Alteraciones en la expresión de genes en cáncer de próstata

Gonzalo Hernández Viedma
Máster universitario en Bioinformática y bioestadística UOC-UB
Estudio genómico del cáncer

Laia Bassaganyas Bars
Jose Antonio Morán Moreno

4 de Junio de 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Alteraciones en la expresión de genes en cáncer de próstata*
Nombre del autor:	*Gonzalo Hernandez Viedma*
Nombre del consultor/a:	*Laia Bassaganyas Bars*
Nombre del PRA:	*Jose Antonio Morán Moreno*
Fecha de entrega (mm/aaaa):	*06/2018*
Titulación:	*Máster universitario en Bioinformática y bioestadística UOC-UB*
Área del Trabajo Final:	*Estudio genómico del cáncer*
Idioma del trabajo:	*Castellano*
Palabras clave	*Próstata, Mutilación, Expresión*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El cáncer de próstata es una enfermedad compleja que supone la neoplasia más frecuente en hombres. Existen tratamientos que tienen como finalidad la inactivación de la ruta de señalización iniciada por el receptor de andrógenos (del que depende el normal funcionamiento y proliferación del tejido prostático normal y tumoral) que tienen una alta tasa de respuesta en pacientes.

Desafortunadamente, para esta terapia de deprivación de andrógenos (ADT), el tumor acaba desarrollando una resistencia, ya sea mutando el receptor de andrógenos (para mantenerlo activo incluso en ausencia de andrógenos) o alterando otras partes de la ruta para poder sobrevivir.

Identificar los cambios de expresión en los genes que se producen conforme el tejido prostático normal evoluciona en tumoral, y éste evoluciona en metástasis nos permitiría identificar las rutas que se ven alteradas en cada etapa, lo que podría permitir desarrollar terapias dirigidas a revertir estas alteraciones.

El objetivo de este trabajo es, mediante el uso de datos públicos de expresión en diversos tejidos prostáticos determinar los cambios sufridos en las diversas etapas de la progresión de la enfermedad. También identificar las rutas biológicas alteradas e identificar las posibles causas (metilaciones aberrantes o variaciones en el número de copia).

Como principales alteraciones se han visto desregulaciones de rutas

relacionadas con el control de la transcripción y de la formación/control de uniones intercelulares.

Abstract (in English, 250 words or less):

Prostate cancer is a complex disease that represents the most frequent neoplasia in men. There are treatments that seek to inhibit the signaling pathway initiated by the androgen receptor (upon which normal and tumoral prostate cells rely to proliferate) that have a very high response in patients.

Unfortunately, for this androgen-deprivation therapy (ADT), the tumor develops a resistance, either by mutating the androgen receptor (to keep it active even in the absence of androgen) or by mutating other parts of the pathway in order to survive.

Identifying the changes in gene expression that takes places as the normal prostate tissue evolves into a primary tumor and finally generates a metastasis is crucial for identifying the pathways that become altered during these changes. Potentially leading to the development of therapies aimed at undoing these changes.

The objective of this work is, through the use of public data, to determine the expression changes in the different stages and trying to identify the leading causes of this changes (abnormal methylation status, copy-number variations…)

The main alterations observed are related to desregulations of pathways involved in control of transcription and formation of cell-to-cell interactions.

Índice

1. Introducción

1.1 Contexto y justificación del Trabajo

Receptores Nucleares

La capacidad de una célula para proliferar, sobrevivir y funcionar de una determinada manera viene determinada en parte por la capacidad de establecer "comunicación" con el entorno en el que se encuentra dicha célula. Esta comunicación se produce mediante la producción y recepción de una serie de mensajes químicos que pueden ser desde moléculas orgánicas pequeñas a macromoléculas grandes en forma de proteínas.

La recepción de estos mensajes comporta una respuesta por parte de la célula. La respuesta viene determinada por el mensaje recibido, la célula puede responder de múltiples formas: proliferando, iniciando procesos de apoptosis, secretando productos al exterior celular… [1].

En la mayoría de las situaciones, parte del proceso de respuesta al estímulo recibido pasa por un cambio en los programas transcripcionales de la célula. Es decir, la señal recibida va a provocar que un grupo determinado de genes vean reducida total o parcialmente su expresión, mientras que otros genes que no se expresaban lo harán. Estos cambios en la expresión génica serán los responsables de la respuesta celular.

Muchos de éstos "mensajes" no pueden entrar dentro de la célula debido a su incapacidad de atravesar la membrana celular, por ello, existen multitud de proteínas ancladas en la membrana celular cuya misión es recibir el mensaje y transmitir la información al interior celular. En el caso que la molécula mensajera pueda entrar dentro de la célula, existe un tipo de receptor que no se encuentran anclados en la superficie exterior de la célula, son los llamados *receptores intracelulares*.

Estos receptores intracelulares o también llamado *nucleares* por el compartimento celular donde actúan, funcionan como factores de transcripción que requieren de la unión a un *ligando* para poder actuar. Controlan un amplio abanico de procesos biológicos como pueden ser la proliferación celular, la diferenciación y la homeostasis celular [2].

El funcionamiento estándar de estos factores de transcripción dependientes de ligando es el siguiente:

-En ausencia de ligando, corepresores unidos al receptor lo mantienen inactivo.

-En presencia de ligando, se reclutan factores coactivadores que permiten al factor de transcripción alterar los patrones de expresión de la célula para poder responder.

Existen múltiples ligandos diferentes, cada uno con su receptor nuclear propio que controlará un determinado proceso biológico (**Figura 1**). A nivel estructural, pero, podemos encontrar una serie de dominios estructurales presentes en todos los receptores [3]:

AF-1 y AF-2: Dominios de activación N-terminal y C-terminal respectivamente.

DBD: *DNA-binding domain.* Dominio de unión al ADN. Región de la proteína responsable de la unión al ADN en las regiones de los genes que serán expresados por la acción del factor de transcripción.

Hinge Region: Región bisagra. Una región flexible que permite los cambios conformacionales requeridos para la activación del receptor.

LBD: *Ligand-binding domain.* Sitio de unión del ligando. Dominio donde la molécula del ligando se unirá al receptor, provocando el cambio conformacional y la activación del receptor.

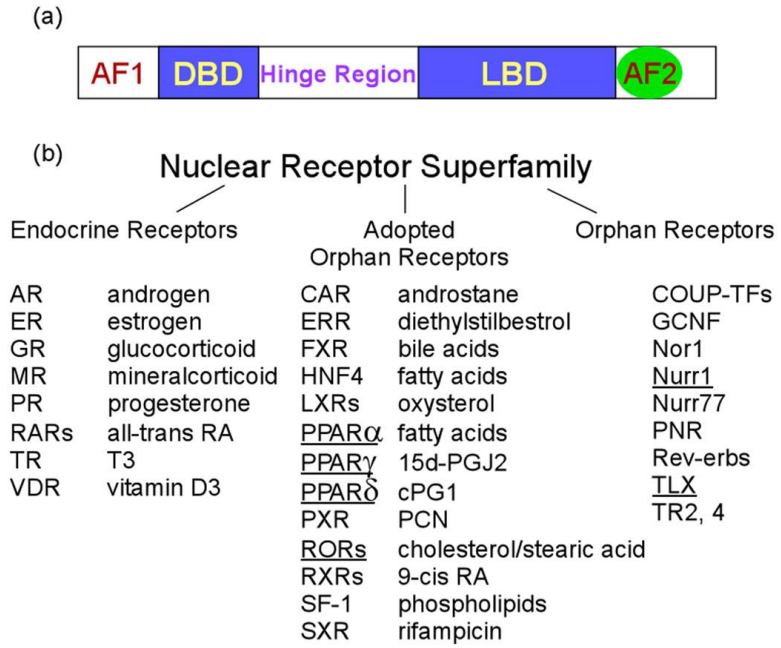


Figura 1. Estructura común de los receptores nucleares y listado de los receptores y sus ligandos si se conocen. *Figura extraída de (Shi, Y. 2007).*

Receptor de andrógenos

Uno de estos receptores nucleares es el llamado receptor de andrógenos (*AR* por su nombre en inglés), también conocido como NR3C4 (*nuclear receptor subfamily 3, group C, gene 4*). Es un miembro de la familia de los receptores nucleares de hormonas esteroideas donde encontramos otros receptores como el: receptor de estrógenos (ER), receptor de glucocorticoides (GR), receptor de progesterona (PR) y el receptor de mineralocorticoides (MR) [4].

El ligando del AR son los andrógenos *testosterona* y la forma modificada de ésta última llamada *5α-dihydrotestosterona* (DHT). El modelo de acción del AR es el siguiente (**Figura 2**) [5]:

-La testosterona circulante entra en la célula donde es transformada en DHT, forma más potente.

-La DHT se une al LBD del receptor lo que permite que sea translocada al interior del núcleo.

-Un dímero de DHT-AR se une a las regiones ARE, secuencias que se encuentran en los promotores de genes regulados por andrógenos.

-La unión de coactivadores al AR y al promotor resultan en la estimulación de la expresión de los genes que responden a andrógenos.

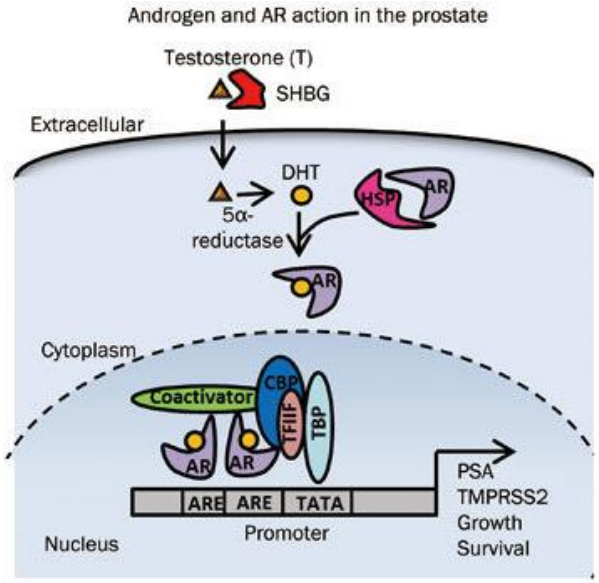


Figura 2. Modelo de funcionamiento del AR en tejido prostático. *Figura extraída de (Eileen Tan et al. 2015)*

Como resultado de la activación del AR, se inicia un programa transcripcional que tiene como resultado la estimulación del crecimiento del tejido prostático, así como la estimulación de la supervivencia celular. No sólo tiene estos efectos, si no que el AR controla la diferenciación sexual masculina durante el desarrollo, por lo que una ausencia de testosterona o la presencia de mutaciones en el AR que provoquen su inactivación causan el *síndrome de insensibilidad a los andrógenos*, cuya sintomatología consiste en la presencia de genitales femeninos, ausencia de próstata y de vello en el pubis y axilas entre otros [6].

El cáncer de próstata

La actividad del AR es una pieza clave en la supervivencia y funcionamiento normal de la próstata, ya que estimula el crecimiento y la capacidad de sobrevivir de las células que la forman. Por este motivo no es de extrañar la importancia de las alteraciones en la ruta de señalización iniciada por el AR en la supervivencia y progresión del cáncer de próstata.

El cáncer de próstata fue el tipo de tumor con el número de diagnósticos más elevado en España en 2015 [7] con un número de diagnósticos superior a los 33.000 casos (**Figura 3**). También es el tipo de tumor con mayor probabilidad de aparición en poblaciones de más de 69 años (**Tabla 1**).

Figura 3. Incidencia en hombres de los 10 tumores más frecuentes. *Figura extraída de REDECAN.*

HOMBRE

TIPO DE CÁNCER	0-39	0-49	0-59	0-69	0-79	0-84
Colorrecto	0,05	0,28	1,19	3,43	7,41	9,96
Estómago	0,02	0,11	0,38	0,98	2,11	2,90
Hígado	0,02	0,09	0,33	0,83	1,69	2,14
Labios, cavidad oral y faringe	0,03	0,23	0,80	1,54	2,29	2,64
Leucemia	0,17	0,24	0,37	0,69	1,34	1,76
Linfoma no Hodgkin	0,13	0,28	0,51	0,85	1,43	1,76
Próstata	0,00	0,06	0,96	4,66	11,12	14,46
Pulmón	0,04	0,40	1,85	4,63	8,77	11,02
Riñón	0,02	0,11	0,35	0,71	1,25	1,54
Vejiga	0,02	0,16	1,02	3,04	6,00	7,95
Todos los tumores (excl. tumores cutáneos no melanoma)	1,14	3,25	10,25	23,96	42,17	50,90

Tabla 1. Probabilidad de aparición de los distintos tipos tumorales en hombres en las distintas franjas de edad. *Tabla extraída de REDECAN.*

A diferencia de otros tipos de cáncer, cuya incidencia se ha mantenido más o menos constante en las últimas décadas en hombres, la incidencia de los tumores prostáticos no ha hecho más que aumentar desde 1993 hasta 2015 (en mayor medida que los cánceres colorrectales que también han visto incrementada considerablemente su incidencia) como puede observarse en la **Figura 4**. La incidencia de este tipo de tumores se ha triplicado aproximadamente desde 1993, lo que convierte al cáncer prostático en una patología de alta incidencia que tiene un gran interés desde el punto de vista sanitario.

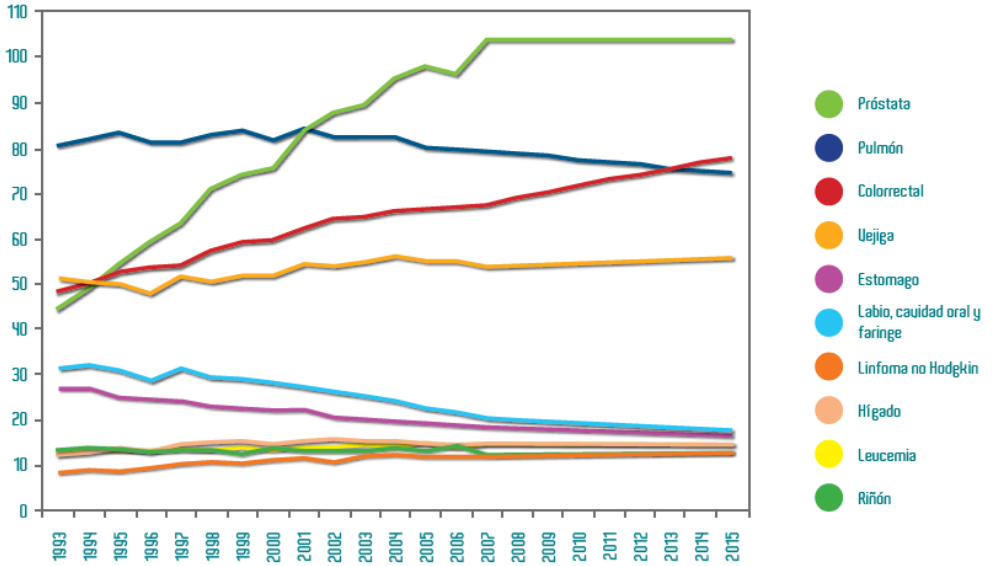


Figura 4. Estimación de la incidencia anual de distintos tipos tumorales en hombres desde 1993 hasta 2015 donde se observa el acusado incremento en la incidencia de los tumores prostáticos. *Figura extraída de REDECAN.*

Tratamientos en el cáncer de próstata

Dada la importancia que para la supervivencia de las células prostáticas (normales y tumorales) supone el correcto funcionamiento de la ruta de señalización iniciada por el AR [8], no es de extrañar que la primera línea de ataque a nivel terapéutico sea tratar de interferir con el normal funcionamiento de dicha ruta.

La terapia de deprivación de andrógenos, o ADT por sus siglas en inglés (*Androgen-deprivation therapy*) tiene como objetivo reducir los niveles en suero de testosterona del paciente a niveles que simulan la castración (<50 ng/dL) [9] y hoy en día es la primera terapia usada para el tratamiento de tumores prostáticos tanto localizados como metastáticos [10].

A pesar de que la tasa de respuesta a la ADT es alta, con el tiempo (inevitablemente), aparecerá en el tumor una resistencia a la castración generada por la ADT [11-13]. Dicha resistencia mayoritariamente aparece en el plazo de un año del inicio de la ADT.

Estos tumores resistentes a la ADT se conocen como *Castration-resistant prostate cáncer* o CRPC. Tratamientos alternativos pueden bloquear otras partes de la ruta del AR (más allá de reducir los niveles de testosterona), pero con el tiempo el CRPC desarrollará resistencias también a estos tratamientos de segunda generación.

A nivel metastático, el cáncer de próstata tiene un gran tropismo por el tejido óseo [14], generando múltiples lesiones metastáticas en el sistema esquelético (**Figura 5**) lo que añade morbilidad a las etapas más avanzadas del cáncer de próstata.

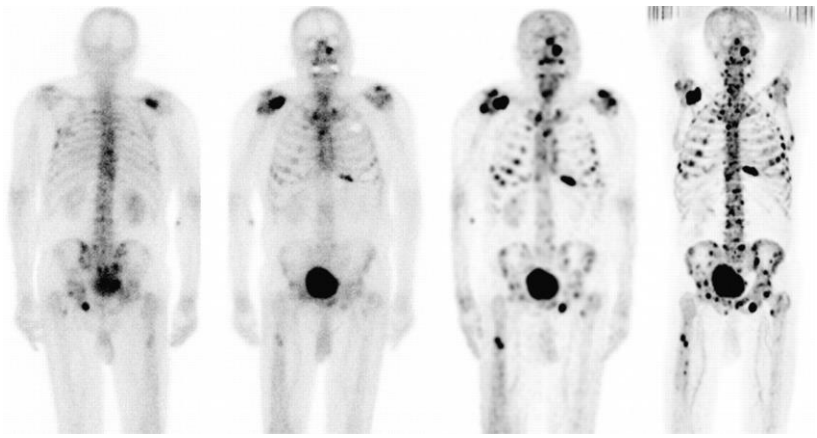


Figura 5. Paciente con cáncer de próstata que presenta múltiples metástasis óseas (puntos negros. El mismo paciente ha sido analizado con distintas técnicas de imagen. En todas ellas se observan las metástasis óseas. *Imagen extraída de (Even-Sapir E. et al, 2006).*

Aunque las terapias que tienen como objetivo la inhibición de la ruta del AR siguen siendo las más usadas para el manejo del cáncer de próstata, en los últimos años se han ido abriendo otras posibilidades fuera del AR.

Recientemente se ha demostrado que el tratamiento con inhibidores de la proteína PARP1 (PARPi) de CRPC con defectos en rutas de reparación del ADN es letal sintético, es decir, estos inhibidores matan selectivamente las células que portan dichos defectos en reparación [15]. En un estudio de 2015, se trató con Olaparib (un inhibidor de PARP1 aprobado por la FDA y la EMA para tumores de ovario y mama con mutaciones germinales en genes BRCA) a pacientes con CRPC que ya no respondían a los tratamientos estándar para el CRPC. Lo que observaron en este estudio fue que aquellos pacientes (16 de 49 totales) cuyos tumores portaban mutaciones en genes como BRCA1/2, ATM, genes de la ruta Fanconi y CHEK2 respondieron favorablemente al tratamiento con Olaparib [16].

El hecho de que existan subgrupos de pacientes que respondan bien a una determinada terapia (en este caso los PARPi) en función de la genética del tumor hace que sea extremadamente interesante la secuenciación de los tumores de los pacientes para poder predecir respuestas al tratamiento de forma previa, lo que implicaría usar la estrategia más adecuada y por lo tanto mejorar la supervivencia de los pacientes.

Alteraciones de la expresión genética

Múltiples enfermedades humanas (y animales) tienen como mecanismo de inicio o progresión la alteración en la expresión de determinados genes. Por ejemplo, en el caso del desarrollo tumoral es frecuente que se pierda o reduzca la expresión de uno o varios de los genes conocidos como *supresores de tumores*, que tienen como función principal la de limitar la proliferación/supervivencia celular o que se sobreexpresen los llamados *protooncogenes*, cuya función es estimular la supervivencia y proliferación celular [17].

Los mecanismos de alteración de estos genes son variados, podemos tener translocaciones que provocan la activación constitutiva de nuestro *protooncogen*, como puede ser la translocación entre el cromosoma 9 y el 22 que provoca la fusión de los genes BCR y ABL, siendo el gen ABL un *protooncogen* que en la forma fusionada con

BCR se encuentra constitutivamente activado, provocando la aparición de la leucemia mieloide crónica [18].

Otra forma de alterar la expresión de genes es mediante la modificación del estatus de metilación del promotor. Un promotor hipermetilado conllevará un silenciamiento del gen, aunque la secuencia codificante no se encuentre alterada, siendo este mecanismo una forma de inactivar a los genes supresores de tumores. Mientras que la hipometilación de las regiones promotoras conlleva la expresión (o incremento) de un protooncogen. Siendo este tipo de alteraciones una causa en múltiples ejemplos de carcinogénesis [19].

Una manera de alterar los niveles de expresión de los genes sin tener que recurrir a la alteración de la secuencia o al estatus de metilación de las regiones promotoras sería lo que se conoce como *Copy Number Variation* o CNV, es decir, variaciones en el número de copias. En esta situación, tenemos duplicaciones o deleciones de regiones que contienen uno o varios genes completos. Esto provoca que tengamos diferencias de expresión de los genes afectados: si tenemos amplificaciones de un gen (por ejemplo, pasando de 2 a 3 o más copias) tendremos un incremento de expresión del gen en cuestión. Existen múltiples casos de enfermedades y tumores causados por amplificaciones de genes [20].

En este escenario, si solamente analizáramos la secuencia del gen, veríamos una secuencia *wild-type* y descartaríamos erróneamente alteraciones en nuestro gen. Lo mismo pasaría si analizamos el patrón de metilaciones del promotor.

Estudios de las alteraciones en expresión

Cuando trabajamos con cambios en los patrones de expresión de genes disponemos de diversas metodologías para determinar en primer lugar, los posibles cambios de expresión ocurridos. Y en segundo lugar el mecanismo de aparición de estas diferencias (cambios en metilación, amplificaciones…).

Si queremos determinar cambios de expresión en uno o pocos genes, podemos recurrir a una técnica conocida como *Real Time-PCR.* Para determinar la expresión de un gen determinado primero deberemos aislar el RNA mensajero del cultivo celular de interés y generar un cDNA que usaremos como materia para el ensayo. El proceso resumido es el siguiente: mediante amplificación (y gracias a la presencia de un

fluorocromo) determinaremos la cantidad de un determinado RNA mensajero en nuestra célula, lo que nos permitirá comparar los niveles de expresión entre diferentes condiciones.

Si queremos obtener una imagen global de los cambios en el transcriptoma del cultivo entre diferentes situaciones (tratamientos o líneas celulares, por ejemplo) tenemos que recurrir a técnicas *hightroughput* como son los *arrays* de expresión genética.

Un *array* es un soporte físico donde se sitúan en posiciones determinadas un fragmento de ácido nucleico o *sonda*. Esta sonda puede ser sintetizada *in situ* o fijada en el soporte. Cada sonda contiene la secuencia complementaria a una determinada región de interés (gen).

El siguiente paso es hacer hibridar nuestras muestras (que estarán marcadas con un fluorocromo) con el *array*. Las sondas retendrán aquellos fragmentos cuya secuencia sea complementaria a la de la sonda.

Para llevar a cabo la detección tan solo tenemos que detectar la intensidad de la señal del fluorocromo en todas las posiciones del *array* donde tenemos depositada cada una de las sondas. Esta intensidad es proporcional a la cantidad de fragmentos presentes en nuestra muestra. A mayor intensidad de señal, más cantidad de fragmentos hay (en este caso implica mayor cantidad de RNA mensajero por lo que tenemos mayor expresión conforme va incrementándose la señal detectada).

Para estudiar el estatus de las metilaciones de un promotor de interés, usaremos un procedimiento conocido como *bisulfite conversión*. En este proceso, el ADN es tratado con bisulfito de sodio a 65ºC. Este tratamiento convierte las citosinas en uracilo, pero sólo funciona con las citosinas que están desmetiladas, las citosinas metiladas quedan intactas.

Si amplificamos mediante PCR y secuenciamos el ADN antes y después del tratamiento con bisulfito, veremos que al tratar tenemos timinas (en el lugar donde hemos generado uracilos con el bisulfito) que en el ADN antes del tratamiento leíamos una citosina. Esta comparación nos permitirá detectar las citosinas que estaban metiladas en la muestra original (que se habrán transformado en timinas en la muestra post-bisulfito).

De forma similar a como pasaba con el estudio de los niveles de expresión, si trabajamos con pocos genes podemos realizar el proceso uno a uno. Si tenemos múltiples genes o lo que queremos es hacer un estudio global de las metilaciones en

nuestra muestra, deberemos recurrir de nuevo al array de metilaciones, que funcionan de forma relativamente similar a los de expresión, pero el tipo de sondas usadas en éste caso son diferentes.

Para la detección de CNV en nuestras muestras, también disponemos de diversas opciones. Desde hace décadas, el uso de *arrays* ha permitido la detección de este tipo de alteraciones. Las amplificaciones causarían un incremento en la señal detectada mientras que las deleciones causarán una reducción de la misma [21].

La reducción en costes de la secuenciación de siguiente generación (o *Next Generation Sequencing* o *NGS*), ha permitido estudiar los CNV mediante el estudio de las secuencias de nuestras muestras. El uso de NGS ha permitido estudiar los CNV con una mayor precisión y resolución que mediante el uso de *arrays* [22].

