PROYECTO BIOINFORMÁTICA

“INTEGRATION OF METHYLATION AND EXPRESSION DATA”

Alfons Navarro Ponz
Máster de Bioinformática y Bioestadística 2014-2015
Tutor: Álex Sánchez Pla

Tabla de contenido

Introducción

El término epigenética hace referencia a todos aquellos mecanismos que intervienen en la regulación de la expresión génica sin producir cambios en la secuencia de nucleótidos, y que son potencialmente heredables[1]. Éste término fue descrito por primera vez por Conrad Waddington en 1939[2] y desde entonces se ha convertido en uno de los campos en expansión más prometedores en el contexto de la investigación biomédica[3]. Podemos agrupar de forma amplia las modificaciones epigenéticas en tres grandes categorías: la metilación del ADN, las modificaciones de histonas y el posicionamiento nucleosómico[4].
La metilación del ADN ha sido la modificación epigenética más estudiada en humanos. Ésta se basa en la adición de un grupo metilo a una citosina y suele producirse principalmente en el contexto de los dinucleótidos CpG. Además, los dinucleótidos CpG tienden a agruparse en clústeres denominados islas CpG, que se definen como regiones de más de 200 bases con un contenido de G+C de al menos el 50% y una ratio de la frecuencia de CpG observadas versus las estadísticamente esperadas de como mínimo del 0.6[5]. Aproximadamente el 60% de los promotores de genes humanos están asociados con islas CpG[6] y en general su metilación se asocia con silenciamiento génico (Figura 1). La metilación del ADN puede inhibir la expresión génica mediante diversos mecanismos que se resumen en la Figura 1.

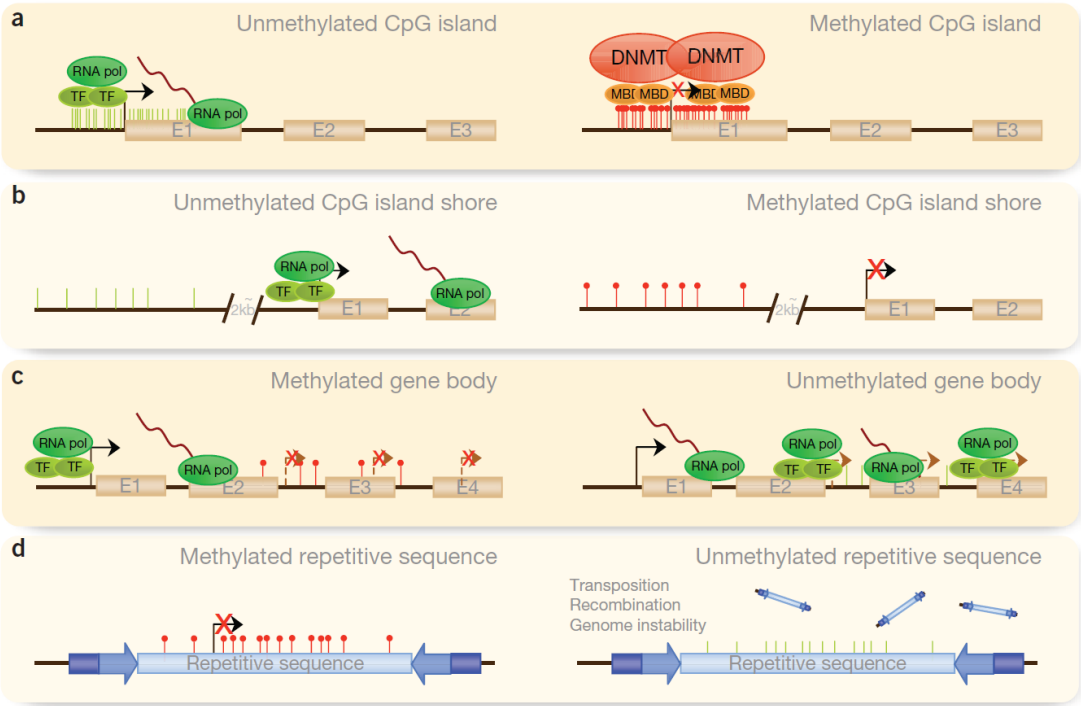


Figure 1. Patrones de metilación del ADN [4].

Es interesante destacar que la metilación del ADN no ocurre de forma exclusiva en las islas CpG. Recientemente, se han definido los denominados “CpG islands shores”, que son aquellas regiones con poca densidad de dinucleótidos CpG que

están situados cerca de una isla CpG(~2kb). La metilación de éstas regiones se ha visto fuertemente asociada a represión transcripcional (Figura 1b)[7]. Por otro lado, cuando la metilación se da en el cuerpo del gen, parece ser que se asocia a activación transcripcional (mucho menos frecuente), ya que se ha propuesto que puede estar relacionada con regulación de la eficiencia de elongación de la transcripción entre otros mecanismos (Figura 1C). Finalmente, se ha observado que la metilación no está asociada sólo a regulación de la transcripción génica, ya que se ha observado que una importante fracción de CpGs metiladas se encuentran situadas en regiones ricas en elementos repetitivos (Figura 1d) donde juegan un papel protector de la integridad genómica evitando la reactivación de secuencias de tipo transposón[8].

En los últimos años, gracias a la implementación de nuevas tecnologías como la secuenciación masiva del genoma, o las mejoras en las plataformas de arrays de metilación, se han realizado grandes avances en el campo del estudio de la metilación. A pesar de ello, todavía son necesarias muchas mejoras en las tecnologías utilizadas y especialmente en las herramientas de análisis de las grandes cantidades de datos obtenidos.

Utilizando datos obtenidos mediante las diferentes plataformas de "high-throughput", diversos estudios han abordado el estudio integrado de expresión y metilación[9-13], con el objetivo de mejorar la comprensión de regulación de la expresión de los genes.

A nivel de metodologías empleadas para la integración de los datos de expresión génica y de metilación, diferentes autores han probado distintos métodos. En el trabajo de Bell y colaboradores[14] utilizan la correlación de Spearman para analizar la relación entre grado de metilación (Illumina Human Methylation 27 DNA Analysis Bead Chips) y de expresión génica (RNAseq), y consiguen identificar correlaciones negativas entre la metilación de algunos promotores y su respectiva expresión génica. En otro estudio, van Eijk KR y colaboradores, utilizan un modelo de regresión lineal multivariado, donde el nivel de expresión génica es la variable dependiente y el nivel de metilación es la variable independiente. La edad y el género se utilizan como co-variables. En este estudio identifican tanto correlaciones positivas como negativas entre la expresión génica y el grado de metilación. En un estudio más reciente de Plume JM y colaboradores utilizan tanto el coeficiente de correlación de Pearson, como el rango de correlación de Spearman. En otro estudio, de Seungyeul y colaboradores, realizan un test de Causalidad para identificar genes regulados por metilación[15], y acaban identificando entre otros al gen *EPAS1* como un gen regulado por metilación con función importante en la enfermedad pulmonar obstructiva crónica.

En otro estudio de Taskesen E y colaboradores en un estudio en leucemia mieloide aguda, los autores realizan una discretización de las variables de metilación y de expresión para identificar genes regulados por metilación. Para ello categorizan los diferentes genes en tres grupos de acuerdo a su expresión: genes up-regulados, genes down-regulados y genes que no cambian. Al mismo tiempo el grado de metilación se agrupa en: hipermetilado, hipometilado o no cambio. En ambos casos se comparaban las muestras de leucemia con muestras controles CD34+. A continuación integran ambos datasets y generan un mapa de calor con diferentes

colores para los diferentes escenarios posibles, de forma que se pueden visualizar aquellos genes infraexpresados cuya metilación está hipermetilada, etc.
Hasta el momento la mayoría de trabajos que han integrado los datos de metilación y de expresión génica han encontrado pocas correlaciones entre el grado de metilación del promotor y la expresión génica. Algunos autores, hipotetizan que esto es debido a que la mayoría de métodos de análisis no capturan la diversidad de patrones de metilación existentes, ya que estos pueden variar de gen a gen[12]. El método más común para caracterizar los cambios de metilación entre dos muestras utiliza una ventana deslizante (que muestra una región concreta del genoma, cuyo tamaño se ha decidido de forma arbitraria) para identificar regiones diferencialmente metiladas (DMR). Un gen con una DMR hipermetilada cerca de su promotor se asume que tendrá una reducción de su expresión, mientras que un gen con una DMR hipometilada, tendrá aumentada su expresión. En la práctica, el coeficiente de correlación de Pearson entre el grado de metilación de la DMR y la expresión de su gen asociado está alrededor de -0.3[16]. Se cree que no se pueden conseguir mejores valores de correlación debido entre otras coas a al ruido producido por los errores experimentales, a las mezcla de poblaciones celulares en los tejidos, a las variaciones en el número de copias, etc.

Rhee y colaboradores[17] integran los datos de expresión génica y de metilación utilizando el coeficiente de correlación de Pearson para identificar correlaciones negativas. En este trabajo, como innovación a los trabajo previos, los autores analizan por separado diferentes regiones: CGIs, CGI shores, regiones promotoras, primer exón y segundo exón (Figura 2).

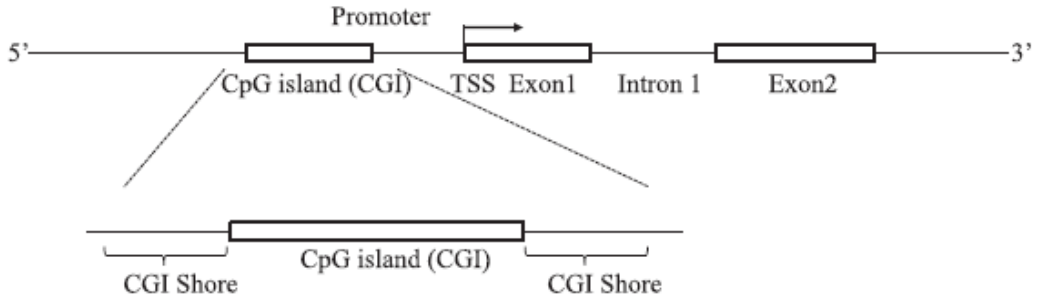


Figura 2. Regiones genómicas para estudiar la metilación[17].

En otro trabajo, VanderKraats y colaboradores[12] utilizan otra aproximación, para también no tener en consideración únicamente la información asociada a la región promotora. Estos autores representan la metilación diferencial como una curva interpolada, o como una firma, y a continuación identifican grupos de genes con una firma que presente una forma similar y cuyos cambios de expresión se correspondan con los cambios de metilación (Figura 3). Los autores representan como se puede ver en la Figura 3, la metilación diferencial para un área fijada alrededor de cada sitio de inicio de trancripción de cada gen (TSS) como una curva o firma de metilación para cada región determinada. A continuación realizan una medida de similaridad de curva mediante el uso de la distancia de Fréchet, para

comparar las firmas de metilación diferencial entre todos los genes. A continuación utilizando técnicas de clusterización no supervisada agrupan las firmas de acuerdo a su forma e identifican que clústers de genes exhiben cambios significativos en su expresión. En base a los resultados obtenidos, los autores concluyen que la simple caracterización de la región promotora del gen como metilada o desmetilada es totalmente insuficiente, ya que observan que la consideración del conjunto de cambios en la metilación en toda la región cercana al promotor permite observar diferentes patrones que correlacionan con los cambios de expresión génica. Finalmente comentan que a pesar de que han probado varios tamaños de ventana alrededor del TSS, la mayoría de patrones de correlación se han encontrado a una distancia de 5kb del TSS, excepto para aquellos asociados con los "long hypomethylated domains".

