UNIVERSITAT DE BARCELONA

Metilación del ADN en posiciones individuales del genoma humano dentro de contextos epigenéticos regionales

Víctor Barrera Burgos

Metilación del ADN en posiciones individuales del genoma humano dentro de contextos epigenéticos regionales

Memoria presentada por

Víctor Barrera Burgos

para optar al grado de

Doctor

por la Universidad de Barcelona.

Tesis realizada bajo la dirección del Dr. Miguel Ángel Peinado en el Instituto de Medicina Predictiva y Personalizada del Cáncer.

Adscrita al programa de Genética 2008/2009
del departamento de Genética de la Facultad de Biología,
Universidad de Barcelona.

Director	Tutora	
Miguel Á. Peinado	Gemma Marfany	Víctor Barrera

Barcelona, Noviembre de 2015.

A mi madre y mi hermana por acompañarme en el camino.
A mi padre por enseñarme a andarlo.

Ahora que este viaje llega a su fin, me sorprendo escribiendo emocionado la sección que deliberadamente he dejado para el final. Aun siendo la última, la considero la más especial ya que aunque el nombre que figura como autor de la obra es el mío, ésta no habría sido posible sin las muchas personas que durante este largo trayecto me han ofrecido su inestimable ayuda. Así que quiero dar las gracias a toda la gente que tanto me ha apoyado en la realización de esta tesis doctoral. Espero no dejarme a nadie pero perdonadme si lo hago, ya que como bien sabéis, ha llovido bastante desde mis primeros años.

En primer lugar, quiero agradecerle a mi director de tesis, Miguel Ángel Peinado, la oportunidad de realizar este trabajo en su laboratorio y sobre todo, por atreverse con el reto de dirigir una tesis bioinformática. De él he aprendido a lograr salir de la caja de los datos y darles un mayor contexto y sentido. Gracias también por enseñarme la paciencia necesaria que ha necesitado la elaboración de este proyecto y por tener siempre la puerta abierta, incluso a kilómetros de distancia.

En segundo lugar, agradecer a Gemma Marfany, mi tutora de tesis, el iniciarme en el camino de la investigación. Desde las clases de Genética Molecular a acogerme en su laboratorio como alumno interno, me proporcionaste el espíritu y curiosidad que deben mover a un científico. Ahora que mi rumbo se ha desviado hacia la bioinformática, aquellas PCRs parecen muy lejanas pero sin ellas y tu apoyo esta tesis no sería una realidad.

Una mención con mucho cariño a todos mis compañeros del laboratorio del MAP-Lab por crear un ambiente tan agradable que hacía que cada día no pareciera trabajo. A Inês, nuestra portuguesa, que siempre nos amenizaba las tardes con su música. A Marta, por esa actitud tan positiva y esa risa tan contagiosa. A Regi, por tener siempre lista una palabra amable. A Berta, por su paciencia enseñándome a trabajar con cultivos celulares. A Mònica, por su gran sabiduría de mami. A Llorenç, por su sentido del humor y dotes musicales. A Elvira, por sus pausas dramáticas durante frases y su mirada al infinito. A Raquel y Yaiza, las últimas incorporaciones, por recordarme constantemente la ilusión de los primeros años. A Quim, por esas grandes tardes de ping-pong y por tener la tenacidad para cumplir su sueño pese a la adversidad. Quiero recordar especialmente a mis dos compis bioinfo del grupo, Sergi y Anna, por unas tardes tan divertidas y ser de los pocos que entendían nuestro raro idioma. Gracias por tan buenas ideas y apoyo durante

el tiempo que compartimos. Izaskun, gracias porque, aunque hemos coincidido muy poco tiempo, desde el principio me ofreciste tu ayuda y este ofrecimiento se alargó incluso cuando yo ya no estaba allí. Dejo para el final a dos personas del grupo por su particular impacto. Mireia, gracias por esas charlas de despacho en las que discutíamos ciencia y no ciencia. Me ayudaste en momentos de incertidumbre y siempre me diste ánimos para continuar. Te deseo lo mejor con los recientes cambios que han aparecido en tu vida. Mar, ya debes estar acostumbrada a tener un trocito dedicado para ti en cada tesis pero es que logras influir en todas las personas que trabajan a tu lado. Aunque en raras ocasiones coincidíamos en temas laborales (excepto con la real-time al principio) no logro recordar una tarde en la que no charláramos. Gracias por apoyarme tanto dentro como fuera del laboratorio. Espero que la vida te sonría porque realmente te lo mereces.

No me puedo olvidar en esta sección de un selecto grupo de amig@s bioinformátic@s. A pesar de la distancia que nos separa, y en algún caso hablamos de muchos miles de kilómetros, seguís siendo uno de mis pilares de confianza. Xavi, Judith, Lorena y Aida, me habéis dado fuerzas en momentos muy difíciles, cuando todo era muy oscuro. Me habéis hecho reír hasta que me dolieran las costillas. Hemos compartido momentos muy frikis y siempre he sentido que éramos una familia. Xavi, gracias por tu magnifico punto de vista sobre la vida, capaz de convertir montañas de problemas en arena. Lorena, gracias por ser un referente científico. Por tener una mentalidad clara y ayudarme a no perder el foco. Aida, agradecerte especialmente tus consejos y ayuda en los difíciles pasos para empezar de cero en una nueva ciudad. Judith, mi sysadmin preferida. Tengo que resumir con un "gracias por todo" ya que hay mucho que agradecer pero por lo menos voy a decir que gracias por las terapias gratuitas de psicología, por las lecciones de Linux, por tu dedicación por los demás y por seguir insistiendo para que acabara esta tesis.

Un cambio de trabajo en una ciudad distinta siempre resulta algo complicado. No obstante, mi nuevo grupo de Bioenergía en Repsol me ha ayudado muchísimo y ahora siento que todos somos una gran pandilla. De entre tod@s, mencionar a dos personas que realmente han hecho que todo fuera más fácil. Anita, quien iba a decir que en aquel desayuno casual empezaría una amistad. Y aunque me des la tabarra con el "jamón dulce", el "biquini" o los "cruasanes fermentados" o que creas que Ironman es mejor que Batman, espero poder compartir muchos "maratones frikis" más. Pamela, los dos sabemos que esta tesis no estaría acabada sin tu ayuda. Muchas gracias por

todo el soporte que me has dado y hacer de revisora aunque el tema te pillara lejos. Entre los tres formamos nuestro pequeño reducto friki en Repsol (aunque Pamela quiera negarlo) y espero que sigamos así durante mucho tiempo.

También quiero agradecer a una persona que me acompañó durante gran parte de este trayecto. Vero, aunque ahora nuestras vidas hayan tomado direcciones distintas, estuviste a mi lado durante una importante etapa compartiéndolo todo. Te deseo lo mejor allá donde te lleven tus pasos.

Llegando ya al final de esta larga lista, quiero dar las gracias a dos amigos que más que amigos, son familia. Héctor, porque a pesar de la distancia te tuve a mi lado reconfortándome y dándome fuerzas. Por no cansarte y desistir cuando yo no paraba de repetir siempre lo mismo. Muchas gracias por ayudarme a salir del pozo. Nacho, quizá el caso más claro de amigo siendo familia. Nos conocemos desde hace muchos años y hemos vivido grandes historias alrededor de una mesa y fuera de ella. Contigo puedo decir que me salió un crítico en los dados y ahora tengo la suerte de tenerte como cuñado.

Y finalmente, a mi familia. A mi madre y a mi hermana, porque cuando el momento más trágico de nuestras vidas llegó, no dejasteis que me perdiera. Patri, gracias por ser una hermana tan maravillosa que me hace sentir tan orgulloso. Mamá, porque aunque te llevaste la peor parte seguiste dándolo todo por nosotros. Gracias a las dos por tanto que me habéis dado.

Y a ti Papá, desde donde me estés observando, gracias por hacer de mí quien soy. Te marchaste demasiado pronto y no pudiste ver el final de lo que de tantas veces hablamos. Ahora lo acabo y te lo dedico. Espero que lo disfrutes.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

sncRNA	*short non coding RNA*
5caC	5-carboxicitosina
5fC	5-formilcitosina
5hmC	5-hidroximetilcitosina
5mC	5-metilcitosina
A	Adenina
ADD	Dominio *ATRX-DNMT3-DMT3L*
ADN	Ácido desoxiribonucleico
ADS-Adi	*ADSC derived Adipocytes*
ADSC	*Adipose Stem Cells*
AIMS	*Amplification of Inter-Methylated Sites*
AME	*Analysis of Motif Enrichment*
AP-PCR	*Arbitrarily Primed PCR*
ARN	Ácido ribonucleico
BER	*Base Excision Repair*
C	Citosina
CAP-Seq	*CXXC Affinity Purification Sequencing*
ChIP-Seq	*Chromatin Immunoprecipitation Sequencing*
COBRA	*COmbined Bisulfite Restriction Analysis*
CTCF	*CCCTC-binding Factor*
dbSNP	*database of Single Nucleotide Polymorphism*
DMR	*Differentially Methylated Region*
DNMT	*DNA Methyltransferase*
FF-iPSC	*Induced Pluripotent Stem Cells derived from Foreskin Fibroblasts*
FT	Factor de transcripción
G	Guanina
H1	*H1 human Stem Cells*
H3K29	Lisina 29 de la histona H3
H3K27	Lisina 27 de la histona H3
H3K4	Lisina 4 de la histona H3
HDAC	*Histone deacetylases*
HP1	*Heterochromatin Protein 1*
HTF	*HpaII Tiny Fragments*
IAP	*Intracisternal A particle*

ICGC	*International Cancer Genome Consortium*
ICR	*Imprinting Control Region*
IHEC	*International Human Epigenome Consortium*
IMR90	*Fetal lung fibroblasts IMR90*
IMR90-iPSC	*Induced Pluripotent Stem Cells derived from Fetal lung Fibroblasts IMR90*
KAT	*Histone Lysine Acetyltransferase*
KDM	*Histone Lysine Demethylase*
KMT	*Histone Lysine Methyltransferase*
LINE	*Long Interspersed Nuclear Element*
lncRNA	*long non coding RNA*
LSH	*Lymphoid-Specific Helicase*
LTR	*Long Terminal Repeat*
mDIP	*Methyl-DNA ImmunoPrecipitation*
MeDIP	*Methylated DNA ImmunoPrecipitation*
MeDIP-Seq	*Methylated DNA ImmunoPrecipitation Sequencing*
MEME	*Multiple Em for Motif Elicitation*
MRE-Seq	*Methyl-sensitive Restrition enzyme Sequencing*
MSP	*Methylation Specific PCR*
MS-HRM	*Methylation-Specific High Resolution Melting*
N	Nucleótido
NGS	*Next generation sequencing*
NIH	*National Institute of Health*
NDR	*Nucleosome Depleted Region*
pb	Pares de bases
PCNA	*Proliferating Cell Nuclear Antigen*
PCR	*Polymerase Chain Reaction*
PRC2	*Polycomb Repressive Complex 2*
piRNAs	*Piwi interacting* RNAs
ROC	*Receptor Operating Characteristic*
RP-HPLC	*Reverse Phase High Performance Liquid Chromatography*
RRBS	*Reduced Representation Bisulfite Sequencing*
SAM	s-adenosilmetionina
SD	*Standard Deviation*
SINE	*Short Interspersed Nuclear Elements*
SNP	*Single Nucleotide Polymorphisms*
T	Timina

TCGA	*The Cancer Genome Atlas*
TDG	*Thymine-DNA glycosylase*
TET	*Ten Eleven Translocation dioxigenase*
TLC	*Thin Layer Chromatography*
TSS	*Transciption Start Site*
UHRF1	*Ubiquitin-like, containing PHD and RING finger domains*
WGBS	*Whole Genome Bisulfite Sequencing*
XIC	*X chromosome Inactivation Center*
XIST	*X-inactive specific transcript*
β_C	Coeficiente de metilación de la isla CpG
β_H	Coeficiente de metilación de la diana HpaII

1-Introducción

1.1-Epigenética

Los días 15 y 16 de febrero de 2001 se publicaron, en la revista *Nature* y *Science* respectivamente, los trabajos sobre la secuenciación inicial del genoma humano (1, 2), cuya versión se consideró definitiva en 2004. No obstante, a día de hoy sigue sin resolverse la pregunta sobre cómo se dirige la expresión espacio-temporal de los genes (3). Todas las células del cuerpo humano contienen una copia idéntica del genoma pero aun así, existen más de 200 tipos celulares distintos. Cómo se controla que un conjunto limitado de genes se expresen con distintos niveles en diferentes tipos celulares resulta la pregunta crítica (4).

La secuencia primaria del genoma es la base para entender cómo se lee el programa genético. Sin embargo, existe una capa de información adicional, denominada "epigenética", que permite regular cómo se manifiesta en distintas fases del desarrollo, tejidos o enfermedades (5). Esta información epigenética se almacena mediante diversos mecanismos: modificaciones químicas tanto de la secuencia nucleotídica como de las proteínas que empaquetan el genoma, secuencias de ARN no codificantes y redes de factores de transcripción regulatorios (6–10). De esta forma, mientras la secuencia primaria del genoma humano está conservada en todos los tipos celulares, la configuración epigenómica de cada célula es distinta, dando lugar a una expresión génica y funciones biológicas diferenciales (4, 11–13).

En los últimos quince años, el interés por el estudio de los mecanismos epigenéticos ha crecido de forma exponencial, como podemos observar por el número de trabajos publicados (**Figura 1**), impulsado por los rápidos avances tecnológicos. Durante este tiempo, importantes consorcios internacionales han surgido para enfrentarse a los retos de los análisis epigenómicos (4)(**Tabla 1**).

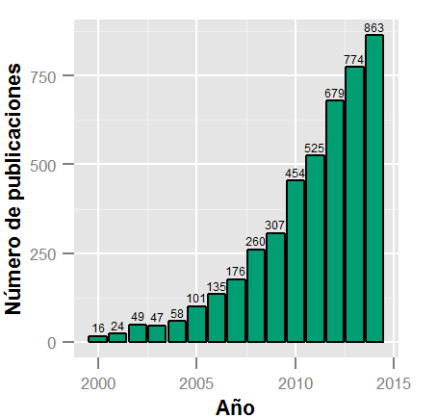


Figura 1: Número de publicaciones por año sobre epigenética
Número de artículos indexados en PubMed que contienen en su título o resumen las palabras "epigenetics" y/o "epigenomics". (http://www.ncbi.nlm.nih.gov/pubmed)

Tabla 1: Consorcios epigenómicos internacionales

Proyecto	Inicio	Afiliaciones	Objetivos	Referencia
Encyclopedia of DNA Elements (ENCODE)	2003	NIH	Estudio de la configuración epigenómica de 100 tejidos humanos primarios y líneas celulares.	(14)
The Cancer Genome Atlas (TCGA)	2006	NIH	Metilomas de ADN de 1000 muestras de cáncer.	(15)
Roadmap Epigenomics Project	2008	NIH	Estudio de la configuración epigenómica de cientos de células primarias normales y células madre.	(16)
International Cancer Genome Consortium (ICGC)	2008	15 países, incluye a TCGA	Perfiles de metilación del ADN de miles de muestras de 50 tipos de cáncer.	(17)
International Human Epigenome Consortium (IHEC)	2010	7 países, incluye a BLUEPRINT y Roadmap Epigenomics	1000 Epigenomas de 250 tipos celulares	(18)

Adaptada de Rivera y Ren (4).

La primera secuenciación del genoma humano, identificó únicamente un 1,5% de la secuencia como codificante de proteína, quedando por resolver la funcionalidad del 98,5% restante, al que en algunas ocasiones se ha denominado ADN "basura" (1, 2). Gracias a investigaciones en el campo de la epigenética, que han permitido generar un número importante de mapas epigenómicos, se ha inferido que casi la mitad del genoma puede realizar actividades bioquímicas específicas y tiene potenciales funciones de regulación (14, 19).

El término "epigenética", acuñado originalmente por Conrad Waddington en 1942, describía los cambios heredables de un fenotipo celular que eran independientes de alteraciones en la secuencia del ADN (20). No obstante, en la actualidad no existe una definición que no resulte ambigua y exenta de debate (21, 22). En el presente trabajo, se entenderá como proceso epigenético aquél en el que se afecta de forma estable la expresión génica mediante mecanismos que no implican alterar la secuencia primaria nucleotídica. Así mismo, un estado epigenético será la configuración de las marcas de la cromatina y el ADN utilizadas por este proceso (23).

1.2-Mecanismos y marcas epigenéticas

Para entender la regulación epigenética es necesario visualizar el ADN de los organismos eucariotas no de forma libre en cada una de nuestras células, sino formando parte de un complejo macromolecular, llamado cromatina, asociado a unas proteínas denominadas histonas. La función principal que se le asigna a la cromatina es la condensación y empaquetado del material genético (24). No obstante, es también el sustrato de los modificadores epigenéticos y su papel en la regulación transcripcional resulta fundamental.
La unidad funcional básica de la cromatina es el nucleosoma, que consiste en un octámero de histonas alrededor del cual se enrollan 147 pares de bases de ADN, dando aproximadamente 1,8 vueltas. El octámero lo conforman dos copias de cada una de las histonas H2A, H2B, H3 y H4. Tradicionalmente, se han definido dos formas o estados de la cromatina:

* Heterocromatina. Se encuentra altamente condensada y localizada sobre todo en la periferia del núcleo. Contiene principalmente genes inactivos.
* Eucromatina. Ligeramente compactada y abierta, contiene la mayoría de genes activos.

La cromatina exhibe una gran plasticidad debida a un conjunto de modificaciones químicas que se producen en las histonas y el ADN. Estas alteraciones influyen tanto en las propiedades físico-químicas de la cromatina como en su estructura, determinando su accesibilidad y funcionalidad. Entre estas modificaciones destaca la metilación del ADN, que evita la unión de factores de transcripción mediante impedimento estérico, provocando el silenciamiento genético. Adicionalmente, la metilación está implicada en el reclutamiento de proteínas represoras capaces de reconocer las posiciones metiladas gracias a dominios de reconocimiento específicos (25). Por otro lado, las modificaciones en las colas de las histonas tienen un efecto en la compactación de los nucleosomas y en la unión de éstos al ADN, provocando que la cromatina se enrolle o desenrolle en función de los grupos químicos añadidos. Un ejemplo es la acetilación de las lisinas, que neutraliza la carga básica de éstas reduciendo la fuerza de unión al ADN y haciéndolo más accesible a factores de transcripción. Otra actividad crítica de las modificaciones de las colas de histonas es el reclutamiento o bloqueo, en función de la modificación, de proteínas que actúan sobre la cromatina. Estas proteínas presentan actividades enzimáticas que modifican la cromatina

permitiendo procesos tales como la transcripción, replicación o reparación del ADN (26).

En la regulación de la cromatina y los procesos epigenéticos intervienen otros mecanismos de actuación a parte de la metilación y las marcas de histonas. Existen ARNs no codificantes que regulan la expresión de determinados genes de una forma específica mientras que otros, se encargan de mantener silenciados a los elementos repetitivos durante procesos de desmetilación global. Por otro lado, los complejos remodeladores de la cromatina son los encargado de desplazar los nucleosomas distancias cortas, gracias a su actividad ATPasa. Este movimiento crea zonas accesibles para que determinados factores de transcripción puedan unirse (27, 28). Adicionalmente, los nucleosomas pueden incluir variantes estructurales de las histonas, que provocan consecuencias funcionales. La variante H2A.Z, por ejemplo, reduce la estabilidad del nucleosoma y se asocia a zonas específicas del genoma dotando a la cromatina de una conformación más abierta (24).

A continuación se describirán las principales marcas epigenéticas. Debido al especial interés de la metilación del ADN en el presente trabajo, ésta se tratará con mayor profundidad. Por otro lado, Los complejos remodeladores de la cromatina y las variantes de histonas no serán tratados. Para una mayor información referirse a Felsenfeld y Groudine (24).

A lo largo del texto se indican diversas modificaciones de las colas de las histonas. Para facilitar la lectura se indican las posiciones modificadas más frecuentes:

- H3K4: Lisina 4 de la histona H3.
- H3K9: Lisina 9 de la histona H3
- H3K27: Lisina 27 de la histona H3

1.2.1-Metilación del ADN

La metilación del ADN es una de las modificaciones epigenéticas más estudiadas y cuyo mecanismo de acción es más conocido. Resulta la principal modificación epigenética en mamíferos (8, 29) y está relacionada con el silenciamiento transcripcional. Se produce gracias a la acción de enzimas con actividad metiltransferasa que catalizan la transferencia de un grupo metilo desde un cosustrato, SAM (s-adenosilmetionina), a la posición 5' de la citosina

(**Figura 2**). La encontramos conservada en los reinos eucariotas de animales, plantas y hongos (29, 30). En mamíferos se encuentra restringida principalmente a la citosina del dinucleótido CpG (31, 32). La simetría de esta secuencia, teniendo en mente la cadena complementaria, ayuda a la conservación de la metilación a lo largo de las divisiones celulares.

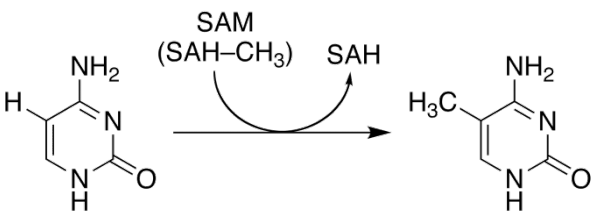


Figura 2: Estructura de la citosina y 5-metilcitosina
Un grupo metilo se transfiere desde un dador, SAM (s-adenosilmetionina), a la posición 5' de la citosina. Fuente: Wikipedia (33).

1.2.1.1-Establecimiento y mantenimiento de la metilación

Tres enzimas conservadas, *DNMT1*, *DNMT3A* y *DNMT3B* (*DNA methyltransferase* 1, 3A y 3B respectivamente), son las responsables de la implantación y mantenimiento de la metilación del ADN (34, 35). Existe otra variante, *DNMT3L*, homólogo estructural de *DNMT3A* y *DNMT3B* que no presenta actividad catalítica pero participa en los procesos de metilación formando complejos con las *DNMT*s activas (36, 37).

La primera en ser identificada y aislada fue la *DNMT1* (38). Esta enzima, encargada de la propagación simétrica de la metilación durante la mitosis, reconoce la cadena naciente opuesta a una posición metilada y transfiere un grupo metilo a la citosina del nuevo dinucleótido CpG. La *DNMT1* se expresa de forma constitutiva en células que se están dividiendo y es más abundante durante la fase S del ciclo celular (39), momentos en los que se produce una mayor síntesis de ADN. Su especificidad se debe a las interacciones que establece con el *PCNA* (*Proliferating Cell Nuclear Antigen*), que la recluta hacia sitios de replicación del ADN, y *UHRF1* (*Ubiquitin-like, containing PHD and RING finger domains*), que se une al ADN hemimetilado (40–42)(**Figura 3**). Esta unión es selectiva hacia el ADN parental de forma que *DNMT1* obtiene la orientación adecuada para metilar únicamente la cadena naciente (43, 44). Esta preferencia estructural evita que se produzcan actividades espurias no deseadas (45).

El complejo ternario *PCNA-UHRF1-DNMT1* forma parte de un complejo aún mayor que contiene varias enzimas asociadas a cromatina. Estas enzimas promueven la degradación de *DNMT1*, vía proteosomal, a menos que existan otras marcas asociadas a heterocromatina (46, 47). De esta manera únicamente se metila el nuevo ADN en las mismas posiciones que el parental, asegurando la fidelidad y precisión para el mantenimiento de la configuración de la metilación del ADN(12).

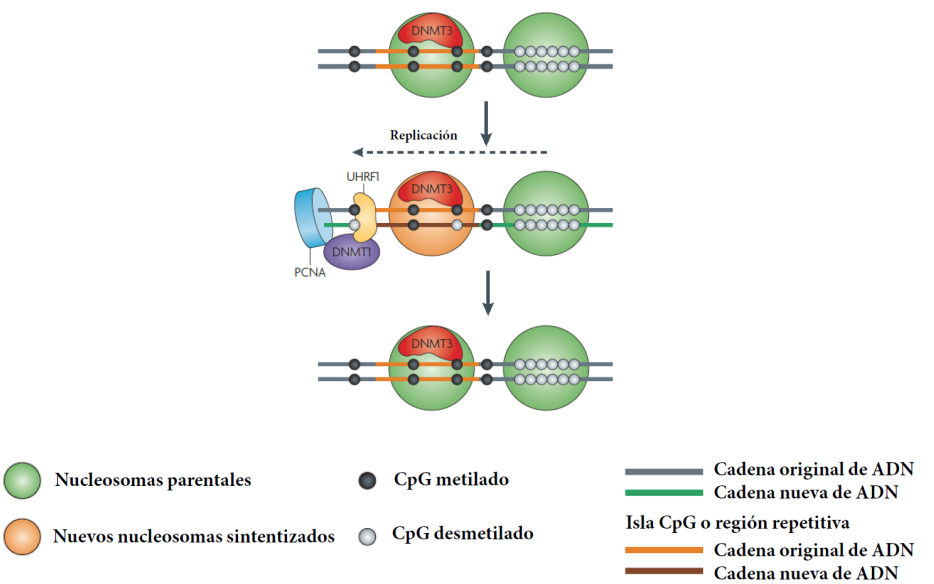


Figura 3: Mantenimiento de la metilación por acción de la DNMT1
El complejo PCNA-UHRF1-DNMT1 se encarga de metilar el ADN hemimetilado. Aunque DNMT1 es la principal enzima de mantenimiento de la metilación, las enzimas DNMT3s participan también la metilación del DNA naciente.
Adaptada de Jones y Liang (48).

Sin embargo, la enzima *DNMT1* tiene una actividad metiltransferasa *de novo* muy reducida debido a su baja afinidad y actividad catalítica sobre ADN desmetilado (45, 49). En este caso, son las enzimas *DNMT3A* y *DNMT3B* las que se han identificado como las encargadas de esta tarea. Interesantemente, se ha observado que ambas participan también en el mantenimiento de la metilación (48)(**Figura 4**).
Experimentos de *knock*-out de *DNMT3A* y *DNMT3B* han mostrado su indispensable papel en el establecimiento de los patrones de metilación en las fases iniciales del desarrollo (34, 35, 50). Estas enzimas contienen un dominio denominado *ADD* (*ATRX-DNMT3-DNMT3L*) que reconoce la histona H3 no modificada. A su vez, este dominio es inhibido por marcas epigenéticas de

activación, tales como la metilación de H3K4 y la presencia de la variante H2A.Z de la histona H2 (51–54).

Por otro lado, *DNMT3A* y *DNMT3B* interaccionan con otros silenciadores epigenéticos también implicados en la heterocromatinización del ADN. Ambas forman un complejo con G9A, una dimetiltransferasa de H3K9, y el remodelador nucleosomal *LSH* (*Lymphoid-specific helicase*). G9A estabiliza y acelera la represión (55, 56) mientras que *LSH* se encarga de proporcionar nucleosomas para el silenciamiento (57–60) (**Figura 4**). No obstante, se ha observado que la unión de G9A es suficiente para iniciar la metilación del ADN (55, 56). Este hecho sugiere un modelo donde la metilación de H3K9 inicia la heterocromatinización de una región y la metilación del ADN asegura el silenciamiento de larga duración (12).

