UNIVERSIDAD NACIONAL AUTONÓMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES
ACATLÁN

ANÁLISIS COMPARATIVO DE ALGORITMOS DE NORMALIZACIÓN EN DATOS DE MICROARREGLOS DE ALTA DENSIDAD

TESIS

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

PRESENTA

IVÁN IMAZ ROSSHANDLER

ASESOR: MTRO. VICTOR JOSÉ PALENCIA GÓMEZ

ASESOR EXTERNO: DRA. CLAUDIA RANGEL ESCAREÑO

Fecha: Mayo 2012

AGRADECIMIENTOS

Esta tesis va dedicada a mis padres Ana Rosshandler y Jorge Imaz, y abuelas Berta Fernández, Dolores Fernández y Berta Lira. Así mismo, agradezco a todos los profesores de la FES Acatlán y la Dra. Claudia Rangel Escareño por haberme apoyado en el desarrollo de este trabajo y permitirme incursionar en la Biología Computacional.

RECONOCIMIENTOS

Este trabajo fue desarrollado con el apoyo del Instituto de Ciencia y Tecnología del Distrito Federal a través del fondo de fomento al uso de tecnologías de punta en la investigación científica y tecnológica del Gobierno del Distrito Federal 2011 (ICYTDF 046/2011), dentro del proyecto PIUTE10-104.

JUSTIFICACIÓN

La tecnología de microarreglos ofrece un mayor detalle y mayor alcance en los estudios moleculares. Esto a su vez se traduce en volúmenes masivos de datos, provenientes de un número realmente restringido de muestras. Por tanto, el diseño de nuevos algoritmos que se adapten a un esquema de muchas variables pocas repeticiones ha sido crucial en el campo de la investigación genómica. Al mismo tiempo el volumen de datos representa un reto para quienes los almacenan y los procesan implicando un estudio detallado de la complejidad de cada algoritmo a utilizar.

La normalización, es una etapa fundamental del análisis de datos de microarreglos ya que permite realizar las comparaciones requeridas entre las diferentes muestras. Sin embar-

go, no existe un método de normalización que garantice un desempeño óptimo en todo tipo de microarreglos o muestras. Entonces, el investigador debe desarrollar un criterio de selección que le permita elegir el método que conlleve a mejores resultados para un determinado experimento.

Este proyecto tiene por finalidad realizar un análisis comparativo de algunos de los métodos de normalización más importantes y con ello, servir de referencia para aquellos investigadores sin experiencia en esta tecnología. Además, cabe mencionar que la mayor parte de la información relevante al tema se encuentra en inglés, de forma que un documento de esta índole en español resulta ser de utilidad para la incorporación de nuevos investigadores a esta tecnología en el país.

PLANTEAMIENTO DEL PROBLEMA

Dada la complejidad del diseño de microarreglos, el análisis de estos datos no es una tarea sencilla y no existe un método que sea un estándar de oro. Es así, que el investigador se ve obligado a desarrollar un procedimiento adecuado a cada experimento específico.

Durante el curso del análisis, la normalización de los datos juega un papel determinante. Sin embargo, existe una amplia gama de algoritmos de normalización, cada uno con diversas ventajas y desventajas de tal manera que la elección de alguno de éstos puede ser complicada, principalmente para los investigadores sin experiencia en esta tecnología o sin formación académica en áreas cuantitativas. Es por tanto, que este trabajo pretende realizar un análisis sistemático selectivo sobre los algoritmos de normalización más eficientes de acuerdo a la bibliografía y el desarrollo de un caso de aplicación.

HIPÓTESIS

No existe un algoritmo de normalización óptimo para todos los posibles experimentos. Sin embargo, el método de Cuantiles es muy eficiente de acuerdo a los criterios de selección

más comunes, es el más usado en la literatura y por tanto podría ser el estandar de oro en análisis de datos de microarreglos de expresión.

OBJETIVO GENERAL

Determinar el método de normalización más eficiente para el análisis de datos de microarreglos en una muestra de ratón transgénico $K14E6$, utilizada para el estudio del perfil de expresión diferencial en piel y cérvix.

OBJETIVOS ESPECÍFICOS

- Comprender los fundamentos moleculares de los experimentos con microarreglos y del análisis de datos provenientes de éstos, haciendo enfasis en el proceso de normalización y sus implicaciones en los resultados. Las fases del análisis involucradas son las siguientes:
  - Pre-procesamiento
    - Control de calidad
    - Corrección de fondo
    - Normalización
    - Generación de niveles de expresión
  - Clasificación
  - Predicción
- Utilizando los recursos de *bioconductor*, generar el código en R necesario para desarrollar el análisis de los datos del experimento seleccionado a través de la generación de diferentes tipos de gráficos.
- Por medio de la generación de diferentes tipos de gráficos MA, de distribución y de cajas comparar el desempeño de los diferentes algoritmos de normalización, siempre teniendo en cuenta el comportamiento de los datos crudos.

- Una vez obtenidos los genes diferencialmente expresados y sus respectivos mapas de calor, relacionar los resultados con base en los diferentes métodos de normalización usados. Los métodos bajo estudio son los siguientes:
    - Cuantiles
    - Loess cíclico
    - Contraste
    - Conjuntos invariantes
    - Qspline
    - Constante

Índice general

Capítulo 1

Introducción

1.1. Genómica

La decodificación del genoma humano, marcó el inicio de una nueva era en la investigación genética. Al mismo tiempo, en gran parte gracias a los avances tecnológicos, se han desarrollado técnicas cada vez más finas y precisas que permiten un análisis más completo de los intrincados procesos de expresión y regulación del genoma. Así, se evolucionó desde el estudio de genes aislados de un organismo al de un conjunto de éstos, dando cabida a una nueva disciplina: la genómica.

1.2. El impacto de la ciencia físico-matemática sobre el origen de la biología molecular

1.2.1. Los primeros pasos

La gran migración de intelectuales hacia los Estados Unidos e Inglaterra durante los años treinta, junto al viraje de muchos físicos teóricos hacia la biología, sobre todo después de la segunda guerra mundial, constituyó el recurso humano que contribuyó al cambio del enfoque de los problemas y soluciones de esta ciencia[1]. La teoría de la relatividad y la mecánica cuántica proporcionaron avances increíbles en el periodo entre las dos guerras

mundiales, principalmente bajo la dirección de Niels Bohr en Copenhague[2]. Para algunos, era convincente que los problemas no resueltos de la física sobre la materia inerte eran pocos. Sin embargo, los problemas sobre estructuras y procesos biológicos eran aún un enigma, por lo que se tornó un campo novedosamente atractivo para la investigación científica.

En 1932 Niels Bohr formuló la idea de que algunos fenómenos biológicos no podían ser explicados totalmente en términos de conceptos físicos convencionales, del mismo modo que la teoría cuántica del átomo no podía ser deducida de la mecánica clásica[2]. Esta idea fue llevada al campo de la genética por uno de sus discípulos, Max Delbruck, un joven alemán que emigró a Estados Unidos; en 1935 publicó un artículo titulado *Sobre la naturaleza de las mutaciones y la estructura de los genes*, en el cual señalaba algunas propiedades de lo vivo, en especial, de las moléculas genéticas que no podían ser deducidas a partir de la física y la química tradicionales[3]. Por ejemplo, la incapacidad para describirlas en términos absolutos (ya que cada gen es diferente y está interrelacionado con otros). A Delbruck, le llamaba principalmente la atención la estabilidad de dichas moléculas a lo largo de millones de años, pese a la gran cantidad de reacciones metabólicas que se llevan a cabo en la célula.

Estas ideas lograron una amplia difusión con la publicación en 1945 del libro *¿Qué es la vida?* escrito por el físico Erwin Schrodinger[4]. Éste tuvo una notable influencia sobre los precursores de la biología molecular. Schrodinger afirmaba que a pesar de que los organismos eran inmensos comparados con los átomos, no había ninguna razón por la cual no obedecieran a leyes físicas exactas. Según él, el problema real que requería una explicación era la herencia y la pregunta específica ¿cómo es que los genes, que no son muy grandes comparados con los átomos, resisten las fluctuaciones tan grandes a las que están sujetos? Postulaba que los genes eran capaces de preservar su estructura porque el cromosoma en el que se encontraban era un cristal aperiódico, o molécula formada por isómeros repetitivos que conformarían un código genético similar al código Morse. Esta especulación, derivada de una forma cuántica de pensar (probabilística, en términos

de estabilidad atómica) fue el primer acercamiento a una posible solución cuantitativa del problema de los genes. Además, propuso que si bien estas moléculas podrían no contravenir las leyes de la la física y la química, sí era posible que su estudio condujera al descubrimiento de nuevas leyes y junto a ello, nuevas herramientas matemáticas.

1.2.2. El ADN contiene la información genética

En 1939 Delbruck junto a Emory Ellis en el Instituto Tecnológico de California, trabajó en la demostración del ciclo de vida por etapas de un bacteriófago. Publicaron el artículo sobre *La multiplicación en un paso* de los bacteriofagos en cultivos bacterianos, dejando claro que a una escala molecular había muchas preguntas por hacer y responder, que la genética clásica ni siquiera había postulado[5]. En 1940, pasó a la Universidad de Vandervilt, donde continuó la investigación acerca de la relación fago-bacteria, pues representaba un modelo sencillo para estudiar los problemas de la reproducción y la herencia. A finales de 1940, conoció a S.E. Luria que también trabajaba con fagos. Rápidamente y en conjunto, desarrollaron experimentos que fueron parte determinante en el desarrollo de la biología molecular. Más adelante, el microbiólogo Alfred Hershey se uniría al equipo para dar forma al llamado *grupo del fago*[1]. En 1943, restringiéndose a un modelo con la cepa B de *Escherichia Coli* y los siete fagos *T*, mediante la *prueba de fluctuación*[7], demostraron que las mutaciones aparecían espontáneamente y entonces podían ser seleccionadas darwinianamente; además, éste fue el comienzo de una serie de aplicaciones metodológicas de tipo estadístico y matemático a la biología.

El acercamiento a la naturaleza química del gen provino de las investigaciones de tipo médico y bacteriológico, que en 1944 Avery, MacLeon y Mcarty publican en las conclusiones de su trabajo sobre transformación de neumococos, en las que por primera vez se atribuye al ADN el papel de transmisor de la información genética. Sin embargo, aunque su investigación dio acercamiento original y suficiente al problema, sus conclusiones no fueron totalmente aceptadas por la comunidad científica. Un factor determinante fue que las técnicas químicas de aislamiento y purificación de ese entonces, siempre dejaban en duda si el ADN no estaría contaminado con pequeñas cantidades de proteína, las cua-

les efectuarían la transformación de los neumococos ya que se creía que la información genética estaba en dichas moléculas. Estos resultados fueron aceptados hasta 1952 gracias al *experimento de la licuadora* desarrollado por dos miembros del grupo del fago, Martha Chase y Alfred Hershey, utilizando las nuevas técnicas de isótopos radioactivos desarrolladas por los físicos durante la segunda guerra mundial[6].

1.2.3. La solución a la estructura del ADN

Desde 1920, se creía que el ADN era una estructura tetranucleótida repetida y monótona. En 1951, motivado por una fotografía del ADN presentada en un congreso en Nápoles, Watson, convencido del caŕacter regular de la molécula y por tanto, de la posibilidad de que fuera el cristal aperiódico transimsor de la información genética del que hablaba Schrodinger; llega al laboratorio Cavendish dirigido por Sir Lawrence Bragg para aprender cristalografía[6]. La cristalografía es una técnica utilizada para inferir la disposición espacial de los átomos en una molécula a partir de un patrón de rayos X desviados por el cristal hacia una placa fotográfica. En su arribo a Cavendish, Watson se une a Crick, un físico con amplios conocimientos de cristalografía interesado en la estructura de alfa hélice de las proteínas desarrollada por L. Pauling[8], que lo llevó a crear una teoría cristalográfica helicoidal[9]. En 1953, después de un análisis exhaustivo, Watson y Crick lograron solucionar la estructura del ADN, ajustada a las evidencias empíricas y a las necesidades funcionales de éste (figura 1.1).

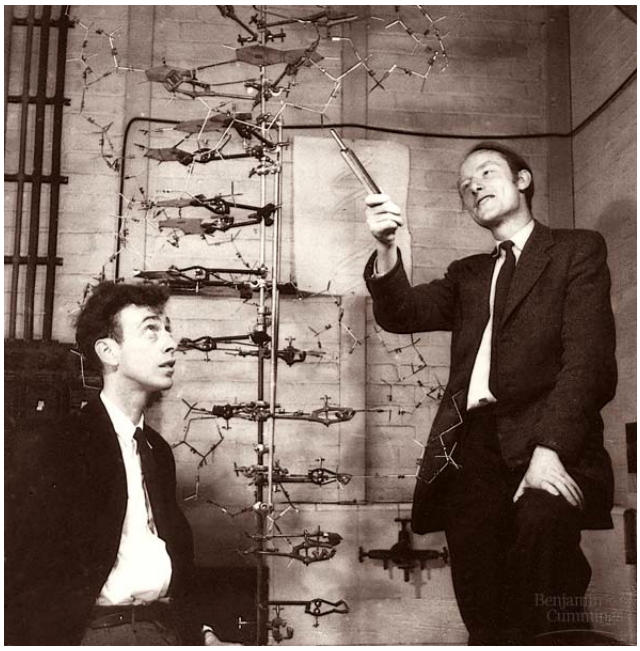

Figura 1.1: **Retrato de James Watson (izquierda) y Francis Crick (derecha) junto al modelo tridimensional de la molécula de ADN propuesto en** 1953 **por ellos mismos**. Este trabajo les hizo acreedores del Premio Nobel de Fisiología en 1962. Imagen de Pepquímic.

1.3. El ADN y el ARN

Watson y Crick descubrieron que la molécula de ADN (ácido desoxirribonucleico) está formada por dos cadenas alargadas que se enrollan formando una doble hélice, análogamente a una larga escalera de caracol. Las cadenas o lados de la escalera, están constituidas por un ácido ortofosfórico y un hidrato de carbono o azúcar denominado desoxirribosa que se alternan a través de ésta. Las bases nitrogenadas, dispuestas en parejas, unidas por enlaces de puente de hidrógeno, representan los escalones. Cada base está unida a una molécula de azúcar y ligada por dos o tres enlaces de hidrógeno a su base complementaria localizada en la cadena opuesta.

Las bases nitrogenadas del ADN se pueden clasificar en dos grupos: las bases púricas, adenina (A) y guanina (G), y las bases pirimidínicas, timina (T) y citosina (C). En la doble hélice, la adenina siempre se vincula con la timina mediante dos puentes de hidrógeno y la guanina con la citosina mediante tres. A cada unidad compuesta por el grupo fosfato, el azúcar desoxirribosa y una de las bases nitrogenadas se le denomina nucleótido. El orden de la secuencia de las bases es muy importante, ya que en él, reside la información contenida en el código genético. La orientación viene dada en el sentido 5'-3' 3'-5'; el 5' representa el extremo terminal del fosfato y el 3' el extremo final del átomo de carbono de la desoxirribosa. Las dos cadenas de la doble hélice tienen sentidos opuestos: mientras una va en sentido 5'-3', la otra lo hace en sentido 3'-5'. Es por ello que se refiere al ADN, como una doble hélice antiparalela[10].

La estructura química de los nucleótidos es lo que se denomina estructura primaria y la forma en doble hélice, que gira en sentido contrario a las agujas del reloj, es decir, hacia la izquierda o levógira, es la estructura secundaria. A su vez, la molécula de ADN sería muy larga para estar contenida en el pequeño núcleo celular de manera que se compacta en una estructura terciaria[10].

En el humano, existen 25 moléculas diferentes de ADN: 24 de ellas constituyen el ADN presente en los núcleos de las células y están contenidas en las 22 unidades de cada pareja de cromosomas, más el cromosoma X y el cromosoma Y; su conjunto se denomina genoma nuclear. La molécula de ADN restante se encuentra en las mitocondrias. El ADN presente en cada cromosoma contiene los genes de cada cromosoma[10].

Si se hace referencia a la estructura primaria, el ARN o ácido ribonucleico se diferencia del ADN en dos cosas[13]:

- El azúcar del nucleótido es la ribosa y no la desoxirribosa.

- Entre las bases nitrogenadas, no aparece la Timina y es sustituida por el Uracilo.

En la célula sobresalen tres tipos de ARN, teniendo cada uno de ellos una función distinta[13]:

1. ARN mensajero (ARNm): Es ARN lineal, de una sola cadena y contiene la información copiada del ADN para sintetizar una proteína. Se forma en el núcleo celular, sale del núcleo y se asocia a los ribosomas, donde se construye la proteína. Cada tres nucleótidos de ARNm forman un codón, cada codón se traduce en un sólo aminoácido (con algunos casos de redundancia), una señal de inicio o una de parada. Así, la secuencia de aminoácidos de la proteína está configurada a partir de la secuencia de los nucleótidos del ARNm.

2. ARN ribosomal (ARNr): El ARN ribosomal, está unido a proteínas de carácter básico, formando en conjunto los ribosomas. Los ribosomas son las estructuras celulares donde se ensamblan aminoácidos para formar proteínas, a partir de la información que transmite el ARN mensajero. Hay dos tipos de ribosomas, el que se encuentra en células procariotas, en el interior de mitocondrias y cloroplastos, y el que se encuentra en el hialoplasma o en el retículo endoplásmico de células eucariotas.

3. ARN de transferencia (ARNt): El ARN de transferencia es un ARN no lineal, en el que se pueden observar tramos de doble hélice intracatenaria, es decir, entre las bases que son complementarias, dentro de la misma cadena. Estas estructuras se estabilizan mediante los puentes de hidrógeno que se establecen entre las bases. Su función es transportar los aminoácidos a los ribosomas.

1.4. El genoma humano

El Proyecto Genoma Humano es una iniciativa internacional que se inició en 1990 y se completó en 2003. El proyecto de 3 billones de dólares fue coordinado por el Departamento de Energía de E.U.A y el Instituto Nacional de la Salud del mismo país. Los objetivos del mismo fueron determinar la secuencia completa de 3 billones de bases, identificar todos

los genes humanos y hacerlo accesible para estudios biológicos posteriores. Asunto que implica el desarrollo de herramientas para el análisis de datos, transferir las tecnologías relacionadas al sector privado y dirigir los aspectos éticos, legales y sociales. El consenso actual predice aproximadamente $20,000 - 25,000$ genes, pero no todos los científicos del genoma están de acuerdo. Durante este proyecto, la secuenciación se realizó paralelamente en modelos de organismos selectos tales como *Escherichia coli* para ayudar al desarrollo de la tecnología y la interpretación de la función de los genes humanos[11].

Importantes contribuciones a este proyecto fueron llevadas a cabo por Celera Genomics. Una empresa estadounidense fundada en mayo de 1998 por Applera Corporation y J. Craig Venter, con el objetivo primario de secuenciar y ensamblar el genoma humano en plazo de tres años. Para ello utilizaron el método *Shotgun*, basado en la rotura del ADN en múltiples trozos, su clonación y búsqueda de solapamientos con aplicaciones bioinformáticas. En el año 2001 presentaron en la revista Science su primer esbozo, 5 genomas de diferentes etnias, entre ellos, se encontraba el de su director, Craig Venter[12].

El trabajo de secuenciación del ADN en su última versión representa un logro enorme, según algunos expertos podría compararse en importancia con el desarrollo de la tabla de los elementos y como en la mayoría de los grandes avances científicos, aún queda mucho trabajo para lograr el potencial completo de la realización. Además, a las exploraciones del genoma humano se han sumado proyectos de genomas de muchos otros organismos, generando datos cuyo volumen y complejidad no tienen precedentes en la biología. Tecnologías de escala genómica se han vuelto necesarias para comparar genomas enteros, conjuntos de ARN expresado o proteínas, familias de genes provenientes de un gran número de especies, la variación entre individuos y las clases de elementos regulatorios de genes.

Derivado del conocimiento de las principales secuencias de ADN, se puede definir el curso de la investigación biológica en las próximas decadas y es requerida la experiencia y creatividad de equipos de biólogos, químicos, ingenieros, matemáticos y científicos de la

computación, entre otros. A continuación, se mencionan algunos de los principales retos que todavía no se resuelven a pesar de tener la secuencia del ADN en la mano[11]:

- Número, ubicación exacta y funciones de los genes.
- Regulación génica.
- Organización de la secuencia de ADN.
- Estructura y organización cromosómica.
- Tipos de ADN no codificador, cantidad, distribución, contenido de información y funciones.
- Coordinación de eventos de la expresión génica, síntesis de proteínas y aquellos posteriores a la traducción.
- La interacción de las proteínas complejas en máquinas moleculares.
- Predicción contra determinación experimental de la función del gen.
- La conservación evolutiva entre los organismos.
- La conservación de proteínas (estructura y función)
- Proteomas (contenido de proteína total y función) en los organismos.
- Correlación de SNPs (variaciones de ADN de una base entre los individuos) con la salud y la enfermedad.
- Predicción de susceptibilidad a la enfermedad con base en la variación de secuencias de genes.
- Los genes involucrados en rasgos complejos y enfermedades multigénicas.
- Biología de sistemas complejos, incluidos los consorcios microbianos útiles para la restauración ambiental.
- El microbioma humano.
- La genética del desarrollo, genómica.

1.5. El dogma central de la biología molecular

En el año de 1953, se adoptó la hipótesis de que el ADN cromosomal funciona como molde para las moléculas de ARN, el cual, subsecuentemente se dirige hacia el citoplasma donde determina el orden de los aminoácidos durante la síntesis de proteínas. En 1956 Francis Crick se refirió a esta ruta metabólica para el flujo de la información genética como el dogma central (figura 1.2)[13]:

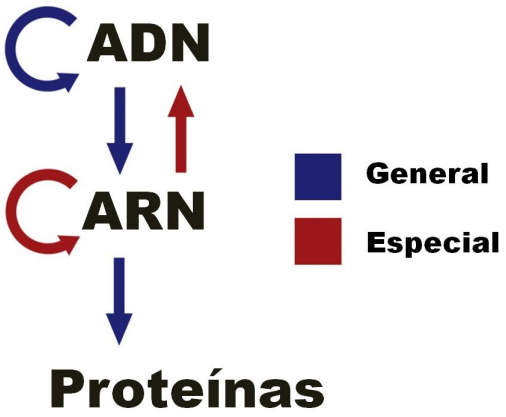

Figura 1.2: **Dogma Central de la Biología Molecular**. La relación general, indica el proceso concebido por primera vez en 1953. La relación especial, denota aquellos casos que se descubrieron posteriormente.

Las flechas de color azul, indican la dirección propuesta para la transferencia de la información genética. La flecha que rodea al ADN implica que esta molécula es el molde para su propia replicación, la que está entre el ADN y ARN índica que la síntesis del ARN (proceso que es llamado transcripción) se realiza directamente a partir de un molde de ADN, así mismo, la síntesis de proteínas (llamada traducción) utiliza un molde de ARN. La traducción es un proceso unidireccional y por lo tanto, nunca ocurre en sentido contrario. Sin embargo, mientras que la idea de que las proteínas nunca sirven de molde para la síntesis de ARN o ADN se mantiene hasta la fecha, se han encontrado casos en los que las cadenas de ARN sirven como molde para la secuencia complementaria de ADN o

ARN; ya que estos son relativamente pocos, el dogma central propuesto hace más de 50 años mantiene su esencia.

Capítulo 2

Tecnología para la Investigación Genómica: Los microarreglos

Una de las técnicas más utilizadas en el campo de la genómica consiste en el uso de microarreglos. Estos son soportes sólidos en los cuales se encuentran inmovilizados en un área pequeña decenas de miles de genes de manera ordenada y pueden ser laminillas de vidrio, cuarzo o de polímeros como silicón y nylon. El ADN fijado en los microarreglos puede ser impreso, depositado o sintetizado directamente sobre la superficie sólida en lugares específicos llamados *spots* o celdas de sondeo y generalmente son oligonucleótidos de 25 pares de bases por cada celda, que representan parte o todos los genes identificados en un organismo. Los microarreglos permiten usar de forma amplia la creciente información de secuencias del genoma para medir de forma paralela y cuantitativa la expresión de los genes a través de su ARNm, con base en el principio de hibridación que las moléculas de ácidos nucleicos obedecen y en el hecho de que los genes activos copian su información por medio de este tipo de ARN.

En 1994 Affymetrix desarrolló los primeros arreglos de alta densidad por medio de técnicas fotolitográficas. Arreglos de menor tamaño se idearon ulteriormente constituyendo los microarreglos de ADN. Los microarreglos modernos pueden contener incluso la cantidad de genes contenidos en el genoma humano junto a algunos otros como los correspon-

dientes controles. De esta forma es posible el estudio del genoma completo de diversas especies en superficies de 1.28cm^2 (figura 2.1). A pesar de su reciente introducción, esta tecnología ha experimentado un gran auge al proveer de una herramienta que permite realizar avances rápidos en el conocimiento de diversas patologías desde el punto de vista de la biología molecular.

Figura 2.1: **Microarreglo Affymetrix para la detección de niveles expresión de ARNm correspondiente al genoma humano**. Imagen de Affymetrix.

Un experimento con microarreglos Affymetrix, utiliza moléculas de ácidos nucleicos (ADN, ARNm u oligonucleótidos) que funcionan como sondas que hibridan con sus cadenas complementarias provenientes de la muestra en estudio, las cuales, son llamados blancos. Comúnmente, estos blancos se generan a partir de ARNm de la célula o del tejido de interés. Una gran ventaja de esta tecnología es que requiere poco material genético para el diseño del microarreglo.

A continuación se describen de forma breve y con ilustraciones las etapas de un experimento con microarreglos Affymetrix. El primer paso, consiste en fijar una cadena corta de

ADN en la superficie de cristal del microarreglo. Estas cadenas son las denominadas sondas (figura 2.2). La cadena está formada por sólo 25 bases y se verifica que no esté presente en ningún otro gen, de esta manera, cuando una molécula de ARNm (blanco) se une a la sonda, se deduce que el gen se ha expresado.

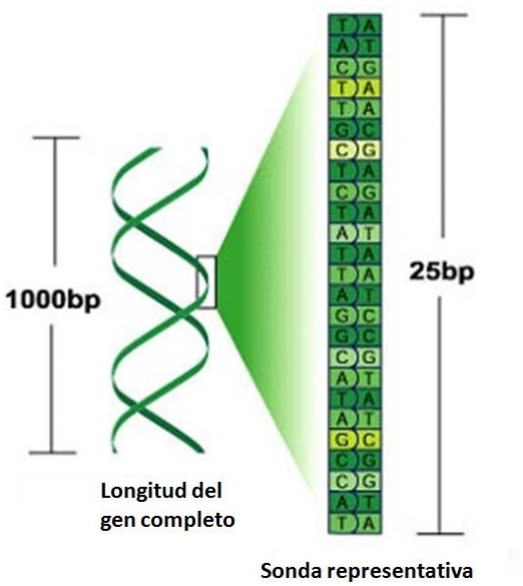

Figura 2.2: **Esquema representativo de una de sonda de ADN**. La doble hélice de ADN corresponde al gen completo, mientras que la respectiva sonda a una fracción de éste. *bp* = Pares de bases. Imagen de Affymetrix.

Las sondas buscan un alineamiento perfecto con el ARNm que se trata de detectar. Por cada una de las sondas para alineamiento perfecto (*perfectmatch*), se coloca una modificada (*mismatch*) en el que la base central se modifica para que no haya un alineamiento perfecto con el blanco objetivo. Esto porque en teoría sólo se adhieren a las sondas las secuencias complementarias perfectas pero en la práctica se producen errores de adherencia no específicos. Se añade la sonda modificada para cuantificar este tipo de errores. La figura 2.3 muestra un esquema representativo de estos oligonucleótidos.

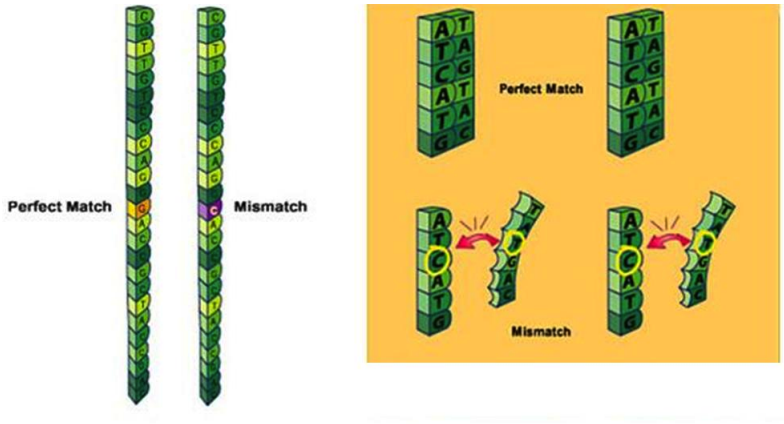

Figura 2.3: **Esquema representativo del producto de hibridación de las sondas *PM* y *MM***. La sonda PM hibrida perfectamente mientras que la MM no lo hace. Imagen de Affymetrix.

La superficie de un microarreglo es como un tablero de ajedrez gigante que ha sido comprimido considerablemente de tamaño. Cada uno de los cuadrados del tablero contiene un único tipo de sonda. Cada sonda se construye molécula a molécula usando el mismo tipo de tecnología que la que se utiliza para construir semiconductores de computadora. A su vez, dichas moléculas se van construyendo base a base (ver figura 2.4).

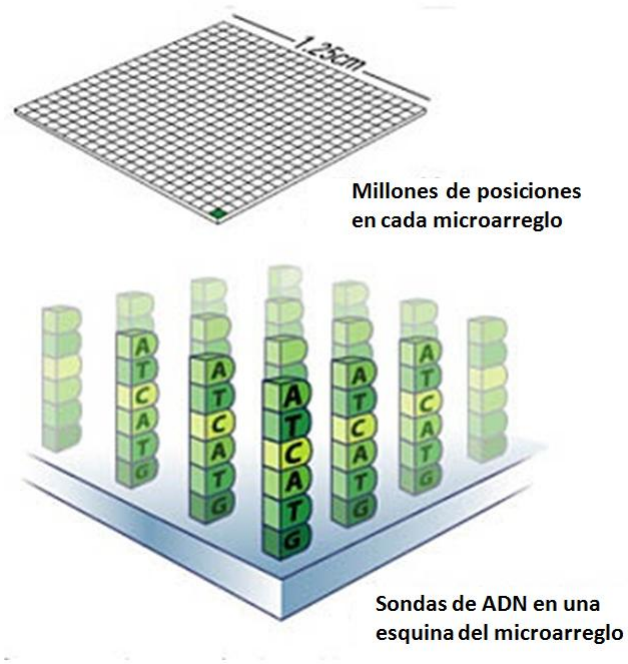

Figura 2.4: **Esquema representativo de un microarreglo**. Para efectos de simplicidad se redujó el número de bases de las sondas. El conjunto de sondas mostrado sólo está contenido en un cuadrado de los miles presentes en el tablero (celdas de sondeo). Imagen de Affymetrix.

Una vez diseñado el microarreglo para medir la expresión de ARNm, se procede a la extracción el dicho material genético, se amplifica por *PCR* (Reacción en cadena de la polimerasa) y se marca con moléculas de biotina. *PCR* es una técnica de biología molecular a través de la cual se puede obtener un gran número de copias de un fragmento de ADN particular, partiendo de un mínimo; en teoría basta una única copia de ese fragmento original, o molde. Las muestras preparadas de ARNm se lavan sobre el microarreglo por un periodo de 14 a 16 horas. El número de moléculas implicado en el proceso es enorme. Si la secuencia de bases de los blancos de ARNm corresponde a la de la sonda, habrá un alineamiento perfecto(figura 2.5).

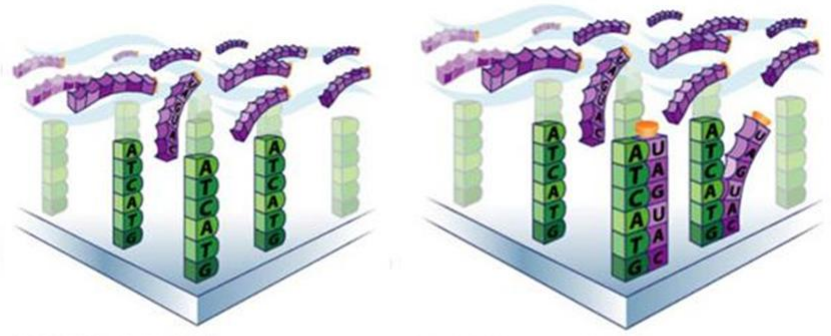

Figura 2.5: **Esquema representativo del proceso de hibridación en microarreglos**. Los blancos son los oligonucleótidos morados, las sondas los verdes y las copas naranjas las marcas de biotina. Imagen de Affymetrix.

Por último, el microarreglo es escaneado para producir una imagen representativa de la cantidad de moléculas que hibridaron en los diferentes conjuntos de sondas de acuerdo a la intensidad de señal detectada. A partir de este momento, comienza la fase de pre-procesamiento que tiene por objetivo preparar los datos para el análisis estadístico-matemático que se explicará más adelante.

Existen tres clases de sondas que pueden ser utilizadas para el diseño de microarreglos [14]: dos son genómicas (pueden detectar pérdida, ganancia de genes o mutaciones puntuales del ADN) y una es transcriptómica (mide los niveles de expresión de ARNm). La diferencia entre cada una de ellas es el tipo de ADN que se inmoviliza en los platos, de manera que dependiendo del objetivo de estudio se elige el tipo de microarreglo adecuado. Las siguientes, son algunas aplicaciones de los diferentes microarreglos en el campo de la medicina.

2.1. En la detección de cambios en la expresión génica

Cuando se desea determinar un cambio en el nivel de expresión de cierto gen, puede realizarse un análisis de expresión génica en microarreglos. El ARN inmovilizado por hibridación es ARNm de genes conocidos. Este proviene de células de tejidos sanos (control) y enfermos (tratamiento). Si un gen se sobreexpresa en cierta enfermedad, mayor cantidad de ARNm hibridará en el *spot* que representa el gen afectado; por consiguiente las intensidades de señal serán disímiles entre el grupo bajo estudio y el grupo control. Una vez que se caractericen los genes involucrados en ciertas enfermedades se podrá determinar si la persona tiene el patrón de expresión génica relacionada con alguna enfermedad, y en el caso ideal, se logrará un diagnóstico y tratamiento oportunos.

Los microarreglos de expresión génica también pueden ser usados para determinar cambios en esta durante un periodo de tiempo, por ejemplo, durante el ciclo celular. Esto representa una importante herramienta en la investigación en cáncer, ya que se podrían identificar nuevos marcadores cancerígenos con propósitos diagnósticos. Adicionalmente, proteínas que se expresan en las células tumorales podrían convertirse en el blanco de nuevas estrategias terapéuticas. Además, dichos microarreglos también pueden utilizarse para desarrollar medicamentos nuevos y determinar si estos afectan las expresión de ciertos genes en células cancerígenas.

2.2. En la detección de ganancia o pérdida de material genético

Un ejemplo de esta aplicación es nuevamente el cáncer, donde ciertas pérdidas o ganancias cromosómicas están relacionadas con su progresión y los patrones de estos cambios son importantes para establecer un pronóstico clínico. Con el uso de microarreglos y el análisis de sus resultados se podría predecir cuáles regiones cromosómicas contienen genes que inician y mantienen procesos tumorales. Los resultados pueden ser visualizados en diagramas jerárquicos referidos como modelos en árbol de progresión del tumor. Dichos

diagramas permiten detectar ganancias y pérdidas o cambios en el número de copias de un gen particular involucrado en determinada patología. En este tipo de microarreglos se usan grandes porciones de ADN genómico inmovilizado en el plato y cada *spot* de ese ADN representa una región cromosómica conocida. La mezcla a hibridar contendrá ADN genómico marcado, proveniente de tejido enfermo y tejido sano. Si el número de copias del gen de interés ha aumentado, una gran cantidad de ADN de la muestra hibridará en los *spots* del microarreglo que representan el gen involucrado, mientras que de la muestra proveniente de tejido sano sólo una pequeña cantidad de ADN hibridará en el mismo punto. Dicha diferencia se reflejará en intensidades disímiles que serán detectadas por el escáner y los programas de lectura.