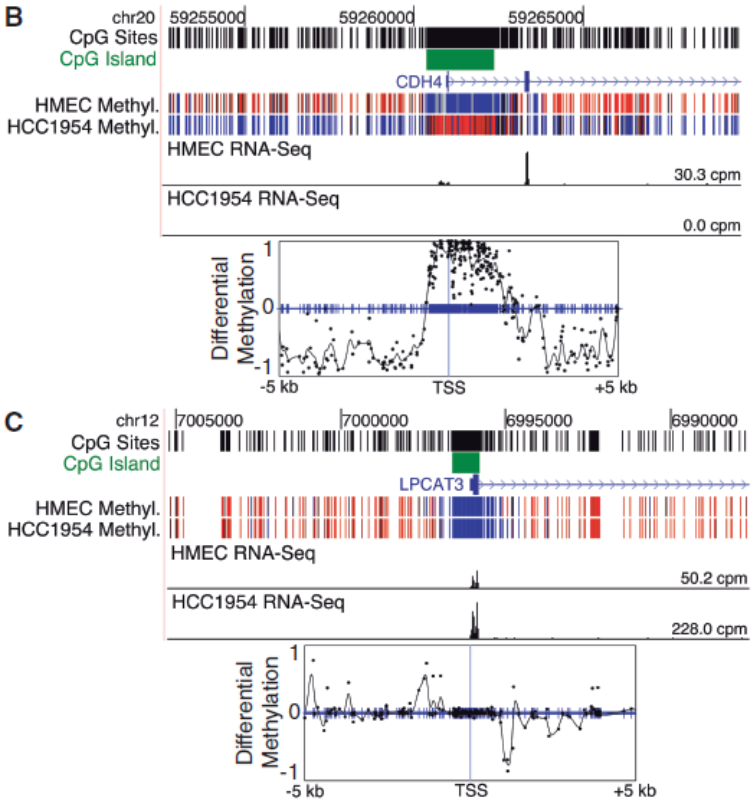


Figura 3. Ejemplo de firma de metilación obtenida en el trabajo de VanderKraats y colaboradores[12].

En otro trabajo de Liu y Qiu[13], exploran la identificación de genes diferencialmente metilados mediante la utilización de "Conditional mutual information". En este trabajo los autores se basan en la premisa de que si un gen está controlado por metilación su expresión es baja cuando está metilado, pero por otro lado cuando está desmetilado su expresión puede ser tanto baja como alta, dependiendo de otros factores, y esto puede variar dependiendo del tipo celular estudiado. Por lo tanto, según esta premisa, al analizar tanto la metilación como la expresión en múltiples tipos celulares, aquellos genes que se regulen por metilación

presentaran una forma en "L", cuando representemos un gráfico con la metilación en el eje de abscisas y la expresión en el eje de ordenadas (Figura 4). Mediante la identificación de genes que presentan esta forma en "L" los autores consiguen identificar genes regulados por metilación. La mayor limitación de esta metodología es que se necesita analizar la expresión y la metilación en muchas muestras de origen diferente para conseguir ver el perfil en "L".

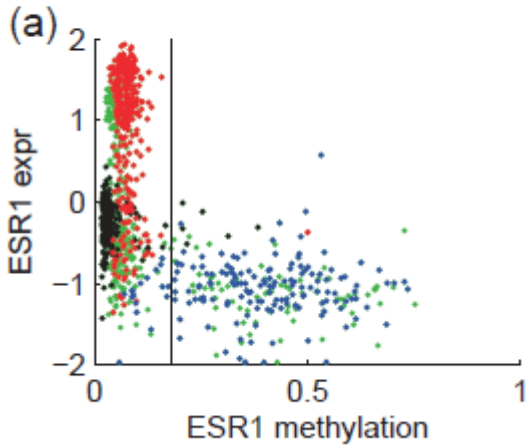


Figura 4. Ejemplo de gen controlado por metilación con patrón en forma de "L". Los diferentes colores indican diferentes tipos de paneles de líneas celulares cancerosas[13].

Actualmente existen diferentes herramientas para analizar e interpretar los datos de metilación, e incluso existen algunas herramientas para poder realizar la integración de ambos tipos de datos (Revisado en[16]). Según el origen de los datos, que pueden provenir de una plataforma basada en secuenciación masiva o basada en arrays de metilación, necesitaremos unas u otras herramientas bioinformáticas para poder realizar los pasos que van desde el procesado a la identificación de DMR. En la Tabla 1, se muestran algunos programas para realizar el pre-procesamiento de los datos de metilación. En el presente proyecto en el que los datos de metilación provenían de arrays 450k de Illumina, hemos utilizado el package `minfi` para el preprocesado de los datos de metilación.

El objetivo del presente proyecto es abordar el tema de la integración de los datos de expresión génica y de metilación, con el objetivo de identificar genes regulados por metilación. Para ello se identificarán genes diferencialmente expresados entre dos condiciones, se realizará el estudio del patrón de metilación de la región donde está incluido cada gen y se analizará si existen diferencias significativas en el grado de metilación de esa región entre los dos grupos estudiados.
Como ejemplo de utilización del pipeline generado, hemos utilizado el estudio de genes diferencialmente expresados entre muestras de tejido tumoral de pacientes con cáncer de pulmón de célula no pequeña (CPCNP) del subtipo histológico adenocarcinoma (ADK) en comparación con los de subtipo histológico escamoso (SCC), de los cuales se encuentran disponibles los datos de metilación y de expresión génica en la base de datos TCGA (http://cancergenome.nih.gov).

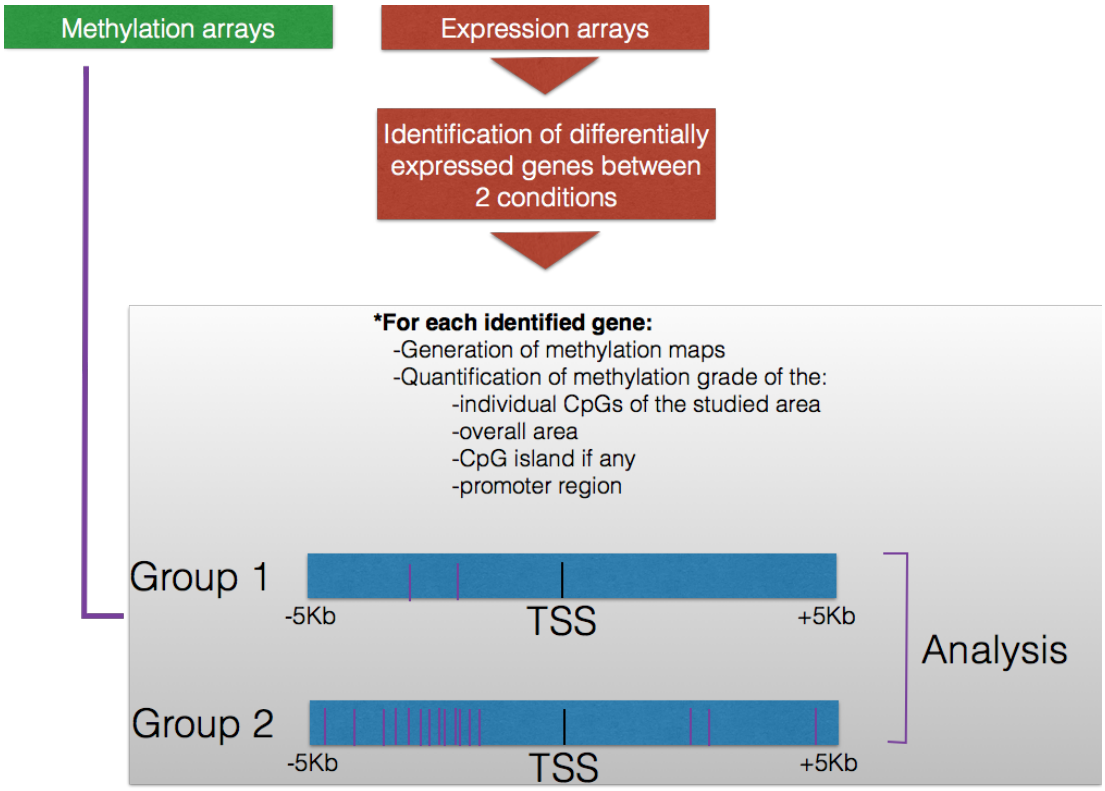
Table 1 | **Software tools for the analysis and interpretation of DNA methylation data**

Software	Description	URL	Refs
Processing bisulphite-sequencing data			
B-SOLANA	Bisulphite aligner for processing bisulphite-sequencing data obtained in the two-base encoding of ABI SOLiD sequencers	http://code.google.com/p/bsolana	40
Bismark	Probably the most widely used three-letter bisulphite aligner; supports both Bowtie (fast, gap-free alignment) and Bowtie 2.0 (sensitive, gapped alignment)	http://www.bioinformatics.babraham.ac.uk/projects/bismark	28
Bis-SNP	Variant caller for inferring DNA methylation levels and genomic variants from bisulphite-sequencing reads that have been aligned by other tools	http://epigenome.usc.edu/publicationdata/bissnp2011	35
BRAT	Highly configurable and well-documented three-letter bisulphite aligner	http://compbio.cs.ucr.edu/brat	29,30
BS-Seeker	Basic three-letter bisulphite aligner based on Bowtie	http://pellegrini.mcdb.ucla.edu/BS_Seeker/BS_Seeker.html	31
BSMAP	Probably the most widely used wild-card bisulphite aligner	http://code.google.com/p/bsmap	21
GSNAP	Wild-card bisulphite aligner included in a widely used general-purpose alignment tool	http://share.gene.com/gmap	22
Last	Recent and well-validated wild-card bisulphite aligner included in a general-purpose alignment tool	http://last.cbrc.jp	23
MethylCoder	Three-letter bisulphite aligner that can be used with either Bowtie (high speed) or GSNAP (high sensitivity)	https://github.com/brentp/methylcode	32
Pash	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://brl.bcm.tmc.edu/pash	24
RMAP	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://www.cmb.usc.edu/people/andrewds/rmap	25
RRBSMAP	Variant of BSMAP that is specialized on reduced-representation bisulphite sequencing (RRBS) data	http://rrbsmap.computational-epigenetics.org	26
segemehl	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://www.bioinf.uni-leipzig.de/Software/segemehl	27
Processing bisulphite microarray data			
ComBat	R script for correcting known or suspected batch effects using an empirical Bayes method	http://www.bu.edu/jlab/wp-assets/ComBat	52
Illumina BeadScan	Machine control and image processing software for Illumina Infinium microarray scanners	http://www.illumina.com/support/array/array_instruments/beadarray_reader.ilmn	
Illumina GenomeStudio	Graphical tool for data normalization, analysis and visualization of Illumina Infinium microarrays (and other genomic data types)	http://www.illumina.com/software/genomestudio_software.ilmn	
isva	R package for batch effect correction using an algorithm that is based on singular value decomposition	http://cran.r-project.org/web/packages/isva	50
methylumi	R/Bioconductor package for Infinium data normalization and general data handling	http://www.bioconductor.org/packages/release/bioc/html/methylumi.html	
minfi	R/Bioconductor package for Infinium data normalization, analysis and visualization	http://www.bioconductor.org/packages/release/bioc/html/minfi.html	
RnBeads	R package providing a software pipeline for Infinium data normalization, quality control, exploratory visualization and differentially methylated region (DMR) identification	http://rnbeads.computational-epigenetics.org	
SVA	R/Bioconductor package for correcting batch effects that are directly inferred from the data using surrogate variable estimation	http://www.bioconductor.org/packages/release/bioc/html/sva.html	53

Tabla 1. Representación parcial de la Tabla 1 de [16].