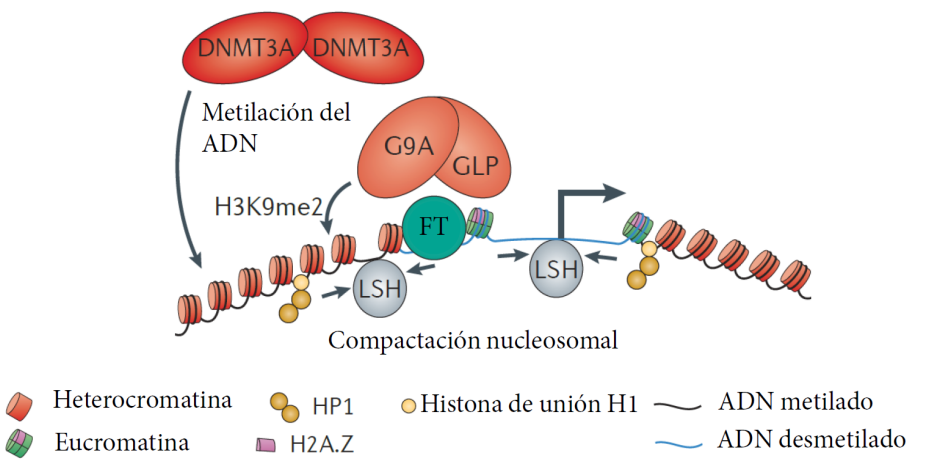


Figura 4: Metilación y silenciamiento de promotores

Factores de transcripción represores (FT) dirigen el reclutamiento del remodelador nucleosomal *LSH*, la histona de unión H1, la *HP1* (*Heterochromatin Protein 1*), la H3K9 metiltransferasa [se muestra el complejo G9A-GLP (G9A-like-protein)] y las DNMT3.

Adaptada de Smith y Meissner (12).

Tan importante es que se produzca la metilación de regiones que deben silenciarse como mantener desmetiladas las zonas activas. Durante mucho tiempo se especuló sobre la existencia de una desmetilación activa o si, por el contrario, únicamente se producía de forma pasiva a través de las divisiones celulares. Ahora se ha podido demostrar que existe un mecanismo activo de desmetilación pero éste requiere de la división celular y de los sistemas de reparación del ADN que finalmente provocan la escisión de la citosina metilada (61, 62) (**Figura 5**). Las enzimas principales implicadas en este mecanismo son las dioxigenasas *Ten Eleven Translocation* (*TET*). Deben su nombre a que fueron descubiertas por su implicación en leucemia mieloide

causada por una translocación del gen *TET1*, situado en el cromosoma 10, con el gen *MLL*, en el cromosoma 11 (63, 64). Estas proteínas son capaces de oxidar la metilcitosina a hidroxi-, formil- o carboxi-citosina. Una vez formados, estos intermediarios facilitan la desmetilación principalmente de dos formas. Por un lado, bloquean el complejo *DNMT1/UHRF1* evitando el mantenimiento de la metilación. La posterior división acaba provocando una desmetilación pasiva (65). Por otro lado, la carboxicitosina y la formilcitosina son el sustrato para la *TDG* (*Thymine-DNA glycosylase*) y el sistema de reparación del ADN por escisión de base (*BER*) que concluyen con la citosina desmetilada (66–70)(**Figura 5**).

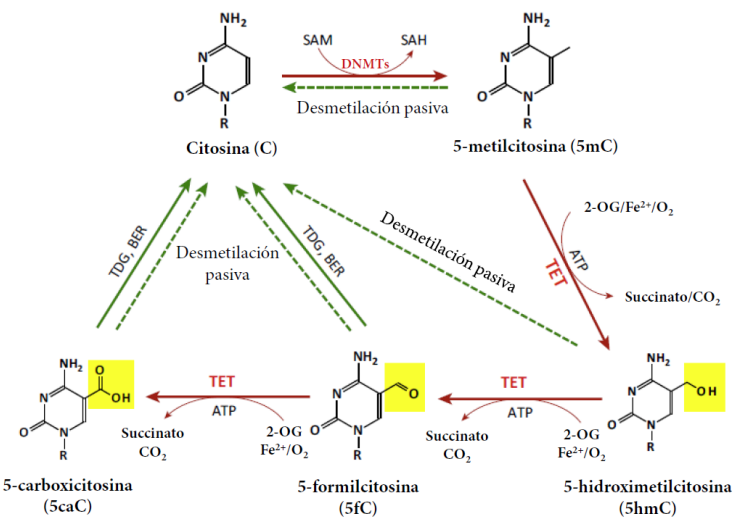


Figura 5: Mecanismos de desmetilación de la 5-metilcitosina
Sucesivas rondas de oxidación de la 5-metilcitosina por parte de las proteínas *TET* dan lugar a los intermediarios: 5- hidroxi-, formil- o carboxi-citosina. Estos intermediarios se reconvierten a citosina mediante mecanismos pasivos o por la acción de la *TDG* y el mecanismo de reparación del ADN por escisión de base (*BER*).
Adaptada de Huang y Rao (65).

Por otra parte, *TET1* se une a promotores asociados a islas CpG (descritas en el siguiente apartado) de genes tanto constitutivos como de desarrollo. Esta unión se produce incluso cuando no hay metilación, sugiriendo que *TET1* puede unirse a complejos más grandes que actúan como correctores epigenéticos (71, 72).

Las actividades de las enzimas *DNMT* y *TET*, junto a los complejos de los que forman parte, aseguran el mantenimiento y propagación de la configuración epigenética de la metilación.

1.2.1.2-Distribución de la metilación

Aunque la metilación de la citosina es una marca epigenética ampliamente conservada, su distribución en el genoma de distintos grupos de organismos presenta configuraciones muy diferentes (**Tabla 2**). En efecto, el patrón de metilación encontrado en el grupo de los vertebrados, con una metilación global a lo largo de todo el genoma, excepto en unas zonas denominadas islas CpG (descritas a continuación) (6, 73), no puede extrapolarse al resto de eucariotas. En hongos encontramos metilados únicamente los elementos repetitivos (74). En los animales invertebrados, el patrón más frecuente es una metilación en mosaico, con zonas metiladas interrumpidas por dominios libres de metilación (75, 76) (**Figura 6**). En mamíferos, el 70-80% de las citosinas en contexto CpG están metiladas. Finalmente, los niveles de metilación más altos los encontramos en algunas especias de plantas, con hasta un 50% de todas las citosinas, no únicamente en contexto CpG, del genoma metiladas (77).

Tabla 2: Patrones de metilación en varios filos eucariotas

	Patrón general	Secuencia metilada	Metilación de transposones	Metilación de cuerpos génicos
Plantas	Mosaico	CG, CNG y CNN	Sí	Sí
Hongos	Mosaico	CNN	Sí	No
Invertebrados: insectos	Mosaico	CG (casos de CT y CA)	No determinado	Sí
Invertebrados: deuterostomas	Mosaico	CG	Sí	Sí
Vertebrados	Global	CG	Sí	Sí

Adaptada de Suzuki y Bird (29).

En el presente trabajo estudiaremos la metilación en humanos. Por este motivo, la configuración que describimos es la común al grupo de los vertebrados, específicamente el de los mamíferos. Como ya se ha comentado, éstos presentan un patrón de metilación global. No obstante, la distribución de las citosinas metiladas es asimétrica. En el genoma, el dinucleótido CpG se encuentra infrarrepresentado. Aparece únicamente con una frecuencia del 20% de la que se esperaría por la composición de las bases (1). El motivo principal es la desaminación espontánea de la citosina y en especial, de la 5-metilcitosina, que produce un cambio de base a timina. Si se repara incorrectamente se da lugar a una mutación del tipo transición con un efecto general del incremento en los dinucleótidos TpG y CpA a cambio de los CpG (78).

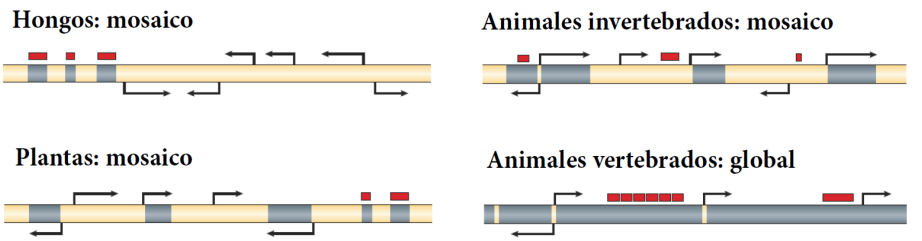


Figura 6: Distribución de la metilación en plantas, hongos y animales
En la imagen se muestran en gris las zonas metiladas y en amarillo las desmetiladas. Las cajas rojas representan transposones. En hongos se observa una distribución en mosaico donde la metilación afecta principalmente a estos elementos. Por otro lado, la distribución en mosaico de plantas y animales invertebrados muestra, principalmente, regiones desmetiladas salpicadas por zonas metiladas. A diferencia de estos casos, en vertebrados la metilación es global excepto en ciertas regiones, como islas CpG, y los elementos repetitivos se encuentran metilados así como cuerpos génicos y ADN intergénico.
Adaptada de Suzuki y Bird (29).

Existen, no obstante, unas regiones en el ADN que representan entre un 1 y 2% del genoma, con una frecuencia cercana a la esperada. A estas zonas se las conoce como islas CpG (CGi) y se encuentran asociadas a las regiones promotoras del 60-70% de los genes del genoma humano (1, 79–82). Aunque de forma más minoritaria, también se pueden encontrar dentro del propio cuerpo del gen (83) (**Figura 7**) o incluso en desiertos génicos. Las islas CpG tienen un papel muy importante en la regulación de la expresión génica y por ello tienden a estar protegidas contra la metilación y su eliminación por mutación. La gran mayoría de las islas en el genoma de células somáticas permanecen desmetiladas (11).
Fueron identificadas tras digerir enzimáticamente ADN genómico de ratón con la enzima de restricción HpaII, cuya diana de reconocimiento es CCGG (84). Posteriormente, tras la secuenciación del genoma humano se utilizaron métodos computacionales para predecirlas en función de las características estructurales observadas (80, 85). Estas características son:

- Tamaño igual o superior a 200 pares de bases.
- Contenido G + C igual o superior al 50%.
- La frecuencia de CpG (observado/esperado) igual o superior a 0,6.

Esta definición no está exenta de debate y han aparecido otras modificando los valores para los parámetros, los métodos computacionales de análisis (86) e incluso mediante el uso de técnicas experimentales adicionales (87). No obstante, la definición de Gardiner-Garden y Frommer (85) ha sido la utilizada ampliamente incluyendo el consorcio internacional del proyecto ENCODE (88).

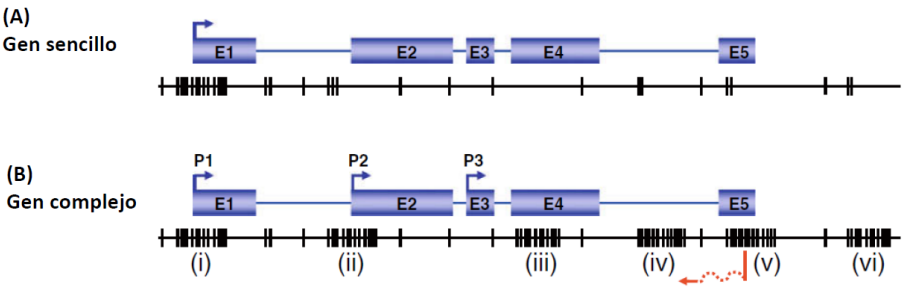


Figura 7: Distribución de islas CpG

En la imagen se representan los dinucleótidos CpG mediante líneas verticales negras y las islas pueden apreciarse como agrupaciones densas de éstas. En muchos genes la isla se encuentra asociada únicamente su región promotora (A). Otros pueden presentar estructuras más complejas como el (B), donde se observan varias islas (ii-iv) asociadas a promotores alternativos (P2-P3) e incluso una asociada a un transcrito antisentido (v).
Adaptada de Illingworth y Bird (89).

En aquellos genes activos que presentan islas CpG en sus *TSSs* (*Transcription Start Sites*) puede observarse, a parte de la isla desmetilada, la región promotora desprovista de nucleosomas, *NDR* (*Nucleosome Depleted Region*). Ooi *et al* (52) mostraron que para darse la metilación *de novo,* DNMT3A forma una estructura con DNMT3L en forma de caja que acomoda un nucleosoma. Una zona desprovista de nucleosomas evitaría la unión de este complejo y por tanto, la metilación. Además, esta región *NDR* se encuentra flanqueada por nucleosomas que contienen la variante de histona H2A.Z y que están marcados con trimetilación de H3K4 (90). Tal y como se ha comentado en el apartado anterior, estas marcas en las zonas flanqueantes son también inhibidoras del complejo de metilación. Adicionalmente, la proteína TET se encuentra presente. Ésta tendría como función corregir cualquier 5-metilcitosina que se produjera. Estos mecanismos se encargan de garantizar el mantenimiento del estado desmetilado (**Figura 8**). Por otro lado, en la represión de estos genes con islas CpG en sus zonas promotoras, se han observado varios tipos de mecanismos de inactivación, como por ejemplo la metilación de H3K27 por parte del complejo represor Polycomb. Además, en aquellos que deben silenciarse de forma permanente, sus islas CpG asociadas a los promotores son metiladas (11). Ambos mecanismos se comentarán en la sección del papel de la epigenética en el desarrollo.

Por el contrario, los genes cuyas zonas promotoras no contienen islas CpG no muestran un patrón de metilación claro y no se conoce el papel que la metilación representa. Existen estudios que asocian la metilación a silenciamiento (91) y otros, todo lo contrario (92).

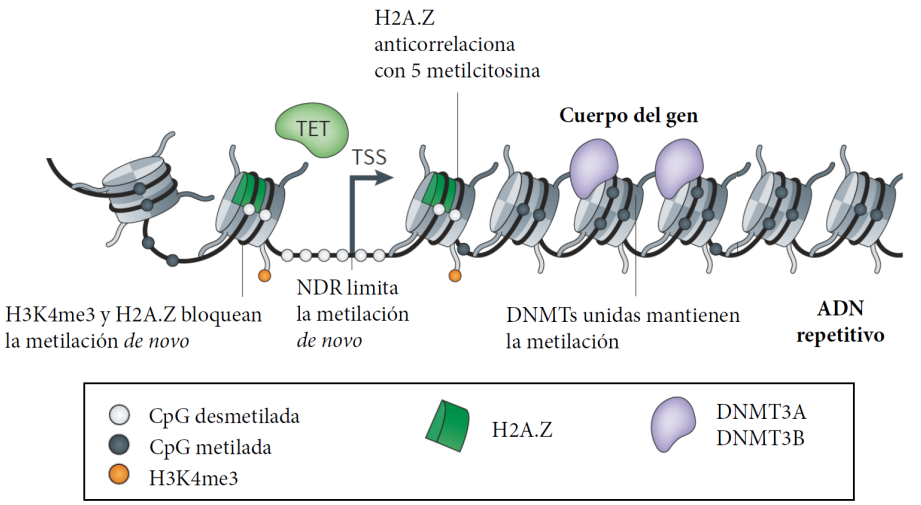


Figura 8: Metilación en promotores con islas CpG y cuerpos génicos
En los promotores con islas CpG el estado de desmetilación se mantiene mediante varios mecanismos. La presencia de marcas de activación, como la trimetilación de H3K4, y de la variante de histona H2A.Z evitan la unión del complejo de metilación. Adicionalmente, la falta de nucleosomas en la región *NDR* evita la actuación de las *DNMT3s*. Finalmente, proteínas de la familia TET se encargan de corregir cualquier metilación espuria. Por otro lado, el ADN se mantiene metilado en los cuerpos génicos y en el ADN repetitivo. En el primer caso debido a la unión de *DNMTs* y en el segundo, por la actuación de mecanismos que involucran *piwi-RNAs* (*piRNAs*) (comentado en secciones posteriores).
Adaptada de Jones (11).

A diferencia de los promotores, los cuerpos génicos tienden a tener poca cantidad de dinucleótidos CpG y estar poblados de elementos repetitivos. No obstante, aunque no se conoce su función, existen también ejemplos de islas CpG en estas regiones. Se ha propuesto que pueden tratarse de promotores "huérfanos" utilizados durante las primeras fases del desarrollo. La protección del estado desmetilado durante esa fase evitaría su mutación a timinas de las citosinas y la desaparición de la isla (93).

1.2.1.3-Funciones de la metilación

La función principal asociada a la metilación es el silenciamiento transcripcional. En los genes regulados mediante isla CpG, una vez ésta se encuentra metilada y los nucleosomas ensamblados, la transcripción no puede iniciarse (94–96). No obstante, este efecto represor ocurre únicamente cuando la metilación afecta a la región promotora. En los cuerpos génicos se observa el comportamiento contrario. Las islas presentes dentro del gen pueden encontrarse totalmente metiladas sin que la elongación del transcrito se vea afectada. Incluso se ha observado que está metilación está asociada

positivamente a la transcripción (83). La posible función de esta metilación dentro del gen sigue sometida a debate. No obstante, existen fuertes evidencias de que esta metilación resulta un mecanismo de silenciamiento de los elementos repetitivos (97). Evitando la transcripción de éstos, se facilitaría la del gen (83).

Se conocen varios procesos donde el silenciamiento asociado a metilación juega un papel determinante:

1. Inactivación del cromosoma X: Las hembras de los mamíferos presentan dos cromosomas X mientras que los machos, únicamente uno. Esto genera un problema de compensación de dosis que la naturaleza ha resuelto inactivando una de las copias en hembras. La elección del cromosoma X a inactivar se produce al azar (98) y están implicados un conjunto de ARNs no codificantes finamente regulados. Éstos se encuentran en una región conocida como *XIC* (*X chromosome Inactivation Center*) (99) y el más destacado es *XIST* (*X-inactive specific transcript*) (100). La selección del cromosoma X a inactivar ocurre en las primeras fases de desarrollo embrionario (101, 102). Esta inactivación se produce mediante la progresiva pérdida de las modificaciones activadoras de las histonas como la metilación de H3K4, reclutamiento del complejo *PRC2* (*Polycomb Repressive Complex 2*) y la acumulación de marcas epigenéticas represivas como la trimetilación de H3K27 y la dimetilación de H3K9 (103–106). Finalmente, la región es metilada estableciendo un bloqueo completo de la expresión (107). Esta configuración de la metilación debe mantenerse a través de las divisiones celulares para evitar la reactivación del cromosoma X inactivo. La enzima *DNMT1*, presentada en el apartado anterior, se encarga del mantenimiento de este estado silenciado.

2. Imprinting génico: De forma parecida a la inactivación del cromosoma X, existe un conjunto de genes donde uno de los dos alelos se inactiva. No obstante, a diferencia del cromosoma X donde la selección es estocástica, la selección es debida al origen, paterno o materno, del alelo. Es decir, algunos genes únicamente expresaran el alelo materno, mientras el paterno está silenciado, y viceversa (108). En humanos, existen alrededor de 150 genes, entre codificantes de proteínas y ARN no codificantes, con este sistema de control de su

expresión (109). Estos genes tienden a estar agrupados formando clústeres que se encuentran regulados por una sección denominada *ICR* (*Imprinting Control Region*) (110). El estado de metilación del *ICR* determina la expresión de los genes que forman el clúster. Éste debe ser estable y heredable a través del desarrollo (111–113). No obstante, no existe un único mecanismo definido, sino que depende de cada *ICR* y un estado metilado no siempre indica un silenciamiento (111). Un ejemplo son el par de genes H19 e Igf2, ambos controlados por el mismo *ICR*. Cuando éste se encuentra desmetilado, caso materno, se permite la unión de CTCF, una proteína aisladora, que bloquea el acceso de *enhancers* al gen Igf2 y éste no se expresa. En cambio, cuando el *ICR* se encuentra metilado, origen paterno, CTCF no puede unirse e Igf2 puede expresarse. Adicionalmente, la metilación del *ICR* se extiende hasta H19 bloqueando su transcripción (114–117). Por otro lado, existen otros mecanismos de regulación que actúan sobre ARNs no codificantes que a su vez controlan la expresión de otros genes (118).

3. Represión de transposones: Se calcula que aproximadamente el 40% del genoma de los mamíferos está constituido por estos elementos repetitivos. Existen tres clases principales: *LINE*s, *SINE*s y *LTR*s *(long interspersed nuclear elements*, *short interspersed nuclear elements* y *long terminal repeats*, respectivamente) (119).
   Los elementos repetitivos pueden alterar la estructura del genoma de dos formas principalmente:
   * Por disrupción génica. Al transponerse puede insertarse en un gen y destruir su función. Son ejemplos la inserción de un transposón en el factor IX sanguíneo provocando hemofilia o el cáncer de mama por inserción en el gen BRCA1 (120, 121)
   * Por recombinación homóloga. El elevado número de copias de estos elementos los convierte en centros de recombinación homóloga a lo largo de todo el genoma. Esta recombinación pude conllevar duplicaciones, deleciones, amplificaciones y translocaciones. Se conocen ejemplos como la distrofia muscular de Duchene o casos de cáncer de colon (120, 121).

   Adicionalmente, la mayoría de estos elementos repetitivos contienen promotores fuertes que, aparte de facilitar su transcripción, pueden alterar la expresión de genes cercanos. La célula mantiene a estos elementos reprimidos mediante una fuerte metilación (12). Ésta ni

siquiera desaparece en los procesos de reprogramación epigenética, mencionados más adelante en la sección del papel de la epigenética en el desarrollo, gracias a la acción de los *piRNA*.

4. Metilación específica de tejido: Durante las primeras fases del desarrollo se generan diferentes tipos celulares a partir de una única célula ancestral (el cigoto). Durante este proceso, se establece un patrón específico de expresión génica para cada tipo de tejido. Se ha observado, por ejemplo, que la ausencia de metilación inhibe por completo la diferenciación de células madre (122). En cambio, durante la diferenciación, genes pluripotentes como OCT4 y NANOG, y específicos de la línea germinal se bloquean mediante metilación (12).
   Cada tejido presenta conjuntos específicos de genes que son metilados en sus regiones promotoras. Experimentos con *knock-outs* condicionales para la *DNMT1* en células madre hematopoyéticas quiescentes muestran una desregulación en las proporciones de células diferenciadas entre tejido mieloide y linfoide (123, 124). Esto demuestra el importante papel de la metilación en la especificación del linaje celular (125–127).

Existe no obstante, un importante debate respecto al silenciamiento transcripcional mediante metilación. Éste hace referencia al orden en el que se producen los distintos eventos. Es decir, si la metilación es la causante directa del silenciamiento o por el contrario, actuaría como una especie de cerradura una vez ya estuviera la región silenciada. Un estudio relacionado con la inactivación del cromosoma X muestra que la metilación aparece una vez ya se ha inactivado (128). Adicionalmente, estudios en cáncer muestran que promotores ya silenciados por *Polycomb* son más proclives a ser metilados (91, 129–131). Por el contrario, estudios de diferenciación de células hematopoyéticas muestran un papel iniciador más que de refuerzo del silenciamiento (132). Actualmente, el grueso de las evidencias apuntan a que la metilación añade un nivel más de estabilidad represiva y no resulta el primer paso del silenciamiento génico (11).

1.2.2-Modificaciones postraduccionales de las histonas

Las histonas son proteínas muy conservadas evolutivamente debido a su importancia crítica en el empaquetamiento del ADN. Estructuralmente se distinguen dos dominios: una parte central globular más organizada y otra totalmente desestructurada, denominada cola, correspondiente al extremo N-terminal. Tanto la sección globular como la cola son el posible sustrato para más de 130 modificaciones postraduccionales. Además, se han descrito más de 700 isoformas de histonas distintas en células humanas (133, 134). Las modificaciones de las colas de las histonas son consideradas como uno de los mecanismos epigenéticos más versátiles. Existen más de 12 posiciones donde pueden producirse estas modificaciones, tales como la acetilación, fosforilación, ubiquitinación y la mono-, di- o tri-metilación. Estos cambios químicos tienen papeles tanto de activación como de silenciamiento de la transcripción, facilitando o dificultando el acceso al ADN o sirviendo como punto de unión o exclusión para otros complejos proteicos (26). Dicha variabilidad ofrece un abanico regulatorio que conduce a muchos autores a hablar de un "código de histonas" donde la combinación de ciertas modificaciones daría lugar a distintas funciones biológicas (135). El estudio de los mapas genómicos ha permitido establecer asociaciones entre ciertas combinaciones y determinada anotación funcional (**Tabla 3**). De esta forma, y mediante estudios bioinformáticos de integración de datos y *machine-learning*, se han podido realizar predicciones sobre el comportamiento de la cromatina en función de los resultados de experimentos de inmunoprecipitación y secuenciación de las marcas cromatínicas (*ChIP-Seq*) (19, 136).

Tabla 3: Marcas cromatínicas distintivas en elementos genómicos

Anotación Funcional	Marcas de Histonas	Referencias
Promotor	H3K4me3	(137–139)
Promotor bivalente/latente	H3K4me3/H3K27m3	(140)
Cuerpo génico transcrito	H3K36me3	(141)
***Enhancer* (activo y latente)**	H3K4me1	(142)
***Enhancer* latente del desarrollo**	H4K4me1/H3K27me3	(143, 144)
***Enhancer* activo**	H4K4me1/H3K27ac	(143–145)
Regiones reprimidas por Polycomb	H3K27me3	(140, 146)
Heterocromatina	H3K9me3	(147)
Modificaciones: me1, monometilación; me3, trimetilación; ac, acetilación.		

Adaptada de Rivera y Ren (4).

1.2.3-ARNs no codificantes

Aunque únicamente una pequeña porción, aproximadamente el 1-2% del genoma de mamíferos, tiene potencial codificante de proteínas, alrededor del 70-90% del genoma se transcribe en algún momento del desarrollo (148). Esta inmensa cantidad de ARN no codificante constituye un verdadero transcriptoma que se ha revelado como un sistema de regulación de la expresión génica. Existen diversas clasificaciones de estos ARN no codificantes. No obstante, únicamente se comentarán aquellos que tienen relación con la regulación epigenética.

* *piRNAs* (*Piwi interacting RNAs*): Son los miembros más numerosos de y mejor estudiados de la familia de los *sncRNA* (*short non coding RNA*) (149), con un tamaño de entre 24-35 nucleótidos. Su función es silenciar transcripcionalmente los genes de retrotransposones mediante una degradación de su ARN mensajero o reclutando proteínas que metilan su secuencia genómica. Tienen un papel fundamental en las células de la línea germinal (150–152) dado que durante el proceso de gametogénesis se produce una desmetilación global. En esta situación, los elementos repetitivos se convierten en puntos de recombinación homóloga o pueden transponerse e insertarse en otros puntos del genoma causando graves alteraciones.

* *lncRNAs* (*long non coding RNAs*): Con una longitud superior a 200 nucleótidos, se estima que su número se encuentra entre 10.000 y 20.000 por genoma de mamífero (148, 153–155). Se encuentran principalmente en el núcleo, donde actúan en la mayoría de ocasiones uniéndose a complejos proteicos epigenéticos como el complejo represor Polycomb 2 (*PRC2*). El ARN otorga especificidad al complejo y lo guía hacia la región sobre la que realizar las modificaciones. A diferencia de los factores de transcripción que tienen una diana de reconocimiento de 6-8 nucleótidos, los *lncRNA* pueden llegar a reconocer centenares de bases por lo que se consigue una especificidad única. Dada esta alta afinidad por una secuencia determinada y el acercamiento y bloqueo físico entre el complejo y la región se puede seleccionar entre los distintos alelos de un gen (156).

1.3-Regulación epigenética en el desarrollo

Las células primordiales de un organismo tienen el potencial de convertirse en cualquier tipo celular. A medida que se van diferenciando, su programa génico queda más definido y bloqueado. Estos cambios se producen sin afectar a la secuencia genética sino mediante mecanismos epigenéticos, como los descritos en el punto anterior, y factores de transcripción específicos de tejido, a su vez también regulados por marcas epigenéticas.
Durante el desarrollo podemos definir 3 fases donde se producen eventos de reprogramación epigenética (**Figura 9**). La primera de ellas se produce tras la fecundación. En ese momento, las células madre pluripotentes únicamente deben expresar un subconjunto de genes determinados entre los que se incluyen factores de transcripción que mantienen el estado de pluripotencia, tales como OCT4 y NANOG. Los genes relacionados con la diferenciación y el desarrollo se mantienen reprimidos mediante el denominado grupo *Polycomb* que metila H3K27 (157–159). Este mecanismo produce una represión temporal y flexible de estos genes dado que será necesario su posterior reactivación durantre la diferenciación. En este sentido, se han observado genes relacionados con el desarrollo que presentan dominios bivalentes que contienen tanto marcas de activación, metilación de H3K4, como de inactivación, metilación de H3K27 (140, 158).