1.2 Objetivos del Trabajo

El objetivo de este Trabajo de Final de Máster tiene como objetivo responder a las siguientes preguntas referentes al cáncer de próstata:

1. *¿Qué cambios a nivel de expresión génica existen entre el tejido normal, el tumoral localizado (o primario) y el metastático?*
2. *Asociado al objetivo 1: ¿Qué rutas, y en qué sentido, se ven alteradas en las distintas etapas de la progresión tumoral?*
3. *¿Existen cambios en los patrones de expresión entre tejido normal y tumoral prostático? ¿Coinciden con genes diferencialmente expresados?*
4. *De los genes diferencialmente expresados, ¿sufren variaciones en el número de copias (CNV) que puedan explicar los cambios de expresión?*

1.3 Enfoque y método seguido

Dada la imposibilidad de trabajar con muestras de pacientes no analizadas previamente, el enfoque de este trabajo de final de máster va a ser centrado en datos de experimentos de expresión génica y de metilaciones realizadas en otros laboratorios y depositadas en repositorios *online* de libre acceso.

La metodología a seguir va a ser aplicar *scripts* de R creados en parte para este trabajo y en parte adaptados de los paquetes de R disponibles como pueden ser *limma*, *minfi*…En el apartado 2 de este trabajo (**Metodología y datos**) se detallarán con mucho más detalle los datos usados, así como la metodología seguida paso por paso (el código completo de R usado se detallará en el apartado de **Anexo**.

1.4 Planificación del Trabajo

Dado que éste trabajo se ha basado completamente en datos ya generados disponible *on-line*, los únicos recursos disponibles son los siguientes:

-Acceso a internet: Para la obtención de los datos requeridos.

-Ordenador: Ejecutar el programa R y acceder a los datos

En cuanto a la planificación temporal del trabajo, la **Figura 6** detalla de forma aproximada la distribución de las distintas tareas asignadas:

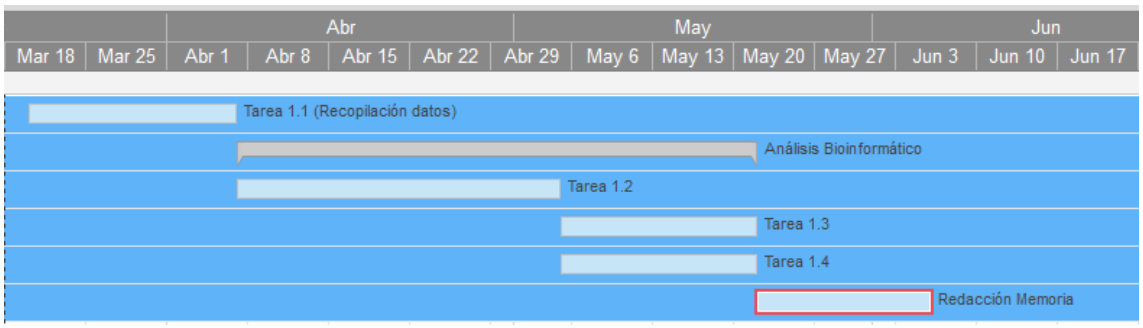


Figura 6. Marco temporal de las distintas etapas de elaboración del TFM.

Al tiempo de finalizar la Tarea 1.1 (*Recopilación de datos*) y empezar la etapa del análisis bioinformático corresponde con los objetivos asignados a la PEC1.

Al finalizar la etapa de la Tarea 1.2 (*Análisis Inicial de los datos*) correspondía a los objetivos de la PEC2. Que consistían en analizar las alteraciones de expresión en las muestras.

Al completar las Tareas 1.3 y 1.4 se corresponde con los resultados de la PEC3. Análisis final de los datos de metilación y CNV. (Esta última etapa no se completó del todo al finalizar la PEC3).

1.5 Breve sumario de productos obtenidos

El producto final de éste TFM no es un fármaco ni un protocolo de análisis de datos. En su lugar tendremos un compendio de alteraciones en expresiones (posiblemente asociadas a alteraciones en determinadas rutas biológicas) que estarán explicadas (o no) por metilaciones aberrantes en promotores y/o por amplificaciones/deleciones de genes.

De este listado de alteraciones se podrá, en un futuro, estudiar formas de revertirlas con el objetivo de reducir la agresividad de ciertos tumores prostáticos o de causar la reversión total/parcial de los mismos.

1.6 Breve descripción de los otros capítulos de la memoria

En el apartado 2 (**Metodología y datos**) se describirá con más detalle los datos usados para completar los objetivos de este TFM. ¿De dónde surgen los *datasets*? ¿Cuántas muestras hay? ¿Qué tipo de muestras tenemos?...También explicaremos de forma resumida los diferentes pasos usados con R para el análisis de los datos. También mostraremos los resultados parciales de cada etapa del análisis.

En el apartado 3 (**Resultados y discusión**) se detallarán los resultados finales de los diferentes *pipelines* descritos en el apartado 2. También discutiremos los resultados obtenidos en función de la bibliografía existente referente a las distintas etapas del cáncer prostático.

En el apartado 4 (**Conclusiones**) sintetizaremos los resultados en unas pocas ideas clave, será el mensaje para llevarse a casa de éste TFM. También discutiremos las posibles implicaciones de estas conclusiones en posibles futuras líneas de investigación terapéutica.

2. Metodología y datos

Dada la imposibilidad de trabajar con muestras biológicas de pacientes no analizados previamente, he recurrido a los repositorios de datos existentes en internet. Seguidamente detallaré la metodología y los datos usados para completar los objetivos marcados en el apartado anterior.

Todos los datos han sido analizados usando distintos paquetes disponibles del programa R. El código íntegro de R usado para cada uno de los pasos se puede encontrar en los **Anexos**.

Obtención de genes diferencialmente expresados

Para determinar el grupo de genes que ven variada su expresión según progresa la enfermedad, he utilizado datos de expresión de *arrays* donde se comparan muestras de tejido prostático ***Normal***, Tejido tumoral localizado (***primario***) y tejido proveniente de ***metástasis***.

Los datos los he obtenido de la web *Gene Expression Omnibus* (https://www.ncbi.nlm.nih.gov/geo/) de donde he descargado el *dataset* con el código **GSE6919**. Este *dataset*, tiene el título *"Expression Data from Normal and Prostate Tumor Tissues"* y fue depositado en la base de datos en 2007, pero se ha ido actualizando hasta 2016.

En el *dataset* tenemos datos de **504 muestras**, repartidas de la siguiente manera:

-233 de tejido normal prostático. **52 muestras** provenientes de tejido normal de donantes de órganos y **181 muestras** de tejido normal adyacente a los tumores analizados. Para simplificar hemos analizado éstas 233 muestras como un único grupo de tejido normal prostático.

-**196 muestras** de tumores primarios.

-**75 muestras** de tumores metastáticos provenientes de diversos tejidos fuera del tejido prostático (pulmones, nodos linfáticos, glándulas adrenales…)

No todas las muestras se analizaron en la misma plataforma (*array*), por lo que tendremos que tener en cuenta este dato a la hora de realizar el análisis. Sabemos que se usaron tres *arrays* diferentes: *GPL92*, *GPL93* y *GPL8300*.

Esta información está contenida en los archivos descargados y ya nos viene separada según plataforma. Solo tendremos que tener la plataforma usada en cada caso para poder asignar correctamente cada sonda del *array* al gen correcto.

Una vez escogido el *dataset* que vamos a analizar, procederemos al análisis de las muestras. Para ello hacemos uso del paquete *GEOQuery* de R. Con este paquete podremos descargarnos desde R los datos asociados a un determinado *dataset* de interés, en nuestro caso el *dataset* mencionado anteriormente.

```
> prostate
$`GSE6919-GPL8300_series_matrix.txt.gz`
ExpressionSet (storageMode: lockedEnvironment)
assayData: 12625 features, 171 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM152804 GSM152805 ... GSM187527 (171 total)
  varLabels: title geo_accession ... Tumor stage:ch1 (40 total)
  varMetadata: labelDescription
featureData
  featureNames: 1000_at 1001_at ... AFFX-YEL024w/RIP1_at (12625 total)
  fvarLabels: ID GB_ACC ... Gene Ontology Molecular Function (16 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL8300
```

Figura 7. Ejemplo del objeto de R creado al leer los datos del *dataset* **GSE6919** con el paquete *GEOQuery*.

Como resultado, obtenemos un objeto de R que hemos llamado *prostate* que es del tipo *ExpressionSet*. Este objeto es un compendio de otros objetos donde obtendremos más información y datos referentes a las muestras analizadas. Jerárquicamente contiene los datos de los tres *arrays* usados en el *dataset*. En la **Figura 7** solo se aprecian los datos de una de las plataformas (**GLP8300**).

Seguidamente separaremos los datos de las tres plataformas para analizarlos de forma individual. Llamaremos a estos objetos **pGLP92**, **pGLP93** y **pGLP8300** respectivamente. En realidad, estos archivos contienen listas que a su vez contienen listas.

Estos tres objetos contienen cuantiosa información, pero la relevante para el análisis son los resultados de expresión. Para llegar a los datos de expresión debemos entrar dentro de la lista llamada *assayData*, y dentro de ella encontraremos una tabla llamada *exprs* que contendrá los datos crudos de cada una de las sondas del *array* para cada una de las muestras analizadas (**Figura 8**).

Seguidamente aplicamos el logaritmo en base dos a todos los datos de expresión y tendremos unos datos donde podremos llevar a cabo la comparación. Pero primero recodificamos la variable *Description* que contiene el tipo de muestra analizada para

tener los tres grupos que hemos mencionado al inicio de esta sección (**Normal**, **T.Primario** y **Metástasis**).

```
> expGPL92[1:5,]
         GSM152822 GSM152823 GSM152824 GSM152825 GSM152826 GSM152827 GSM152828 GSM152829 GSM152830 GSM152831 GSM152832
41880_at      44.9     123.8      16.6     104.3      88.1      90.8     102.6      22.3      19.1      50.6      78.1
41881_at      28.8      61.9      17.5      41.8      74.6     171.5     147.0      95.9      59.2      41.8      74.5
41882_at       5.4       5.2      47.2       5.2       6.5       7.7       6.5       8.8       5.5      37.0       5.2
41883_at       8.8       9.1      18.0       6.1      15.5      25.7      20.5      11.0       7.4       5.2      11.6
41884_at       9.1      45.7      28.9      17.3      39.1      33.1      17.0      57.0      33.2      13.1      61.2
```

Figura 8. Datos crudos de expresión de las sondas (*filas*) para cada muestra (*columnas*) para el caso del *array* GLP92. *Resultados parciales*.

Para llevar a cabo el estudio de expresión diferencial usaremos el paquete *limma* de R. Nos interesa comparar la expresión entre los distintos grupos de muestras, y lo haremos por parejas, por lo que tendremos tres estudios diferentes para cada plataforma. Compararemos:

-Tejido Normal *versus* Tumor primario.

-Tejido Normal *versus* Metástasis.

-Tumor primario *versus* Metástasis.

Como resultado tendremos unas tablas como la que vemos en la **Figura 9**. La columna llamada *PROBEID* es el nombre de la sonda analizada, la columna *logFC* es el logaritmo del cambio de expresión de la sonda (en este caso comparando tejido **Normal** y **Tumoral Primario**). El resto de las columnas son los valores estadísticos usados en la comparación. Dado que estamos en una situación en la que estamos haciendo comparaciones múltiples (en cada grupo comparado hay múltiples muestras), el valor que nos interesa es el **valor p ajustado**, que lo encontramos en la columna de nombre *adj.P.Val*.

```
> tt92PvsN[1:5,]
          PROBEID     logFC        t      P.Value    adj.P.Val        B
38726_at 38726_at 0.8363611 6.620646 4.338422e-10 5.477257e-06 12.107262
922_at     922_at 0.7010436 5.757031 3.828516e-08 2.416751e-04  8.109948
39011_at 39011_at 0.3980952 5.560678 1.004695e-07 4.228091e-04  7.250551
954_s_at 954_s_at 0.4835693 5.137915 7.445960e-07 1.886794e-03  5.469489
1218_at   1218_at 0.7036947 5.137146 7.472452e-07 1.886794e-03  5.466336
```

Figura 9. Ejemplo del resultado obtenido de la comparación de expresión con el paquete *limma*.

Como *cutoff* para escoger las sondas que muestran una ***expresión diferencial significativa*** escogeremos las sondas que tengan un valor p ajustado inferior a 0.05.

Una vez determinadas las sondas significativas transformaremos el nombre de la sonda del *array* en el correspondiente nombre del gen asociado. Para hacer esta

conversión debemos tener en cuenta la plataforma en la que se ha analizado. Para cada plataforma deberemos usar un *database* diferente "hgu95b.db", "hgu95c.db" y "hgu95av2.db" que corresponden a las plataformas **GLP92**, **GLP93** y **GLP8300** respectivamente. Al final el proceso obtendremos una tabla como la expuesta en la **Figura 10**, donde tenemos el nombre del gen (SYMBOL) y el valor del cambio de expresión.

```
> gdeMN[1:5,]
  SYMBOL      logFC
1  ACBD6 -0.6489128
2 ACOT13  0.8044555
3 ADAM12  0.7627432
4   ADM5  0.5869478
5   AGO1  0.4675194
```

Figura 10. Resultado de la anotación de los nombres de los genes partiendo del nombre de las sondas del *array*.

En este punto, ya tenemos el resultado de los genes que varían su expresión en función del tejido que estamos analizando. También resulta interesante estudiar si estos grupos de genes están asociados a ciertas *funciones biológicas* o a determinadas *funciones moleculares*. Por ejemplo, podríamos observar un grupo de genes que reducen su expresión al progresar de tejido normal a tumoral primario que están relacionados con, por ejemplo, el ciclo celular. De esta forma podríamos decir que la progresión de tejido de próstata normal a tumoral primario comporta una alteración del control del ciclo celular.

Estos términos son los llamados *Gene Ontology (GO)*, que nos permiten agrupar múltiples genes a las mismas categorías funcionales. Para ello usaremos el paquete *clusterProfiler* de R cuya función *enrichGO* nos permitirá determinar si hay términos GO sobrerrepresentados en nuestro *set* de genes diferencialmente expresados.

Respecto a los términos GO los podemos dividir en tres grandes categorías: *Biologial Pathway (BP)*, *Molecular Function (MF)* y *Celular Component (CC)*.

Los términos BP nos dan una idea de la función biológica a la que se encuentra asociado un gen. Un ejemplo de término BP es *"DNA repair"*, que nos indica que los genes clasificados en esta categoría funcionan en rutas de reparación del ADN.

En el caso de los términos MF, nos indica con un poco más de detalla que actividad tiene nuestro gen (o la proteína que codifica si estamos hablando de genes que se traducen). *"Endonuclease"* es un ejemplo de término GO de clase MF, y nos

indicaría que la proteína asociada a dicho gen tiene la capacidad de unirse al ADN y cortarlo.

Por último, los términos CC nos indican en que compartimento celular se encontrará la proteína resultante de la expresión del gen. *"Nuclear"* es un ejemplo de término CC que nos indica que la proteína se encuentra en el núcleo celular.

Obviamente, un mismo gen puede tener asociados múltiples términos GO de cualquiera de las tres clases mencionadas. En la **Figura 11** vemos un resultado parcial del análisis de enriquecimiento.

```
> egoBPmn@result[1:5,]
                ID                                   Description GeneRatio   BgRatio       pvalue   p.adjust
GO:0010810 GO:0010810          regulation of cell-substrate adhesion    20/506 164/11518 3.314986e-05 0.06458119
GO:0042255 GO:0042255                              ribosome assembly    10/506  48/11518 3.549097e-05 0.06458119
GO:0010811 GO:0010811 positive regulation of cell-substrate adhesion    14/506  94/11518 5.713476e-05 0.06458119
GO:0060736 GO:0060736                        prostate gland growth     5/506  11/11518 5.946703e-05 0.06458119
GO:0000027 GO:0000027              ribosomal large subunit assembly     7/506  26/11518 9.611670e-05 0.08350619
               qvalue
GO:0010810 0.06362972
GO:0042255 0.06362972
GO:0010811 0.06362972
GO:0060736 0.06362972
GO:0000027 0.08227590
                                                                                                                  geneID
GO:0010810 ARHGEF7/EPHA3/NDNF/NF1/PTEN/ARPC2/BCL6/CIB1/COL16A1/CRKL/CYR61/ECM2/FBLN1/FGB/FLNA/FOXF1/FZD7/ILK/PTN/SFRP1
GO:0042255                                                                BRIX1/DHX30/MRPL11/MTERF3/RPL10/RPL11/RPL3/RPS10/RPS14/RRS1
GO:0010811                                ARHGEF7/NDNF/ARPC2/CIB1/COL16A1/CRKL/CYR61/ECM2/FGB/FLNA/FOXF1/ILK/PTN/SFRP1
GO:0060736                                                                                 AR/PTEN/CYP19A1/FGFR2/PSAP
GO:0000027                                                                BRIX1/DHX30/MRPL11/RPL10/RPL11/RPL3/RRS1
           Count
GO:0010810    20
GO:0042255    10
GO:0010811    14
GO:0060736     5
GO:0000027     7
```

Figura 11. Ejemplo de un resultado de enriquecimiento de términos GO aplicado a los genes diferencialmente expresados obtenidos en el apartado anterior.

En la columna *ID* tenemos el código numérico del término GO. En la columna *Description* encontramos la descripción de, en este ejemplo, la función biológica asociada al ID. La columna *GeneRatio* nos indica cuantos genes diferencialmente expresados están asociados al término respecto al total de genes diferencialmente expresados. La columna *BgRatio* nos indica la proporción de genes de todo el genoma asociados al mismo término GO en referencia al número total de genes del genoma humano. Las tres siguientes columnas son estadísticos usados durante el proceso. La columna *geneID* nos proporciona el nombre de los genes asociados a cada término (observamos que algunos genes se encuentran asociados a más de un término GO como comentamos más arriba). Por último, la columna *Count* nos indica cuantos genes de nuestro *set* están asociados a cada término GO.

Estudio de metilaciones

Para el estudio de las metilaciones hemos usado el paquete *minfi* de R. Esta parte del análisis de datos ha sido más exigente en cuanto a recursos informáticos necesarios para el procesado debido a la gran cantidad de datos que se han tenido que manejar. Para ello, hemos recurrido al paquete de R llamado *doParallel*, que nos ha permitido trabajar con más de un *core* de nuestro ordenador. También hemos tenido que ampliar la RAM disponible para ser usada por R mediante la creación de un archivo de paginación de mayor tamaño en el disco duro del ordenador donde hemos llevado a cabo el análisis de los datos de metilación.

De la misma forma que con los datos de expresión, los datos de metilaciones los hemos obtenido de la web del GEO. En este caso, el *dataset* usado tiene el código **GSE112047**. Este *dataset* publicado en marzo de 2018 tiene como título *Genome-Wide DNA methylation analysis of tumor and adjacent normal prostate tissues*. Las muestras están analizadas en la plataforma de Affymetryx **GPL13534** (Illumina HumanMethylation450 BeadChip).

Este *dataset* está formado por **47 muestras** obtenidas de tejidos parafinados, de las cuales 16 de ellas corresponden a muestras tumorales y 31 de ellas a tejido normal prostático. La situación ideal sería haber realizado los estudios de metilación diferencial en las mismas muestras donde hemos estudiado los cambios de expresión. Debido a la imposibilidad de realizar el trabajo de esta manera, no nos ha quedado más remedio de trabajar con datos generados con muestras de origen diferente, lo que nos limitará a la hora de extraer conclusiones de estos.

Para el análisis de los datos de metilación decidimos partir desde el punto más cercano al análisis de las muestras posible. Dado que para éste *dataset* están disponibles los archivos *idat*, que son los archivos resultantes de la lectura de los *arrays* de metilación, hemos decidido realizar el análisis desde el principio (lo que incluye la normalización de los datos, así como el control de calidad de los mismos).

Primero nos descargamos los archivos *idat* de todas las muestras del *dataset*, lo que suponen aproximadamente **565 Mb** de información. En este caso, para cada muestra tenemos dos archivos *idat*, uno para el canal verde y otro para el canal rojo.

Con la función *read.metharray.exp* del paquete *minfi* podremos leer todos los archivos *idat* que forman parte de nuestro *dataset*.

Partiendo de los datos crudos, el primer paso es analizar la calidad de la señal de las sondas para determinar si hay alguna muestra de baja calidad que deberíamos eliminar de nuestro estudio. En la **Figura 12** vemos el gráfico de control donde vemos representadas cada una de las muestras separadas en función de si son muestras provenientes de tejido normal o de tejido tumoral. Vemos que los gráficos se parecen, y que no hay muestras que descartar, por lo que continuaremos adelante con el análisis.

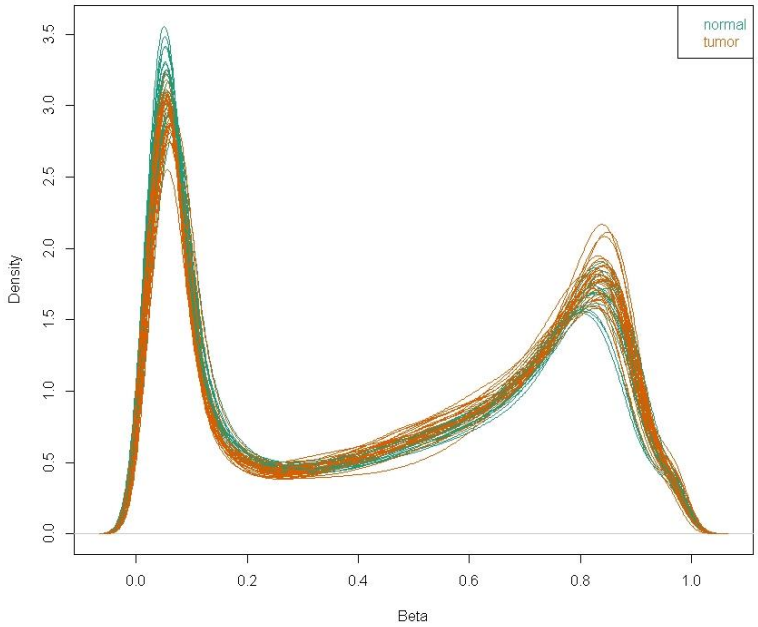


Figura 12. Gráfico con el control de calidad de las muestras analizadas en el *array* de metilaciones. Vemos que podemos proceder con el análisis con la totalidad de las muestras.

Después de normalizar los datos, usaremos la función *bumphunter* del paquete *minfi* para detectar las regiones diferencialmente expresadas (*DMR*). Esta función usa los valores de *Beta*, a partir de un valor *cutoff*, para detectar sondas que están sobrerrepresentadas. Es decir, estas sondas indicarían la presencia de regiones metiladas en el ADN. En nuestro caso el valor del *cutoff* usado es de 0.4.

Para que *bumphunter* funcione, tenemos que proporcionarle un modelo, es decir, ¿Qué vamos a comparar? En nuestro caso compararemos tejido normal de próstata con tejido tumoral (en esta ocasión no sabemos el tipo de tumor que tenemos). La función calculará una serie de permutaciones con las que realizará un cálculo estadístico para determinar aquellas sondas detectadas diferencialmente entre las muestras.

En la **Figura 13** se muestran los resultados parciales:

```
> dmrs2$table
     chr     start       end     value      area cluster indexStart indexEnd L clusterL p.value  fwer p.valueArea
58 chr20  44098387  44098724 0.4247198 1.2741594  121221     456796   456798 3       15    0.00 0.000        0.00
85  chr7 143579665 143579698 0.4880959 0.9761919  183092     213117   213118 2        4    0.00 0.000        0.00
79  chr6  29974858  29974863 0.4366941 0.8733882  163270     159843   159844 2       58    0.00 0.000        0.00
99  chrX 154842689 154842718 0.4248599 0.8497198  204667     485090   485091 2       11    0.00 0.000        0.00
3   chr1  46914023  46914033 0.4242248 0.8484496    7637      18763    18764 2        6    0.00 0.000        0.00
```

Figura 13. Ejemplo de los resultados obtenidos al aplicar la función *bumphunter*.

Cada fila corresponde con una DMR detectada. La tabla de resultados nos proporciona las coordenadas genómicas de su ubicación (número de cromosoma y posición de inicio y fin de la DMR). La columna *value* nos proporciona la media de la diferencia de señal entre los grupos analizados.

Para determinar si nos creemos que esa DMR realmente está diferencialmente metilada, recurrimos a la columna *fwer* que contiene un estadístico calculado que ya viene corregido para comparaciones múltiples (a diferencia de la columna *p.value* que el valor que aparece no está corregido).

Como estándar seleccionaremos aquellas DMRs con un valor de la columna *fwer* inferior a 0.01. Una vez seleccionadas las DMRs candidatas deberemos anotar los genes en las regiones y las compararemos con el listado de genes diferencialmente expresados encontrados en el apartado anterior.

Estudio de CNV

Para el estudio de las CNV, primero tendremos que tener el listado de genes diferencialmente expresados entre tejido **Normal-Primario**, **Normal-Metastático** y **Primario-Metastático**. Una vez tengamos este listado consultaremos los datos disponibles en el portal web ***cBioportal*** (http://www.cbioportal.org/).

cBioportal es un sitio web público desarrollado y mantenido por el *Memorial Sloan Kettering Cancer Center*. En el servidor encontramos **215 estudios** realizados en múltiples tipos de tumores, tanto en términos de tejido afectado como de estado de la enfermedad.

Dependiendo del estudio seleccionado tendremos diversos tipos de información disponible referente a las muestras. En general podemos encontrar información sobre expresión de genes (*RNAseq*), estudio de presencia de mutaciones y estudios de variaciones en el número de copias.

Si nos fijamos en los estudios asociados al cáncer de próstata vemos que existen **16 estudios** diferentes. En el caso de este TFM seleccionamos el estudio llamado ***"Prostate Adenocarcinoma (MSKCC/DFCI, Nature Genetics 2018)"***. Los motivos para seleccionar este estudio son varios:

-Es el más reciente de todos los estudios disponibles, ya que fue publicado en 2018.