Descripción del proyecto

En el presente proyecto hemos desarrollado una *pipeline* de análisis en R para identificar si genes diferencialmente expresados entre dos condiciones o grupos de muestras lo están por diferencias en los patrones de metilación entre los 2 grupos comparados.
Muchas veces en el laboratorio obtenemos listados de genes diferencialmente expresados entre dos grupos de pacientes, entre pacientes y controles, o entre líneas celulares tratadas con un determinado fármaco o inhibidas con un siRNA concreto, entre otras. Una vez obtenidos estos listados de genes diferencialmente expresados entre las 2 condiciones comparadas, la siguiente pregunta es a qué se deben las diferencias observadas, y en el contexto del presente proyecto, nos preguntamos si los genes diferencialmente expresados identificados podrían explicarse por diferencias en su patrón de metilación.
Para contestar a ésta pregunta y en base a los antecedentes en relación a la integración de los datos de expresión y metilación, nos planteamos el siguiente diseño:



El tamaño del área de trabajo alrededor del TSS va a ser una variable que se pueda modificar de forma que el investigador pueda explorar varias posibilidades en relación a cada gen. Por defecto el tamaño escogido será de 5kb, ya que en el trabajo de VanderKraats y colaboradores[12], definen que es el tamaño que mejor funciona en la mayoría de casos, excepto para el análisis de regiones reguladas por *long hypomethylated domains* en que sería necesario valorar un área mayor.

Materiales y métodos

Muestras y datos

Con el objetivo de identificar genes diferencialmente expresados regulados por metilación entre los pacientes con cáncer de pulmón del subtipo histológico adenocarcinoma (TCGA: Lung adenocarcinoma) y carcinoma escamoso (TCGA: Lung squamous cell carcinoma) utilizamos datos de expresión génica y de metilación obtenidos del proyecto TCGA (http://cancergenome.nih.gov). En el proyecto existen datos de metilación y de expresión obtenidos por diferentes plataformas, pero para el presente proyecto trabajamos con aquellas muestras que tienen disponibles datos de metilación obtenidos con Illumina Infinium HumanMethylation450 BeadChip, y datos de expresión obtenidos con Agilent 244 Custom Gene Expression G4502A-07.
En total trabajamos con 32 muestras de ADK y 155 muestras de SCC.

Preprocesamiento de los datos de expresión génica

TCGA permite descargarse los datos a 3 niveles diferentes que van desde el nivel 1, que son los datos sin procesar, al nivel 3 que son los datos preprocesados y normalizados de forma estándar. En el caso de los arrays de expresión utilizados que son de tipo custom, la mejor opción fue descargarse los datos de nivel 3. Estos datos se descargan en forma de un fichero para cada muestra de tipo `.txt` delimitado por tabulaciones, que contiene los nombres de los genes y su valor de expresión ya normalizado. Una vez descargados los ficheros procedemos a crear un `data.frame` que contenga los datos de expresión de todas las muestras con las que se va trabajar. Una vez generado el `data.frame`, procedemos a crear un objeto `expressionSet`, con el cual poder realizar posteriormente todos los análisis necesarios. Creamos también un fichero que identifica el subtipo histológico de cada muestra, y lo cargamos como `phenoData` del objeto `expressionSet`.
El script que realiza el procesamiento de los datos de expresión se puede ver en el Anexo 1.

Preprocesamiento de los datos de metilación

En el caso de los arrays de metilación, estos no son de tipo custom, por lo que nos da mayor libertad descargarnos los datos de nivel 1 de TCGA, que en este caso son archivos de tipo `.idat` que contiene los datos brutos del array de metilación. Los arrays Infinium Human Methylation450 Bead Chip nos permiten interrogar 485.000 CpGs por muestra y cubre el 99% de los genes de RefSeq, con un promedio de 17CpGs por región génica distribuidos a lo largo del promotor, la región 5'UTR, el primer exón, el cuerpo del gen y la región UTR 3'. Además, cubre el 96% de las islas CpGs, y también tiene cobertura de las regiones flanqueantes de las islas (p.e. shores).

El preprocesado de los arrays lo realizamos utilizando el package `minfi` de Bioconductor. Cargamos los datos mediante la función `read.450()` y preprocesamos los arrays con la función `preprocessIllumina()` del package `IlluminaHumanMethylation450kmanifest`. Para obtener las coordenadas de las diferentes CpG incluidas en el array en relación con el genoma utilizamos la función `mapToGenome()` del package `IlluminaHumanMethylation450kanno.ilmn12.hg19`, que contiene los datos de anotación. A continuación procedemos a obtener los valores de metilación y a trabajar con los datos disponibles. El script que realiza el procesamiento de los datos de metilación se puede ver en el anexo 2.

Obtención de coordenadas genómicas de los genes y posiciones de las islas CpG

Una vez identificados los genes diferencialmente expresados, se procede a obtener las localizaciones genómicas de dichos genes para poder analizar la metilación de dicha región. Para ello se utiliza el package `biomaRt`, y las anotaciones de Ensembl.

Para la obtención de las coordenadas de las islas CpG, se utilizan los datos proporcionados por el laboratorio del Dr. Irizarry (http://rafalab.jhsph.edu/CGI/model-based-cpg-islands-hg19.txt) creados usando el método descrito en el trabajo de Wu y colaboradores[18] donde identifican las islas CpG utilizando los modelos de Markov.

Análisis estadístico

Para la identificación de genes diferencialmente expresados se han utilizado los modelos lineales implementados en el package `limma` de Bioconductor.
Para el estudio de las diferencias en los niveles de metilación entre los dos grupos en una región determinada se ha realizado una T-Student. Cuando se han realizado múltiples comparaciones (por ejemplo cuando se han valorado las distintas CpG presentes en la región individualmente) se ha realizado una corrección para comparaciones múltiples utilizando la función `p.adjust()` del package `stats`. Se ha realizado el ajuste con la opción `fdr` (*false discovery rate*).
El presente estudio se ha realizado utilizando RStudio versión 0.99.446 con R 3.2.0 en el SO OS X Yosemite versión 10.10.3.

Resultados

Identificación de genes diferencialmente expresados entre ADK y SCC

El análisis de los datos de expresión génica con modelos lineales (limma) nos permite identificar los 10 genes diferencialmente expresados más significativos entre las muestras de ADK y de SCC. A continuación podemos ver los datos de los 10 genes identificados:

	logFC	AveExpr	t	P.Value	adj.P.Val	B
FMO5	3.336272	-0.53763764	15.62023	7.456085e-36	4.584746e-32	70.69338
DSC3	-5.053050	3.60958601	-14.48509	1.819152e-32	5.592983e-29	63.03728
KRT5	-5.530429	1.62465971	-14.42316	2.788079e-32	5.714632e-29	62.61796
ACOX2	2.636997	-0.78858467	12.87407	1.229727e-27	1.890398e-24	52.11060
UCK2	-1.718109	-0.19925312	-12.56811	1.013036e-26	1.244159e-23	50.03794
SPP2	2.086642	0.82080000	12.54182	1.214011e-26	1.244159e-23	49.86005
KRT6C	-4.946082	2.96231854	-12.45944	2.140363e-26	1.750064e-23	49.30268
ST3GAL5	2.298681	0.08203573	12.45046	2.276876e-26	1.750064e-23	49.24190
KCNK5	2.334337	-0.88435740	12.26554	8.120324e-26	5.547986e-23	47.99200
IL1F7	2.389384	0.68469786	12.24457	9.378528e-26	5.766857e-23	47.85040

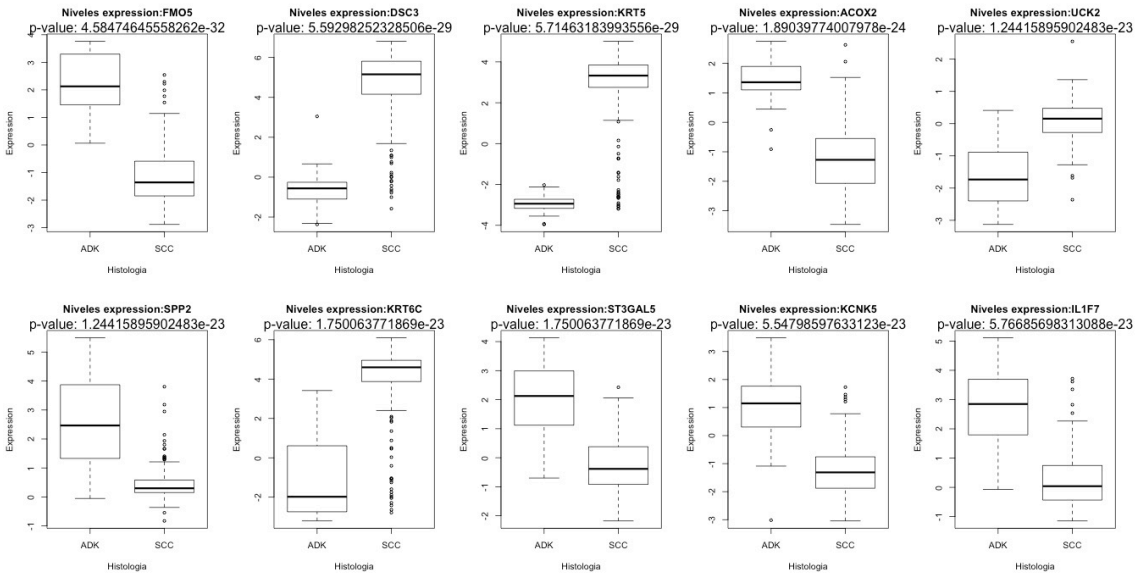


Figura 5. Boxplots mostrando los niveles de expresión de los 10 genes identificados en los dos subtipos histológicos de CPCNP. Se muestra la p-value ajustada obtenida en el análisis realizado con *limma*.

Si realizamos un *hierarchical cluster* utilizando los valores de expresión de los 10 genes identificados, vemos que todos los pacientes del subtipo ADK se agrupan en el mismo clúster (Figura 6, pacientes en rojo).

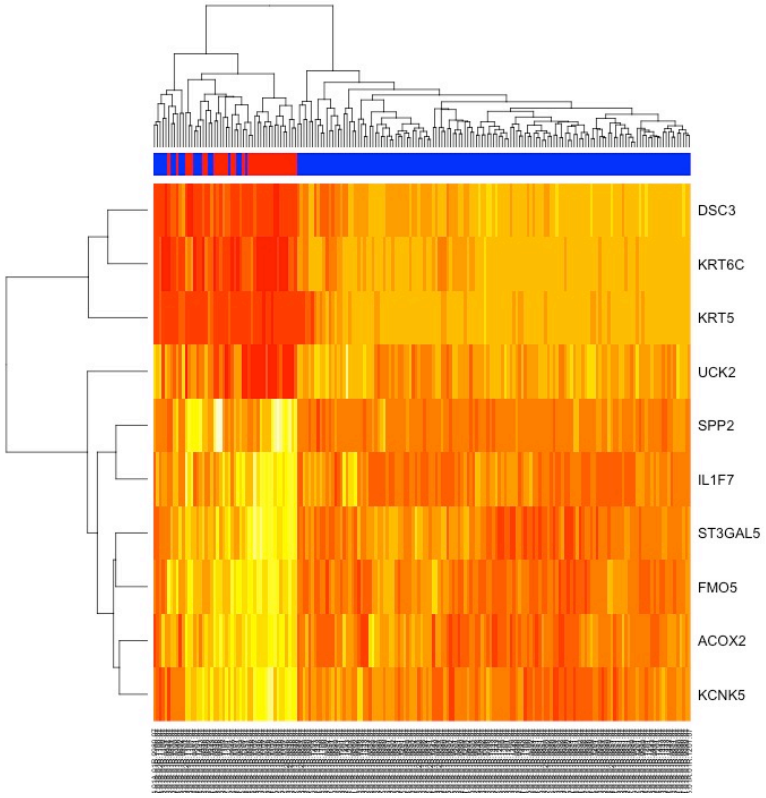


Figura 6. *Hierarchical cluster* utilizando los valores de expresión de los 10 genes identificados. En rojo se indican los pacientes del subtipo ADK y en azul los pacientes del subtipo SCC.