Durante la diferenciación celular nos encontramos la segunda fase de reprogramación epigenética. Para ello, las células deben mantener los genes relacionados con la pluripotencia silenciados. No obstante, los mecanismos de represión deben ser más firmes para evitar una desdiferenciación, como ocurre en algunos cánceres. Por ese motivo, se desmetila H3K4, se metila H3K27 y, lo más importante, se metila el ADN, una marca epigenética más estable (160). Por el contrario, los genes asociados a la diferenciación se desmetilan y se añaden marcas de activación como la metilación de H3K4.

La última fase de reprogramación se produce durante la generación de las células primordiales germinales, aquellas que luego daran lugar a los gametos. Esto es debido a que es necesaria una reescritura de toda programación epigenética para la siguiente generación (161, 162).

No obstante, existen otros tipos de secuencias que requieren una regulación particular, los retrotransposones y los genes imprintados, mencionados en el apartado anterior. Los primeros deben mantenerse siempre silenciados y por

eso, no se desmetilan en ninguna fase de reprogramación (163). Esto es gracias a la actuación principalmente de los *piRNA*, aunque otros mecanismos también participan en el silenciamiento de estos elementos. Los segundos, genes con una expresión parental específica, únicamente deben expresar un alelo parental determinado. La configuración epigenética de estos genes tan solo se una reprograma durante la generación de gámetos (161, 164).

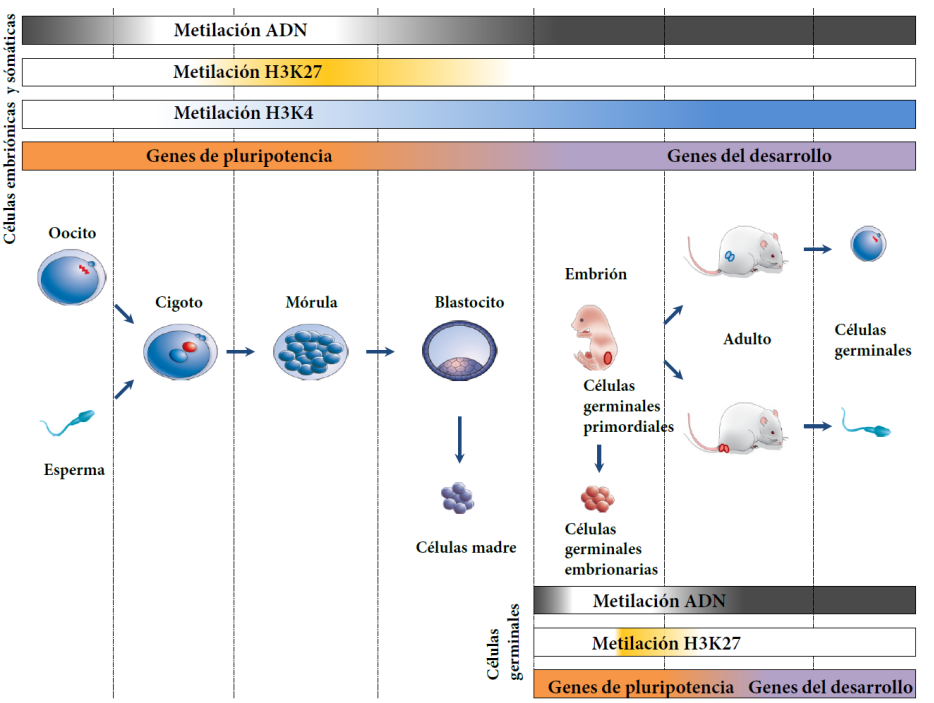


Figura 9: Regulación epigenética durante el desarrollo de un mamífero

En las primeras fases del desarrollo se elimina la metilación del ADN. Adicionalmente, se inicia la expresión de los genes encargados de la pluripotencia mientras que los genes asociados al desarrollo se mantienen reprimidos. En cambio, una vez iniciada la diferenciación de células pluripotentes, como las células madre, los genes encargados de la pluripotencia se reprimen, potencialmente de forma permanentemente, mediante la metilación del ADN. Por otro lado, los genes de desarrollo comienzan a expresarse incrementándose la metilación de H3K4.

Por otro lado, en los procesos iniciales de desarrollo de las células germinales primordiales, la metilación del ADN y las marcas de histonas represoras, como la metilación de H3K27 son eliminadas. Los genes de pluripotencia se reexpresan durante este periodo. Los genes imprintados se desmetilan y los genes de desarrollo se expresan posteriormente.

En estos procesos, las marcas flexibles, como la metilación de H3K27, permiten un silenciamiento temporal de los genes de desarrollo mientras que la metilación del ADN representa un silenciamiento estable como en el caso de genes imprintados, transposones y genes asociados a pluripotencia.

Adaptada de Reik (163).

1.4-Epigenética y ambiente

Muchas de las definiciones utilizadas para describir la "epigenética" incluyen la heredabilidad mitótica y meiótica. No obstante, se trata de un punto muy controvertido. Las células pertenecientes a un mismo linaje celular comparten patrones de expresión y configuración epigenética. Aun así, para algunas de las marcas epigenéticas, como las modificaciones de histonas, no se ha podido demostrar que exista ningún mecanismo de transmisión entre divisiones celulares (10). Sin embargo, la metilación del ADN, en el caso de la inactivación del cromosoma X en hembras, muestra un patrón de herencia claro (**Figura 10**). Por este motivo, se han buscado definiciones, como la utilizada en este documento, que no requieren del requisito de heredabilidad para definir un carácter epigenético.
Existe otro tipo de herencia sobre la que se debate, la transgeneracional. Ésta describe que las distintas configuraciones epigenéticas se transmiten a través de las generaciones de individuos y no únicamente entre sus tejidos. Resulta muy importante el estudio de estos fenómenos de herencia debido al efecto que tiene el ambiente sobre nuestro paisaje epigenético.

Existen numerosos ejemplos donde el ambiente tiene un claro efecto en el desarrollo. En varias especies de organismos la determinación sexual no se produce por una configuración genética, sino en función de la temperatura durante la embriogénesis (165). Otro ejemplo es la necesidad de algunas plantas a una exposición al frío para poder florecer. Este proceso, conocido como vernalización, tiene sus bases en cambios epigenéticos debidos a la temperatura (166). La dieta también tiene un papel fundamental en el desarrollo. En las abejas, existen dos clases de hembras: las obreras y las reinas. El dimorfismo no se debe a una diferenciación genética sino a la alimentación. Aquellas larvas únicamente alimentadas con jalea real serán las futuras reinas (167).

Algunos efectos ambientales pueden llegar a percibirse en generaciones posteriores a las que recibieron el estímulo. Así, por ejemplo, el cuidado materno en ratas modifica la respuesta al estrés de la primera y segunda generaciones. Se ha comprobado que aquellas crías que reciben menos atenciones por parte de sus madres, tienen un incremento en el silenciamiento mediante metilación del receptor de glucocorticoides en el hipocampo, haciéndolas más susceptibles a estrés (168).

Figura 10: Patrón de inactivación del cromosoma X
La decisión estocástica sobre el cromosoma X a inactivar, durante las primeras fases del desarrollo, se transmite en las divisiones celulares posteriores. En la imagen puede observarse la distribución de cada uno de los dos cromosomas X, teñidos con distintos marcadores.
Imagen reproducida con permiso del *Walter & Eliza Hall Institute of Medical Research* (*WEHI*). (169)

En humanos, se han observado diferencias epigenéticas entre gemelos monocigóticos, genéticamente iguales, incluso desde el momento del nacimiento (170). Estas diferencias, presentes también en los patrones de metilación, se incrementan con el paso de los años indicando un importante efecto del ambiente (171). También se han observado efectos transgeneracionales relacionados con épocas de hambruna. En estos casos, la cantidad de alimentos que recibían los abuelos y la fase del desarrollo en la que se encontraban, tenían efectos en la predisposición a ciertas enfermedades, como la obesidad y la diabetes, de los nietos (172). No obstante, estos estudios en humanos no pueden realizarse en las condiciones adecuadas que permitan asegurar que se trata de un evento transgeneracional y han recibido bastantes críticas. Sin embargo, sí que se han podido realizar experimentos con organismos modelos que demuestran estos efectos intergeneracionales.

Se han realizado numerosos estudios con el modelo murino agutí amarillo viable (A^{vy}). El alelo *wild-type* del gen proporciona un color marrón al ratón. Los ratones A^{vy} presentan un retrotransposón, de la clase *IAP* (*Intracisternal A particle*), situado cerca del gen agutí. Este retrotransposón presenta un promotor fuerte que puede actuar sobre el gen agutí dando lugar a ratones amarillos y con una mayor incidencia de obesidad, diabetes y tumores (23, 173). La regulación del transposón mediante mecanismos epigenéticos, su metilación en este caso, determina la expresión del mismo (174). Cuando se encuentra desmetilado, los ratones resultan amarillos. En cambio, cuando está metilado los ratones son marrones, denominados pseudo-agutís. Aparece

también un fenotipo moteado en función de si una mayor o menor proporción de células metilan el retrotransposón (**Figura 11**).

Estudios de la transmisión de este fenotipo a la descendencia han mostrado unos interesantes resultados. Cuando es el padre el que transmite el fenotipo es indistinto que éste sea amarillo o pseudo-agutí, se mantienen las proporciones previamente observadas en otros experimentos. Sin embargo, cuando la madre es quien transmite el alelo, se observa un mayor porcentaje de ratones amarillos cuando ésta es amarilla que cuando es pseudo-agutí (175). Esto sugiere algún fallo en los mecanismos para borrar las marcas epigenéticas introducidas durante la oogénesis. Se está produciendo por tanto, herencia transgeneracional vía los gámetos femeninos (173). Experimentos adicionales descartaron efectos maternos tales como el ambiente uterino (175). De forma muy interesante, las proporciones de crías amarillas se ven reducidas cuando se alimenta a la madre con dietas ricas en grupos metilo (como el ácido fólico o la vitamina B12), genisteína (proveniente de la soja) o etanol (174, 176–179). Se han encontrado otros alelos parecidos a este caso en ratones y siempre asociados a la inserción de algún elemento repetitivo (180, 181). En humanos, se han encontrado alelos donde la estación anual en el momento de la concepción, definía su estado epigenético (182).

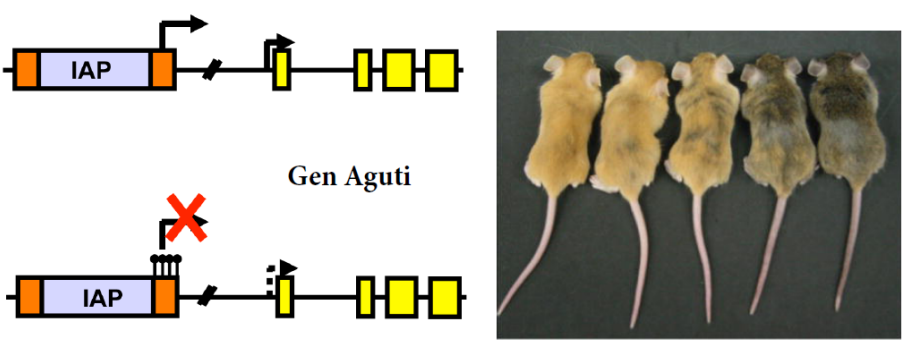


Figura 11: Regulación de la expresión del gen agutí
Derecha: El retrotransposón IAP tiene un promotor fuerte que si no está metilado (arriba) actúa incrementando la expresión del gen. Al metilarse (abajo), el gen tendrá su expresión normal dando lugar al fenotipo pseudo-agutí.
Izquierda: Imagen de los fenotipos posibles.
Adaptada de Dolinoy (179).

En resumen, se ha observado herencia epigenética transgeneracional en plantas y en algunos animales. No obstante, en mamíferos estos eventos son raros y debidos a la reprogramación de la línea germinal (183).

1.5-Epigenética y cáncer

Debido a la importancia de las marcas epigenéticas en procesos clave como la transcripción, reparación y replicación del ADN, las alteraciones de estos reguladores de la cromatina tienen drásticas consecuencias capaces de inducir y mantener varios tipos de cánceres (20). A continuación se comentan las alteraciones epigenéticas más estudiadas en los procesos tumorales.

1.5.1-Alteraciones relacionadas con las histonas

Las modificaciones postraduccionales de las colas de las histonas tienen un papel crítico en la estructura de la cromatina a la vez que regulan el reclutamiento, o exclusión, de proteínas con actividades enzimáticas remodeladoras. En numerosos tipos de cáncer puede observarse una desregulación tanto en las proteínas encargadas de la adición o eliminación de estas marcas como de las proteínas que las reconocen y se unen (20).

Las modificaciones de histonas más estudiadas en relación con procesos tumorales son la acetilación y la metilación. La acetilación se produce únicamente en lisinas, neutralizando su carga positiva y provocando una apertura de la cromatina haciéndola más accesible a factores de transcripción y a la maquinaria transcripcional. En concordancia, varios análisis de su distribución muestran que esta modificación se encuentra asociada a zonas activas como promotores y *enhancers* (142, 184). En su regulación intervienen dos familias de enzimas: las *KATs* (*histone lysine acetyltransferases*), encargadas de añadir el grupo acetilo a las lisinas, y las *HDACs* (*histone deacetylases*), que lo eliminan. El comportamiento aberrante de enzimas pertenecientes a ambas familias se ha estudiado en diversos tipos de cáncer. Las *KATs* fueron las primeras enzimas descritas que modificaban las histonas y su relación con el cáncer se estableció rápidamente (185). Por otro lado, la actividad represora de las *HDACs*, al desacetilar las histonas, se ha asociado a numerosos ejemplos de leucemia (186) y su inhibición en células tumorales provoca bloqueo del crecimiento, diferenciación y apoptosis (186, 187).

A diferencia de la acetilación, la metilación puede darse en argininas, lisinas e histidinas sin afectar a la carga general de la molécula. De entre éstas, la más estudiada es la metilación de la lisina, que puede encontrarse mono-, di-, o tri-metilada. Como en el caso de la acetilación, la desregulación de su adición por parte de la familia de proteínas *histone lysine methyltransferase* (*KMTs*) o su

eliminación mediante *histone lysine demethylases* (*KDMs*) se encuentra asociada a procesos tumorales (20, 188, 189). El descubrimiento de las *KDMs* alteró la percepción de la metilación de histonas como una marca estable y estática (190).

En general, las modificaciones de las colas de las histonas muestran un cambio dinámico que establece un complejo marco regulatorio dependiente del contexto celular. Interesantemente, se han descubierto situaciones donde marcas aparentemente excluyentes, indicadoras de fenómenos antagónicos de activación y represión, pueden observarse conjuntamente. Éste es el caso de los dominios bivalentes. Éstos presentan metilación de H3K27, asociada a represión, y metilación de H3K4, una marca de activación (137, 159). En células madre se ha observado que estos dominios se correlacionan con niveles bajos de expresión en genes relacionados con el desarrollo y linaje celular. Cuando estas células se diferencian, los dominios bivalentes mantienen únicamente una de las dos marcas. Por tanto, este mecanismo de dominios bivalentes permitiría a las células madre preservar su pluripotencia mediante el mantenimiento de una baja expresión de factores implicados en la diferenciación (26, 140, 159). Lo que resulta más interesante, es el descubrimiento de que el ADN de muchos genes que presentan dominios bivalentes se encuentra frecuentemente hipermetilado en cánceres (129, 191).

1.5.2-Alteraciones de la metilación

Por su importancia en la regulación de la transcripción genética, las alteraciones en la metilación tienen efectos graves en los procesos de desarrollo. Cuando los patrones de esta marca epigenética se ven afectados, se presentan patrones aberrantes de expresión que cambian el comportamiento celular. Así, los primeros estudios epigenéticos en tumores se centraron en el análisis de la metilación global. Inicialmente se encontraron dos situaciones con consecuencias opuestas: la hipometilación global y la hipermetilación de islas CpG (192, 193).

Hipermetilación de islas CpG

Se trata del evento más importante en el origen de muchos cánceres (192). Desde el momento en que se descubrió el primer gen hipermetilado en cáncer en 1987 (194), el número de genes silenciados por este mecanismo no ha parado de aumentar. Los genes afectados se encuentran en todas las vías celulares conocidas y algunos autores plantean patrones de metilación

específicos de tumores o hipermetilomas (195). Los estudios mediante secuenciación de nueva generación (*NGS*) del estado de la metilación muestran que entre un 5-10% de islas CpG en promotores se ven metiladas en varios tipos de cáncer (20). Esta desregulación incluye también la expresión de ARNs no codificantes, algunos de los cuales participan en el proceso tumorigénico (196).

Hipometilación

La pérdida de metilación del ADN en los dinucleótidos CpG fue la primera anormalidad genética identificada en células tumorales. La hipometilación puede afectar básicamente a dos regiones genómicas distintas: una parte mayoritaria del genoma formada por secuencias y elementos repetitivos y una porción minoritaria de islas CpG asociadas a genes silenciados del cromosoma X inactivo, genes imprintados o metilados específicos de tejido. En el primer caso se trata de una hipometilación global asociada a inestabilidad genética mientras que en el segundo hablamos de hipometilación regional y se encuentra relacionada con la reactivación transcripcional de genes silenciados durante el desarrollo y diferenciación.

Se conocen varios genes en diferentes tipos de cáncer que se desmetilan durante el proceso tumoral produciéndose una ganancia de función, como el gen HRAS en cáncer gástrico (197, 198) o el locus IGF2/H2 en tumores de Wilm´s, un caso de pérdida del imprinting (193, 199). En la actualidad sabemos que la hipometilación global se da en muchos tipos de tumores diferentes (200).

Así, como se ha mencionado, durante los procesos tumorales pueden producirse alteraciones en las diferentes piezas que configuran el paisaje epigenético. La determinación del estado de cada una de las marcas epigenéticas descritas resulta crítica para un correcto diagnóstico y pronóstico, así como para servir de base para en el desarrollo de nuevos fármacos epigenéticos.

1.6-Técnicas de análisis del estado de metilación

El interés surgido tras los primeros descubrimientos sobre el papel de la metilación en la expresión génica y el desarrollo (201, 202), promovieron la creación de nuevas técnicas para su caracterización. Éstas se basaban en alguna de las siguientes aproximaciones: separación cromatográfica, uso de

anticuerpos específicos, digestión mediante enzimas de restricción o conversión química mediante bisulfito (**Figura 12**).

La *RP-HPLC* (*Reversed-Phase High Performance Liquid Chromatography*), fue una de las primeras técnicas aplicadas. A principio de los años 80, esta técnica fue utilizada para la separación de las citosinas metiladas y desmetiladas y su posterior cuantificación, a distintas longitudes de onda (203). El método se mejoró más adelante con la inclusión de espectrometría de masas (204).

Durante el mismo periodo se lograron obtener anticuerpos específicos para la metilcitosina (205). Mediante su utilización se logró inmunoprecipitar y enriquecer la muestra en secuencias metiladas. Su uso fue adaptado años más tarde para el estudio de cambios globales en la metilación (206). Las técnicas basadas en estos anticuerpos han sido determinantes en la caracterización del estado de metilación durante las distintas fases del desarrollo embrionario (207).

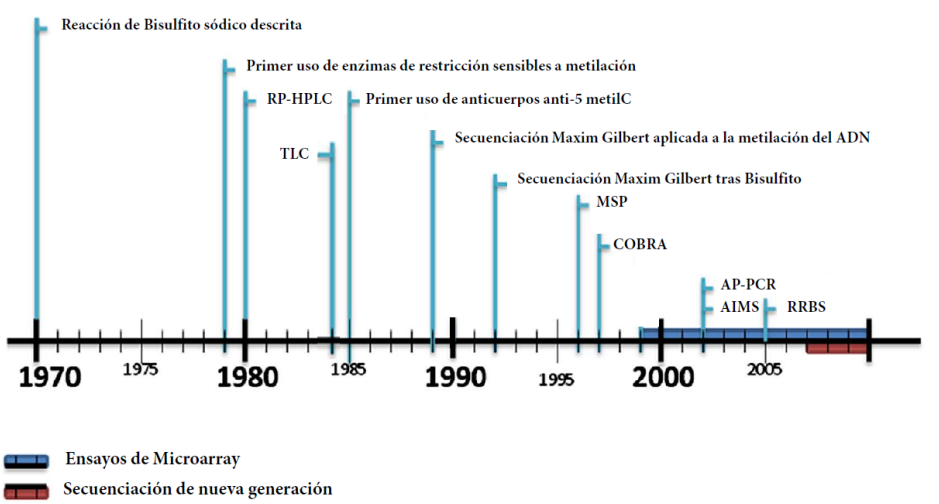


Figura 12: Principales tecnologías de análisis de metilación del ADN

Aparición de las tecnologías más relevantes en el análisis de la metilación del ADN. La mayoría se basan en el uso de anticuerpos, digestión enzimática o conversión química mediante bisulfito. El uso de microarrays y plataformas de secuenciación de nueva generación permiten la aplicación de estas metodologías de una forma masiva.

Adaptada de Parle-Mcdermott y Harrison (208)

Paralelamente, durante los primeros años de la década de los 80, se desarrollaron otras metodologías basadas en unas importantes proteínas, las enzimas de restricción. Estas enzimas, ampliamente utilizadas en biología molecular, son el mecanismo de defensa de las bacterias contra secuencias

víricas invasoras, como los fagos. Reconocen secuencias específicas normalmente palindrómicas, denominadas motivos, y cortan la cadena de ADN. No obstante, para proteger su propio material genético, las bacterias añaden una marca de protección a sus cadenas de ADN, la metilación (209). A pesar de ello, existen algunas enzimas que no se ven afectadas y son capaces de cortar la secuencia. Esto permite que, dada la alta redundancia existente entre muchos motivos reconocidos por estas proteínas, se puedan encontrar pares de ellas que reconocen la misma diana pero difieren en la sensibilidad a la metilación. El uso de estos pares, denominados isoesquizómeros, permitió a finales de los 70, distinguir entre regiones metiladas y desmetiladas utilizando los enzimas *HpaII* y *MspI* (210). Posteriormente, Bestor *et al.* (211) logran, en 1984, separar en dos fracciones las citosinas y las metilcitosinas de una muestra mediante *TLC* (*Thin Layer Chromatography*), tras una digestión con isoesquizómeros.

Con la ayuda de metodologías como la *RT-HPLC* y *TLC* se pudieron obtener ratios globales de metilación permitiendo la comparación del contenido de esta marca entre distintitos organismos (212, 213). No obstante, su aplicación queda limitada al no resultar válidas para la determinación específica de regiones. A pesar de ello, en la actualidad han recobrado importancia en la detección de niveles de hidroximetilcitosina (214, 215).

Por otro lado, la potencia de las técnicas basadas en enzimas de restricción se vio incrementada por la incorporación de otra metodología revolucionaria, la *PCR* (*Polymerase Chain Reaction*) (216). La *PCR* permite la amplificación de regiones delimitadas por las dianas de corte de las endonucleasas. Mediante la selección de pares de enzimas sensibles y no sensibles a la metilación, se pueden generar muestras enriquecidas en el tipo de fragmento que se desee estudiar. Aprovechando esta sinergia surgieron técnicas como la *AP-PCR* (*arbitrarily primed PCR*) que permitió obtener perfiles de metilación tanto de muestras normales como tumorales (217, 218) Otro método parecido, aunque más efectivo, es *AIMS* (*Amplification of Inter-Methylated Sites*) (219). El uso de un par de isoesquizómeros donde uno de ellos dejaba el punto de corte romo y el otro cohesivo, permite la adición de adaptadores específicos que incrementan la eficiencia de la amplificación mediante *PCR*. Se pueden obtener así, patrones de bandas característicos del metiloma de tejido normal y tumoral. La "huella dactilar" de cada muestra puede ser entonces comparada y las bandas diferenciales, aisladas y analizadas.

1.6.1-Conversión mediante bisulfito

Sin duda, el análisis de la metilación vivió una revolución con el establecimiento de una nueva metodología, la diferenciación entre citosinas y metilcitosinas gracias a una conversión química. Basándose en estudios de los 70 (220), se determinó que la citosina sometida a un tratamiento con bisulfito sódico, derivaba en uracilo y aunque la 5-metilcitosina podía transformarse también, lo hacía a una velocidad mucho más lenta (221)(**Figura 13**). Esta diferencia de reactividad permite distinguir entre los dos estados epigenéticos tras secuenciación (222). De esta forma, la elusiva marca epigenética queda fijada en forma de cambio genético.

La alteración en secuencia introducida permite la diferenciación y amplificación selectiva, hecho que aprovecha la *MSP* (*Methylation Specific PCR*) (223). Además, estas modificaciones en la secuencia pueden representar nuevas dianas para enzimas de restricción El análisis *COBRA* (*COmbined Bisulfite Restriction Analysis*) permite determinar la metilación según el patrón de bandas obtenido tras la digestión (224). Adicionalmente, el cambio a uracilo (timina tras la *PCR*) altera las propiedades de fusión de la región. Esto permite establecer los niveles de metilación con técnicas como la *MS-HRM* (*Methylation-Specific High Resolution Melting*) (225, 226).

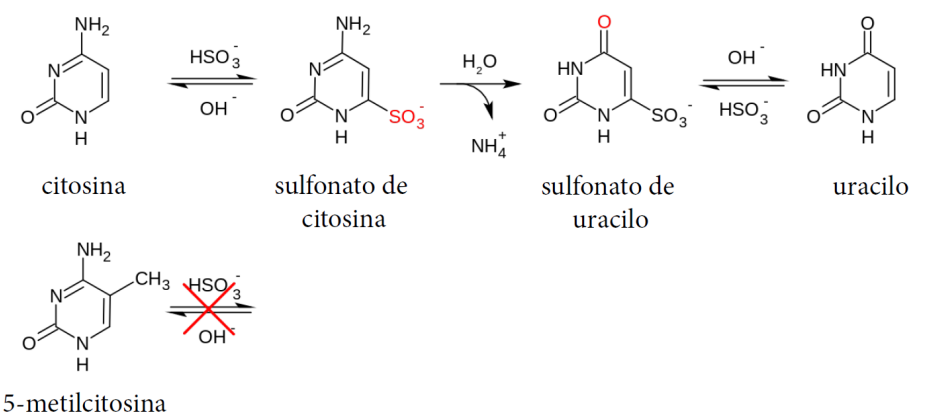


Figura 13: Conversión química mediante bisulfito
La citosina se transforma en uracilo tras el tratamiento con bisulfito mientras la 5-metilcitosina muestra una mayor resistencia al cambio.
Fuente: Wikimedia Commons (227).

1.6.2-Métodos de alto rendimiento

Las técnicas anteriores fueron vitales en la determinación del estado de metilación de numerosos genes en diversas patologías. No obstante, su alcance estaba limitado al estudio de regiones específicas y no permitían determinar la distribución global de la metilación. Esta limitación se vio superada con el desarrollo de métodos de alto rendimiento: en primer lugar los microarrays y en años más recientes, la secuenciación de nueva generación.

1.6.2.1-Microarrays

Esta nueva tecnología se desarrolló durante los años 90 y desde entonces, ha permitido el estudio de miles de posiciones genómicas simultáneamente. Los microarrays de ADN utilizados en los análisis de metilación, se construyen adhiriendo polinucleótidos sobre una superficie sólida. Éstos actúan como sondas sobre las que hibridan sus correspondientes secuencias complementarias. Por tanto, si la muestra contiene dicha secuencia en cantidades suficientes, se fijará al microarray y podrá ser detectada.

