la expresión de esta proteína está inhibida en varias líneas celulares de cáncer, lo que sugiere la posible función supresora de tumores de esta proteína. Se ha encontrado relación de este gen con la metástasis en el cáncer de mama hacia el cerebro [37].
- UCP2: Las enfermedades asociadas con UCP2 son la obesidad y la hipoglucemia. Entre sus rutas relativas son la glucosa / energía del metabolismo y el metabolismo. Se ha encontrado cierta relación con el cáncer de mama [82].
- EFHD1: La proteína que codifica muestra aumento de la expresión durante la diferenciación neuronal. Se han encontrado estudios que relacionan este gen con el cáncer de mama [53].
- RGS2: Una enfermedad asociada con el gen RGS2 es el síndrome de Rieger. Ya se había reportado relación entre este gen y el cáncer de mama [96].
- IFI16: Las enfermedades asociadas con IFI16 incluyen herpes simplex y la esclerosis sistémica cutánea difusa. Se han encontrado estudio que demuestran relación entre el gen y la enfermedad [43].
- ZBTB16: Las enfermedades asociadas con ZBTB16 incluyen defectos esqueléticos, hipoplasia genital y disminución mental. No se han encontrado estudios que relacionen este gen con el cáncer de mama.
- DNER: No se han encontrado relaciones entre la enfermedad y dicho gen.
- GNA15: Las enfermedades asociadas con GNA15 incluyen la tos ferina y la enfermedad infecciosa bacteriana primaria. Se ha hallado relación entre este gen y la enfermedad estudiada [72].
- PNLRP3: Las enfermedades asociadas con PNLRP3 incluyen carcinoma hepatocelular. No se ha encontrado relación entre el cáncer de mama y este gen.

- KRT6B: Las enfermedades asociadas con KRT6B incluyen paquioniquia congénita 4 y paquioniquia congénita 2. Se ha encontrado relación entre este gen y la enfermedad en estudios [20].
- BNC1: Las enfermedades asociadas con BNC1 incluyen empiema subdural. Se han hallado relaciones entre este gen y el cáncer de mama [26].
- FBP1: Las enfermedades asociadas con el gen FBP1 son la deficiencia de fructosa-1, bisfosfataza-6 y la hipoglucemia. Se ha encontrado relación entre la enfermedad y este gen [73].
- RAB38: Se ha encontrado relación entre el gen y la enfermedad [24].
- TSPYL5: Participa en la modulación del crecimiento celular y la respuesta celular a la radiación gamma. Se han encontrado estudios que relacionan este gen con la enfermedad [64].
- NMU: Una enfermedad asociada con NMU incluye leiomioma traqueal. No se han encontrado aparentemente estudios que relacionen dicho gen con la enfermedad.
- MAOA: Las enfermedades asociadas con MAOA incluyen el síndrome de Brunner y trastornos del comportamiento relacionados con la MAOA. Se han encontrado relaciones entre este gen y el cáncer de mama [82].
- EVA1C: No se ha encontrado relación entre este gen y la enfermedad estudiada.
- GPR87: La proteína codificada por este gen ha demostrado que se sobreexpresa en el carcinoma de células escamosas de pulmón. Hay estudios que proponen relación entre este gen y la enfermedad [29].
- CPVL: Una enfermedad asociada con CPVL incluye vitreoretinopatía exudativa. No hay estudios que relacionen este gen con el cáncer de mama.
- CBS: La enfermedad asociada con la CBS incluye homocistinuria debido a la deficiencia de CBS. No se ha encontrado relación entre dicho gen y la enfermedad.
- CASP1: Las enfermedades asociadas con CASP1 incluyen la viruela y la legionelosis. Se han encontrado relación entre este gen y la enfermedad [29].
- FAM83A: Probable proto-oncogén que funciona en el receptor del factor de crecimiento epidérmico. Se han encontrado estudios que relacionan este gen con el cáncer de mama [77].

- SDPR: Se han encontrado relaciones entre este gen y el cáncer de mama [97].
- MSLN: Las enfermedades asociadas con MSDN incluyen el cáncer de ovario y el mesotelioma pleural maligno. Estudios revelan relaciones entre este gen y la enfermedad [63].
- DFNA5: Las enfermedades asociadas con DFNA5 incluyen sordera, autosómica dominante, pérdida auditiva no sindrómica y sordera. Se ha encontrado relación en estudios entre la enfermedad y el gen [59].
- TNNT2: Las enfermedades asociadas con TNNT2 incluyen artritis distal y artrogriposis múltiple congénita. No se han encontrado relaciones entre dicho gen y la enfermedad.
- IRX4: La enfermedad asociada con Irx4 incluye fibrilación auricular familiar.
- TFCP2L1: Las funciones relacionadas con este gen incluyen la actividad del factor de transcripción, unión a ADN específica de secuencia y la actividad de represor transcripcional. Se han encontrado relaciones entre este gen y la enfermedad [80].
- BEX2: Una enfermedad asociada con BEX2 incluye el cáncer de mama como se puede en la plataforma GeneCards [32].
- BIRC3: Las enfermedades asociadas con BIRC3 incluyen el linfoma, malta, somático y linfomas de células B. Se han encontrado relaciones entre este gen y la enfermedad a estudiar [44].
- INHBB: Se ha demostrado que regula la proliferación de células del estroma gonadal negativamente y que tienen actividad supresora de tumores. Guarda relación con la enfermedad a estudiar según algunos estudios [27].
- SLC26A2: Las enfermedades asociadas con SLC26A2 incluyen acondrognésis y displasia epifisaria múltiple. No hay estudios que relacionen dicho con la enfermedad.
- C3orf14: Parece ser un posible gen relacionado con el cáncer de mama según los estudios [75].
- ACOT4: Existe estudios que relacionan dicho gen con la enfermedad estudiada [36].