A continuación obtengo mediante el uso de `bioMart()`, las coordenadas de los genes identificados para pasárselas al programa de estudio de la metilación:

	hgnc_symbol	chromosome_name	start_position	end_position
1	ACOX2	3	58505136	58537319
2	DSC3	18	30990008	31042815
3	FMO5	1	147175351	147243050
4	IL37	2	112912971	112918882
5	KCNK5	6	39188973	39229450
6	KRT5	12	52514575	52520687
7	KRT6C	12	52468516	52473785
8	SPP2	2	234050679	234077134
9	ST3GAL5	2	85839144	85889014
10	UCK2	1	165827531	165911618

Estudio del grado de metilación en los genes identificados diferencialmente expresados entre ADK y SCC

Para realizar el estudio de la posible regulación por metilación de cada uno de los genes estudiados se han realizado representaciones gráficas donde se incluye un esquema de la región seleccionada (-5Kb del inicio del gen - +5Kb del final del gen), con información sobre las posibles islas CpG presentes, las sondas incluidas en el array para dicha región y el grado de metilación para cada una de ellas (Figura 7).

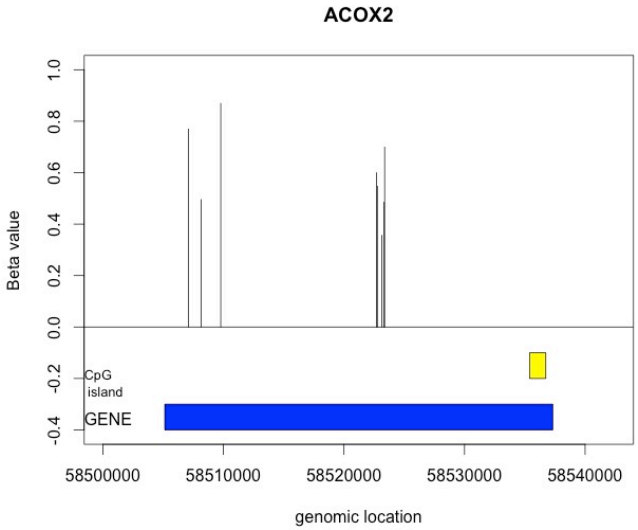


Figura 7. Representación de la región incluida para el gen ACOX2 de una muestra. Las cajas amarillas indican las islas CpG presentes en la región. La caja azul indica la posición del gen. Cada barra vertical es una sonda que cuantifica una CpG específica, y su grado de metilación está indicado por la altura de la barra mediante el Beta value (B-value=1-Completamente Metilado; B-value=0: completamente desmetilado).

Para el estudio del grado de metilación de los genes seleccionados hemos analizado si existen diferencias significativas en los niveles de metilación de las CpG de la región para las cuales tenemos información(es decir, que exista sonda en el array) a 3 niveles diferentes:

1. Cuantificación de los niveles globales de la región seleccionada. Para ello hacemos promedio de los valores de todas las sondas y valoramos si existen diferencias significativas entre ADK y SCC.
2. Cuantificación de los niveles de metilación de las sondas situadas en lo que consideramos la zona promotora (hasta 2Kb antes de donde encontramos el inicio de transcripción del gen [TSS]).
3. Cuantificación de los niveles de metilación de las posibles islas CpGs que existan en la región seleccionada.

A continuación muestro varios ejemplos de los resultados obtenidos en cada una de las comparaciones y al final se muestra una tabla resumen de todos los genes estudiados. Debido a las características de cada región estudiada, ha habido algunas comparaciones que no se han podido realizar en algunos genes.

Con el objetivo de mostrar todas las muestras analizadas en una sola figura para cada gen, los valores de metilación de cada sonda se muestran con puntos en lugar de con columnas, y en rojo se indican las muestras de tipo SCC y en negro las del tipo ADK (Figura 8).

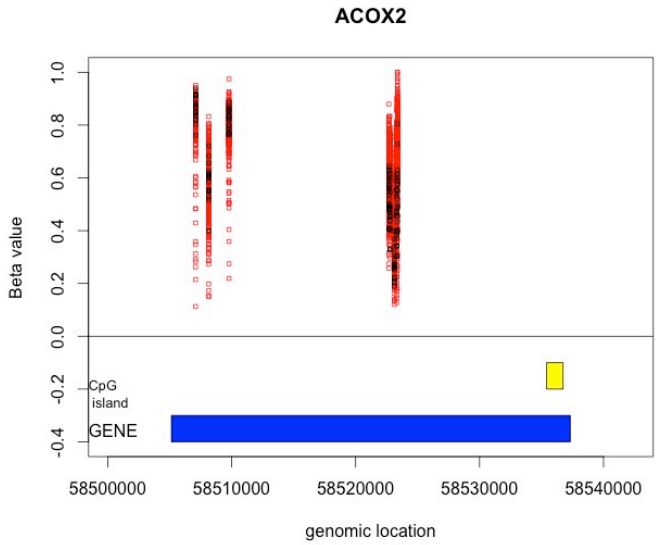


Figura 8. Mapa de metilación de la región seleccionada para el gen ACOX2.

En la siguiente página podemos ver los mapas de metilación de todos los genes estudiados (Figura 9). Se puede observar que cada gen tiene sus peculiaridades, y a pesar de que hemos escogido arbitrariamente un tamaño de 5Kb alrededor del gen, probablemente habría que estudiar cada caso particularmente con el objetivo de ver si otros tamaños de región podrían aportar mayor información. Además, vemos que el número de sondas y su distribución en cada uno de los genes varía considerablemente, probablemente uno de los motivos es que la distribución de CpGs en los diferentes genes no es la misma, pero por otro lado, en los arrays de metilación no todas las regiones están igual de cubiertas por sondas, lo que sería otra explicación a las diferencias observadas entre genes.

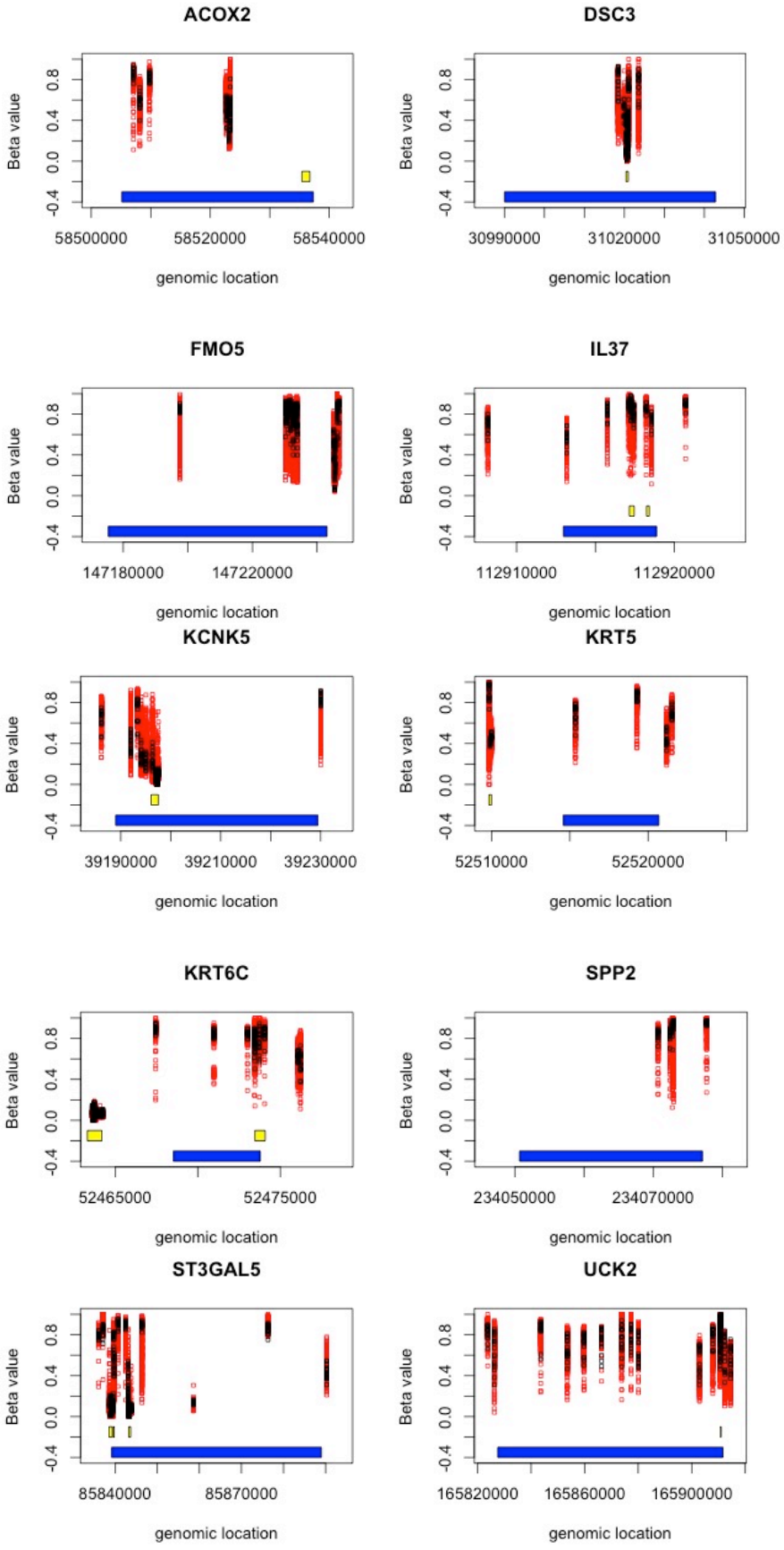


Figura 9. Mapas de metilación de los diferentes genes estudiados.

Para cada uno de los genes analizamos también si existen diferencias significativas entre los niveles de metilación en ADK y SCC para cada una de las sondas incluidas en la región de estudio. En la Figura 10 podemos ver las diferentes sondas representadas en la región del gen ACOX2, y como se observa podemos distinguir tres patrones: sondas en las que no observamos diferencias significativas entre ADK y SCC (p.e. la sonda cg16869862); sondas cuyo grado de metilación está infraregulado en ADK vs SCC (p.e. la sonda cg02259384); y sondas con un grado de metilación sobreregulado en ADK vs SCC (p.e. la sonda cg23652987). En este caso tenemos que recordar que el gen ACOX2 estaba sobreexpresado en ADK vs SCC, por lo que para considerar que podría estar regulado por metilación lo ideal sería encontrar que la mayoría de sondas están inframetiladdas en ADK vs SCC. En este ejemplo, 5/8 (62.5%) sondas están inframetiladas en ADK, 2/8 (25%) están sobremetiladas, y en 1/8 no existen diferencias significativas.