Los microarrays representan una tecnología de paralelización de análisis y por tanto, expanden las técnicas anteriores. Se han desarrollado aplicaciones que tienen como base los distintos enfoques ya presentados: uso de enzimas de restricción, inmunoprecipitación y tratamiento con bisulfito.

El uso combinado de los microarrays con enzimas de restricción sigue el siguiente esquema general: (i) las endonucleasas cortan el ADN en función de su estado de metilación, (ii) una PCR posterior permite enriquecer la muestra y añadir algún tipo de marcador (radioactivo o lumínico) y finalmente, (iii) se depositan en el microarray. Aquellas secuencias que encuentren una sonda, hibridarán y darán señal. En uno de los primeros ejemplos se utilizó esta metodología para estudiar la hipermetilación de islas CpG en células tumorales de cáncer de mama. El diseño utilizado permitía únicamente la amplificación de las regiones metiladas. Los microarrays utilizados contenían secuencias pertenecientes a 276 islas CpG (228). De esta forma se podía analizar y comparar el estado de éstas en varias muestras tumorales y sus controles. El bajo número de islas representaba una limitación pero a lo largo de los años, otros diseños fueron ampliando el número de regiones interrogadas hasta llegar a los modernos microarrays que cubren centenares de miles de secuencias.

Los métodos de inmunoprecipitación vivieron una revitalización con la llegada de los microarrays. El modo de trabajar sigue un esquema parecido al anterior. Sin embargo, el enriquecimiento se produce gracias a la acción de los anticuerpos específicos. Técnicas como *MeDIP* y *mDIP* (*Methylated DNA ImmunoPrecipitation* y *Methyl-DNA InmunoPrecipitation* respectivamente) capturan el ADN metilado tras su fragmentación, por sonicación o por vía enzimática, y lo marcan con un fluoróforo. La comparación con ADN genómico control, marcado con otro fluoróforo, permite, tras la hibridación en un microarray, observar los cambios de híper e hipometilación (229, 230).

Aunque los microarrays pueden analizar miles de posiciones al mismo tiempo, la transformación por bisulfito reduce la variabilidad de secuencias posibles. Esto hace muy difícil el diseño de sondas que capturen regiones únicas del genoma (231). Aunque aparecieron variantes para intentar solventar este problema, su aplicación quedaba restringida a genomas con características específicas (232). No obstante, Illumina ha desarrollado una nueva metodología, Illumina Infinium Methylation, que permite superar esta limitación. Para ello, utiliza unos microarrays, Illumina Beadchips, con un diseño diferente al tradicional. En este caso, las sondas cubren la región a analizar pero se detienen tras el dinucleótido CpG que queremos interrogar. Además, para cada posición, existen un par de tipos de sondas. En uno de ellos el final de la sonda acaba en CG y en otro, en CA. El primero es para el caso en el que el dinucleótido CpG estuviera metilado originalmente y el tratamiento con bisulfito no le hubiera afectado. El segundo en cambio, capturaría el caso de desmetilado que daría con una transformación a timina en la región complementaria. Tras la hibridación, se extiende la secuencia mediante el uso de una polimerasa. Esto únicamente ocurrirá donde haya complementariedad correcta en la posición CpG. Utilizando nucleótidos marcados se puede determinar el estado de metilación original. El primer Beadchip para metilación en humanos fue el HumanMethylation27K que cubría 27578 lugares CpG. Sin embargo, el último modelo, el 450K, supera los 450.000, permitiendo interrogar posiciones distribuidas por todo el genoma tanto dentro como fuera de islas CpG.

Estos métodos ofrecen una resolución altísima, llegando a la determinación de la metilación a nivel de base. No obstante, únicamente se pueden analizar aquellas posiciones cubiertas por alguna sonda. Además, el diseño de sondas únicas para regiones genómicas no es una tarea trivial ya que se producen

muchas uniones inespecíficas. Por estas razones, las principales limitaciones de estas técnicas radican en el número de sondas y su distribución.

1.6.2.2-Secuenciación de nueva generación

En 2004 Life Science sacó al mercado el primer secuenciador de nueva generación, basado en una nueva tecnología, la pirosecuenciación. Durante los siguientes años se desarrollaron numerosas plataformas con distintas aproximaciones para el análisis masivo (**Figura 14**). Esta nueva capacidad de procesamiento de muestras permite, a día de hoy, la secuenciación de genomas humanos completos en busca de alteraciones.

De forma similar a los microarrays, la secuenciación de nueva generación permite ampliar los límites de planteamientos anteriores y adaptar técnicas ya existentes a un modelo de alto rendimiento.

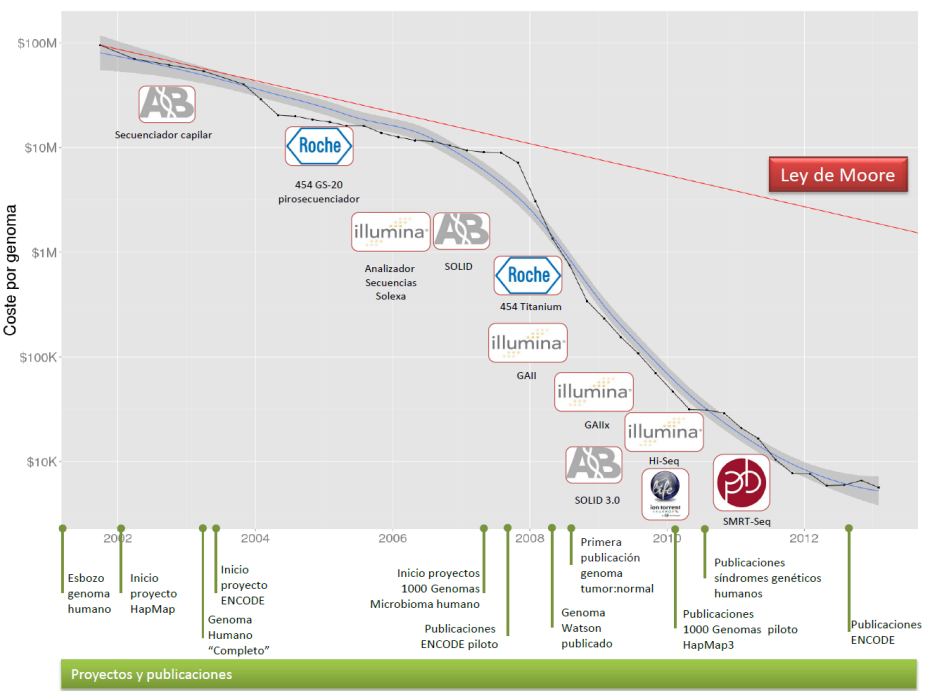


Figura 14: Evolución de los costes y las tecnologías de secuenciación de nueva generación
A lo largo de la última década han aparecido las principales tecnologías de *NGS* que han permitido que los costes por genoma humano secuenciado se reduzcan considerablemente (se compara con la ley de Moore). Importantes proyectos de secuenciación se han podido realizar gracias a estos avances. Datos obtenidos de http://www.genome.gov/sequencingcosts/.

La aplicación que más potencial ha demostrado para el análisis de metilación es la secuenciación de todo el genoma tras la transformación por bisulfito. Esta metodología, al igual que los microarrays de Illumina, ofrece una resolución a nivel de base. No obstante, a diferencia de éstos, el análisis no está limitado a unas regiones determinadas. El estudio del metiloma completo de *A. thaliana* mediante este método descubrió nuevos lugares de metilación al compararlo con resultados de microarrays (233). En 2009, Lister *et al.* publicaron el primer mapa de metilación de células humanas con este nivel de resolución (234). Se han estudiado numerosos modelos y líneas celulares adicionales y recientemente, el proyecto internacional Roadmap Epigenomics ha mostrado los resultados del análisis de más de un centenar de epigenomas humanos, la metilación de los cuales ha sido determinada, en gran medida, mediante secuenciación de nueva generación tras bisulfito (19).

Aunque se considera a este método el "gold standard" del estado del arte actual y los precios de secuenciación bajan año tras año (**Figura 14**), su aplicación no es realista para laboratorios que no dispongan de amplios medios económicos. Por este motivo, se han desarrollado otras metodologías que reducen el volumen de secuencias a analizar. *MeDIP-Seq* consiste en la secuenciación directa de la fracción inmunoprecipitada por la técnica *MeDIP*. De forma similar, la secuenciación masiva tras una digestión mediante enzimas de restricción es la base de la metodología *MRE-Seq* (*Methyl-sensitive Restriction Enzyme Sequencing*). Estas metodologías pueden considerarse complementarias al analizar distintas regiones del genoma (235). También han sido utilizadas por el proyecto *Roadmap Epigenomics* (19). Utilizando bisulfito, se adapta una metodología anterior, la *RRBS* (*Reduced representation bisulfite sequencing*) (236), a los nuevos métodos masivos. En este caso la reducción del número de secuencias se consigue mediante una digestión enzimática y una selección por tamaño.

La nueva generación de técnicas analíticas permite la obtención de una gran cantidad de datos epigenéticos. Conjuntamente a todas estas metodologías se han desarrollado numerosos procedimientos bioinformáticos que permiten su análisis. En el presente trabajo se utilizan estas técnicas computacionales para el análisis de datos provenientes de experimentos de nueva generación.

2-Objetivos

La metilación del ADN tiene un papel clave en la regulación epigenética y por tanto, su análisis es de gran interés para entender procesos biológicos tan importantes como la replicación del DNA, la expresión génica, la diferenciación celular o las bases moleculares de muchas enfermedades, incluyendo el cáncer. Como hemos visto en la introducción, se han desarrollado numerosas técnicas para el estudio de la metilación del ADN, tanto a nivel regional como global y genómico. El análisis de los 28 millones de CpGs en el genoma humano presenta un alto coste y dificultad, por lo que la mayoría de estudios a escala genómica utilizan diseños que reducen la complejidad del análisis. Para ello, se analiza únicamente una pequeña porción de las regiones genómicas de interés, una o unas pocas CpGs, y se extrapolan los resultados a la totalidad del elemento genético, por ejemplo una isla CpG. Aunque la mayoría de los estudios hacen una validación de la estrategia en unos cuantos loci, ninguno ha realizado una validación global. Con la aparición de los primeros datos de células humanas obtenidos mediante *WGBS* (*Whole Genome Bisulfite Sequencing*) (234, 237) es posible, por primera vez, evaluar la exactitud de los diferentes abordajes de complejidad reducida. Asimismo, es posible identificar CpGs cuya metilación tiene un comportamiento autónomo respecto a las adyacentes, lo que se podría interpretar como que tienen una funcionalidad propia (por ejemplo, son un lugar de unión de un factor de transcripción, cuya unión se regula por metilación de la CpG concreta).

En este escenario y partiendo de datos generados por WGBS en el genoma humano, se plantean los siguientes objetivos:

1. **Determinación de la variabilidad de la metilación dentro de las islas CpG.**

2. **Evaluación del valor informativo de la metilación de las dianas HpaII (CCGG) localizadas en islas CpGs respecto a la metilación global de la mismas.**

3. **Identificación de posiciones CpG con una metilación anómala con respecto a su entorno genómico cercano (metilación opuesta respecto a las CpGs adyacentes).**

4. **Caracterización de las propiedades funcionales de las posiciones CpG con metilación anómala.**

3-Materiales y Métodos

3.1-Evaluación a escala genómica de sitios CpG individuales como representantes del estado de metilación de islas CpG

El proceso de estudio (**Figura 15**) del valor informativo de las posiciones CpG individuales HpaII dentro de las islas CpG se divide en los siguientes bloques:

- Descarga y preprocesado de las lecturas de metilación.
- Obtención de las coordenadas y secuencias de las islas CpG y dianas HpaII.
- Cálculo del coeficiente de metilación de islas CpG y dianas HpaII.
- Análisis de homogeneidad de la metilación y correlación entre islas CpG y dianas HpaII.

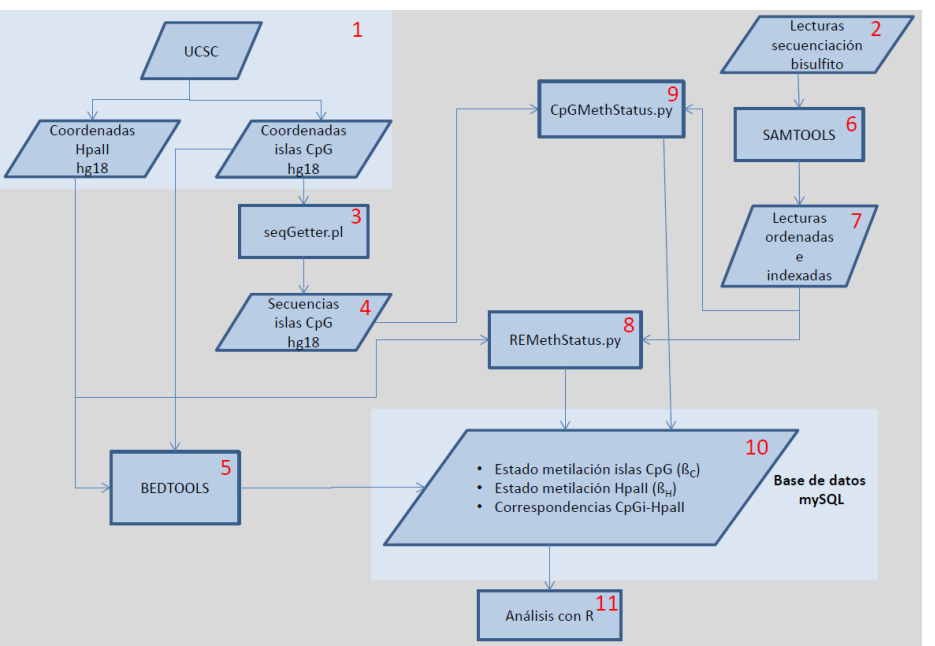


Figura 15: Esquema general del estudio de representatividad

1. Coordenadas de dianas HpaII e islas CpG obtenidas de las bases de datos del *UCSC Genome Browser* (versión hg18, http://genome.ucsc.edu). **2**. Lecturas de bisulfito descargadas de los estudios originales (234, 237). **3**. Script de Perl para la obtención de las secuencias de las islas CpG. **4**. Archivo de secuencias de las islas CpG. **5**. Uso de la función "intersectBED" del paquete de programas BEDTOOLS para asignar cada HpaII a su correspondiente isla CpG. **6**. Uso de las utilidades de SAMTOOLS para procesar las lecturas de bisulfito de archivos BAM a archivos BAM ordenados e indexados. **7**. Archivos BAM con información de las lecturas. **8**. Script de Python para el cálculo de la metilación de HpaII. Asigna valores de metilación ($ß_H$) a cada HpaII tomando como datos de entrada las coordenadas de las dianas (1) y las lecturas de bisulfito procesadas (7). **9**. Script de Python para el cálculo de la metilación de las islas CpG. Se reporta el coeficiente de metilación medio ($ß_C$) y la desviación estándar para cada isla CpG. Toma como datos de entrada las secuencias de las islas (4) para la búsqueda del dinucleótido CpG y las lecturas de bisulfito procesadas (7). **10**. Almacenamiento de los datos en una base de datos MySQL **11**. Análisis estadístico.

3.1.1-Obtención y procesado de los datos de metilación del ADN

Los datos de metilación utilizados en este trabajo provienen, en forma de lecturas, de dos estudios de secuenciación masiva de genomas de varias líneas celulares tras su tratamiento con bisulfito (234, 237). Una vez transformado el genoma, la secuenciación de alto rendimiento proporciona datos de metilación con una resolución a nivel de base nucleotídica.
Las líneas celulares analizadas corresponden a células madre embrionarias humanas H1, fibroblastos fetales pulmonares IMR90, células madre femeninas adiposas (ADSC) y adipocitos derivados de ADSC (ADS-Adi). En conjunto, estos estudios generaron 1,16, 1,18, 1,10 y 1,13 mil millones de lecturas para H1, IMR90, ADSC y ADS-Adi respectivamente (234, 237). Es importante mencionar que originalmente, las lecturas de ADSC y ADS-Adi se encontraban apareadas. Es decir, cada una tenía asociada otra, formando un par que cubría regiones muy próximas en el genoma. No obstante, en nuestro estudio fueron separadas y tratadas individualmente para obtener mayor profundidad en el análisis. Estas cuatro líneas fueron utilizadas en todos los análisis del estudio. Adicionalmente, en el análisis de correlación del estado de metilación entre las posiciones HpaII y las islas CpG, se estudiaron otras líneas celulares: células madre pluripotentes inducidas derivadas de ADSC (ADS-iPSC), tres líneas de iPSC derivadas de fibroblastos de prepucio (FF-iPSC 6.9, FF-iPSC 19.7, FF-iPSC 19.11), células madre embrionarias humanas H9, e iPSC derivadas de fibroblastos IMR90 (IMR90-iPSC) (237).

Según los estudios, todas las lecturas fueron mapeadas en el genoma humano de referencia (NCBI versión 36/hg18) utilizando el programa Bowtie (238). Las lecturas fueron descargadas de:

- http://neomorph.salk.edu/human_methylome/data.html
  (H1 e IMR90)
- http://neomorph.salk.edu/ips_methylomes/data.html
  (resto de líneas celulares)

El formato de las lecturas originales no permitía su análisis directo. Para su uso, fueron, en primer lugar, modificadas a un formato similar a SAM (239). Para ello se simularon los campos informativos que faltaban. Posteriormente, fueron transformadas al formato binario BAM (239) que permite su compresión para un uso rápido y eficiente. Finalmente, mediante el programa

SAMTOOLS (239), se ordenan e indexan. Los índices generados sirven de datos de entrada para los scripts de análisis (**Figura 15**).

3.1.2-Obtención de coordenadas y secuencias de las islas CpG y dianas HpaII

Las islas CpG utilizadas se ciñen a la definición clásica, tal y como se describió en la introducción (85):

- Tamaño igual o superior a 200 pares de bases.
- Contenido G + C igual o superior al 50%.
- La frecuencia de CpG (observado/esperado) igual o superior a 0,6.

Las coordenadas de las islas, así como la secuencia del genoma humano se descargaron del *UCSC Genome Browser*, versión hg18 (240). Adicionalmente, se estudiaron dos conjuntos de islas adicionales, unas anotadas experimentalmente (241) y otras mediante criterios bioinformáticos (86).

Las posiciones de las dianas HpaII y las secuencias de las islas CpG fueron obtenidas escaneando la secuencia genoma humano mediante scripts escritos en Perl. Cada diana HpaII fue asignada a su correspondiente isla CpG utilizando la función "intersectBed" del paquete de software BEDTOOLS (242).
En este estudio, únicamente se tuvieron en cuenta aquellas lecturas que solapaban total o parcialmente las islas CpG indicadas (**Tabla 4**).

3.1.3-Cálculo del coeficiente de metilación

El código de los scripts utilizados en este trabajo, así como archivos importantes, puede encontrarse en el repositorio *online*
github.com/vbarrera/thesis

Debido a la transformación por bisulfito, las citosinas desmetiladas se observan como timinas (T) en las lecturas, mientras que las citosinas metiladas, no se transforman (C). Por ello, el coeficiente de metilación se define como el ratio entre el número de citosinas y el número total de citosinas y timinas [nº C/(nºC+nºT)] (**Figura 16**). Esta definición es equivalente a la del valor ß utilizado en los microarrays de metilación (243) y

el rango de sus valores oscila entre 0 (no metilación) a 1 (totalmente metilado). Por otro lado, el dinucleótido CpG contiene dos citosinas, una en cada cadena, y las lecturas pueden cubrir de forma diferente cada una de ellas (p.ej. si todos las lecturas que cubren un CpG son de 5' a 3', únicamente se obtiene información de una de ellas, en este caso, la cadena +). Por este motivo, cada citosina del dinucleótido CpG se trató como una posición genómica individual y se le asignó su propio coeficiente de metilación.

El coeficiente de metilación se calculó para las islas CpG ($ß_C$) y para las dianas de restricción de HpaII ($ß_H$) mediante scripts de Python (CpGMethStatus.py y REMethStatus.py).

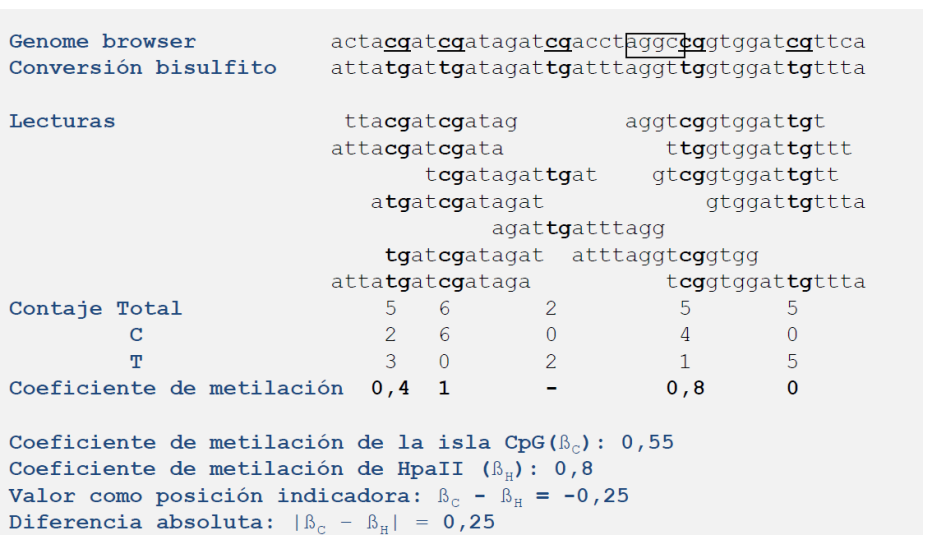


Figura 16: Ejemplo de cálculo de los coeficientes de metilación

Los dinucleótidos CpG están subrayados y la diana HpaII enmarcada. Los cálculos se realizaron para las dos cadenas de ADN (únicamente se muestra una en el ejemplo) e integrados en una única medida para cada CpG. El ejemplo contiene 4 posiciones CpG informativas y una no informativa. Se calculan los valores de los distintos parámetros utilizados en el estudio. Se muestra una posición HpaII discordante, estando ella hipermetilada (0,8) respecto a su correspondiente isla CpG (0,55).

En primer lugar, el script de Python escanea la secuencia de la isla CpG buscando el motivo CG. Una vez encontrado, le asigna los correspondientes nucleótidos a esa posición de acuerdo a las lecturas ya indexadas (**Figura 16**), utilizando la librería de Python pysam. En este proceso, se tienen en cuenta ambas cadenas, Watson y Crick, del ADN. Para evitar sesgos en el cálculo del coeficiente de metilación, las posiciones con menos de 5 lecturas fueron descartadas al no considerarse suficientemente informativas. Adicionalmente, se añadió un filtro extra para eliminar del análisis posibles posiciones mutadas

a otros nucleótidos. Las posiciones cuya suma C+T era inferior a 5, también fueron descartadas. Una vez aplicados estos filtros, se obtuvo la misma cobertura que la mostrada en análisis previos (244).
Cierto número de islas CpG contenían más de una diana HpaII en su secuencia (**Tabla 4**). En estos casos, se utilizó el valor medio de la metilación del conjunto de dianas HpaII. Por otra parte, el valor de una posición como indicadora del estado de metilación se calculó como la diferencia de su valor ($ß_H$), o media de valores en caso de ser varias, con el valor de metilación de su isla CpG correspondiente ($ß_C$).

Para cada isla CpG se obtuvieron los valores de la media y la desviación estándar (SD) teniendo en cuenta tan solo las posiciones válidas. Únicamente las islas con un nivel aceptable de informatividad fueron utilizadas. Para ser incluidas, un mínimo del 25% de sus dinucleótidos CpG debían tener al menos 5 lecturas. Esto resultó en un número de islas CpG que oscilaba entre el mínimo de 10.693 islas CpG informativas en la línea celular H1 y más de 26.000 en las muestras ADS-Adi. El número total de dianas HpaII incluidas en el análisis se encuentra entre 32.153 en las células H1 y 250.000 en ADS-Adi (**Tabla 4**). Cabe destacar que pueden observarse importantes diferencias en el número de islas CpG y dianas HpaII informativas entre los pares H1/IMR90 y ADSC/ADS-Adi. Estas diferencias se deben a la distinta cobertura de lecturas existente entre los dos estudios. No obstante, los resultados obtenidos no muestran ningún sesgo por este motivo.
Adicionalmente y con la intención de comprobar la robustez de la posición indicadora cuando se dispone de un bajo número de lecturas cubriendo las islas, se realizó un segundo análisis con islas CpG cuyo único requisito era tener al menos una posición informativa. En estas condiciones, más de 21.000 islas contenían al menos una diana HpaII, cubierta por al menos 5 lecturas (**Tabla 4**).

Por otro lado, para evaluar posibles sesgos debidos a la distribución bimodal de la metilación, se realizó un estudio de aleatoriedad. Éste consistió en una redistribución al azar de todos los valores de metilación dentro de las islas CpG. En primer lugar, se obtuvieron y almacenaron los valores de metilación de todas las posiciones CpG válidas dentro de islas CpG. Posteriormente, en cada isla CpG del estudio, se substituían los valores de metilación de sus posiciones CpG por valores al azar del conjunto almacenado. De esta forma, se generó un catálogo virtual de islas CpG que se correspondían en tamaño y cobertura a las originales. Los nuevos valores de metilación de las posiciones

CpG se utilizaron para calcular su correspondiente ß$_C$. Las dianas HpaII hipotéticas fueron seleccionadas al azar entre las posiciones CpG de cada isla. Esta simulación se realizó únicamente con los datos de H1 e IMR90.

3.1.4-Análisis de homogeneidad y correlación

Los valores de metilación de las islas CpG y las dianas HpaII, junto a datos estructurales y descriptivos como las coordenadas genómicas y los ratios O/E fueron almacenados en una base de datos MySQL. Esto permite un acceso rápido a los datos, tanto de valores de metilación como a qué isla CpG corresponde cada dinucleótido CpG (**Figura 15**, **Figura 17**).

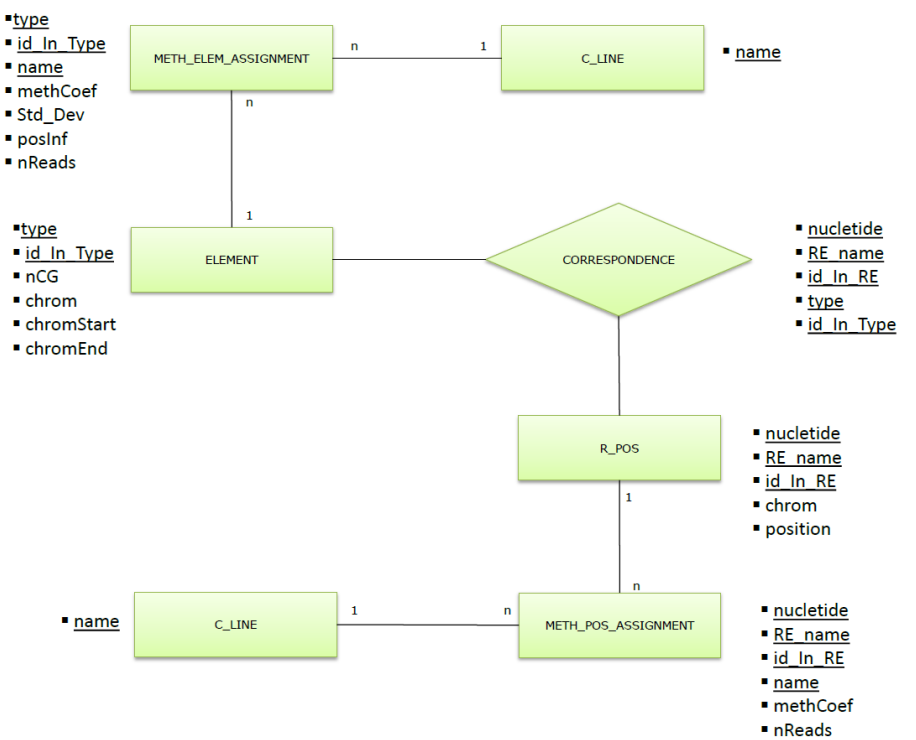


Figura 17: Esquema de la base de datos utilizada
Tablas utilizadas en la base de datos MySQL. Se muestran los atributos de cada una y las relaciones entre tablas. Los atributos primarios de cada tabla aparecen subrayados.