-Este estudio tiene disponibles los datos de **1013 muestras** con datos de CNV disponibles para todas ellas. El doble del segundo estudio más grande disponible sobre el cáncer de próstata.

Desde la propia web de *cBioportal* se pueden consultar los datos de CNV para genes individuales o para listados de genes. Para automatizar los estudios de CNV una vez tengamos los resultados de los estudios de expresión diferencial, usaremos *scripts* del programa R para hacer las consultas en el estudio seleccionado.

Para ello instalaremos el paquete de R llamado ***cgdsr***. Este paquete nos permitirá establecer una conexión vía internet con el servidor de *cBioportal*. Una vez establecida la conexión podremos consultar el listado de los estudios disponibles y seleccionar el que nos interese.

Una vez determinado el estudio a recuperar, obtendremos un *ID* asociado al estudio, con el que podremos recuperar el identificador asociado al tipo de datos genéticos que nos interesa estudiar. En nuestro caso el *ID* del estudio es ***"prad_p1000"*** y dentro de los datos disponibles para este estudio nos quedaremos con el identificador asociado a los datos de CNV, que es ***"prad_p1000_cna"***. Estos datos los obtenemos con las funciones *getCancerStudies* y *getCaseLists* respectivamente.

Una vez determinado el estudio y el tipo de datos que seleccionar, deberemos recuperar los datos de las muestras, para ello usaremos la función *getProfileData*. A esta función le proporcionaremos como argumentos el *ID* del estudio, el identificador de los datos de CNV y un listado con los genes de los que queremos recuperar datos.

En la **Figura 14** vemos el resultado parcial de los datos de CNV obtenidos.

	ACTB	ACTB.1	ACTB.2	ADD1	API5	ARHGDIA	BCR
X2746	0	0	0	0	0	NaN	NaN
X7951	0	0	0	0	0	NaN	NaN
X10361	0	0	0	0	0	NaN	NaN
X10362	0	0	0	-1	1	NaN	NaN
X10363	-1	-1	-1	-1	0	NaN	NaN

Figura 14. Ejemplo de los datos y del formato de estos, obtenidos con la función getProfileData de los valores de CNV de los genes analizados.

Los resultados del estudio de CNV son números enteros que pueden ser los siguientes: [-2, -1, 0, 1, 2]. Estos valores están asociados al tipo de CNV detectado para el gen en cuestión:

-***Deep Deletion:*** Deleción bialélica del gen, asociada al valor de **-2**.

-***Shallow Deletion:*** Deleción monoalélica del gen, asociada al valor de **-1**.

-***No alteration:*** Sin deleciones detectables, asociada al valor de **0**.

-***Gain:*** Amplificación parcial del gen, asociada al valor de **1**.

-***Amplification*** Amplificación de alto nivel del gen, asociada al valor de **2**.

Una vez obtenidos estos datos determinaremos la proporción de cada una de las posibilidades del CNV para cada gen. Para simplificar la manipulación de los datos, reclasificaremos los valores de la siguiente forma:

-***Deleción:*** Todos aquellos casos con valores de CNV de -2 o -1.

-***Amplificación:*** Todos aquellos casos con valores de CNV de 1 o 2.

Una vez identificadas las proporciones de CNV, intentaremos determinar si existe correlación entre estos valores y los valores de expresión obtenidos en el primer punto de los resultados: *¿los genes sobreexpresados tienen mayores niveles de amplificaciones? ¿Los genes subexpresados tienen mayores niveles de deleciones?*

3. Resultados y discusión

Genes diferencialmente expresados

Como resultado inicial de los análisis de expresión diferencial hemos encontrado que:

-**52 genes** diferencialmente expresados entre tejido normal y tumoral primario.

-**568 genes** cuyas expresiones varían entre tejido normal y tumoral metastático.

-**3201 genes** con expresión diferencial entre tejido tumoral primario y metastático.

Si nos centramos en el sentido de la variación de la expresión (**Figura X**), vemos que en la mayoría de los casos al comparar tejido normal-primario es que los genes se sobreexpresan al progresar de tejido normal a tumoral (con unas pocas excepciones que se ve reducida la expresión).

En el caso de las otras dos comparaciones no hay un claro decantamiento por la sobreexpresión o la reducción. Lo que sí que parece observarse es que en ambos casos la variación en los niveles de expresión es muy heterogénea, tal y como se desprende de las amplias barras de error que aparecen en los gráficos de la comparación tejido normal-metastático y tumor primario-metastático.

Si graficamos los valores individuales de *logFC* para las tres comparaciones (**Figura 15 y 16**) observamos más claramente la tendencia mostrada previamente. En el caso de la comparación Normal-Primario, tenemos 8 genes que se encuentran subexpresados en el tumor primario en comparación con el tejido normal ($logFC<0$). Mientras que la gran mayoría de genes en esta comparación se encuentran en la región de la sobreexpresión ($logFC>0$)

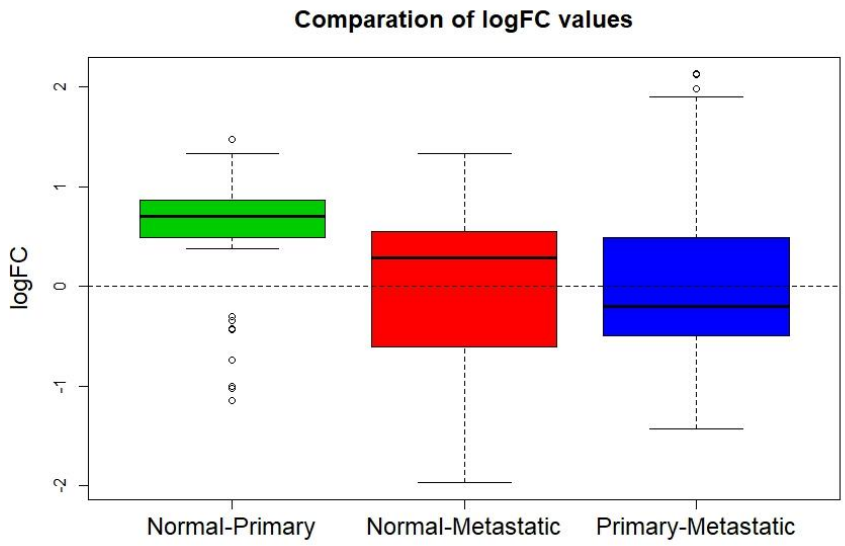


Figura 15. Comparación visual de los valores de diferencia de expresión entre las tres comparaciones realizadas donde se observa que la aparición del tumor primario provoca un incremento mayoritario en la expresión. En el resto de los eventos tenemos genes que se sobreexpresan y genes que se subexpresan.

En cambio, en el caso de la comparación del tejido Normal con el Metastático vemos que, aunque sigue apreciándose una mayor cantidad de genes en la región de la sobreexpresión, el número de genes subexpresados es mucho más grande que en el caso anterior.

Por último, en la comparación entre el tumor primario y el metastático, visualmente no se aprecian grandes diferencias entre los genes presentes en las dos regiones del gráfico.

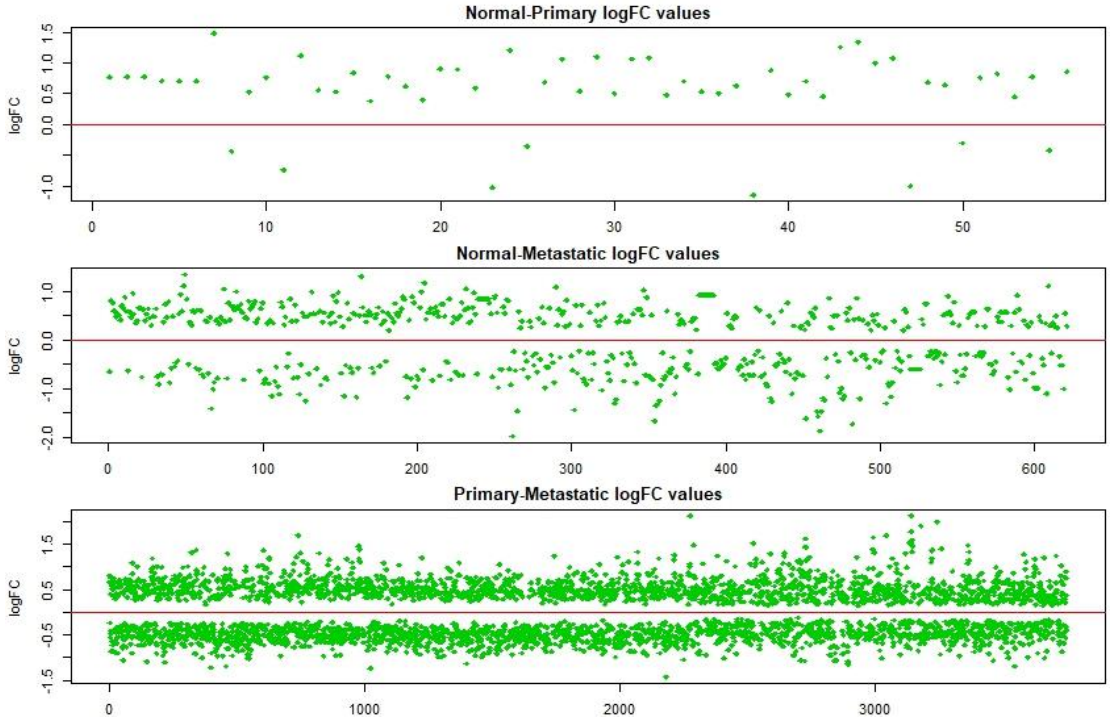


Figura 16. Valores individuales de cambios de expresión para todos los genes diferencialmente expresados hallados en este trabajo. La línea roja marca la frontera entre sobreexpresión y subexpresión. Se puede observar las distribuciones de valores comentadas en la *Figura 15*.

Una pregunta interesante que nos hacemos al ver estos genes diferencialmente expresado es la siguiente: *¿existen genes que alteran su expresión de forma diferente entre las diferentes comparaciones de tejidos?*

Comprobamos que existen **25 genes** que encontramos diferencialmente expresados en todas las comparaciones que hemos realizado. Es decir, estos genes alteran su expresión tanto cuando el tejido normal progresa a tumor primario, como al progresar a tumor metastático.

Tal y como vemos en la **Figura 17**, mayoritariamente el sentido de la alteración es opuesto entre la comparación *Normal_Primario* y entre *Primario_Metástasis*, lo que podría indicar que la acción de dichos genes podría ser beneficiosa para la aparición del tumor, pero perjudicial para la progresión a metástasis (o viceversa). También existen algunos genes cuya alteración va en la misma dirección en ambas comparaciones. Esto podría indicar un efecto positivo (o negativo) en la progresión entre las distintas fases del cáncer de próstata.

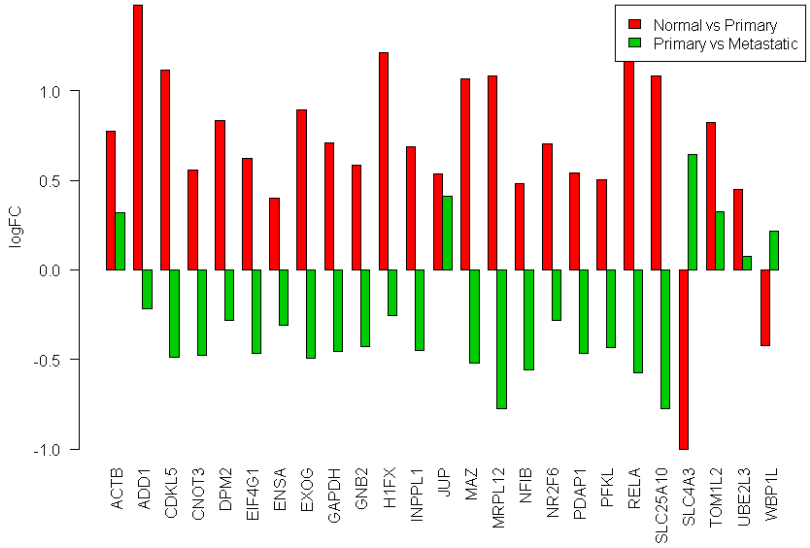


Figura 17. Valores de expresión diferencial de los genes encontrados en la comparación ente tejido normal-primario (rojo) y tumor primario-metastático (verde) donde se observan claramente que muchos genes alteran su expresión en sentidos diferentes dependiendo de la etapa de la progresión tumoral.

De este grupo de genes, sólo tenemos **4 casos:** *ACTB*, *JUP*, *TOM1L2* y *UBE2L3*.

ACTB: Gen de la β-actina. Proteína que forma parte del citoesqueleto y que se encuentra desregulada en muchos tumores y está asociada a una mayor invasividad del tumor [23].

JUP: Gen de la plakoglobina. Proteína implicada en la formación de las uniones intermedias célula-célula y que se ha visto que altos niveles de expresión está asociada a una mayor mortalidad en pacientes con adenocarcinomas pulmonares [24].

TOM1L2: Gen que codifica la proteína "*target of myb1 like 2 membrane trafficking"*. Implicada en el tráfico de vesículas. No existe un aparente *link* entre *TOM1L2* y la aparición/progresión tumoral.

UBE2L3: Gen que codifica para una enzima E2 requerida para la ubiquitinación de otras proteínas. Se han visto fusiones de *UBE2L3* con el gen *KRAS* en un subgrupo de tumores metastáticos de próstata [25] y se sabe que es un coactivador de los receptores de hormonas esteroideas [26].

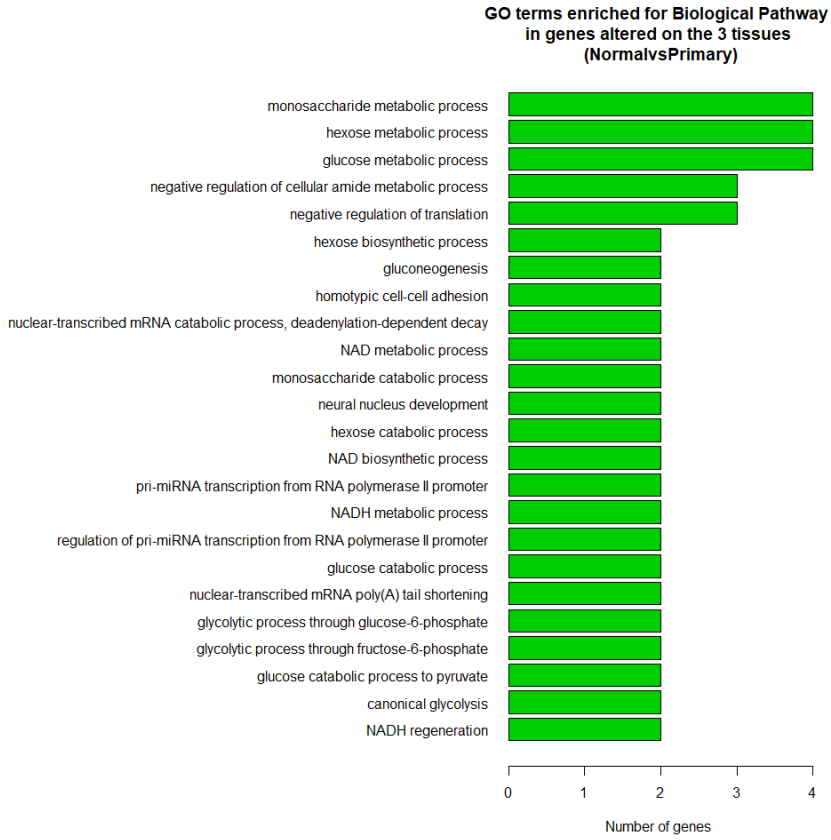
Normalmente, las alteraciones no se reducen a simples cambios en los niveles de expresión de un determinado gen. Frecuentemente lo que se ve alterado es un grupo de rutas o procesos biológicos (formados por decenas de genes implicados) que conllevan

un cambio en la funcionalidad y respuesta celular. Por ello, ahora pasaremos de estudiar genes individuales a determinar que rutas biológicas están alteradas en nuestros diferentes tejidos. Al tener genes que se alteran entre la comparación *NormalvsPrimario* y en la *PrimariovsMetastático*, la alteración de la ruta podría no ir en la misma dirección. Por ello no sólo miraremos las rutas enriquecidas, si no también miraremos el sentido de la alteración en las dos comparaciones.

Como vemos en la **Figura 18**, si nos centramos en las rutas biológica alteradas, vemos que entre la comparación *NormalvsPrimario* y *PrimariovsMetastático* sólo hay una ruta que difiere en el sentido de la alteración. En la primera comparación está claramente sobreexpresada mientras que en la segunda tenemos el mismo número de genes sub y sobreexpresados.

Vemos existen un gran número de rutas biológicas que, aunque diferentes tienen en común que controlan partes del metabolismo celular (metabolismo glucosa, monosacáridos, NADH…) así como control de la expresión.

Estos resultados nos enfatizan el hecho de que la aparición y posterior progresión tumoral van más allá de simples inactivaciones de rutas de reparación o señalización. El metabolismo energético de la célula tumoral se ve afectado, debido a las demandas energéticas y metabólicas asociadas con la naturaleza tumoral de la célula.



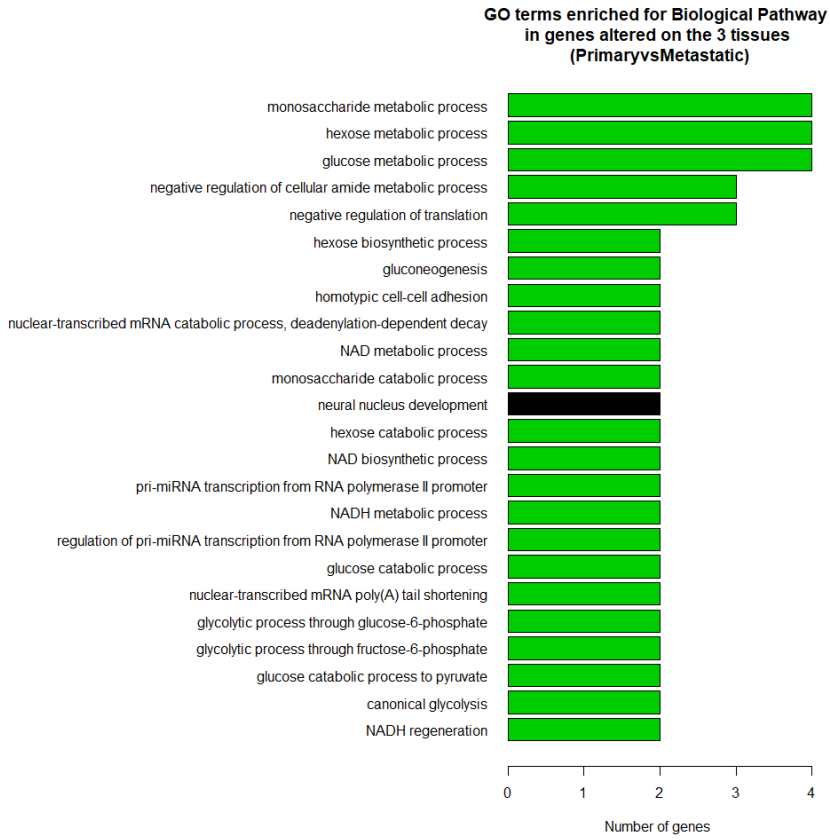


Figura 18. Estudio de enriquecimiento de términos GO asociados a ruta biológica en el grupo de genes alterados en todos los tejidos. Color verde indica que hay más genes sobreexpresados que subexpresados. Color negro significa que hay igualdad en el número de genes sobre/subexpresados.

Si nos fijamos en la **Figura 19**, vemos que en términos de funciones moleculares, la alteración entre *NormalvsPrimario* y *PrimariovsMetastático* es bastante diferente:

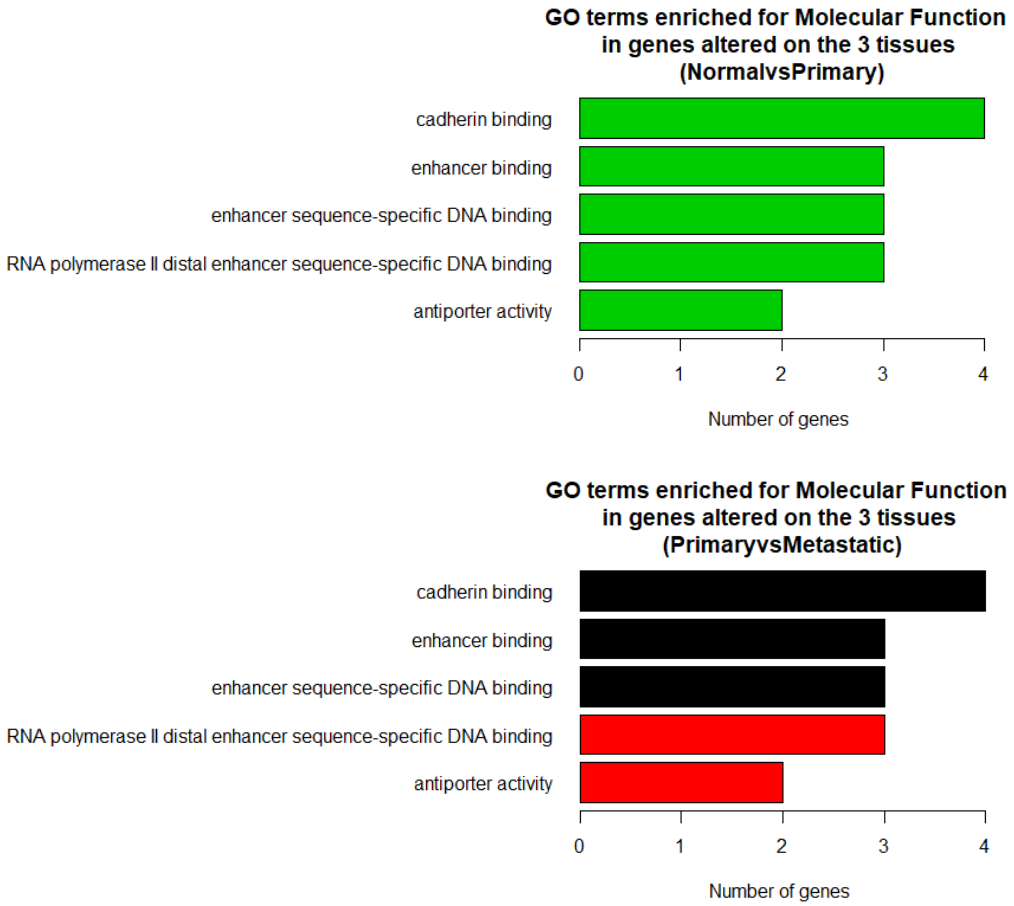


Figura 19. Estudio de enriquecimiento de términos GO asociados a función molecular en el grupo de genes alterados en todos los tejidos. Color verde indica que hay más genes sobreexpresados que subexpresados. Color negro significa que hay igualdad en el número de genes sobre/subexpresados. Color rojo significa que la mayoría de genes asociados a la ruta está subexpresados.

En la primera comparación, todas las rutas están sobreexpresadas, mientras que en la segunda, o están subexpresadas o no parece que haya una alteración clara de las mismas.

El grupo de rutas que en la segunda comparación no aparecen con sub o sobreexpresadas nos estarían indicando que aunque tengamos expresiones diferenciales de los genes asociados, a nivel de ruta no se ven cambios, ya que dicho cambio habría tenido lugar durante la transformación del tejido normal en tumoral primario.

Destacan las funciones moleculares que están asociadas con la transcripción (unión a *enhancers* en general). Destaca también el número de genes asociados a la unión a las *cadherinas*, proteínas imprescindibles para la formación de uniones adherentes entre células. De las funciones moleculares asociadas a la transcripción, deducimos que este grupo de genes identificados básicamente funcionarían controlando los perfiles de transcripción de otros genes lo que le permitiría a la célula adaptar su metabolismo (entre otras funciones que no hayamos detectado) a las diferentes circunstancias del tejido.

La asociación con las *cadherinas* nos indicaría que también estarían regulando en parte la forma celular y la relación con las células vecinas, posiblemente modificando las uniones adherentes de la célula tumoral con sus vecinas. Esta última función molecular se ve reforzada al estudiar el enriquecimiento de términos GO asociados a *Compartimento Celular*.

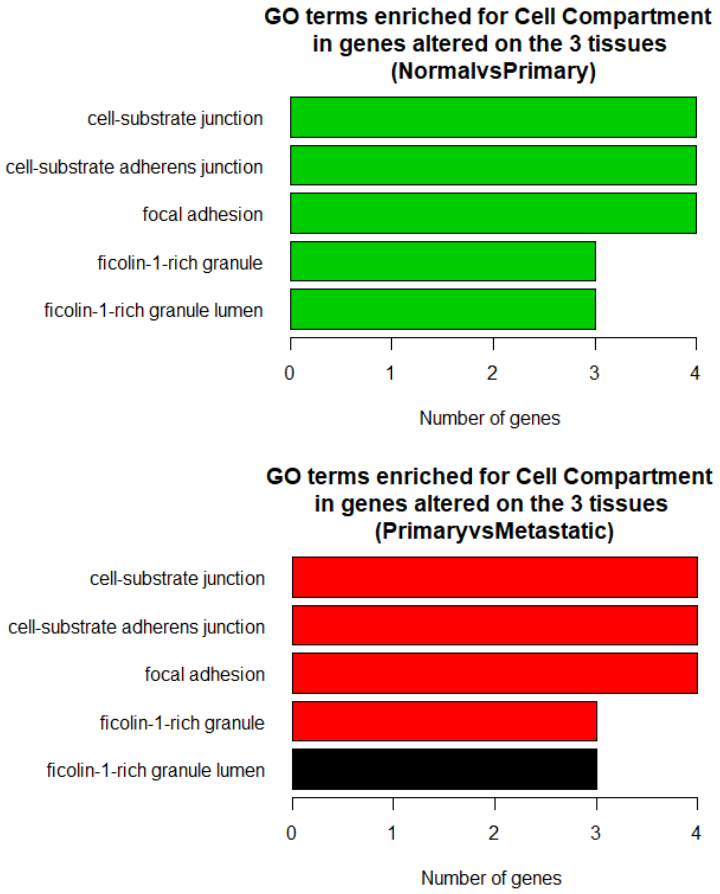


Figura 20. Estudio de enriquecimiento de términos GO asociados a compartimento celular en el grupo de genes alterados en todos los tejidos. Color verde indica que hay más genes sobreexpresados que subexpresados. Color negro significa que hay igualdad en el número de genes sobre/subexpresados. Color rojo significa que la mayoría de genes asociados a la ruta están subexpresados.