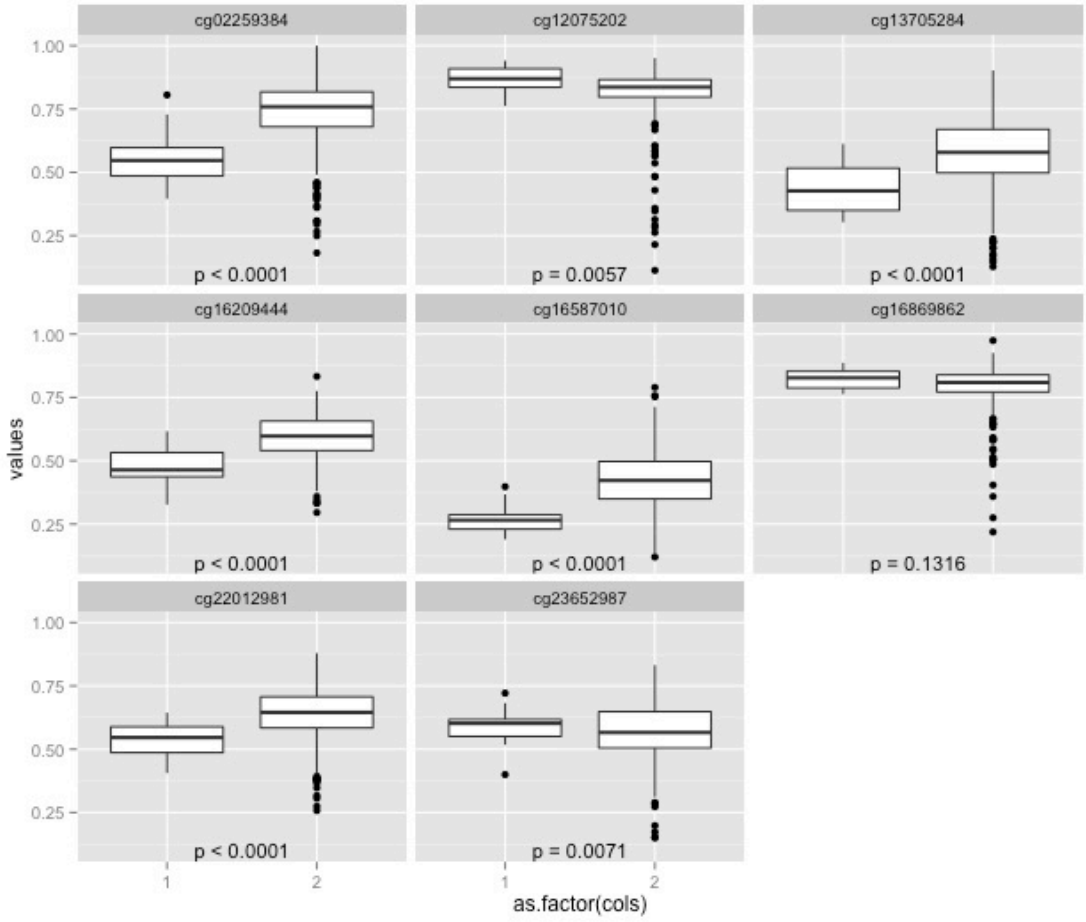


Figura 10. Comparación del grado de metilación en ADK vs SCC para cada una de las sondas incluidas en la región de estudio del gen ACOX1. En el eje de las x, el 1 significa ADK y el 2 significa SCC.

El estudio de los valores de metilación sonda a sonda nos muestran que no todas las islas CpG de una región pueden relacionarse de la misma manera con la expresión génica, y por lo tanto no todas serían valorables para ser incluidas en la correlación con los niveles de expresión.

1. Cuantificación de los niveles globales de metilación de la región seleccionada

Una de las primeras medidas del grado de metilación que hemos analizado es la cuantificación conjunta de todas las sondas incluidas en la región de estudio en cada uno de los grupos. En la Figura 11, podemos ver que cuando cuantificamos de esta forma el grado de metilación para el gen ACOX2 observamos que hay niveles significativamente más bajos de metilación en el grupo ADK que en el grupo SCC. Esto cuadraría con la observación de niveles más altos de expresión del gen ACOX2 en los ADK.

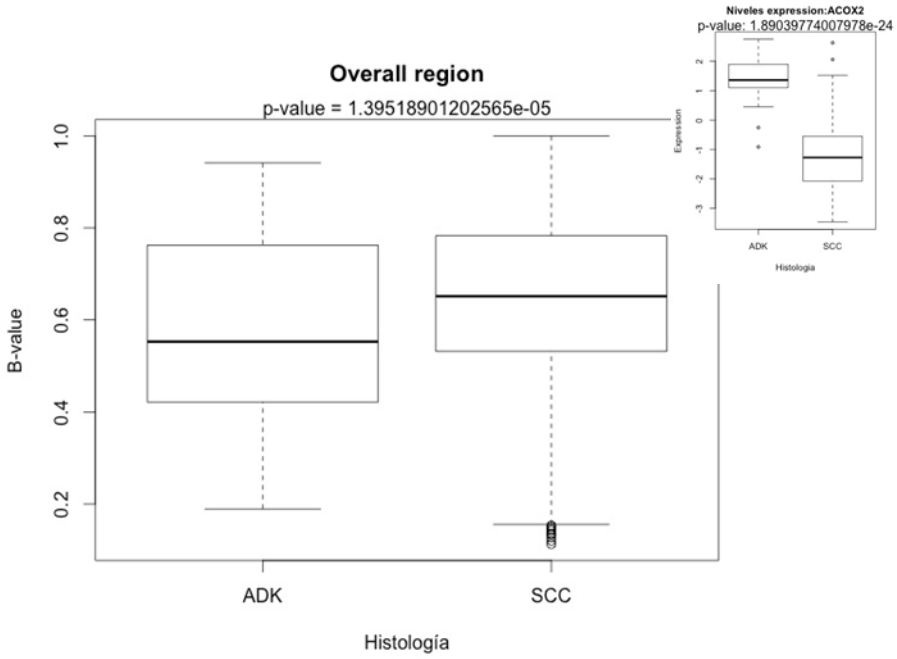


Figura 11. Boxplots mostrando los niveles de metilación para ADK y SCC teniendo en cuenta todas las sondas incluidas en la región de estudio para ACOX2. En la parte superior derecha podemos observar que los niveles de expresión del gen ACOX2 son inversos al patrón de metilación.

2. Cuantificación de los niveles de metilación de las sondas situadas en lo que consideramos la zona promotora (hasta 2Kb antes de donde encontramos el inicio de transcripción del gen [TSS]).

A continuación analizamos el grado de metilación de sondas situadas en la zona promotora del gen. Considerando como zona promotora 2kb upstream del inicio de transcripción (TSS), pero este valor se puede cambiar al ejecutar el script. Como podemos ver en la Figura 8, en esa localización no encontramos ninguna sonda que podamos evaluar en el caso del gen ACOX2, por lo tanto a continuación muestro el resultado obtenido con el gen KCNK5 que sí que presenta sondas en esa región, aunque tenemos que ajustar un poco el tamaño a 3Kb, ya que la sonda se encuentra situada entre 3 y 2 Kb (Figura 12).

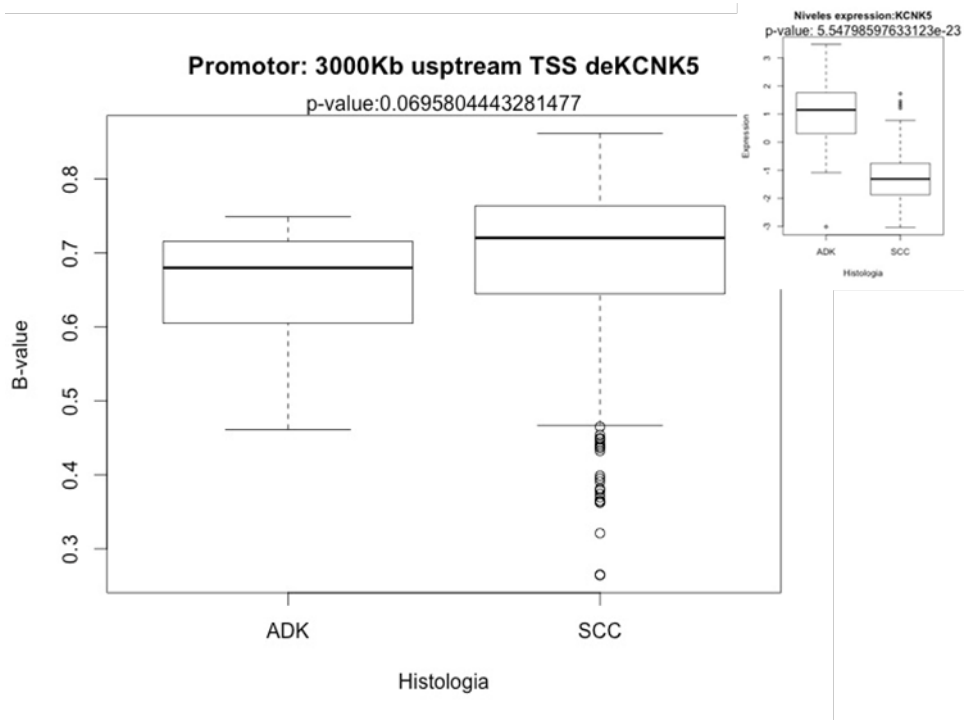


Figura 12. Boxplots mostrando los niveles de metilación para ADK y SCC teniendo en cuenta todas las sondas incluidas a 3Kb del TSS del gen KCNK5. En la parte superior derecha podemos observar que los niveles de expresión del gen KCNK5 son inversos al patrón de metilación observado en la zona promotora.

3. Cuantificación de los niveles de metilación de las posibles islas CpGs que existan en la región seleccionada.

A continuación analizamos aquellas sondas situadas sobre una isla CpG (pueden ser varias). Al realizar este análisis hemos de tener varias consideraciones en cuenta:

a. Podemos no tener ninguna isla CpG en la región de estudio (p.e. el caso del gen FMO5)
b. Podemos tener islas CpGs pero no sondas situadas sobre ellas (p.e. el caso del gen ACOX2)
c. La localización de la isla CpG puede influenciar en su papel en la regulación de la expresión génica (p.e. el caso del gen KRT6C).

A continuación mostramos los valores de cuantificación de las 2 islas CpG presentes en el gen KRT6C (Figura 13-14). Como podemos ver en la Figura 13, la primera isla está situada en la región 5’ del gen cerca de su zona promotora y por tanto quizá está sería la que podría tener mayor influencia en la regulación de la transcripción. La segunda isla CpG se encuentra situada en la región 3’ del gen.

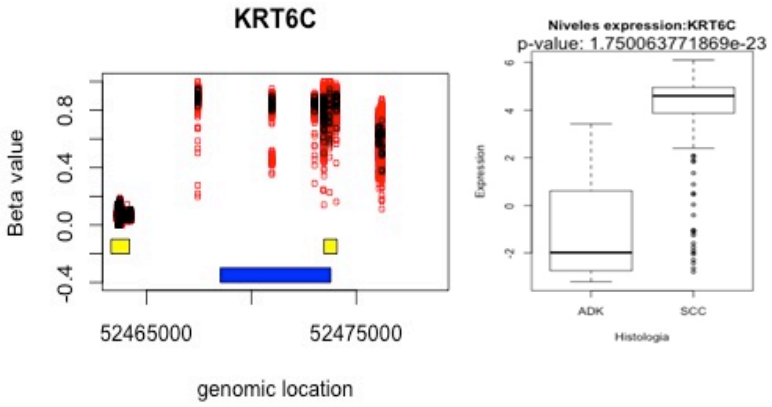


Figura 13. Mapa de metilación del gen KRT6C donde podemos observar la presencia de dos islas CpGs y boxplot mostrando los niveles de expresión.

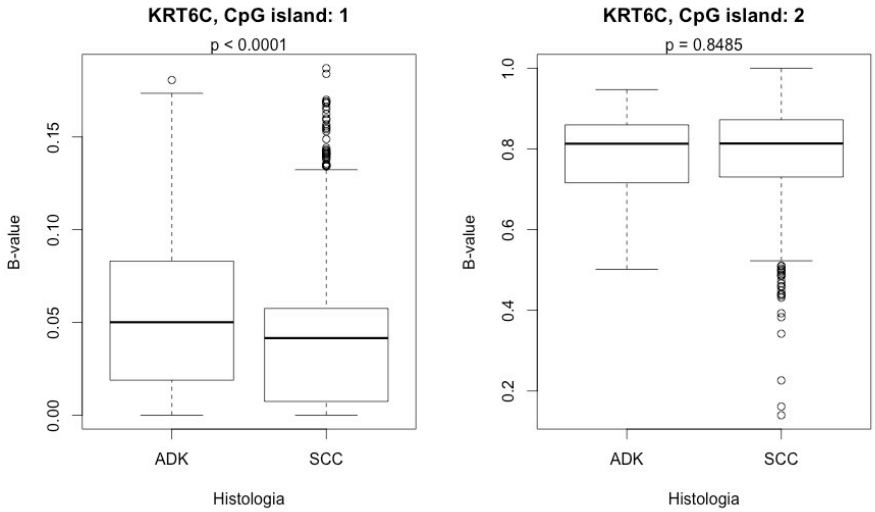


Figura 14. Boxplots mostrando los niveles de metilación en ADK y SCC para cada una de las 2 islas CpGs presentes en la región del gen KRT6C. La isla 1 está situada a 5' del gen y la isla 2 está situada a 3' del gen.