Por otro lado, los cálculos de homogeneidad y correlación, así como los gráficos derivados, se realizaron utilizando el software estadístico R. El modo *jitter* fue utilizado en los gráficos de dispersión para facilitar su visualización, dada la gran cantidad de puntos. El análisis completo se realizó tanto con el subconjunto de islas CpG con alta cobertura como con todas las islas que

contuvieran al menos una lectura (**Tabla 4**). La mayoría de los análisis produjeron resultados idénticos o muy parecidos, tanto al utilizar el subconjunto filtrado como el que no.

Tabla 4: Número de lecturas, islas CpG y dianas HpaII informativas consideradas en este estudio

Líneas celulares	H1		IMR90		ADS-Adi		ADSC	
Nº de lecturas totales [1]	1.154.658.045		1.183.875.099		1.131.768.326		1.098.572.398	
Nº de lecturas consideradas [2]	2.490.724		3.378.352		8.755.104		8.162.058	
	Cobertura Alta[3]	**Todas** [4]	**Cobertura Alta** [3]	**Todas** [4]	**Cobertura Alta** [3]	**Todas** [4]	**Cobertura Alta** [3]	**Todas** [4]
Nº de islas CpG informativas	10.693	25.770	17.094	26.617	26.092	26.742	26.019	26.719
Nº de dianas HpaII informativas	32.153	68.622	77.417	113.343	251.516	252.521	249.053	250.110
Islas CpG con dianas HpaII informativas[5]	9.271	21.002	15.776	24.038	25.220	25.567	25.141	25.521
Islas CpG con 1 diana HpaII informativa	2.154	5.838	2.320	3.849	1.624	1.784	1.612	1.774
Islas CpG con 2 dianas HpaII informativas	2.168	4.925	2.773	4.263	2.253	2.327	2.262	2.366
Islas CpG con >2 dianas HpaII informativas	4.949	10.239	10.683	15.926	21.343	21.456	21.267	21.381

[1]De Lister *et al.* (234, 237).

[2]Solapantes con islas CpG.

[3]Islas CpG con >25% de sitios CpG informativos. Un sitio CpG se considera informativo cuando está cubierto por al menos 5 lecturas.

[4]Todas las islas CpG consideradas con al menos una posición CpG informativa (≥5 lecturas informativas).

[5] Número de islas CpG con una diana HpaII: 26.508

3.2-Identificación a nivel genómico de secuencias regulatorias mediante detección de estados de metilación anómalos en sitios CpG

El esquema general (**Figura 18**) de la identificación de secuencias regulatorias puede dividirse en los siguientes bloques:

- Descarga y preprocesado de las lecturas de metilación.
- Obtención de la secuencia del genoma de referencia así como las coordenadas de elementos genómicos de interés.
- Obtención de los perfiles de metilación, identificación de posiciones anómalas y comparación entre distintas líneas.
- Determinación de las características epigenéticas de las regiones que contienen las posiciones anómalas.

3.2.1-Obtención de los datos de metilación y genómicos

Los datos referentes al estado de metilación a nivel genómico pertenecen a células madre embrionarias humanas H1 y fibroblastos fetales pulmonares IMR90, ambas analizadas mediante secuenciación de alto rendimiento tras tratamiento por bisulfito (234). Tal y como se ha mencionado en el bloque anterior, estos estudios generaron 1,16 y 1,18 mil millones de lecturas para H1 e IMR90 respectivamente. Las lecturas fueron mapeadas en el genoma humano de referencia (NCBI versión 36/hg18) utilizando el programa Bowtie (238). Fueron descargadas de:

- http://neomorph.salk.edu/human_methylome/data.html

De forma similar al trabajo anterior, las lecturas fueron procesadas a un formato similar a SAM y transformadas al formato binario BAM (239). Posteriormente, fueron ordenadas e indexadas utilizando el programa SAMTOOLS (239). Los índices generados sirven de datos de entrada para los scripts de análisis.
Adicionalmente se descargaron, de la misma fuente, los valores de expresión de 26.430 transcritos (que corresponden a 17.980 genes) obtenidos mediante *mRNA-Seq* (**Figura 18**).

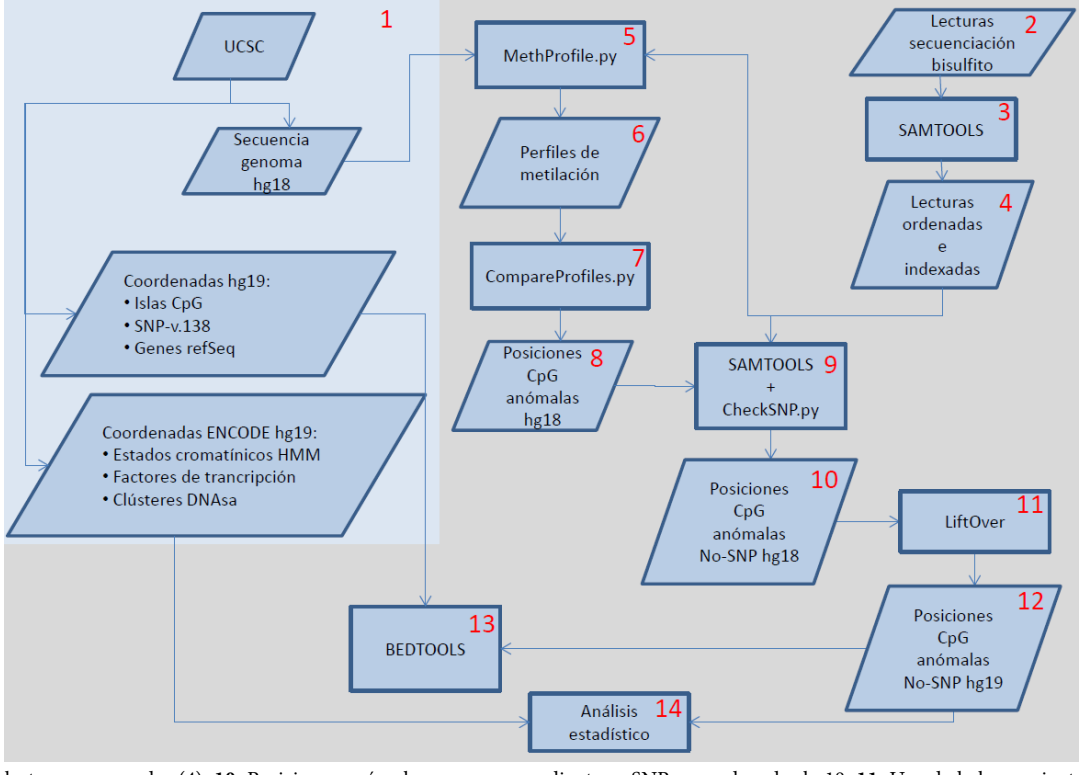


Figura 18: Esquema general del estudio de identificación de secuencias regulatorias

1. Secuencia del genoma, coordenadas de elementos genómicos (islas CpG, genes refSeq, SNPs v.138) y coordenadas de estudios del proyecto ENCODE (Estados cromatínicos basados en modelos de Markov ocultos, zonas de unión de factores de transcripción y clústeres de zonas hipersensibles a DNAsa) obtenidas de las bases de datos del *UCSC Genome Browser* (versión hg18 para el genoma y hg19 para el resto de sets de información, http://genome.ucsc.edu). **2**. Lecturas de bisulfito descargadas del estudio original (234). **3**. Uso de las utilidades de SAMTOOLS para procesar las lecturas de bisulfito de archivos SAM a archivos BAM ordenados e indexados **4**. Archivos BAM con información de las lecturas. **5**. Script de Python para la generación de los perfiles. Utiliza como datos de entrada la secuencia del genoma (1) y las lecturas procesadas (4). **6**. Perfiles de metilación. **7**. Script de Python que realiza la comparación de patrones entre perfiles (6). **8**. Posiciones anómalas, coordenadas hg18. **9**. Uso de la función "pileup" de SAMTOOLS y procesado con script de Python para comprobar si las posiciones anómalas son realmente SNPs o no. Utiliza como datos de entrada las coordenadas de las posiciones anómalas (8) y las lecturas procesadas (4). **10**. Posiciones anómalas no correspondientes a SNPs, coordenadas hg18. **11**. Uso de la herramienta liftOver, del *UCSC Genome Browser*, para el cambio de coordenadas a hg19. **12**. Posiciones anómalas no correspondientes a SNPs, coordenadas hg19. **13**. Uso de las funciones "intersectBED" y "closestBED" del paquete de programas BEDTOOLS para asignar a cada posición anómala su correspondiente elemento genómico (isla, gen o SNP) que lo contenga o el más cercano. **14**. Análisis estadístico de enriquecimiento.

Las islas CpG utilizadas se ciñen la definición clásica (85) (ver bloque anterior para detalles). Sus coordenadas, junto a las de los genes pertenecientes a *refseq* (245) y las de los *SNPs* de la versión 138 (*allSNP*) de la base de datos dbSNP (246), se descargaron del *UCSC Genome Browser*, versión hg19 (240). Adicionalmente, se obtuvieron de la misma fuente, las posiciones de estados cromatínicos definidos mediante modelos de Markov (247), clústeres de hipersensibilidad a DNAsa y de los lugares de unión de 161 factores de transcripción en 91 tipos celulares, obtenidos mediante *ChIP-seq*. Estos tres últimos conjuntos de datos pertenecen al proyecto ENCODE (14).

Dado que las lecturas del estudio original habían sido mapeadas con la versión del genoma hg18, se utilizó dicha versión en la generación de los perfiles genómicos. No obstante, para los análisis de enriquecimiento se utilizaron datos más modernos, versión hg19, que contienen mayor número de elementos. Para poder cruzar ambos conjuntos de datos, una vez obtenidas las posiciones anómalas, se realizó un cambio de coordenadas entre las versiones de ensamblajes genómicos hg18 y hg19, mediante la utilidad del *UCSC Genome Browser*, *liftOver*.

3.2.2-Generación de perfiles cuaternarios genómicos

El código de los scripts utilizados en este trabajo, así como archivos importantes, puede encontrarse en el repositorio *online*
github.com/vbarrera/thesis

Los perfiles de metilación para los genomas de las líneas celulares H1 e IMR90 se generaron mediante un script de Python (MethProfile.py). Aprovechando la diferencia en secuencia que se produce por la transformación por bisulfito, se definió el coeficiente de metilación como [nº C/(nºC+nºT)] (ver bloque anterior para detalles). El script de Python escanea el genoma buscando el motivo CG. Una vez encontrado, calcula el coeficiente de metilación en función de los nucleótidos de las lecturas, utilizando la librería de Python pysam y teniendo en cuenta ambas cadenas del ADN. Posteriormente, asigna un valor de un código de 4 posibilidades en función del valor del coeficiente de metilación (**Figura 19**):

- 1 si el coeficiente de metilación es igual o superior a 0,8
- 0 si el coeficiente de metilación es igual o inferior a 0,2

- X si el coeficiente de metilación se encuentra entre 0,2 y 0,8
- N si no hay suficiente lecturas informativas.

Para evitar falsos positivos, la búsqueda de posiciones individuales anómalas requiere un alto grado de informatividad en la determinación del coeficiente de metilación. Por ese motivo, en este estudio se descartaron aquellas posiciones con menos de 10 lecturas. Adicionalmente, para minimizar el número posiciones anómalas debidas a cambios en la secuencia no debidos a metilación, las posiciones cuya suma C+T sea inferior a 10, fueron también descartadas.

De esta forma se obtiene para cada cromosoma un perfil de metilación, en forma de cadena de texto, formado por valores de un código cuaternario. Cabe mencionar no obstante, que únicamente se procesaron los cromosomas autosómicos. Dada la especial regulación epigenética que tienen los cromosomas X e Y durante el desarrollo, no se les incluyó en el análisis.

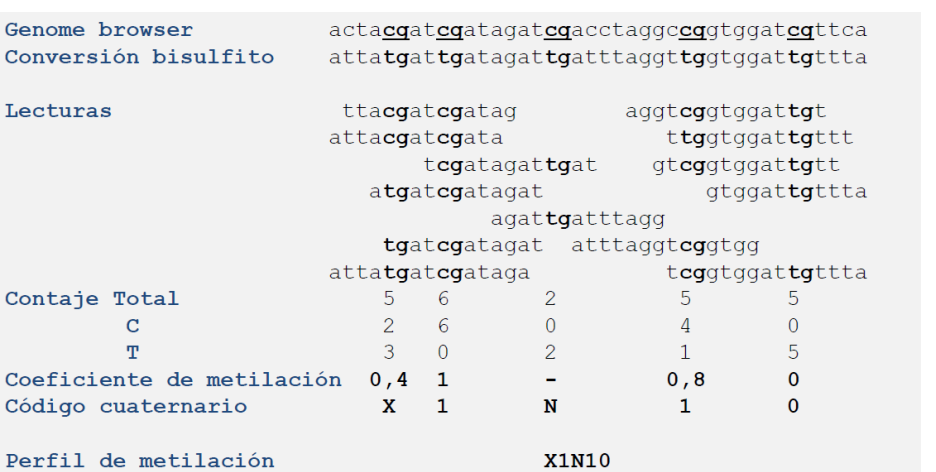


Figura 19: Ejemplo de obtención del perfil de metilación

Los dinucleótidos CpG están subrayados. Los cálculos se realizaron para las dos cadenas de ADN (únicamente se muestra una en el ejemplo) e integrados en una única medida para cada CpG. En este ejemplo se muestra, por simplicidad, un caso para un filtro de 5 y no de 10. Bajo estas premisas, se muestran 4 posiciones CpG informativas y una no informativa. Se obtiene el coeficiente de metilación según [nºC/(nºC+nºT)]. Posteriormente se asigna, a cada sitio CpG, un valor del código cuaternario en función del coeficiente de metilación (1 si Coef.≥0,8; 0 si Coef. ≤0,2; X si 0,8>Coef. <0,2 y N si no es informativa)

3.2.3-Búsqueda de secuencias anómalas

Una vez obtenidos los perfiles de cada cromosoma, éstos fueron analizados mediante un script de Python (profileFrequencies.py) para obtener de cada

uno, la frecuencia de aparición de secuencias ininterrumpidas formadas por valores de metilación extremos (totalmente desmetilados, 0, o totalmente metilados, 1). El script recorre el perfil posición a posición identificando las secuencias que se encuentran limitadas por valores de X o N. Se obtiene así una relación de las diferentes secuencias junto a la cantidad de veces que aparecen en los perfiles de metilación de H1 e IMR90 (**Figura 20**).

Adicionalmente, se realizó un estudio de aleatoriedad para poder estudiar posibles sesgos en los resultados, debidos a la distribución bimodal de la metilación (ver trabajo anterior para detalles). Para ello, en primer lugar, se obtuvieron todos los valores de 0s y 1s de cada perfil. Posteriormente, éstos fueron reasignados al azar entre las posiciones originales que contenían 0s y 1s. De esta forma, se obtienen unos perfiles de metilación virtuales con una distribución aleatoria de la metilación.

Tras el análisis de la distribución de los tamaños y su frecuencia, se seleccionó un tamaño de secuencia de 11 sitios CpG y se identificaron como posiciones anómalas de interés aquellas posiciones centrales que presentan un comportamiento totalmente opuesto a los sitios CpG flanqueantes:

- Desmetilación en entorno metilado 11111011111.
- Metilación en entorno desmetilado 00000100000.

La primera secuencia presenta 8.872 y 4.882 apariciones en H1 e IMR90 respectivamente. Por otro lado, la segunda, no aparece en el perfil de H1 y en IMR90, únicamente en 11 ocasiones.

```
Perfil de metilación              X1N10001XXX01XN10NXN11100XN01N10X00XNN
Secuencias                         1 10001   01  10   11100  01 10 00

Secuencia      N°apariciones          Longitud N°apariciones
1                      1                 1          3
01                     2                 2          2
10                     2                 5          1
00                     1
10001                  1
```

Figura 20: Ejemplo del procesado de los perfiles

Los perfiles son escaneados en busca de secuencias formadas por sitios CpG con estados totalmente metilados (1) o desmetilados (0). Se obtiene así una relación de todas las secuencias presentes y la frecuencia de las longitudes.

3.2.4-Comparación de las posiciones anómalas entre líneas celulares

Se comparó el estado de metilación de las posiciones anómalas entre las líneas H1 e IMR90 para identificar posiciones con una posible función reguladora putativa. Para ello, se utilizó un script de Python (compareProfiles.py). El script escanea un primer perfil (A) buscando la secuencia deseada. Diversas instancias solapantes de una misma secuencia son mostradas individualmente (p.ej. En **111110<u>11111</u>**<u>011111</u> aparece la secuencia 11111011111 en dos ocasiones, en negrita y subrayado). Una vez identificadas las coordenadas de la secuencia dentro del perfil (A), se interroga al segundo perfil (B) utilizando esas coordenadas para determinar qué secuencias aparecen (**Figura 21).**

```
Perfil de metilación H1         0011XXNN11111011111XNN11111011111NXXN010

Posiciones                                  [9-19]        [23-33]

Perfil de metilación IMR90      0011XXNN11111011111XNN11111111111NXXN010

Posiciones      Comparación H1>IMR90
[9-19]          11111011111>11111011111
[23-33]         11111011111>11111111111
```

Figura 21: Ejemplo de comparación de perfiles

En este ejemplo, el perfil (A) es el correspondiente a la línea celular H1. En primer lugar se obtienen las coordenadas, dentro de ese perfil para la secuencia deseada, en este caso 11111011111. Una vez obtenidas, se obtienen las secuencias correspondientes a esas coordenadas en el perfil (B), correspondiente a la línea celular IMR90.

Una vez identificadas las posiciones anómalas, se realizó un análisis con las lecturas originales para descartar posibles SNPs que explicaran el valor anómalo de metilación observado. Para ello se utilizó la función "pileup" del programa SAMTOOLS (239) que permite obtener los valores de las lecturas de cada posición. Debido a que la mutación puede aparecer en ambos nucleótidos del dinucleótido CpG, el script de Python (checkTableSNP.py) los analiza por separado. Para saber si se trata o no de una mutación y no un cambio provocado por el bisulfito, se analiza el nucleótido complementario a la citosina de interés. Un número elevado de adeninas (A) indica que la timina (T) observada es debida a un cambio polimórfico. En cambio, si se observa un valor elevado de guaninas (G), la obtención de la timina es debida a la transformación por bisulfito de una citosina (C) desmetilada.

3.2.5-Determinación de las características epigenéticas de las regiones con posiciones anómalas

Para encontrar los elementos genómicos más cercanos a las posiciones, se utilizaron las funciones "intersectBed" y "closestBed" del paquete de software BEDTOOLS (242).

Por otro lado, para los cálculos y gráficos derivados, se utilizó el software estadístico R. En los análisis de enriquecimiento en elementos genómicos (estados cromatínicos, sitios de hipersensibilidad o lugares de unión de factores de transcripción) se realizaron test de permutaciones mediante el paquete *regioneR* (248). Dicho paquete genera, en cada permutación, una nueva población de análisis del mismo tamaño que la original, tomando las muestras de una población control. Una vez generada, analiza el número de veces que las posiciones intersecan con los elementos genómico de interés. En cada test de enriquecimiento se realizaron 5.000 permutaciones. De esta forma, se genera la distribución poblacional del número de intersecciones. Ésta se utiliza posteriormente para establecer la significancia de la población original en el análisis. En los test realizados, la población control utilizada consistió en todas las posiciones que cambian su estado en la comparación H1 e IMR90, un total de 1.017.790 posiciones. En la realización de estos test se utilizó una versión del genoma hg19, *BSgenome.Hsapiens.UCSC.hg19.masked*, que tiene las regiones repetitivas, como los centrómeros, enmascaradas. De esta forma, se evitan posibles sesgos dada la intensa metilación de estas regiones heterocromatínicas.

Finalmente, los análisis de lugares de unión de factores de transcripción se realizaron con el paquete de software MEME (249). Se utilizaron las funcionalidades AME (*Analysis of Motif Enrichment*"), MEME (*Multiple Em for Motif Elicitation*) y Tomtom. La primera permite analizar enriquecimiento en motivos de factores de transcripción entre dos conjuntos de secuencias. La base de datos de motivos utilizadas fue JASPAR (250). La segunda funcionalidad permite descubrir posibles motivos comunes en un conjunto de secuencias. Finalmente, Tomtom permite comparar los motivos descubiertos con bases de datos de motivos para ver si comparte parecido con alguno. Para este fin se ha utilizado la base de datos HOCOMOCO v9 (251). Dicha base de datos contiene 426 motivos redundantes de *homo sapiens* para 401 factores de transcripción.

4-Resultados

4.1-Evaluación a escala genómica de sitios CpG individuales como representantes del estado de metilación de islas CpG

4.1.1-Caracterización de la heterogeneidad de la metilación del ADN en las islas CpG

La exploración y caracterización de la metilación de todas las dianas CpG de una región requiere el uso de metodologías costosas en términos de tiempo y dinero. Por este motivo, en la mayoría de estudios donde se determina la metilación del ADN, se utilizan técnicas de análisis basadas en la determinación de uno o unos pocos sitios CpG. La principal hipótesis de trabajo que se asume es la homogeneidad de la metilación en la región a analizar. Dada la importancia que tiene esta asunción como pilar de estas metodologías, nuestro primer objetivo es la caracterización del estado de metilación a lo largo de nuestras regiones de interés, las islas CpG.

Mediante el script CpGMethStatus.py se obtuvo el valor del coeficiente de metilación, $ß_C$, de todas las islas CpG del estudio. Podemos observar que los valores se distribuyen de forma bimodal, con la mayoría de los valores cercanos a 0 (totalmente desmetilada) o 1 (totalmente metilada) y algunos valores intermedios (**Figura 22**). Los resultados muestran un enriquecimiento en islas CpG metiladas en células H1, en comparación con el resto de líneas celulares, tal y como se había indicado en estudios anteriores (234). No obstante, esta diferencia desaparece al utilizar en el análisis todas las islas CpG cuya única restricción es tener al menos una lectura (**Figura 22**). Esto indica un sesgo dependiente de metilación en la muestra H1 ya que, aparentemente, las islas CpG metiladas tienen mayor cobertura de lecturas que las desmetiladas.

El valor $ß_C$ de una isla CpG representa el valor general de su metilación. No obstante, es necesario estudiar la metilación que presentan el conjunto de sitios CpG que la componen para determinar su variabilidad y comprobar si $ß_C$ es una medida estable. Podemos representar la heterogeneidad de la metilación de una isla CpG como la desviación estándar (*SD*) del coeficiente de metilación que existe entre los dinucleótidos CpG contenidos en dicha isla (**Figura 23**). Si obtenemos esos valores para todas las islas del estudio, la

mayor variabilidad se observa en una pequeña población de islas CpG con valores intermedios de metilación (**Figura 23**). Este hecho indica que los valores intermedios se deben a estados de metilación alternantes (entre totalmente desmetilado, 0, y totalmente metilado ,1) de las posiciones individuales y no a un valor intermedio homogéneo entre todos los dinucleótidos CpG. Para analizar si la desviación estándar dependía del valor de metilación total de la isla, se exploró la distribución de la variabilidad en función de los niveles de metilación. En base a los resultados anteriores (**Figura 23**), se esperaba que aquellas islas CpG altamente metiladas o desmetiladas mostraran una menor heterogeneidad. Así se pudo confirmar en las islas CpG metiladas, con los niveles más bajos de variabilidad interna (**Figura 24**). Sorprendentemente, las islas desmetiladas exhibían una alta homogeneidad en las células H1 pero una distribución más amplia en el resto de muestras, indicando un perfil de metilación más relajado. Las islas CpG metiladas de forma intermedia mostraron los mayores niveles de variabilidad interna, pero al igual que en el caso anterior, en las células H1 era inferior al resto (**Figura 24**).

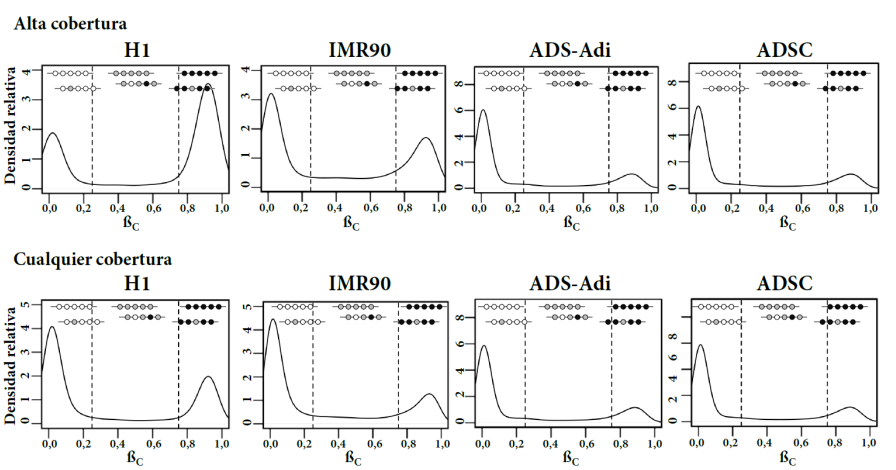


Figura 22: Distribución del coeficiente de metilación de islas CpGs con alta cobertura (paneles superiores) y todas las islas sin importar cobertura (paneles inferiores) en las cuatro líneas celulares

Las líneas verticales de puntos delimitan las áreas del gráfico conteniendo islas CpG desmetiladas ($ß_C$<0,25) y metiladas ($ß_C$>0,75). Perfiles ilustrativos de la metilación del ADN en islas CpG se representan en el gráfico mediante diagramas *lollipop*, donde cada circulo vacío representa sitios CpG desmetilados, los grises representan metilación parcial y los negros, sitios completamente metilados.