En la **Figura 20**, vemos que estamos en una situación parecido a lo ocurrido con los términos GO asociados a función molecular. En la primera comparación, todas las rutas enriquecidas están sobreexpresadas, mientras que en la segunda, o se subexpresan o no aparece una alteración clara.

Llama la atención el hecho de que en la segunda comparación las rutas que se subexpresan están básicamente asociadas con interacciones célula-célula, hecho que podría favorecer la migración de las células metastáticas.

Independientemente de la comparación, vemos que los términos que aparecen están básicamente asociado a uniones celulares (adhesiones focales, unión adherente…) lo que nos indica la importancia de la alteración en la interacción célula-célula en el desarrollo tumoral.

En primer lugar, hemos estudiado las alteraciones comunes al cáncer de próstata, pero también resulta interesante analizar las alteraciones específicas de fase: *¿Qué alteraciones vemos cuando una célula normal se tumoriza? ¿Qué alteraciones vemos cuando el tumor primario es capaz de metastatizar?*

Comparación tejido Normal con tejido tumoral primario

Primero miraremos que pasa cuando una célula normal se tumoriza, *¿qué alteraciones observamos?*

En este primer paso habíamos detectado **52 genes** alterados. Si nos fijamos en la **Figura 21**, vemos un gran número de rutas biológicas alteradas. Vemos que en la mayoría de los casos, las rutas alteradas se encuentren sobreexpresadas (color verde). Tenemos tres rutas en las que no podemos decir que haya sobre o subexpresión, ya que tenemos el mismo número de genes sobre o subexpresados en la ruta. Pero si nos fijamos en el nombre de éstas rutas no parecen ser relevantes para el tumor (están asociadas con el oocito y la meiosis femenina).

La ruta con mayor número de genes asociada es la llamada *"cell cycle phase transition"*. No sorprende este resultado, ya que las alteraciones en el control del ciclo celular es un evento típico de la aparición del tumor. Si esta ruta no se viera afectada, muy probablemente la célula no llegara a tumorizarse. El hecho de que la ruta esté sobreexpresada (y asumiendo que esto se correlaciona con una mayor actividad de dicha ruta) implica que el proceso de transformación de una célula normal en una tumoral

comporta una desregulación del ciclo celular (en este caso estimulando la capacidad de la célula de progresar por el ciclo y dividirse).

Seguidamente, con 5 genes asociados, tenemos tres rutas biológicas asociadas con el metabolismo de los monosacáridos/glucosa…Cambios seguramente orientados a satisfacer las necesidades energéticas crecientes de la célula tumoral. Como ya hemos visto este evento estaría asociado no tanto a la etapa de la enfermedad como a la presencia o no de ella.

Con menos genes asociados, tenemos otras rutas asociadas con el metabolismo energético y con el metabolismo asociado a la transcripción/traducción.

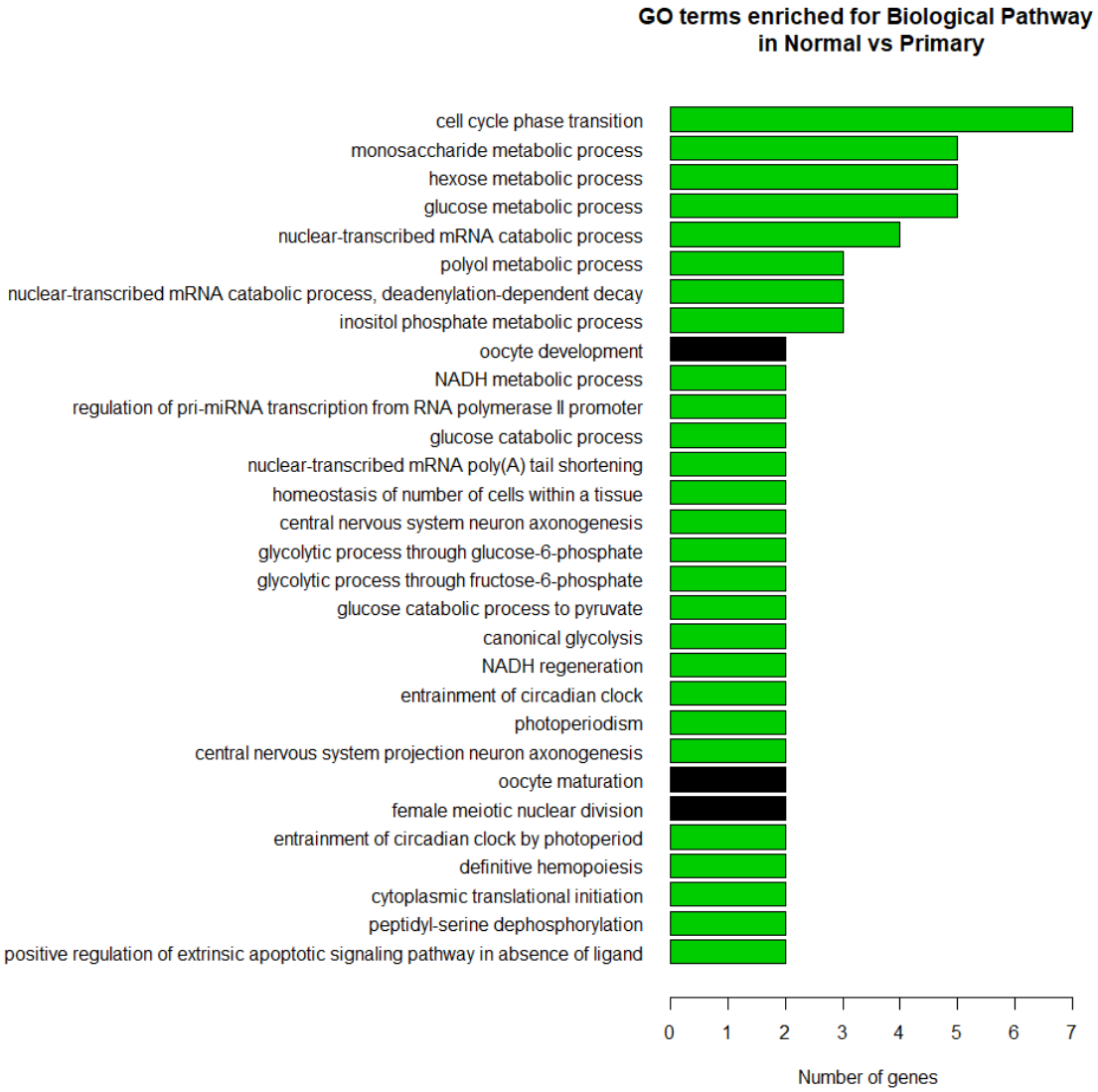


Figura 21. Estudio de enriquecimiento de términos GO asociados a ruta biológica en la comparación de tejido normal *versus* tumoral primario. Color verde indica que hay más genes sobreexpresados que subexpresados. Color negro significa que hay igualdad en el número de genes sobre/subexpresados.

Si nos fijamos en las funciones moleculares llevadas a cabo por los productos de los genes diferencialmente expresados (**Figura 22**), vemos que todas ellas están sobreexpresadas en el tejido tumoral primario. La mitad de las funciones moleculares enriquecidas están asociadas a la unión al ADN, indicando la importancia de las alteraciones en transcripción y remodelación de cromatina asociadas a la transformación tumoral de una célula sana. También existen 3 funciones asociadas con la interacción célula-célula. Y, por último, tenemos dos funciones asociadas con la defosforilación de proteínas (*phosphatase binding* y *phosphoric ester hydrolase activity*). Esta última

función es relevante, ya que la fosforilación (o ausencia de fosforilación) de proteínas controla la actividad de muchas proteínas implicadas en procesos importantes como ciclo celular, apoptosis, metabolismo, transcripción…

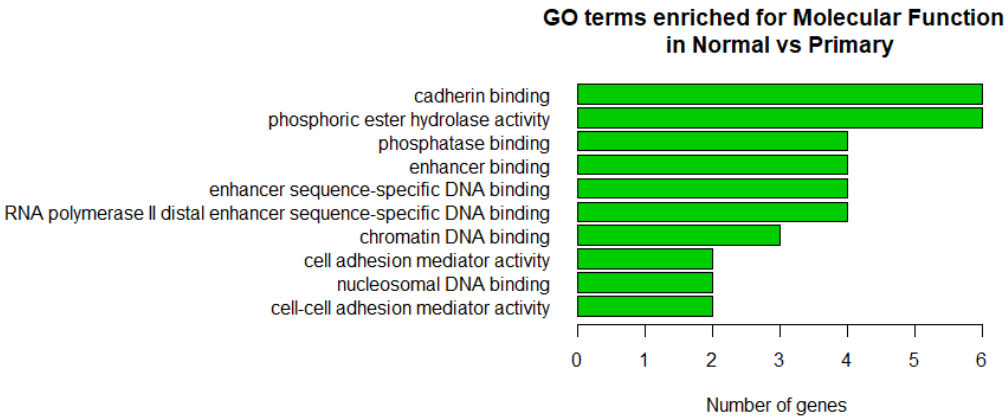


Figura 22. Estudio de enriquecimiento de términos GO asociados a función molecular en la comparación de tejido normal *versus* tumoral primario. Color verde indica que hay más genes sobreexpresados que subexpresados.

Los resultados observados en el estudio de las funciones moleculares se ven reflejados cuando estudiamos la localización celular de las proteínas producidas por los genes estudiados. En la **Figura 23**, vemos que sólo hemos detectado dos ubicaciones celulares y que en ambos casos los genes asociados a estas localizaciones están mayoritariamente sobreexpresados. Tenemos seis genes asociados a los cromosomas (lo que se deduce de las funciones moleculares asociadas a la unión al ADN) y en segundo lugar tenemos la localización en las uniones adherentes entre células.

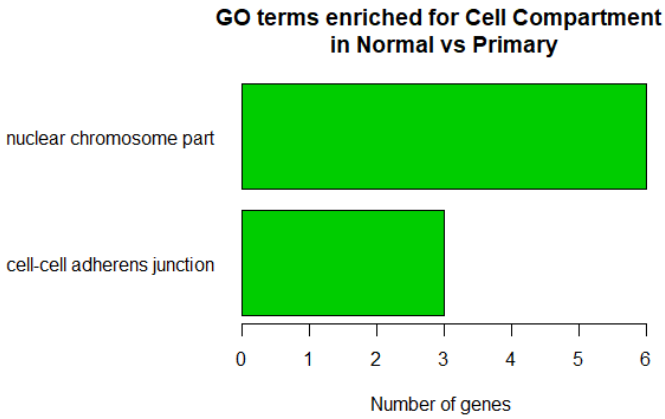


Figura 23. Estudio de enriquecimiento de términos GO asociados a compartimento celular en la comparación de tejido normal *versus* tumoral primario. Color verde indica que hay más genes sobreexpresados que subexpresados.

Comparación tejido Normal con tejido tumoral metastático

En este caso tenemos **568 genes** alterados al comparar el tejido normal con el tejido metastático. En la **Figura 24**, vemos las rutas biológicas alteradas en el tejido metastático en comparación con el tejido normal.

En esta comparación vemos que existen rutas cuyos genes asociados están mayoritariamente sobreexpresados (verde), mientras que hay otras rutas que se encuentran mayoritariamente subexpresadas (rojo) y en tercer lugar tenemos rutas en las que existe un equilibrio entre genes sobre y subexpresados.

Tenemos dos rutas mayoritariamente enriquecidas: la angiogénesis y la *regulación negativa de la modificación de proteínas* y ambas se encuentran sobreexpresadas.

La *angiogénesis*, es un proceso vital para la supervivencia de los tumores, ya que se estimula la creación de nuevos vasos sanguíneos para que el tumor pueda recibir suficiente sangre (y de esta forma recibir nutrientes) para mantener su crecimiento. En línea con la sobreexpresión de la ruta que promueve la *angiogénesis*, vemos que tenemos una subexpresión de los genes asociadas a la ruta *negative regulation of angiogénesis*. También tenemos la ruta *regulation of angiogenesis*, que en nuestro caso no parece estar sobre o subexpresada. Con esta combinación de alteraciones, el tumor metastático es capaz de inducir la generación de nuevos vasos sanguíneos, lo que permitirá el crecimiento de la metástasis.

La tercera ruta más enriquecida en esta comparación es la respuesta a hormonas esteroideas que también la encontramos sobreexpresada. De la biología del cáncer de próstata sabemos que una piedra angular en el desarrollo y supervivencia de los tumores es el AR, por lo que esperamos ver algún tipo de alteración en estos genes implicados en la ruta de respuesta a hormonas esteroideas.

Relacionada con la ruta anterior, tenemos la ruta llamada *celular response to steroid hormone stimulus*. Dicha ruta se encuentra sobreexpresada y en conjunción con la ruta anterior podría sostener el crecimiento y supervivencia tumorales.

También destaca el enriquecimiento de las rutas asociadas a ERK1/2, una ruta de señalización asociada a la estimulación de la supervivencia y proliferación celulares [27]. Tenemos que la ruta de la cascada ERK1/2 está subexpresada, mientras que la ruta que regula esta cascada no tiene una alteración clara. De forma similar a lo que ocurría con el control del ciclo celular en la comparación anterior, en esta ruta podríamos estar

subexpresando los controladores negativos de la ruta, mientras sobreexpresamos los reguladores positivos. Podríamos hipotetizar que esta combinación de alteraciones provoca una mayor actividad de ERK1/2 de forma no controlada, lo que estimularía la supervivencia del tumor.

Una ruta interesante es la llamada *positive regulation of cell adhesion*. Esta ruta es la cuarta más enriquecida y no tiene una alteración clara. Como sabemos, las células metastáticas tienen que moverse desde el tumor primario hasta un nuevo nicho. Para realizar este proceso deben modificar sus interacciones con las células que la rodean. Para ello deberán desregular las proteínas encargadas de formar estas uniones. Por éste motivo no vemos una alteración clara de la ruta, ya que habrá partes que la célula sobreexpresará y otras que subexpresará.

De las **83 rutas enriquecidas** que hemos encontrado, a parte de las mencionadas anteriormente, encontramos procesos metabólicos, de control de la diferenciación, procesos relacionados con el control de la apoptosis y morfogénesis entre otras. Pero un gran número de ellas tienen asociados un reducido número de genes.

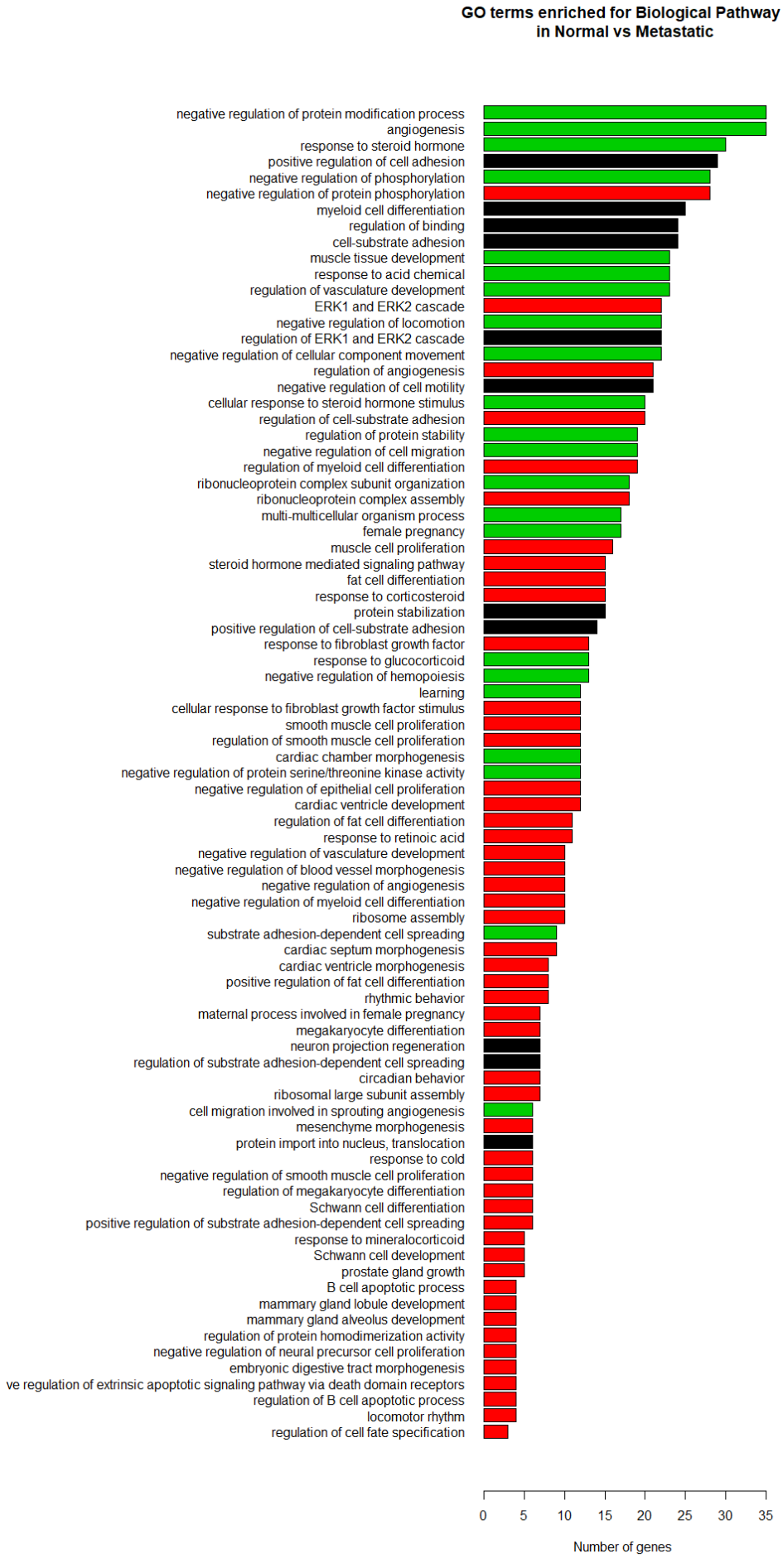

GO terms enriched for Biological Pathway
in Normal vs Metastatic
negative regulation of protein modification process
angiogenesis
response to steroid hormone
positive regulation of cell adhesion
negative regulation of phosphorylation
negative regulation of protein phosphorylation
myeloid cell differentiation
regulation of binding
cell-substrate adhesion
muscle tissue development
response to acid chemical
regulation of vasculature development
ERK1 and ERK2 cascade
negative regulation of locomotion
regulation of ERK1 and ERK2 cascade
negative regulation of cellular component movement
regulation of angiogenesis
negative regulation of cell motility
cellular response to steroid hormone stimulus
regulation of cell-substrate adhesion
regulation of protein stability
negative regulation of cell migration
regulation of myeloid cell differentiation
ribonucleoprotein complex subunit organization
ribonucleoprotein complex assembly
multi-multicellular organism process
female pregnancy
muscle cell proliferation
steroid hormone mediated signaling pathway
fat cell differentiation
response to corticosteroid
protein stabilization
positive regulation of cell-substrate adhesion
response to fibroblast growth factor
response to glucocorticoid
negative regulation of hemopoiesis
learning
cellular response to fibroblast growth factor stimulus
smooth muscle cell proliferation
regulation of smooth muscle cell proliferation
cardiac chamber morphogenesis
negative regulation of protein serine/threonine kinase activity
negative regulation of epithelial cell proliferation
cardiac ventricle development
regulation of fat cell differentiation
response to retinoic acid
negative regulation of vasculature development
negative regulation of blood vessel morphogenesis
negative regulation of angiogenesis
negative regulation of myeloid cell differentiation
ribosome assembly
substrate adhesion-dependent cell spreading
cardiac septum morphogenesis
cardiac ventricle morphogenesis
positive regulation of fat cell differentiation
rhythmic behavior
maternal process involved in female pregnancy
megakaryocyte differentiation
neuron projection regeneration
regulation of substrate adhesion-dependent cell spreading
circadian behavior
ribosomal large subunit assembly
cell migration involved in sprouting angiogenesis
mesenchyme morphogenesis
protein import into nucleus, translocation
response to cold
negative regulation of smooth muscle cell proliferation
regulation of megakaryocyte differentiation
Schwann cell differentiation
positive regulation of substrate adhesion-dependent cell spreading
response to mineralocorticoid
Schwann cell development
prostate gland growth
B cell apoptotic process
mammary gland lobule development
mammary gland alveolus development
regulation of protein homodimerization activity
negative regulation of neural precursor cell proliferation
embryonic digestive tract morphogenesis
ve regulation of extrinsic apoptotic signaling pathway via death domain receptors
regulation of B cell apoptotic process
locomotor rhythm
regulation of cell fate specification
0
5
10
15
20
25
30
35
Number of genes

Figura 24. Estudio de enriquecimiento de términos GO asociados a ruta biológica en la comparación de tejido normal *versus* tumoral metastático. Color verde indica que hay más genes sobreexpresados que subexpresados. Color negro significa que hay igualdad en el número de genes sobre/subexpresados. Color rojo indica que hay más genes subexpresados que genes sobreexpresados.

En la **Figura 25**, vemos que la principal función que se encuentra enriquecida es la *unión al promotor de la RNApolII*. Esta función está asociada a la transcripción de los genes que se traducirán en proteínas. Pero vemos que los genes asociados a dicha ruta están subexpresados.

También vemos múltiples rutas asociadas a la señalización celular (*protein tyrosine/serine/threonine phosphatase*, *SMAD binding* y *R-SMAD binding* (todas sobreexpresadas), *receptor signaling complex* (subexpresada)). Por último (aunque supone la segunda ruta más enriquecida) tenemos la unión al *rRNA* que se encuentra subexpresada. Éste ácido ribonucleico está asociado a los ribosomas y por lo tanto la unión de estos genes al rRNA podría modular la traducción de genes a nivel de ribosoma.

El hecho de que las rutas sobreexpresadas en el tumor metastático sean rutas asociadas a la señalización indica la importancia de las alteraciones a nivel de señalización para el desarrollo tumoral

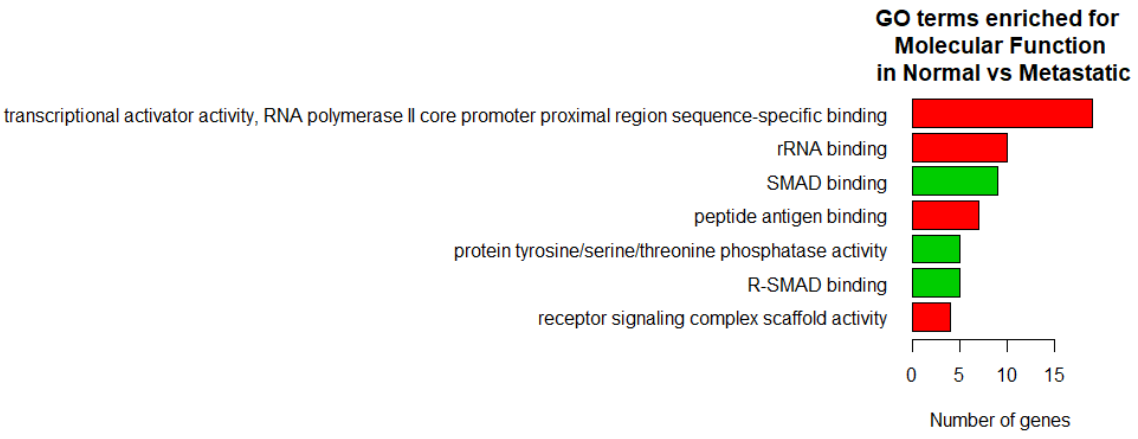


Figura 25. Estudio de enriquecimiento de términos GO asociados a función molecular en la comparación de tejido normal *versus* tumoral metastático. Color verde indica que hay más genes sobreexpresados que subexpresados. Color rojo indica que hay más genes subexpresados que genes sobreexpresados.

En la **Figura 26**, volvemos a ver la importancia de las alteraciones en las uniones celulares. Hecho que se refleja en qué términos asociados con tipos de unión célula-célula son los más enriquecidos (los tres primeros). Vemos que tenemos una subexpresión de todos ellos, indicando que la célula metastática reducirá o alterará el número y tipo de éstas interacciones en comparación con una célula normal.

También, vemos que tenemos una subexpresión de los genes asociados con la localización en el retículo endoplasmático y en el aparato de Golgi, orgánulos celulares asociados a la síntesis de proteínas. Aunque no sea de las más enriquecidas, nos llama la atención el compartimento celular denominado *cytoplasmic stress granule*. Estas estructuras citoplasmáticas están formadas por proteínas y ARN mensajero y se forman en respuesta a diversos tipos de estrés celular [28]. La subexpresión de ésta ruta de los gránulos de estrés nos indica que el tumor debe alterar su funcionamiento, con el objetivo de poder sobrevivir.