2.3. En la detección de mutaciones en el ADN

Cuando se utilizan los microarreglos para detectar mutaciones o polimorfismos en una región genómica, el material colocado en diferentes *spots* puede diferir solamente en uno o en unos cuantos nucleótidos. El tipo de mutación que suele utilizarse para este tipo de análisis es denominado SNP (*Single Nucleotide Polymorphism*). Los SNPs, son pequeñas variaciones que pueden ocurrir en el ADN y que a diferencia de otras mutaciones, deben estar presentes en al menos el $1\,\%$ de la población para ser consideradas como tales. En este tipo de experimento se necesita ADN genómico derivado de una muestra normal para ser usado en la mezcla de hibridación.

Una vez que se ha establecido que un SNP se relaciona con una enfermedad de interés, se puede usar esta tecnología para detectar quienes tienen o son susceptibles de tener la enfermedad. Si el ADN genómico de un individuo se agrega a un microarreglo cargado de varios SNPs, el ADN muestra (blanco) hibridará con mayor frecuencia con los SNPs asociados al individuo de estudio. Estos puntos del microarreglo tendrán mayor intensidad y se demostrará que dicha persona es susceptible o tiene determinada enfermedad.

Capítulo 3

Bases matemáticas en el análisis de datos de microarreglos

3.1. Antecedentes

La investigación genómica actual es un campo multidisciplinario por excelencia donde biólogos, matemáticos, estadísticos y computólogos trabajan en equipo. Las nuevas tecnologías ofrecen un detalle y alcance más elevado en los estudios moleculares; esto a su vez se traduce en volúmenes masivos de datos. El uso de microarreglos permite estudiar miles de genes en forma simultánea, es decir, cada uno genera información de miles de variables (expresión de genes en el caso de este trabajo). Dada la complejidad de su diseño, el análisis de estos datos no es una tarea sencilla y no existe un método que sea un estándar de oro. El diseño de algoritmos requiere de adecuaciones a un contexto de muchas variables pocas muestras comprometiendo las consideraciones esenciales de la ley fuerte de los grandes números de Kolmogorov, a que los genes no se pueden considerar como variables independientes e idénticamente distribuidas y a que no siempre la distribución de los datos sigue una curva Gaussiana que nos permita hacer otro tipo de consideraciones. El esquema de muchas variables pocas repeticiones implica que los métodos estadísticos más utilizados sean en base a modelos Bayesianos, ya que permiten incorporar conocimiento *a priori* (tal vez conocimiento cualitativo), los resultados se aproximan a la estimación de máxima verosimilitud conforme n aumenta, los intervalos de confianza tienen un significa-

do más intuitivo (son llamados conjuntos de credibilidad), tienen la habilidad de encontrar más cantidades de interés y es relativamente fácil establecer y estimar modelos complejos. Al mismo tiempo el volumen de datos representa un reto para quienes los almacenan y procesan, siendo necesario un estudio detallado de la complejidad de cada algoritmo a utilizar.

La capacidad de estudiar el genoma completo de un organismo permite responder y generar una enorme cantidad de preguntas en un único experimento. Sin embargo, a pesar del uso de sofisticado equipo, detallados protocolos experimentales y una amplia gama de algoritmos matemáticos diseñados específicamente para el análisis de los datos obtenidos, las fuentes de variación son muchas y muy diversas. Para garantizar la reproducibilidad de estos experimentos, así como su relevancia biológica, deben considerarse todas las posibles fuentes de variación. En general dichas fuentes de variación se pueden clasificar de la siguiente manera:

- **Manufactura del microarreglo.** En los experimentos con microarreglos es fundamental tener en cuenta la variación que el efecto de lote introduce en los resultados. Aunque se recomienda utilizar un solo lote para cada experimento, en ocasiones el número de microarreglos necesarios sobrepasa la cantidad de los contenidos en el lote, en estos casos se debe revisar el diseño del experimento para asignar los microarreglos de manera que el efecto sobre los resultados sea el menor posible. También puede suceder que el análisis esté orientado a grupos de muestras obtenidas y procesadas por diferentes laboratorios. En este sentido, existen algoritmos creados específicamente para tratar el efecto de lote. En relación al diseño del microarreglo, se ha demostrado que la localización de las sondas en la superficie del microarreglo provoca un efecto espacial que introduce un sesgo sobre la señal de intensidad[15].

- **Selección de muestras biológicas.** La selección de muestras usualmente es una de las fuentes de variación más notable. Hay una enorme cantidad de factores a considerar. Dicha selección depende en gran parte del objetivo principal del experimento por lo que la pregunta biológica a responder debe ser clara. Algunos ejemplos comunes son la edad, el género, los hábitos de vida, el tipo y grado de la enferme-

dad bajo estudio e incluso la raza de la persona. También existen microarreglos para bacterias como *Escherichia coli* o bien en ocasiones las muestras provienen de líneas celulares específicas. En estos casos las condiciones de cultivo, el tiempo en el cual se extrae el material genético y la aplicación efectiva de los diferentes tratamientos, juegan un papel fundamental.

- **Protocolo experimental.** Durante el protocolo experimental la preparación del ARNm y la hibridación de éste en el microarreglo pueden considerarse como pasos críticos. En el primero, son factores relevantes la cantidad de material genético, su degradación, el *pooling* y algunas condiciones ambientales tales como el efecto del ozono sobre los marcadores fluorecentes Cy3-Cy5 en microarreglos de dos colores[16]. En el segundo, principalmente la hibridación no específica y la temperatura se suman a las fuentes de variación.

- **Procesamiento de Imagenes.** Si se considera que esta etapa parte desde el escaneo hasta la transformación de la imagen obtenida en un archivo de pixeles. Las fuentes de variación son debidas a aspectos como la calidad y calibración del escáner, el ruido de fondo (causado por la hibridación no específica), el re-escaneo, la alineación de la imagen al *grid* (red de coordenadas de las celdas de sondeo) y los mismos algoritmos para su procesamiento.

- **Algoritmos matemáticos para el análisis de expresión.** En la actualidad, para cada fase del análisis de microarreglos (Pre-procesamiento, Clasificación y Predicción) existe una amplia gama de algoritmos matemáticos. Cada uno debe aplicarse cuidadosamente con el fin de no perder información biológica relevante y no existe un diagrama de flujo establecido para ello. Los principales algoritmos dentro del análisis de expresión génica son para la corrección de fondo, normalización, selección de genes diferencialmente expresados, agrupamiento y aprendizaje estadístico.

3.2. Fases del análisis de datos de microarreglos

El siguente esquema resume las fases del análisis de datos de microarreglos (figura 3.1). El analista debe conocer a detalle los aspectos técnicos de cada experimento con el objetivo

de elegir el curso apropiado en dicho análisis.

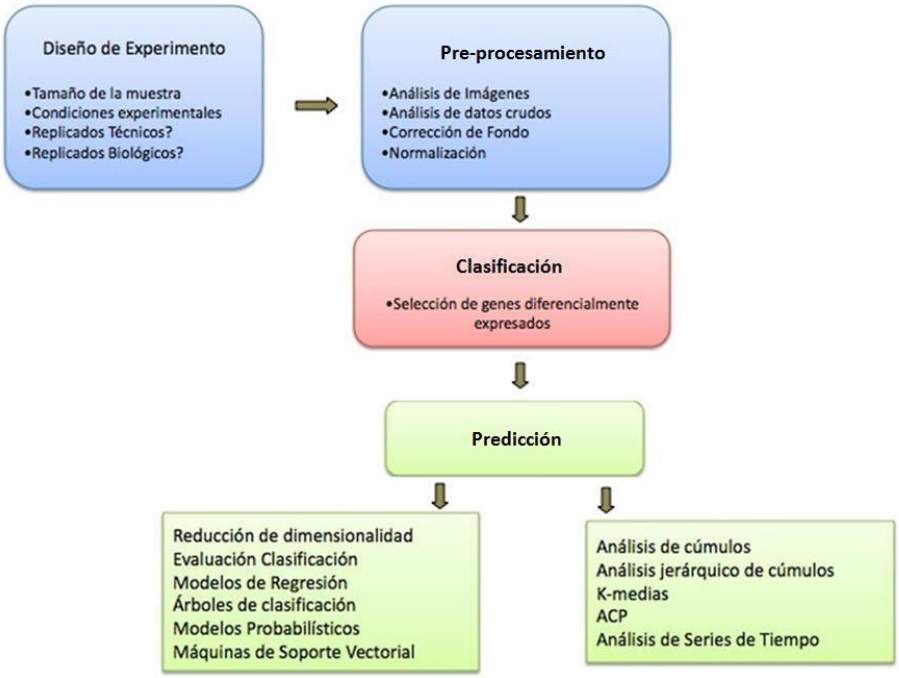

Figura 3.1: **Esquema representativo de las fases del análisis de datos de microarreglos**.

3.2.1. Fase I: Pre-procesamiento

El punto de partida para garantizar un buen análisis depende de la calidad del microarreglo y la preparación adecuada de los datos para adquirir una medida cuantitativa confiable de expresión génica que permita hacer las comparaciones de interés. Cada impresión de material genético debe ser completa, ordenada y adquirir una medida cualitativa total de calidad. Tanto la hibridación, la detección de presencia o ausencia de expresión génica y la existencia de celdas de sondeo vacías o fuera de sus límites depende de un buen manejo experimental. Una vez garantizado lo anterior, se da inicio al análisis de datos de microarreglos. Los pasos involucrados en la primera fase llamada pre-procesamiento de

datos se muestran a continuación:

1. Control de calidad
2. Corrección de fondo
3. Normalización
4. Generación de niveles de expresión

3.2.2. Fase II: Clasificación

Posteriormente, se deben identificar dentro de los resultados del experimento, los genes que presentan cambios significativos. Lo anterior se realiza por medio de la filtración e identificación de genes activos, donde esto último con frecuencia se basa en métodos bayesianos.

3.2.3. Fase III: Predicción

La predicción consiste en la busqueda de una firma de expresión génica que prediga conjuntos de clases. La idea de identificar *clusters* o particiones de genes, reside en que aquellos que tienen patrones de expresión cercanos contienen mecanismos regulatorios iguales y eventualmente podrían crearse mapas de control de transcripción exactos. Adicionalmente, los genes corregulados pueden ayudar a la asignación de funciones a genes redescubiertos.

3.3. Pre-procesamiento

En los últimos años los microarreglos Affymetrix han sido unas de las plataformas genómicas más populares. Estos microarreglos contienen sondas que son sintetizadas *in situ* en obleas de cuarzo. La síntesis de luz dirigida es llevada a cabo pasando nucleótidos de adenina, guanina, citosina o timina que contienen un grupo protector fotosensible sobre

la oblea. Máscaras litográficas son usadas para bloquear o bien transmitir luz sobre una locación específica en el arreglo. El emparejamiento de los nucleótidos sólo ocurrirá en las regiones iluminadas. Este proceso es repetido con los diferentes nucleótidos de tal forma que la sonda es construida usando una base por ciclo. La mayor ventaja de esta tecnología es el gran número de celdas de sondeo en el arreglo, el cual aumenta la redundancia de tal manera que cada transcrito es oportunamente detectado por 22 diferentes secuencias de oligonucleótidos, conocido como conjunto de sondas. Dentro del conjunto de sondas, 11 de las secuencias son designadas para empalmar perfectamente con el blanco del transcrito por lo que son llamadas *perfect match* o PM , así mismo sus parejas, las cuales difieren de la sonda de referencia en una base en el centro de la secuencia son llamadas *mismatch* o MM (ver capítulo 2). Generalmente, entre 16 y 20 de estos pares de sondas representan un gen. Este diseño implica que la intensidad registrada por múltiples sondas es usada para detectar cada transcrito y el grado de hibridación no específica puede ser estimado a partir de las sondas MM. El resultado de manejar 22 sondas por transcrito significa que el microarreglo más largo en 2006, contiene un exceso de $1{,}3$ millones de celdas de sondeo de 1μm de diámetro y detecta 54 mil transcritros[24]. Affymetrix, en su algoritmo MAS 5.0 utiliza como medida de expresión génica la siguiente señal[25]:

$$\text{Señal} \sim \frac{\sum_{k=1}^{N} w_k (PM_k^{intensity} - MM_k^{intensity})}{\sum_{k=1}^{N} w_k}, \tag{3.1}$$

donde N es el número de estos pares que determinan un gen en particular y w_k es el peso asignado a cada celda k obtenida a partir de sus respectivos pares, el cual esta dado por el siguiente esquema binario,

$$w_k = \begin{cases} 1 & \text{PM} > \text{MM} \\ 0 & \text{PM} < \text{MM}, \end{cases} \tag{3.2}$$

de forma que se eliminan celdas con una mejor hibridación para la sonda MM y se garantiza que la diferencia promedio tenga un valor positivo. Esto tiene por objetivo proveer

medidas de expresión reproducibles. Sin embargo, en la práctica los MM no representan sólo ruido y al usar la ecuación 3.2 como medida de expresión génica se suele perder información relevante. Los últimos microarreglos desarrollados por Affymetrix (el *Human Gene 1.0 ST* por ejemplo) ya no contienen sondas MM. En la actualidad, generalmente se utiliza el algoritmo RMA (Robust Multiarray Average) para estimar el ruido de fondo, normalizar y generar la correspondiente medida de expresión génica[17].

Con el objetivo de producir un espectro continuo en un rango de valores numericamente más fácil de manejar, a las intensidades de señal PM y MM se les aplica una transformación logarítmica con base 2,

$$\text{Señal} = \log_2(PM). \tag{3.3}$$

Computacionalmente, esta transformación es conveniente ya que por propiedades de los logaritmos todas las operaciones de multiplicación pueden expresarse como sumas.

Además del material genético bajo estudio, se incluyen sondas de control, normalizadoras o calibradoras. Estas sondas son dirigidas hacia genes en los que no se espera un cambio significativo en su expresión bajo las condiciones del experimento. Es deseable incluir el mayor número de sondas control como sea posible. Ejemplos de estos controles pueden ser los *BioB*, *BioC* y *BioD* que representan genes dentro de la ruta metabólica para la síntesis de biotina de *Escherichia Coli* o *Cre recombinase* que es tomado del bacteriofago $P1$[25]. Debido a que estos controles son agregados en forma idéntica, se espera que las intensidades sean iguales o por lo menos muy similares. Esto básicamente significa que en un gráfico de dispersión de dos microarreglos, los puntos que representan a estos controles de hibridación deben caer a lo largo de la diagonal de 45 grados. A partir de los controles se puede deducir el punto de corte o umbral para definir expresión diferencial, así como determinar el rango dinámico para los niveles de expresión de los genes que en general es un intervalo con valores de 6 a 12.

3.3.1. Corrección de ruido de fondo

La señal de intensidad es detectada cuando la sonda muestra u oligo de DNA motivo de estudio se adhiere a su correspondiente contraparte sonda blanco, presente desde el diseño del microarreglo). A este proceso se le llama hibridación (ver capítulo 2). Sin embargo, durante este proceso puede darse el caso de hibridación no específica, es decir, se emite una señal pero las dos sondas no se complementaron correctamente. También es posible que el lavado del material genético no haya retirado todas aquellas sondas que no hibridaron y se emita una señal cuando no debería existir. Ambos casos son fuentes de ruido para la estimación de la señal de intensidad y su posterior análisis. El objetivo de realizar la corrección de ruido de fondo es detectar la señal falsa y removerla, pero existe una complicación en este procedimiento, el ruido de fondo varía de transcrito en transcrito y por encima de esto también por microarreglo. En general, dicho ruido se remueve restando el valor medio por lo que tendrá efecto positivo cuando la estimación sea muy buena pero una enorme afectación si la dispersión entre las muestras es grande. Esto hace que la estimación de ruido sea una alternativa mas que una regla. Existen diversos algoritmos para remover el ruido de fondo[19][18], el más usado es el RMA (Robust Multiarray Average)[17] cuya principal característica es el uso de una variable que indica distribución espacial tratando de compensar por hibridación débil en algunas posiciones físicas dentro del microarreglo.

3.3.2. Normalización

La necesidad de normalización surge cuando se lidia con experimentos que involucran múltiples arreglos. Existen dos amplias caracterizaciones que podrían ser usadas de acuerdo al tipo de variación esperada cuando se comparan arreglos: Variación de interés y variación oscura[26]. Por ejemplo, diferencias grandes en el nivel de expresión de un gen en un tejido enfermo y uno sano es una variación de interés. Sin embargo, los niveles de expresión observados incluyen cierto ruido introducido durante el proceso del experimento, ésta podría ser clasificada como variación oscura. Ejemplos de esta variación surgen debido a diferencias en la preparación de la muestra o en la producción y procesamiento

de arreglos. El propósito de la normalización es tratar con esta variación obscura.

Affymetrix ha atacado el problema de la normalización proponiendo que las intensidades deben ser escaladas para que cada arreglo tenga el mismo valor promedio. La normalización de Affymetrix es desarrollada sobre una integración de los valores de expresión, sin embargo, este enfoque no funciona particularmente bien en casos donde hay relaciones no lineales entre arreglos. Con el fin de resolver este problema, se han desarrollado muchos otros métodos de normalización bajo diferentes enfoques. Sin embargo, la mayoría de éstos dependen de la elección de un arreglo de referencia, asunto que puede tornarse problemático ya que no se puede garantizar que el arreglo elegido no contenga un determinado sesgo.

Bolstad, Irizarry, Astrand y Speed, proponen tres métodos de normalización que no dependen de arreglos de referencia[27]. Estos métodos realizan la normalización de los datos a nivel de intensidad de sonda para todas las sondas de todos los arreglos, por lo que son llamados métodos de información total. En general no se tratan los PM y MM por separado, en lugar de ello siempre se consideran ambas como señales de intensidad que deben ser normalizadas.

3.4. Clasificación

Sean x_{ki} y x_{kj}, las señales de intensidad de cada microarreglo $i, j = 1, 2, ...n$ y M_k la diferencia entre los valores logarítmicos de expresión para cada sonda $k = 1, 2, ...p$. Por propiedades de logaritmos, M_k se puede expresar como el cociente

$$M_k = \log_2 \frac{x_{ki}}{x_{kj}}. \tag{3.4}$$

Además de las razones mencionadas en la sección de pre-procesamiento, la transformación logarítmica de base 2 se aplica para producir un espectro continuo de valores para genes

diferencialmente expresados y así tratar los genes sobre o sub expresados equivalentemente, recordando que los logarítmos tratan a los números y sus inversos simétricamente: $\log_2(1) = 0$, $\log_2(2) = 1$, $\log_2(\frac{1}{2}) = -1$, $\log_2(4) = 2$ y $\log_2(\frac{1}{4}) = -2$. De acuerdo a esto último, un gen sobre expresado por un factor de 2 tiene un $\log_2(cociente)$ de 1, un gen sub expresado por un factor de 2 tiene un $\log_2(cociente)$ de -1 y un gen expresado en un nivel constante (con cociente de expresión igual a 1) tiene un $\log_2(cociente)$ de 0[25]. Es importante señalar que antes de calcular la ecuación 3.4 las señales de intensidad deben ser comparables entre sí, es decir, deben tener la misma distribución, lo que se garantiza a través de un proceso de normalización. Los algoritmos de normalización se revisarán con mayor detalle en su respectivo capítulo.

De acuerdo al enfoque frecuentista, los estadísticos que podrían utilizarse para seleccionar los genes diferencialmente expresados son el promedio (la diferencia entre los valores logarítmicos de expresión) y su forma estandarizada, ésto es $M_k = \log_2(\frac{x_{ki}}{x_{kj}})$ y $t_k = \frac{M_k}{SE_k}$ respectivamente, donde $k = 1, ..., p$ representa el número de sonda para cualesquiera dos microarreglos $i, j = 1, ..., n$ con señales de intensidad x_{ki} y x_{kj}; $SE_k = \frac{s_k}{\sqrt{n}}$ es el error estandar de M_k, y s_k es la desviación estandar de M_k.

Asumiendo que los datos ya están normalizados, para cada gen $g = 1, ...N$ se desea conocer si el vector de $M_k - values$, M_g, provee evidencia para sugerir que el verdadero valor de M de ese gen es diferente de cero. Sin embargo, hay algunos problemas con estos estadísticos. Por ejemplo, una media grande puede surgir a consecuencia de un solo valor atípico, cosa que ocurre frecuentemente en este contexto. Además, un valor grande de t puede ser provocado por un denominador pequeño (SE) incluso aunque la media sea por sí misma pequeña. Si se tiene en cuenta que de un gran número de genes en un experimento con microarreglos la mayoría tendrá errores estandar muy pequeños y algunos de estos tendrán una media pequeña, tampoco esta t sería confiable. Tusher propuso un refinamiento de la t para tratar las respectivas dificultades mencionadas a través de la adición de un término constante a su denominador para evitar que este sea muy pequeño[21]. Se sugiere un factor a_0, igual al percentil 90 de los errores estandar de todos los genes de

tal forma que se busca el valor absoluto de la expresión $S_g = \frac{M_g}{s_g + a_0}$[23].

Una alternativa posible, esta basada en una aproximación empírica de Bayes. Datos de todos los genes en un conjunto de replicados son combinados para estimar los parámetros de una distribución previa. Dichos parámetros estimados son entonces combinados a nivel de expresión génica, con medias y desviaciones estandar para formar un estadístico B el cual es un *Bayes log posterior odds*. Este puede ser usado para determinar si hay expresión diferencial. El estadístico B ha resultado ser más eficiente en este tipo de problemas que esta t refinada[20].

3.4.1. El estadístico B

Los argumentos que justifican este *Bayes log posterior odds* llamado B son complejos, por lo que no se profundizará a detalle en su desarrollo y para una mejor comprensión se recomienda revisar a fondo la referencia [20]. Nótese que el desarrollo que se presentará en este trabajo, difiere al del artículo de referencia en el tipo de microarreglo de aplicación. En la referencia se manejan microarreglos de dos canales y en este documento microarreglos Affymetrix de un canal, esto implica algunos cambios en la notación del desarrollo matemático que se presenta en esta sección.

Sea N el número de genes en un experimento de microarreglos, n el número de replicados para cada gen y $M_{ij_1j_2} = \log_2(\frac{x_{ij_1}}{x_{ij_2}})$ donde $i = 1,...N,\ j_1, j_2 = 1,...n$ y $j_1 \neq j_2$, la diferencia entre los valores logarítmicos de expresión. Se considera que $M_{ij_1j_2}$ es una variable aleatoria que se distribuye normalmente con media μ_i y varianza σ_i^2 (la varianza es encontrada empíricamente), independiente e identicamente distribuida,

$$M_{ij_1j_2} \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2) \quad \text{para toda i.} \tag{3.5}$$

Sea I, la variable que índica si el gen $g = 1,...,N$ está diferencialmente expresado ($\mu_g \neq 0$). Para cada gen g se desea conocer $P(I_g = 1 \mid M_{ij_1j_2})$ (La mayoría de los genes tienen $\mu_g = 0$, pero una pequeña proporción p de estos tienen $\mu_g \neq 0$, indicada por $I_g = 1$ como opuesto a $I_g = 0$) y equivalentemente, el *log odds B*,

$$B_g = \log \frac{P(I_g = 1 \mid M_{ij_1j_2})}{P(I_g = 0 \mid M_{ij_1j_2})}. \tag{3.6}$$

Entonces, $P(I_g = 1 \mid M_{ij_1j_2}) > P(I_g = 0 \mid M_{ij_1j_2})$ si y sólo si $B_g > 0$ (Los parámetros (μ_g, σ_g^2) son tratados como realizaciones independientes identicamente distribuidas de parámetros con cierta distribución *a priori*). Ahora, utilizando el teorema de Bayes, bajo el supuesto de independencia sobre todos los genes,

$$B_g = log \frac{\frac{Pr(I_g=1)Pr(M_{ij_1j_2}|I_g=1)}{\sum_{g=1}^{N} Pr(I_g=1)Pr(M_{ij_1j_2}|I_g=1)}}{\frac{Pr(I_g=0)Pr(M_{ij_1j_2}|I_g=0)}{\sum_{g=1}^{N} Pr(I_g=0)Pr(M_{ij_1j_2}|I_g=0)}}$$

$$B_g = log \frac{p}{1-p} \frac{Pr(M_{ij_1j_2} \mid I_g = 1)}{Pr(M_{ij_1j_2} \mid I_g = 0)} \tag{3.7}$$

$$B_g = log \frac{p}{1-p} \frac{Pr(M_g \mid I_g = 1) \prod_{i \neq g} Pr(M_i \mid I_g = 1)}{Pr(M_g \mid I_g = 0) \prod_{i \neq g} Pr(M_i \mid I_g = 0)}$$

$$= log \frac{p}{1-p} \frac{Pr(M_g \mid I_g = 1)}{Pr(M_g \mid I_g = 0)}, \tag{3.8}$$

donde M_g es el vector de las n medidas para el gen g y p es la proporción de genes diferencialmente expresados en el experimento, $p = Pr(I_i = 1)$ para toda $i = 1,2,...N$. Entonces, es necesario calcular las funciones de densidad $f_{I_i=1}(M_i)$ y $f_{I_i=0}(M_i)$.

Dado que se tienen pocas muestras y muchas variables (genes), para usar todo el conocimiento que se tiene de las medias y las varianzas se colecta la información obtenida del conjunto completo de genes en la estimación de sus respectivas distribuciones conjuntas *a priori*. Se asume una distribución gamma para $\frac{1}{\sigma_i^2}$ y normal para μ_i dada σ_i^2. Esto es una distribución conjugada *a priori* que permite calcular B_i de forma explícita[22]. Para ν grados de libertad y los parámetros de escala $a > 0$, $c > 0$, se propone $\tau_i = \frac{na}{2\sigma_i^2}$ y se supone que,

$$\tau_i \sim \Gamma(\nu, 1), \tag{3.9}$$

$$\mu_i|\tau_i = \begin{cases} 0 & \text{Si } I_i = 0 \\ N(0, \frac{cna}{2\tau_i}) & \text{Si } I_i = 1, \end{cases} \tag{3.10}$$

para toda $i = 1, .., N$. El parámetro c expresa dependencia entre los *priors* para μ_i y τ_i y es necesario para los cálculos. De forma que las funciones de densidad son:

$$f_{\tau_i} = \frac{1}{\Gamma(\nu)} \tau^{\nu-1} e^{-\tau_i}, \tag{3.11}$$

$$f_{I_i=1}(\mu_i|\tau_i) = (2\pi)^{-\frac{1}{2}} c^{-\frac{1}{2}} \left(\frac{na}{2\tau_i}\right)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{2\tau_i}{cna}\mu_i^2}, \tag{3.12}$$

$$f_{I_i=0}(\mu_i) = \delta(0), \tag{3.13}$$

$$\begin{aligned} f(M_i|\mu_i, \tau_i) &= (2\pi)^{-\frac{n}{2}} \left(\frac{na}{2\tau_i}\right)^{-\frac{n}{2}} e^{-\frac{1}{2}\frac{2\tau_i}{cna}\sum_i (M_{ij_1j_2} - \mu_i)^2} \\ &= (2\pi)^{-\frac{n}{2}} \left(\frac{na}{2\tau_i}\right)^{-\frac{n}{2}} e^{-\frac{1}{2}\frac{2\tau_i}{cna}\left(\sum_i^N (M_{ij_1j_2} - M_i)^2 + n(M_i - \mu)^2\right)}, \end{aligned} \tag{3.14}$$

$$f_{I_i=1}(M_i) = \int_0^\infty \int_{-\infty}^\infty f_{I_i=1}(M_i, \mu_i, \tau_i) d\mu_i d\tau_i$$

$$= \int_0^\infty \int_{-\infty}^\infty f(M_i|\mu_i, \tau_i) f_{I_i=1}(\mu_i|\tau_i) d\mu_i d\tau_i, \tag{3.15}$$

$$f_{I_i=0}(M_i) = \int_0^\infty \int_{-\infty}^\infty f_{I_i=0}(M_i, \mu_i, \tau_i) d\mu_i d\tau_i$$

$$= \int_0^\infty f(M_i|\mu_i = 0, \tau_i) f(\tau_i) d\tau_i, \tag{3.16}$$

La integración de las funciones de densidad conjunta es desarrollada identificando la distribución normal *posterior* de $\mu_i|\tau_i$, $N(\frac{nc}{1+nc}M_i, \frac{a}{2\tau i}\frac{nc}{1+nc})$, para el caso $I_i = 1$ y la distribución gamma *posterior* de τ_i, $\Gamma(\nu + \frac{n}{2}, 1 + \frac{1}{a}(s_i^2 + \frac{M_i^2}{1+nc}))$, $s_i^2 = \frac{1}{n-1}\sum_i (M_{ij_1j_2} - M_i)^2$. Entonces, al integrar

$$f_{I_i=1}(M_i) = \frac{\Gamma(\nu + \frac{n}{2})}{\Gamma(\nu)}(2\pi)^{-\frac{n}{2}}(1+nc)^{-\frac{1}{2}}\left[1 + \frac{1}{a}(s_i^2 + \frac{M_i^2}{1+nc})\right]^{-(\nu+\frac{n}{2})}, \tag{3.17}$$

$$f_{I_i=0}(M_i) = \frac{\Gamma(\nu + \frac{n}{2})}{\Gamma(\nu)}(2\pi)^{-\frac{n}{2}}\left[1 + \frac{1}{a}(s_i^2 + M_i^2)\right]^{-(\nu+\frac{n}{2})}. \tag{3.18}$$

Por tanto para un gen g, de 3.8 se tiene

$$B_g = \log\frac{p}{1-p}\frac{1}{\sqrt{1+nc}}\left(\frac{a + s_g^2 + M_g^2}{a + s_g^2 + \frac{M_g^2}{1+nc}}\right)^{\nu+\frac{n}{2}}. \tag{3.19}$$

Aquí a y ν son hiperparámetros en la distribución *a priori* gama inversa para las varianzas, y c es un hiperparámetro en la distribución *a priori* normal de las medias que no son cero. En particular, s_g^2 es en este caso la suma de cuadrados sobre n en lugar de $n-1$. La única parte génica específica de B, es el último cociente, el cual es siempre ≥ 1 dado

que $1/(1+nc) < 1$. Se deduce que al incrementar la expresión génica (y por tanto M_g^2), incrementa B_g, y más si la varianza es pequeña. Si M_g^2 lo es también, a asegura que el cociente no pueda ser expandido por una muy pequeña varianza.

Hay cuatro parámetros en el modelo de B: p, ν, a y c. Desafortunadamente no es sencillo calcularlos. Por tanto, se ajusta p y se estima $\nu, a|p$ y $c|p, v, a$. Los parámetros ν y a son tales que $\tau_i = \frac{na}{2\sigma_i^2} \sim \Gamma(\nu, 1)$. Así, se usan las varianzas estimadas para estimar ν y a por el método de momentos. La dificultad en la estimación de c es que sólo se considera en la distribución de los genes diferencialmente expresados $(I_i = 1)$ y no se sabe cuales son. Una opción es comparar la función de densidad normal observada de los promedios $(M_i)_{i \in T}$, donde T es la proporción superior dada p con respecto a B, con la función de densidad normal observada de todos los promedios M_i; lo que conduce a la estimación de c. En ausencia de una estimación satisfactoria, p es ajustada a un valor sensible tal como $0{,}01$ o $0{,}001$, pero en consecuencia no se puede usar el umbral $B = 0$ para seleccionar los genes diferencialmente expresados.

Si se considera sólo la información del parámetro p, el *Bayes log posterior odds* se expresa,

$$B = \log \frac{p}{1-p}, \tag{3.20}$$

de forma que para cada valor de B, se calculan las probabilidades con la fórmula

$$p = \frac{e^B}{e^B + 1}. \tag{3.21}$$

Estas probabilidades pueden ser usadas como referencia para el analista de acuerdo a la siguiente tabla:

B	p	Nivel de confianza
0	0.50	50%
1	0.7310586	73.1%
1.5	0.8175745	81.8%
2	0.880797	88.1%
3	0.9525741	95.3%
6	0.9975274	99.8%

3.5. Predicción

Una base natural para organizar los datos de expresión génica es agrupar los genes con patrones similares de expresión. El primer paso para ello, es la adopción de una descripción matemática de similitud. Las medidas de similitud utilizadas pueden ser muy variadas, al igual que los métodos de agrupamiento, cada una de ellas tiene ventajas y desventajas. Aunque diversos métodos de *clustering* pueden organizar de forma útil tablas de medidas de expresión de genes, el resultado ordenado sigue siendo una colección masiva de números difícil de asimilar. Por lo tanto, siempre se combinan los métodos de *clustering* con representaciones gráficas de los datos primarios, asignando a cada punto un color que refleja cuantitativa y cualitativamente las observaciones experimentales originales. El producto final, es una representación del complejo de datos de expresión génica que a través de una organización estadística mostrada gráficamente, permite a los biólogos asimilar y explorar los datos de una manera intuitiva y natural.

Para identificar genes con perfiles de expresión similar, los mapas de calor resultan ser una herramienta muy útil. Estos mapas muestran una relación de agrupamiento tanto entre genes (renglones) como entre microarreglos (columnas).

3.5.1. Algunas medidas de similitud usadas en el análisis de expresión

- **Coeficiente de correlación de Pearson:** Es un concepto estadístico que mide el grado de relación entre dos variables que varían conjuntamente. Sin embargo, es

susceptible a ser sesgado por datos atípicos y asume los siguientes supuestos:

1. X e Y son variables aleatorias. Luego, no existe una variable explicativa y otra explicada.
2. La población de la cual se extrae la muestra es normal bivariada.
3. Existe una relación lineal entre las variables, la cual está medida por el coeficiente de correlación poblacional definido como:

$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(X-\mu_y)]}{\sqrt{E[(X-\mu_x)^2]E[(X-\mu_y)^2]}}, \tag{3.22}$$

   donde $-1 \leq \rho \leq 1$ y (X,Y) es una variable que se distribuye normalmente.

En el análisis de datos de microarreglos es común centrar los perfiles de expresión génica de manera que tengan media de 0 y desviación estandar de 1, para simplificar la ecuación anterior a:

$$\rho = \sum_{i=1}^{n} x_i y_i. \tag{3.23}$$

Además, esta medida requiere ser convertida a una medida de distancia, por ejemplo,

$$d(X,Y) = 1 - \rho(X,Y)^2. \tag{3.24}$$

Sin embargo, dado que algunos de los supuestos no siempre se cumplen y para datos de series de tiempo este centrado pierde la noción natural de que los genes son fuerte o débilmente regulados, el uso del coeficiente de correlación en el estudio de los perfiles de expresión de genes es motivo de discusión.

- **Correlación de Spearman:** Es una medida de correlación no paramétrica que es robusta para datos atípicos y ello representa una ventaja para el análisis de datos de microarreglos. Se basa en reemplazar los valores originales de ambas variables por números enteros positivos, comenzando del 1 en adelante, de forma que correspondan a su ordenamiento de menor a mayor magnitud (Rangos). Para ello, los valores reales de cada una de las variables son ordenados de menor a mayor, por separado y

reemplazados por los rangos[33]. Dado que la correlación de Spearman siempre utiliza la misma escala para los rangos de las observaciones de X e Y $(1,2,3,4...,n)$, el análisis se puede hacer mediante la siguiente fórmula:

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)} \tag{3.25}$$

Donde r_s es el coeficiente de correlación de Spearman, D^2 el cuadrado de las diferencias entre X e Y y N el número de parejas.

De la misma forma en la que el coeficiente de correlación debe ser transformado a una medida de distancia, r_s debe de serlo también. Además, es importante tener en cuenta que el método de Spearman esta limitado a calcular correlación pero sólo entre dos variables. Así mismo, tampoco permite hacer regresiones, es decir, no se puede modelar la variable respuesta Y con varios predictores en forma simultánea o ver la influencia de un predictor sobre otro.

- **Distancia euclidiana:** Esta medida de similitud, se basa en la distancia entre dos puntos calculada por medio del teorema de Pitágoras. Para los datos de expresión génica se extiende esta idea a muchas dimensiones.

Un problema con la distancia euclidiana es que no varía de acuerdo a la escala, es decir, dos perfiles de expresión con la misma forma pero diferentes magnitudes aparentarán ser muy distantes. Una posible solución a este problema es centrar los datos. Por otro lado, cuando dos genes tienen un perfil de expresión similar pero uno es regulado fuertemente y el otro débilmente, éstos aparentarán estar correlacionados pero la distancia euclidiana reflejará la diferencia entre patrones siendo una medida más realista.

Observaciones

- En perfiles de expresión opuestos, la correlación de Spearman tiene la habilidad de localizar correlaciones negativas mientras que la distancia euclideana las separa en el dendograma. Si los perfiles son similares, la correlación de Spearman puede producir distancias de cero, eliminando algunas estructuras finas en el *cluster*.
- La distancia euclidiana tiende a producir distancias más largas que las correlaciones, por lo que los clusters están más separados y se visualizan mejor.
- En la práctica se usan combinaciones entre algunas de las medidas de correlación y la distancia euclidiana.