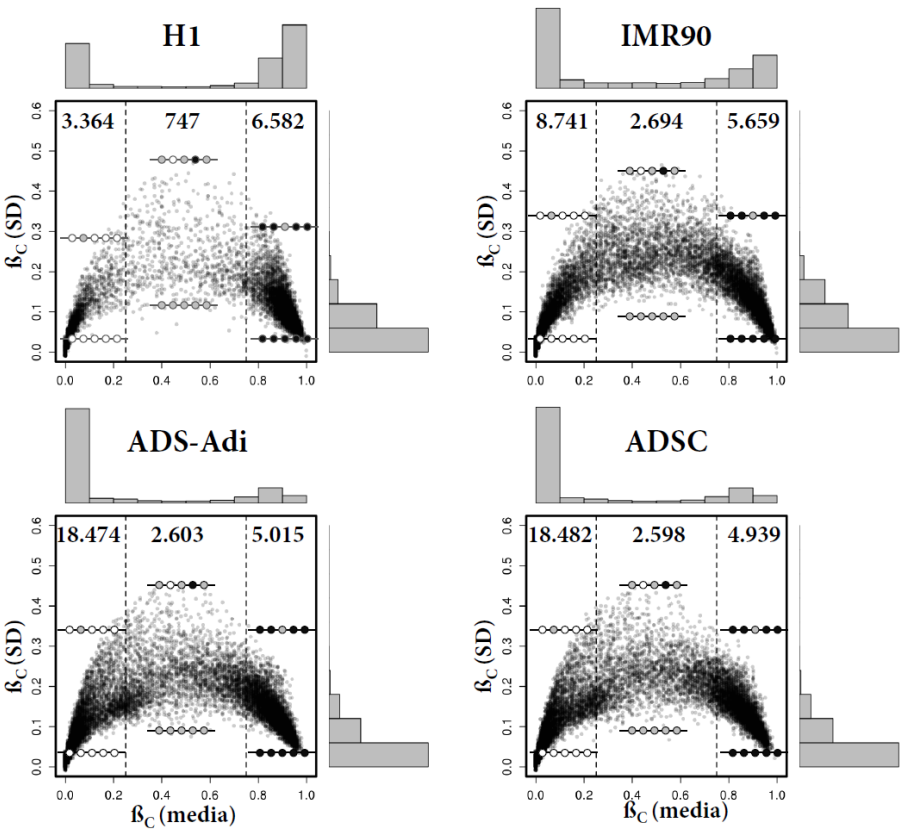


Figura 23: Homogeneidad de la metilación de los sitios CpG en las islas CpG

La media de todas los sitios CpG informativos que se localizan dentro de cada isla CpG (Coeficiente de metilación de la isla CpG, ßC) se representa contra la desviación estándar (*SD*) para las cuatro líneas celulares analizadas en este estudio. Las líneas verticales de puntos delimitan las áreas del gráfico que contienen islas CpG desmetiladas (ßC<0,25) y desmetiladas (ßC>0,75). Se muestran el número de puntos representados en cada área del gráfico e histogramas de su distribución en ambos ejes. Perfiles ilustrativos de la metilación del ADN en islas CpG se representan en el gráfico mediante diagramas *lollipop*, donde cada circulo vacío representa sitios CpG desmetilados, los grises representan metilación parcial y los negros, sitios completamente metilados. Únicamente se muestran islas CpG con alta cobertura.

En conjunto, estos resultados indican una alta homogeneidad en los perfiles de metilación de las islas CpG, especialmente los de las islas metiladas. Los datos obtenidos permiten afirmar que los valores de metilación total de las islas CpG, ßC, representan de una forma fiable los valores medios de metilación de los dinucleótidos que contienen. No obstante, las islas altamente desmetiladas y, especialmente, las intermediamente metiladas, exhiben diferentes niveles de heterogeneidad. Este hecho puede sugerir que, para un pequeño número de islas, los dinucleótidos CpG individuales pueden no ser representativos de su perfil global de metilación.

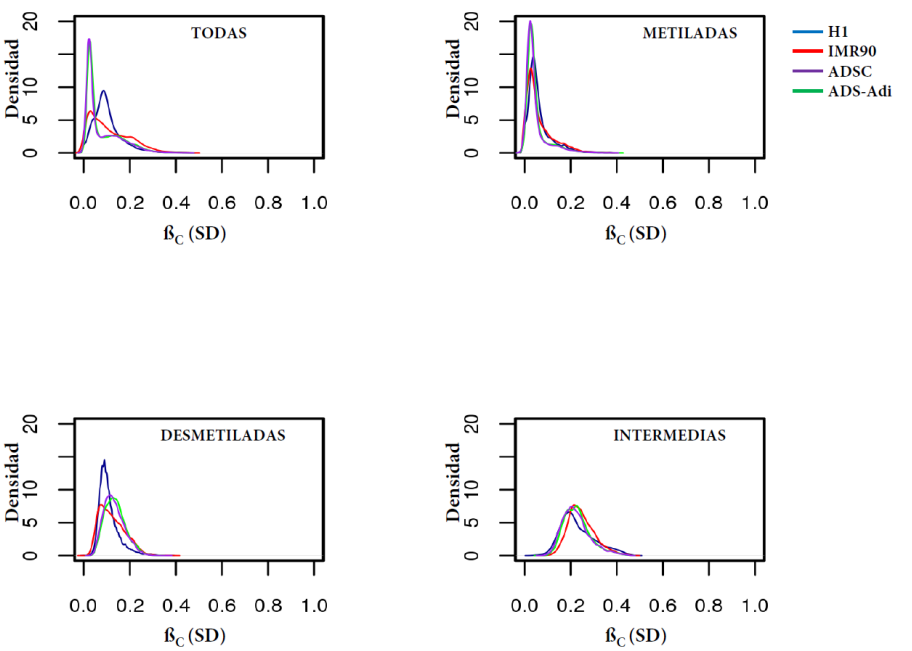


Figura 24: Gráficos de densidad de la desviación estándar (*SD*) del coeficiente de metilación de las islas CpG

Densidad de la distribución de la desviación estándar en grupos de islas CpG definidos por su coeficiente de metilación: METILADAS $ß_C > 0,75$; DESMETILADAS, $ß_C < 0,25$; INTERMEDIAS, $ß_C$: 0,25 a 0,75.

4.1.2-Metilación de las dianas HpaII del ADN como representantes de la metilación de la isla CpG

Una vez determinada la homogeneidad de las islas, el siguiente paso para dar validez a las técnicas habituales de análisis de la metilación es estudiar el valor predictivo de un sitio CpG individual contenido en una isla CpG. Se seleccionó la diana de restricción HpaII (CCGG) debido a su amplio uso en técnicas específicas de *locus* y de escala genómica.

Mediante el script REMethStatus.py se determinó el coeficiente de metilación de las dianas HpaII, $ß_H$, del estudio. Un total de 31.153, 77.417, 249.053 y 251.516 dianas HpaII localizadas dentro de las islas CpG preseleccionadas eran informativas en células H1, IMR90, ADSC y ADSC-Adi respectivamente (**Tabla 4**). La simetría de la metilación del ADN en las dianas HpaII se confirmó comparando los coeficientes de metilación de las dos cadenas, calculados por separado. En este estudio únicamente aquellas dianas HpaII

con 5 o más lecturas informativas en al menos una cadena se consideraron con alta cobertura y fueron incluidas en el análisis.

Con los datos de metilación de las dianas HpaII y sus correspondientes islas CpG se analizó su asociación. A nivel global, se observó una excelente correlación entre el coeficiente de metilación de HpaII, $ß_H$, y su correspondiente isla CpG (H1, r = 0,96, P<10^{-15}; IMR90, r = 0,93, P<10^{-15}; ADSC, r = 0,94, P<10^{-15}; ADS-Adi, r = 0,94, P<10^{-15}) (**Figura 27**). También se obtuvieron valores altos de correlación (r > 0,94) para el resto de líneas celulares (**Figura 28**). Análisis posteriores de los datos revelaron que, en las islas CpG metiladas de todas las líneas celulares analizadas, las dianas HpaII tienden a estar hipermetiladas (diferencias entre los coeficientes de metilación, $ß_C$ - $ß_H$ <0) al compararlas con el coeficiente global de metilación de su respectiva isla CpG (**Figura 25**). Por el contrario, no se observa este sesgo en islas CpG desmetiladas o intermediamente metiladas.

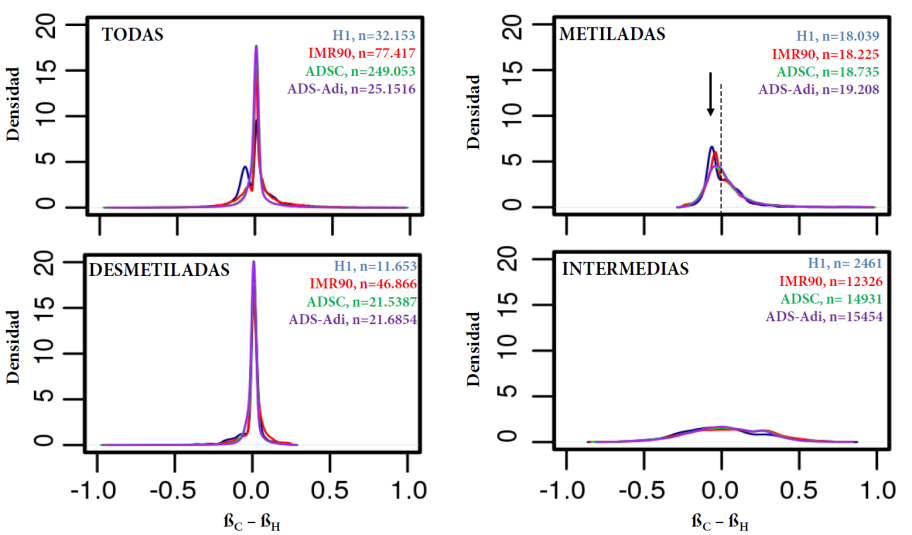


Figura 25: Gráficos de densidad de la diferencia entre el coeficiente de metilación de la isla CpG y su respectivo coeficiente de metilación de HpaII

Se observa una distribución bimodal en las células H1 (azul) sugiriendo que una subpoblación de sitios HpaII tiende a estar hipermetilada, comparada con su isla CpG. Una exploración posterior de la distribución en función del estado de metilación de la isla CpG (METILADAS, DESMETILADAS, INTERMEDIA) reveló una ligera hipermetilación en los sitios HpaII (comparados con sus respectivas islas CpG) en las islas CpG metiladas (METILADAS, ver flecha).

Con la intención de comprobar si los resultados de correlación no eran debidos a la mayor presencia de ciertos niveles de metilación en las muestras, se realizó una simulación computacional (ver materiales y métodos). La

aleatorización de los sitios CpG de todas las islas CpG virtuales mostró que no había correlación entre las dianas HpaII hipotéticas y sus correspondiente islas CpG (r = 0,30 y r = 2,22 con los datos de H1 e IMR90 respectivamente), demostrando que la correlación observada no se debe a las distribución bimodal de los niveles de metilación del ADN (**Figura 26**).

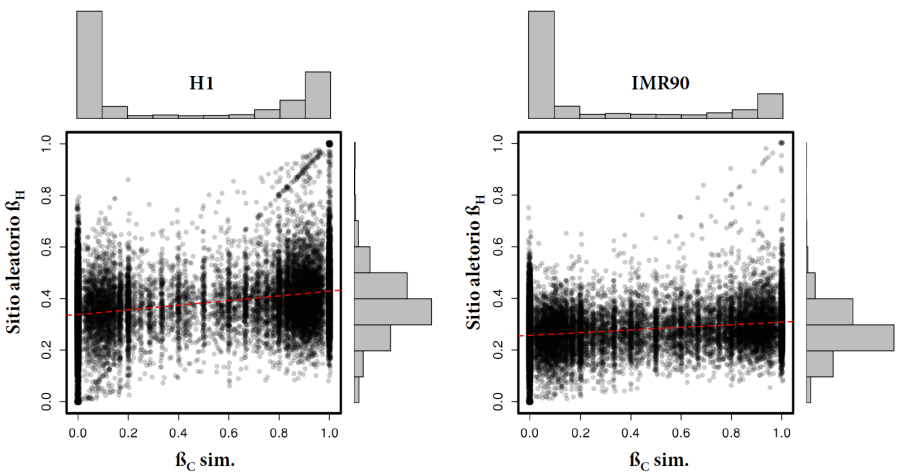


Figura 26: Estudio de aleatoriedad con islas CpG virtuales
Gráficos de correlación entre los coeficientes de metilación de dianas de restricción HpaII simuladas (sitio aleatorio $ß_H$) e islas CpG simuladas (sim $ß_C$) utilizando datos relativos a las dos líneas celulares H1 e IMR90. Se ha añadido una línea que representa el modelo lineal.

4.1.3-Extensión a otras definiciones de islas CpG

Los resultados mostrados hasta el momento se corresponden con el análisis de las islas CpG anotadas en el *UCSC Genome Browser*, una fuente de referencia en información genómica, que se ajustan a la definición clásica (85). No obstante, otros estudios han propuesto nuevos métodos y criterios para definir las islas CpG. Con la intención de ampliar nuestros resultados a otras definiciones de islas CpG, se han considerado dos de estos estudios, uno basado en la captura experimental de islas (93) y otro en la aplicación de modelos de Markov ocultos (86). En ambos casos, se mantiene una alta correlación (**Figura 29**). Es importante destacar que el conjunto de islas CpG identificadas experimentalmente (93), pero que no estaban anotadas en el *UCSC Genome Browser*, mostraban una correlación similar a pesar de contener una proporción en islas CpG metiladas mayor que las islas que se ceñían a la definición clásica (**Figura 29**).

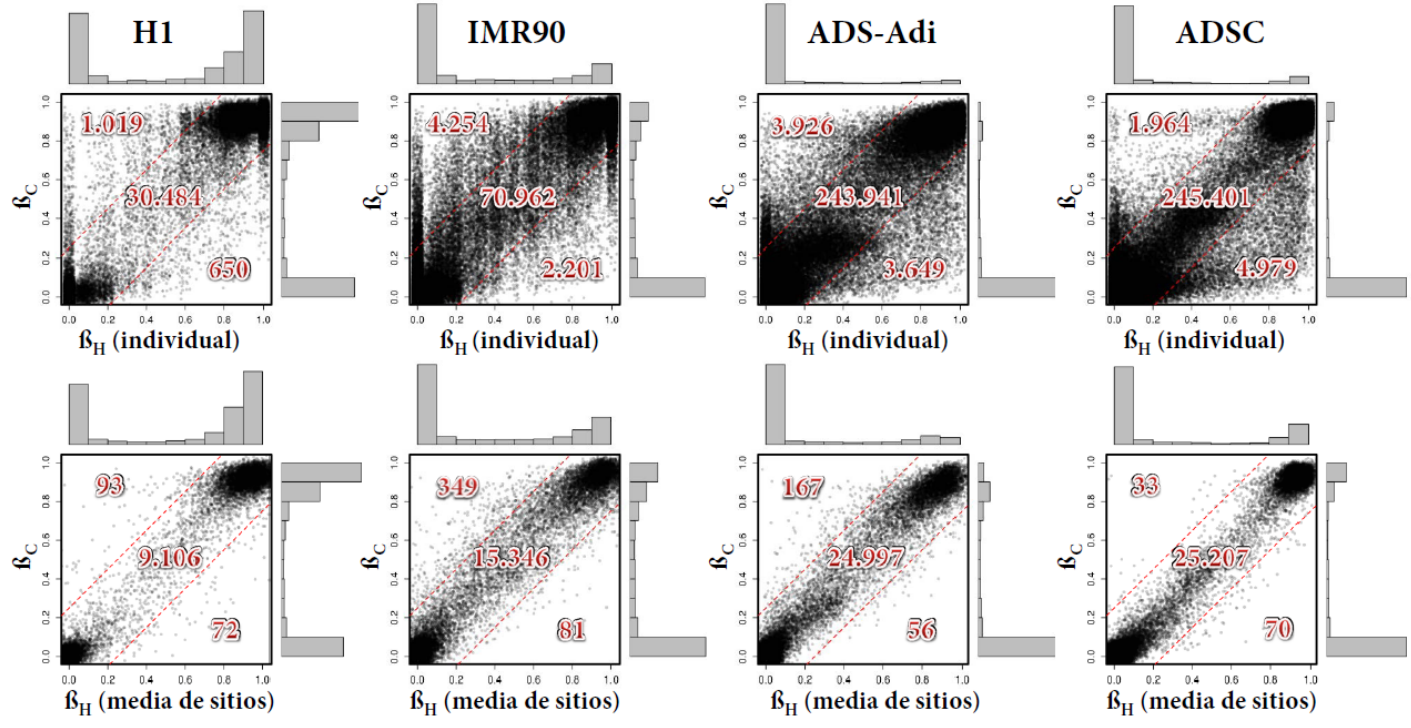


Figura 27: Gráficos de correlación para los coeficientes de metilación entre las HpaII ($ß_H$) y sus correspondientes islas CpG ($ß_C$) en muestras H1, IMR90, ADS-Adi y ADSC

Las líneas de puntos delimitan las áreas con diferencias >0,25 entre los coeficientes de metilación del sitio HpaII y su correspondiente isla CpG. Se muestran el número de puntos representados en cada área del gráfico e histogramas de su distribución en ambos ejes. Los paneles superiores muestran la correlación para sitios HpaII individuales con sus respectivas islas CpG. Los paneles inferiores muestran las mismas correlaciones pero comparando la media de todos los sitios HpaII en una isla CpG dada. El número de islas CpG informativas para cada línea celular se muestra en la **Tabla 4**.

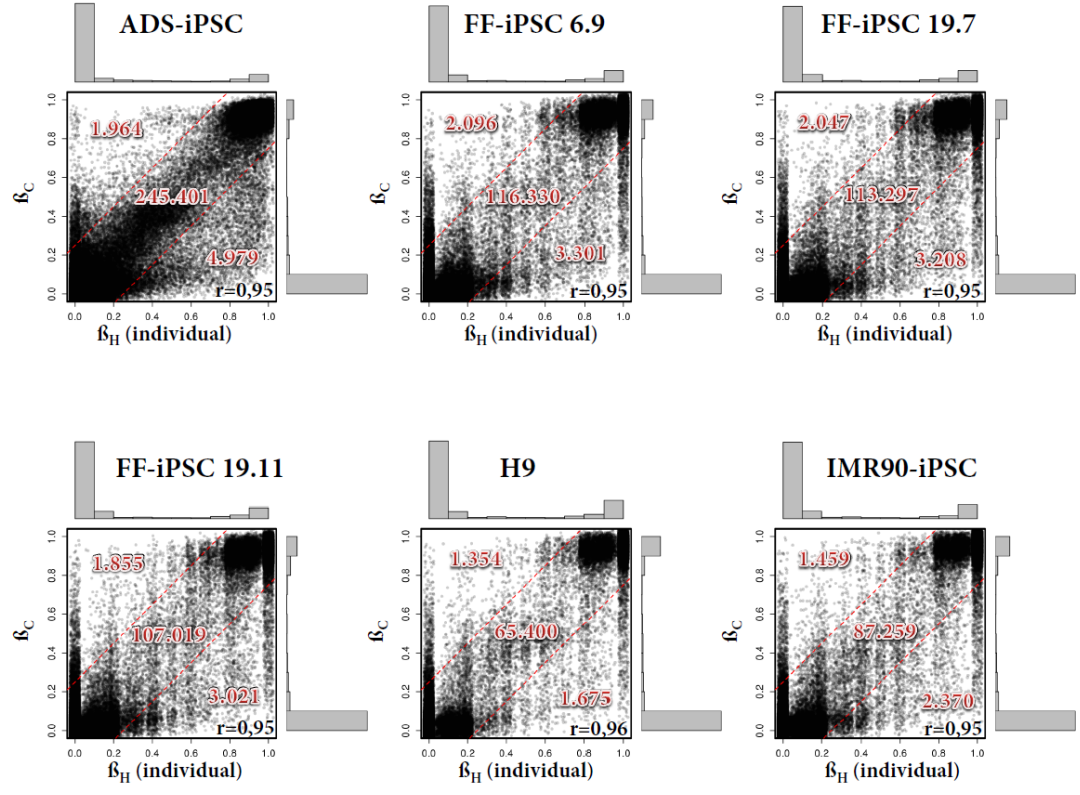


Figura 28: Gráficos de correlación para los coeficientes de metilación entre las HpaII ($ß_H$) y sus correspondientes islas CpG ($ß_C$) para ADS-iPSC, FF-iPSC, H9 y IMR90-iPSC

Las líneas de puntos delimitan las áreas con diferencias >0,25 entre los coeficientes de metilación del sitio HpaII y su correspondiente isla CpG. Se muestran el número de puntos representados en cada área del gráfico e histogramas de su distribución en ambos ejes. Únicamente las islas CpG con alta cobertura fueron consideradas.

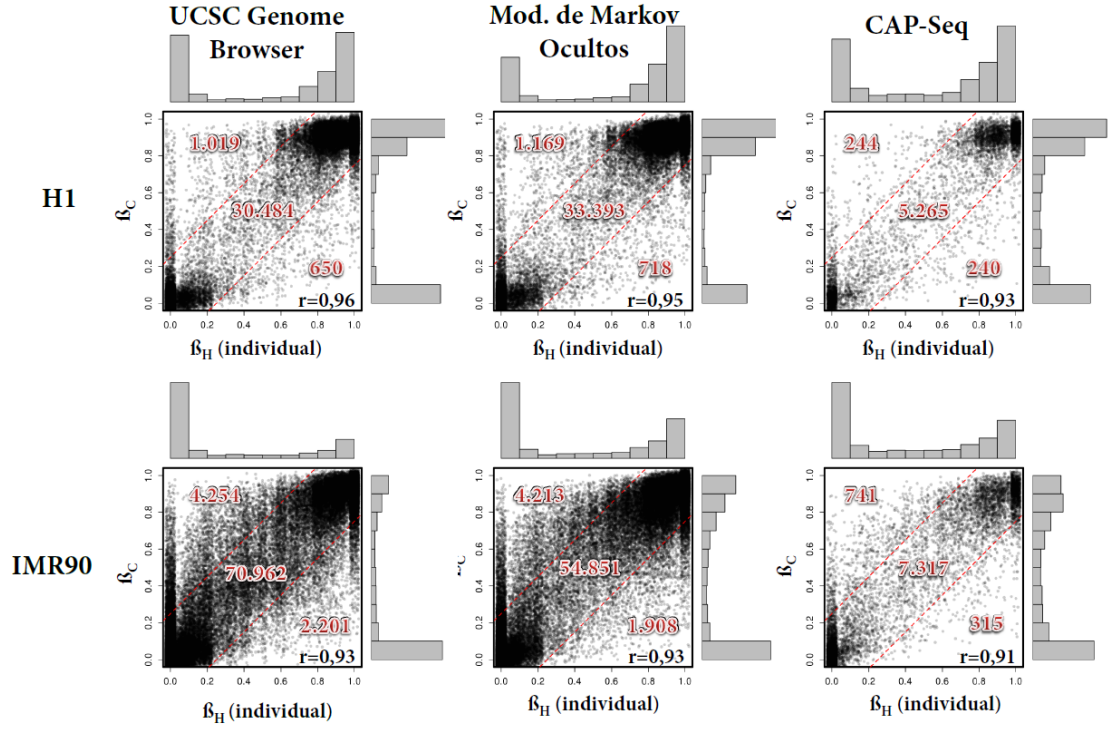
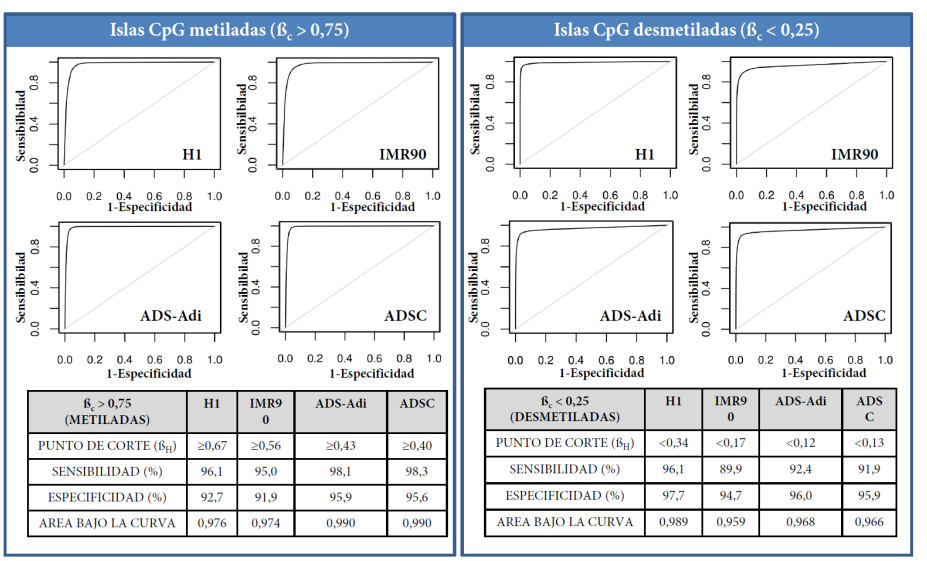


Figura 29: Gráficos de correlación para los coeficientes de metilación entre las HpaII ($ß_H$) y sus correspondientes islas CpG ($ß_C$), según varios criterios de definición, en las líneas celulares H1 e IMR90

Correlación entre ($ß_H$) y ($ß_C$) en islas CpG definidas por el criterio clásico (*UCSC Genome Browser*), modelos de Markov ocultos (86), y detectadas mediante *CAP-Seq* (93). En este último caso, únicamente se muestran las islas CpG no incluidas en la lista del *UCSC Genome Browser*. Las líneas de puntos delimitan las áreas con diferencias >0,25 entre los coeficientes de metilación del sitio HpaII y su correspondiente isla CpG. Se muestran el número de puntos representados en cada área del gráfico e histogramas de su distribución en ambos ejes. Únicamente las islas CpG con alta cobertura fueron consideradas.

4.1.4-Valor predictivo de la diana HpaII

En muchos estudios de metilación del ADN los datos se indican como marcas binarias (metilado/desmetilado). Por ese motivo, se realizó una evaluación cualitativa del valor predictivo de las dianas HpaII. Mediante un análisis de curva de Característica Operativa del Receptor (*ROC*) se definieron los puntos de corte óptimos y la exactitud en las predicciones de la metilación de la isla CpG, en función de la metilación de la diana HpaII (**Figura 30**). En todos los casos el área bajo la curva estaba por encima de 0,95 indicando un alto valor diagnóstico. El punto de corte de $ß_H$ para islas CpG metiladas ($ß_C$>0,75) oscilaba entre ≥0,40 y ≥0,67 y para islas CpG desmetiladas ($ß_C$<0,25), entre ≥0,34 y ≥0,12. En todos los casos, la sensibilidad y la especificidad eran ≥90% (**Figura 30**).



ß$_c$ > 0,75 (METILADAS)	H1	IMR90	ADS-Adi	ADSC
PUNTO DE CORTE (ß$_H$)	≥0,67	≥0,56	≥0,43	≥0,40
SENSIBILIDAD (%)	96,1	95,0	98,1	98,3
ESPECIFICIDAD (%)	92,7	91,9	95,9	95,6
AREA BAJO LA CURVA	0,976	0,974	0,990	0,990

ß$_c$ < 0,25 (DESMETILADAS)	H1	IMR90	ADS-Adi	ADSC
PUNTO DE CORTE (ß$_H$)	<0,34	<0,17	<0,12	<0,13
SENSIBILIDAD (%)	96,1	89,9	92,4	91,9
ESPECIFICIDAD (%)	97,7	94,7	96,0	95,9
AREA BAJO LA CURVA	0,989	0,959	0,968	0,966

Figura 30: Análisis del valor predictivo del coeficiente de metilación de HpaII ($ß_H$) respecto al coeficiente de metilación de su respectiva isla CpG ($ß_C$) mediante curva de Característica Operativa del Receptor (*ROC*)

Para realizar el análisis, se considera que unas isla está metilada cuando el valor medio de todos los coeficiente de metilación de los sitios CpG que contiene, $ß_C$ >0,75. Por otro lado, se la considera desmetilada cuando $ß_C$ <0,25.

Adicionalmente, se realizó una evaluación de la cantidad de sitios discordantes tomando la diferencia absoluta entre los dos coeficientes, ($|ß_C - ß_H|$), como referencia de concordancia entre las dianas HpaII y sus correspondientes islas CpG. Si se establece un criterio arbitrario de ($|ß_C - ß_H|$) >0,25 como valor límite de concordancia, 5,2% (1.169 de 32.153), 8,3% (6.455

de 77.417), 3,13% (7.802 de 249.053) y 3,01% (7.575 de 251.516) de las dianas HpaII mostraron resultados discordantes en las células H1, IMR90, ADSC y ADS-Adi respectivamente (**Figura 27**). Cuando el valor límite se establece en >0,5, la proporción de datos no coincidentes se reduce a 0,9% (300 de 32.153), 1,3% (1.034 de 77.417), 0,52% (1.288 de 249.053) y 0,50% (1.249 de 251.516) respectivamente.