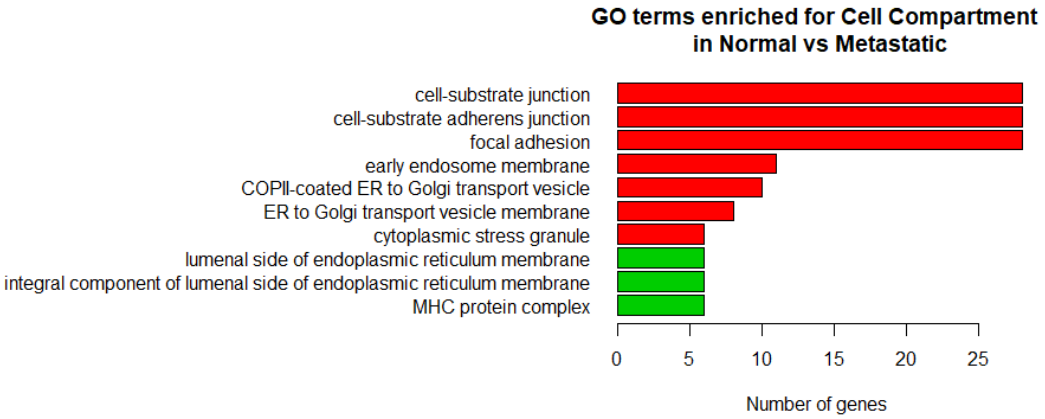


Figura 26. Estudio de enriquecimiento de términos GO asociados a compartimento celular en la comparación de tejido normal *versus* tumoral metastático. Color verde indica que hay más genes sobreexpresados que subexpresados. Color rojo indica que hay más genes subexpresados que genes sobreexpresados.

Comparación tejido tumoral primario con tejido tumoral metastático

En este caso tenemos **3201 genes** diferencialmente expresados entre el tumor primario y el metastático. En este caso, la mayoría de las rutas están sobreexpresadas, pero en el caso de las rutas con mayor número de genes asociadas tenemos múltiples casos de rutas subexpresadas.

Dentro de estas rutas subexpresadas encontramos la ruta *negative regulation of cell migration* y rutas de respuesta al TGF-β.

Si nos centramos en que ruta biológica es la más enriquecida en este grupo de genes, vemos que dicha ruta es la denominada *cell-substrate adhesión* (**Figura 27**). Lo que indica que la capacidad de la célula de unirse al sustrato (entendiendo en este contexto sustrato como la matriz extracelular) es de vital importancia para la progresión de tumor primario a metastático. Esto es esperable en un tumor metastático, ya que para que pueda existir dicha metástasis la célula ha de ser capaz de abandonar el tumor primario, pasar al torrente sanguíneo, salir de dicho torrente y establecer un nuevo nicho en otro tejido diferente al prostático donde deberá sobrevivir y proliferar.

Observamos multitud de rutas asociadas a la morfogénesis de varios tejidos como la glándula prostática (sobreexpresadas), mamaria (subexpresadas), diversos tejidos del corazón, diferenciación de osteoblastos…Estas rutas alteradas podrían explicar en parte el tropismo de las células metastáticas (recordemos que las muestras de metástasis analizadas provienen de diversos tejidos fuera de la próstata).

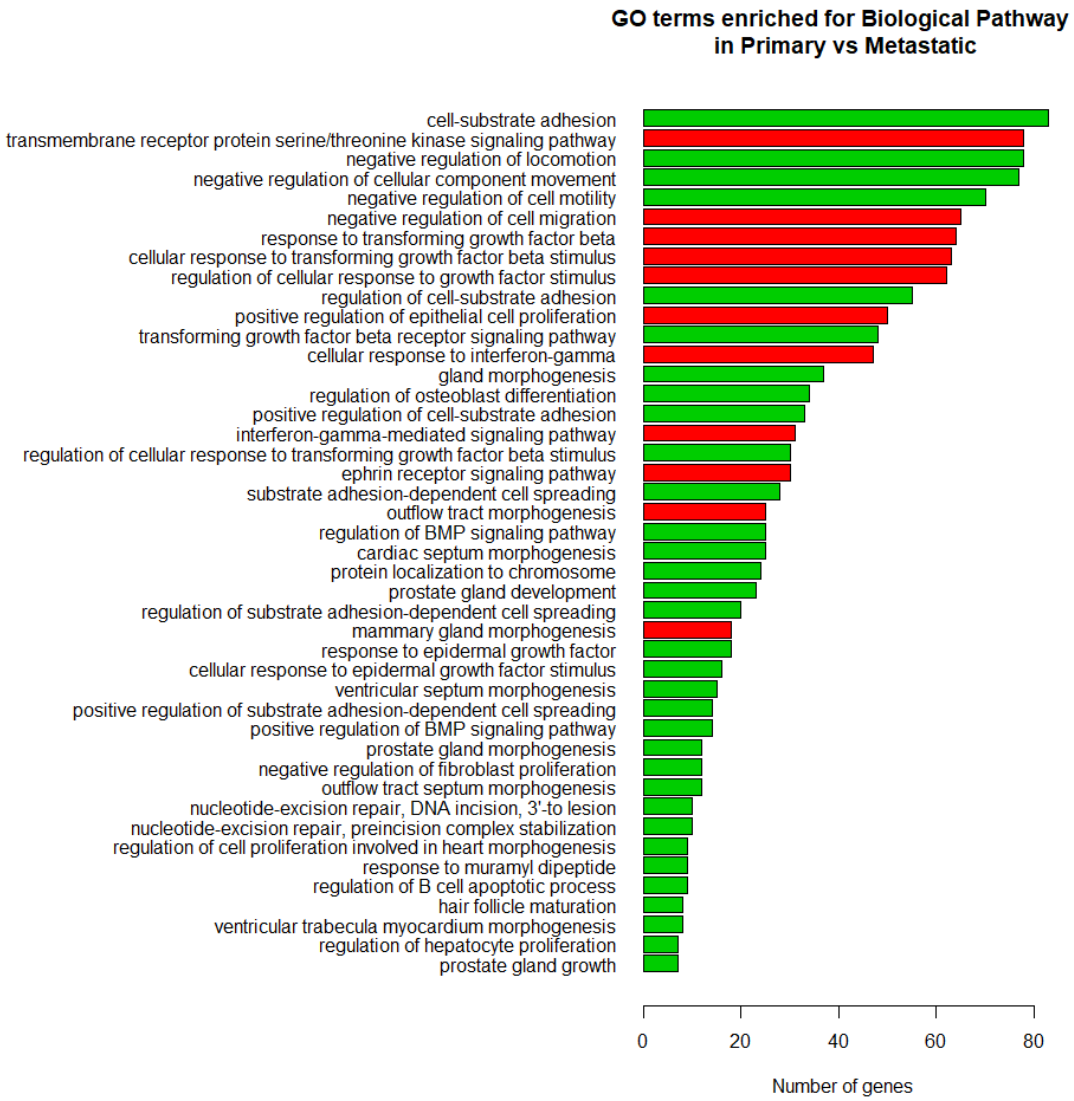


Figura 27. Estudio de enriquecimiento de términos GO asociados a ruta biológica en la comparación de tejido tumoral primario *versus* tumoral metastático. Color verde indica que hay más genes sobreexpresados que subexpresados. Color rojo indica que hay más genes subexpresados que genes sobreexpresados.

Si nos centramos en las funciones moleculares (**Figura 28**), vemos que mayoritariamente se ve alterada la función de la actividad transcripcional mediada por co-represores, seguida de lejos por la unión de represores de los factores de transcripción de la RNApolII, ambas rutas sobreexpresadas en el tejido metastático.

También tenemos alteraciones en ciertas rutas de señalización (*R-SMAD binding*) que se encuentra sobreexpresada y en la unión de péptido antigénicos (¿posible mecanismo de evasión inmunológica del tumor?) que se encuentra subexpresada.

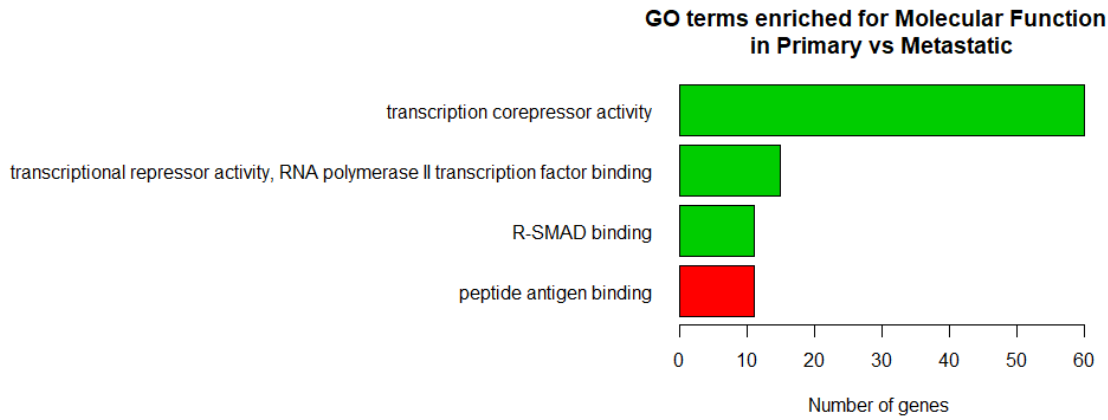


Figura 28. Estudio de enriquecimiento de términos GO asociados a función molecular en la comparación de tejido tumoral primario *versus* tumoral metastático. Color verde indica que hay más genes sobreexpresados que subexpresados. Color rojo indica que hay más genes subexpresados que genes sobreexpresados.

Respecto al estudio de los compartimentos celulares (**Figura 29**), todos los compartimentos donde vemos enriquecimiento están sobreexpresados. Vemos un mayor enriquecimiento en genes asociados con las caveolas. Las caveolas son unas invaginaciones de la membrana celular asociadas con la endocitosis y la transducción de señales [29]. De forma similar al apartado anterior, volvemos a ver un enriquecimiento en genes asociados con los gránulos de estrés citoplasmáticos, de nuevo indicando la existencia de un estrés en la célula.

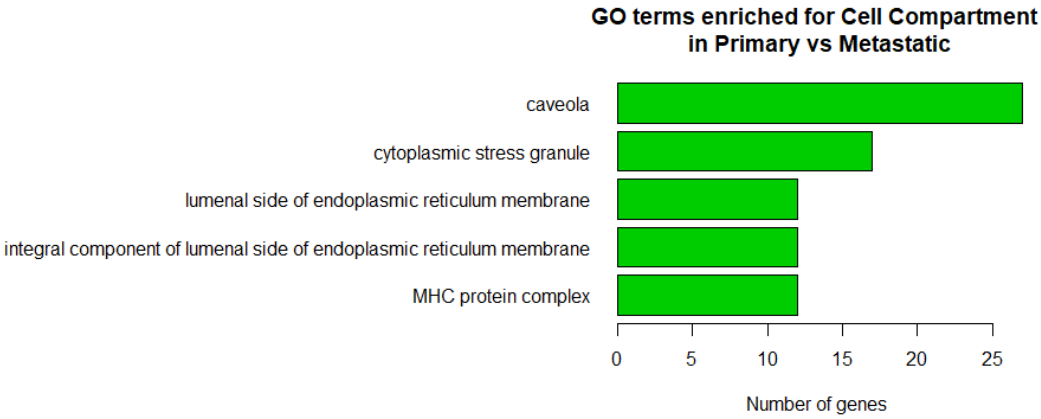


Figura 29. Estudio de enriquecimiento de términos GO asociados a compartimento celular en la comparación de tejido tumoral primario *versus* tumoral metastático. Color verde indica que hay más genes sobreexpresados que subexpresados.

De nuevo aparecen asociaciones con el retículo endoplasmático, indicando así alteraciones en la síntesis de proteínas, así como en el tráfico de proteínas entre el exterior-interior celular.

Regiones diferencialmente metiladas

De los resultados descritos previamente, podemos deducir que muchas de las alteraciones en los niveles de expresión de los genes identificados procederán de actividades aberrantes en los mecanismos de transcripción (activación de *enhancers*, silenciamiento de represores o viceversa), seguramente debido a activaciones inapropiadas de rutas que promuevan la proliferación y/o supervivencia celular. Pero una alternativa a este mecanismo de acción serían las alteraciones en los patrones de metilación de los promotores de los genes.

Después de aplicar la función *bumphunter* a los datos normalizados de estudios de metilación hemos encontrado, usando un *cutoff* de 0.4 y un valor límite de *fwer* de 0.01, que entre tejido prostático normal y tumoral existen **99 DMRs**.

El tamaño de la DMR se calcula restando a la posición final de la región la posición inicial y obtenemos que:

-El tamaño medio de es de **8 pares de bases**.

-El tamaño máximo detectado es de **338 nucleótidos**.

-La mediana es de **1 nucleótido** (indicando metilaciones de una única base).

Si hacemos un recuento de los tamaños (**Figura 30**) vemos que de las 99 DMRs, 90 son metilaciones de una única base, el resto son casos únicos de diferentes tamaños de DMR.

```
> table(sigDMRs$size)

  0   5   6   8  10  29  33 113 186 337
 90   1   1   1   1   1   1   1   1   1
```

Figura 30. Distribución de los tamaños de las regiones diferencialmente metiladas detectadas con la función *bumphunter*. Se observa que la mayoría de las regiones están formadas por un único nucleótido diferencialmente metilado.

Si nos fijamos en la posición de éstas DMRs (**Figura 31**) vemos que el cromosoma 14 es el que contiene menos DMRs mientras que el 10 y el 12 son los que tienen más.

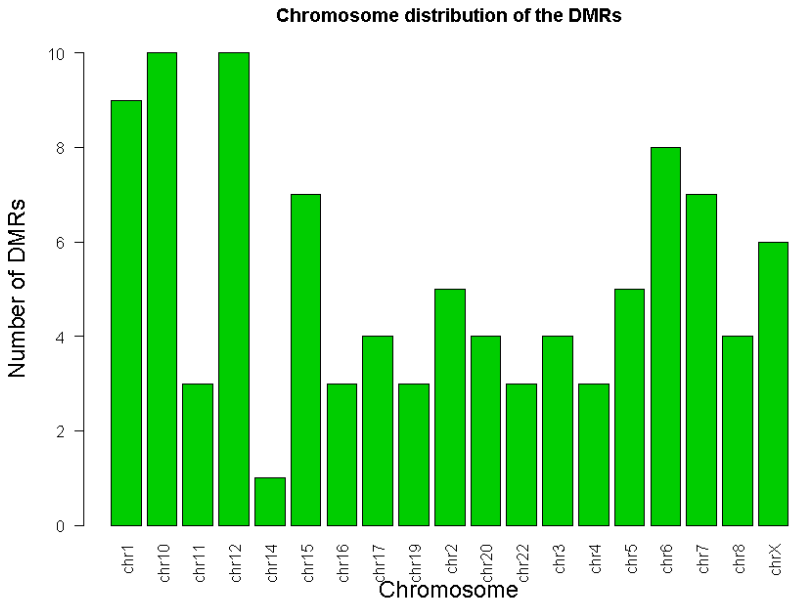


Figura 31. Figura donde se observa en que cromosomas encontramos las DMRs detectadas y cuantas tenemos por cromosoma.

Si nos centramos en las DMRs cuyo tamaño es superior a un par de bases y analizamos en que cromosomas se encuentran, vemos que estas DMRs están asociadas a seis de los cromosomas humanos: 1, 6, 7, 15, 20 y el X (**Figura 31**).

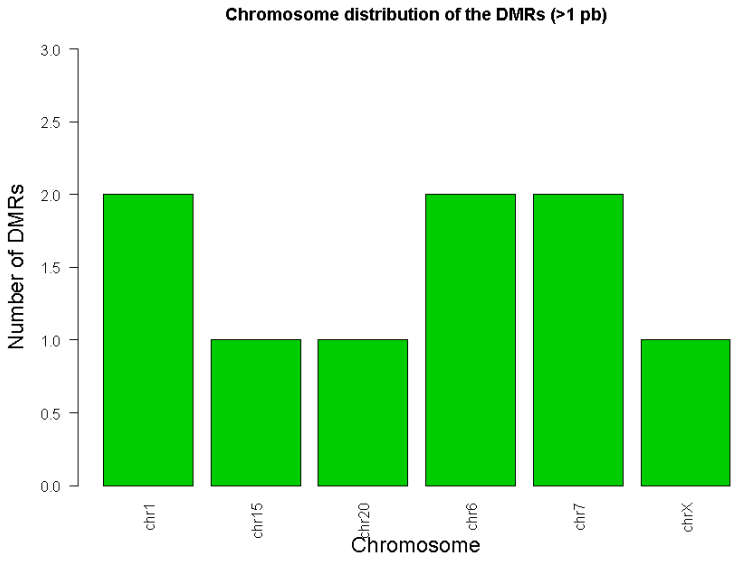


Figura 32. Localización de las DMRs con un tamaño superior al nucleótido.

Si nos centramos en las diferencias en metilación observadas entre el tejido normal y el tumoral (**Figura 33**) vemos que independientemente del tamaño de la DMR

analizada, vemos que estas regiones están más metiladas en el tumor que en el tejido normal, ya que los valores de la columna *value* producida por *bumphunter* son positivos.

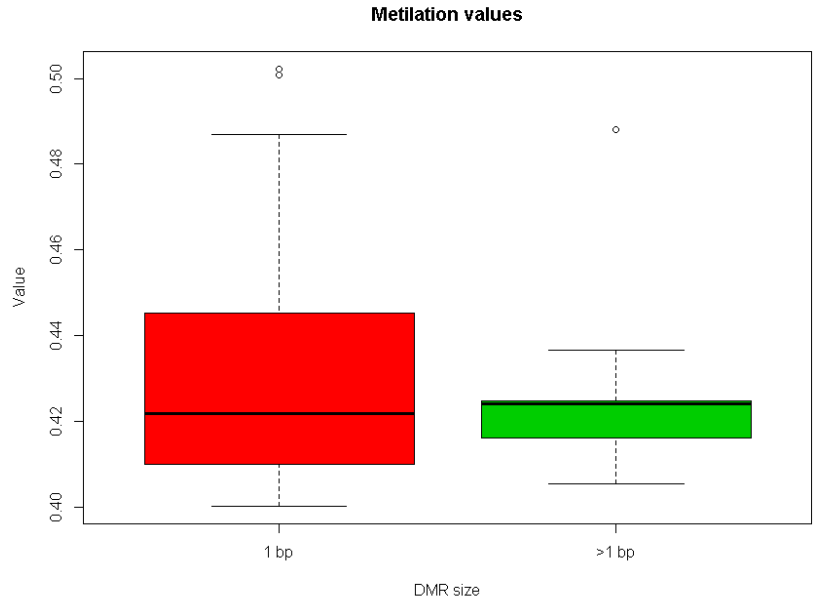


Figura 33. Las regiones diferencialmente metiladas están hipermetiladas en el tejido tumoral, independientemente del tamaño de la DMR ya que los valores detectados son todos positivos.

Después de anotar los genes asociados a las DMRs detectadas obtenemos que:

-Existen **132 genes** asociados a las DMRs formadas por una única base.

-Existen **9 genes** asociados a las DMRs más largas.

-Sólo existe un caso (el del gen ***CCDC181***) que esté asociado tanto a una DMR de 1 bp como a una de más envergadura.

Curiosamente, el gen ***CCDC181*** se ha visto asociado al riesgo de recurrencia del cáncer de próstata después de una prostatectomía radical, asociado a una hipermetilación en este gen (entre otros) [30]. Curiosamente, el gen ***KLF8***, que hemos detectado asociado una DMR de 1 bp también aparece hipermetilado y asociado a un mayor riesgo de recurrencia en el mismo artículo donde se menciona a ***CCDC181***.

Ahora compararemos el listado de genes asociado a estas DMRs con el listado de genes diferencialmente expresados que hemos encontrado en el apartado anterior:

-*Comparación tejido Normal-Primario:* No se han encontrado genes diferencialmente expresados que estén asociados a ninguna DMR detectada.

-*Comparación tejido Normal-Metastático:* Sólo se ha detectado un gen (***GALNT10***) que se encuentra sobreexpresado en tejido metastático asociado a una DMR de 1 bp. Este gen codifica para un miembro de las *GalNAc polypeptide N-acetylgalactosaminyltransferasas*, que se ha observado recientemente que está asociada con la formación de metástasis [31] y con una peor supervivencia y recurrencia en algunos tipos de cáncer [32].

-*Comparación tejido Primario-Metastático:* En esta ocasión hemos encontrado **6 genes** diferencialmente expresados asociados a DMRs de 1 bp. Y **2 genes** asociados a DMRs de tamaño superior a 1 bp.

De los seis genes asociados a las DMR de 1bp tenemos el caso de los genes ***C16orf72***, ***GALNT10***, ***PIK3CB*** y ***SLC25A37***, que se encuentran subexpresados. La metilación encontrada en el tejido tumoral metastático podría ser la explicación a esta reducción en la expresión.

Destaca el gen ***GALNT10***, detectado también en la comparación Normal-Metastático, pero en esa ocasión el dicho gen está sobreexpresado, justo lo contrario que en la comparación Primario-Metastático.

En el bando opuesto tenemos los genes ***ANK3*** y ***HIVEP3*** que se encuentran sobreexpresados en tejido metastático.

En el caso de los genes asociados a DMRs de tamaño superior a 1 bp tenemos los genes ***GRM1*** (subexpresado) y ***SLC30A4*** (sobreexpresado).

C16orf72: Sin información en *pubmed* de la posible función de este gen.

PIK3CB: Gen que codifica para una quinasa del tipo *phosphoinositide 3-kinase*. Se han detectado altos niveles de expresión en algunos tipos de tumores [33].

SLC25A37: Gen que codifica para un transportador localizado en la membrana mitocondrial interna. Requerido también para la homeostasis del hierro. No existen *links* con cáncer en *pubmed*.

ANK3: Codifica para un miembro de la familia de las ankirinas, proteínas involucradas en el citoesqueleto celular. Juega un papel importante en la movilidad

celular, la proliferación, el contacto célula-célula…Se ha visto sobreexpresado en tumores de próstata localizados [34].

HIVEP3: Codifica para un factor de transcripción que es capaz de regular la actividad transcripcional de NF-kappa-β. Se sabe que acentúa la actividad de la ruta de señalización iniciada por TGF-β [35], siendo de sobra conocido el papel del TGF-β en la proliferación de células tumorales [36].

GRM1: Receptor de glutamato expresado principalmente en las neuronas. La expresión de este receptor en células mamarias se ha visto que origina una transformación maligna de las células [37].

SLC30A4: Codifica para un transportador de zinc. Está reportada como subexpresada en tumores de próstata en comparación con el tejido normal [38].

Hemos observado que sí que existen genes cuya expresión es diferencial entre los tres tejidos analizados que encontramos asociados a regiones que están diferencialmente metiladas. Existen varios ejemplos de genes (por ejemplo, ***C16orf72***, ***GALNT10***, ***PIK3CB*** y ***SLC25A37***) cuyos resultados de expresión y metilación van en la misma dirección, es decir, tenemos una subexpresión y una hipermetilación de la región, lo que cuadra con los efectos reportados de la metilación.

Por otro lado, tenemos otros genes cuyos resultados son contradictorios, se sobreexpresan en una región hipermetilada. Esta discrepancia podría deberse al hecho de que las muestras donde se han analizado las metilaciones no son las mismas que en las que se ha analizado la expresión.

Análisis de la Variación en el número de copias (CNV)

Por último, analizamos si existen amplificaciones o deleciones de los genes diferencialmente expresados que pudieran explicar los cambios observados, más allá de los cambios que pueda haber en dichas regiones.

En la **Figura 34** tenemos un resumen general de la frecuencia de eventos de deleción o de amplificación genética en los tres grupos de genes diferencialmente expresados detectados (*NormalvsPrimario*, *NormalvsMetástasis* y *PrimariovsMetástasis*).

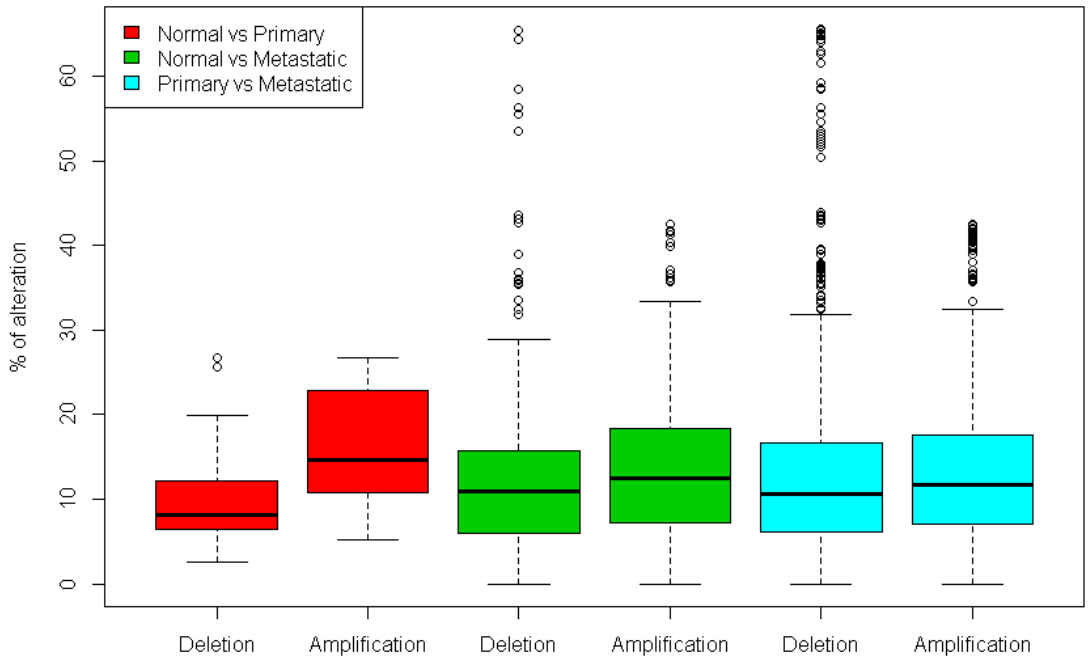


Figura 34. Frecuencias de deleciones o amplificaciones en los genes diferencialmente expresados en las tres comparaciones realizadas en el primer punto de este trabajo.