3.5.2. Agrupamiento jerárquico

Probablemente las técnicas de agrupamiento jerárquico sean las más utilizadas para analizar perfiles de expresión génica. En general, consisten en ordenar los genes o perfiles de expresión en diagramas de árbol llamados dendogramas. En este diagrama, los perfiles similares se encontrarán cerca. Hay dos aproximaciones básicas para generar *clusters* jerárquicos[34]:

- **Aglomerativo:** Comienza con cada unidad como *cluster* y en cada paso, une los pares más cercanos de estos. Se requiere definir una noción de proximidad de agrupamiento.
- **Divisivo:** Empieza con un *cluster* que incluye a todas las unidades y en cada paso divide un *cluster* hasta que sólo restan unidades individuales de estos. En este caso, se necesita decidir que *cluster* dividir en cada paso y como hacer dicha división.

El agrupamiento jerárquico puede simplificar grandes volúmenes de información y revela grupos de genes similares que pueden entonces ser estudiados con mayor profundidad.

Algoritmo

En general, el algoritmo de agrupamiento jerárquico utilizado consta de cuatro pasos:

1. Busca a partir de una matriz de distancias los genes o *clusters* más cercanos.

2. Une dichos genes o *clusters* en el dendograma para formar un nuevo *cluster*.

3. Calcula la distancia entre el nuevo *cluster* formado y los otros genes y *clusters* (*Linkage*).

4. Regresa al paso uno y repite el procedimiento hasta que todos los genes y *clusters* están ligados, o bien, forman un único *cluster*.

El tercer paso del algoritmo puede realizarse por tres principales métodos[34]:

- *Single linkage:* Tiende a producir un efecto de encadenamiento ya que un solo gen es agregado a un *cluster* a la vez.
- *Complete linkage:* Produce *clusters* pequeños, compactados y muy bien definidos (siempre y cuando los *clusters* iniciales estén bien definidos).
- *Average linkage:* Es un intermediario entre los dos anteriores y por lo tanto tiende a funcionar bien en muchas aplicaciones de microarreglos.

Un algoritmo para agrupamiento de datos de expresión génica

En el algoritmo que se presenta a continuación, la métrica de similitud génica utilizada es una forma del coeficiente de correlación de Pearson y la distancia euclidiana. Este coeficiente de correlación tiene la siguiente forma:

Sea G_i el nivel de intensidad (bajo una transformación logarítmica base 2) para el Gen G en la condición (microarreglo) i. Para dos genes cualquiera X y Y observados sobre una serie de N condiciones, un valor de similitud puede ser calculado como sigue[36]:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\phi_Y}\right)$$

donde

$$\phi_G = \sqrt{\sum_{i=1}^{N}\frac{(G_i - G_{offset})^2}{N}}$$

Cuando G_{offset} se establece en la media de las observaciones de G, entonces ϕ_G se vuelve la desviación estándar de G y $S(X,Y)$ es exactamente igual al coeficiente de correlación de Pearson de las observaciones de X y Y. Los valores de G_{offset} que no son el promedio sobre las observaciones en G, son usados cuando hay un supuesto estado sin cambio o de referencia representado por el valor de G_{offset}, contra aquellos cambios que son analizados.

El algoritmo de *clustering* jerárquico usado se basa estrechamente en el método de average-linkage[37]. El obejtivo de este algoritmo es calcular un dendograma que reúne todos los elementos en un solo árbol. Para cualquier conjunto de n genes, un matriz de similitud triangular superior es calculada usando la métrica descrita, la cual contiene valores de similitud para todos los pares de genes. La matriz es escaneada para identificar el valor máximo (que representa el par de genes más similares), un nodo es creado uniendo estos dos genes y utilizando el respectivo promedio de las observaciones, un perfil de expresión génica es calculado para dicho nodo (los valores perdidos son omitidos y los dos elementos unidos son ponderados por el número de genes que contiene). La matriz de similitud es actualizada con este nuevo nodo reemplazando los dos elementos unidos y el proceso es repetido $n-1$ veces hasta que resta un solo elemento.

3.5.3. Agrupamiento por *K-means*

Este método es una técnica de *clustering* particional que intenta encontrar un número de *clusters* especificados por el usuario (K), los cuales son representados por sus respectivos centroides[34]. Por tanto, difiere del algoritmo jerárquico en tres aspectos principales. El número de *clusters* tiene que ser especificado por adelantado. No hay jerarquía o relación entre *clusters* ni entre genes o microarreglos dentro de los *clusters*. Basicamente, sólo son grupos con perfiles de expresión génica similares. El algoritmo comienza colocando aleatoriamente genes o muestras dentro de *clusters*. Por tanto, corridas diferentes de este algoritmo pueden arrojar resultados ligeramente diferentes.

Algoritmo

En general, el algoritmo de agrupamiento por *K-means* consta de seis pasos:

1. Se escoge un número de *clusters* denotado por k.
2. Aleatoriamente, se asigna cada perfil de expresión a uno de los k *clusters*.
3. Se calculan los centroides de cada uno de los k *clusters*.
4. Para cada perfil en turno, se calcula la distancia entre él y cada uno de los centroides de los k *clusters*.
5. En caso de que el perfil sea más cercano a un *cluster* diferente de aquel al que actualmente pertenece, se asigna dicho perfil al nuevo *cluster* y se calculan de nuevo los centroides de ambos grupos.
6. Regresa al paso cuatro hasta que ningún perfil cambie de *cluster*.

Para escoger un buen valor de k hay dos posibles alternativas. La primera, es totalmente empírica y consiste en probar diversos valores y escoger la que mejor se ajuste de acuerdo a nuestra percepción y conocimiento del problema. La segunda, se apoya una técnica de escalamiento multidimensional (MDS), que lleva a cabo una reducción de la dimensionalidad que además resulta ser buena para la visualización. Se mide la distancia entre perfiles usando algún método de los descritos y posteriormente, se intenta localizar los perfiles en un espacio de dos o tres dimensiones de tal manera que las distancias son dentro de lo posible, las más cercanas a las distancias medidas entre los perfiles en el espacio de dimensiones superiores.

Capítulo 4

Algoritmos de Normalización

4.1. Métodos de información total

4.1.1. Normalización por Loess Cíclico

Este enfoque está basado en los gráficos MA, donde M es la diferencia entre los valores logarítmicos de expresión y A es el promedio de estos mismos[28]. En un gráfico MA con datos normalizados debe observarse una nube de puntos dispersos alrededor del eje $M = 0$. Es aplicado a las señales de intensidad de dos arreglos simultáneamente[27].

Para cualesquiera dos arreglos i, j con señales de intensidad x_{ki} y x_{kj} donde $k = 1, ..., p$ representa el número de sonda, se calculan $M_k = \log_2(\frac{x_{ki}}{x_{kj}})$ y $A_k = \frac{1}{2}\log_2(x_{ki}x_{kj})$. Con estos datos, una curva normalizada sobre el gráfico MA es ajustada usando Loess, un método de regresión local. Los ajustes en que se basa la curva de normalización son $\hat{M}_k$, y por tanto, el ajuste de normalización respectivo es $M'_k = M_k - \hat{M}_k$. Las señales de intensidad ajustadas están dadas por $x'_{ki} = 2^{A_k + \frac{M'_k}{2}}$ y $x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$. Las curvas de normalización pueden ser generadas en la computadora usando un rango de conjuntos invariantes de sondas.

Para lidiar con más de dos arreglos, el método puede ser extendido al contemplar todas las distintas combinaciones por pares; de manera que la normalización es llevada a cabo recordando el ajuste de cada uno de los dos arreglos pareados. Una vez revisados todos los pares de arreglos, se tiene un conjunto de ajustes que pueden ser aplicados al conjunto de arreglos, se realiza, y se repite nuevamente el proceso. Generalmente, son necesarias sólo una o dos iteraciones completas a través de todas la combinaciones para lograr buenos resultados. Dado que este método trabaja de manera pareada, es algo tardado.

Loess es un método de regresión local relativamente moderno, basado en otros más clásicos, tales como la regresión lineal y no lineal por mínimos cuadrados[29]. Combina la simplicidad de dicha regresión lineal con la flexibilidad de la no lineal. Hace esto ajustando modelos simples en subconjuntos localizados de los datos para construir una función que describe la parte determinística de la variación en los datos punto por punto. De hecho, una de las principales atracciones de este método es que no es requerido un análisis de los datos para especificar una función global que ajuste un modelo a los datos, si no sólo para ajustar segmentos de éstos.

Una desventaja de este método y de otros similares, es el aumento en el consumo de recursos computacionales, por lo que han sido diseñados para usar conscientemente la actual capacidad de cómputo con la mayor ventaja posible para alcanzar los objetivos no alcanzados fácilmente por los enfoques tradicionales.

Este método fue propuesto originalmente por Cleveland en 1979 y posteriormente desarrollado por Cleveland y Devlin en 1988[29], específicamente denota un método que es descrito como una regresión polinomial ponderada y local. En cada punto en el conjunto de datos, un polinomio de bajo grado es ajustado a un subconjunto de datos, con valores de una variable explicativa cercanos al punto cuya respuesta está siendo estimada. El polinomio es ajustado usando mínimos cuadrados ponderados, dando más peso a los puntos cercanos al punto cuya respuesta está siendo estimada y menos conforme éstos se

alejen. El valor de la función de regresión para el punto es entonces obtenido al evaluar el polinomio local usando los valores de la variable explicativa para esos puntos de los datos.

Los subconjuntos de datos usados para cada ajuste por mínimos cuadrados ponderados en Loess, son determinados por un algoritmo de identificación de vecinos cercanos. Este algoritmo necesita una entrada específica llamada parámetro de suavizamiento que determina cuantos de los datos son usados para ajustar cada polinomio local. El parámetro de suavizamiento, q, es un número entre $(\frac{d+1}{n})$ y 1, donde d denota el grado de polinomio local. El valor de q es la proporción de los datos usada en cada ajuste. El conjunto de datos usado en cada ajuste por mínimos cuadrados ponderados se compone de nq puntos (redondeando al entero más próximo) cuyos valores de la variable explicativa son más cercanos a los puntos en los cuales la respuesta está siendo estimada.

q es llamado parámetro de suavizamiento porque controla la flexibilidad de la función de regresión de Loess. Valores altos de q producen las funciones más estables que se mueven menos en respuesta a fluctuaciones en los datos. Una q más pequeña es más cercana a la función de regresión que conformará a los datos y una muy pequeña no es deseable. Sin embargo, dado que la función de regresión eventualmente empezará a capturar el error aleatorio en los datos. Usualmente los valores del parámetro de suavizamiento se fijan entre un rango de $0{,}25$ y $0{,}5$ para la mayoría de las aplicaciones de Loess.

Los polinomios locales ajustados a cada subconjunto de los datos casi siempre son de primer o segundo grado. Usando un polinomio de grado cero, Loess se convierte en una media móvil ponderada. En algunas situaciones modelos simples pueden funcionar aunque posiblemente no siempre aproximen a la función subyacente como se desea. Polinomios de mayor grado trabajarían eficazmente en teoría, pero el campo de los modelos no es el fuerte de Loess. Éste, está basado en ideas que implican que cualquier función puede ser aproximada con buena precisión en una pequeña vecindad por un polinomio de bajo grado y que modelos simples pueden ajustar los datos fácilmente. Polinomios de alto gra-

do tienden a sobreajustar los datos en cada subconjunto y ser numéricamente inestables, provocando dificultades en la precisión del proceso computacional.

Como ya se mencionó, la función de peso proporciona más peso a los puntos cercanos al punto cuya respuesta está siendo estimada y menos conforme estos se alejen. El uso de los pesos está basado en la idea de que puntos cercanos entre sí en el espacio de la variable explicativa, son más propensos a ser relacionados de una forma simple que aquellos que están más separados. Siguiendo esta lógica, puntos que son más propensos a seguir el modelo local influyen mejor en la estimación de los parámetros locales.

La función tradicional de peso usada para Loess es la siguiente:

$$w = \begin{cases} (1-|x|^3)^3 & \text{para} \quad |x| < 1 \\ 0 & \text{para} \quad |x| \geq 1. \end{cases} \tag{4.1}$$

Sin embargo, alguna otra función de peso que satisfaga las propiedades listadas en el trabajo de Cleveland (1979) podría ser usada. El peso para un específico punto en algún subconjunto localizado de datos es obtenido al evaluar la función de peso sobre la distancia entre ese punto y el punto de estimación, después, establecer la escala de manera que la distancia máxima absoluta en todos los puntos en el subconjunto sea exactamente una.

4.1.2. Normalización por Cuantiles

El objetivo de este método es hacer que la distribución de las señales de intensidad para cada microarreglo en un conjunto de éstos sea la misma. Este método parte de la idea de que en un gráfico $Q-Q$ se muestra que la distribución de dos vectores de datos es la misma si éste define una recta diagonal y no lo es, en cualquier otro caso. Este concepto puede ser extendido a n dimensiones, por tanto, si todos los vectores de datos tienen la misma distribución, entonces, al graficar los cuantiles en n dimensiones se obtiene una

línea recta alrededor de la línea dada por el vector unitario $(\frac{1}{\sqrt{n}},...,\frac{1}{\sqrt{n}})$. Esto sugiere que es posible hacer un conjunto de datos con la misma distribución si se proyectan los puntos del gráfico $Q-Q$ de dimensión n sobre la diagonal[27].

Sea $q_k = (q_{k1},...,q_{kn})$ donde $k = 1,...p$ es el vector del k-ésimo cuantil para todos los n arreglos de longitud p, $q_k = (q_{k1},...,q_{kn})$ y $d = (\frac{1}{\sqrt{n}},...,\frac{1}{\sqrt{n}})$ la diagonal unitaria. Para lograr que todos los cuantiles se sitúen alrededor de la diagonal, se considera la proyección de q sobre d,

$$proj_d q_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj},...,\frac{1}{n}\sum_{j=1}^{n} q_{kj}\right). \tag{4.2}$$

Esto implica que se puede asignar a cada arreglo la misma distribución al tomar la mediana de los cuantiles y sustituirla como el valor de los datos de partida en el conjunto original. De aquí, la realización del siguiente algoritmo para normalizar un conjunto de vectores de datos a través de la asignación de una misma distribución.

1. Dados n arreglos de longitud p, Se forma $\overline{X}$ de dimensión pxn donde cada arreglo es una columna.
2. Se ordena cada columna de $\overline{X}$ para obtener $\overline{X}_{sort}$.
3. Se toman las medias de las filas de $\overline{X}_{sort}$ y se asignan a cada elemento de $\overline{X}'_{sort}$.
4. Obtener $\overline{X}_{normalized}$ al reordenar cada columna de $\overline{X}'_{sort}$ para tener el mismo orden que la original $\overline{X}$.

El método Cuantiles es una caso específico de la transformación $x'_i = F^{-1}(G(x))$, donde G se estima a través de la distribución empírica de cada microarreglo y F usando la distribución empírica de los cuantiles promediados de la muestra. Extensiones de este método pueden ser implementadas cuando F^{-1} o G son estimadas más suavemente.

Un problema con este método podría ser que fuerza los valores de los cuantiles a ser iguales. Esto sería más problemático en las colas de las distribuciones donde es posible que una sonda tenga el mismo valor a través de todos los microarreglos. Sin embargo, en la práctica dicho problema no tiene mucha relevancia.

4.1.3. Normalización por Contraste

Así como el método Loess, el método Contraste está basado en los gráficos MA[27][30]. Se parte de un conjunto de k arreglos para ser normalizados, cada arreglo es representado por n intensidades de señal.

Sea la matriz Y de tamaño nxk aquella que denota las intensidades de estos arreglos. Entonces, el elemento de la fila i y la columna j de Y es la sonda de interés no normalizada i en el arreglo j.

En el primer paso, estas intensidades una vez con la aplicación del logaritmo son transformadas usando una matriz M ortonormal de tamaño kxk (las filas de la matriz M son vectores unitarios mutuamente ortogonales),

$$Z = [x, y_1, ..., y_{k-1}] = log(Y)M', \tag{4.3}$$

además, la primera fila de M contiene siempre un vector con 1 en sus elementos multiplicado por $\sqrt{\frac{1}{k}}$, mientras que las otras filas son un conjunto de contraste ortonormal. M es una matriz de transformación y se aplica un cambio de base donde los renglones de M forman la nueva base denotada como base alternativa durante el resto de la explicación del método La transformación logarítmica que se hacen antes del cambio de base hace que el error de varianza se más homogeneo.

Usando la base alternativa, se estima la curva de normalización usando la primera columna de Z como predictor para la columna $2,..,k$ de esta misma matriz $y_1,...,y_{k-1}$. Teniendo en cuenta que el conjunto ortonormal de contrastes no es único. Por tanto, el método para estimar la curva debe ser invariante con respecto a la elección de contraste. A razón de lograr esto, se estima una curva usando un modelo de regresión local (Loess) para cada vector y_i. Para que la curva sea sensible a datos atípicos, se usa un *redescending M estimator* con la raíz cuadrada de la función de peso como se lleva a cabo en la función de *R* para Loess, pero con una importante modificación. Si $\hat{y}_1,...,\hat{y}_{k-1}$ son los vectores de los valores estimados, se toma a $\hat{\varepsilon}$ como la distancia euclidiana entre las filas de las $n(k-1)$ matrices $[y_1,...,y_{k-1}]$ y $[\hat{y}_1,...,\hat{y}_{k-1}]$ (vectores unitarios mutuamente ortogonales),

$$\hat{\varepsilon} = \sqrt{\sum_{i=1}^{k-1} (\hat{y}_i - y_i)^2}. \tag{4.4}$$

Entonces en cada iteración, el mismo conjunto de pesos robustos es usado para cada uno de los $k-1$ vectores de contraste y esos pesos son invariantes a la elección del conjunto ortonormal de contrastes, de manera que la curva ajustada es igualmente invariante a la elección del conjunto ortonormal de contrastes.

La normalización se puede llevar a cabo, sustrayendo a la curva de normalización representada con la matriz $[x,\hat{y}_1,...,\hat{y}_{k-1}]$ los valores estimados usando la base alternativa y posteriormente regresando a la base original a través de la matriz M. En este caso la intensidades normalizadas serían

$$exp\{[x, y_1 - \hat{y}_1, ..., y_{k-1} - \hat{y}_{k-1}]\, M\}, \tag{4.5}$$

pero esto resulta en un proceso de normalización no suave, en el sentido de que las intensidades iguales en un arreglo posterior a la normalización pueden no serlo después. No obstante, se puede seguir usando la curva de normalización con la base alternativa. La matriz $[x,\hat{y}_1,...,\hat{y}_{k-1}]$ es una representación de la curva con la base alternativa y la

matriz $[x,0,...,0]$ es una representación de lo que la curva debería de ser después de la normalización. Por lo tanto, el mapeo

$$[x, y_1 - \hat{y}_1, ..., y_{k-1} - \hat{y}_{k-1}] \longrightarrow [x,0,...,0], \tag{4.6}$$

define la transformación que hace el trabajo de equilibrar el contraste para la base alternativa. Además, el mapeo

$$exp\{[x, y_1 - \hat{y}_1, ..., y_{k-1} - \hat{y}_{k-1}]\ M\} \longrightarrow exp\{[x,0,...,0]\ M\}, \tag{4.7}$$

define la misma transformación pero para la base original y en escala antilogarítmica. Esta transformación forma una función $F : R^k \to R^k$ que renglón por renglón normaliza la matriz de intensidades Y. En consecuencia, si $F((x_1,...,x_k)) = (f_1(x_1),...,f_k(x_k))$, f_j es la función que normaliza el renglón j.

Nótese que:

$$[x,0,...,0]\ M = \frac{1}{\sqrt{k}}\ [x,x,...,x]$$

donde $\frac{1}{\sqrt{k}}$ es $\overline{\log(Y)}$, es decir, la media a través de los renglones de $\log(Y)$. Por tanto, este procedimiento normaliza a la escala determinada por medio de $exp\{\overline{\log(Y)}\}$, o sea, la media geométrica de los arreglos.

4.2. Métodos de información parcial

4.2.1. Normalización por Constante

En este método se escoge un arreglo de referencia de entre n arreglos. Los $n-1$ arreglos restantes son normalizados en base al arreglo seleccionado.

Sea $\bar{x}_{base}$ la intensidad media de el arreglo seleccionado y $\bar{x}_i$ con $(i = 1, 2, ...n)$ la intensidad media de los otros arreglos. Entonces, se calculan las intensidades normalizadas de la siguiente manera:

$$x'_i = \beta_i x_i, \tag{4.8}$$

donde $\beta_i = \frac{\bar{x}_{base}}{\bar{x}_i}$ para $i = 2, 3, ..., n$.

Los métodos lineales son predominantemente medias de normalización de microarreglos. En escala, la aproximación lineal simple asume una relación lineal que pasa por el origen. Fuerza las distribuciones de los arreglos a tener la misma tendencia central (media aritmética, media geométrica, mediana), lo que puede ser realizado por un factor de escala y ha sido uno de los métodos seleccionados por Affymetrix[32].

4.2.2. Normalización por Conjunto invariante

Este método, lleva a cabo la normalización de un grupo de microarreglos utilizando otro como referencia común. Esta relación de normalización puede ser interpretada como una curva en el gráfico de dispersión de dos microarreglos, en donde el arreglo de referencia es representado por el eje y y el arreglo a ser normalizado en el eje x.

La normalización debe basarse únicamente en las sondas con valores de intensidad que pertenecen a los genes no diferencialmente expresados, que generalmente no se conocen. No obstante, se espera que la sonda de un gen no diferencialmente expresado tenga rangos de intensidad similares. Estas sondas son identificadas a través de un proceso iterativo llamado Conjunto invariante[31], el cual presumiblemente está compuesto de puntos de genes no diferencialmente expresados. Específicamente, se empieza con puntos de todas las sondas PM. Si la proporción de los rangos de diferencia de los puntos (PRD, que es la diferencia absoluta de rangos en dos arreglos dividida entre el número total de sondas n) es suficientemente pequeña, el punto es mantenido en el conjunto,

$$PRD_{ki} = \left\| \frac{r_{k,i} - r_{k,base}}{K} \right\|, \tag{4.9}$$

donde $r_{k,j}$ es el rango de la señal del $k_{\text{ésimo}}$ PM en el arreglo i y $r_{k,base}$ el rango de la señal de el mismo PM en el arreglo de referencia. El rango de proporción de la diferencia de puntos (PRD) es calculado para toda sonda $k = 1,2,..K$.

Empíricamente se puede establecer, por ejemplo, un umbral $PRD < 0{,}003$ cuando el promedio de intensidades de rangos es pequeño en los dos arreglos, mientras que $PRD < 0{,}007$ cuando es grande. A partir de ambos se interpola un nuevo umbral intermedio y se repite el proceso iterativamente hasta que el conjunto no cambia. Una vez formado dicho conjunto invariante, es usado como base para ajustar una curva de normalización, la cual pasa a través de la mediana de las intensidades de las sondas seleccionadas.

4.2.3. Normalización por Qsplines

En este método se escoge un arreglo de referencia ν, el cual es calculado a partir de la media geométrica de cada sonda sobre los n arreglos,

$$\nu_k = \left(\prod_{i=1}^{n} x_{ki} \right)^{\frac{1}{n}}, \tag{4.10}$$

donde $i = 1,2,...,n$ son los microarreglos y $k = 1,2,...,K$ las sondas. De manera que el arreglo de referencia está compuesto por información de todos los arreglos analizados.

Para la normalización, el método usa los cuantiles de las señales de los arreglos y blancos (x y ν) para estimar *B-splines* de suavizamiento[32]. De cada arreglo y el vector ν se toman 100 cuantiles, q_i y q_ν (percentiles). Cada q_ν, q_i es usado para ajustar una función de *spline* cúbica, $s_i = f(q_\nu, q_i)$, donde f es la función generadora de *splines* que ajusta los parámetros de un spline cúbico natural (*B-spline*). Los parámetros de los *splines* se

ajustan para cada intervalo entre cuantiles consecutivos. La función de interpolación de *splines* definida sobre el $j_{\text{ésimo}}$ intervalo, es definida por los parámetros a_{ij} y y_{ij} en la ecuación 4.11. Para un *spline* cúbico no suavizado, $y_{ij} = q_{vj}$,

$$s_j(x) = a_{ij_1} + a_{ij_2}(x - y_{ij}) + a_{ij_3}(x - y_{ij})^2 + a_{ij_4}(x - y_{ij})^3 \tag{4.11}$$

Los *splines* son entonces usados como funciones de normalización señal-dependientes en las señales de x.

Las señales de los blancos pueden provenir de otro arreglo o pueden ser medias calculadas a partir de múltiples arreglos. Los *splines* son una elección natural y robusta que permite la representación de casi cualquier relación suave y además funciona bien si los datos tienen una relación lineal. Usando la información de los cuantiles es más sencillo el problema del ajuste pues lo evita directamente ajustando los datos por pareo, lo cual en general requiere de robustas técnicas de regresión[32].

Capítulo 5

Aplicación: Proyecto real

Nombre del proyecto.

Perfil de expresión diferencial de genes en piel y cérvix del ratón transgénico $K14E6$ mediante el uso de microarreglos de alta densidad[35].

5.1. Introducción

Para estudiar la progresión neoplásica inducida por VPH, se expresó la región temprana ($E6$) del virus del papiloma humano tipo 16 ($VPH16$) en las células basales del epitelio (piel, cérvix, vagina, etc.) en un ratón transgénico, usando el promotor de la queratina humana $K14$.

En el ratón transgénico $K14E6$, se encontró un aumento en la proliferación de las células de la piel (por lo tanto un engrosamiento en ésta), además de la aparición espontánea de tumores de piel en el ratón adulto, en su mayoría malignos. Sin embargo, la expresión de la oncoproteína E6 en el ratón $K14E6$ no fue suficiente para inducir cáncer en el tracto reproductivo (cérvix y vagina).

Los diferentes fenotipos que induce $E6$ en piel del ratón transgénico $K14E6$ incitan a plantear diferentes preguntas táles como, ¿qué otras funciones además de inducir la degradación de $p53$ realiza $E6$? ¿Qué genes están involucrados en el efecto diferencial de $E6$ en piel y cérvix del ratón $K14E6$? Para responder a estas preguntas, se piensa realizar un perfil de expresión de genes en piel y cérvix del ratón transgénico $K14E6$ mediante el uso de microarreglos.

Lo que se pretende realizar con los microarreglos, es encontrar los genes diferencialmente expresados en piel del ratón $E6$ con respecto al cérvix que puedan explicar su fenotipo. Además se tienen que comparar los genes de piel FvB con los de piel $E6$ para descartar los genes que permanezcan sin cambio. Lo mismo se pretende realizar con el cérvix para ambos ratones. Por último se pide realizar una comparación entre piel y cérvix del ratón $E6$ y después compararlos con genes de piel y cérvix de FvB. En resumen las comparaciones que se realizarían serían las siguientes:

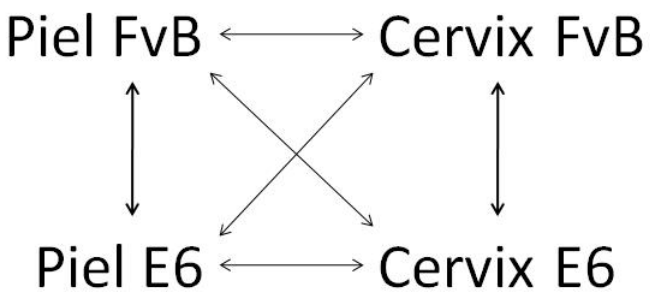

Figura 5.1: contrastes a realizar en el perfil de expresión. Piel FvB vs Piel E6, Piel FvB vs Cérvix FvB, piel FvB vs Cérvix E6, Piel E6 vs Cérvix FvB, Piel E6 vs Cérvix E6, Cérvix FvB vs Cérvix E6.

5.2. Datos

Los datos utilizados para el proyecto provienen de ARNm de dos tejidos distintos, en dos diferentes modelos murinos (ratón). En total cuatro grupos de tratamientos. A su vez cada uno de éstos contiene dos replicados técnicos (misma muestra ARNm en dos microarreglos) y dos replicados biológicos (diferente muestra de ARNm en dos microarreglos). En total, 16 muestras. Los microarreglos Affymetrix utilizados son $Mouse430a2$, las secuencias se seleccionaron del banco de datos GenBank de NCBI, dbEST y RefSeq. Los conjuntos de sondas que determinan la expresión de cada gen están formados por $16-20$ pares de oligonucleótidos para la detección de transcritos. El diseño de la tecnología Affymetrix incluye aproximadamente $14,000$ genes de ratón perfectamente bien caracterizados, controles y *Housekeeping genes* distribuidos a lo largo de todo el microarreglo, estos controles pueden o no tener interés biológico pero son de radical importancia

por diferentes razones. Antes de procesar el microarreglo, se verifica la calidad de las muestras de ARNm.

Una vez que los microarreglos son escaneados, la imagen obtenida en el experimento de microarreglos se almacena en archivos con formato *.CEL*. El resultado del escáner de la consola Affymetrix es sometido a un control de calidad, después, los archivos son capturados con el paquete estadístico R versión $2.11.1\ (2010-05-31)$ utilizando los recursos de *bioconductor*; un proyecto para el desarrollo de fuentes de recursos y herramientas de software libre para el análisis y la comprensión de datos genómicos. Por medio de *bioconductor*, Affymetrix provee información de secuencias y anotaciones de carácter genómico. En este experimento, las anotaciones requeridas para iniciar el análisis son las siguientes:

Modelo biológico: CINVESTAV-RATÓN TRANSGÉNICO $K14E6$.
Anotaciones: Mouse 430 A $2{,}0$.
Tamaño: $22{,}690$ genes por arreglo.
Controles: Desde la sonda $22{,}627$ a la $22{,}690$.
Controles de Hibridación: $BioB$, $BioC$, $BioD$, Cre.

5.3. Análisis

5.3.1. Pre-procesamiento

La fase de pre-procesamiento es crucial para la confiabilidad en los resultados ya que involucra la observación a detalle de los datos crudos. Es en este paso donde se tomarán las decisiones de los algoritmos para corrección y/o normalización. El análisis de datos crudos se realiza a nivel intensidad de señal por sonda. Primero nos enfocamos en el panorama general de todos los datos involucrados en el proyecto mediante curvas de distribución (Figura 4.2) y con otra perspectiva los gráficos de cajas (Figura 4.3). Ambas nos permiten detectar sesgos, verificar calidad de replicados y nos servirán para medir el efecto de corrección y normalización. Por otro lado, nos enfocamos en los controles de hibridación porque precisamente son diseñados para detectar problemas durante el

proceso de hibridación en el microarreglo y un error a este nivel conlleva a una posible eliminación de la muestra en el análisis (ver sección 3.3.). La corrección del ruido de fondo se refiere a la detección y de ser el caso, eliminación de señal falsa o bien no específica (ver sección 3.3.1.). Finalmente, la aplicación de algunos de los principales algoritmos de normalización, tema principal de esta tesis.

Exploración de datos crudos

La figura 5.2 y 5.3 muestran la distribución de datos crudos de los 16 microarreglos utilizados en el experimento. La gran mayoría de los genes tiene niveles de intensidad bajos, mientras que los extremos de la cola superior de las distribuciones contienen aquellos diferencialmente expresados. Se observa un ligero desplazamiento horizontal entre las distribuciones así como una densidad heterogénea. En el gráfico de cajas también se expone la necesidad de normalización. Si remarcamos las diferencias entre las distribuciones tanto entre los replicados técnicos como biológicos, las más notables corresponden al replicado biológico 1 de piel E6, al replicado biológico 2 de cérvix E6 e incluso al replicado técnico de piel FvB. La figura 5.4 y 5.5 en conjunto con las anteriores, proveen un panorama más claro en relación a estas variaciones (ver replicado biológico 2 de piel E6). La razón de estos comportamientos reside en las fuentes de variación oscura mencionadas (ver sección 3.3.2.). Los gráficos MA, con frecuencia son utilizados como guía para algunos métodos de normalización. Dado que M es la diferencia entre los valores logarítmicos de expresión y A es el promedio de estos mismos, $M = \log_2(x_{ki}) - \log_2(x_{kj})$ y $A = \frac{\log_2(x_{ki}) + \log_2(x_{kj})}{2}$ donde $k = 1, ..., p$ representa el número de sonda para cualesquiera dos microarreglos $i, j = 1, ..., n$ con señales de intensidad x_{ki} y x_{kj}, en datos normalizados un gráfico M contra A debe contener la mayor densidad de puntos dispersos en forma simétrica alrededor del eje $M = 0$. Como estos gráficos se crean a partir de la información de un par de microarreglos, en este trabajo únicamente se presentan algunas comparaciones de importancia. La figura 5.6, muestra como la nube de datos dispersos evidencía relaciones no lineales (Muestra 1 y 2 o 7 y 8) o bien diferencias de intensidad que se alejan de $M = 0$ por un factor de escala (Muestra 13 y 14 o 15 y 16).

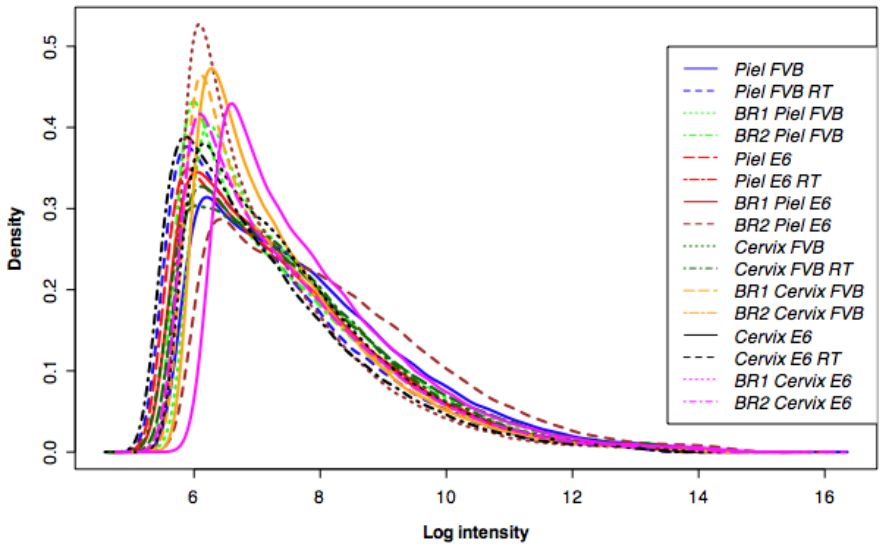

Figura 5.2: **Función de densidad de datos crudos (intensidad de señal a nivel de sonda) de todos los microarreglos utilizados para el análisis**. Se aplicó a los datos una transformación logarítmica con base 2. *RT*=Replicado técnico, *BR*1=Replicado biológico 1, *BR*2=Replicado biológico 2.

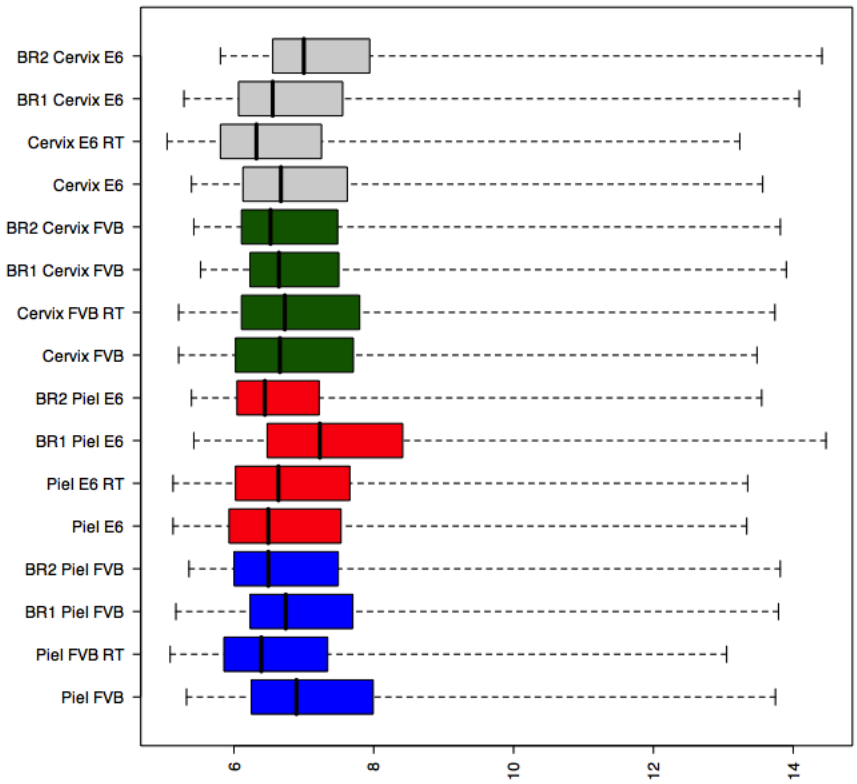

Figura 5.3: **Gráfico de cajas de los datos crudos de los** 16 **microarreglos utilizados para el análisis**. La medida de tendencia central es la mediana de las intensidades de señal a nivel de sonda de cada microarreglo. *RT*=Replicado técnico, *BR*1=Replicado biológico 1, *BR*2=Replicado biológico 2.