Las técnicas de análisis de la metilación que utilizan enzimas de restricción pueden beneficiarse de las múltiples ocurrencias que se dan en una misma isla CpG. Por ese motivo, se analizó la mejora que suponía la inclusión de varias posiciones HpaII. En aquellas islas con más de una diana HpaII informativa, la exactitud de la predicción de $ß_C$ mejoró utilizando la media de todos los sitios HpaII en lugar de uno. Tal y como se esperaba, la media de $ß_H$ muestra una mejor correlación con la metilación de la isla CpG (r>0,98) (**Figura 27**, **paneles inferiores**) y la diferencia entre $ß_C$ y $ß_H$ de dianas individuales se redujo dramáticamente al utilizar la media de la metilación de HpaII (Figura **31**). A nivel cualitativo los puntos discordantes (aquellos cuya diferencia en los coeficientes de metilación es superior a 0,25) fueron reducidos a 1,8, 2,7, 0,95, y 0,89% en H1, IMR90, ADSC y ADS-Adi, respetivamente. Al utilizar un límite de 0,50, estos valores se reducen a 0,2, 0,1, 0,05 y 0,04%.

A pesar de la baja proporción de sitios discordantes, se observó una alta recurrencia, especialmente en las muestras con alta cobertura (**Figura 32**). Por ejemplo, la comparación de 3 muestras mostró un solapamiento del 42% para 2 o más muestras y un 16% para las 4 muestras (diferencia absoluta >0,25). Esto representa un enriquecimiento extraordinario ya que el número total de sitios HpaII informativos es de más de 250.000 y el número de sitios discordantes por muestra está alrededor del 3%. Si la discordancia estuviera distribuida de forma aleatoria se debería esperar aproximadamente 7 sitios HpaII discordantes en las 3 muestras frente a las 2.268 observadas. Esto representa un enriquecimiento superior a 300 veces lo esperado.

En conjunto, estos resultados indican que la medida de metilación del ADN en las dianas HpaII (de forma individual o agrupada) es un buen indicador del estado de metilación de la isla CpG, especialmente cuando se utiliza más de una diana por isla.

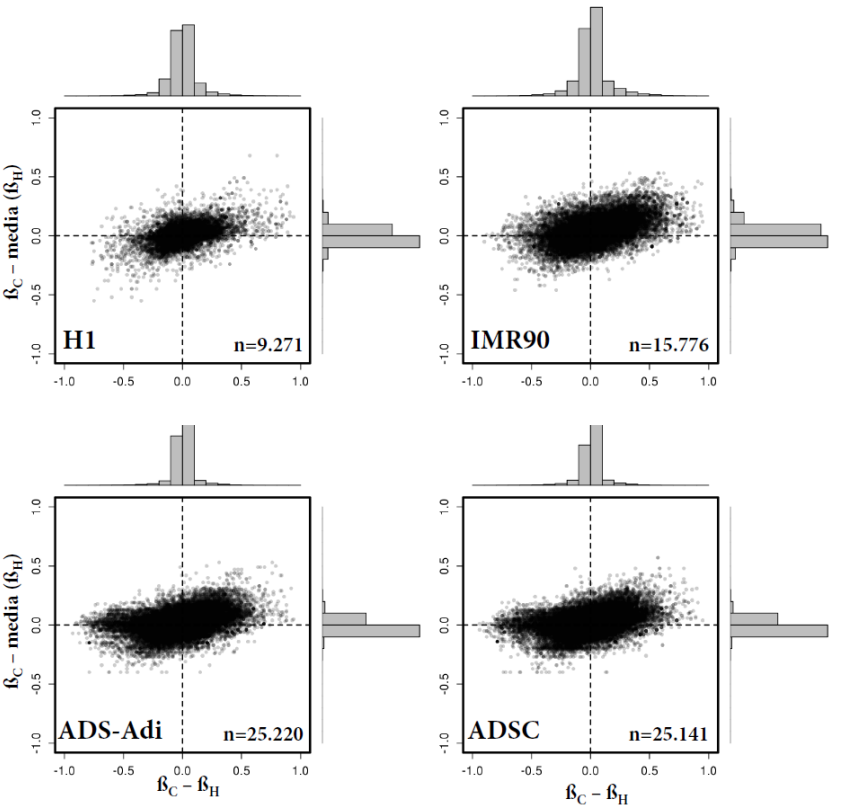


Figura 31: Mejora en el valor predictivo del coeficiente de metilación de HpaII al usar la media de todos los sitios HpaII dentro de una isla

La diferencia de metilación entre la isla CpG y el sitio HpaII (eje X) muestra una distribución más amplia al compararla con la utilización de la media de todas las dianas HpaII (eje Y) contenidas dentro de la isla CpG

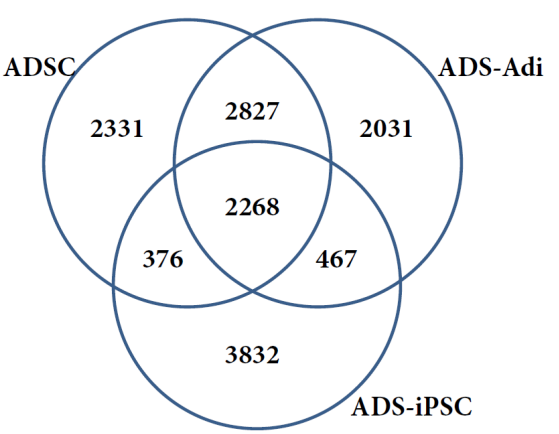


Figura 32: Recurrencia de las dianas HpaII discordantes en tres líneas celulares

Los diagramas de Venn muestran el número de sitios con valores de metilación discordantes (diferencia absoluta >0,25) entre la diana HpaII y el coeficiente de metilación medio de su respectiva isla CpG. El número de sitios HpaII informativos es aproximadamente de 250.000 para todas las muestras.

4.1.5-Características de los sitios discordantes

Para obtener mayor información sobre los determinantes de la metilación atípica en las dianas HpaII discordantes (aquellos puntos con una diferencia absoluta entre $ß_C$ y $ß_H$ >0,25) se exploraron algunas características genómicas: (i) distancia del sitio HpaII al extremo más cercano de la isla CpG; (ii) tamaño de la isla; (iii) número total de dinucleótidos CpG en la isla y (iv) el ratio CG Observado/Esperado de la isla). Se estudiaron de forma separadas los sitios hipometilados de los hipermetilados (en relación a la metilación de la correspondiente isla CpG) (**Figura 33**). En los análisis no se observaron diferencias significativas en la distribución de estas características, con la excepción de un ligero incremento en sitios discordantes (tanto hipo como hipermetilados) en islas de gran tamaño y con mayor número de dinucleótidos CpG (**Figura 33**), especialmente en células H1.

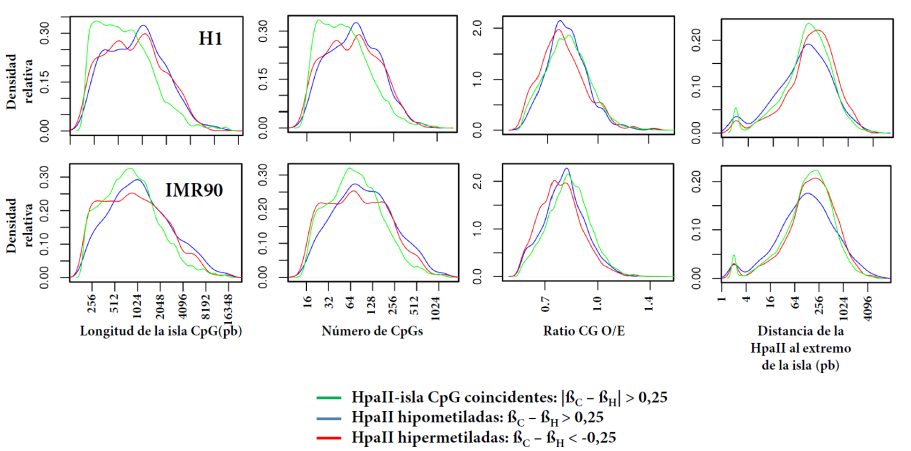


Figura 33: Características genómicas de las islas CpG en función de su concordancia con su sitio HpaII

Evaluación de características estructurales de la isla CpG para las líneas H1 e IMR90 (longitud, número de CpGs, ratio CG observado/esperado y distancia de HpaII a su extremo más cercano).

Finalmente, se analizó la posible aparición de *SNPs* (*single nucleotide polymorphisms*) afectando al dinucleótido de la diana HpaII que pudiera explicar las discrepancias. Los datos sobre *SNPs* fueron descargados de dbSNP versión 135 en ftp.ncbi.nih.gob/snp. Alrededor del 3% de todos los sitios HpaII contienen polimorfismos que afectan a la detección de la metilación. Las dianas HpaII discordantes exhiben un significativo elevado enriquecimiento en *SNPs* (hasta el 9%) comparado con los sitios concordantes (**Tabla 5**). Esto sugiere que una fracción de las discordancias puede ser debida

a variaciones genéticas que pueden ser confundidos con cambios epigenéticos debido a la ambigüedad que genera el método de secuenciación con bisulfito.

Tabla 5: Frecuencia de *SNPs* en sitios HpaII

Línea celular	**H1**	**IMR90**
Concordantes[1] (SNP/total)	1.025/30.484	1.810/70.962
	3,4%	2,5%
Discordantes[2] (SNP/total)	159/1.669	310/6.415
	9,5%	4,8%
Test de Fisher	$p<10^{-15}$	$p<10^{-15}$

Clasificación en función de la diferencia de metilación respeto a su respectiva isla CpG.

[1]Concordante: La diferencia absoluta entre el coeficiente de metilación de la isla CpG y el coeficiente de metilación del sitio HpaII <0,25.

[2]Discordante: La diferencia absoluta entre el coeficiente de metilación de la isla CpG y el coeficiente de metilación del sitio HpaII >0,25. Los datos sobre SNPs fueron obtenidos de dbSNP versión 135 de ftp.ncbi.nih.gov/snp

4.2-Identificación a nivel genómico de secuencias regulatorias mediante detección de estados de metilación anómalos en sitios CpG

4.2.1-Identificación de posiciones anómalas

El estudio anterior ha mostrado la alta homogeneidad existente en la metilación de las islas CpG. En estas regiones genómicas, cada posición CpG individual tiene un coeficiente de metilación similar a sus posiciones CpG adyacentes. Sin embargo, en otros escenarios, no es raro observar diferencias de metilación importantes entre dos CpGs próximas. Un ejemplo claro son las CpGs situadas en las zonas limítrofes de las islas CpGs. En este caso, las diferencias pueden explicarse por las transiciones epigenéticas que se producen entre dos regiones genómicas con distinta funcionalidad. Por otro lado, aunque de una manera mucho menos frecuente, también es posible observar CpGs cuyo estado de metilación difiere de las posiciones adyacentes, pero sin una relación aparente con elementos genómicos. Las posiciones discordantes del estudio anterior que no corresponden a *SNPs* representan un ejemplo.

Dado el interés que presentan estas posiciones discordantes, nuestro siguiente objetivo fue identificar aquellos dinucleótidos CpG con una metilación anómala respecto a su entorno. En primer lugar, se transformaron los genomas a perfiles de metilación mediante el script MethProfile.py, de acuerdo a las lecturas originales (ver Materiales y Métodos). Tal y como se describe también en la sección de Materiales y Métodos, hemos definido, de forma arbitraria, una CpG anómala como aquella posición central dentro de una secuencia de 11 CpGs que presenta un coeficiente de metilación contrario al resto de CpGs (**Figura 34**). Bajo esta definición se seleccionaron las secuencias 11111011111 y 00000100000. Se identificaron 13.765 apariciones de estas secuencias en conjunto para las líneas H1 e IMR90. Estas apariciones se distribuyen en 8.872 y 4.882 casos de la secuencia 11111011111 y 0 y 11, de la secuencia 00000100000, en H1 e IMR90 respectivamente. Cabe destacar que estas diferencias de frecuencias entre H1 e IMR90 para la misma secuencia no se deben a la cantidad de lecturas, dado que éstas presentan valores similares (1,16 y 1,18 mil millones para H1 e IMR90 respectivamente).

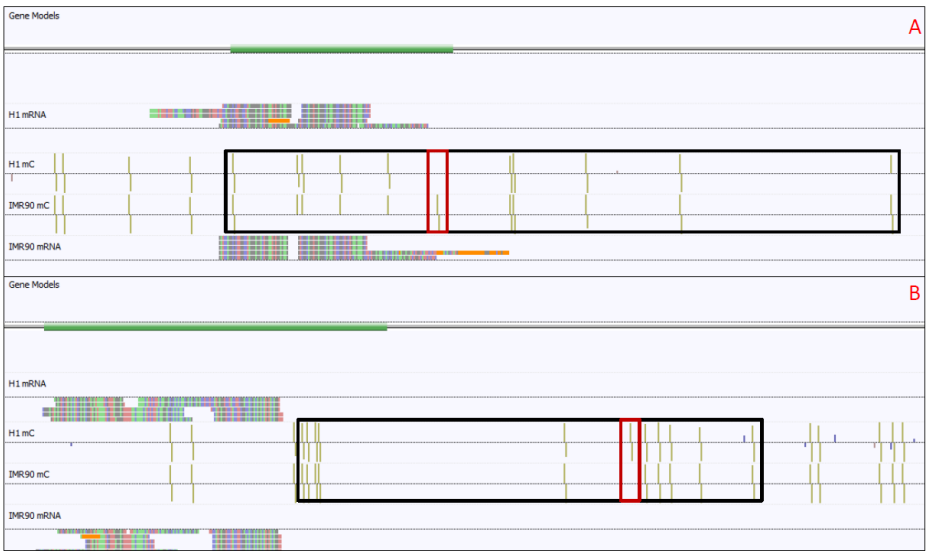


Figura 34: Patrones anómalos de metilación y comparación entre líneas celulares
Imágenes obtenidas del navegador http://neomorph.salk.edu/human_methylome/browser.html (234).
Se muestran dos regiones genómicas (**A** y **B**) comparando las secuencias de metilación en H1 e IMR90. Los rectángulos horizontales verdes representan genes y los multicolor, las lecturas de experimentos de *RNA-Seq*. Las líneas verticales verde claro corresponden a los valores de metilación de citosinas en contexto CpG. Se enmarcan en negro las secuencias de 11 nucleótidos y en rojo, las CpGs centrales. Las secuencias anómalas de metilación son aquellas que presentan estados de metilación divergentes en la CpG de la posición central respecto al resto de dinucleótidos CpG flanqueantes (11111011111 en las líneas celulares H1 e IMR90 en **A**-y **B** respectivamente). La comparación de estas secuencias entre las líneas H1 e IMR90 permite encontrar las posiciones de cambio. En las imágenes se observa la metilación de la posición central de H1 a IMR90 (**A**), situada dentro de un gen, y la desmetilación (**B**), fuera del gen.

Para evaluar si las frecuencias observadas se podrían explicar por una distribución al azar de las posiciones anómalas, se generaron 10 perfiles de metilación virtuales de H1 e IMR90, redistribuyendo aleatoriamente todos los valores de 1s y 0s entre las posiciones originales (ver Materiales y Métodos). Todos los sets de datos virtuales mostraron, de forma consistente, un incremento de la secuencia 11111011111 respecto a los datos obtenidos experimentalmente, tanto en H1 como en IMR90 (**Figura 35**) con unos valores medios de 37.555 y 25.386 apariciones de la secuencia 11111011111 en H1 e IMR90 respectivamente. Este hecho parece indicar una selección negativa de estas posiciones anómalas. Por otro lado, la secuencia 00000100000 no se observó en ninguna de las reconstrucciones virtuales.

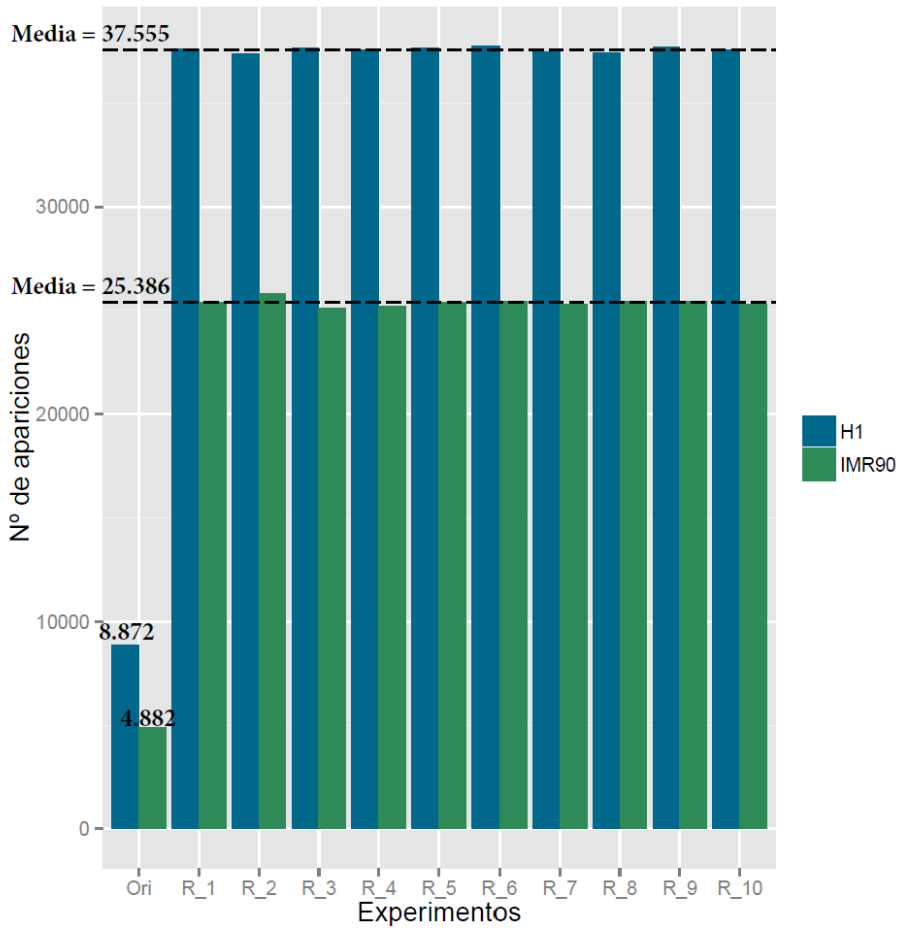


Figura 35: Experimentos de aleatoriedad de apariciones del patrón 11111011111
Se realizaron 10 simulaciones de aleatoriedad de los valores 1 y 0 en las líneas H1 e IMR90. Para ello se reubicaron los valores 1 y 0 de los perfiles de H1 e IMR90 de forma aleatoria y se obtenía el número de veces que se generaba la secuencia 11111011111.

4.2.2-Comparación entre líneas celulares

La determinación de las marcas epigenéticas en una misma región para distintas líneas celulares ha permitido encontrar zonas con metilación diferencial, como las *DMRs*, asociadas a cambios en la regulación génica. Por ese motivo, se compararon las apariciones de la secuencia 11111011111 entre las líneas celulares H1 e IMR90. No se analizó la secuencia 00000100000 dada su frecuencia extremadamente baja. Se observaron un total de 553 casos con cambios en el estado de metilación de la posición central (**Figura 34**) y 495 donde ésta se mostró invariante (**Tabla 6**). De esta forma se identificaron posiciones candidatas que pueden representar zonas putativas de regulación.

Tabla 6: Comparación de la secuencia 11111011111 y cambios de la posición central en H1 e IMR90

H1	↔	IMR90	Nº de casos
11111**0**11111		11111**1**11111	237
11111**0**11111		11111**0**11111	495
11111**1**11111		11111**0**11111	316

Ya se ha mostrado en los resultados anteriores que algunos cambios de metilación detectados pueden ser debidos a alteraciones de la secuencia (*SNPs*). Por este motivo, se estudió la presencia de polimorfismos en las posiciones anómalas. Mediante el script checkTableSNP.py se utilizaron las lecturas originales para identificar posibles *SNPs* en estas posiciones. Una vez analizados, se compararon los resultados con los *SNPs* de la base de datos dbSNP versión 138 (**Tabla 7**). El 100% de las posiciones identificadas como *SNPs* mediante las lecturas se correspondían con *SNPs* reportados en la base de datos y fueron descartadas. Por otro lado, se identificaron 297 posiciones como No *SNPs* que se utilizaron en los análisis descritos a continuación. De éstas, 187 aparecían en la base de datos como posibles *SNPs*. Dado que los polimorfismos no aparecen en toda la población, el uso de las lecturas originales permite un mejor cribado. Por otro lado, la alta prevalencia de polimorfismos que se observa en estas posiciones se debe a la alta tasa de mutación del dinucleótido CpG que lo convierte en un sitio altamente polimórfico.

Tabla 7: Identificación de *SNPs* mediante lecturas

	SNP **H1 > IMR90**			***No SNP*** **H1 > IMR90**		
	0 > 1	**0 > 0**	**1 > 0**	**0 > 1**	**0 > 0**	**1 > 0**
Experimental	204	432	115	33	63	201
dbSNP v.138	204	204	115	32	63	92

Comparación entre la clasificación *SNP/NoSNP* generada mediante las lecturas (Experimental) y los *SNPs* de la base de datos dbSNP versión 138 de ftp.ncbi.nih.gov/snp.

Tal y como se ha indicado, se entiende una posición anómala como aquella que presenta un valor de metilación contrario al de sus posiciones adyacentes. No obstante, debido a la infrarrepresentación del dinucleótido CpG en el genoma humano, la distancia entre éstas podría ser demasiado elevada para considerarlas dentro de una misma región regulatoria. Por este motivo, se obtuvieron, para las 297 posiciones anómalas candidatas, las distancias de todos los dinucleótidos CpG a la posición central (**Figura 36**). Los valores de la mediana para las posiciones más alejadas (-5 y 5 respecto a la posición

central) se encuentran entre 1 Kb y 1.5 Kb. Interesantemente, las posiciones más cercanas (-1 y 1) se sitúan a ~80 pb para los casos 0 > 1 y 0 > 0 y a ~250 pb en 1 > 0 (**Figura 36**). Este resultado destaca la importancia de estas posiciones anómalas como indicadoras de zonas de especial regulación dado que resulta poco probable que los complejos encargados de la metilación no establezcan el mismo patrón en posiciones tan cercanas.

En conjunto, nuestros análisis nos han permitido obtener las coordenadas de 297 posiciones anómalas en los perfiles de metilación de las líneas celulares H1 e IMR90. Hemos podido comprobar, gracias a las lecturas originales, que estas anomalías no son debidas a cambios polimórficos a pesar de que muchas de ellas se encuentran frecuentemente asociadas a *SNPs*.

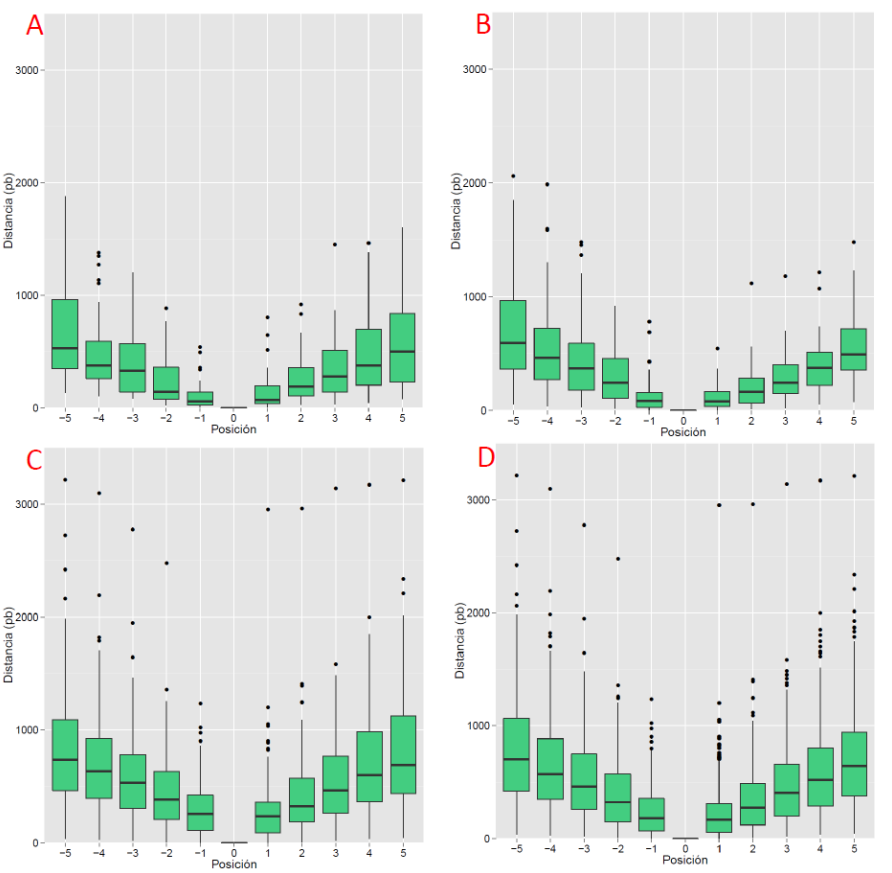


Figura 36: Gráficos Boxplot de las distancias entre dinucleótidos CpG de los patrones comparados

Se muestra la distribución de las distancias de los dinucleótidos CpG que conforman el patrón a la posición central. Se analizan los grupo; **A**: 11111011111 > 11111111111; **B**: 11111011111 > 11111011111; **C**: 11111111111 > 11111011111; **D**: todo.

4.2.3-Caracterización de las posiciones anómalas

Una vez identificadas las posiciones anómalas, el siguiente objetivo fue estudiar si éstas se encontraban en regiones cromatínicas con características diferenciales. En primer lugar, se intersecaron las posiciones anómalas con las coordenadas de islas CpG debido a la importancia de estas regiones. No obstante, únicamente se observó un solapamiento con una isla en el grupo 1 > 0.
Una vez descartado un posible efecto sobre estos elementos regulatorios se decidió analizar la configuración cromatínica que envuelve a las posiciones anómalas. Para ello, se realizaron test de enriquecimiento mediante permutaciones con el paquete *RegioneR* (248) del software estadístico R. En el análisis se comparó la población de posiciones anómalas con una población control formada por todas las posiciones (1.017.790) que cambiaban de estado (de metilado a desmetilado o viceversa) entre las dos líneas celulares. De este cálculo se excluyeron por tanto, las CpGs que en alguna de las dos líneas celulares analizadas resultaban no informativas (menos de 10 lecturas, n=7.879.984) o presentaban niveles de metilación intermedios (entre 0,25 y 0,75, n=8.640.766) o ambas características, una en cada línea (n=978.544). Estos filtros permiten utilizar una población control más parecida a la formada por las posiciones anómalas. Para cada análisis de enriquecimiento se realizaron 5000 permutaciones.

Para estas posiciones se realizaron diversos análisis de enriquecimiento en las siguientes características cromatínicas.

4.2.3.1-Accesibilidad del ADN

Como ya se ha descrito en la introducción, en el genoma humano existen distintas configuraciones epigenéticas que definen regiones con diferentes actividades funcionales. Las zonas de mayor actividad en el genoma se corresponden con una estructura abierta y accesible. Por esta razón, en primer lugar se evaluó la presencia de las posiciones anómalas en zonas accesibles del ADN. Para ello, se testó su asociación a regiones hipersensibles a DNAsa I (**Figura 37**). La DNAsa I es una enzima que corta únicamente el ADN desnudo (no asociado a proteínas como por ejemplo las histonas) indicando zonas de accesibilidad para proteínas efectoras. Los resultados mostraron comportamientos opuestos entre el grupo 0 > 0, donde se observó empobrecimiento (**Figura 37-B,** p-val = 0,015 y z-score=2,00), y el grupo de 1

> 0 donde se encontró un claro enriquecimiento de estas posiciones en regiones hipersensibles a la DNAsa I (**Figura 37-C** p-val = 0,0002 y z-score=11,06). El grupo de 0 > 1 muestra un ligero enriquecimiento aunque no es significativo (**Figura 37-A** p-val = 0,076 y z-score=1,24).