Un análisis estadístico mediante una *t-student* indica que no hay diferencias significativas ($p>0.05$) con la excepción de la comparación entre las amplificaciones en los genes identificados en el grupo *NormalvsPrimario* en comparación con el grupo *PrimariovsMetastático*. Del gráfico vemos que en el primer grupo tenemos unos niveles ligeramente superiores en amplificaciones con un valor de ***p=0.03817***.

Aunque en general no podemos decir que hay diferencias entre el comportamiento de los genes de los tres grupos analizados, llama la atención que en los genes diferencialmente expresados en tejido metastático *versus* normal o primario tenemos conjuntos de genes con niveles altos (>30% de los casos) tanto de amplificaciones como de deleciones que no se observan en la comparación tumor primario con tejido normal.

Pero ¿existe una correlación entre la frecuencia del CNV y el cambio de expresión observado? Para ello graficaremos los valores del *logFC* en el análisis de expresión diferencial obtenido en el primer punto *versus* a la frecuencia del CNV que acabamos de calcular.

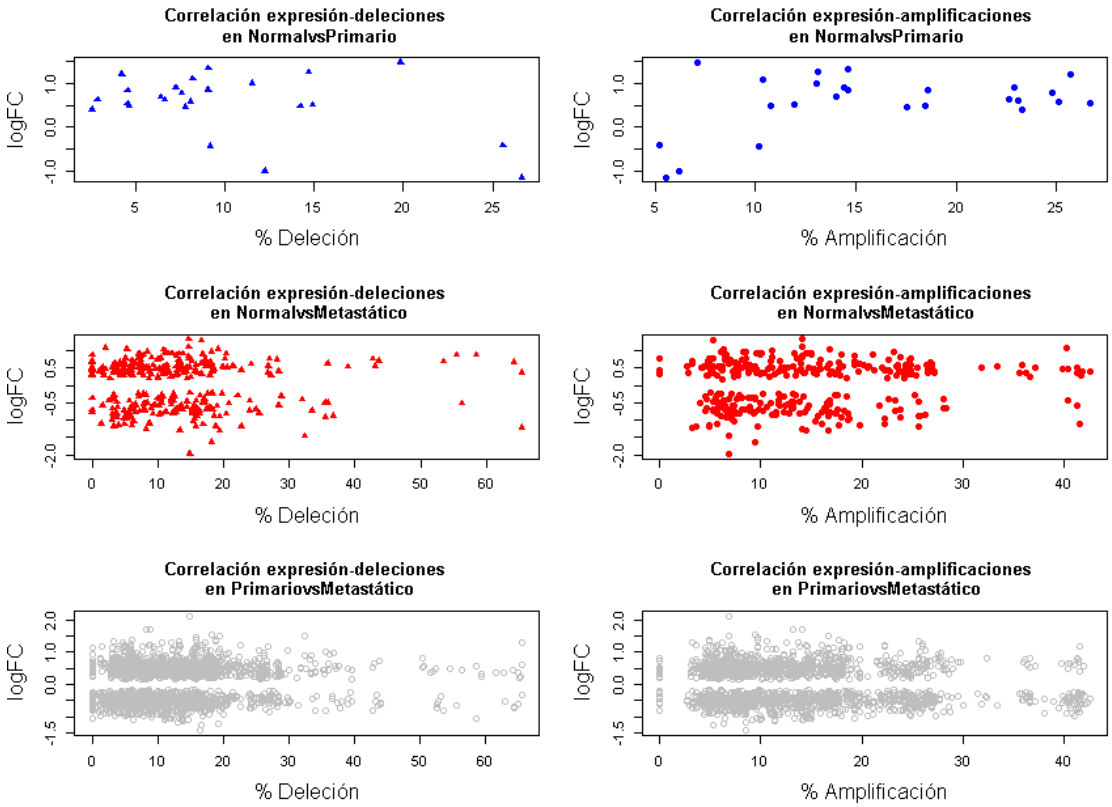


Figura 35. Correlaciones entre los porcentajes de deleción/amplificación y los cambios a nivel de expresión de los genes diferencialmente expresados.

En la **Figura 35** se muestran los valores del *logFC* de los genes diferencialmente expresados frente a la frecuencia de deleciones o amplificaciones observadas en los estudios de CNV.

A primera vista no se observa ningún tipo de relación entre mayores niveles de deleción y una mayor reducción en la expresión. Tampoco observamos una relación entre más frecuencia de amplificaciones y una mayor sobreexpresión. Esto nos indica que las amplificaciones/deleciones no son un mecanismo *universal* para alterar la expresión génica durante la progresión del cáncer de próstata.

No sólo no parece existir correlación, si no que en genes que se encuentran altamente delecionados en los estudios de CNV se observan sobreexpresados.

La misma falta de correlación se observa si analizamos por separado los genes sobreexpresados y los que se encuentran subexpresados (**Figura 36**).

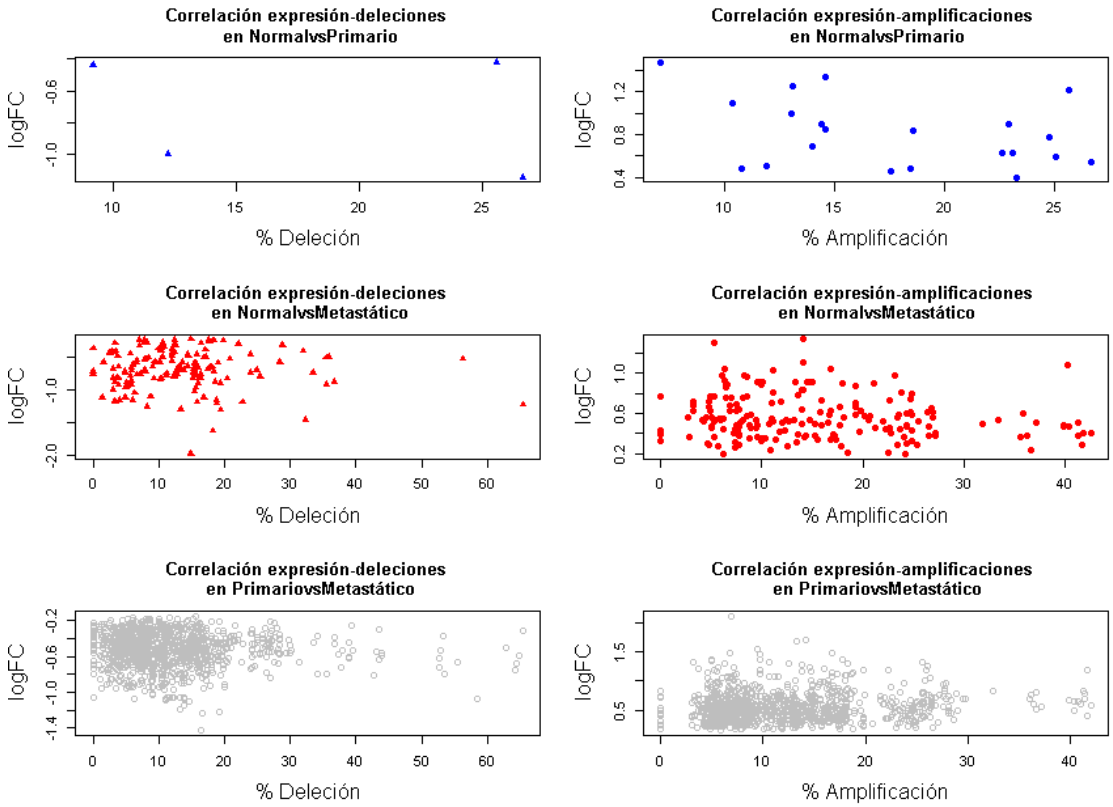


Figura 36. Correlaciones entre los porcentajes de deleción/amplificación y los cambios a nivel de expresión de los genes diferencialmente expresados. En las gráficas de la izquierda se representan solo los genes subexpresados en función de sus porcentajes de deleción, mientras que a la derecha se representan los genes sobreexpresados.

El hecho de que existan genes que están amplificados y aun así vean reducida su expresión (y *viceversa*), nos indica que hay otros factores en juego que determinan los cambios de expresión sufridos por el gen. El gen amplificado podría sufrir un proceso de silenciamiento genético mediante metilaciones de sus regiones promotoras, por ejemplo. Y al contrario, un gen parcialmente delecionado, podría verse desmetilado para incrementar sus niveles de expresión.

A nivel de comprobación, resulta interesante graficar las proporciones de deleciones de los genes en comparación con los niveles de amplificaciones detectados. En la **Figura 37** se muestran dichos gráficos. En general se observa que la mayoría de los genes pueden presentar tanto deleciones o amplificaciones (esto se debe a la heterogeneidad de los pacientes, ya que algunos pueden presentar amplificaciones y otras deleciones).

Llama la atención (sobre todo en la comparación *NormalvsMetastático* y en *PrimriovsMetastático*) que existen casos de genes que presentan valores variables de amplificaciones, pero no existen deleciones, y *viceversa*.

Los genes que nunca se encuentran delecionados podrían ser los llamados genes *house-keeping* cuyas mutaciones serían letales para la célula o genes que son vitales para la supervivencia celular tumoral (genes esenciales en este contexto particular pero que no son genes *hous-keeping*). El hecho de que este grupo de genes sólo se encuentre amplificado podría indicar que son genes que estimularían la proliferación/supervivencias celulares, serían genes candidatos a ser *protooncogenes*.

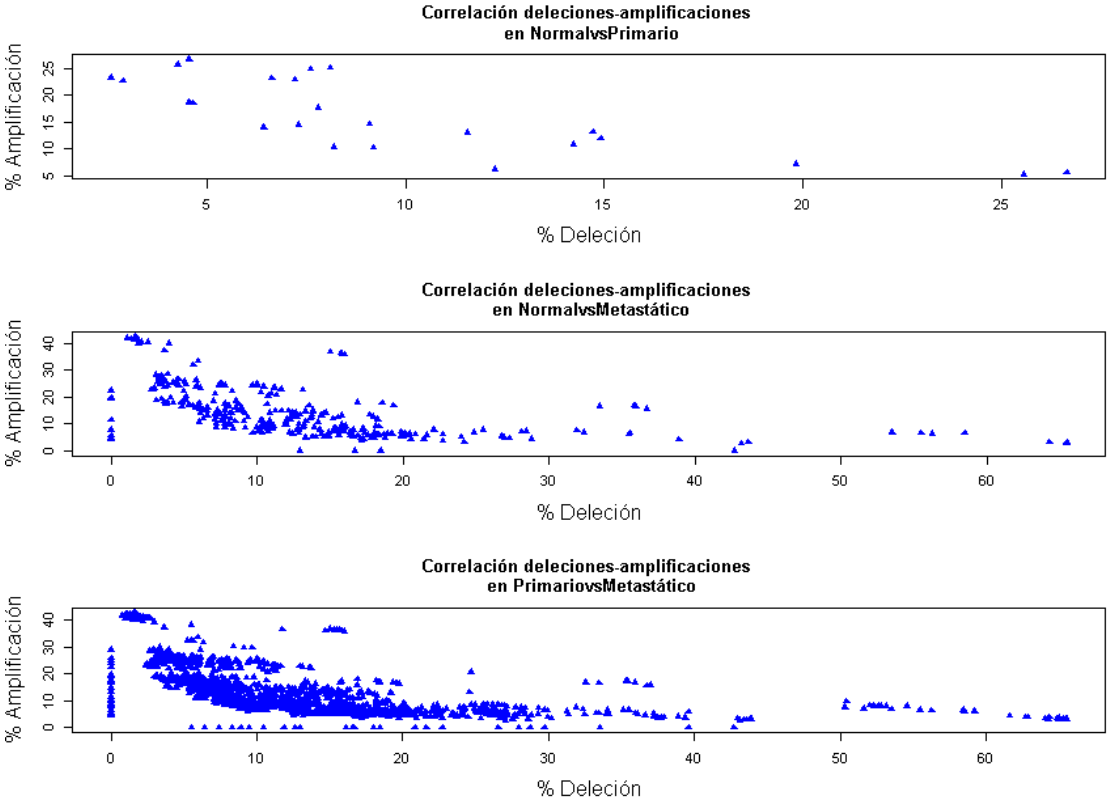


Figura 37. Correlación entre las deleciones y amplificaciones de los genes analizados.

En cambio, los genes que solo se encuentran delecionados podrían ser genes supresores de tumores que aportarían alguna ventaja selectiva a la célula con dichas deleciones.

En la comparación entre tejido normal y tumoral primario no existen genes con este comportamiento (existen genes con valores bajos de amplificaciones pero que no

llegan a 0). Mientras que en las otras comparaciones sí que encontramos unos cuantos genes con este comportamiento:

-*Normal vs Metastático:* Existen **4 genes** donde no hay individuos con amplificaciones: ***CBL***, ***NF1***, ***PIK3R3*** y ***PTEN***. También existen **9 genes** donde no hay deleciones: ***AR***, ***AURKB***, ***BCL6***, ***CRKL***, ***FGFR2***, ***FLT1***, ***JUN***, ***RIT1*** y ***U2AF1***.

-*Primario vs Metastático:* Existen **26 genes** donde no hay individuos con amplificaciones: ***ARID1A, ARID1B, ARID2, AXIN2, CASP8, CBL, CDK12, CDKN1B, CHEK2, CTCF, ERCC3, ERCC4, FAT1, NF1, PIK3R1, PIK3R3, PTCH1, PTEN, RASA1, RNF43, SDHC, SMAD3, SMAD4, STK11, TCF7L2*** y ***TET2***. También existen **33 genes** donde no hay deleciones: ***AR, AURKA, AURKB, AXL, BCL6, CCND2, CRKL, DDR2, ERBB2, ESR1, ETV1, FGF19, FGFR2, FGFR3, FLT1, GNA11, GNAS, HRAS, IGF1, IGF1R, JUN, MAPK1, NOTCH1, PIK3CB, PRKCI, RIT1, ROS1, SF3B1, SMO, STAT3, TMPRSS2, U2AF1*** y ***YAP1***.

Un gen que se repite en ambas comparaciones y que en ambas nunca está delecionado es el gen ***AR***. Dicho gen codifica para el receptor de andrógenos que es una pieza clave para la supervivencia del tejido prostático normal y/tumoral, lo que es coherente con el hecho de no encontrarlo delecionado, pero si amplificado.

Otro caso similar, pero en el que sólo se observan deleciones, es el caso del gen ***PTEN***. Es un gen del que se han observado deleciones en aproximadamente el 18% de los casos (según el estudio) y cuta pérdida está asociada a un mayor riesgo de muerte [39].

En el caso del gen ***JUN*** y ***HRAS***, que se encuentran amplificados en ambas comparaciones, tenemos un claro ejemplo del tipo de genes que podríamos esperar en esta categoría, los *proto-oncogenes*. *JUN* es un factor de transcripción cuya sobreexpresión se ha visto asociada con múltiples tipos de tumores [40]. *HRAS* es una proteína miembro de la familia *RAS* de protooncogenes con la capacidad de unirse a GTP y GDP y que posee actividad GTPasa. Se ha observado que los niveles altos de esta proteína están asociados con mayor mortalidad en algunos tipos de tumores [41].

CHEK2, ***ERCC3*** y ***ERCC4*** son proteínas involucradas en procesos de reparación del ADN que vemos que nunca se encuentran amplificadas en la comparación *PrimariovsMetastático*. Tiene sentido que existan deleciones de dichos

genes, lo que generaría en la célula tumoral una inestabilidad genómica que podría suponer una ventaja para la misma.

También tenemos ejemplos de proteínas implicadas en rutas de señalización como ***SMAD3*** y ***SMAD4***, que nunca se observan amplificaciones y ***IGF1, IGF1R***, ***MAPK1, NOTCH1*** y ***STAT3*** que nunca se observan delecionados. Todo esto podría estar indicando el efecto negativo que la función normal de las primeras proteínas podría tener sobre la supervivencia/proliferación celulares (de ahí que sólo detectemos deleciones en los estudios de CNV). Mientras que en el caso del segundo grupo de proteínas implicadas en señalización podría estar asociado con esencialidad o al menos efectos positivos en las células tumorales, lo que explicaría que no viéramos deleciones, pero si amplificaciones.

4. Conclusiones

Seguidamente pasaremos a resumir las principales conclusiones obtenidas de los resultados de este trabajo:

1. El número de genes diferencialmente expresados se incrementa contra más avanzado se encuentra el tumor (52 *versus* 568 genes diferencialmente expresados en la comparación con el tejido normal del tumoral primario y metastático respectivamente)
2. El número de genes alterados a nivel de expresión entre el tejido tumoral primario y el metastático es muy grande (3201 genes). Este número sorprende por su magnitud ya que entre el tejido normal y metastático solo hemos detectado 568 genes con diferencias de expresión.
3. En la mayoría de los casos, los genes diferencialmente expresados están asociados con funciones de transcripción o de regulación de las interacciones intercelulares.
4. El número y tamaño de DMRs es relativamente bajo: 99 DMRs de las cuales 90 consisten en un único par de bases.
5. Todas las DMRs se encuentran hipermetiladas en el tejido tumoral respecto al tejido normal.
6. De los genes asociados a las DMRs halladas, sólo nueve de ellos nos aparecen como diferencialmente expresados.
7. De estos nueve genes, sólo cuatro tienen una reducción de expresión y están en regiones hipermetiladas.
8. La frecuencia de CNV de tipo deleción/amplificación no varía entre los genes analizados y NO se corresponde con los cambios de expresión observados.
9. Un único tipo de alteración (metilación o CNV) no parece ser suficiente para explicar las variaciones de expresión detectadas.

Respecto a las limitaciones de estos análisis. Hemos analizado tres parámetros diferentes (expresión, metilación y CNV) recurriendo a los datos de tres estudios diferentes de carácter público.

Esto hace que sea complicado extraer conclusiones definitivas de los mismos si intentamos justificar los cambios de expresión mediante los cambios de metilación y/o los cambios de CNV. Esto se debe a que las muestras analizadas en cada uno de los estudios son diferentes (tanto en número como en estado de la enfermedad). Esto introduce variables de confusión que complican el análisis y el llegar a unas conclusiones sólidas.

Si se pudiera seguir con este trabajo en el futuro, el camino ideal sería reclutar muestras de tejidos prostáticos normales y tumorales de diverso grado y en dichas muestras analizar la expresión, la metilación y la presencia o no de CNV. También pudiera ser interesante incluir estudios mutacionales. Para ello se deberían secuenciar todos los exones de las muestras, si encontramos deleciones/inserciones dentro de los genes o mutaciones puntuales que generasen un codón STOP serían una posible explicación para las reducciones de expresión.

5. Glosario

AR = *Androgen Receptor* (Receptor de andrógenos)

AF-1 y AF-2*:* Dominios de activación N-terminal y C-terminal respectivamente presentes en los receptores nucleares.

DBD: *DNA-binding domain.* Dominio de unión al ADN presente en los receptores nucleares.

LBD: *Ligand-binding domain.* Sitio de unión del ligando presente en los receptores nucleares.

DHT: 5α-dihydrotestosterona

ADT: *Androgen-Deprivation Therapy*. Terapia de deprivación de andrógenos.

CRPC: *Castration-resistant Prostate Cancer*. Cáncer de próstata resistente a la castración.

CNV: *Copy Number Variations*. Variación en el número de copias.

NGS: *Next Generation Sequencing*.

GO: *Gene Ontology.*

BP: *Biological Pathway*. Función biológica.

MF: *Molecular Function*. Función molecular.

CC: *Celular Component*. Componente o Compartimento celular.

DMR: Región Diferencialmente Metilada.

6. Bibliografía

[1] Duronio RJ, Xiong Y. *Signaling pathways that control cell proliferation.* Cold Spring Harb Perspect Biol. 2013 Mar 1;5(3):a008904. doi: 10.1101/cshperspect.a008904.

[2] Evans RM. *The nuclear receptor superfamily: a rosetta stone for physiology.* Mol Endocrinol. 2005 Jun;19(6):1429-38.

[3] Shi, Y. *Orphan Nuclear Receptors in Drug Discovery*. Drug Discov Today. 2007 Jun; 12(11-12): 440–445. doi: 10.1016/j.drudis.2007.04.006

[4] Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schutz G, Umesono K, et al. *The nuclear receptor superfamily: the second decade.* Cell 1995; 83: 835–9.

[5] Eileen Tan, MH. Jun Li, H, Xu, E. Melcher,K. & Yong,E. *Androgen receptor: structure, role in prostate cancer and drug Discovery*. Acta Pharmacologica Sinica volume 36, pages 3–23 (2015) doi:10.1038/aps.2014.18

[6] Brinkmann AO. *Molecular basis of androgen insensitivity. Mol Cell Endocrinol. 2001 Jun 20;179(1-2):105-9.*

[7] REDECAN. *Cancer Incidence in Spain 2015*. Clin Transl Oncol. DOI 10.1007/s12094-016-1607-9)

[8] Heinlein CA, Chang C. *Androgen receptor in prostate cancer*. Endocr Rev 2004; 25: 276–308.

[9] Bubley GJ, Carducci M, Dahut W, Dawson N, Daliani D, Eisenberger M, Figg WD, Freidlin B, Halabi S, Hudes G, Hussain M, Kaplan R, Myers C, Oh W, Petrylak DP, Reed E, Roth B, Sartor O, Scher H, Simons J, Sinibaldi V, Small EJ, Smith MR, Trump DL, Wilding G. *Eligibility and response guidelines for phase II clinical trials in androgen-independent prostate cancer: recommendations from the Prostate-Specific Antigen Working Group.* J Clin Oncol. 1999;17(11):3461–3467.

[10] Jones CU, Hunt D, McGowan DG, Amin MB, Chetner MP, Bruner DW, Leibenhaut MH, Husain SM, Rotman M, Souhami L, Sandler HM, Shipley WU. *Radiotherapy and short-term androgen deprivation for localized prostate cancer.* N Engl J Med. 2011 Jul 14; 365(2):107-18.

[11] Heidenreich, A., Bastian, P.J., Bellmunt, J. et al. *EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer*. Eur Urol. 2014; 65: 467–479

[12] Karantanos, T., Evans, C.P., Tombal, B., Thompson, T.C., Montironi, R., and Isaacs, W.B. *Understanding the mechanisms of androgen deprivation resistance in prostate cancer at the molecular level*. Eur Urol. 2015; 67: 470–479

[13] Gravis, G., Boher, J.M., Joly, F. et al. *Androgen deprivation therapy (ADT) plus docetaxel versus ADT alone in metastatic non castrate prostate cancer: impact of metastatic burden and long-term survival analysis of the randomized phase 3 GETUG-AFU15 trial*. Eur Urol. 2016; 70: 256–262

[14] Even-Sapir E, Metser U, Mishani E, Lievshitz G, Lerman H, Leibovitch I. *The detection of bone metastases in patients with high-risk prostate cancer: 99mTc-MDP Planar bone scintigraphy, single- and multi-field-of-view SPECT, 18F-fluoride PET, and 18F-fluoride PET/CT.* J Nucl Med. 2006 Feb;47(2):287-97.

[15] Teply, B.A. and Antonarakis, E.S. *Treatment strategies for DNA repair-deficient prostate cancer*. Expert Rev Clin Pharmacol. 2017; 10: 889–898

[16] Mateo, J., Carreira, S., Sandhu, S. et al. *DNA-repair defects and olaparib in metastatic prostate cancer*. N Engl J Med. 2015; 373: 1697–1708

[17] Morris LG, Chan TA. *Therapeutic targeting of tumor suppressor genes*. Cancer. 2015 May 1;121(9):1357-68. doi: 10.1002/cncr.29140.

[18] Kurzrock R, Kantarjian HM, Druker BJ, Talpaz M. *Phildelphia Chromosome-positve leukemias: from basic mechanisms to molecular therapeutics*. Ann Intern Med 2003;138:819–30.

[19] Leal Rojas P1, Anabalón Rodríguez L, García Muñoz P, Tapia Escalona O, Guzmán González P, Araya Orostica JC, Villaseca Hernández M, Roa Strauch JC. *[Promoter hypermethylation gene patterns in gynecological tumors]*. *Med Clin (Barc). 2009 Mar 21;132(10):371-6. doi: 10.1016/j.medcli.2008.05.022.*

[20] Libermann TA, Nusbaum HR, Razon N, Kris R, Lax I, Soreq H, Whittle N, Waterfield MD, Ullrich A, Schlessinger J. *Amplification and overexpression of the EGF receptor gene in primary human glioblastomas.* J Cell Sci Suppl. 1985; 3():161-72.

[21] Carter NP .*Methods and strategies for analyzing copy number variation using DNA microarrays.* Nat Genet. 2007 Jul; 39(7 Suppl):S16-21.

[22] Meyerson M, Gabriel S, Getz G. *Advances in understanding cancer genomes through second-generation sequencing.* Nat Rev Genet. 2010 Oct; 11(10):685-96.

[23] Guo C1, Liu S, Wang J, Sun MZ, Greenaway FT. *ACTB in cancer.*Clin Chim Acta. 2013 Feb 18;417:39-44. doi: 10.1016/j.cca.2012.12.012. Epub 2012 Dec 22.

[24] He X, Zhou T, Yang G, Fang W, Li Z, Zhan J, Zhao Y, Cheng Z, Huang Y, Zhao H, Zhang L.*The expression of plakoglobin is a potential prognostic biomarker for patients with surgically resected lung adenocarcinoma.*] Oncotarget. 2016 Mar 22;7(12):15274-87. doi: 10.18632/oncotarget.7729.