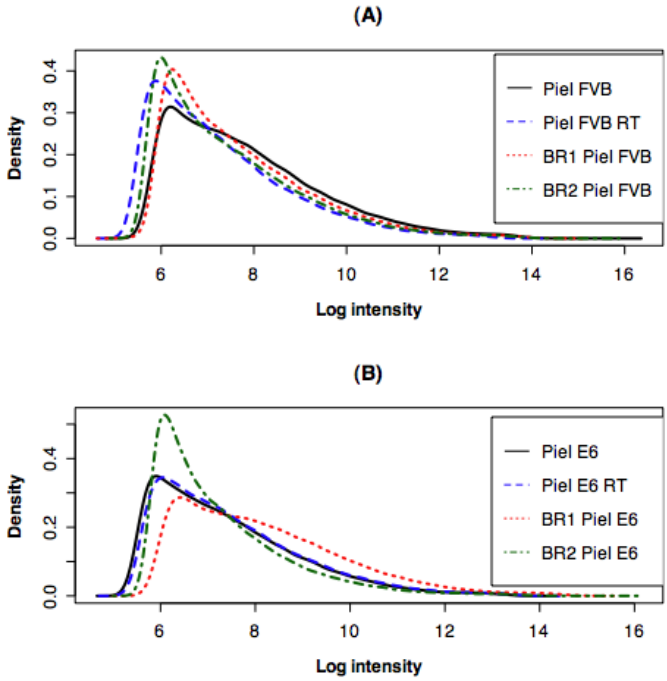

Figura 5.4: **Función de densidad de datos crudos por tipo de tejido**. Los datos de ambas gráficas provienen de las muestras de ARNm de células de piel. (A) Microarreglos de tejido de piel FvB (muestra), piel FvB *RT* (replicado técnico), piel FvB *RB*1 (replicado biológico 1) y piel FvB *RB*2 (replicado biológico 2). (B) Microarreglos de tejido de piel E6, piel E6 *RT*, piel E6 *RB*1 y piel E6 *RB*2.

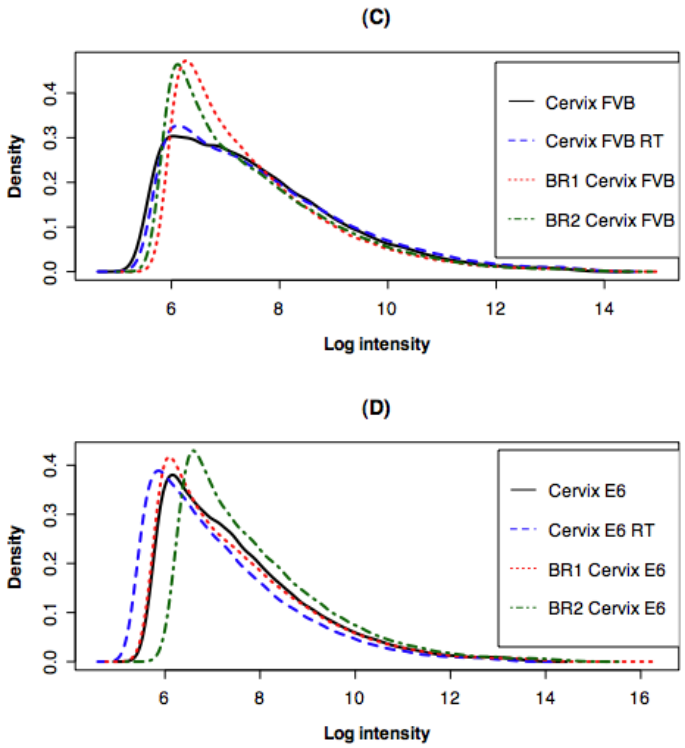

Figura 5.5: **Función de densidad de datos crudos por tipo de tejido**. Los datos de ambas gráficas provienen de las muestras de ARNm de células de cérvix. (C) Microarreglos de tejido de cérvix FvB (muestra), cérvix FvB *RT* (replicado técnico), cérvix FvB *RB*1 (replicado biológico 1) y cérvix FvB *RB*2 (replicado biológico 2). (D) Microarreglos de tejido de cérvix *E*6, cérvix E6 *RT*, cérvix *RB*1 y cérvix E6 *RB*2.

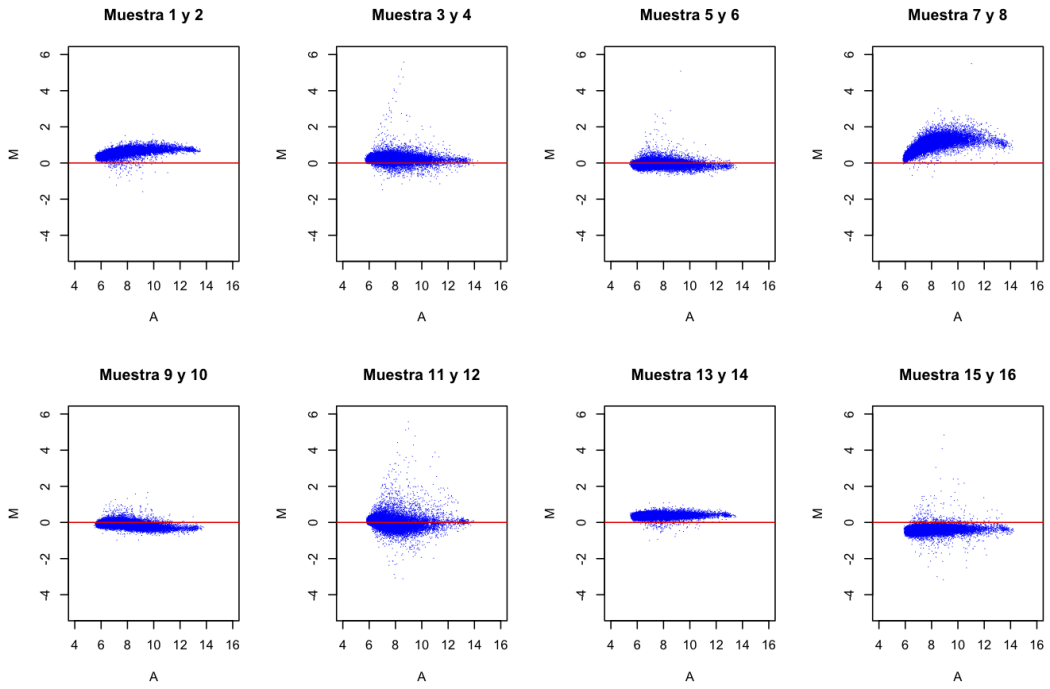

Figura 5.6: **Gráficos *MA* de datos crudos**. *M* es la diferencia de expresión en valores logarítmicos y *A* es el promedio de dichos valores. Las comparaciones presentadas en los encabezados de cada gráfico corresponden a los tratamientos y replicados de acuerdo a las siguientes anotaciones: Muestra 1=Original Piel FVB, muestra 2= Replicado técnico Piel FVB, muestra 3= Replicado biológico 1 Piel FVB, muestra 4= Replicado biológico 2 Piel FVB, muestra 5= Original Piel E6, muestra 6= Replicado técnico Piel E6, muestra 7= Replicado biológico 1 Piel E6, muestra 8= Replicado biológico 2 Piel E6, muestra 9= Original Cérvix FVB, muestra 10= Replicado técnico Cérvix FVB, muestra 11= Replicado biológico 1 Cérvix FVB, muestra 12= Replicado biológico 2 Cérvix FVB, muestra 13= Original Cérvix E6, muestra 14= Replicado técnico Cérvix E6, muestra 15= Replicado biológico 1 Cérvix E6, muestra 16= Replicado biológico 2 Cérvix E6.

Controles de hibridación

Como ya se comentó anteriormente, los controles de hibridación proveen información importante durante la fase de pre-procesamiento. En los gráficos de dispersión que se pre-

sentan, se debe verificar que los respectivos puntos se distribuyan sobre la diagonal. Bajo esta referencia se pueden detectar problemas de hibridación (intensidad de señal ausente o relativamente muy baja) y en general dificultades experimentales.

Según los siguientes cuatro gráficos, el error respecto a la diagonal es mayor entre la muestra y sus replicados biológicos y menor cuando se compara con su replicado técnico. La comparación entre los replicados biológicos de cérvix E6, es evidencia de una relación que se aleja de lo esperado (figura 5.10). Debido a que las combinaciones por pares pueden llegar a ser demasiadas, se grafican las parejas que desde el punto de vista del diseño experimental sean más importantes.

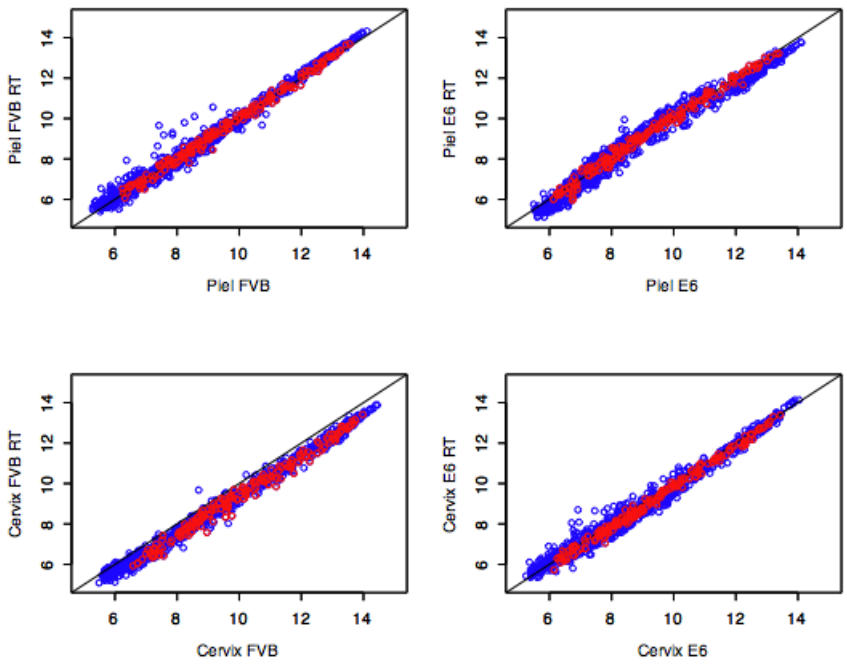

Figura 5.7: **Controles de hibridación BioB, BioC, BioD y Cre de los cuatro tipos de muestra comparados con sus replicados técnicos**. Los datos graficados corresponden a la totalidad de los controles del microarreglo. Los puntos rojos representan los controles de hibridación.

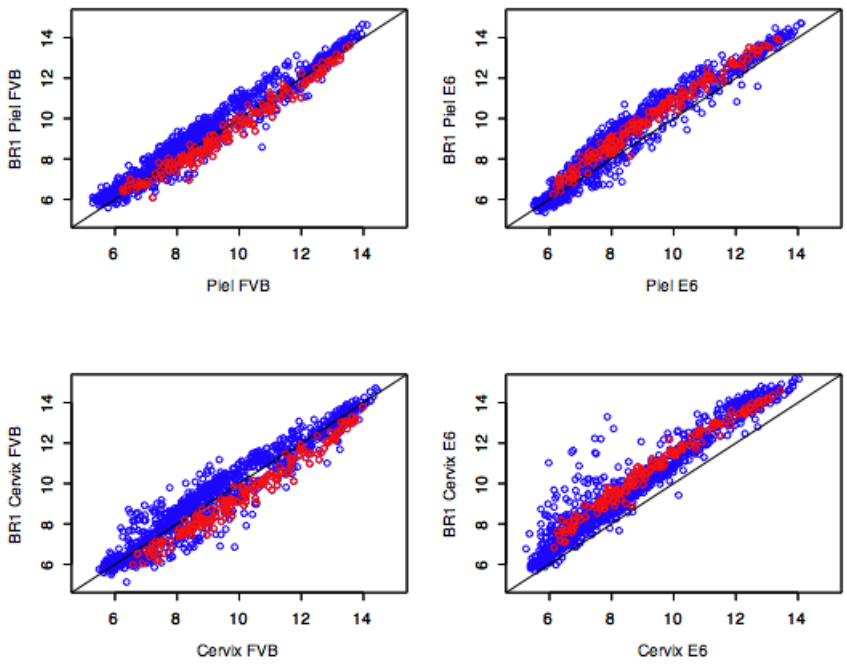

Figura 5.8: **Controles de hibridación BioB, BioC, BioD y Cre de los cuatro tipos de muestra comparados con los replicados biológicos** 1.

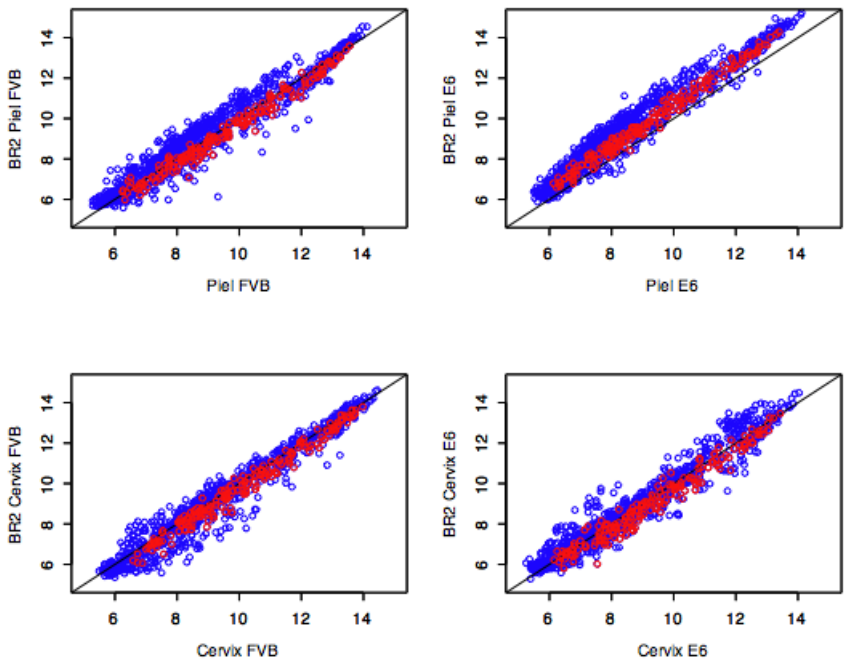

Figura 5.9: **Controles de hibridación BioB, BioC, BioD y Cre de los cuatro tipos de muestra comparados con los replicados biológicos** 2.

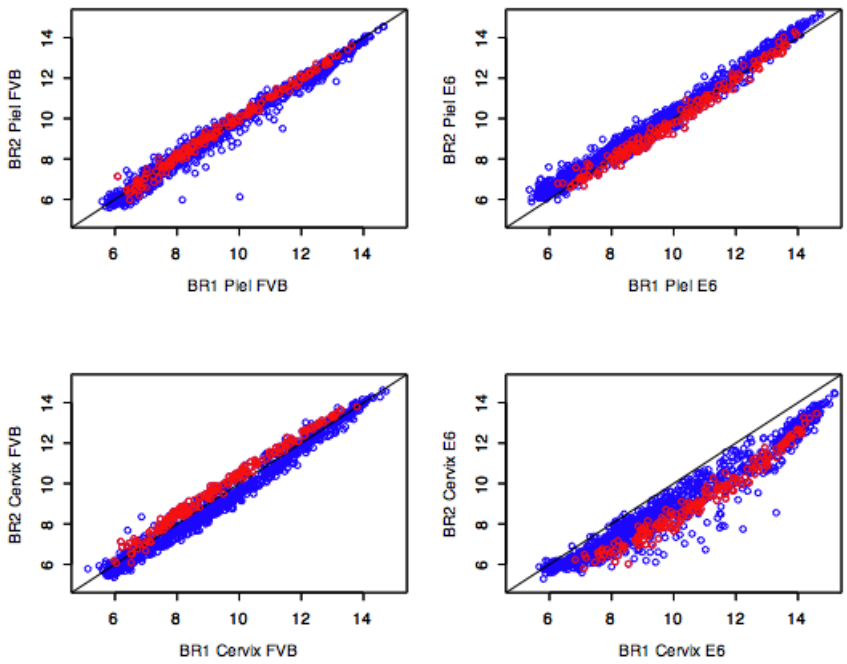

Figura 5.10: **Controles de hibridación BioB, BioC, BioD y Cre de los cuatro tipos de muestra. La comparación es entre los replicados biológicos** 1 **y** 2.

Normalización

El análisis de los diferentes procesos de normalización se dividió en tres secciones, la primera dedicada a los métodos de información total, la segunda a los de información parcial y la última a su aplicación combinada con un algoritmo de corrección de ruido de fondo. Se contempló el efecto sobre los datos después de normalizar por los seis métodos elegidos (Cuantiles, Loess, Contraste, Qsplines, Conjuntos invariantes y Constante), el de aplicar el algoritmo de Cuantiles por separado a los grupos de contraste para posteriormente utilizar cualquier otro método para normalizar el conjunto completo de datos y el de aplicar los algoritmos de corrección de fondo sobre los datos normalizados directamente por los seis métodos que se están comparando.

Normalización por métodos de información total

De acuerdo a los gráficos presentados a continuación, los tres métodos de información propuestos para este análisis reducen notablemente las diferencias entre las distribuciones de los datos (figura 5.11). La gráfica A y B de la misma figura, se diferencía por la forma en la que se aplicó el método de Cuantiles (ver figura 5.13). En estas dos, los datos aún se mantienen a nivel de intensidad de señal a diferencia de C y D que ya se manejan como valores de expresión génica. La figura 5.12, presenta las distribuciones de los datos normalizados en dos etapas, primero se aplicó el método de Cuantiles a los grupos a comparar y consecuentemente Loess, Contraste y nuevamente Cuantiles. No se incluyó el gráfico de distribución por el método de Cuantiles directamente porque la resolución no permite distinguirlo del gráfico A de la figura 5.12. Aunque en términos númericos sí hay cierta variación, es de hecho relativamente pequeña.

Los gráficos de cajas de la figura 5.13, corresponden al mismo proceso de normalización que se realizó en el gráfico A y B de la figura 5.11. Debido a que los diferentes grupos de muestras conservan demasiada variación entre sí, no es recomendable realizar las comparaciones en este punto. Sin embargo, ambos gráficos ayudan a identificar aquellos grupos de microarreglos que se desvían de forma notable del comportamiento general, la presencia de efectos de lote o bien sirven como base para una segunda normalización no tan radical como la de Cuantiles, asunto que en ocasiones es deseable incluso bajo el conocimiento de que un número grande de muestras va de la mano con la eficiencia del método de normalización en cuestión. En estos gráficos (figura 5.11 y 5.13), se observa que los replicados biológicos de piel E6 y cérvix E6 presentan una mayor desviación del comportamiento en relación a los demás grupos de muestras, pero se espera que dicha variación obscura sea corregida al completar la fase de pre-procesamiento.

A partir de los datos normalizados mostrados en la figura 5.14 ya es posible continuar con la elaboración de los contrastes. En general, los datos clasificados como atípicos en cada caja conforman a los genes candidatos a estar diferencialmente expresados. La figura 5.15, contiene los gráficos de cajas de los datos normalizados en dos etapas al igual que

en el a figura 5.12. En ellos, se observa que la distribución de los datos se modifica poco en relación a las normalizaciones directas. La aplicación de un proceso de normalización en dos etapas debe tener una razón experimental que la sustente y por tanto, marcar cierta diferencia en cuanto a los genes encontrados en los correspondientes contrastes, de lo contrario dicha aplicación no es de utilidad y sólo disminuye el poder del primer método utilizado.

Los gráficos correspondientes a la figura 5.16 y 5.17 se presentan para hacer notar el empate casi perfecto entre las distribuciones, producto de la normalización por Cuantiles entre los diferentes tipos de replicados. Por último, se muestran algunos gráficos MA de los datos normalizados por los diferentes métodos estudiados (figura 5.18, 5.19 y 5.20). En los tres métodos de información total bajo análisis, se observan desviaciones de la línea horizontal $M = 0$ (en valores grandes de A se aprecia con mayor facilidad). En los gráficos correspondientes al método de Loess cíclico (figura 5.19) la variación es notablemente mayor. Así mismo, la variación más amplia se continúa presentando en las comparaciones que involucran a los replicados biológicos.

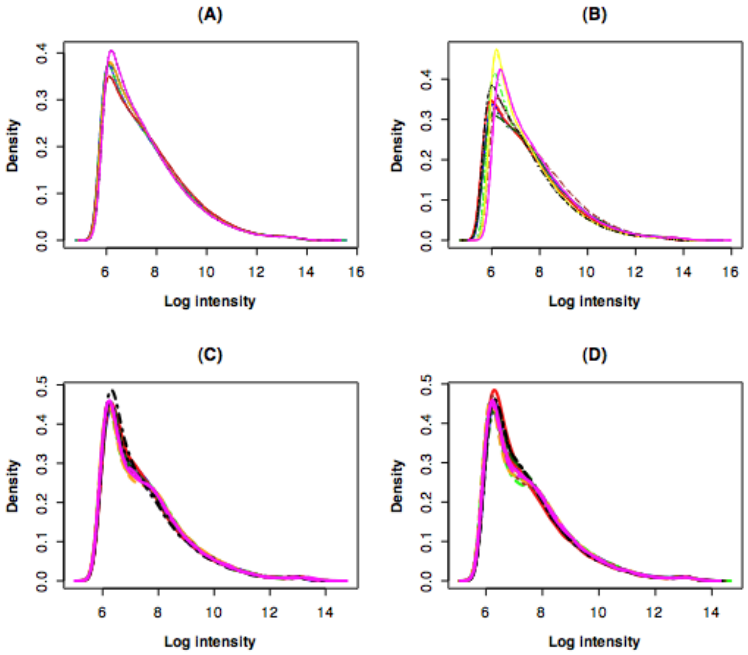

Figura 5.11: **Función de densidad de los datos normalizados por métodos de información total**: (A) Normalización por el método de Cuantiles entre replicados de tejido. (B) Normalización por el método de Cuantiles entre **Muestra-RT** y **RB1-RB2**. (C) Normalización por el método de Loess cíclico. (D) Normalización por el método de Contraste.

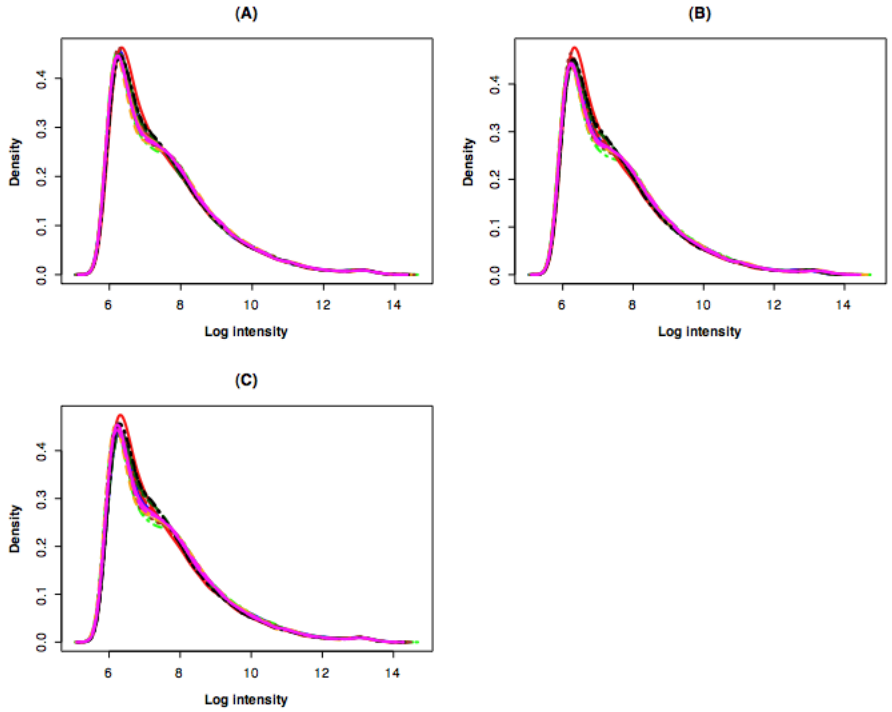

Figura 5.12: **Función de densidad de los datos normalizados por métodos de información total después de normalizar por Cuantiles entre replicados de tejido**: (A) Normalización por el método de Cuantiles. (B) Normalización por el método de Loess cíclico. (C) Normalización por el método de Contraste.

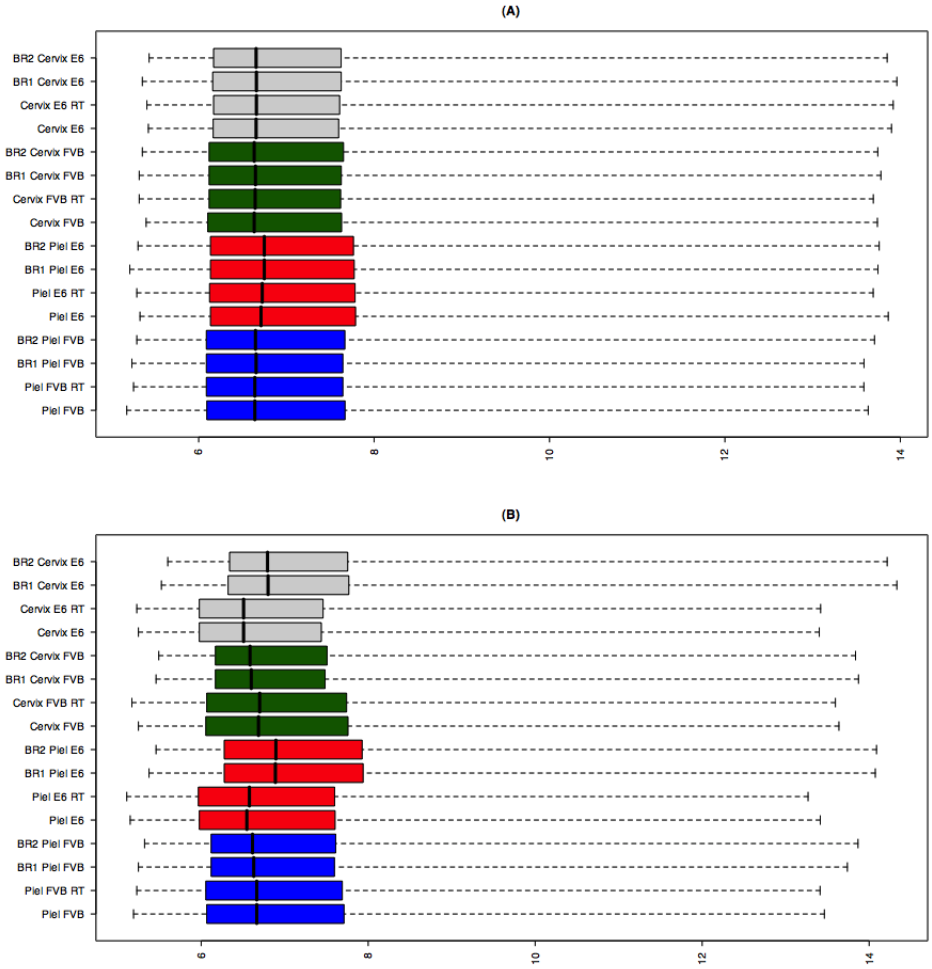

Figura 5.13: **Gráficos de cajas de los datos normalizados por el método de Cuantiles aplicado de dos diferentes formas**: (A) Normalización por el método de Cuantiles. (B) Normalización por el método de Cuantiles entre **Muestra-RT** y **RB1-RB2**.

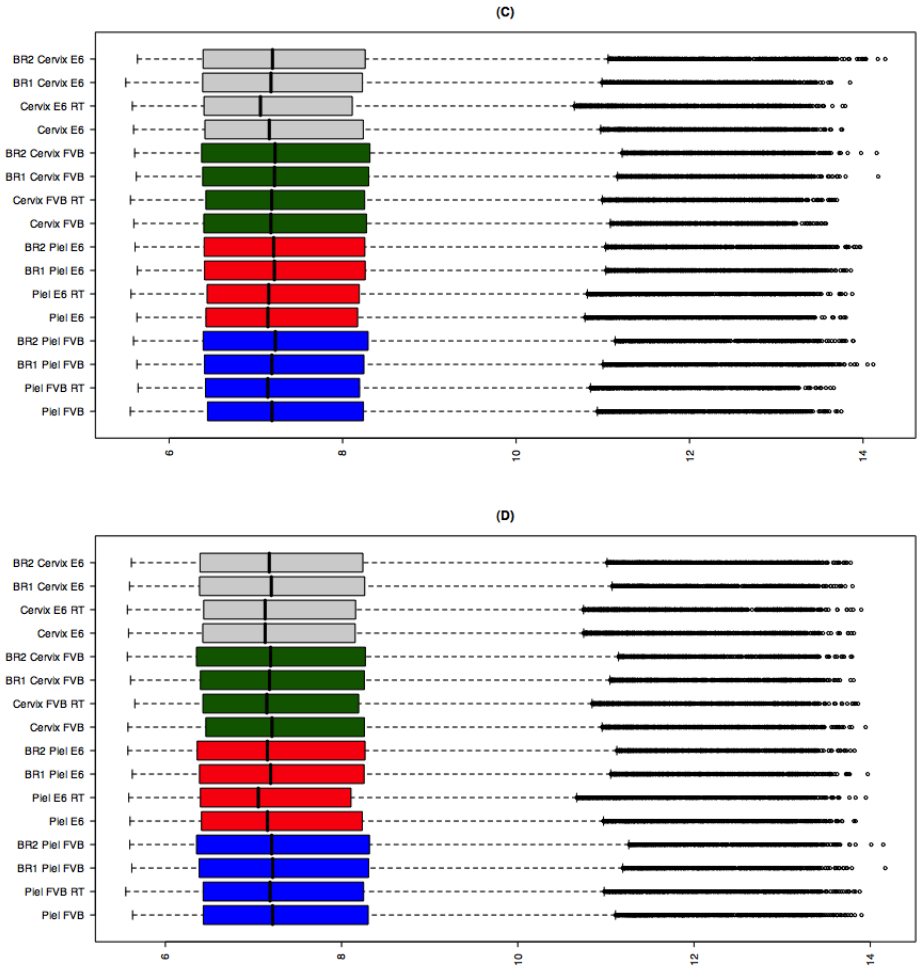

Figura 5.14: **Gráficos de cajas de los datos normalizados por los métodos de información total**: (C) Normalización por el método de Loess cíclico. (D) Normalización por el método de Contraste.

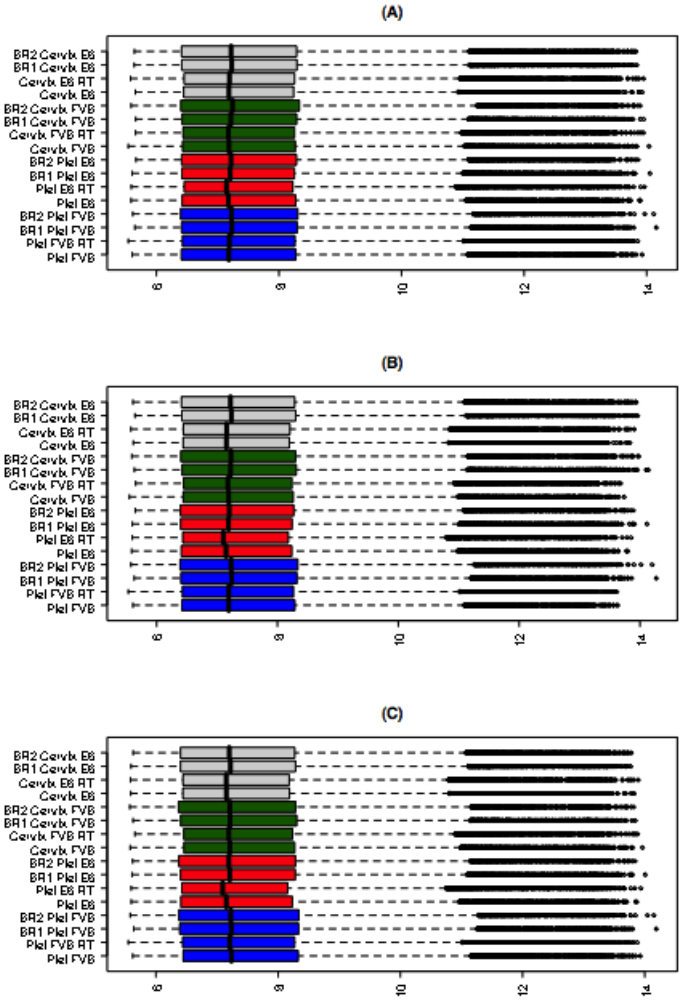

Figura 5.15: **Gráficos de cajas de los datos normalizados por los métodos de información total después de normalizar por Cuantiles los replicados de cada tejido**: (A) Normalización por el método de Cuantiles. (B) Normalización por el método de Loess cíclico. (C) Normalización por el método de Contraste.

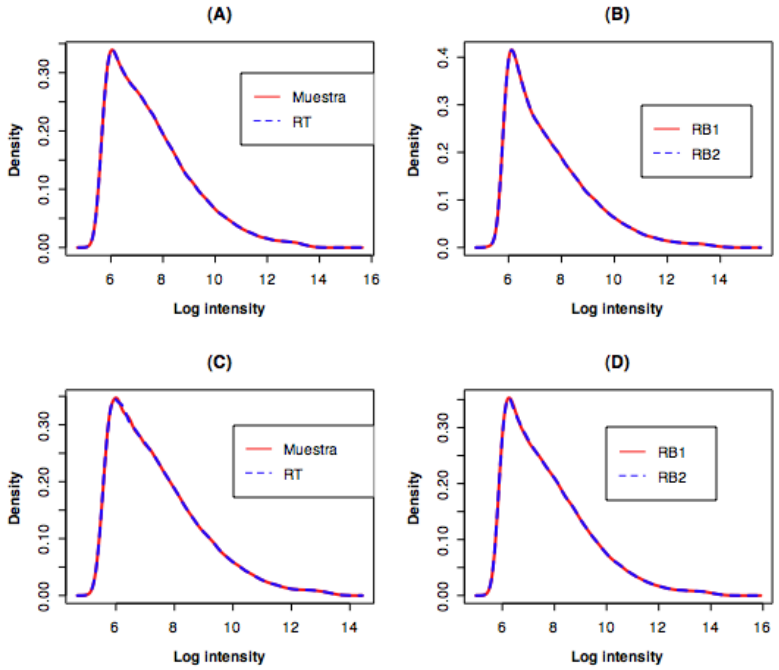

Figura 5.16: **Funciones de densidad de los datos de piel normalizados por Cuantiles**: (A) Tejido de piel FvB normalizado por Cuantiles con su *RT*. (B) Tejido de piel FvB *RB*1 normalizado por Cuantiles con *RB*2. (C) Tejido de piel E6 normalizado por Cuantiles con su *RT*. (D) Tejido de piel E6 *RB*1 normalizado por Cuantiles con *RB*2.