Como observamos en la Figura 14, la isla CpG situada en la región 5' del gen sí que presenta diferencias significativas entre los 2 grupos estudiados que correlacionan de forma negativa con el grado de expresión del gen, pero en cambio no se observan diferencias cuando analizamos la isla CpG situada a 3' del gen.

Resumen de los resultados de los diferentes análisis realizados a los 10 genes

A continuación en la Tabla 2 se muestra un resumen de los análisis realizados con los diferentes genes y de la conclusión obtenida en relación a si las diferencias observadas entre ambos grupos son o no debidas a diferencias en el patrón de metilación.

Gen	Nivel de expresión en ADK	Grado de metilación región ADK	Grado de metilación Promotor ADK (2kb)	Grado de metilación Islas CpGs ADK	Conclusión: Regulado por metilación
ACOX2	Alto	Bajo	NA	NA	SI
DSC3	Bajo	Alto	NA	Alto	SI
FMO5	Alto	Alto	NA	NA	NO
IL37	Alto	Alto	NA	Alto	NO
KCNK5	Alto	=	NA-bajo a 3000	=	SI
KRT5	Bajo	Alto	NA	Alto	SI
KRT6C	Bajo	=	=	Isla 1: Alto Isla 2: =	SI
SPP2	Alto	Alto	NA	NA	NO
ST3GAL5	Alto	=	Alto	Isla 1: Alto 2 y 3:=	NO
UCK2	Bajo	Alto	=	Alto	SI

Tabla 2. Resumen resultados de los análisis realizados. NA, significa que no se ha podido realizar dicho análisis. La referencia Alto/Bajo en ADK siempre es en comparación con los niveles de SCC. = significa que no existen diferencias significativas.

Ejemplos de genes cuya expresión no parece estar regulada por metilación en los grupos estudiados

En los apartados previos hemos mostrado los diferentes análisis realizados para determinar si los genes identificados diferencialmente expresados pueden explicarse por diferencias en su patrón de metilación en los 2 grupos. Los casos utilizados como ejemplos son casos que consideramos que su regulación es por metilación. A continuación muestro algunos ejemplos de genes que consideramos que su diferencia no se explica por metilación, bien porque no veamos diferencias significativas, bien porque veamos una correlación directa con los niveles de metilación que no explicarían las diferencias observadas.

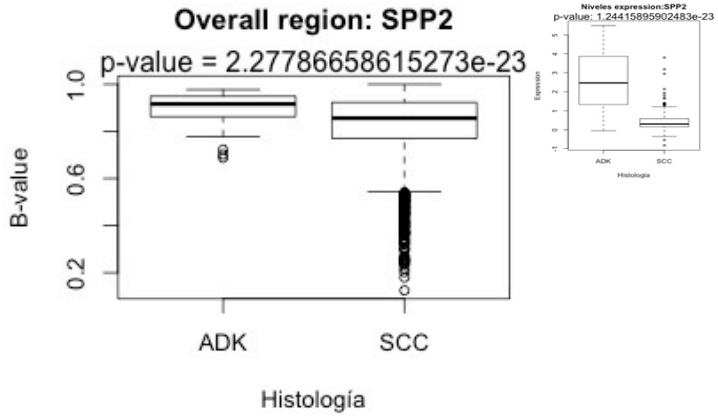


Figura 15. Los niveles de metilación de la región del gen SPP2 en ADK son altos al igual que los niveles de expresión.

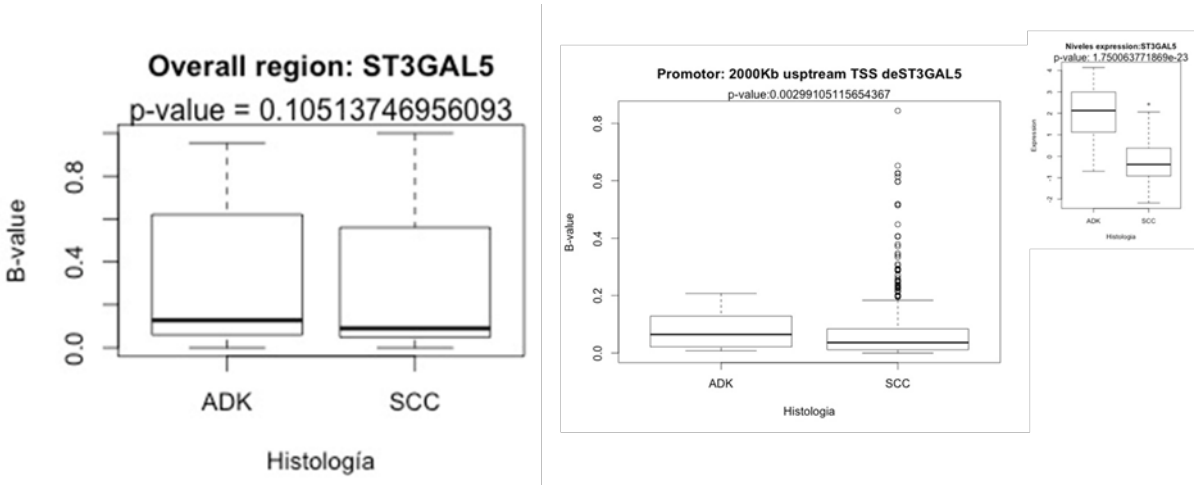


Figura 16. El Gen ST3GAL5 no muestra diferencias globales de metilación entre ADK y SCC, y cuando subanalizamos el promotor este presenta correlación directa.

Discusión

En el presente proyecto hemos integrado datos de expresión génica y datos de metilación con el objetivo de identificar genes diferencialmente expresados entre dos grupos de muestras (pacientes con CPCNP del subtipo ADK y pacientes con el subtipo SCC) cuya expresión diferencial pudiera explicarse por metilación del ADN. El principal problema que nos encontramos cuando estudiamos metilación con técnicas de *high throuput* es que no siempre es posible obtener un listado de genes diferencialmente metilados, como se realizaría en un análisis estándar de expresión, sino que lo que se suelen identificar son grandes regiones diferencialmente metiladas(*"Differential methylated regions"*)[19]. Cuando nos disponemos a explorar una región pequeña, en nuestro caso la correspondiente a un gen, nos encontramos con diferentes problemas a la hora de poder decir si dicho gen se encuentra metilado o no. Entre ellos, el hecho de que no todas las CpG tienen la misma función en relación con la regulación transcripcional de dicho gen. De ahí la pregunta clave: ¿Qué CpG debo estudiar para cada gen que sean informativas del grado de expresión de dicho gen? Con el objetivo de responder a esta pregunta múltiples autores han estudiado diferentes regiones y han empleado múltiples soluciones[9-13]. En nuestro caso, hemos realizado una aproximación que combina varias de las aproximaciones realizadas previamente en la literatura. En primer lugar hemos analizado la metilación de una región que incluye al gen y 5Kb por encima y 5Kb por debajo de este, y hemos analizado las diferencias cuantitativas observadas en relación a todas las sondas disponibles en el array de metilación para esta región entre los dos grupos estudiados. Además hemos realizado dos mediciones más, una relacionada con la zona promotora del gen y otra relacionada con las posibles islas CpGs presentes en la región. Aquí nos encontramos que cada gen no tiene la misma representación de sondas, lo que complica en algunos casos la obtención de conclusiones. Nuestro objetivo ha sido integrar las tres mediciones para poder decidir si las diferencias observadas entre ambos grupos son atribuibles a diferencias en el grado de metilación de los genes. De esta forma hemos identificado 6 genes de los 10 estudiados cuya diferencia vendría explicada por diferencias en los patrones de metilación, aunque habría que realizar una validación experimental para poder validar las observaciones computacionales. Cuando estudiamos si existen evidencias en la literatura de si los 6 genes identificados se han descrito como regulados por metilación en otras patologías o en CPCNP, nos encontramos en que algunos de ellos ya se habían validado previamente, lo que nos refuerza las conclusiones obtenidas en el presente estudio. Por ejemplo, el gen UCK2 se ha observado hipermetilado en líneas celulares de cáncer de tiroides[20]. El gen DSC3 se ha visto previamente regulado por metilación en cáncer de adenocarcinoma de esófago[21]. Además, hemos encontrado una patente en la que se incluyen los genes KCNK5 y KRT5 como genes cuya metilación puede ser utilizada como biomarcador cancerígeno[22]. Además, el gen KRT5 también se describe diferencialmente metilado en células epiteliales de la vía aérea en comparación con células de sangre periférica[23]. Con todo ello vemos que los genes que hemos identificado regulados por metilación en ADK y SCC, son genes en los que

previamente se ha descrito que la metilación juega un papel en la regulación de su expresión génica en otras patologías.
Finalmente comentar que es muy difícil estandarizar una metodología para escanear de forma automática todo el genoma para identificar aquellos genes regulados por metilación cuya expresión correlacione, por el simple hecho de que cada región tiene sus propias características. En el presente estudio, cuando hemos realizado los mapas de metilación que nos permiten visualizar la región de estudio hemos podido observar que cada gen tenía sus propias características y que necesitaba de un análisis pormenorizado. Nos hemos encontrado genes con muchas sondas disponibles a lo largo de su superficie, otros con islas CpG cerca o encima, otros sin islas CpG, etc. A pesar de que lo ideal, desde nuestro punto de vista seria poder analizar las diferentes regiones por separado y obtener respuestas similares para cada una de ellas en relación a la expresión observada, ya hemos visto que en nuestro caso solo en 3 genes de los 10 estudiados hemos podido aplicar los 3 análisis (el global, el del promotor y el de las islas CpGs). Además, en algunos casos hemos tenido que tomar decisiones como la de incrementar el tamaño del área considerada como promotor para poder incluir sondas limitantes que nos permitieran el análisis (KCNK5), etc.

En conclusión podemos decir que el análisis integrado de la metilación actualmente todavía necesita de un estudio detallado/individualizado a pesar de que con procedimientos automáticos de análisis podemos acotar el número de genes con los que trabajar, o como en el presente proyecto hemos realizado, se podría reducir el número de genes a estudiar mediante selección previa de candidatos diferencialmente expresados que entre otras cosas nos permitan aumentar el poder estadístico de los análisis y utilizar técnicas de análisis más simplificadas para identificar aquellos genes modulados por metilación.