[25] Wang XS1, Shankar S, Dhanasekaran SM, Ateeq B, Sasaki AT, Jing X, Robinson D, Cao Q, Prensner JR, Yocum AK, Wang R, Fries DF, Han B, Asangani IA, Cao X, Li Y, Omenn GS, Pflueger D, Gopalan A, Reuter VE, Kahoud ER, Cantley LC, Rubin MA, Palanisamy N, Varambally S, Chinnaiyan AM. *Characterization of KRAS rearrangements in metastatic prostate cancer.* Cancer Discov. 2011 Jun;1(1):35-43. doi: 10.1158/2159-8274.CD-10-0022. Epub 2011 Jun 1.

[26] Verma S, Ismail A, Gao X, Fu G, Li X, O'Malley BW, Nawaz Z. *The ubiquitin-conjugating enzyme UBCH7 acts as a coactivator for steroid hormone receptors.* Mol Cell Biol. 2004 Oct;24(19):8716-26.

[27] Wortzel, I., Seger, R. *The ERK Cascade.* Genes Cancer. 2011 Mar; 2(3): 195–209. doi: 10.1177/1947601911407328

[28] Sheinberger J, Shav-Tal Y. *mRNPs meet stress granules.* FEBS Lett. 2017 Sep;591(17):2534-2542. doi: 10.1002/1873-3468.12765. Epub 2017 Aug 8.

[29] Anderson, RG. *The caveolae membrane system*. Annu. Rev. Biochem. 67: 199-225.1998.doi:10.1146/annurev.biochem.67.1.199

[30] Haldrup C, Mundbjerg K, Vestergaard EM, Lamy P, Wild P, Schulz WA, Arsov C, Visakorpi T, Borre M, Høyer S, Orntoft TF, Sørensen KD. *DNA methylation signatures for prediction of biochemical recurrence after radical prostatectomy of clinically localized prostate cancer. J Clin Oncol. 2013 Sep 10;31(26):3250-8. doi: 10.1200/JCO.2012.47.1847. Epub 2013 Aug 5.*

[31] Liu L, Xiong Y, Xi W, Wang J, Qu Y, Lin Z, Chen X, Yao J, Xu J, Guo J. *Prognostic role of N-Acetylgalactosaminyltransferase 10 in metastatic renal cell carcinoma.* Oncotarget. *2017 Feb 28;8(9):14995-15003. doi: 10.18632/oncotarget.14786.*

[32] Wu Q, Yang L, Liu H, Zhang W, Le X, Xu J. *Elevated Expression of N-Acetylgalactosaminyltransferase 10 Predicts Poor Survival and Early Recurrence of Patients with Clear-Cell Renal Cell Carcinoma.* Ann Surg Oncol. 2015 Jul;22(7):2446-53. doi: 10.1245/s10434-014-4236-y. Epub 2014 Nov 13.

[33] Karlsson T, Krakstad C, Tangen IL, Hoivik EA, Pollock PM, Salvesen HB, Lewis AE. *Endometrial cancer cells exhibit high expression of p110β and its selective*

inhibition induces variable responses on PI3K signaling, cell survival and proliferation. Oncotarget. 2017 Jan 17;8(3):3881-3894. doi: 10.18632/oncotarget.13989.

[34] Wang T, Abou-Ouf H, Hegazy SA, Alshalalfa M, Stoletov K, Lewis J, Donnelly B, Bismar TA. *Ankyrin G expression is associated with androgen receptor stability, invasiveness, and lethal outcome in prostate cancer patients.* J Mol Med (Berl). 2016 Dec;94(12):1411-1422. Epub 2016 Aug 18.

[35] Yakovich AJ, Jiang B, Allen CE, Du J, Wu LC, Barnard JA. *ZAS3 accentuates transforming growth factor β signaling in epithelial cells.* Cell Signal. 2011 Jan;23(1):105-13. doi: 10.1016/j.cellsig.2010.08.009. Epub 2010 Aug 21.

[36] Zhou C, Li J, Lin L, Shu R, Dong B, Cao D, Li Q, Wang Z. *A targeted transforming growth factor-beta (TGF-β) blocker, TTB, inhibits tumor growth and metastasis.* Oncotarget. 2018 Feb 24;9(33):23102-23113. doi: 10.18632/oncotarget.24562. eCollection 2018 May 1.

[37] Teh JL, Shah R, La Cava S, Dolfi SC, Mehta MS, Kongara S, Price S, Ganesan S, Reuhl KR, Hirshfield KM, Karantza V, Chen S. *Metabotropic glutamate receptor 1 disrupts mammary acinar architecture and initiates malignant transformation of mammary epithelial cells.* Breast Cancer Res Treat. 2015 May;151(1):57-73. doi: 10.1007/s10549-015-3365-8. Epub 2015 Apr 10.

[38] Beck FW, Prasad AS, Butler CE, Sakr WA, Kucuk O, Sarkar FH. *Differential expression of hZnT-4 in human prostate tissues.* Prostate. 2004 Mar 1;58(4):374-81.

[39] Cuzick J, Yang ZH, Fisher G, Tikishvili E, Stone S, Lanchbury JS, Camacho N, Merson S, Brewer D, Cooper CS, Clark J, Berney DM, Møller H, Scardino P, Sangale Z; Transatlantic Prostate Group. *Prognostic value of PTEN loss in men with conservatively managed localised prostate cancer.* Br J Cancer. 2013 Jun 25;108(12):2582-9. doi: 10.1038/bjc.2013.248. Epub 2013 May 21.

[40] Zhong JT, Wang HJ, Yu J, Zhang JH, Wang SF, Yang X, Su W. *Correlations of the expressions of c-Jun and Egr-1 proteins with clinicopathological features and prognosis of patients with nasopharyngeal carcinoma.* Cancer Biomark. 2017;19(2):213-220. doi: 10.3233/CBM-161710.

[41] Wan X, Liu R, Li Z. *The Prognostic Value of HRAS mRNA Expression in Cutaneous Melanoma.* Biomed Res Int. 2017;2017:5356737. doi: 10.1155/2017/5356737. Epub 2017 Nov 19.

7. Anexos

Seguidamente se detalla el código de R usado para la ejecución de los análisis contenidos en este trabajo.

Código usado para el estudio de expresión diferencial

#Descarga y lectura de los datos del dataset usando el paquete GEOquery:

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("GEOquery")
library(GEOquery)
prostate <- getGEO("GSE6919")
```

#Separación de los datos según la plataforma utilizada para su análisis individual:

```
pGPL8300 <- prostate[1]
pGPL92 <- prostate[2]
pGPL93 <- prostate[3]
```

#Extracción de los datos de expresión y transformación con logaritmo base 2:

```
expGPL93 <- pGPL93$`GSE6919-GPL93_series_matrix.txt.gz`@assayData$exprs
expGPL92 <- pGPL92$`GSE6919-GPL92_series_matrix.txt.gz`@assayData$exprs
expGPL8300 <- pGPL8300$`GSE6919-GPL8300_series_matrix.txt.gz`@assayData$exprs
logGPL93 <- log2(expGPL93)
logGPL92 <- log2(expGPL92)
logGPL8300 <- log2(expGPL8300)
```

#Añadimos los datos sobre el origen de las muestras para poder identificarlos y agruparlas según el tejido de origen:

```
muestras93 <- colnames(logGPL93)
muestras93 <- cbind.data.frame(muestras93,pGPL93$`GSE6919-GPL93_series_matrix.txt.gz`@phenoData@data$title)
colnames(muestras93) <- c("Muestra","Descripción")
muestras92 <- colnames(logGPL92)
muestras92 <- cbind.data.frame(muestras92,pGPL92$`GSE6919-GPL92_series_matrix.txt.gz`@phenoData@data$title)
colnames(muestras92) <- c("Muestra","Descripción")
muestras8300 <- colnames(logGPL8300)
muestras8300 <- cbind.data.frame(muestras8300,pGPL8300$`GSE6919-
GPL8300_series_matrix.txt.gz`@phenoData@data$title)
colnames(muestras8300) <- c("Muestra","Descripción")

muestras92$Descripción <- as.character(muestras92$Descripción)
muestras92$Descripción[grepl("Normal",muestras92$Descripción)] <- "Normal"
```

```
muestras92$Descripción[grepl("Metastatic",muestras92$Descripción)] <- "Metástasis"
muestras92$Descripción[grepl("Tumor",muestras92$Descripción)] <- "T.Primario"
n <- muestras92[muestras92$Descripción=="Normal",]
t <- muestras92[muestras92$Descripción=="T.Primario",]
m <- muestras92[muestras92$Descripción=="Metástasis",]
muestras92 <- rbind.data.frame(n,t,m)

muestras93$Descripción <- as.character(muestras93$Descripción)
muestras93$Descripción[grepl("Normal",muestras93$Descripción)] <- "Normal"
muestras93$Descripción[grepl("Metastatic",muestras93$Descripción)] <- "Metástasis"
muestras93$Descripción[grepl("Tumor",muestras93$Descripción)] <- "T.Primario"
n <- muestras93[muestras93$Descripción=="Normal",]
t <- muestras93[muestras93$Descripción=="T.Primario",]
m <- muestras93[muestras93$Descripción=="Metástasis",]
muestras93 <- rbind.data.frame(n,t,m)

muestras8300$Descripción <- as.character(muestras8300$Descripción)
muestras8300$Descripción[grepl("Normal",muestras8300$Descripción)] <- "Normal"
muestras8300$Descripción[grepl("Metastatic",muestras8300$Descripción)] <- "Metástasis"
muestras8300$Descripción[grepl("Tumor",muestras8300$Descripción)] <- "T.Primario"
n <- muestras8300[muestras8300$Descripción=="Normal",]
t <- muestras8300[muestras8300$Descripción=="T.Primario",]
m <- muestras8300[muestras8300$Descripción=="Metástasis",]
muestras8300 <- rbind.data.frame(n,t,m)
```

#Usando el paquete limma realizaremos el modelado para llevar a cabo la comparación por parejas de tejidos:

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

#Modelado lineal de GPL92

```
grupoMvsP <- as.factor(muestras92$Descripción!="Normal")
grupoMvsN <- as.factor(muestras92$Descripción!="T.Primario")
grupoPvsN <- as.factor(muestras92$Descripción!="Metástasis")
designMvsP<-model.matrix(~grupoMvsP)
designMvsN<-model.matrix(~grupoMvsN)
designPvsN<-model.matrix(~grupoPvsN)
fitMvsP<-lmFit(logGPL92, designMvsP)
fitMvsP<-eBayes(fitMvsP)
tt92MvsP <- toptable(fitMvsP, coef=2,number = nrow(fitMvsP))
tt92MvsP <- subset(tt92MvsP,tt92MvsP$adj.P.Val<0.05)
fitMvsN<-lmFit(logGPL92, designMvsN)
fitMvsN<-eBayes(fitMvsN)
tt92MvsN <- toptable(fitMvsN, coef=2,number = nrow(fitMvsN))
tt92MvsN <- subset(tt92MvsN,tt92MvsN$adj.P.Val<0.05)
fitPvsN<-lmFit(logGPL92, designPvsN)
fitPvsN<-eBayes(fitPvsN)
tt92PvsN <- toptable(fitPvsN, coef=2,number = nrow(fitPvsN))
tt92PvsN <- subset(tt92PvsN,tt92PvsN$adj.P.Val<0.05)
```

#Modelado lineal de GPL93

grupoMvsP <- as.factor(muestras93$Descripción!="Normal")
grupoMvsN <- as.factor(muestras93$Descripción!="T.Primario")
grupoPvsN <- as.factor(muestras93$Descripción!="Metástasis")
designMvsP<-model.matrix(~grupoMvsP)
designMvsN<-model.matrix(~grupoMvsN)
designPvsN<-model.matrix(~grupoPvsN)
fitMvsP<-lmFit(logGPL93, designMvsP)
fitMvsP<-eBayes(fitMvsP)
tt93MvsP <- toptable(fitMvsP, coef=2,number = nrow(fitMvsP))
tt93MvsP <- subset(tt93MvsP,tt93MvsP$adj.P.Val<0.05)
fitMvsN<-lmFit(logGPL93, designMvsN)
fitMvsN<-eBayes(fitMvsN)
tt93MvsN <- toptable(fitMvsN, coef=2,number = nrow(fitMvsN))
tt93MvsN <- subset(tt93MvsN,tt93MvsN$adj.P.Val<0.05)
fitPvsN<-lmFit(logGPL93, designPvsN)
fitPvsN<-eBayes(fitPvsN)
tt93PvsN <- toptable(fitPvsN, coef=2,number = nrow(fitPvsN))
tt93PvsN <- subset(tt93PvsN,tt93PvsN$adj.P.Val<0.05)

#Modelado lineal de GPL8300

grupoMvsP <- as.factor(muestras8300$Descripción!="Normal")
grupoMvsN <- as.factor(muestras8300$Descripción!="T.Primario")
grupoPvsN <- as.factor(muestras8300$Descripción!="Metástasis")
designMvsP<-model.matrix(~grupoMvsP)
designMvsN<-model.matrix(~grupoMvsN)
designPvsN<-model.matrix(~grupoPvsN)
fitMvsP<-lmFit(logGPL8300, designMvsP)
fitMvsP<-eBayes(fitMvsP)
tt8300MvsP <- toptable(fitMvsP, coef=2,number = nrow(fitMvsP))
tt8300MvsP <- subset(tt8300MvsP,tt8300MvsP$adj.P.Val<0.05)
fitMvsN<-lmFit(logGPL8300, designMvsN)
fitMvsN<-eBayes(fitMvsN)
tt8300MvsN <- toptable(fitMvsN, coef=2,number = nrow(fitMvsN))
tt8300MvsN <- subset(tt8300MvsN,tt8300MvsN$adj.P.Val<0.05)
fitPvsN<-lmFit(logGPL8300, designPvsN)
fitPvsN<-eBayes(fitPvsN)
tt8300PvsN <- toptable(fitPvsN, coef=2,number = nrow(fitPvsN))
tt8300PvsN <- subset(tt8300PvsN,tt8300PvsN$adj.P.Val<0.05)

#Conversión de los ID de sonda a nombres de genes:

library(dplyr)
source("https://bioconductor.org/biocLite.R")
biocLite("hgu95b.db")
biocLite("hgu95c.db")
biocLite("hgu95av2.db")
library(hgu95b.db)
library(hgu95c.db)
library(hgu95av2.db)

#Convertiremos los ID de las sondas de Affymetrix a nombre de genes

```
a <- rownames(tt92MvsP)
tt92MvsP <- cbind.data.frame(PROBEID=a,tt92MvsP)
a2 <- rownames(tt92MvsN)
tt92MvsN <- cbind.data.frame(PROBEID=a2,tt92MvsN)
a3 <- rownames(tt92PvsN)
tt92PvsN <- cbind.data.frame(PROBEID=a3,tt92PvsN)

g92MP <- select(hgu95b.db, as.vector(tt92MvsP$PROBEID), c("SYMBOL"))
g92MN <- select(hgu95b.db, as.vector(tt92MvsN$PROBEID), c("SYMBOL"))
g92PN <- select(hgu95b.db, as.vector(tt92PvsN$PROBEID), c("SYMBOL"))

a <- rownames(tt93MvsP)
tt93MvsP <- cbind.data.frame(PROBEID=a,tt93MvsP)
a2 <- rownames(tt93MvsN)
tt93MvsN <- cbind.data.frame(PROBEID=a2,tt93MvsN)
a3 <- rownames(tt93PvsN)
tt93PvsN <- cbind.data.frame(PROBEID=a3,tt93PvsN)

g93MP <- select(hgu95c.db, as.vector(tt93MvsP$PROBEID), c("SYMBOL"))
g93MN <- select(hgu95c.db, as.vector(tt93MvsN$PROBEID), c("SYMBOL"))
g93PN <- select(hgu95c.db, as.vector(tt93PvsN$PROBEID), c("SYMBOL"))

a <- rownames(tt8300MvsP)
tt8300MvsP <- cbind.data.frame(PROBEID=a,tt8300MvsP)
a2 <- rownames(tt8300MvsN)
tt8300MvsN <- cbind.data.frame(PROBEID=a2,tt8300MvsN)
a3 <- rownames(tt8300PvsN)
tt8300PvsN <- cbind.data.frame(PROBEID=a3,tt8300PvsN)

g8300MP <- select(hgu95av2.db, as.vector(tt8300MvsP$PROBEID), c("SYMBOL"))
g8300MN <- select(hgu95av2.db, as.vector(tt8300MvsN$PROBEID), c("SYMBOL"))
g8300PN <- select(hgu95av2.db, as.vector(tt8300PvsN$PROBEID), c("SYMBOL"))
```

#Creamos una función fara fusionar los datos

```
juntar <- function(x,y){
        l1 <- nrow(x)
        l2 <- nrow(y)
        logFC <- c()
        Symbol <- c()
        for(i in 1:l1){
                for(j in 1:l2){
                        if(x[i,1]==y[j,1]){
                                logFC <- append(logFC,x[i,2])
                                Symbol <- append(Symbol,y[j,2])
                        }
                }
        }
        result <- cbind.data.frame(SYMBOL=Symbol,logFC=logFC)
```

```
        print(result)
}


gdeMP <- rbind.data.frame(juntar(tt92MvsP,g92MP),juntar(tt93MvsP,g93MP),juntar(tt8300MvsP,g8300MP))
gdeMP <- na.omit(gdeMP)
gdeMP <- gdeMP[order(gdeMP$SYMBOL),]
rownames(gdeMP) <- c()

gdeMN <- rbind.data.frame(juntar(tt92MvsN,g92MN),juntar(tt93MvsN,g93MN),juntar(tt8300MvsN,g8300MN))
gdeMN <- na.omit(gdeMN)
gdeMN <- gdeMN[order(gdeMN$SYMBOL),]
rownames(gdeMN) <- c()

gdePN <- rbind.data.frame(juntar(tt92PvsN,g92PN),juntar(tt93PvsN,g93PN),juntar(tt8300PvsN,g8300PN))
gdePN <- na.omit(gdePN)
gdePN <- gdePN[order(gdePN$SYMBOL),]
rownames(gdePN) <- c()

#Añadimos a los resultados si el gen está subexpresado o sobreexpresado

gdeMN <- cbind.data.frame(gdeMN,status=c(1))
gdePN <- cbind.data.frame(gdePN,status=c(1))
gdeMP <- cbind.data.frame(gdeMP,status=c(1))

gdeMN$status[gdeMN$logFC<0] <- 2
gdeMN$status[gdeMN$logFC>0] <- 3

gdePN$status[gdePN$logFC<0] <- 2
gdePN$status[gdePN$logFC>0] <- 3

gdeMP$status[gdeMP$logFC<0] <- 2
gdeMP$status[gdeMP$logFC>0] <- 3

#Código para el gráfico de comparación de valores de logFC entre grupos

boxplot(gdePN$logFC,gdeMN$logFC,gdeMP$logFC,col=c(3,2,4),ylab="logFC",main="Comparation of logFC
values",cex.lab=1.5,cex.main=1.5)
abline(h=0,col=1,lty=2)
axis(1,at=1:3,labels=c("Normal-Primary","Normal-Metastatic","Primary-Metastatic"),cex.axis=1.5)

#Graficos de cada uno de los valores de logFC para cada caso

par(mfrow=c(3,1),mar=c(2, 4, 2, 2))
plot(gdePN$logFC,col=3,pch=18,ylab="logFC",main="Normal-Primary logFC values")
abline(h=0,col=2)
plot(gdeMN$logFC,col=3,pch=18,ylab="logFC",main="Normal-Metastatic logFC values")
abline(h=0,col=2)
plot(gdeMP$logFC,col=3,pch=18,ylab="logFC",main="Primary-Metastatic logFC values")
```

#Código para la detección de genes diferencialmente expresados que se repiten entre los diferentes tejidos analizados

```
PN_MP <- merge(gdePN,gdeMP,by="SYMBOL")
PN <- PN_MP[,1:2]
MP <- cbind.data.frame(PN_MP$SYMBOL,PN_MP$logFC.y)
genesRepe <- readXL("C:/Users/hergo/Dropbox/Master Bioinformatica TFM/Versión 2/genesRepe.xlsx", rownames=FALSE, header=TRUE, na="", sheet="Hoja1", stringsAsFactors=TRUE)

barplot(t(genesRepe[-1]), names.arg=genesRepe[,1], beside=TRUE,col=c(2,3),las=2,ylab="logFC")
legend("topright",legend = c("Normal vs Primary","Primary vs Metastatic"),fill = c(2,3))
```

#Código para el análisis de enriquecimientos de términos GO

```
source("http://bioconductor.org/biocLite.R")
biocLite('biomartr')
biocLite("mygene")
library(mygene)
library(biomartr)
biocLite("clusterProfiler")
library(clusterProfiler)
library(org.Hs.eg.db)
biocLite("clusterProfiler")
biocLite("org.Hs.eg.db")
library(clusterProfiler)
library(org.Hs.eg.db)
```

#Primero convertimos los símbolos de los genes al ID de Entrez para poder hacer el estudio de enriquecimiento

```
genesPN = bitr(gdePN$SYMBOL, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
genesMN = bitr(gdeMN$SYMBOL, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
genesMP = bitr(gdeMP$SYMBOL, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
genesrepetidos = bitr(listagenesrepe, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")
```

#Realizamos el estudio de enriquecimiento de GO usando el pakage "DOSE" y llevamos a cabo el estudio para términos GO relacionados con ruta biológica (BP), compartimento celular (CC) y función molecular (MF).

```
data(geneList, package="DOSE")
egoBPpn <- enrichGO(gene         = genesPN$ENTREZID,
                    universe     = names(geneList),
                    OrgDb        = org.Hs.eg.db,
                    ont          = "BP",
                    pAdjustMethod = "BH",
                    pvalueCutoff  = 0.01,
                    qvalueCutoff  = 0.05,
                    readable     = TRUE)
egoCCpn <- enrichGO(gene         = genesPN$ENTREZID,
                    universe     = names(geneList),
                    OrgDb        = org.Hs.eg.db,
                    ont          = "CC",
                    pAdjustMethod = "BH",
```

```
            pvalueCutoff  = 0.01,
            qvalueCutoff  = 0.05,
            readable      = TRUE)
egoMFpn <- enrichGO(gene          = genesPN$ENTREZID,
            universe      = names(geneList),
            OrgDb         = org.Hs.eg.db,
            ont           = "MF",
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.01,
            qvalueCutoff  = 0.05,
            readable      = TRUE)
```

#Genes diferencialmente expresado normal-metastatico

```
data(geneList, package="DOSE")
egoBPmn <- enrichGO(gene          = genesMN$ENTREZID,
            universe      = names(geneList),
            OrgDb         = org.Hs.eg.db,
            ont           = "BP",
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.01,
            qvalueCutoff  = 0.05,
            readable      = TRUE)
egoCCmn <- enrichGO(gene          = genesMN$ENTREZID,
            universe      = names(geneList),
            OrgDb         = org.Hs.eg.db,
            ont           = "CC",
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.01,
            qvalueCutoff  = 0.05,
            readable      = TRUE)
egoMFmn <- enrichGO(gene          = genesMN$ENTREZID,
            universe      = names(geneList),
            OrgDb         = org.Hs.eg.db,
            ont           = "MF",
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.01,
            qvalueCutoff  = 0.05,
            readable      = TRUE)
```

#Genes diferencialmente expresado primario-metastatico

```
data(geneList, package="DOSE")
egoBPmp <- enrichGO(gene          = genesMP$ENTREZID,
            universe      = names(geneList),
            OrgDb         = org.Hs.eg.db,
            ont           = "BP",
            pAdjustMethod = "BH",
            pvalueCutoff  = 0.01,
            qvalueCutoff  = 0.05,
            readable      = TRUE)
```

```
egoCCmp <- enrichGO(gene         = genesMP$ENTREZID,
                    universe     = names(geneList),
                    OrgDb        = org.Hs.eg.db,
                    ont          = "CC",
                    pAdjustMethod = "BH",
                    pvalueCutoff  = 0.01,
                    qvalueCutoff  = 0.05,
                    readable     = TRUE)
egoMFmp <- enrichGO(gene         = genesMP$ENTREZID,
                    universe     = names(geneList),
                    OrgDb        = org.Hs.eg.db,
                    ont          = "MF",
                    pAdjustMethod = "BH",
                    pvalueCutoff  = 0.01,
                    qvalueCutoff  = 0.05,
                    readable     = TRUE)
```

#Genes repetidos en los diversos tejidos

```
data(geneList, package="DOSE")
egoBPrepe <- enrichGO(gene         = genesrepetidos$ENTREZID,
                      universe     = names(geneList),
                      OrgDb        = org.Hs.eg.db,
                      ont          = "BP",
                      pAdjustMethod = "BH",
                      pvalueCutoff  = 0.01,
                      qvalueCutoff  = 0.05,
                      readable     = TRUE)
egoCCrepe <- enrichGO(gene         = genesrepetidos$ENTREZID,
                      universe     = names(geneList),
                      OrgDb        = org.Hs.eg.db,
                      ont          = "CC",
                      pAdjustMethod = "BH",
                      pvalueCutoff  = 0.01,
                      qvalueCutoff  = 0.05,
                      readable     = TRUE)
egoMFrepe <- enrichGO(gene         = genesrepetidos$ENTREZID,
                      universe     = names(geneList),
                      OrgDb        = org.Hs.eg.db,
                      ont          = "MF",
                      pAdjustMethod = "BH",
                      pvalueCutoff  = 0.01,
                      qvalueCutoff  = 0.05,
                      readable     = TRUE)
```

#Selección y ordenación de los términos GO significativos con un valor p<0.01