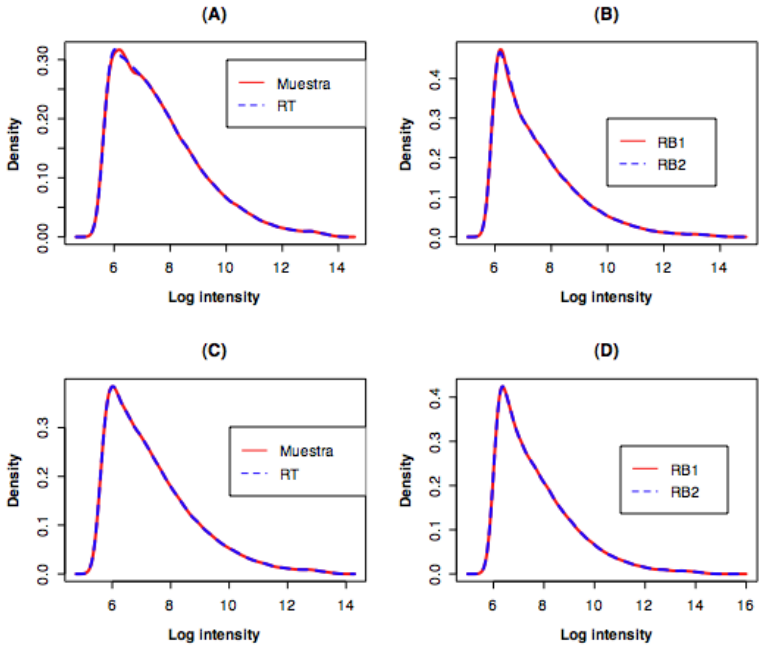

Figura 5.17: **Funciones de densidad de los datos de cérvix normalizados por Cuantiles**: (A) Tejido de cérvix FvB normalizado por Cuantiles con su *RT*. (B) Tejido de cérvix FvB *RB*1 normalizado por Cuantiles con *RB*2. (C) Tejido de cérvix E6 normalizado por Cuantiles con su *RT*. (D) Tejido de cérvix E6 *RB*1 normalizado por Cuantiles con *RB*2.

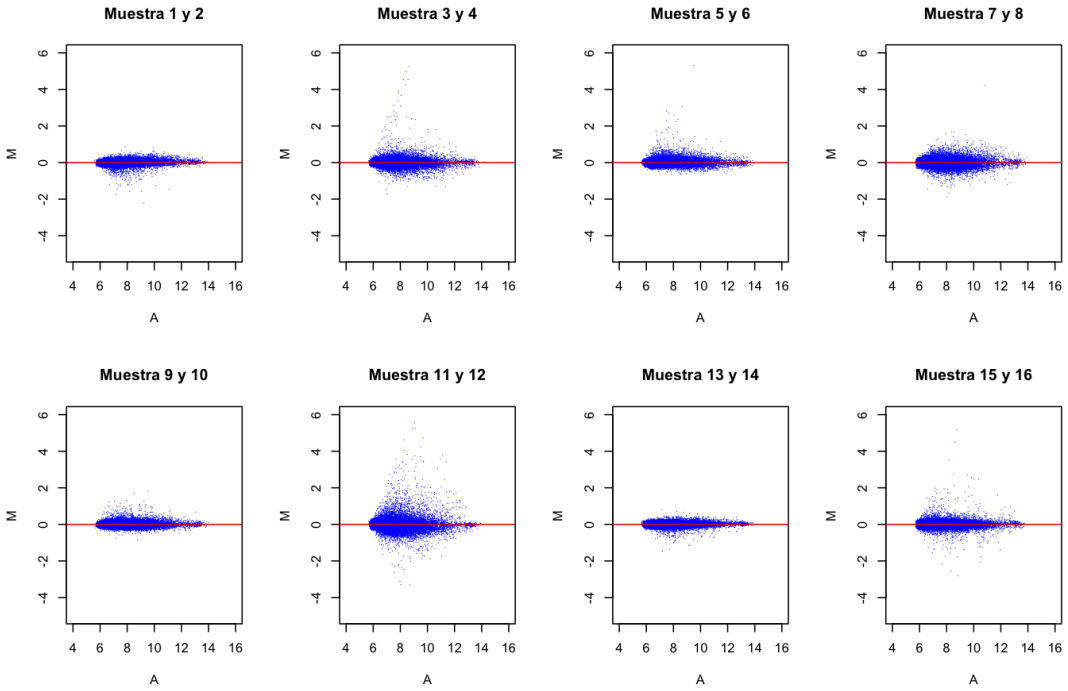

Figura 5.18: **Gráfico MA de los datos normalizados por el método de Cuantiles**. *M* es la diferencia de expresión en valores logarítmicos y *A* es el promedio de dichos valores. Las comparaciones presentadas en los encabezados de cada gráfico corresponden a los tratamientos y replicados de acuerdo a las siguientes anotaciones: Muestra 1=Original Piel FVB, muestra 2= Replicado técnico Piel FVB, muestra 3= Replicado biológico 1 Piel FVB, muestra 4= Replicado biológico 2 Piel FVB, muestra 5= Original Piel E6, muestra 6= Replicado técnico Piel E6, muestra 7= Replicado biológico 1 Piel E6, muestra 8= Replicado biológico 2 Piel E6, muestra 9= Original Cérvix FVB, muestra 10= Replicado técnico Cérvix FVB, muestra 11= Replicado biológico 1 Cérvix FVB, muestra 12= Replicado biológico 2 Cérvix FVB, muestra 13= Original Cérvix E6, muestra 14= Replicado técnico Cérvix E6, muestra 15= Replicado biológico 1 Cérvix E6, muestra 16= Replicado biológico 2 Cérvix E6.

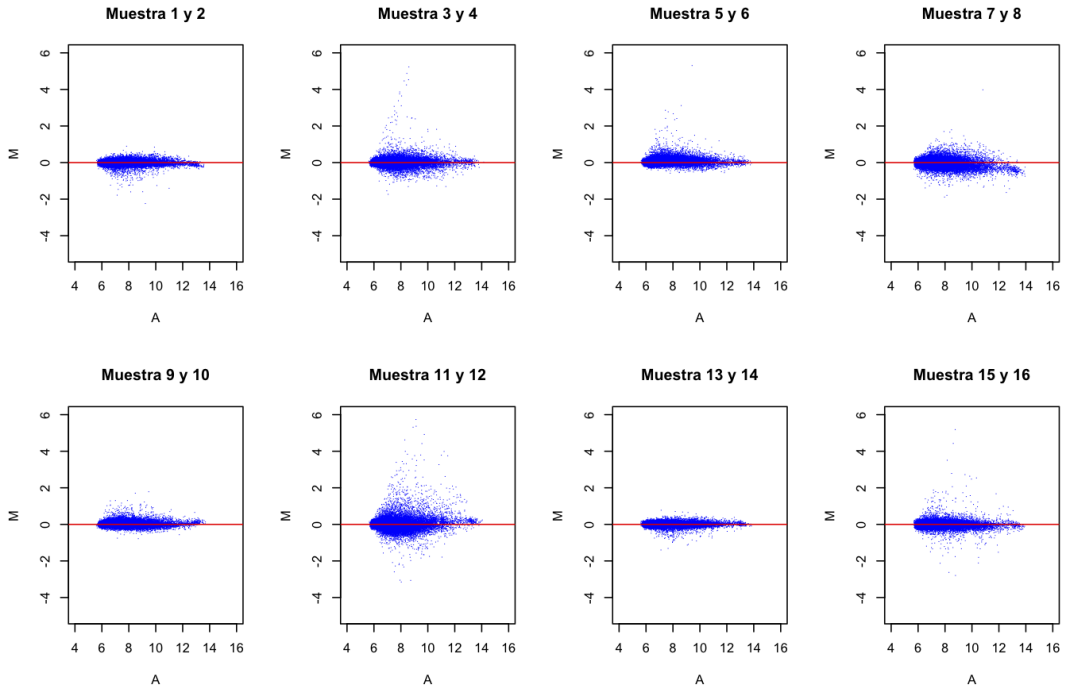

Figura 5.19: **Gráfico MA de los datos normalizados por el método de Loess cíclico**.

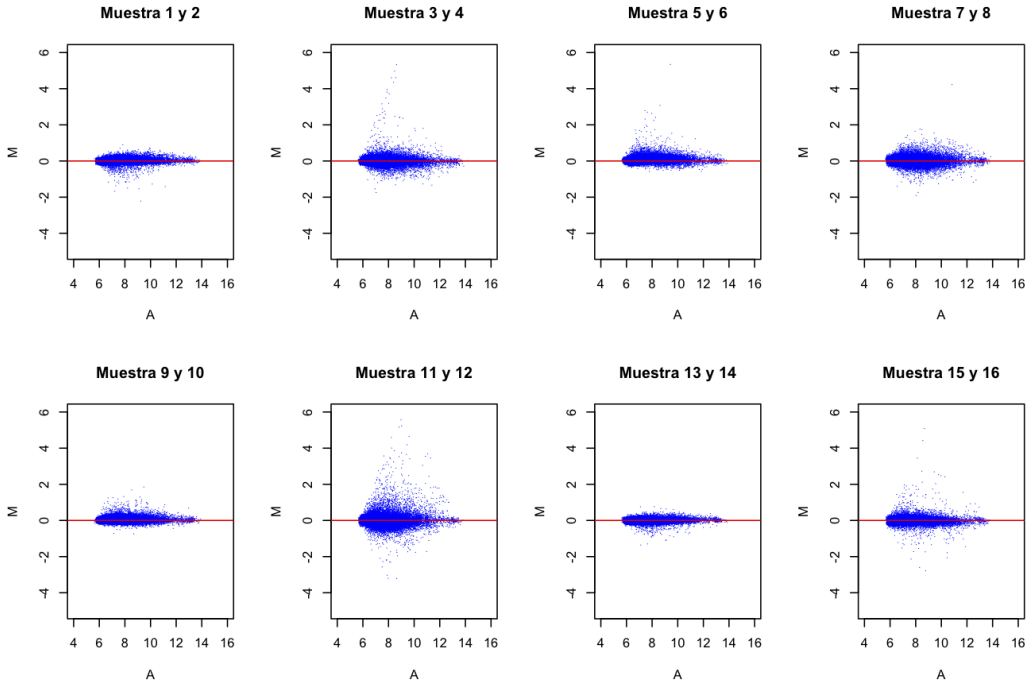

Figura 5.20: **Gráfico MA de los datos normalizados por el método de Contraste**.

Normalización por métodos de información parcial

Los métodos de información parcial que se seleccionaron para el análisis, son los presentados en la figura 5.21. De acuerdo a estos gráficos y los de la figura 5.23, el algoritmo de normalización más débil en el sentido de empatar la distribuciones, resultó ser constante. Al comparar estos últimos gráficos de cajas con los de la figura 5.24, se observa que el método de Qspline y Conjuntos invariantes parece funcionar bien, de manera que las distribuciones de ambos no difieren mucho incluso en relación a las obtenidas con los métodos completos. Los gráficos MA resultantes (figura 5.25, 5.26 y 5.27), muestran ligeramente una mayor dispersión con respecto a la línea horizontal $M = 0$ que en los métodos de información total, siendo las más notables aquellas correspondientes a Constante (figura 5.27). Cabe recalcar que en las comparaciones donde se observa una relación no lineal en los datos crudos, el método de Qspline funciona de forma eficaz (figura 5.26).

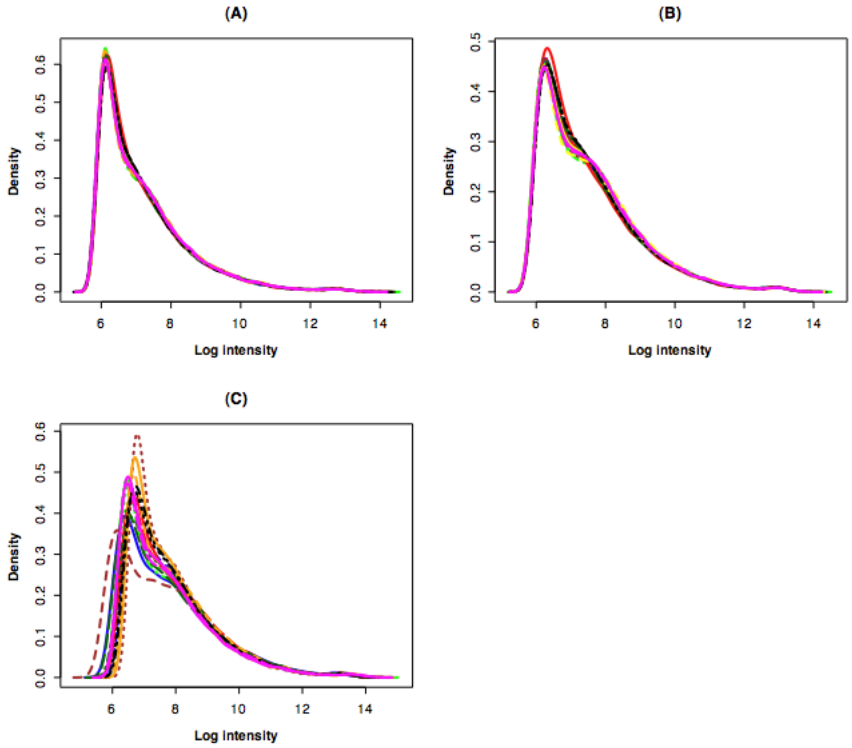

Figura 5.21: **Función de densidad de los datos normalizados por métodos de información parcial**: (A) Normalización por el método de Qspline. (B) Normalización por el método de Conjuntos invariantes. (C) Normalización por el método Constante.

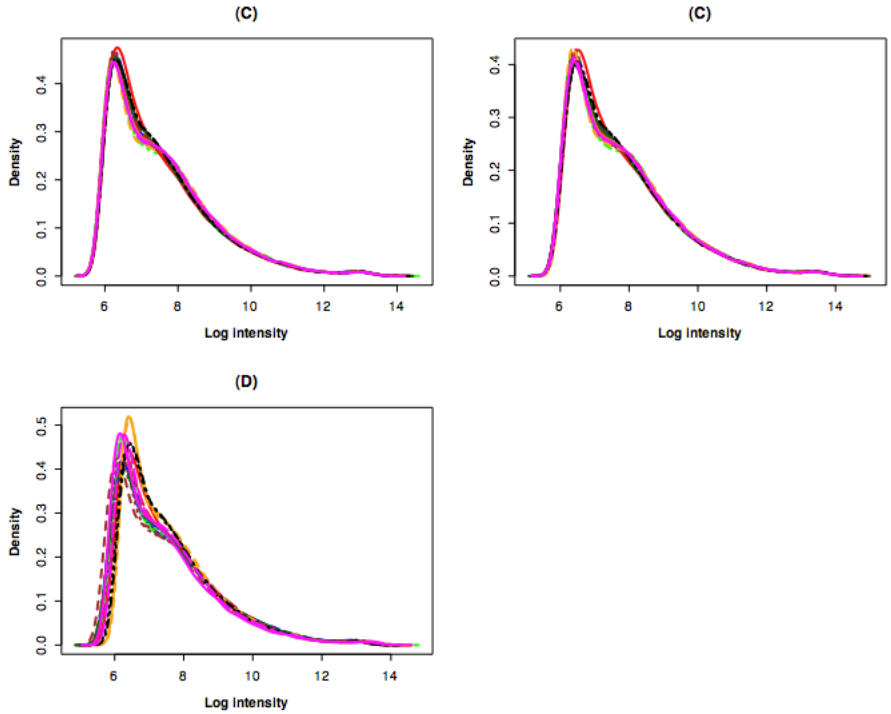

Figura 5.22: **Función de densidad de los datos normalizados por métodos de información parcial después de normalizar por Cuantiles entre replicados de tejido**: (A) Normalización por el método de Qspline. (B) Normalización por el método de Conjuntos invariantes. (C) Normalización por el método Constante.

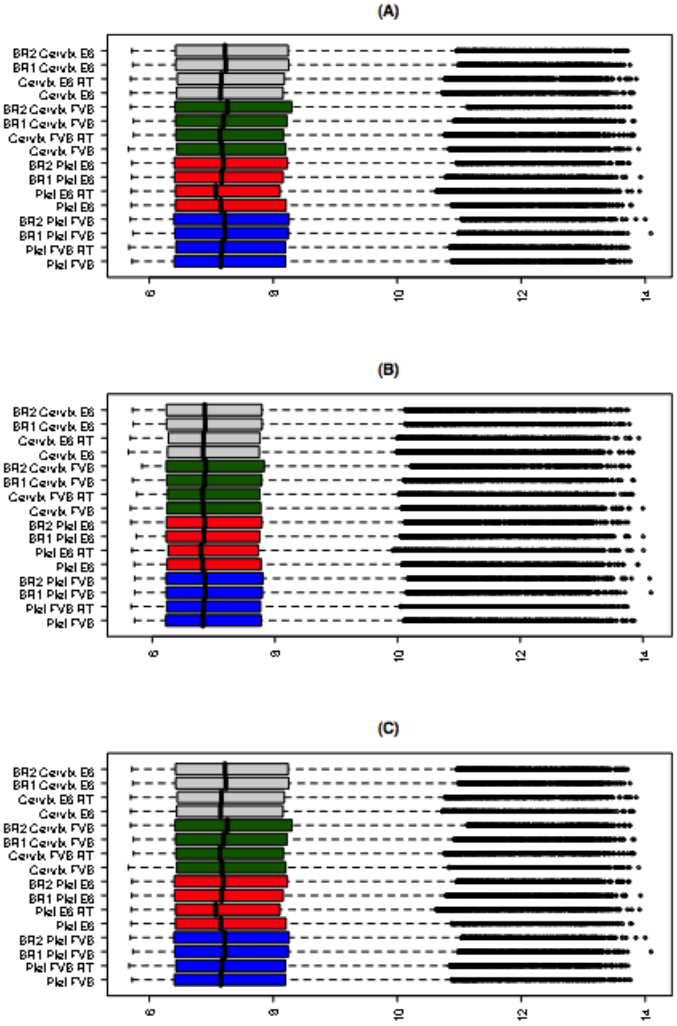

Figura 5.23: **Gráficos de cajas de los datos normalizados por los métodos de información parcial**: (A) Normalización por el método de Qspline. (B) Normalización por el método de Conjuntos invariantes. (C) Normalización por el método Constante.

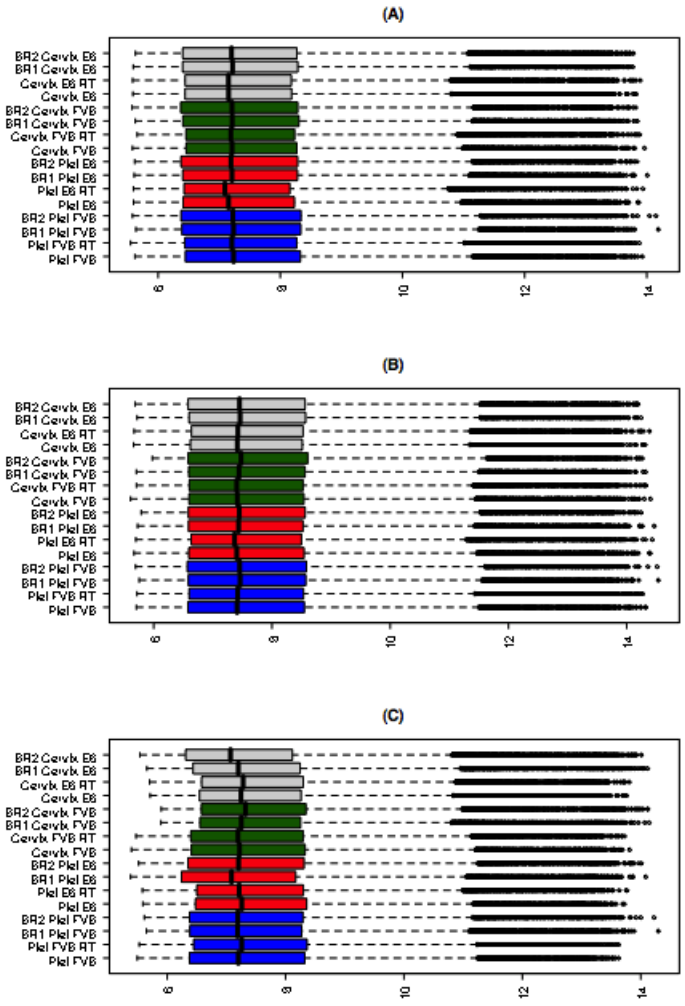

Figura 5.24: **Gráficos de cajas de los datos normalizados por los métodos de información parcial después de normalizar por Cuantiles entre replicados de tejido**: (A) Normalización por el método de Qspline. (B) Normalización por el método de Conjuntos invariantes. (C) Normalización por el método Constante.

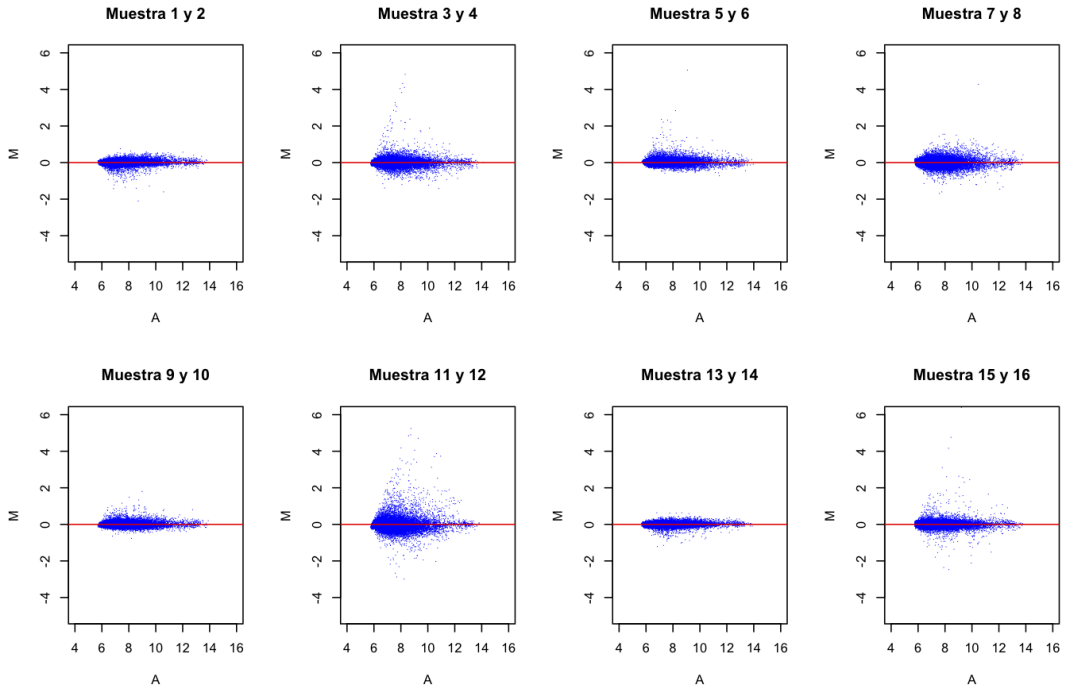

Figura 5.25: **Gráfico MA de los datos normalizados por el método de Conjuntos invariantes**. *M* es la diferencia de expresión en valores logarítmicos y *A* es el promedio de dichos valores. Las comparaciones presentadas en los encabezados de cada gráfico corresponden a los tratamientos y replicados de acuerdo a las siguientes anotaciones: Muestra 1=Original Piel FVB, muestra 2= Replicado técnico Piel FVB, muestra 3= Replicado biológico 1 Piel FVB, muestra 4= Replicado biológico 2 Piel FVB, muestra 5= Original Piel E6, muestra 6= Replicado técnico Piel E6, muestra 7= Replicado biológico 1 Piel E6, muestra 8= Replicado biológico 2 Piel E6, muestra 9= Original Cérvix FVB, muestra 10= Replicado técnico Cérvix FVB, muestra 11= Replicado biológico 1 Cérvix FVB, muestra 12= Replicado biológico 2 Cérvix FVB, muestra 13= Original Cérvix E6, muestra 14= Replicado técnico Cérvix E6, muestra 15= Replicado biológico 1 Cérvix E6, muestra 16= Replicado biológico 2 Cérvix E6.

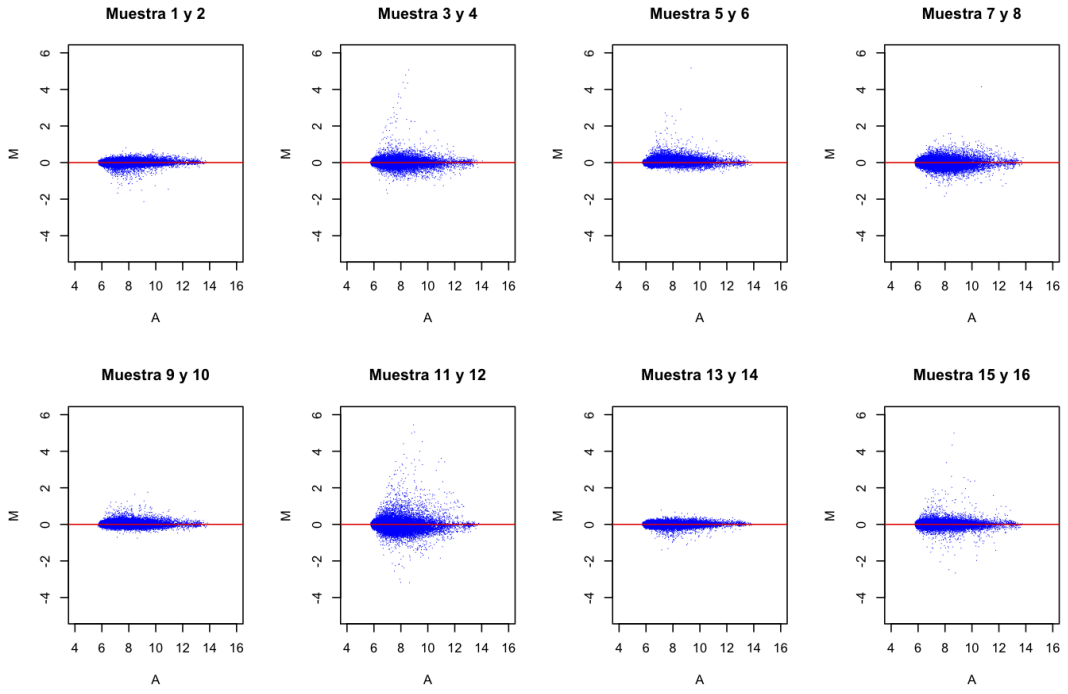

Figura 5.26: **Gráfico MA de los datos normalizados por el método de Qspline**.

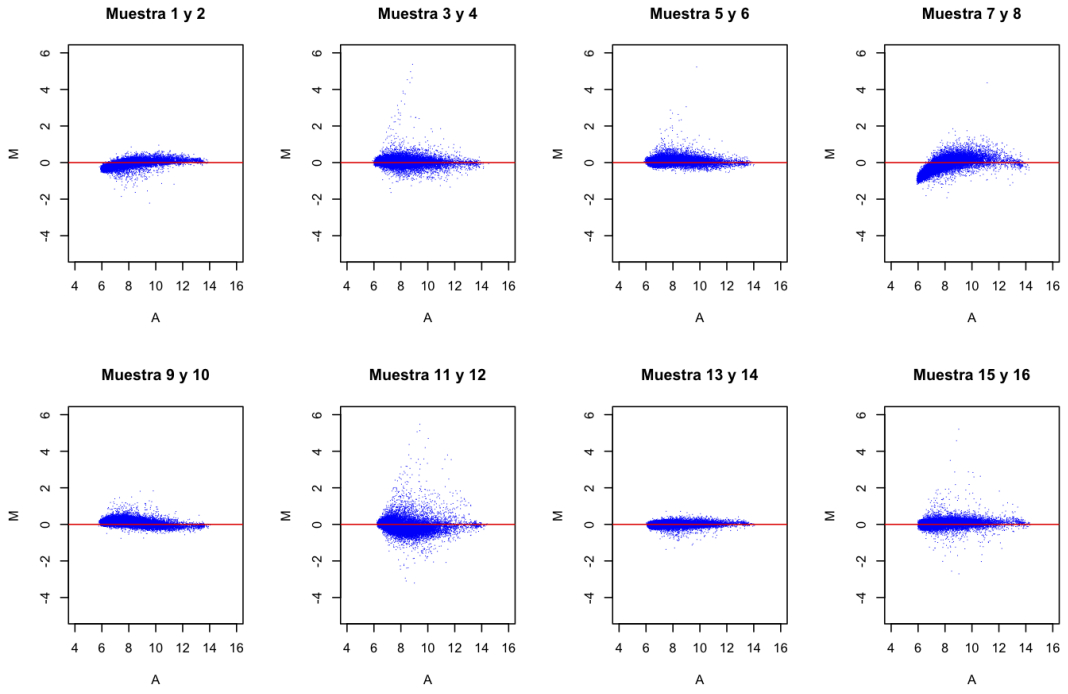

Figura 5.27: **Gráfico MA de los datos normalizados por el método Constante**.

Normalización con corrección de fondo

La figura 5.28 es un acercamiento a una sección del microarreglo que hace evidente la presencia de ruido en la señal y la necesidad de un proceso de normalización. En la imagen del escáner las intensidades de señal de sonda, no están perfectemante acotadas por la red que define las posiciones de dichas sondas provocando interferencia entre señales de diferentes sondas. Además, a pesar de ser replicados técnicos provenientes de un mismo lote la intensidad de señal varía de forma notable tanto en la imagen del escáner como en el archivo $.CEL$ procesado.

La corrección de ruido de fondo, tiene un efecto importante sobre los resultados que se obtienen en las siguientes fases del análisis. Por está razón se realizó una breve comparación

a través de algunos gráficos como los que se han presentado. En la figura 5.29, se observa una importante modificación en las funciones de densidad. Se aprecian algunas diferencias a simple vista pero todas presentan un comportamiento bimodal, lo cual es tema para discutir en relación a la interpretación de los datos.

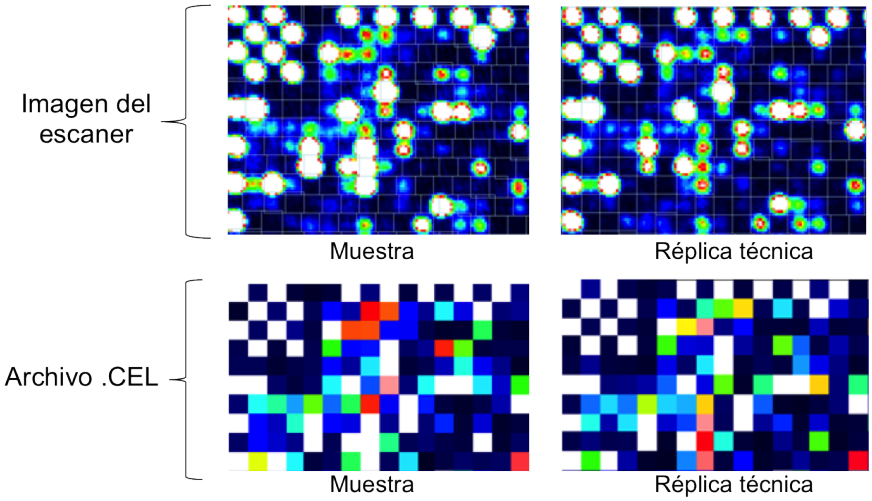

Figura 5.28: **Comparación de una sección de microarreglo correspondendiente a un par de replicados técnicos.** La imagen del escáner, es obtenida por la misma plataforma de procesamiento de microarreglos dentro de la unidad de alta tecnología. El software que incluye la plataforma convierte la imagen del escáner en un archivo de pixeles, donde cada pixel representa un nivel de intensidad de señal de sonda. Los algoritmos de procesamiento de imagenes también proveen diferencias importantes[38].

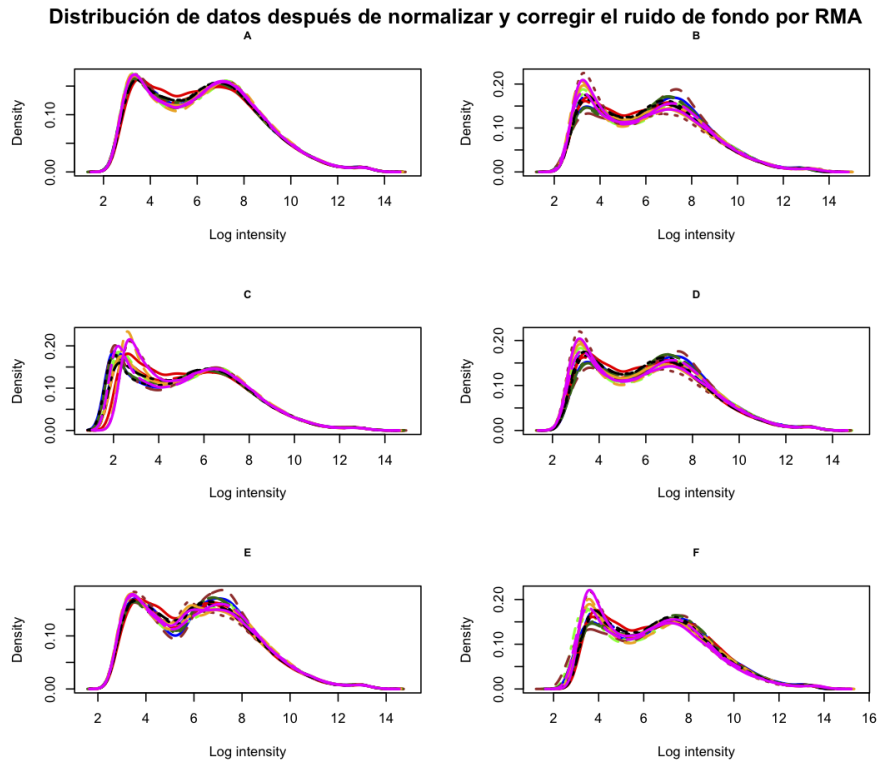

Figura 5.29: **Funciones de densidad de los datos normalizados por los seis métodos bajo comparación y corrección de ruido de fondo**. (A) Método de Cuantiles. (B) Método de Loess cíclico. (C) Método de Conjuntos invariantes. (D) Método de Contraste. (E) Método de Qspline. (F) Método Constante. Se utilizó el algoritmo RMA para la corrección de ruido de fondo.

5.3.2. Clasificación

Modelos lineales para determinar expresión diferencial (LIMMA)

Los métodos que se utilizan para detección de expresión diferencial no obedecen un estandar de oro y el camino a elegir no es trivial[38], en este caso el enfoque es en base a modelos lineales. Limma es una librería de *bioconductor* especialmente diseñada para análisis de expresión diferencial en datos de microarreglos de expresión[39]. Está construi-

da para analizar proyectos, cuya característica es involucrar comparaciones entre muchos transcritos simultáneamente. Se basa en ajustar un modelo lineal a los niveles de expresión de cada gen. Dada la limitante de pocas réplicas para el volumen de variables, se hace uso de métodos que usen información generada durante y entre los genes que se comparan por métodos empíricos de Bayes, lo cual permite que el análisis sea estable aun cuando lidiamos con muy reducido número de microarreglos. Para utilizar limma en el análisis de expresión diferencial, se requiere especificar dos matrices: la matriz de diseño que detalla como esta planteado el experimento, es decir, una matriz de dimensión nxm de unos y ceros, donde n es el número de microarreglos utilizados en el experimento y m la cantidad de grupos propuestos, y la matriz de contrastes que permite que los coeficientes definidos en la matriz de diseño sean combinados en forma de contrastes de interés entre las diferentes fuentes de ARN (por ejemplo casos y controles, entre tratamientos o a través de estadíos en un proyecto de series de tiempos).

Limma resume los resultados del modelo lineal, realiza las pruebas de hipótesis de expresión diferencial y el ajuste del valor de p por múltiples pruebas a través de una variedad de estadísticos. Entre ellos encontramos fold changes, errores estándar, pueba t y valor de p. El estadístico básico usado para análisis de significancia es la prueba t moderada, misma que es calculada para cada sonda y para cada contraste. Su interpretación es tal cual la prueba t, la única diferencia es que los errores estándar han sido moderados a lo largo de los genes generando un valor común a través de un simple modelo Bayesiano que no hace otra cosa que *"pedir prestada"* información a grupos de genes para generar inferencia sobre cada gen en forma individual[40].

Sin embargo, limma despliega tres estadísticos para determinar significancia: valor de p, valor de p ajustado y estadístico B. El valor de p depende de algunas consideraciones sobre normalidad e independencia entre variables que no necesariamente se cumplen en el caso de microarreglos de expresión. El valor de p es ajustado utilizando *Benjamini and Hochberg* FDR (tasa de detección de falsos positivos)[41] que básicamente dice: si todos los genes con un valor de p caen por debajo de un umbral dado (por ejemplo 0.05), éstos son seleccionados como diferencialmente expresados y entonces la proporción de

detección de falsos positivos está controlada a ser menor que ese umbral predeterminado (en nuestro caso $5\,\%$). Ahora bien, otra alternativa es el estadístico B (*lods* o *B-stat*) el cual es ajustado de forma automática por múltiples pruebas asumiendo por *default* o *a priori* que el $1\,\%$ de los genes (o bien otro porcentaje que definamos) estarán diferencialmente expresados. Es común ver que en términos de ranqueo tanto los valores p como el estadístico B ordenarán los genes en forma similar. Las probabilidades calculadas en el estadístico B requieren de un valor inicial en forma de *prior* ($1\,\%$ en este caso), que indica que al menos el $1\,\%$ de los genes estarán diferencialmente expresados. En este sentido algunas personas se sienten más cómodas usando valores p como tal o bien ajustados y no depender de la precisión asumida en el *prior*.