Bibliografía

1. SL Berger, T Kouzarides, R Shiekhattar, A Shilatifard: **An operational definition of epigenetics**. *Genes & development* 2009, **23**:781-783.
2. CH Waddington: **An introduction to modern genetics**. *An Introduction to Modern Genetics.* 1939.
3. M Rodríguez-Paredes, M Esteller: **Cancer epigenetics reaches mainstream oncology**. *Nature medicine* 2011:330-339.
4. A Portela, M Esteller: **Epigenetic modifications and human disease**. *Nature biotechnology* 2010, **28**:1057-1068.
5. M Gardiner-Garden, M Frommer: **CpG islands in vertebrate genomes**. *Journal of molecular biology* 1987, **196**:261-282.
6. S Saxonov, P Berg, DL Brutlag: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:1412-1417.
7. RA Irizarry, C Ladd-Acosta, B Wen, Z Wu, C Montano, P Onyango, H Cui, K Gabo, M Rongione, M Webster: **The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores**. *Nature genetics* 2009, **41**:178-186.
8. M Esteller: **Cancer epigenomics: DNA methylomes and histone-modification maps**. *Nature Reviews Genetics* 2007, **8**:286-298.
9. M Li, C Balch, JS Montgomery, M Jeong, JH Chung, P Yan, TH Huang, S Kim, KP Nephew: **Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer**. *BMC medical genomics* 2009, **2**:34.
10. R Shaknovich, H Geng, NA Johnson, L Tsikitas, L Cerchietti, JM Greally, RD Gascoyne, O Elemento, A Melnick: **DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma**. *Blood* 2010, **116**:e81.
11. CGAR Network: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**:609-615.
12. ND VanderKraats, JF Hiken, KF Decker, JR Edwards: **Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes**. *Nucleic acids research* 2013:gkt482.
13. Y Liu, P Qiu: **Integrative analysis of methylation and gene expression data in TCGA**. In: *Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on*; 2012. 1-4.
14. JT Bell, AA Pai, JK Pickrell, DJ Gaffney, R Pique-Regi, JF Degner, Y Gilad, JK Pritchard: **DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines**. *Genome Biol* 2011, **12**:R10.
15. S Yoo, S Takikawa, P Geraghty, C Argmann, J Campbell, L Lin, T Huang, Z Tu, R Feronjy, A Spira: **Integrative Analysis of DNA Methylation and Gene Expression Data Identifies EPAS1 as a Key Regulator of COPD**. *PLoS genetics* 2015, **11**:e1004898.

16. C Bock: **Analysing and interpreting DNA methylation data**. *Nature Reviews Genetics* 2012, **13**:705-719.
17. J-K Rhee, K Kim, H Chae, J Evans, P Yan, B-T Zhang, J Gray, P Spellman, TH-M Huang, KP Nephew: **Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer**. *Nucleic acids research* 2013, **41**:8464-8474.
18. H Wu, B Caffo, HA Jaffee, RA Irizarry, AP Feinberg: **Redefining CpG islands using hidden Markov models**. *Biostatistics* 2010:kxq005.
19. VK Rakyan, TA Down, DJ Balding, S Beck: **Epigenome-wide association studies for common human diseases**. *Nature Reviews Genetics* 2011, **12**:529-541.
20. P Hou, D Liu, M Xing: **Genome-wide alterations in gene methylation by the BRAF V600E mutation in papillary thyroid cancer cells**. *Endocrine-related cancer* 2011, **18**:687-697.
21. Q Wang, D Peng, S Zhu, Z Chen, T Hu, M Soutto, R Saad, S Zhang, W Ei-Rifai: **Regulation of Desmocollin3 Expression by Promoter Hypermethylation is Associated with Advanced Esophageal Adenocarcinomas**. *Journal of Cancer* 2014, **5**:457.
22. AM Chinnaiyan, MS Dhanasekaran, J Kim: **Dna methylation profiles in cancer**. In. City: Google Patents; 2013.
23. D Stefanowicz, T-L Hackett, FS Garmaroudi, OP Günther, S Neumann, EN Sutanto, K-M Ling, MS Kobor, A Kicic, SM Stick: **DNA methylation profiles of airway epithelial cells and PBMCs from healthy, atopic and asthmatic children**. 2012.

Apéndice 1

```
#---
#Pipeline analisis LEVEL 3 arrays Agilent from TCGA
#---

#Leemos los archivos de expresión de ADK y de SCC de nivel 3

dir_ADK<-
"/Users/anavarroponz/Documents/Proyecto_UOC/Archivos/ADK/Expression_
ADK/Expression-Genes/UNC__AgilentG4502A_07_3/Level_3"

dir_SCC<-
"/Users/anavarroponz/Documents/Proyecto_UOC/Archivos/SCC/Expression_
_SCC/Expression-Genes/UNC__AgilentG4502A_07_3/Level_3"

#guardamos los nombres de los archivos en una variable
temp_ADK<-list.files(path=dir_ADK, pattern="*.txt")
temp_SCC<-list.files(path=dir_SCC, pattern="*.txt")

#file1<-read.table(temp[1], header = TRUE, skip=1, sep="\t")

setwd(dir_ADK)
listOfFiles_ADK <- lapply(temp_ADK, function(x) read.table(x, header
= TRUE, sep="\t"))
setwd(dir_SCC)
listOfFiles_SCC <- lapply(temp_SCC, function(x) read.table(x, header
= TRUE, sep="\t"))



out_ADK<-Reduce(merge, listOfFiles_ADK)
out_SCC<-Reduce(merge, listOfFiles_SCC)

out_conjunto<-merge(out_ADK, out_SCC, by="Hybridization.REF")

array<- data.frame(out_conjunto[,-1], row.names=out_conjunto[,1])
dim(array)
#creo un fichero que indique que muestras son ADK y cuales SCC

num_samples_ADK<-dim(out_ADK)[2]-1
num_samples_SCC<-dim(out_SCC)[2]-1

datos<-array(c(colnames(array), rep(c(1,2),c(32,155))),
dim=c(187,2))
head(datos)


datos<- data.frame(datos[,-1], row.names=datos[,1])
colnames(datos)=c("Type")
summary(datos)


#out <- Reduce(function(x,y) {merge(x,y, by =
"Composite.Element.REF")}, listOfFiles)

#juntado<-merge(temp[1],temp[2])

#name.list<-lapply(temp, read.delim)
```

```
#name.list

#Ahora convertimos el array (tipo data.frame) en un expressionSet
para posteriormente poder realizar diferentes analisis


biocLite("Biobase")
library("Biobase")
#biocLite("convert")
#library("convert")
#as(out, "ExpressionSet")

pData<-datos
all(rownames(pData)==colnames(exprs))
phenoData <- new("AnnotatedDataFrame", data=pData)
out_exprs<-as.matrix(array)
class(out_exprs)<-"numeric"
minimalset<-ExpressionSet(assayData = out_exprs, phenoData =
phenoData)



#Ara que ja tinc un expressionset puc realitzar qualsevol comparació
utilitzant limma o altres
biocLite("affy")
require(affy)
boxplot(exprs(minimalset))

dist(t(exprs(minimalset)))
hclust(minimalset)
biocLite("genefilter")
library("genefilter")
#biocLite("hgu95av2.db")
#library("hgu95av2.db")
#bioclite("org.Hs.eg.db")
#ans <- nsFilter(minimalset)

#construyo el modelo para limma

biocLite("limma")
library("limma")
design<-pData
summary(design)

design<-cbind(ADK=c(rep(1, 32),rep(0, 155)), SCC=c(rep(0,32),
rep(1,155)))
rownames(design)<-rownames(pData)

fit<-lmFit(minimalset, design)
cont.matrix<- makeContrasts(ADKvsSCC=ADK-SCC, levels=design)
fit2<- contrasts.fit(fit, cont.matrix)
fit2<-eBayes(fit2)
resultat<-topTable(fit2,adjust.method="BH")

llista_de_gens<-rownames(resultat)

expression_data_top<-exprs(minimalset[llista_de_gens,])
expression_data_top<-t(expression_data_top)
expression_data_top<-cbind(expression_data_top, datos)


#realizamos boxplots de todos los genes identificados
```

```
par(mfrow=c(2,5))
i<-1
for(i in 1:10){


boxplot(expression_data_top[,i][which(expression_data_top$Type==1)],
expression_data_top[,i][which(expression_data_top$Type==2)],
ylab="Expression", xlab="Histologia", names=c("ADK", "SCC"))
title(paste0("Niveles expression:",
colnames(expression_data_top)[i]))
mtext(paste0("p-value: ",resultat$adj.P.Val[i]))
i<-i+1
}


head(expression_data_top[,1:10])

#pintamos un heatmap utilizando la expresión de los 10 genes
identificados

par(mfrow=c(1,1))

heatmap(data.matrix(expression_data_top[,1:10]))

top10exprs<-data.matrix(t(expression_data_top[,1:10]))

color.map <- function(Type) { if (Type==1) "#FF0000" else "#0000FF"
}
patientcolors <- unlist(lapply(expression_data_top$Type, color.map))
heatmap(top10exprs, ColSideColors=patientcolors)

ncol(t(top10exprs))
length(patientcolors)

class(as.matrix(expression_data_top))
#-------------------------------------------------------------------
--------------------------

#Suposant que ara ja tinc una llista amb els gens significativament
expressats entre 2 grups, ara necessito obtenir la seva localització
cromosómica

library("biomaRt")
#Hem de seleccionar una databse de anotacions per treballar de les
disponibles, de moment triem ensembl
listMarts()
ensembl=useMart("ensembl")
#a continuació triem un dataset dels posibles de ensembl
listDatasets(ensembl)
ensembl = useDataset("hsapiens_gene_ensembl",mart=ensembl)
#posibles filtres per utilitzar en la busqueda
filter<-listFilters(ensembl)
filter
#possibles atributs a obtenir com a resultat de la busqueda
attributes = listAttributes(ensembl)
attributes[1:5,]


#si tinc aquesta llista de gens

#llista<-c("A4GALT","A2M","AAAS")
#llista
```

```
#He cambiado IL1F7 por IL37, que es el nombre oficial actual, si no
no lo encontraba
#llista_de_gens[10]<-"IL37"

info_gens<-getBM(attributes=c("hgnc_symbol","chromosome_name",
"start_position", "end_position"), filters="hgnc_symbol",
values=llista_de_gens, mart=ensembl)
```