```
egoBPpnS <- subset(egoBPpn,egoBPpn@result$pvalue<0.01)
egoCCpnS <- subset(egoCCpn,egoCCpn@result$pvalue<0.01)
egoMFpnS <- subset(egoMFpn,egoMFpn@result$pvalue<0.01)
egoBPmnS <- subset(egoBPmn,egoBPmn@result$pvalue<0.01)
```

```
egoCCmnS <- subset(egoCCmn,egoCCmn@result$pvalue<0.01)
egoMFmnS <- subset(egoMFmn,egoMFmn@result$pvalue<0.01)
egoBPmpS <- subset(egoBPmp,egoBPmp@result$pvalue<0.01)
egoCCmpS <- subset(egoCCmp,egoCCmp@result$pvalue<0.01)
egoMFmpS <- subset(egoMFmp,egoMFmp@result$pvalue<0.01)
egoBPrepeS <- subset(egoBPrepe,egoBPrepe@result$pvalue<0.01)
egoCCrepeS <- subset(egoCCrepe,egoCCrepe@result$pvalue<0.01)
egoMFrepeS <- subset(egoMFrepe,egoMFrepe@result$pvalue<0.01)

egoBPpnS <- egoBPpnS[order(egoBPpnS$Count,decreasing = TRUE),]
egoCCpnS <- egoCCpnS[order(egoCCpnS$Count,decreasing = TRUE),]
egoMFpnS <- egoMFpnS[order(egoMFpnS$Count,decreasing = TRUE),]
egoBPmnS <- egoBPmnS[order(egoBPmnS$Count,decreasing = TRUE),]
egoCCmnS <- egoCCmnS[order(egoCCmnS$Count,decreasing = TRUE),]
egoMFmnS <- egoMFmnS[order(egoMFmnS$Count,decreasing = TRUE),]
egoBPmpS <- egoBPmpS[order(egoBPmpS$Count,decreasing = TRUE),]
egoCCmpS <- egoCCmpS[order(egoCCmpS$Count,decreasing = TRUE),]
egoMFmpS <- egoMFmpS[order(egoMFmpS$Count,decreasing = TRUE),]
egoBPrepeS <- egoBPrepeS[order(egoBPrepeS$Count,decreasing = TRUE),]
egoCCrepeS <- egoCCrepeS[order(egoCCrepeS$Count,decreasing = TRUE),]
egoMFrepeS <- egoMFrepeS[order(egoMFrepeS$Count,decreasing = TRUE),]
```

#Listado de genes de cada término GO para asignar color en función de si hay sobre expresión o subexpresión

```
b1 <- strsplit(egoBPpnS$geneID,"/")
b2 <- strsplit(egoMFpnS$geneID,"/")
b3 <- strsplit(egoCCpnS$geneID,"/")
b4 <- strsplit(egoBPmnS$geneID,"/")
b5 <- strsplit(egoMFmnS$geneID,"/")
b6 <- strsplit(egoCCmnS$geneID,"/")
b7 <- strsplit(egoBPmpS$geneID,"/")
b8 <- strsplit(egoMFmpS$geneID,"/")
b9 <- strsplit(egoCCmpS$geneID,"/")
b10 <- strsplit(egoBPrepeS$geneID,"/")
b11 <- strsplit(egoCCrepeS$geneID,"/")
b12 <- strsplit(egoMFrepeS$geneID,"/")
```

#Función que cruzará los genes asociados con sus valores de expresión diferencial. Determinará si hay más sobreexpresión o subexpresión o si hay mismo número de genes. Asignará un valor (color) a cada término GO asociado que usaremos para colorear las gráficas de enriquecimiento

```
GOexpcol <- function(x,y){
        l <- length(x)
        result <- c()
        for(i in 1:l){
                m <- merge(y,data.frame(SYMBOL=unlist(x[i])),by="SYMBOL")
                m1 <- table(m[3])
                sobre <- m1[1]
                sub <- m1[2]
                if(sobre>sub | is.na(sub)){
                        result <- append(result,3)
```

```
                    }else{
                      if(sobre<sub | is.na(sobre)){
                        result <- append(result,2)
                      }else{
                        result <- append(result,1)
                      }
                    }
                }
                print(result)
        }

        goBPpncol <- GOexp(b1,gdePN2)
        goMFpncol <- GOexp(b2,gdePN2)
        goCCpncol <- GOexp(b3,gdePN2)
        goBPmncol <- GOexp(b4,gdeMN2)
        goMFmncol <- GOexp(b5,gdeMN2)
        goCCmncol <- GOexp(b6,gdeMN2)
        goBPmpcol <- GOexp(b7,gdeMP2)
        goMFmpcol <- GOexp(b8,gdeMP2)
        goCCmpcol <- GOexp(b9,gdeMP2)
        goBPrepecol <- GOexp(b10,gdeMP2)
        goMFrepecol <- GOexp(b11,gdeMP2)
        goCCrepecol <- GOexp(b12,gdeMP2)
```

#Códigos de los gráficos de resultados de enriquecimiento de términos GO

```
        par(mar=c(5,30, 4, 2))
        barplot(egoBPpnS$Count,horiz=TRUE,names.arg=egoBPpnS$Description,las=1,col=goBPpncol,xlab="Number of genes",main="GO terms enriched for Biological Pathway \n in Normal vs Primary")

        par(mar=c(5,30, 4, 2))
        barplot(egoMFpnS$Count,horiz=TRUE,names.arg=egoMFpnS$Description,las=1,col=goMFpncol,xlab="Number of genes",main="GO terms enriched for Molecular Function \n in Normal vs Primary")

        par(mar=c(5,30, 4, 2))
        barplot(egoCCpnS$Count,horiz=TRUE,names.arg=egoCCpnS$Description,las=1,col=goCCpncol,xlab="Number of genes",main="GO terms enriched for Cell Compartment \n in Normal vs Primary")

        par(mar=c(5,30, 4, 2))
        barplot(egoBPmnS$Count,horiz=TRUE,names.arg=egoBPmnS$Description,las=1,col=goBPmncol,xlab="Number of genes",main="GO terms enriched for Biological Pathway \n in Normal vs Metastatic")

        par(mar=c(5,40, 4, 2))
        barplot(egoMFmnS$Count,horiz=TRUE,names.arg=egoMFmnS$Description,las=1,col=goMFmncol,xlab="Number of genes",main="GO terms enriched for \n Molecular Function \n in Normal vs Metastatic")

        par(mar=c(5,30, 4, 2))
        barplot(egoCCmnS$Count,horiz=TRUE,names.arg=egoCCmnS$Description,las=1,col=goCCmncol,xlab="Number of genes",main="GO terms enriched for Cell Compartment \n in Normal vs Metastatic")
```

par(mar=c(5,30, 4, 2))

barplot(egoBPmpS$Count,horiz=TRUE,names.arg=egoBPmpS$Description,las=1,col=goBPmpcol,xlab="Number of genes",main="GO terms enriched for Biological Pathway \n in Primary vs Metastatic")

par(mar=c(5,30, 4, 2))

barplot(egoMFmpS$Count,horiz=TRUE,names.arg=egoMFmpS$Description,las=1,col=goMFmpcol,xlab="Number of genes",main="GO terms enriched for Molecular Function \n in Primary vs Metastatic")

par(mar=c(5,30, 4, 2))

barplot(egoCCmpS$Count,horiz=TRUE,names.arg=egoCCmpS$Description,las=1,col=goCCmpcol,xlab="Number of genes",main="GO terms enriched for Cell Compartment \n in Primary vs Metastatic")

par(mar=c(5,30, 4, 2))

barplot(egoBPrepeS$Count,horiz=TRUE,names.arg=egoBPrepeS$Description,las=1,col=goBPrepePNcol,xlab="Number of genes",main="GO terms enriched for Biological Pathway \n in genes altered on the 3 tissues \n (NormalvsPrimary)")

par(mar=c(5,30, 4, 2))

barplot(egoBPrepeS$Count,horiz=TRUE,names.arg=egoBPrepeS$Description,las=1,col=goBPrepeMPcol,xlab="Number of genes",main="GO terms enriched for Biological Pathway \n in genes altered on the 3 tissues \n (PrimaryvsMetastatic)")

par(mar=c(5,30, 4, 2))

barplot(egoMFrepeS$Count,horiz=TRUE,names.arg=egoMFrepeS$Description,las=1,col=goMFrepePNcol,xlab="Number of genes",main="GO terms enriched for Molecular Function \n in genes altered on the 3 tissues \n (NormalvsPrimary)")

barplot(egoMFrepeS$Count,horiz=TRUE,names.arg=egoMFrepeS$Description,las=1,col=goMFrepeMPcol,xlab="Number of genes",main="GO terms enriched for Molecular Function \n in genes altered on the 3 tissues \n (PrimaryvsMetastatic)")

par(mar=c(5,30, 4, 2))

barplot(egoCCrepeS$Count,horiz=TRUE,names.arg=egoCCrepeS$Description,las=1,col=goCCrepePNcol,xlab="Number of genes",main="GO terms enriched for Cell Compartment \n in genes altered on the 3 tissues \n (NormalvsPrimary)")

par(mar=c(5,30, 4, 2))

barplot(egoCCrepeS$Count,horiz=TRUE,names.arg=egoCCrepeS$Description,las=1,col=goCCrepeMPcol,xlab="Number of genes",main="GO terms enriched for Cell Compartment \n in genes altered on the 3 tissues \n (PrimaryvsMetastatic)")

Código usado para el estudio de regiones diferencialmente metiladas

install.packages("stringi")

library(minfi)

install.packages("doParallel")

library(doParallel)

registerDoParallel(cores = 4)

invisible(utils::memory.limit(160000))

#Lectura de datos descargados en el disco duro

RGSet <- read.metharray.exp("C:/Users/hergo/Dropbox/Master Bioinformatica TFM/Metilacion data")

#Creamos un objeto con el tipo de muestra según el orden del dataset (normal/tumoral).

pheno <- c("normal", "tumor", "tumor", "tumor", "normal", "normal", "tumor", "tumor", "normal", "tumor", "tumor", "tumor", "normal", "normal", "tumor", "tumor", "normal", "tumor", "tumor", "tumor", "normal", "tumor", "tumor", "tumor", "normal", "normal", "tumor", "tumor", "normal", "tumor", "tumor", "normal", "normal", "tumor", "tumor", "tumor", "normal", "tumor", "tumor", "tumor", "normal", "tumor", "tumor", "normal", "tumor", "tumor", "tumor")

MSet <- preprocessRaw(RGSet)
RSet <- ratioConvert(MSet, what = "both", keepCN = TRUE)

densityPlot(MSet, sampGroups = pheno)

#Normalización de los datos entre muestras

GRset.funnorm <- preprocessFunnorm(RGSet)

#Uso de la función bumphunter para detecra las DMRs

designMatrix <- model.matrix(~ pheno)

dmrs2 <- bumphunter(GRset.funnorm, design = designMatrix,cutoff = 0.4, B=1000, type="Beta")

plot(table(dmrs$table$chr),las=2,ylab="Number of dmrs",xlab="Chromosome",main="DMRs in prostate normal vs. tumor",col=2)

tableDMRs <- dmrs2$table
tableDMRs <- cbind.data.frame(tableDMRs,size=tableDMRs$end-tableDMRs$start)
sigDMRs <- subset(tableDMRs,tableDMRs$fwer<=0.01)
loneDMRs <- subset(sigDMRs,sigDMRs$size==0)
largeDMRs <- subset(sigDMRs,sigDMRs$size>0)

#Generación de gráficos para visualización de resultados:
#Distribución de DMR en general en los cromosomas

barplot(table(sigDMRs$chr),las=2,col=3,ylim=c(0,10),cex.lab=1.5, main="Chromosome distribution of the DMRs",ylab="Number of DMRs",xlab="Chromosome")

#Distribución de DMR de 1 bp o mayores en los cromosomas

barplot(table(largeDMRs$chr),las=2,col=3,ylim=c(0,3),cex.lab=1.5, main="Chromosome distribution of the DMRs (>1 pb)",ylab="Number of DMRs",xlab="Chromosome")
barplot(table(loneDMRs$chr),las=2,col=3,ylim=c(0,10),cex.lab=1.5, main="Chromosome distribution of the DMRs (1 pb)",ylab="Number of DMRs",xlab="Chromosome")

*#**"Intensidad" de la metilación entre los dos tipos de DMRs***

boxplot(loneDMRs$value,largeDMRs$value,col=c(2,3),axt="n",main="Metilation values",ylab="Value",xlab="DMR size")
axis(1,at=1:2,labels=(c("1 bp",">1 bp")))

#Obtención de los genes asociados a las DMRs localizadas

```
source("https://bioconductor.org/biocLite.R")
biocLite("biomaRt")
library(biomaRt)
ensembl=useMart("ensembl",dataset="hsapiens_gene_ensembl")

chromosome_name <- sigDMRs$chr
chromosome_name <- gsub("chr", "", paste(chromosome_name))
start <- sigDMRs$start
end <- sigDMRs$end


loneDMRs$query = paste(gsub("chr","",loneDMRs$chr),loneDMRs$start,loneDMRs$end, sep = ":")
largeDMRs$query = paste(gsub("chr","",largeDMRs$chr),largeDMRs$start,largeDMRs$end, sep = ":")


genesloneDMR <- getBM(c("ensembl_gene_id","hgnc_symbol","start_position","end_position","transcript_biotype"),
filters = c("chromosomal_region"), values = list(loneDMRs$query), mart = ensembl)
geneslargeDMR <- getBM(c("ensembl_gene_id","hgnc_symbol","start_position","end_position","transcript_biotype"),
filters = c("chromosomal_region"), values = list(largeDMRs$query), mart = ensembl)

genesloneDMRs = genesloneDMR[which(genesloneDMR$transcript_biotype=="protein_coding"),]
geneslargeDMRs = geneslargeDMR[which(geneslargeDMR$transcript_biotype=="protein_coding"),]


genes <- unique(genesDMRs$hgnc_symbol)
```

#Cruce de datos entre los genes diferencialmente expresados y los genes asociados a las DMR para tratar de identificar genes cuya expresión varía que esté asociados a regiones diferencialmente metiladas

```
merge(gdePN,genesloneDMRs,by="SYMBOL")
merge(gdePN,geneslargeDMRs,by="SYMBOL")

merge(gdeMN,genesloneDMRs,by="SYMBOL")
merge(gdeMN,geneslargeDMRs,by="SYMBOL")

merge(gdeMP,genesloneDMRs,by="SYMBOL")
merge(gdeMP,geneslargeDMRs,by="SYMBOL")
```

Código usado para el estudio de la variación de número de copias

 # Instalació y cargado de todos los paquetes requeridos

```
setwd("C:/Users/ghernandez/Dropbox/Master Bioinformatica TFM")
install.packages("cgdsr")
library(cgdsr)
library(ggplot2)
#library(ggpubr)
library(rmarkdown)
library(knitr)
```

```
library(ggplot2)

# Creamos un objeto CGDS, que nos permitirá conectarnos al servidor
mycgds = CGDS("http://www.cbioportal.org/public-portal/")

# Nos descargamos el listado de estudios disponibles y seleccionaremos el que nos interesa. En nuestro caso el estudio al que queremos acceder tiene el número 162:

listado <- getCancerStudies(mycgds)
a <- 162
prostate <- getCancerStudies(mycgds)[a,]

# Obtenemos las referencias de las muestras de nuestro estudio
mycancerstudy = getCancerStudies(mycgds)[162,1]
mycaselist = getCaseLists(mycgds,mycancerstudy)[3,1]

# Obtenemos el ID de las muestras
mygeneticprofile = getGeneticProfiles(mycgds,mycancerstudy)[1,1]

# Obtenemos los datos de CNV para nuestras muestras y para los genes diferencialmente expresados. Debido al gran número de genes identificados, en el caso de la comparación PrimariovsMetastático tenemos que recabar los datos de CNV en 4 bloques, ya que el servidor solo nos deja explorar un máximo de 1000 genes de forma simultánea:

cnaPNc <- getProfileData(mycgds,unique(gdePN$SYMBOL),mygeneticprofile,mycaselist)
cnaMNc <- getProfileData(mycgds,unique(gdeMN$SYMBOL),mygeneticprofile,mycaselist)
cnaMP0c <- getProfileData(mycgds,unique(gdeMP$SYMBOL)[1:999],mygeneticprofile,mycaselist)
cnaMP1c <- getProfileData(mycgds,unique(gdeMP$SYMBOL)[1000:1999],mygeneticprofile,mycaselist)
cnaMP2c <- getProfileData(mycgds,unique(gdeMP$SYMBOL)[2000:2999],mygeneticprofile,mycaselist)
cnaMP3c <- getProfileData(mycgds,unique(gdeMP$SYMBOL)[3000:3201],mygeneticprofile,mycaselist)

#Eliminamos las columnas de genes sin datos

cnaPNc <- cnaPNc[ , ! apply( cnaPNc , 2 , function(x) all(is.na(x)) ) ]
cnaMNc <- cnaMNc[ , ! apply( cnaMNc , 2 , function(x) all(is.na(x)) ) ]
cnaMP0c <- cnaMP0c[ , ! apply( cnaMP0c , 2 , function(x) all(is.na(x)) ) ]
cnaMP1c <- cnaMP1c[ , ! apply( cnaMP1c , 2 , function(x) all(is.na(x)) ) ]
cnaMP2c <- cnaMP2c[ , ! apply( cnaMP2c , 2 , function(x) all(is.na(x)) ) ]
cnaMP3c <- cnaMP3c[ , ! apply( cnaMP3c , 2 , function(x) all(is.na(x)) ) ]

#Juntamos los cuatro archivos de la comparación PrimariovsMetastático

cnaMPc <- cbind.data.frame(cnaMP0c,cnaMP1c,cnaMP2c,cnaMP3c)

#Creamos una función personalizada que nos permitirá juntar los dos tipos de deleciones y los dos tipos de amplificaciones y calcular el porcentaje respecto al total de casos en los que se tienen datos de CNV (que no siempre son los mismos para todos los genes)

cna <- function(x){
  c <- ncol(x)
  names <- colnames(x)
  tabla <- c()
```

```
  fila <- c()
  for(i in 1:c){
    r <- nrow(na.omit(x[i]))
    fila <- c(((sum(na.omit(x[,i])<0))/r)*100,((sum(na.omit(x[,i])>0))/r)*100)
    tabla <- rbind.data.frame(tabla,fila)
  }
  tabla <- cbind.data.frame(names,tabla)
  colnames(tabla) <- c("SYMBOL","Deleción","Amplificación")
  print(tabla)
}

cnaPNc2 <- cna(cnaPNc)
cnaMNc2 <- cna(cnaMNc)

cnaMP0c2 <- cna(cnaMP0c)
cnaMP1c2 <- cna(cnaMP1c)
cnaMP2c2 <- cna(cnaMP2c)
cnaMP3c2 <- cna(cnaMP3c)

cnaMPc2 <- rbind.data.frame(cnaMP0c2,cnaMP1c2,cnaMP2c2,cnaMP3c2)
cnaMPc2 <- cnaMPc2[cnaMPc2$SYMBOL!="tipo",]

#Recodificado de genes diferencialmente expresados según si son sobre o sub

gdePN2 <- cbind.data.frame(gdePN,variacion=c(1))
gdeMN2 <- cbind.data.frame(gdeMN,variacion=c(1))
gdeMP2 <- cbind.data.frame(gdeMP,variacion=c(1))

gdePN2$status[gdePN2$logFC>0] <- "Sobre"
gdePN2$status[gdePN2$logFC<0] <- "Sub"

gdeMN2$status[gdeMN2$logFC>0] <- "Sobre"
gdeMN2$status[gdeMN2$logFC<0] <- "Sub"

gdeMP2$status[gdeMP2$logFC>0] <- "Sobre"
gdeMP2$status[gdeMP2$logFC<0] <- "Sub"

gdePN2 <- gdePN2[!duplicated(gdePN2), ]
gdeMN2 <- gdeMN2[!duplicated(gdePN2), ]
gdeMP2 <- gdeMP2[!duplicated(gdePN2), ]

#Estudios de posibles correlaciones del CNV con los cambios de expresión

corrPN <- merge(gdePN2,cnaPNc2)
corrMN <- merge(gdeMN2,cnaMNc2)
corrMP <- merge(gdeMP2,cnaMPc2)

boxplot(corrPN$Deleción,corrPN$Amplificación,corrMN$Deleción,corrMN$Amplificación,corrMP$Deleción,corrMP$
Amplificación,col=c(2,2,3,3,5,5),axt="n",ylab="% of alteration")
axis(1,at=1:6,labels=c(rep(c("Deletion","Amplification"),3)))
legend("topleft",legend = c("Normal vs Primary","Normal vs Metastatic","Primary vs Metastatic"),fill = c(2,3,5))
```

#Estudio estadístico de las posibles diferencias en % de deleción/amplificación en los diferentes tejidos analizados

ks.test(corrPN$Deleción,corrMN$Deleción)
ks.test(corrPN$Deleción,corrMP$Deleción)
ks.test(corrPN$Amplificación,corrMN$Amplificación)
ks.test(corrPN$Amplificación,corrMP$Amplificación)
ks.test(corrMN$Deleción,corrMP$Deleción)
ks.test(corrMN$Amplificación,corrMP$Amplificación)

#Gráficos para visualizar posibles correlaciones entre el % de deleciones y amplificaciones con los cambios de expresión. Sin separar los genes en función de sus cambios de expresión.

par(mfrow=c(3,2))
plot(corrPN$Deleción,corrPN$logFC,col=4,pch=17,ylab="logFC",xlab="% Deleción",main="Correlación expresión-deleciones \n en NormalvsPrimario",cex.lab=1.5)
plot(corrPN$Amplificación,corrPN$logFC,col=4,pch=19,ylab="logFC",xlab="% Amplificación",main="Correlación expresión-amplificaciones \n en NormalvsPrimario",cex.lab=1.5)
plot(corrMN$Deleción,corrMN$logFC,col=2,pch=17,ylab="logFC",xlab="% Deleción",main="Correlación expresión-deleciones \n en NormalvsMetastático",cex.lab=1.5)
plot(corrMN$Amplificación,corrMN$logFC,col=2,pch=19,ylab="logFC",xlab="% Amplificación",main="Correlación expresión-amplificaciones \n en NormalvsMetastático",cex.lab=1.5)
plot(corrMP$Deleción,corrMP$logFC,col=8,pch=21,ylab="logFC",xlab="% Deleción",main="Correlación expresión-deleciones \n en PrimariovsMetastático",cex.lab=1.5)
plot(corrMP$Amplificación,corrMP$logFC,col=8,pch=21,ylab="logFC",xlab="% Amplificación",main="Correlación expresión-amplificaciones \n en PrimariovsMetastático",cex.lab=1.5)

#Gráficos para visualizar posibles correlaciones entre el % de deleciones y amplificaciones con los cambios de expresión. Separando los genes en función de si están o no sobreexpresados.

par(mfrow=c(3,2))
plot(corrPN[corrPN$status=="Sub",]$Deleción,corrPN[corrPN$status=="Sub",]$logFC,col=4,pch=17,ylab="logFC",xlab="% Deleción",main="Correlación expresión-deleciones \n en NormalvsPrimario",cex.lab=1.5)
plot(corrPN[corrPN$status=="Sobre",]$Amplificación,corrPN[corrPN$status=="Sobre",]$logFC,col=4,pch=19,ylab="logFC",xlab="% Amplificación",main="Correlación expresión-amplificaciones \n en NormalvsPrimario",cex.lab=1.5)
plot(corrMN[corrMN$status=="Sub",]$Deleción,corrMN[corrMN$status=="Sub",]$logFC,col=2,pch=17,ylab="logFC",xlab="% Deleción",main="Correlación expresión-deleciones \n en NormalvsMetastático",cex.lab=1.5)
plot(corrMN[corrMN$status=="Sobre",]$Amplificación,corrMN[corrMN$status=="Sobre",]$logFC,col=2,pch=19,ylab="logFC",xlab="% Amplificación",main="Correlación expresión-amplificaciones \n en NormalvsMetastático",cex.lab=1.5)
plot(corrMP[corrMP$status=="Sub",]$Deleción,corrMP[corrMP$status=="Sub",]$logFC,col=8,pch=21,ylab="logFC",xlab="% Deleción",main="Correlación expresión-deleciones \n en PrimariovsMetastático",cex.lab=1.5)
plot(corrMP[corrMP$status=="Sobre",]$Amplificación,corrMP[corrMP$status=="Sobre",]$logFC,col=8,pch=21,ylab="logFC",xlab="% Amplificación",main="Correlación expresión-amplificaciones \n en PrimariovsMetastático",cex.lab=1.5)

#Gráficos para visualizar posibles correlaciones entre el % de deleciones y amplificaciones

par(mfrow=c(3,1))

plot(corrPN$Deleción,corrPN$Amplificación,col=4,pch=17,ylab="% Amplificación",xlab="% Deleción",main="Correlación deleciones-amplificaciones \n en NormalvsPrimario",cex.lab=1.5)

plot(corrMN$Deleción,corrMN$Amplificación,col=4,pch=17,ylab="% Amplificación",xlab="% Deleción",main="Correlación deleciones-amplificaciones \n en NormalvsMetastático",cex.lab=1.5)

plot(corrMP$Deleción,corrMP$Amplificación,col=4,pch=17,ylab="% Amplificación",xlab="% Deleción",main="Correlación deleciones-amplificaciones \n en PrimariovsMetastático",cex.lab=1.5)

#Determinación de genes que no se encuentran amplificados en ningún caso del estudio

PNamp <- corrPN[corrPN$Amplificación==0,]

MNamp <- corrMN[corrMN$Amplificación==0,]

MPamp <- corrMP[corrMP$Amplificación==0,]

#Determinación de genes que no se encuentran delecionados en ningún caso del estudio

PNdel <- corrPN[corrPN$Deleción==0,]

MNdel <- corrMN[corrMN$Deleción==0,]

MPdel <- corrMP[corrMP$Deleción==0,]