Comparación de los efectos que diferentes algoritmos de pre-procesamiento tienen sobre la identificación de genes diferencialmente expresados

Una vez normalizados los datos ya es posible realizar las comparaciones de interés entre los microarreglos. Así mismo, un paso fundamental para comparar la eficiencia de los diferentes algoritmos de pre-procesamiento consiste en cuantificar los falsos positivos y negativos que se obtienen bajo un mismo umbral de clasificación. Sin embargo dado el número de hipótesis (genes) que se están probando, en la práctica no es posible cuantificar por completo los falsos negativos y es complicado, costoso y tardado identificar todos los falsos postivos pues cada gen requiere un experimento de validación (además de las variantes en parámetros dentro de cada algoritmo de normalización, los modelos que se usen para determinar expresión diferencial y los umbrales que se determinen para la prueba estadística que se tome en cuenta).

A razón de la complejidad que representa el proceso de validación de resultados, en este trabajo la exploración de los efectos que los diferentes procesos de normalización tienen a nivel de la fase de clasificación, parte de la observación de los gráficos de volcán correspondientes a los contrastes requeridos en este análisis. Esto con el fin de obtener una aproximación visual sobre las diferencias cuantitativas y cualitativas que dichos algoritmos

confieren a la identificación de genes diferencialmente expresados y con ello complementar argumentos que permitan reducir el número de comparaciones consecuentes a través de diagramas de Venn, ya que se espera que para algunos de los algoritmos de pre-procesamiento aplicados la gran mayoría de los genes detectados coincida.

Contrastes con los algoritmos de normalización de información total

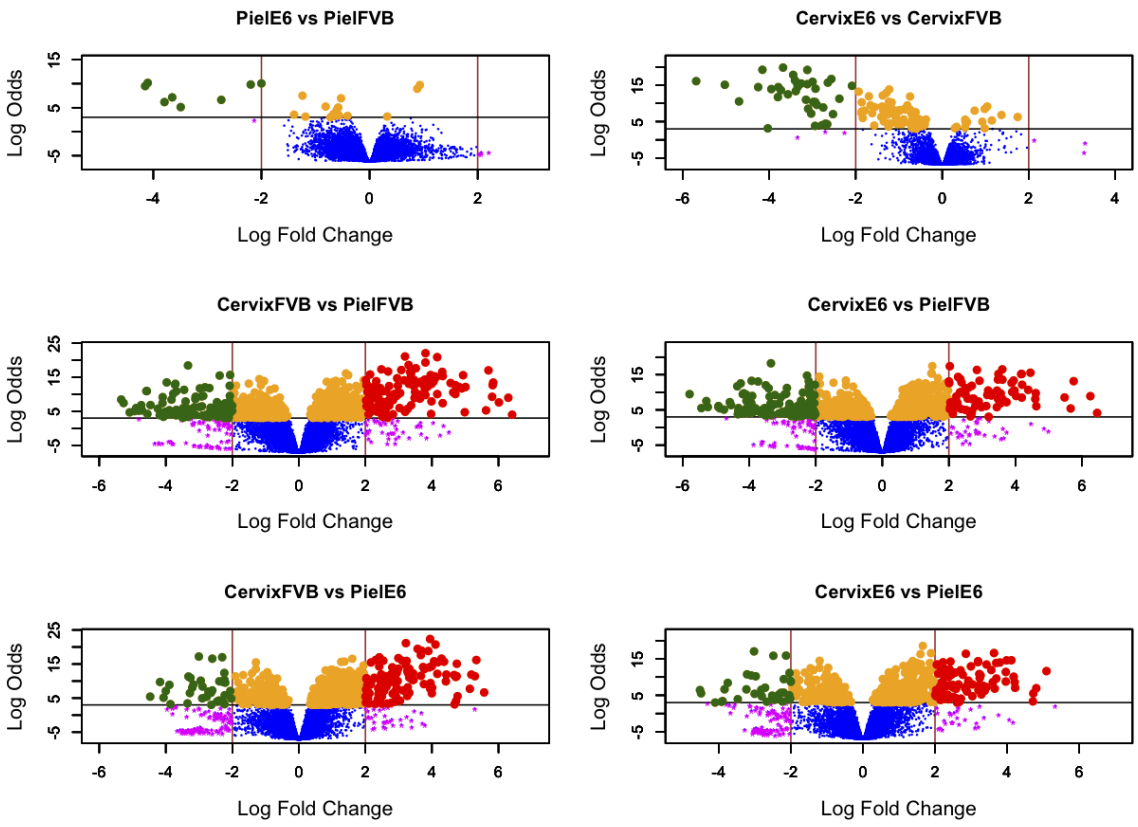

Figura 5.30: **Gráficos de volcán de los datos normalizados por el método de Cuantiles**. Cada gráfico representa el correspondiente contraste mostrado en los encabezados. En el eje de las abscisas se tiene el valor del *LogFold Change* (M) y en el de las ordenadas el estadístico B. Los puntos de color rojo corresponden a aquellos genes diferencialmente sobreexpresados en dado tratamiento con respecto a su contraste, los genes diferencialmente subexpresados corresponden a los puntos verdes, los magenta a los que sobrepasan el umbral de $M = |2|$ pero no al de $B = 3$, los amarillos a los que sobrepasan el umbral de $B = 3$ pero no al de $M = |2|$ y los azules a los que no sobrepasan ambos umbrales. Por ejemplo, en el primer gráfico los puntos verdes corresponden a los genes diferencialmente subexpresados en Piel E6 con respecto a Piel FVB.

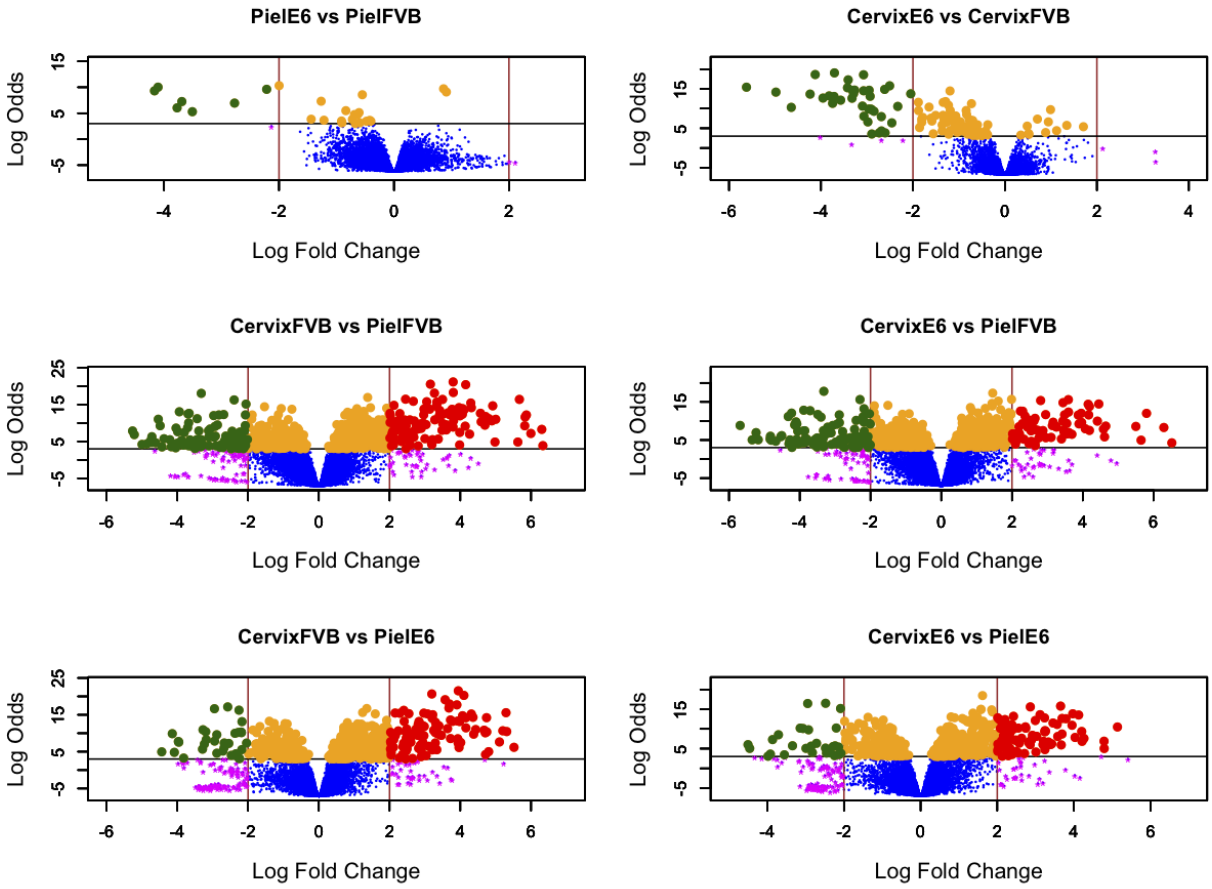

Figura 5.31: **Gráficos de volcán de los datos normalizados por el método de Loess cíclico**.

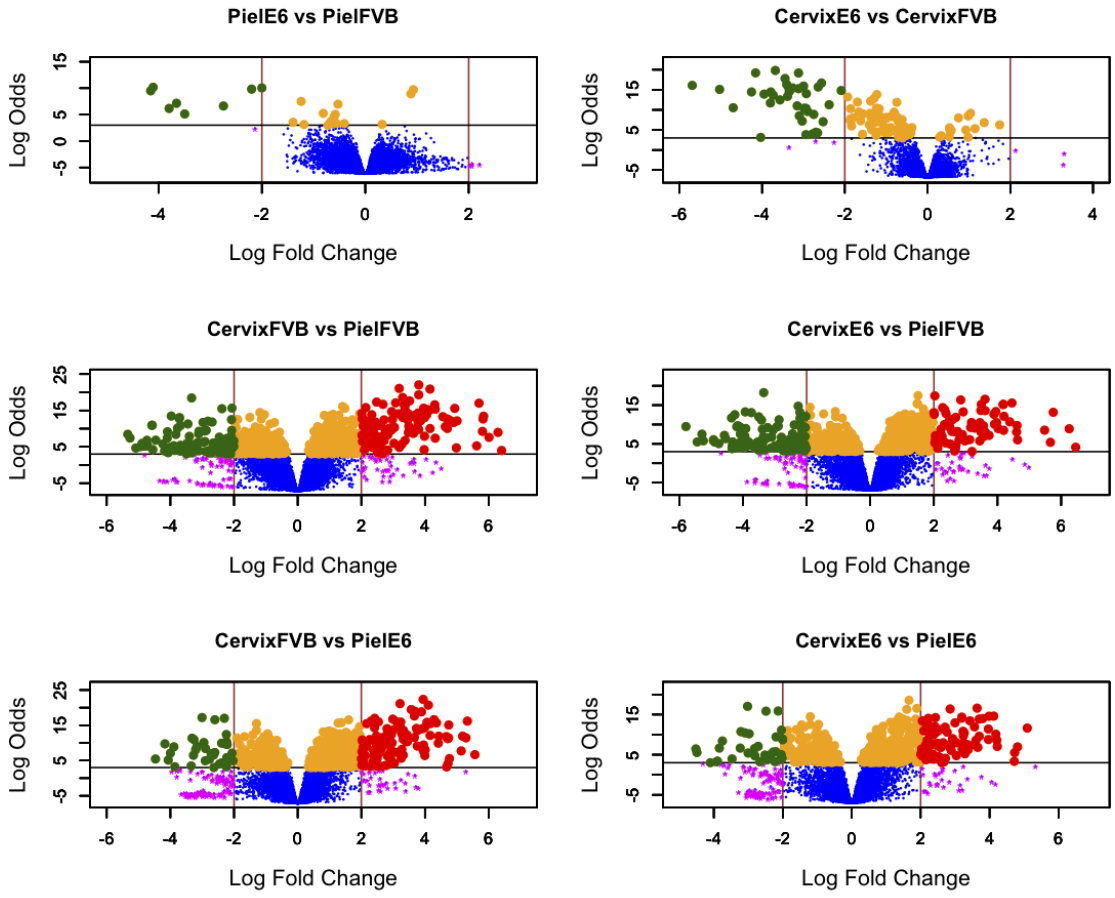

Figura 5.32: **Gráficos de volcán de los datos normalizados por el método de Loess cíclico después de haber normalizado entre grupos por Cuantiles**.

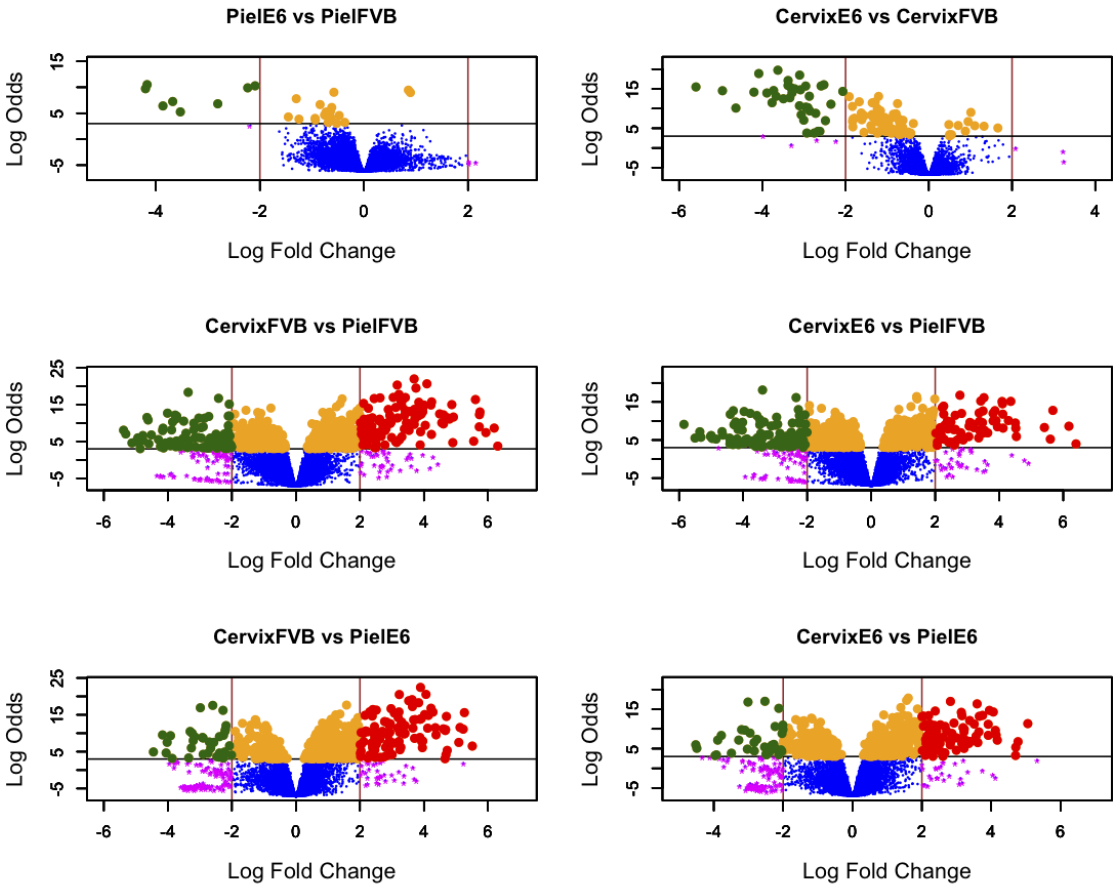

Figura 5.33: **Gráficos de volcán de los datos normalizados por el método de Contraste**.

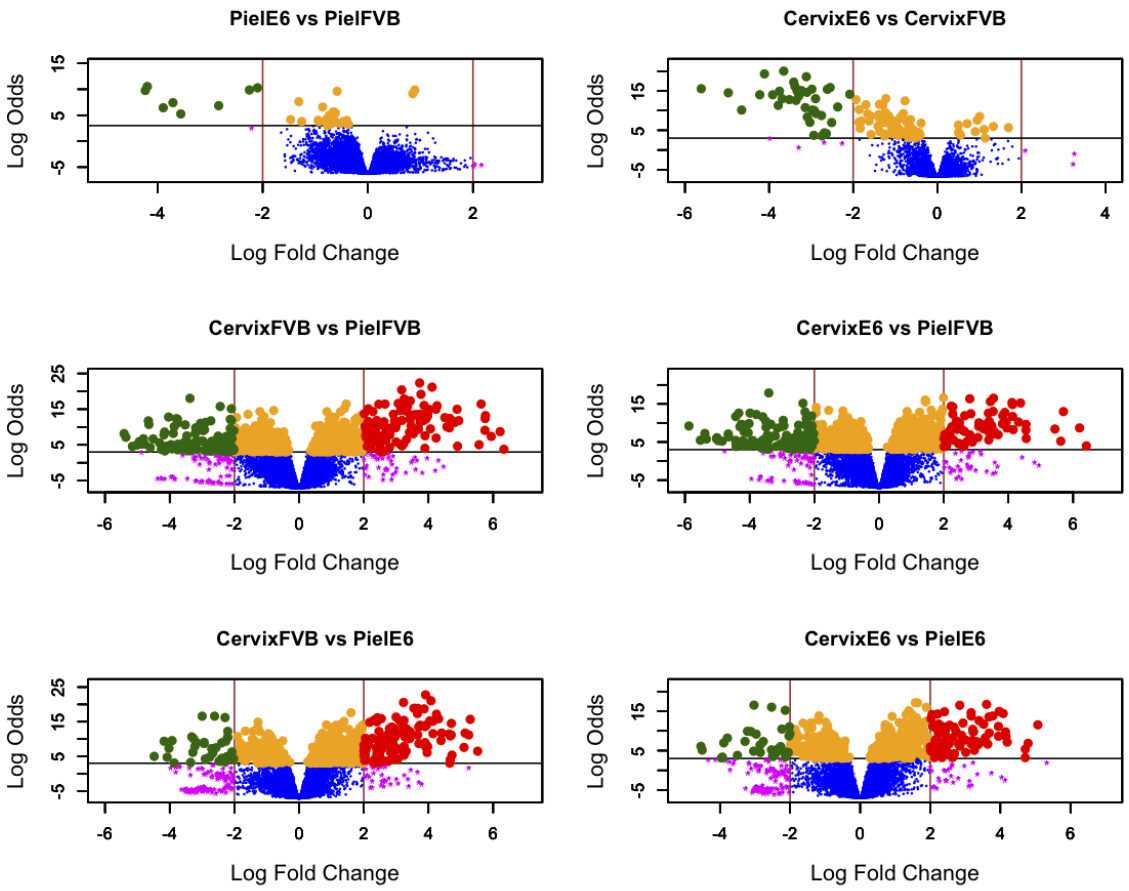

Figura 5.34: **Gráficos de volcán de los datos normalizados por el método de Contraste después de haber normalizado entre grupos por Cuantiles**.

Contrastes con los algoritmos de normalización de información parcial

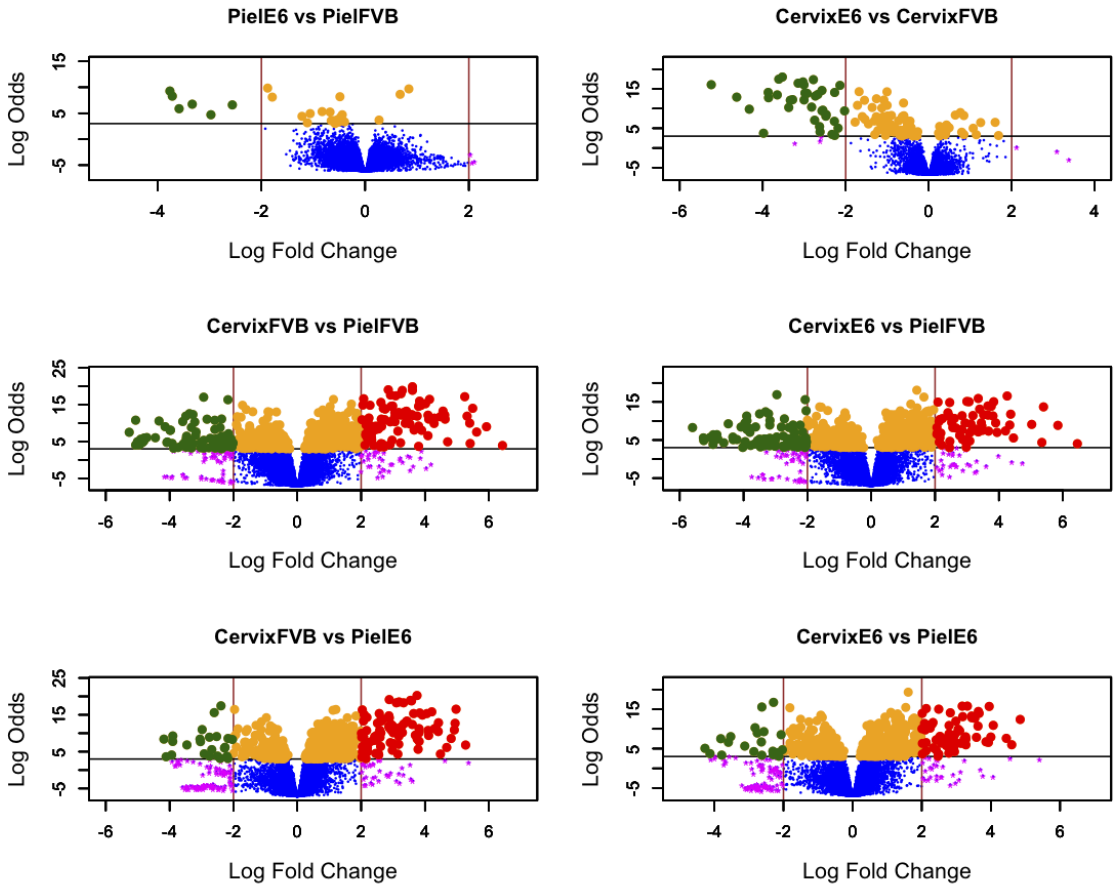

Figura 5.35: **Gráficos de volcán de los datos normalizados por el método de Conjuntos invariantes**.

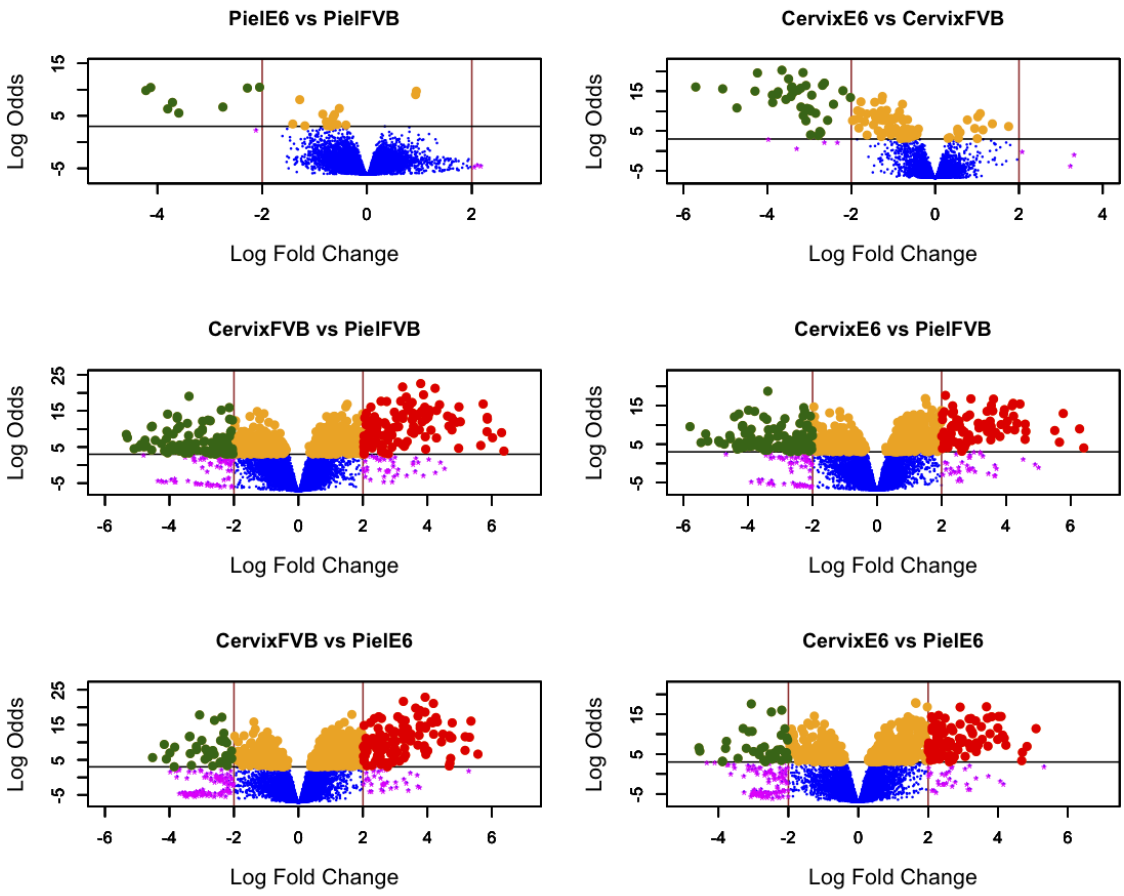

Figura 5.36: **Gráficos de volcán de los datos normalizados por el método de Conjuntos invariantes después de haber normalizado entre grupos por Cuantiles**.

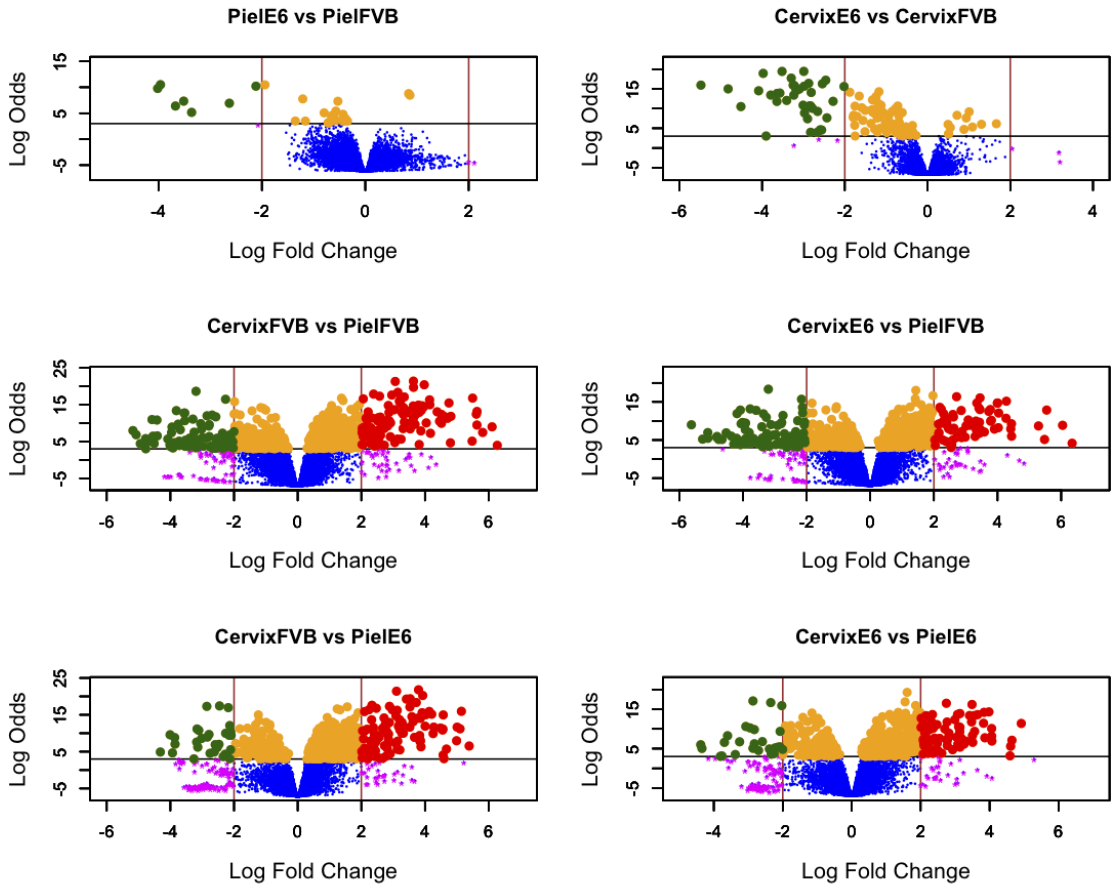

Figura 5.37: **Gráficos de volcán de los datos normalizados por el método de Qspline**.

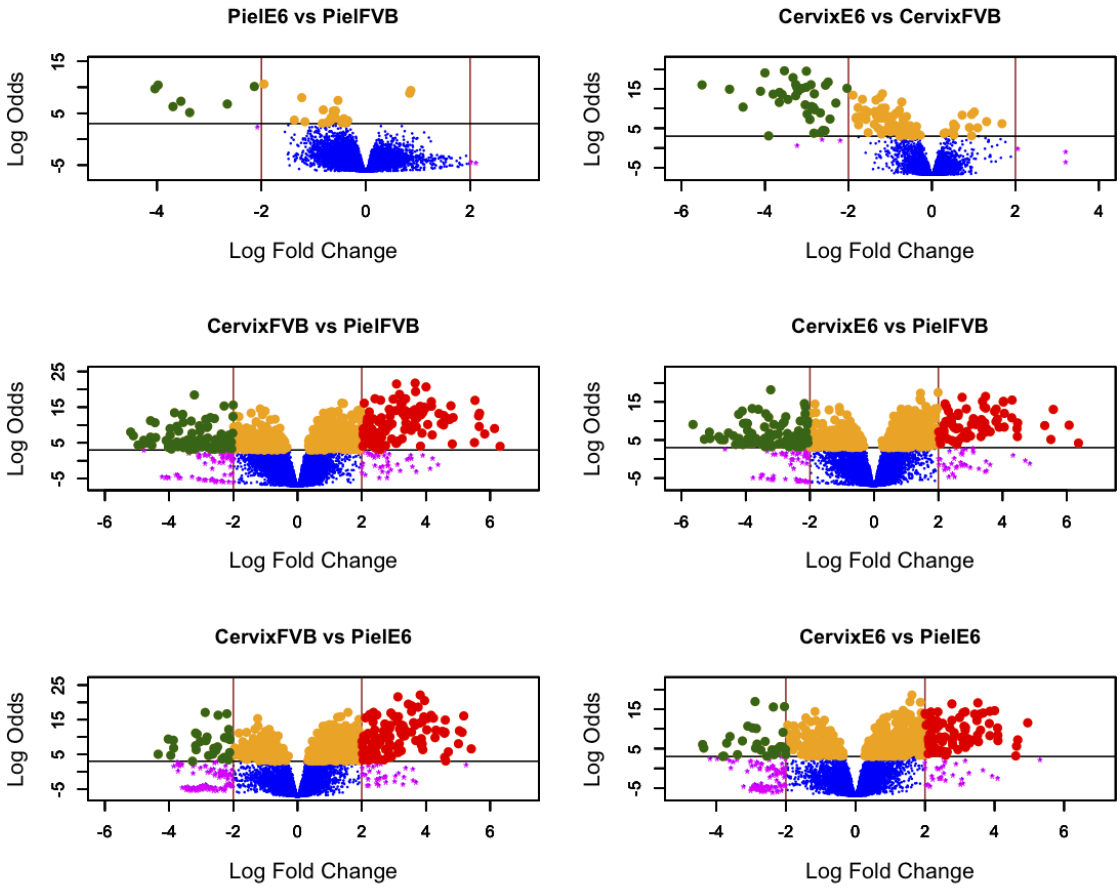

Figura 5.38: **Gráficos de volcán de los datos normalizados por el método de Qspline después de haber normalizado entre grupos por Cuantiles**.

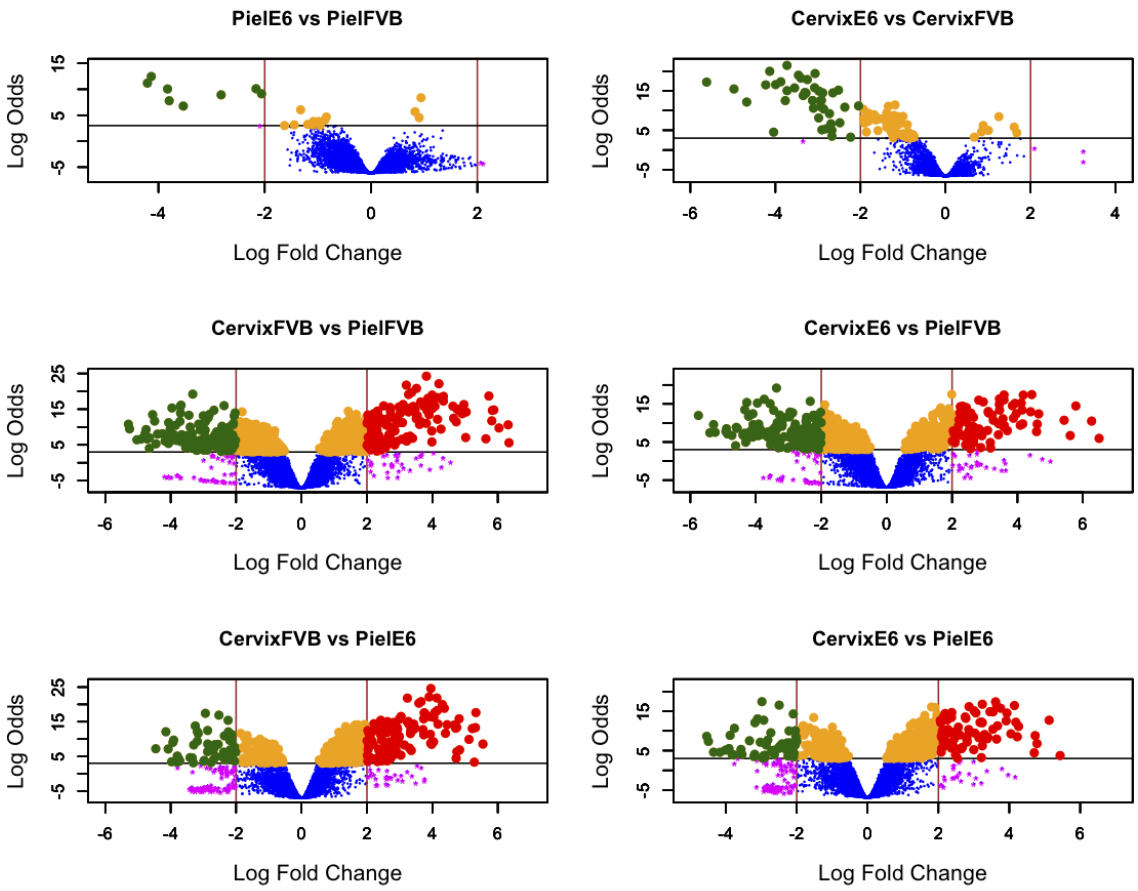

Figura 5.39: **Gráficos de volcán de los datos normalizados por el método lineal**.