Apéndice 2

```
#---
#Integrative analysis of methylation and expression data
#Alfons Navarro Ponz
#Máster de Bioinformática y Bioestadística
#---


#Packages que necesitamos

library(ggplot2)
library("stats")
biocLite("regioneR")
library("regioneR")

illes <- toGRanges("http://rafalab.jhsph.edu/CGI/model-based-cpg-
islands-hg19.txt")



#Cargamos los arrays de metilación de Illumina 450K, datos
#procedentes de TCGA

#Cargamos Bioconductor i minfi
source("http://bioconductor.org/biocLite.R")
biocLite("minfi")
library(minfi)

#path arrays metilació ADK nivell 1 TCGA
path_ADK<-
"/Users/anavarroponz/Documents/Proyecto_UOC/Archivos/ADK/Metilacion_
ADK/DNA_Methylation/JHU_USC__HumanMethylation450/Level_1"

#path arrays metilació SCC nivell 1 TCGA
path_SCC<-
"/Users/anavarroponz/Documents/Proyecto_UOC/Archivos/SCC/Metilacion_
SCC/DNA_Methylation/JHU_USC__HumanMethylation450/Level_1"

#monto fitxer target per ADK

setwd(path_ADK)
files_names_ADK<-list.files(path_ADK)
s0 <- sapply(strsplit(files_names_ADK, split='_', fixed=TRUE),
function(x) (paste0(x[1],"_",x[2])))
#files_names_ADK<-cbind(files_names_ADK,s0,rep(1,
length(files_names_ADK)),file.path(path_ADK, files_names_ADK))
files_names_ADK<-cbind(files_names_ADK,s0,rep(1,
length(files_names_ADK)),file.path(path_ADK))
dim(files_names_ADK)
colnames(files_names_ADK)<-c("File","BaseName", "Type", "path")

#lo mateix: preparo target per SCC
setwd(path_SCC)
files_names_SCC<-list.files(path_SCC)
s1 <- sapply(strsplit(files_names_SCC, split='_', fixed=TRUE),
function(x) (paste0(x[1],"_",x[2])))
dim(files_names_SCC)
files_names_SCC<-cbind(files_names_SCC,s1,rep(2,
length(files_names_SCC)),file.path(path_SCC))
colnames(files_names_SCC)<-c("File","BaseName", "Type", "path")
```

```
#HEM FALTA FUSIONAR ELS ADK AMB ELS SCC
files_names_def<-rbind(files_names_SCC,files_names_ADK)

dim(files_names_def)
#indicamos el directorio de trabajo
#path<-"/Users/anavarroponz/Documents/Proyecto_UOC/609b9eee-eb59-
46e0-b774-
08ffd37d4af8/DNA_Methylation/JHU_USC__HumanMethylation450/prueba"

#setwd(path)

#guardamos los nombres de los archivos con los que vamos a trabajar
y indicamos el csv con las características de las muestras
#files_names<-list.files(path)
#path2<-paste0(path,"/datos2.csv")

#targets<-read.csv(path2, sep=";", as.is=TRUE)  #vigilar con el
separador del csv
#targets<-targets[1:4, 1:3]  #chapucilla...si vengo excel no
problemas
#files_names_def$BaseName<-targets$BaseName[1:2]

targets<- as.data.frame(files_names_def)

targets$Basename <- file.path(targets$path,targets$BaseName)
#targets$Basename <- file.path(path_SCC,targets$BaseName)



targets_sin_repl<-unique(targets$Basename)

target.unic<-unique(targets[,2:3])



#raw data
rgset <- read.450k(targets_sin_repl, verbose=TRUE)
pData(rgset)<-target.unic

#faig un index per ADK i un index per SCC
ADK_index<-which(target.unic$Type==1)
SCC_index<-which(target.unic$Type==2)

table(target.unic$Type)

#pData(rgset)<-targets

#Per accedir a les red and green intensities
dim(getRed(rgset))
dim(getGreen(rgset))

#preprocessament del les raw data

biocLite("IlluminaHumanMethylation450kmanifest")
library("IlluminaHumanMethylation450kmanifest")
mset <- preprocessIllumina(rgset)

#busquem la localització de les illes CpG en el genoma

biocLite("IlluminaHumanMethylation450kanno.ilmn12.hg19")
library("IlluminaHumanMethylation450kanno.ilmn12.hg19")
mset <- mapToGenome(mset)
```

```
#A continuació obtenim els valors de metilació i les localitzacions
de les illes CpG


dim(getBeta(mset,type="Illumina")) ##the argument type="Illumina"
gives us default procedure
head(granges(mset))

#grafica control de calitat
densityPlot(rgset, main = "Beta", xlab = "Beta")
head(getBeta(mset,type="Illumina"))



par(mfrow=c(3,4))



#Ahora ya tenemos los datos de metilación preprocesados y vamos
#centrarnos en generar gráfica con los valores de metilación en
#relación a unas coordenadas que le pasemos:

head(mset)

head(granges(mset)[seqnames(granges(mset))=="chrX"])

head(granges(mset)[ranges(granges(mset))==IRanges(2715017,2715017)])
meth<-mset
dim(meth)
gr<-granges(mset)
length(gr)

#middle<-
granges(mset)[ranges(granges(mset))==IRanges(2715017,2715017)]

#genero un genomic range amb posició miR-34a

par(mfrow=c(2,2))
#*******************************************************#
k<-1 #Aqui le indico la posición en la matriz del gen que quiero
estudiar

posicion_inicial<-info_gens$start_position[k]
posicion_final<-info_gens$end_position[k]
chromo<-paste0("chr",info_gens$chromosome_name[k])
genx<-info_gens$hgnc_symbol[k]
#x<-9217329

#este es el valor por enciam y por debajo del TSS que voy a utilizar
rango<-5000


ini<-posicion_inicial-rango
fin<-posicion_final+rango

middle<-GRanges(seqnames=Rle(chromo), ranges=IRanges(start=ini,
end=fin), strand=Rle(strand("*")))

Index<-gr%over%(middle)
#head(gr(index))
pos<-start(gr)
#pos[Index]
```

```
#par(mfrow = c(2,2))

cols=ifelse(target.unic$Type==1,1,2)

#plot(pos[Index], getBeta(mset[Index,1]), type="h", xlab="genomic
location", ylab="Beta value", ylim=c(-0.4,1), xlim=c(ini, fin),
main=genx)

matplot(pos[Index], getBeta(mset[Index]), pch=0.2, xlab="genomic
location", ylab="Beta value", col=cols, ylim=c(-0.4,1), xlim=c(ini,
fin), main=genx, cex=0.5)

#matplot(pos[Index], getBeta(mset[Index]), type="p", xlab="genomic
location", ylab="difference", col=cols)


#abline(v=posicion_inicial, col="blue")
#abline(v=posicion_final, col="blue")

#mtext(side=1,"GENE", cex=1, col="blue")
rect(xleft=posicion_inicial, ybottom=-0.4, xright=posicion_final,
ytop=-0.3, col="blue")
abline(h=0)
#intento incloure informació illes CpG



Index_CpG<-illes%over%(middle)
pos_CpG<-start(illes)
pos_CpG_end<-end(illes)
CpG.ini<-pos_CpG[Index_CpG] #posiciones de inicio de las islas CpG
CpG.end<-pos_CpG_end[Index_CpG] #posiciones de final de las islas
CpG
#head(Index_CpG)


rect(xleft=CpG.ini, ybottom=-0.2, xright=CpG.end, ytop=-0.1,
col="yellow")



mtext("CpG \n island", side=1, padj=-2.5, adj=0, cex=0.8)

mtext(side=1,"GENE", cex=1, padj=-3, adj=0)
```

Ara anem a calcular si hi han diferencies significatives en relació a metilació en la regió seleccionada entre ADK i SCC

```{r}

cols=ifelse(target.unic$Type==1,1,2)

data<-as.data.frame(getBeta(mset[Index,]))

#traposem el data.frame

datat<-as.data.frame(t(data))
dataf<-stack(cbind(datat))
datat_inf<-cbind(datat,cols)
dataf<-cbind(dataf,cols)
```

```
pvalue<- vector()
i<-1
print(i)
for(i in 1:(length(colnames(datat_inf))-1)){
  print(i)
pvalue[i]<-
t.test(datat_inf[datat_inf$cols==1,i],datat_inf[datat_inf$cols==2,i]
)["p.value"]
pvalue
i<-i+1
print(i)
}

#corrijo las p-values obtenidas para multiples comparaciones
mediante método FDR
pvalue.ajustada<-p.adjust(pvalue,method="fdr")
#p.adjust.methods

pvalue.ajustada<-round(pvalue.ajustada,4)
pvalue.ajustada2<- as.vector(pvalue.ajustada)

pvalue_ajustada3<-vector()
for(i in 1:length(pvalue.ajustada2)){
  if(pvalue.ajustada2[i]<0.0001){ pvalue_ajustada3[i]<-"p < 0.0001"

  }

  else {pvalue_ajustada3[i]<-paste0("p = ", pvalue.ajustada2[i])}
  i<-i+1
}
pvalue_ajustada3

pvalue.f<-cbind(colnames(datat_inf[1:length(colnames(datat_inf))-
1]), pvalue_ajustada3)
pvalue.f<-as.data.frame(pvalue.f)
colnames(pvalue.f)<-c("ind", "pvalue")

plt <- ggplot(dataf)+facet_wrap(~ind) +
geom_boxplot(aes(x=as.factor(cols), y = values)) +
geom_text(data=pvalue.f, aes(x=1.5, y=0.1,label=pvalue), size=4)
plt

#Ahora calculamos estadística región seleccionada, es decir
comparación de medias de todas las sondas presentes en la región

newdata_SCC <- dataf[which(dataf$cols=='2'),]
newdata_ADK<-dataf[which(dataf$cols=='1'),]
boxplot(newdata_ADK$values, newdata_SCC$values, ylab="B-value",
xlab="Histología",names=c("ADK", "SCC"))
title(paste0("Overall region: ",genx))

pvalue_all<-t.test(newdata_ADK$values,newdata_SCC$values)["p.value"]
mtext(paste0("p-value = ",pvalue_all))

#Si consideramos solo aquellas sondas situadas en una isla CpG


#variables a necesitar
CpG.ini #inicio isla CpG vector
CpG.end #fin isla CpG vector
ini<-posicion_inicial-rango #rango gen promotor
pos[Index] #posiciones de las sondas
```

```
sonda_v<-vector()
illa_v<-vector()
sonda_n<-vector()
r<-1

for(e in 1:length(pos[Index])){
    for(i in 1:length(CpG.ini)){
      if (pos[Index][e] >= CpG.ini[i] && pos[Index][e] <=
CpG.end[i]) {
          sonda_v[r]<-pos[Index][e]
          illa_v[r]<-CpG.ini[i]
          sonda_n[r]<-rownames(mset[Index][e])
          r<-r+1
      }

        i=i+1
  }
    e<-e+1
    i<-1
}

#Hacemos bucle par pintar las diferencias para las diferentes islas
CpG presentes

illas_u<-unique(illa_v)

par(mfrow=c(1,1))

dades_illes<-data.frame(sonda_n, illa_v)
colnames(dades_illes)<-c("sonda", "isla_pos")


#pre-calculo p-values, para poder aplicar correccion por multiple
comparison
pvalue_isla_1<- vector()
for(i in 1:length(illas_u)){

sondas_isla1<-subset(dades_illes, dades_illes$isla_pos==illas_u[i])
data_isla_1<-subset(dataf,ind %in% sondas_isla1$sonda)

pvalue_isla_1[i]<-
t.test(data_isla_1$value[which(data_isla_1$cols==1)],data_isla_1$val
ue[which(data_isla_1$cols==2)])["p.value"]

}
pvalue_isla_1<-p.adjust(pvalue_isla_1,method="fdr")
pvalue_isla_2<-vector()
for(i in 1:length(pvalue_isla_1)){
  if(pvalue_isla_1[i]<0.0001){ pvalue_isla_2[i]<-"p < 0.0001"

  }

  else {pvalue_isla_2[i]<-paste0("p = ", round(pvalue_isla_1[i],4))}
  i<-i+1
}

#pvalue_isla_2

par(mfrow=c(1,1))
for(i in 1:length(illas_u)){

sondas_isla1<-subset(dades_illes, dades_illes$isla_pos==illas_u[i])
data_isla_1<-subset(dataf,ind %in% sondas_isla1$sonda)
```

```
boxplot(data_isla_1$value[which(data_isla_1$cols==1)],
data_isla_1$value[which(data_isla_1$cols==2)], ylab="B-value",
xlab="Histologia", names=c("ADK", "SCC"))
title(paste0(genx,", CpG island: ",i ))
mtext(pvalue_isla_2[i])

}


#Miro metilacion de sondas situadas en las zona promotora de mi gen,
considerando como zona promotora 2kb upstream de inicio de TSS

promotor<-3000
posicion_inicial

sonda_r<-vector()

r<-1
for(e in 1:length(pos[Index])){
    #(pos[Index][e] >=(posicion_inicial-promotor)  && pos[Index][e]
<= posicion_inicial)
      if (pos[Index][e] >=(posicion_inicial-promotor)  &&
pos[Index][e] <= posicion_inicial) {
          sonda_r[r]<-rownames(mset[Index][e])
          r<-r+1
      }
    e<-e+1
}

par(mfrow=c(1,1))
data_promotor<-subset(dataf,ind %in% sonda_r)

boxplot(data_promotor$value[which(data_promotor$cols==1)],
data_promotor$value[which(data_promotor$cols==2)], ylab="B-value",
xlab="Histologia", names=c("ADK", "SCC"))
title(paste0("Promotor: ", promotor,"Kb usptream TSS de", genx))
mtext(paste0("p-value:",
t.test(data_promotor$value[which(data_promotor$cols==1)],data_promot
or$value[which(data_promotor$cols==2)])["p.value"]))
```