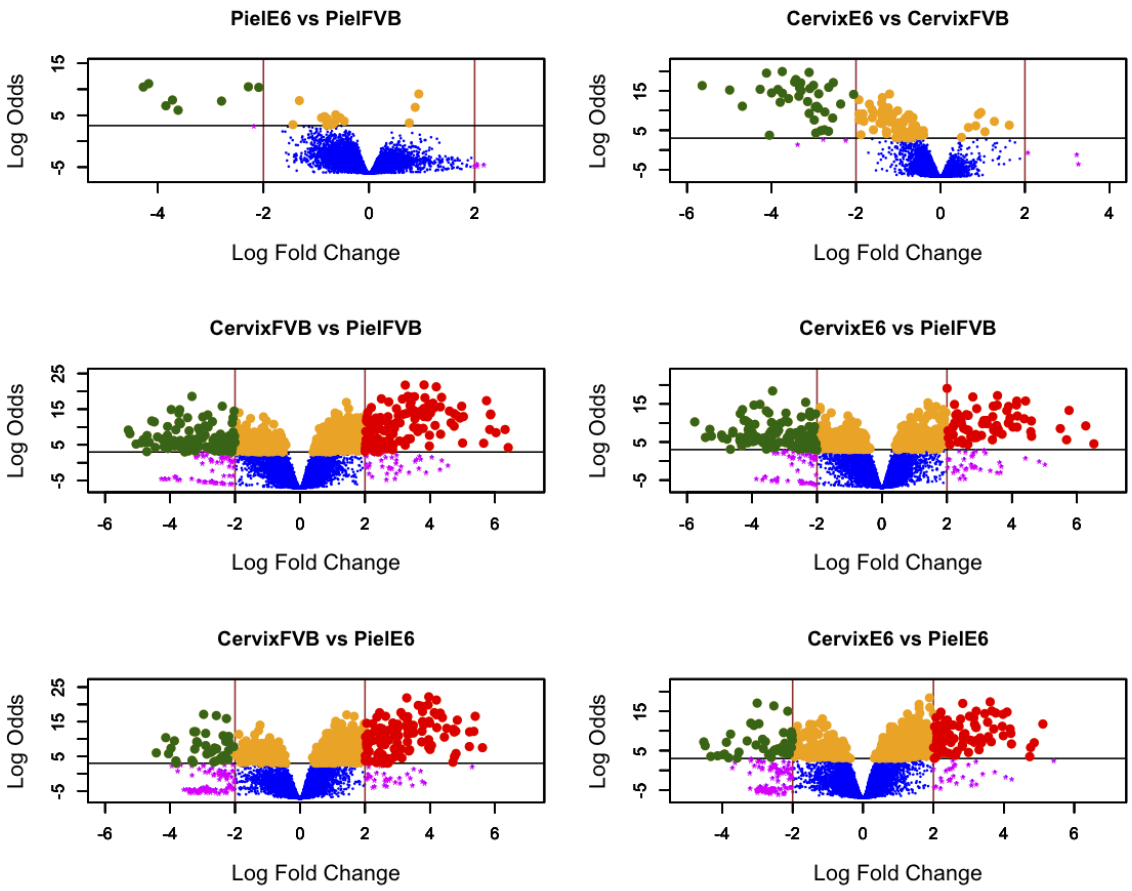

Figura 5.40: **Gráficos de volcán de los datos normalizados por el método lineal después de haber normalizado entre grupos por Cuantiles**.

Contrastes con los seis algoritmos de normalización comparados y corrección de ruido de fondo

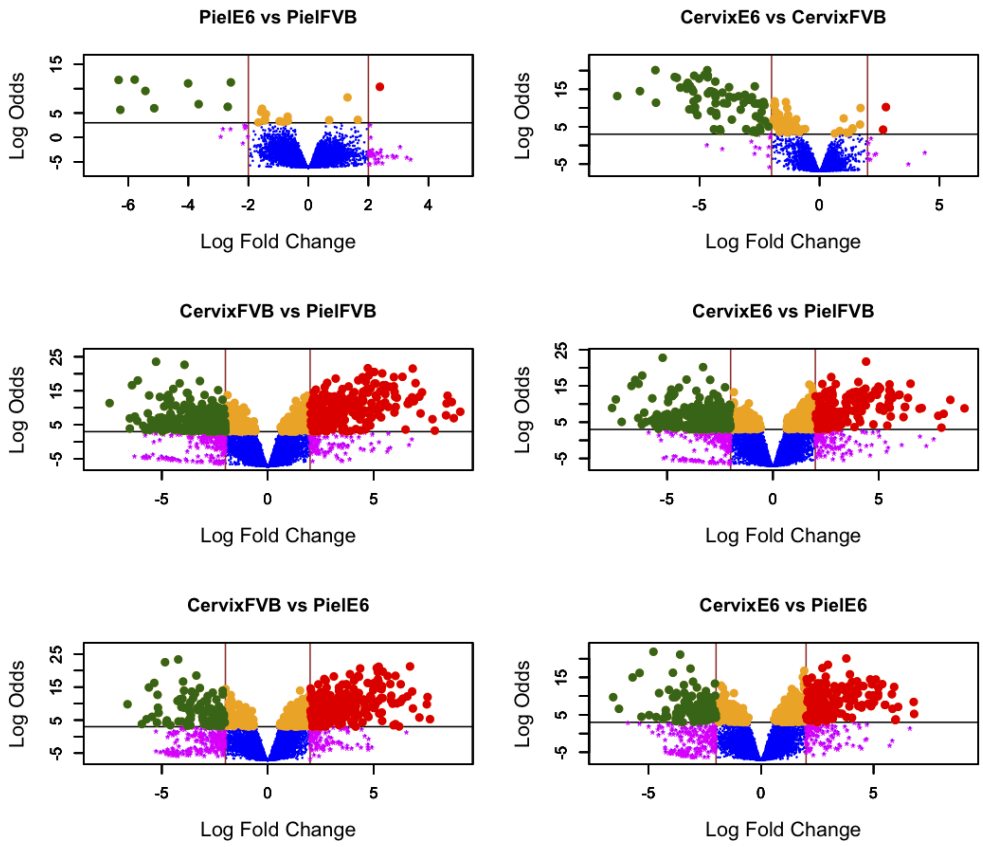

Figura 5.41: **Gráficos de volcán de los datos normalizados por el método de Cuantiles con corrección de ruido fondo (RMA).**

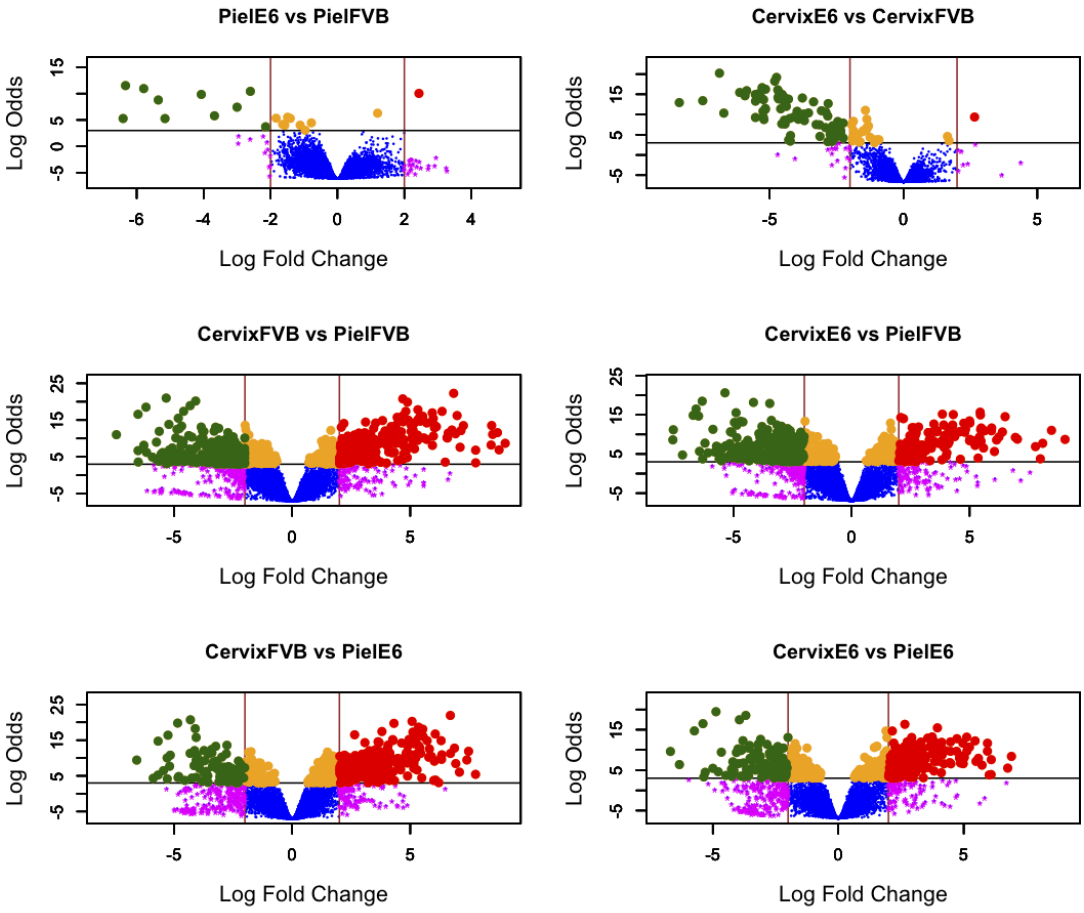

Figura 5.42: **Gráficos de volcán de los datos normalizados por el método de Loess cíclico con corrección de ruido fondo (RMA).**

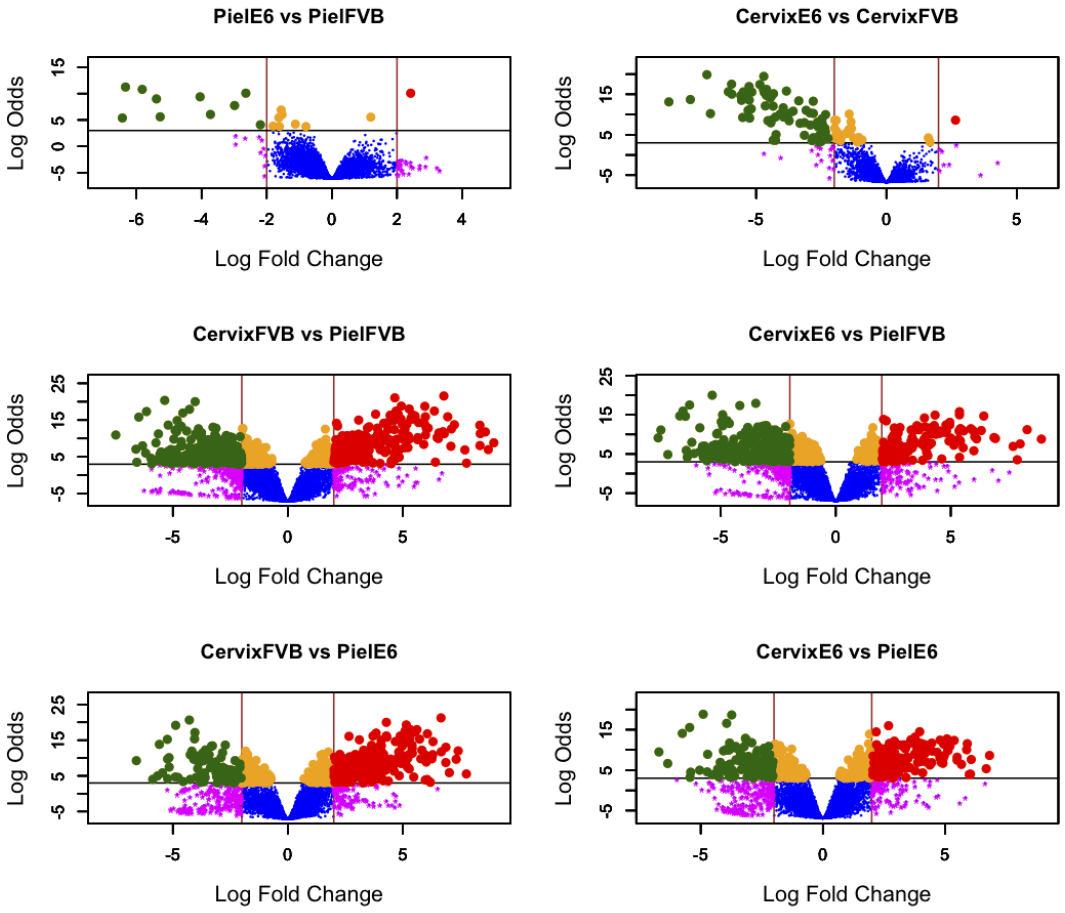

Figura 5.43: **Gráficos de volcán de los datos normalizados por el método de Contraste con corrección de ruido fondo (RMA).**

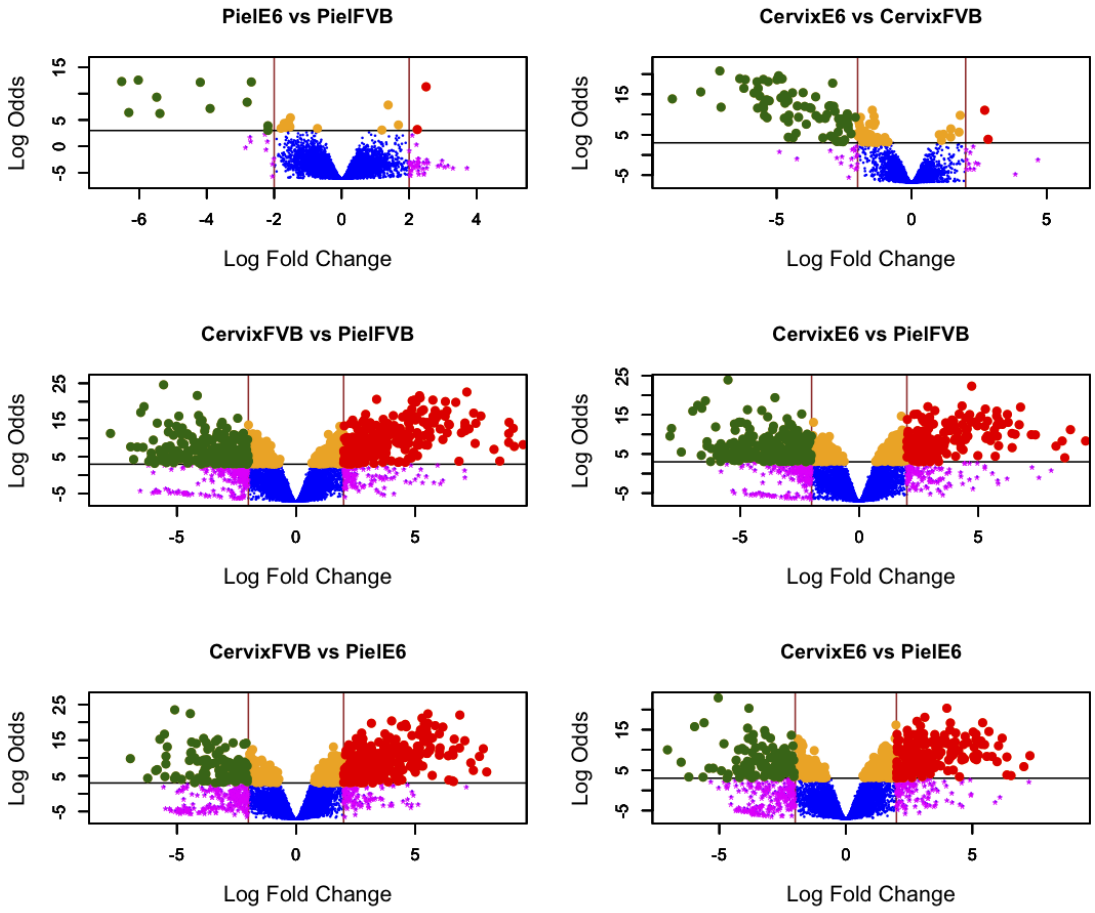

Figura 5.44: **Gráficos de volcán de los datos normalizados por el método de Conjuntos invariantes con corrección de ruido fondo (RMA).**

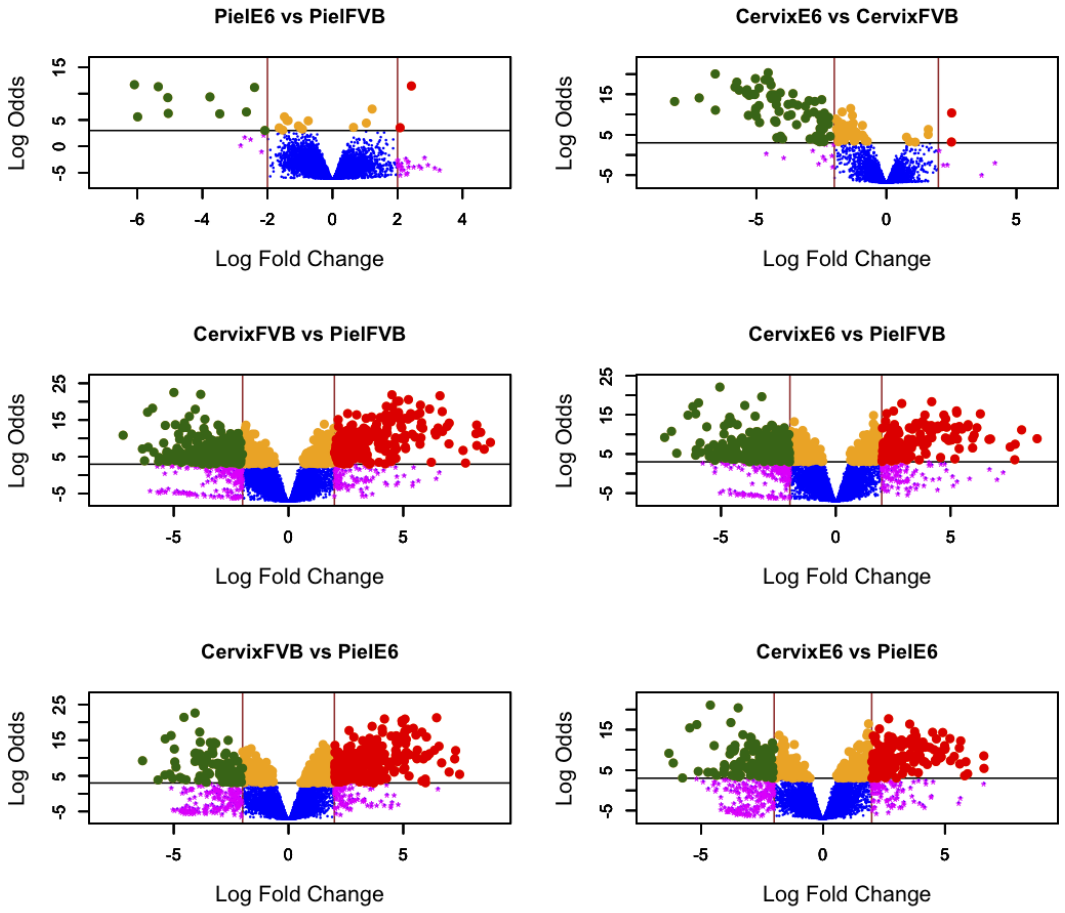

Figura 5.45: **Gráficos de volcán de los datos normalizados por el método de Qsplines con corrección de ruido fondo (RMA).**

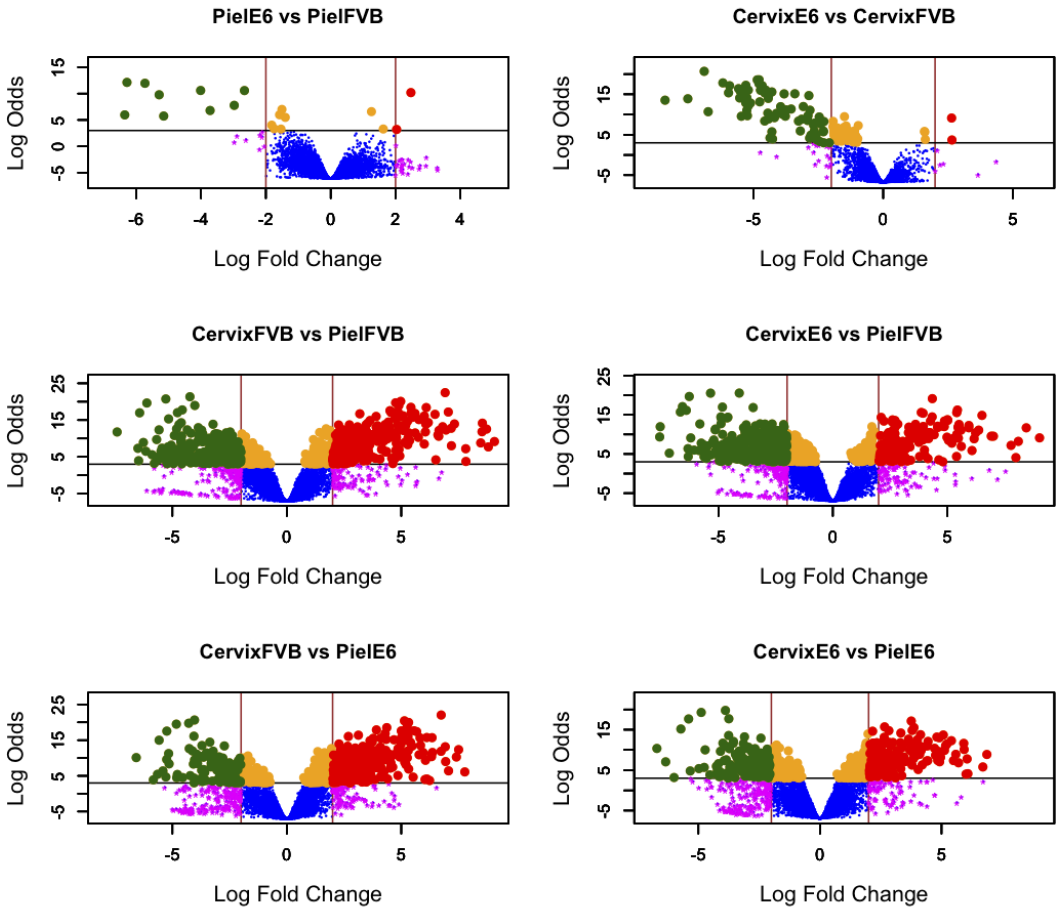

Figura 5.46: **Gráficos de volcán de los datos normalizados por el método Constante con corrección de ruido fondo (RMA).**

Diagramas de Venn

Para hacer comparaciones a tráves de diagramas de Venn se utilizaron los resultados del cuarto contraste (CérvixE6 vs PielFVB) correspondiente a los gráficos de volcán de las figuras 5.30, 5.31, 5.33, 5.35, 5.37, 5.39, 5.41, 5.42, 5.43, 5.44, 5.45 y 5.46.

No se tomarón en cuenta para los diagramas de Venn, aquelllos contrastes en los que se aplicó un algoritmo de pre-procesamiento que incluyerá un proceso de normalización por Cuantiles sobre grupos y otro para el conjunto completo de datos. Esto debido a que la normalización por Cuantiles resultó dominante sin importar el método aplicado con el que se combinara (figuras 5.32, 5.34, 5.36, 5.38 y 5.40).

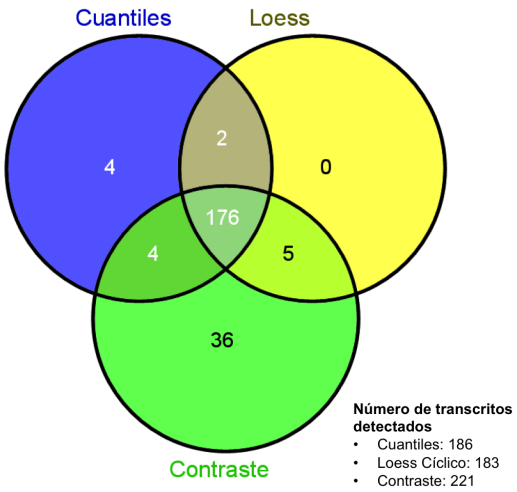

Figura 5.47: **Comparación de algoritmos de información total.** La intersección entre los tres algoritmos conforma el 94,6 % de los genes detectados por Cuantiles, el 96,17 % de los detectados por Loess cíclico y el 79,6 % de los detectados por Contraste.

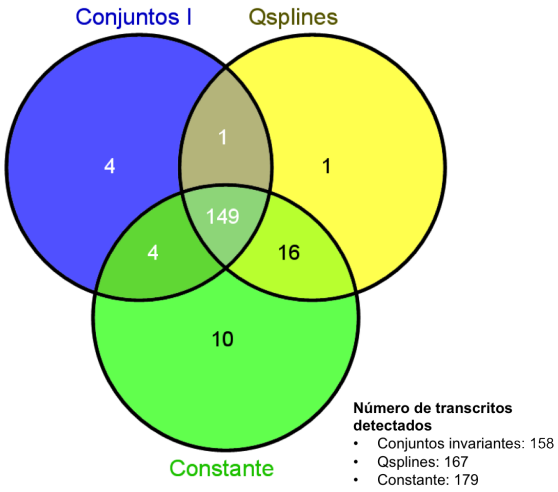

Figura 5.48: **Comparación de algoritmos de informacińparcial.** La intersección entre los tres algoritmos conforma el 94,30 % de los genes detectados por Conjuntos invariantes el 89,22 % de los detectados por Qsplines y el 83,32 % de los detectados por Constante. CI = Conjuntos invariantes.

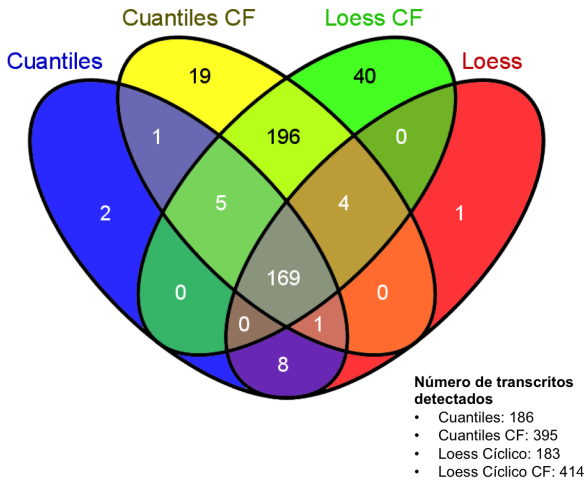

Figura 5.49: **Comparación de algoritmos de información total, efecto de la corrección de ruido de fondo.** La intersección entre los cuatro algoritmos conforma el 90,86 % de los genes detectados por Cuantiles, el 42,78 % de los detectados por Cuantiles CF, el 92,35 % de los detectados por Loess y el 40,82 % de los detectados por Loess CF. Sin embargo, la cantidad de transcritos detectados aumenta considerablemente por efecto de la aplicación del algoritmo de corrección de ruido de fondo. En el caso de Cuantiles el número de transcritos detectados al aplicar corrección de fondo es en proporción el 212,36 % con respecto al mismo algoritmo sin dicha corrección, mientras que en el caso de Loess la misma relación es de un 226,22 %. Además, la intersección entre Cuantiles CF y Loess CF que no se comparte con los respectivos métodos sin CF es de un 49,62 % para el caso de Cuantiles y 47,34 % para el de Loess. CF = Corrección de ruido de fondo.

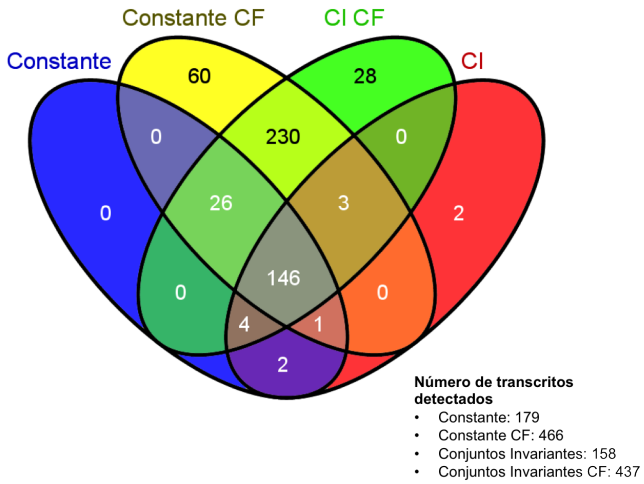

Figura 5.50: **Comparación de algoritmos de información parcial, efecto de la corrección de ruido de fondo.** La intersección entre los cuatro algoritmos conforma el 81,56 % de los genes detectados por Constante, el 31,33 % de los detectados por Constante CF, el 92,40 % de los detectados por Conjuntos invariantes y el 33,41 % de los detectados por Conjuntos invariantes CF. Tal como en los métodos de información total (figrua 5.49), la cantidad de transcritos detectados aumenta considerablemente por efecto de la aplicación del algoritmo de corrección de ruido de fondo. En el caso de Constante el número de transcritos detectados al aplicar corrección de fondo es en proporción el 260,33 % con respecto al mismo algoritmo sin dicha corrección, mientras que en el caso de Conjuntos invariantes la misma relación es de un 276,58 %. Además, la intersección entre Constante CF y Conjuntos invariantes CF que no se comparte con los respectivos métodos sin CF es de un 49,36 % para el caso de Constante y 52,63 % para el de Conjuntos invariantes.

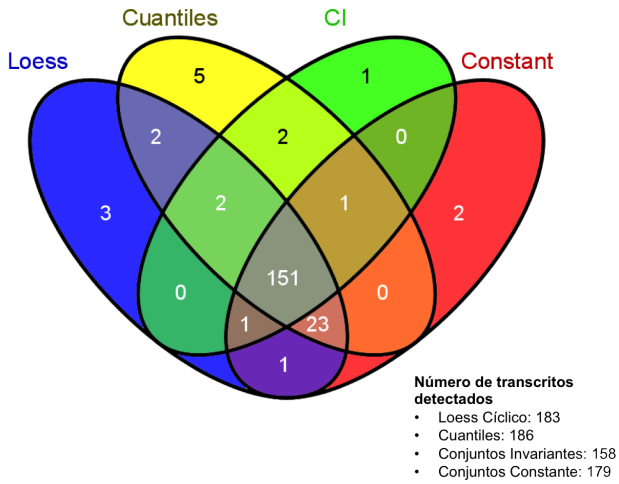

Figura 5.51: **Comparación de algoritmos de información parcial y total.** La intersección entre los cuatro métodos es de 82,5 % con Loess, 81,2 % con Cuantiles, 95,6 % con CI y 84,4 % con Constante. El resto de las particiones contienen un número de genes con una aparente distribución uniforme con excepción de la intersección entre Cuantiles y Constante (23 genes).

5.3.3. Predicción

A los datos seleccionados en la fase anterior como diferencialmente expresados, ahora se les aplica un método de *clustering* no supervisado con el objetivo de identificar los patrones de expresión diferencial correspondientes a algunos de los contrastes realizados.

Mapas de calor con los genes obtenidos en los contrastes con los algoritmos Cuantiles y Constante

En todos los mapas de calor presentados los dendogramas corresponden a un algoritmo de agrupamiento jerárquico con *average linkage*, aunque en distinto orden entre microarreglos tiende a agrupar por tejido (piel y cérvix), tratamiento (FVB y E6) y tipo de replicado

(técnico y biológico), lo cual concuerda con la biología del experimento. El agrupamiento entre genes muestra un poco de inestabilidad a pesar del alto nivel de consistencia entre los métodos (figura 5.51), aunque es similar en la mayoría los contrastes presenta diferencias importantes. Se hablará acerca de ésto en la discusión.

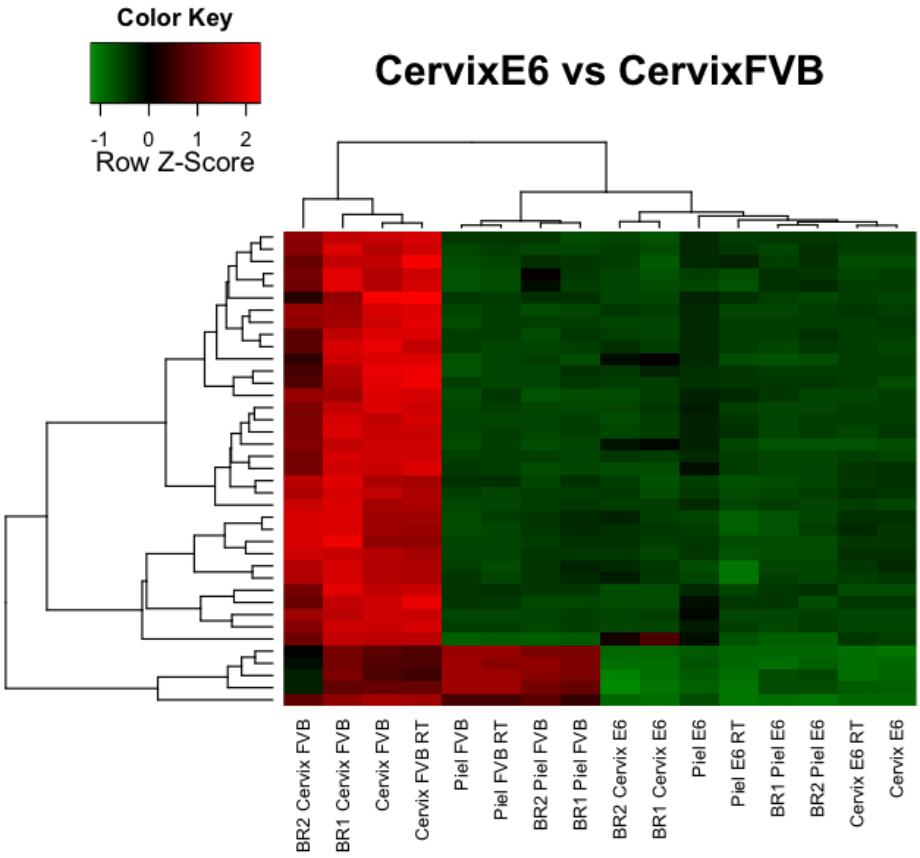

Figura 5.52: **Mapa de calor del contraste CervixE6 vs CervixFVB para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método de Cuantiles**. El agrupamiento jerárquico se aplicó a muestras y genes. La escala de rangos de color se muestra en la esquina superior izquierda de cada gráfico. Cada gen es representado por un sólo renglón de pixeles coloreados, cada microarreglo es representado por una sóla columna. Los diferentes *clusters* pueden identificarse por las agrupaciones de un mismo color, el cambio de color entre cada pixel es proporcional a la medida de similitud calculada.

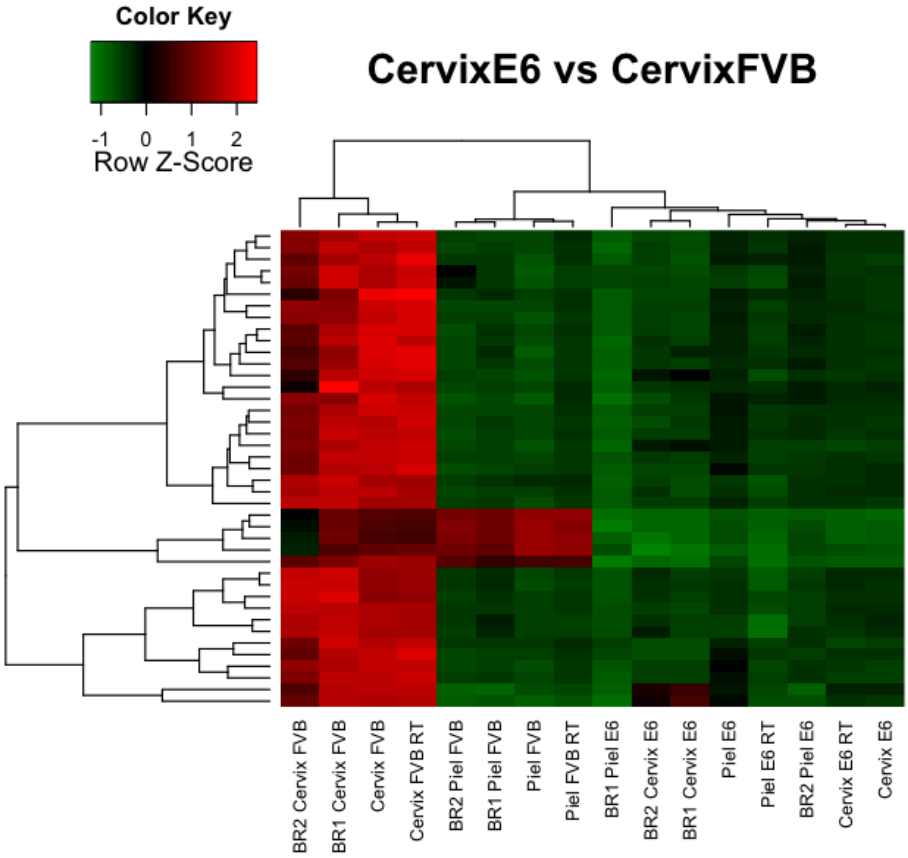

Figura 5.53: **Mapa de calor del contraste CervixE6 vs CervixFVB para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método lineal**.

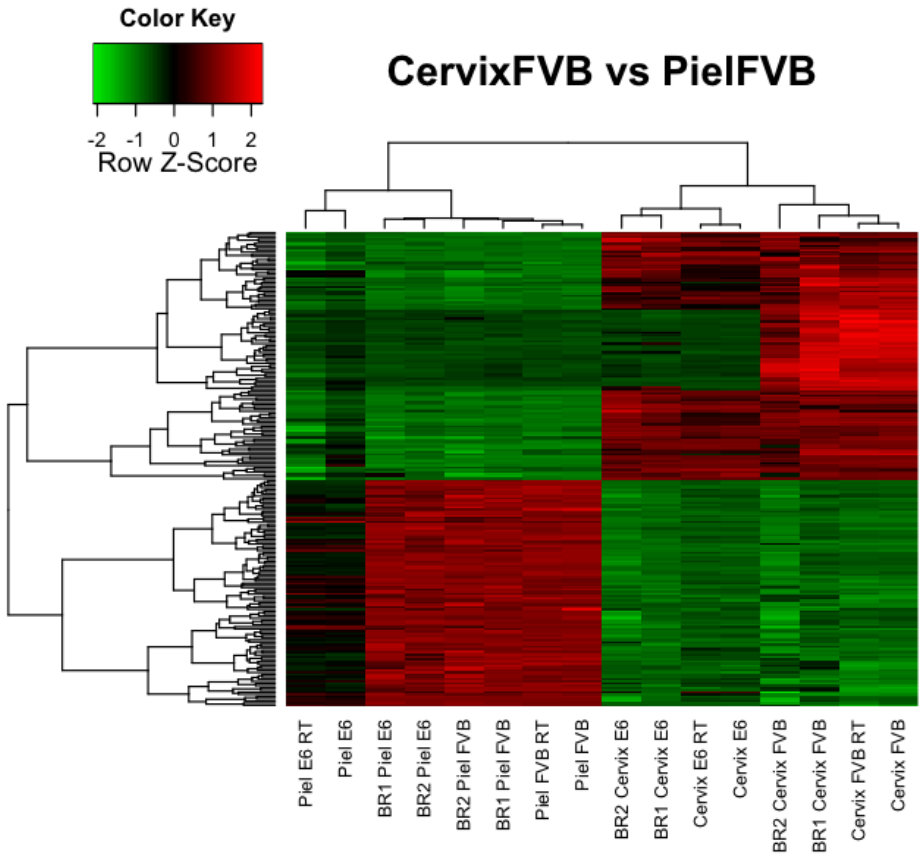

Figura 5.54: **Mapa de calor del contraste CervixFVB vs PielFVB para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método de Cuantiles**.

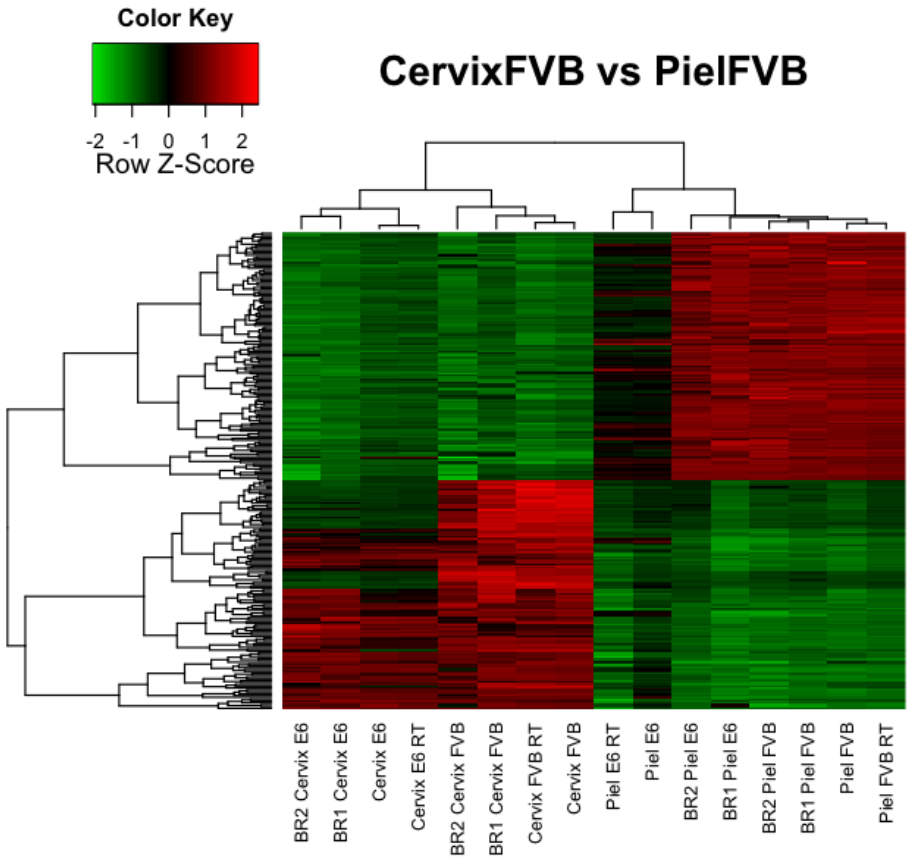

Figura 5.55: **Mapa de calor del contraste CervixFVB vs PielFVB para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método lineal**.

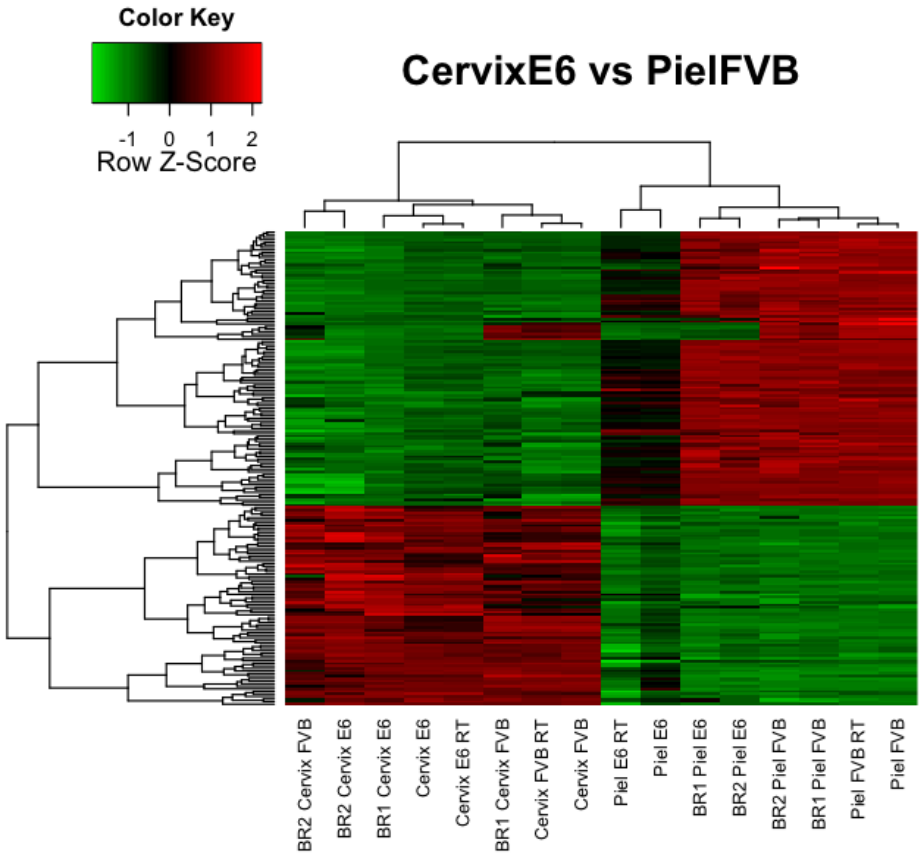

Figura 5.56: **Mapa de calor del contraste CervixE6 vs PielFVB para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método de Cuantiles**.

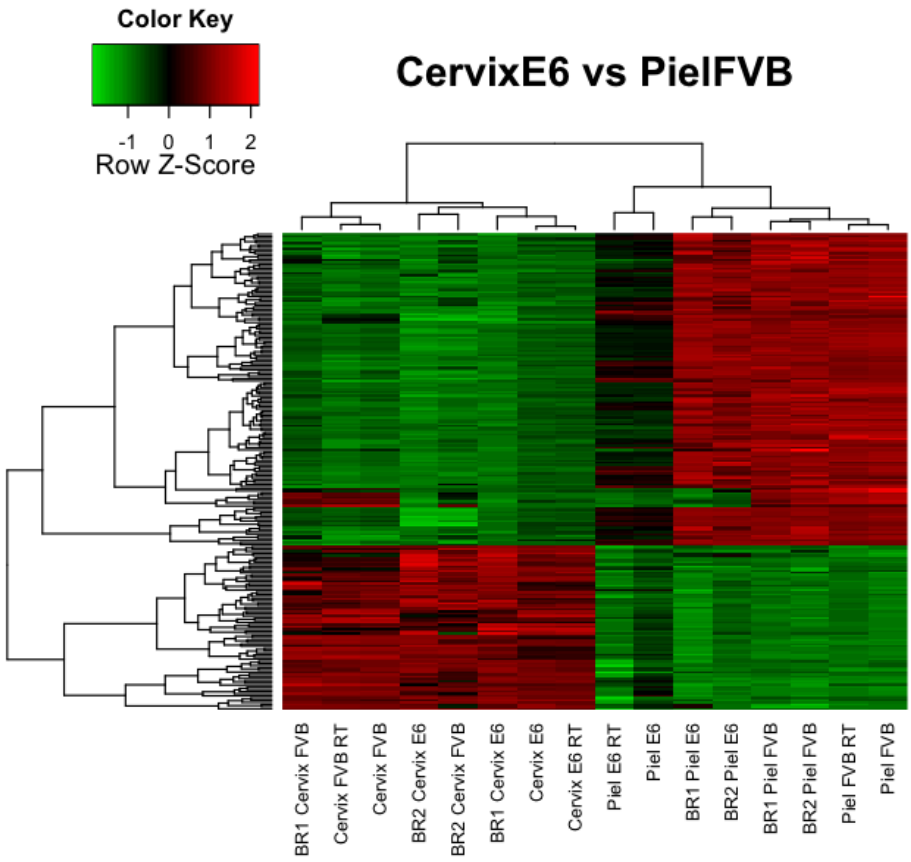

Figura 5.57: **Mapa de calor del contraste CervixE6 vs PielFVB para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método lineal**.

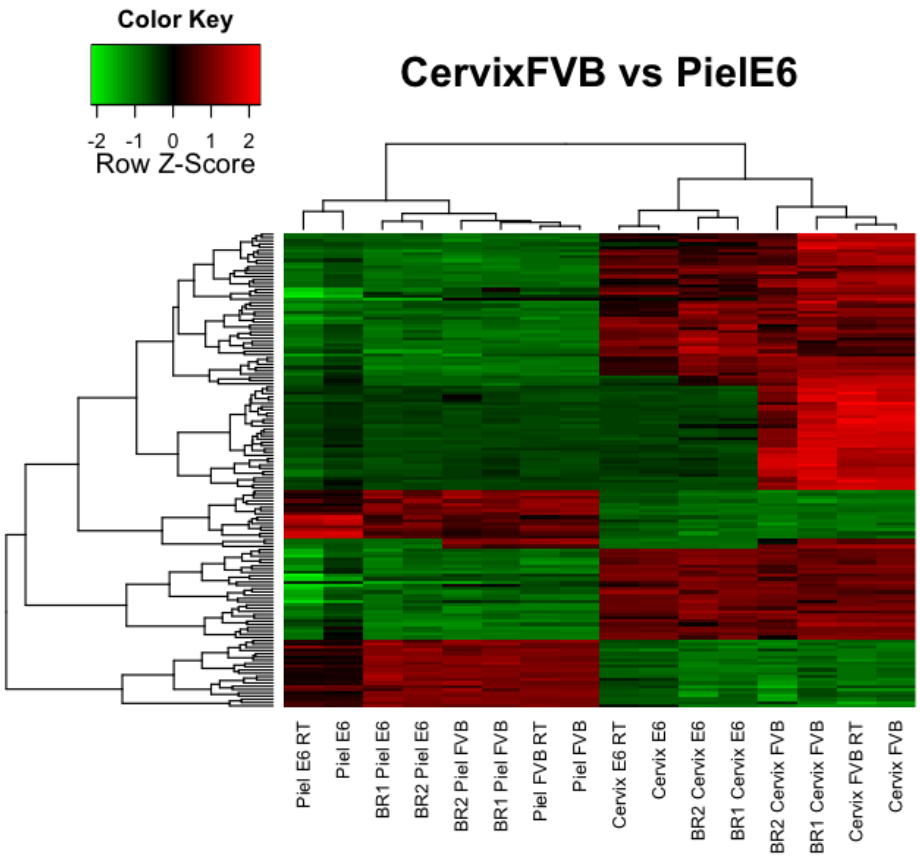

Figura 5.58: **Mapa de calor del contraste CervixFVB vs PielE6 para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método de Cuantiles**.

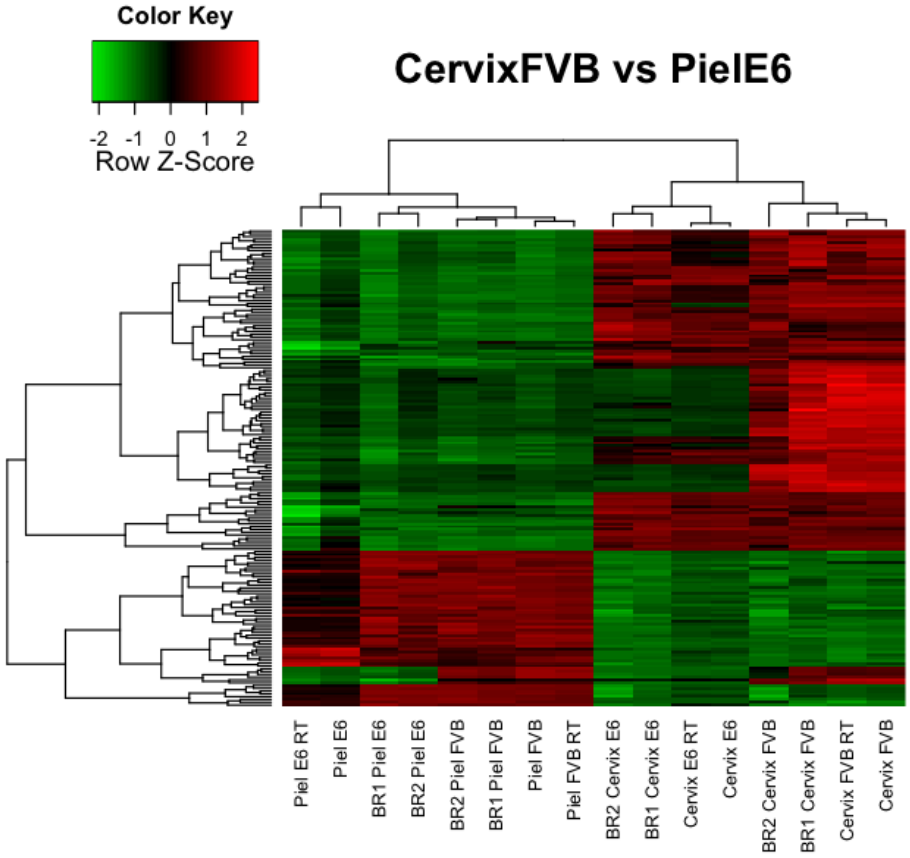

Figura 5.59: **Mapa de calor del contraste CervixFVB vs PielE6 para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método lineal**.

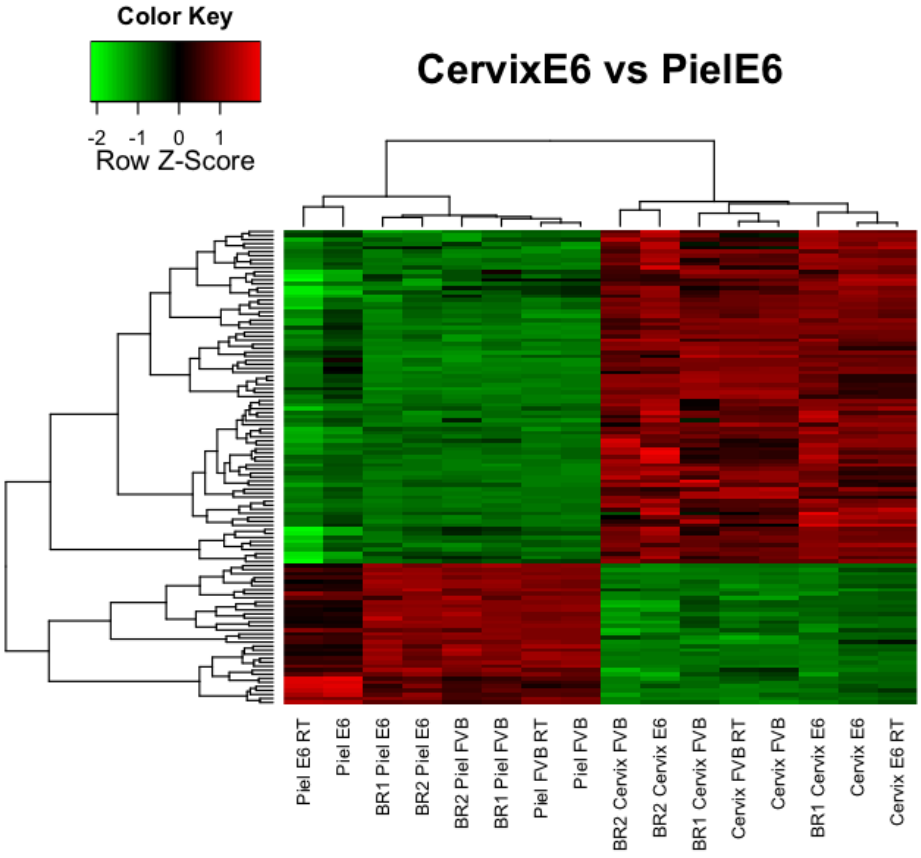

Figura 5.60: **Mapa de calor del contraste CervixE6 vs PielE6 para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método de Cuantiles**.

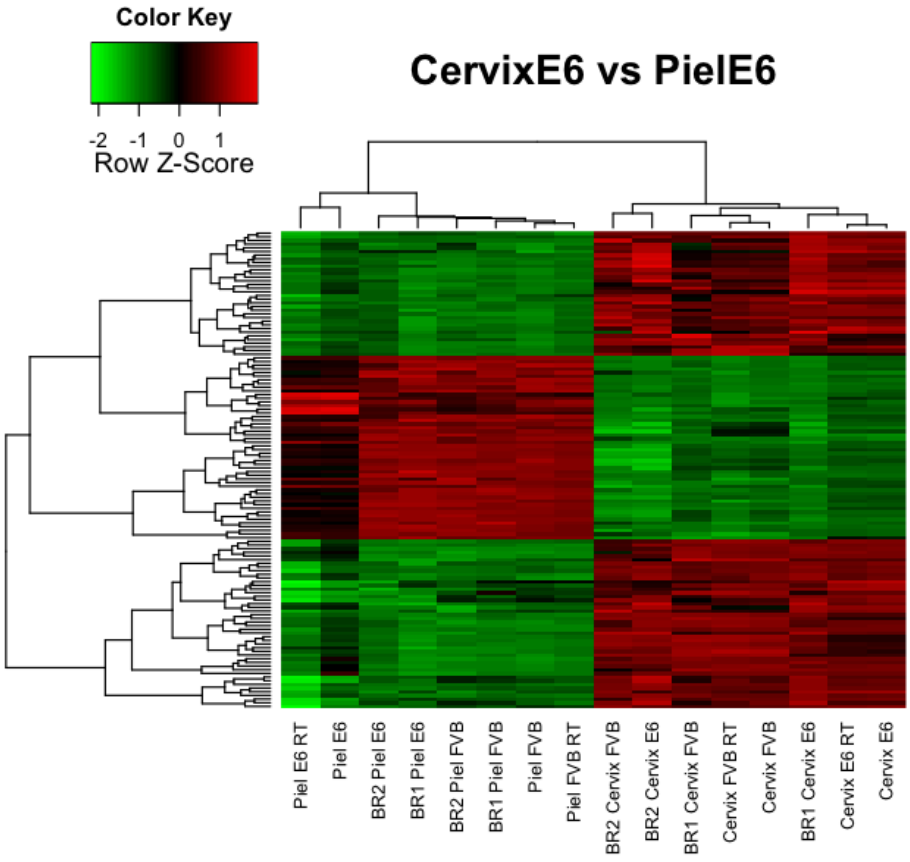

Figura 5.61: **Mapa de calor del contraste CervixE6 vs PielE6 para los datos seleccionados como diferencialmente expresados posteriormente a la normalización por el método lineal**.

5.4. Discusión

Los datos utilizados en este trabajo, resultan ser un caso a favor del uso de microarreglos ya que en muchas ocasiones se obtienen estadísticos de confianza poco favorecidos. En

muchos casos, las diversas fuentes de variación provocan resultados poco satisfactorios tanto cuantitativa como cualitativamente, de ello que se haga tanto énfasis sobre el diseño del experimento pues en muchas ocasiones se critica esta tecnología y a los algoritmos de análisis cuando los resultados obtenidos son pobres y de poca reproducibilidad. En realidad muchos de los problemas son producto de un mal manejo de dichas fuentes de variación y la limitada disponibilidad de muestras como factor que repercute en la robustez estadística de los resultados.

En relación a la fase de pre-procesamiento, los algoritmos de normalización comparados han sido usados y discutidos en un número importante de trabajos. Todos ellos son eficaces cuando se trata de reducir la variación oscura entre las muestras, pero cada uno tiene diversas ventajas y desventajas a tomar en cuenta. Incluso algunos de estos métodos utilizan ideas y técnicas de otros, como el caso de Contraste con Loess por citar alguno (ver capítulo 4). Los métodos basados en información parcial implican la posibilidad de introducir un sesgo en el análisis, en particular Constante tiene la limitación de asumir un comportamiento lineal (figura 5.27), mientras que Qsplines corre el riesgo de sobreajustar el comportamiento de los datos (figura5.26), Conjuntos invariantes es sensible con respecto a la definicón establecida de çonjunto de genes sin variación". Los métodos de información total como Loess Cíclico y Contraste pueden llegar a ser computacionalmente muy demandantes pero suelen preferirse por su misma naturaleza. En apariencia Cuantiles es el método más popular en la actualidad, es computacionalmente rápido, cumple relativamente bien con los criterios de decisión basados en gráficos como los MA (figura 5.18). Sin embargo, debe tenerse en cuenta que a costa de igualar casi perfectamente las distribuciones de los datos ejerce un efecto radical sobre ellos (ver figuras 5.16 y 5.16).

En torno a los procesos de normalización en dos pasos, en específico el propuesto en este trabajo que comienza aplicando Cuantiles por grupos (figuras 5.13). Se debe considerar el costo que representa una disminución en poder de la estimación de los Cuantiles a partir de las distribuciónes de los datos y sobre la variación de estos dentro del conjunto completo. Dadas la características del método su efecto es dominante sobre los procesos

de normalización aplicados después, de las figuras 5.32, 5.34, 5.36, 5.38 y 5.40, se observa que los gráficos de volcán se modifican muy poco en comparación a cuando los métodos se aplicaron directamente (figuras 5.30, 5.31, 5.33, 5.35, 5.37 y 5.39).

En la fase de clasificación, los diagramas de Venn proveen información interesante para la comparación. En la figura 5.47 se obtiene que en promedio Cuantiles y Loess coinciden en $95\,\%$ y Contraste en $79{,}6\,\%$, tomando en cuenta que los dos primeros métodos son más usados en la literatura se sugiere que Contraste podría encontrar un mayor número de falsos positivos. Sin embargo sólo a través de la validación de resultados mediante una técnica alternativa (como PCR) es posible verificar dicha aseveración. Los métodos de información parcial en todos los casos identifican un menor número de genes y su dispersión aparenta ser menor como se observa en la siguiente tabla.

Tipo de algortimo	Desviación estandar	Media
Información parcial	10.53565	196.6667
Información total	21.12660	168

En la figura 5.50 y 5.49 se observa un importante efecto de la corrección de ruido de fondo sobre el número de genes identificados. Una posible explicación de este comportamiento es la distribución binomial que adquieren los datos después de la corrección de ruido de fondo (figura 5.29), en la que se expande el rango dinámico y por lo tanto M (Comparar figuras 5.46 y 5.39 o cualquier otro par análogo de gráficos), generando más genes etiquetados como "diferencialmente expresados" de acuerdo al umbral establecido ($M = 2$). En el caso de los métodos de información total aproximademente el $95\,\%$ de los genes que se detectan en el conjunto sin corrección de fondo también son identificados cuando se hace la corrección y al comparar este efecto entre los dos tipos de algoritmos, contrasta el hecho de que se obtengan más genes en los métodos de información completa cuando no hay correción de fondo pero menos que en los de información parcial cuando sí la hay. Si la corrección de ruido de fondo transforma la distribución de los datos en bimodal, los supuestos de las pruebas estadísticas se comprometen y por tanto es probable que obtengamos un mayor volumen de falsos positivos. Es importante mencionar que en el

caso de los microarreglos de Affymetrix más recientes, se ha visto que el algoritmo RMA tiende a transformar los datos a una distribución como se observa en la figura 5.62.

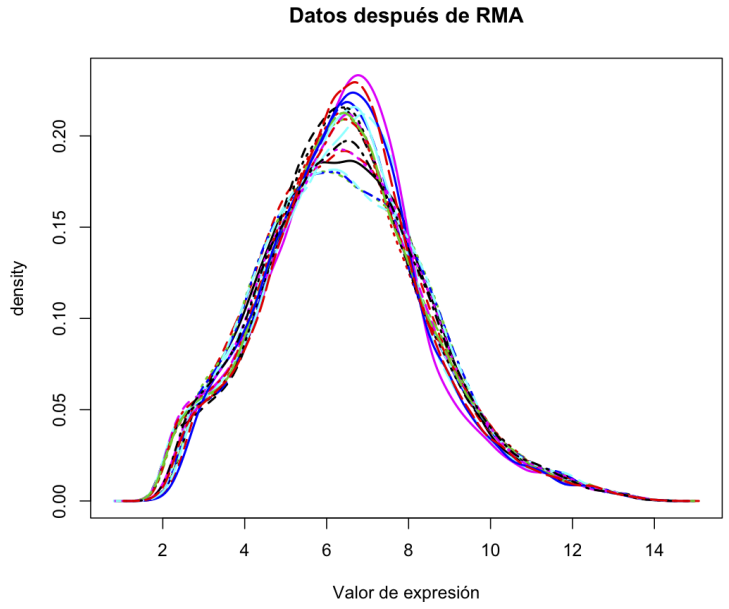

Figura 5.62: **Función de densidad de los datos de un microarreglo *HuGene 1.0 st* después de aplicar el algoritmo RMA**.

La figura 5.51 muestra el nivel de consistencia en la detección de genes diferencialmente expresados por los cuatro algoritmos seleccionados sin corrección de fondo, resalta el hecho de que la intersección de los 4 algoritmos identifica el $82{,}5\,\%$ de genes en relación al total con el método de Loess, $81{,}2\,\%$ con Cuantiles, $95{,}6\,\%$ con Conjuntos invariantes y $84{,}4\,\%$ con Constante. Los resultados podrían usarse para elegir el conjunto de genes en la intersección que por el hecho de aparecer con todos los algoritmos, hacen su detección más robusta desde un punto de vista estadístico.

Para la fase de predicción, se eligieron aleatoriamente los perfiles de expresión diferen-

cial provenientes del proceso de normalización por Cuantiles y Constante para realizar un agrupamiento jerárquico tanto entre muestras como entre genes y compararlo. En algunos contrastes los patrones parecen similares (figuras 5.52 a 5.59), pero en otros como el de Cérvix E6 vs Piel E6 hay diferencias claras (figuras 5.60 y 5.61). Además del efecto que pueda tener el algoritmo de pre-procesamiento aplicado, debe tenerse en cuenta que un método de agrupamiento como el utilizado en este trabajo compromete su estabilidad dentro del contexto de muchas variables pocas repeticiones[42]. Para lidiar con este problema se han desarrollado máquinas de aprendizaje estadístico que evalúan la estabilidad del número de grupos encontrados y sus elementos, en ocasiones estos métodos son exclusivamente diseñados para el análisis de expresión diferencial[43].

Para poder afirmar que alguno de los algoritmos de normalización comparados es el mejor es necesario realizar un análisis estadístico bastante más amplio en el que además se considere una variedad de conjuntos de datos provenientes de diferentes experimentos. No obstante, este trabajo provee fundamentos y ejemplos que se discuten a nivel que permiten a personas con conocimientos básicos de biología, estadística y computación familiarizarse con este tipo de análisis.

5.5. Conclusiones

Cuando se utilizan microarreglos, el diseño de experimento es determinante sobre la relevancia biológica de los resultados finales y un protocolo llevado a cabo con los cuidados pertinentes robustecerá cada paso del análisis. Todos los algoritmos de pre-procesamiento comparados son eficientes bajo el supuesto de que estos dos pasos (diseño de experimento y protocolo experimental) se realicen correctamente.

La elección del algoritmo de pre-procesamiento a utilizar, en principio depende de los recursos computacionales disponibles y el comportamiento de los datos antes y después de la normalización. El análista debe tener en mente la base biológica del experimento durante todas la fases del análisis. Aunque a partir de este trabajo no es posible definir un

estandar de oro, incluso se podría trabajar sólo con la intersección de los seis algoritmos.

Los resultados y la discusión en este trabajo pueden sugerir que los algoritmos de información total son mejores por sus características estadísticas pero se deja entendido que es necesario un análisis más profundo y exhaustivo para concluir que el método de Cuantiles es el mejor y por tanto el más usado.

Bibliografía

[1] Suárez Edna María, Barahona Ana. **(1992)**. Biología Molecular: La determinación de la estructura del ADN. Facultad de Ciencias. UNAM (México). LLULL, vol. 15, 1992, 395-41.

[2] Bohr, N. **(1933)**. Light and Life. Nature, 131, 421.

[3] Mofeeff-Ressovsky, N.W., Vmmer, K.G., Delbruck, M. **(1935)**. Ueber die Natura der Genmutation und der Genstruktur. Nachrichten der gelehrten Gesselschaften der Wissenchaften Math. Phisik KI., Frackgruppe 6(13), 190-245.

[4] Schrodinger, E. **(1944)**. What is Life?: The Physical Aspect of the Living Cell. New York, Cambridge University Press. Traducción en español: 1983, edición 11, Barcelona, Tusquets Editores.

[5] Ellis, E.L., Delbruck, M. **(1939)**. The growth of bacteriophage. J. Gen Physiol., $22(365)$, $631-642$.

[6] Stent, G. S. **(1971)**. Molecular Genetics; an Introductory Narrative. San Francisco: W.H. Freeman.Spanish translation: Stent, G. S. (1981). Genética molecular. Barcelona: Omega.

[7] Luria, S.E., Delbruck, M. **(1943)**. Mutations of Bacteria from virus sensitivity to virus resistance. Genetics, 28, $491-511$.

[8] Pauling, L. **(1960)**. The nature of the chemical bond. 3th edition. Ithaca, New York, Cornell.

[9] Watson, J.D. **(1982)**. La doble hélice. 1a edición, México, Consejo Nacional de Ciencia y Tecnología.

[10] Robert J. Brooker. **(2005)** Genetics: Analysis and Principles. 2th edition, USA. Mc Graw Hill.

[11] U.S Department of Energy Genome Program's-Biological and Environmental Research Information System. **(2008)**. Human Genome Project Information. Última modificación: Agosto 19, 2008. Disponible en http://genomics.energy.gov/

[12] J. Craig Venter-Celera Corporation Web Site. **(2010)**. History of Celera. Disponible en: https://www.celera.com/celera/history

[13] Watson, Baker, Bell, Gann, Levine, Losick. **(2008)**. Molecular Biology of the Gene. 6th edition, USA. Pearson-Benjamin Cummings, Cold Spring Harbor Laboratory Press.

[14] Guido Lastra L., Camila Manrique A. **(2005)**. Microarreglos: herramienta para el conocimiento de las enfermedades. Asociación Colombiana de Reumatologa.

[15] Tristan Mary-Huard, Jean-Jacques Daudin, Stéphane Robin, Frédérique Bitton, Eric Cabannes, Pierre Hilson. **(2004)**. Spotting effect in microarray experiments. BMC Bioinformatics 2004, 5:63.

[16] Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y. **(2003)**. Effects of atmospheric ozone on microarray data quality.

[17] Irizarry R. A., Hobbs Bridget, Collin Francois, Beazer-Barclay Yasmin, J. Antonellis Kristen, Scherf Uwe, Speed P. Terence. **(2003)**. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (2003), 4, 2, pp. 249–264.

[18] Rafael A. Irizarry, Laurent Gautier, and Leslie M. Cope. **2003**. The Analysis of Gene Expression Data: Methods and Software, chapter 4. Spriger Verlag.

[19] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. **2001**. Proceedings of the National Academy of Science U S A, 98:31.

[20] Lonnstedt Ingrid, Speed Terry. **(2001)**. Replicated microarray data. Department of mathematics, Uppsala University. Department of Statistics, University of California, Berkeley. Division of Genetics & Bioinformatics. Walter & Eliza Hall Institute of Medical Research, Melbourne.

[21] V. G. Tusher, R. Tibshirani, and G. Chu. **(2001)**. Significance analysis of microarray applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116-21.

[22] Hartigan, J. A. **(1983)**. Bayes Theory. Springer-Verlag, New York.

[23] Efron, B., Tibshirani, R., Goss, V. and Chu, G. **(2000)**. Microarrays and their use in a comparative experiment. Technical report, Stanfond University.

[24] Gareth Elvidge. **(2006)**. Microarray expression technology: from start. University of Oxford, Genomic group. Farmacogenomics 7(1).

[25] Helen C. Causton, John Quackenbush and Alviz Brazma. **(2003)**. A Begginers Guide Microarray Gene Expression Data Analysis. Blackwell Publishing.

[26] Hartemink, D. Gifford, T. Jaakola and R. Young. **(2001)**. Maximun likelihood estimation of optimal scaling factors for expression array normalization. In SPIE BIOS 2001.

[27] Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. **(2003)**. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193.

[28] Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. **(2002)**. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Stat. Sin., 12(1), 111-139.

[29] W. S. Cleveland and S. J. Devlin **(1988)**. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association, 83:596-610.

[30] Astrand M. **(2003)**. Contrast normalization of oligonucleotide arrays. J Comput Biol; 10: 95-102.

[31] Cheng Li and Wing Hung Wong **(2001)**. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Department of Biostatistics, Harvard School of Public Health, Boston, USA. Department of Statistics, Harvard University, Boston, USA.

[32] Christopher Workman, Lars Juhl Jensen, Hanne Jarmer, Randy Berka, Laurent Gautier, Henrik Bjorn Nielser, Hans-Henrik Saxild, Claus Nielsen, Soren Brunak and Steen Knudsen **(2002)**. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology, 3:research$0048,1-0048,16$ doi:10.1186/gb-2002-3-9-research0048.

[33] Salinas M. **(2007)**. Modelos de regresión y correlación IV. Correlación de Spearman. Cienc Trab. jul-sep;9(25):143:145.

[34] Pang-Ning Tan, Vipin Kumar, Michael Steinbach **(2005)**. Introduction to data mining. Person Education. Cap. 8.

[35] D Mendoza-Villanueva1, J Diaz-Chavez1, L Uribe-Figueroa, C Rangel-Escareão, A Hidalgo-Miranda, S March-Mifsut, G Jimenez-Sanchez, PF Lambert and P Gariglio. **2008**. Gene expression profile of cervical and skin tissues from human papillomavirus type 16 E6 transgenic mice. BMC Cancer, 8:347

[36] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, David Botstein **(1998)**. Cluster analysis and display of genome-wide expression patterns. Department of Genetics and Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305.

[37] R. R. Sokal, C. D. Michener **(1958)**. A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin, Vol. 28 (1958), pp. 1409-1438. Key: citeulike:5845721.

[38] Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. **(2000)**. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report No. 578.

[39] Smyth, G. K. **(2005)**. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York.

[40] Smyth, G. K. **(2004)**. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. Vol. 3, No. 1.

[41] Yoav Benjamini and Yosef Hochberg **(1995)**. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Tesiting. J. Royal Statistical Society. Series B Vol.57, No. 1, 289–300

[42] Ramoni, M., Sebastiani, P., & Kohane, I. S. **(2002)**. Cluster analysis of gene expression dynamics. In Proceedings of the National Academy of Sciences, 99:14, 9121–9126.

[43] Stefano Monti, Pablo Tamayo, Jill Mesirov, Todd Golub. **(2003)**. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning, 52, 91–118. Kluwer Academic Publishers. Manufactured in The Netherlands.