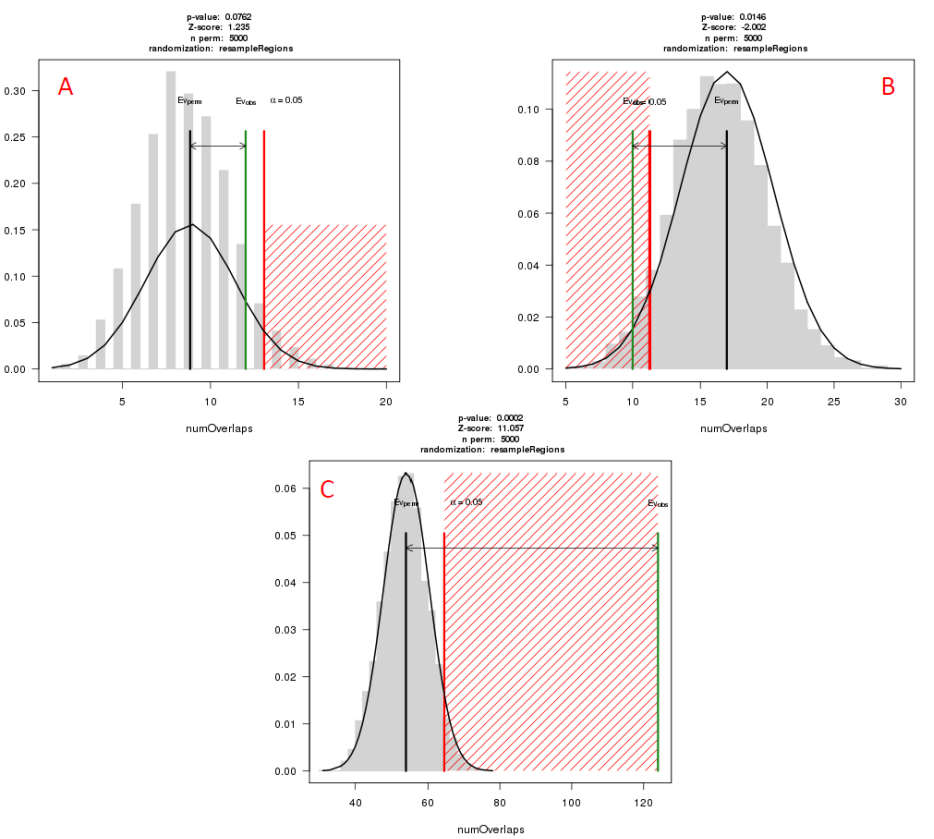


Figura 37: Análisis de enriquecimiento de las posiciones anómalas en regiones hipersensibles a DNAsa

Gráficos del paquete *RegioneR* de enriquecimiento de las posiciones centrales de los patrones anómalos (**A**: 11111011111 > 11111111111; **B**: 11111011111 > 11111011111; **C**: 11111111111 > 11111011111). La distribución se genera mediante 5.000 permutaciones generando grupos del mismo tamaño que la muestra a analizar tomando los elementos de la población control (todas los dinucleótidos cambiantes entre H1 e IMR90). La línea vertical roja marca el punto de la distribución correspondiente al p-valor 0,05 y la línea vertical verde el valor del estadístico observado.

En general, estos resultados indican una tendencia de las posiciones anómalas a estar asociadas a regiones accesibles del ADN, especialmente las posiciones que pertenecen al grupo 1 > 0.

4.2.3.2- Asociación a estados cromatínicos

Por otro lado, el conjunto de marcas de histonas de una región permite identificar secciones funcionales dentro del genoma. Por este motivo, se evaluó la asociación de las posiciones estudiadas con estados de la cromatina. Estos estados provienen de estudios donde se integraron de datos de *ChIP-Seq* mediante modelos de Markov ocultos (247). Se utilizaron los datos de H1 dada la inexistencia de datos para IMR90 en el momento del estudio. De nuevo, mediante *RegioneR*, se obtuvieron los estadísticos de los test de enriquecimiento (**Tabla 8**). De forma interesante, en los tres grupos estudiados se encontró un empobrecimiento en regiones inactivas (**Tabla 8-A**) que, al ser analizadas en detalle, corresponden a heterocromatina (**Tabla 8-B**). Estos datos indican que las posiciones anómalas se sitúan preferentemente en regiones activas del genoma. Es más, los estados asociados a transcripción estaban claramente enriquecidos en los tres grupos (**Tabla 8-A** y **B**). Por otro lado, se observó en el grupo 1 > 0 un empobrecimiento en regiones activadoras, especialmente en zonas activadoras débiles.

Tabla 8: Análisis de enriquecimiento con estados cromatínicos de H1

A	**0 > 1**			**0 > 0**			**1 >0**		
Estado	**pval**	**z_score**	**+/-**	**pval**	**z_score**	**+/-**	**Pval**	**z_score**	**+/-**
Promotor	0,0002	0,603	-	0,0002	0,826	-	0,0002	1,484	-
Enhancer	0,2192	0,415	-	0,0536	1,147	-	0,0010	2,467	-
Insulator	0,0002	0,478	-	0,0736	0,861	+	0,2645	0,274	-
Transcripción	0,0002	9,770	+	0,0002	11,812	+	0,0002	14,530	+
Reprimido	0,0516	1,021	+	0,0002	0,868	-	0,0980	0,859	-
Inactivo	0,0002	7,828	-	0,0002	9,179	-	0,0002	9,869	-
B									
***Enhancer* fuerte**	0,0002	0,370	-	0,0002	0,503	-	0,0002	0,937	-
***Enhancer* débil**	0,2476	0,327	-	0,0676	1,028	-	0,0018	2,267	-
Transcripción	0,0002	11,745	+	0,0002	18,205	+	0,0002	17,225	+
Transcripción débil	0,0002	6,088	+	0,0002	6,128	+	0,0002	9,187	+
Heterocromatina	0,0002	7,735	-	0,0002	8,908	-	0,0002	10,089	-
Repetitivo/CNV	0,0002	0,187	-	0,0002	0,257	-	0,0002	0,491	-

Los valores en +/- indican enriquecimiento (+) o empobrecimiento (-). Se indica el p-valor para evaluar la significancia estadística. La magnitud de z-score indica cuan alejado se encuentra el valor del estadístico del valor medio.

A) Clases amplias de estados cromatínicos.

B) Subgrupos de los estados estadísticamente significativos; (Activador: Activador fuerte y débil; Transcripción: Transcripción y transcripción débil; Inactivo: Heterocromatina y Repetitivo/CNV).

4.2.3.3-Unión de factores de transcripción

Los resultados obtenidos con los análisis anteriores muestran que las posiciones anómalas se sitúan en zonas de la cromatina activa, especialmente asociadas a la transcripción génica. Aparte de las marcas de histonas y la accesibilidad del ADN sabemos que la transcripción se regula, en gran medida, mediante la unión de factores de transcripción. Con el objetivo de comprobar la posible unión de factores de transcripción específicos a las zonas donde se encuentran las posiciones anómalas, se realizaron análisis de enriquecimiento. Los datos provenían de experimentos *ChIP-Seq* de 91 tipos celulares, permitiéndonos explorar una gran cantidad de ambientes regulatorios. Los resultados para los grupos 0 > 1 y 0 > 0 no mostraron enriquecimiento o empobrecimiento en ningún factor de transcripción. En cambio, si aparecieron resultados positivos para el grupo 1 > 0 (**Tabla 9**). Gran parte de los lugares de unión que aparecen enriquecidos pertenecen a factores de transcripción relacionados con la proliferación celular, diferenciación, supervivencia y crecimiento celular.

En el interior de las células se establecen redes de expresión que regulan una gran cantidad de genes de forma coordinada. Los efectores de esa regulación son los factores de transcripción y por tanto, sus lugares de unión pueden utilizarse para estudiar patrones de expresión conjunta de distintos genes. Por ese motivo, se realizó un análisis de clústeres entre las posiciones anómalas y los factores de transcripción que aparecían unidos a estas posiciones. Se utilizó una medida de distancias binarias, donde la presencia de un factor en una posición anómala es un 1 y un 0 en caso contrario, y Ward como método jerárquico para el *clustering*. Interesantemente, al analizar la distribución de los lugares de unión, se observó un clúster en un conjunto de posiciones anómalas (**Figura 38**). Dicho clúster está formado por 17 posiciones anómalas, distribuidas por varios cromosomas, y 10 factores de transcripción. La totalidad de estos factores aparecían enriquecidos en el análisis anterior y corresponden a CEBPB, JUN, FOSL2, JUND, GATA2, POLR2A, FOS, EP300, STAT3 y MYC.

El conjunto de factores FOS, JUN, FOSL2 y JUND se agrupan en homo- y heterodímeros conformando el factor AP-1. Este factor controla una gran cantidad de procesos celulares incluyendo la diferenciación, proliferación, supervivencia, apoptosis, crecimiento, migración celular y transformación (252). Adicionalmente, JUN y FOS, junto a MYC son factores de transcripción

de los que se ha descrito su implicación en la división celular (253). Otros de estos factores, como EP300 y STAT3, están también relacionados con procesos de crecimiento, división celular y apoptosis.

Tabla 9: Análisis de enriquecimiento en lugares de unión de factores de transcripción de las posiciones anómalas del grupo 1 > 0

FT	pval	z-score	+/-
FOS	0,0002	12,110	+
JUN	0,0002	9,618	+
FOSL2	0,0002	8,318	+
JUND	0,0002	5,869	+
FOXA1	0,0002	5,785	+
CEBPB	0,0002	5,711	+
POLR2A	0,0002	5,454	+
EP300	0,0002	5,181	+
NR3C1	0,0006	4,128	+
STAT3	0,0002	4,093	+
GATA3	0,0032	3,097	+
GATA2	0,0050	2,637	+
CHD2	0,0100	2,458	+
RAD21	0,0176	2,065	+
MYC	0,0194	2,054	+
MAFK	0,0294	1,805	+

Los valores en +/- indican enriquecimiento (+) o empobrecimiento (-). Se indica el p-valor para evaluar la significancia estadística. La magnitud de z-score indica cuan alejado se encuentra el valor del estadístico del valor medio

Estos resultados muestran una alta presencia de factores de transcripción en la población de posiciones anómalas relacionados con procesos de expansión celular. El clúster principal está formado por 17 posiciones, un 8,5 % de las posiciones anómalas. No obstante, puede observarse como, para ciertos factores, se genera un clúster secundario de similar tamaño.

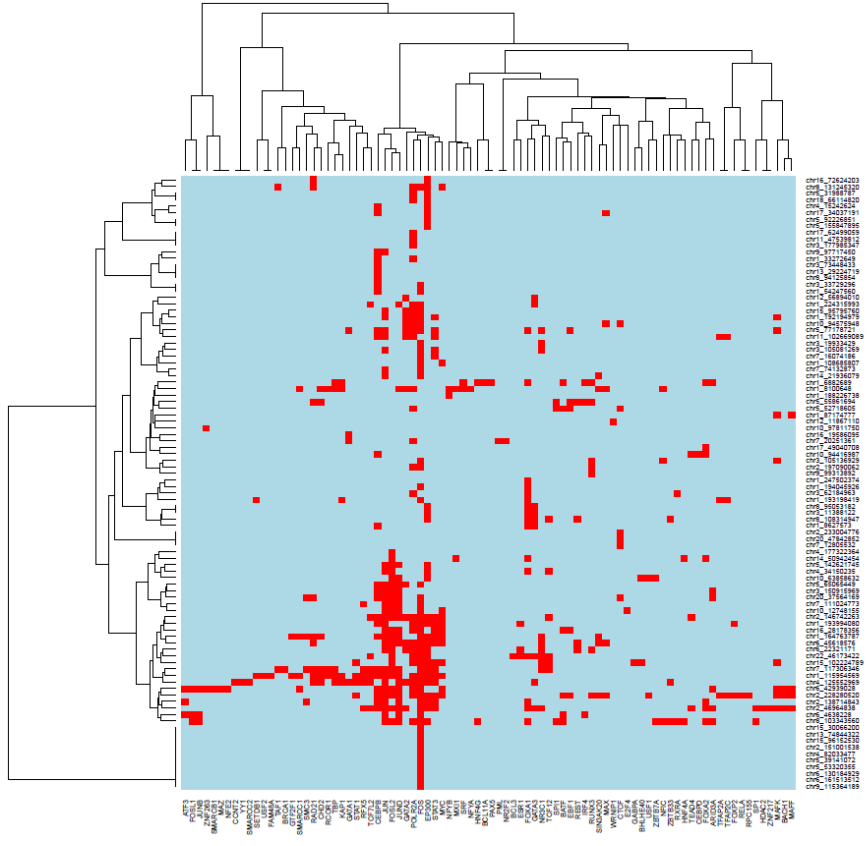


Figura 38: Heatmap del solapamiento entre lugares de unión de factores de transcripción y las posiciones anómalas del grupo 1 > 0

En rojo se muestra la presencia de lugares de unión para factores de transcripción. Las posiciones y lugares de unión de factores de transcripción que no presentaban ningún solapamiento fueron descartadas para el gráfico. La presencia o ausencia de un factor de transcripción en un posición dada puede codificarse como 1 o 0 respectivamente. Por ese motivo se utiliza un cálculo de distancia para datos binarios. El método de *clustering* utilizado es Ward, del tipo jerárquico aglomerativo.

Adicionalmente, se realizó una comparación *in silico* de motivos característicos de lugares de unión de factores de transcripción. Mediante la funcionalidad AME (*Analysis of Motif Enrichment*) del conjunto de paquete de software MEME Suite (249), se evaluó el enriquecimiento en motivos de secuencia asociados a factores de transcripción de la base de datos JASPAR (250). La búsqueda se realizó en las regiones que comprenden desde 100 nucleótidos *upstream* hasta 100 nucleótidos *downstream* de las posiciones anómalas. Como población control se utilizaron todos los dinucleótidos CpG del genoma que cambiaban su estado de metilación (de 1 a 0 o viceversa).

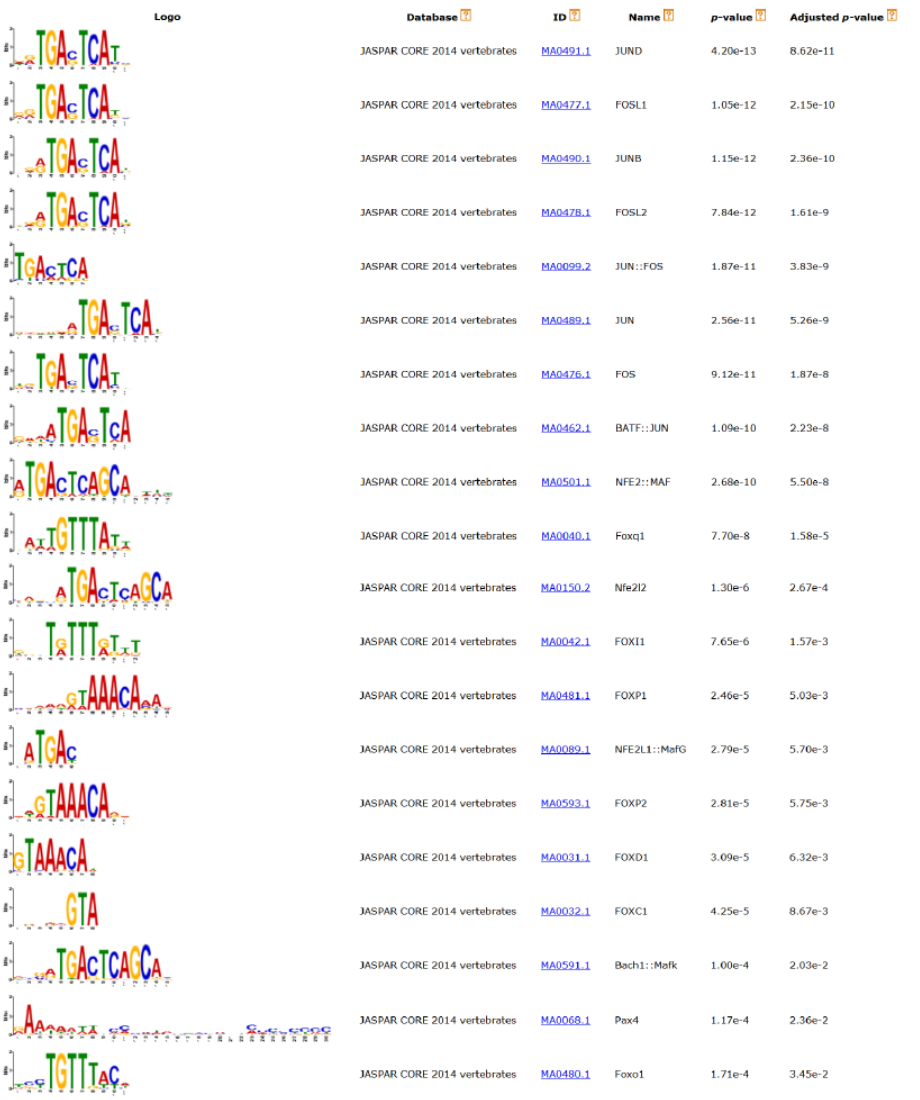
En concordancia con el análisis anterior, en los grupos 0 > 1 y 0 >0 no se observó ningún tipo de enriquecimiento. De manera similar, el grupo 1 > 0 mostró enriquecimiento en motivos de secuencia de factores de transcripción de la familias JUN y FOS (**Figura 39**). Interesantemente, también aparecen lugares de unión para proteínas de la familia FOX. De forma similar a JUN y FOS, esta familia de factores de transcripción está involucrada en crecimiento celular, proliferación y diferenciación. De esta forma se valida tanto *in silico* como experimentalmente, el enriquecimiento en lugares de unión de factores de transcripción críticos en las regiones que rodean a las posiciones anómalas.

Posteriormente, se analizó la existencia de motivos de unión no descritos en las bases de datos. Para ello, se utilizó la funcionalidad MEME (*Multiple Em for Motif Elicitation*) (249). En primer lugar se estudió la existencia de un motivo común en todas las secuencias. Los resultados obtenidos mostraron un único motivo significativo (por debajo del umbral de 0,05 de e-valor) en el grupo de secuencias de 1 > 0 (**Figura 40-A**). Este motivo se analizó mediante la funcionalidad Tomtom (249) que permite la comparación con una base de datos de motivos de ADN en humanos. No obstante, no se obtuvo ningún resultado significativo.

A continuación, se realizó el mismo análisis eliminando el requisito de aparición del motivo en todas las secuencias. Aparecieron entonces diversos motivos significativos en distintos subgrupos de secuencias del grupo 1 > 0 (**Figura 40-B**). Cuando se analizaron los distintos motivos mediante Tomtom, no aparecieron resultados significativos excepto para un motivo común a 12 secuencias. Dicho motivo presenta en su zona central una región de muy alta similitud a la secuencia de unión de PAX5, perteneciente a la familia PAX (**Figura 40-C**). Esta familia de factores de transcripción son importantes

reguladores del desarrollo temprano y contribuyen a la transformación tumoral en caso de alteraciones.

En conjunto, estos resultados muestran que las regiones en las que se enmarcan una parte de las posiciones anómalas son lugares con una importante regulación transcripcional.



Logo	Database	ID	Name	*p*-value	Adjusted *p*-value
	JASPAR CORE 2014 vertebrates	MA0491.1	JUND	4.20e-13	8.62e-11
	JASPAR CORE 2014 vertebrates	MA0477.1	FOSL1	1.05e-12	2.15e-10
	JASPAR CORE 2014 vertebrates	MA0490.1	JUNB	1.15e-12	2.36e-10
	JASPAR CORE 2014 vertebrates	MA0478.1	FOSL2	7.84e-12	1.61e-9
	JASPAR CORE 2014 vertebrates	MA0099.2	JUN::FOS	1.87e-11	3.83e-9
	JASPAR CORE 2014 vertebrates	MA0489.1	JUN	2.56e-11	5.26e-9
	JASPAR CORE 2014 vertebrates	MA0476.1	FOS	9.12e-11	1.87e-8
	JASPAR CORE 2014 vertebrates	MA0462.1	BATF::JUN	1.09e-10	2.23e-8
	JASPAR CORE 2014 vertebrates	MA0501.1	NFE2::MAF	2.68e-10	5.50e-8
	JASPAR CORE 2014 vertebrates	MA0040.1	Foxq1	7.70e-8	1.58e-5
	JASPAR CORE 2014 vertebrates	MA0150.2	Nfe2l2	1.30e-6	2.67e-4
	JASPAR CORE 2014 vertebrates	MA0042.1	FOXI1	7.65e-6	1.57e-3
	JASPAR CORE 2014 vertebrates	MA0481.1	FOXP1	2.46e-5	5.03e-3
	JASPAR CORE 2014 vertebrates	MA0089.1	NFE2L1::MafG	2.79e-5	5.70e-3
	JASPAR CORE 2014 vertebrates	MA0593.1	FOXP2	2.81e-5	5.75e-3
	JASPAR CORE 2014 vertebrates	MA0031.1	FOXD1	3.09e-5	6.32e-3
	JASPAR CORE 2014 vertebrates	MA0032.1	FOXC1	4.25e-5	8.67e-3
	JASPAR CORE 2014 vertebrates	MA0591.1	Bach1::Mafk	1.00e-4	2.03e-2
	JASPAR CORE 2014 vertebrates	MA0068.1	Pax4	1.17e-4	2.36e-2
	JASPAR CORE 2014 vertebrates	MA0480.1	Foxo1	1.71e-4	3.45e-2

Figura 39: Análisis de enriquecimiento *in silico* en motivos de secuencia para lugares de unión de factores de transcripción para el grupo 11111111111 > 11111011111

El grupo de control es el conjunto de todos los dinucleótidos CpG del genoma que cambiaban su estado de metilación (de 1 a 0 o viceversa). Se utilizó la última versión de JASPAR-vertebrados para la búsqueda de motivos. Se muestra el logo, el identificador dentro de la base de datos, el nombre del factor asociado y los p-valores del test (incluyendo el ajustado por Bonferroni).

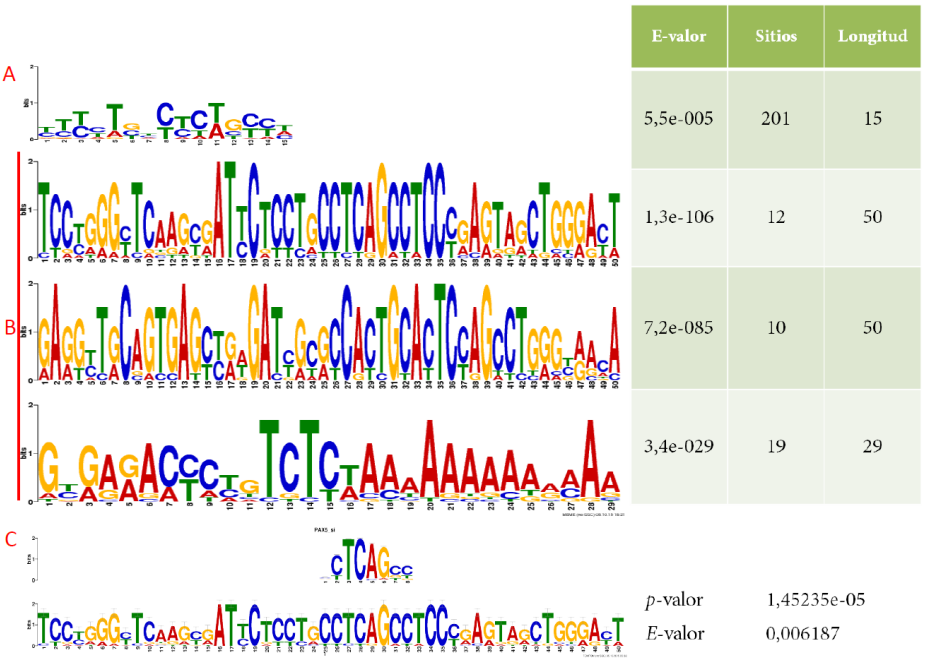


Figura 40: Análisis de motivos mediante MEME y Tomtom para el grupo 11111111111 > 11111011111

La búsqueda de motivos comunes a todas las secuencias indicó un único motivo significativo (A). Este motivo no tiene parecido a otros conocidos en humanos [base de datos HOCOMOCO v9 (251)]. Al permitir la búsqueda en subgrupos de secuencias aparecen tres motivos significativos (B). Uno de ellos presenta un alto parecido al motivo de unión de PAX5 (C) al utilizar la utilidad Tomtom.

4.2.3.4-Expresión de genes cercanos

Estudios anteriores han descrito regiones diferencialmente metiladas (*DMRs*) entre células humanas embrionarias y fibroblastos (254). En estos análisis se observó que el estado de metilación de estas regiones correlacionaba con cambios en la expresión génica. Por este motivo, se estudió si los cambios de metilación detectados en las posiciones anómalas, entre las líneas celulares H1 e IMR90, mostraban alguna correlación con la expresión de genes cercanos. Para ello se identificaron los *TSSs* de genes *refseq* más cercanos a cada posición anómala y se asignó a cada gen los valores de expresión, en H1 e IMR90, del estudio original (234). Posteriormente, se analizaron aquellos cambios significativos a una distancia inferior a 5.000 pb (**Tabla 10**).

En la tabla pueden observarse cambios importantes en genes característicos de fibroblastos, como MMP1, una colagenasa, y ALCAM, implicado en la adhesión celular. En estos genes podemos observar un aumento en la expresión así como una desmetilación de la posición. Esto es especialmente

interesante en el caso de MMP1, dada la cercanía de la posición anómala a su *TSS*.

Tabla 10: Cambios de expresión, entre H1 e IMR90, en genes cercanos a posiciones anómalas del grupo 1 > 0

		Expresión (RPKM)	
Gen	**Distancia al *TSS*(pb)**	**H1**	**IMR90**
MITF	-3.799*	0,67	9,13
ALCAM	4.982	25,11	98,99
MMP1	-266	2,00	40,26
PSMC6	-4.588*	49,12	29,39
SNRPG	-3.794	558,07	223,85
NDUFB1	-1.605	68,36	43,88
SH3GLB1	-4.195*	8,20	19,17
NCLN	-2.462*	21,12	17,27

RPKM: *Reads per Kilobase per Million* = [Nº de lecturas mapeadas]/ ([longitud del tránscrito]/1000)/ ([Nº total de lecturas]/10^6). Datos proporcionados en el estudio original (234).
Distancia al *TSS* (*Transcription Start Site*). Las distancias negativas indican que el TSS se encuentra *upstream* de la posición anómala y las positivas, *downstream*.
* La posición anómala se encuentra dentro del gen.

Es importante mencionar el caso del gen NDUFB1. Éste codifica una subunidad del complejo I de la cadena respiratoria. No obstante, a pesar de que la posición anómala se encuentra más cerca del *TSS* de este gen, ésta se encuentra dentro del cuerpo de otro, CPSF2, de función desconocida pero relacionado con carcinoma de tiroides papilar (255, 256). La expresión de este gen varía de forma similar a NDUFB1, reduciéndose de 32.65 RPKM a 25.45 en el paso de H1 a IMR90. Podemos observar que otros genes de la lista están relacionados a diversas enfermedades, MITF y ALCAM asociados a melanomas (257–259), SNRPG asociada a lupus eritematoso (260) y SH3GLB1 a Alzheimer (261).

5-Discusión

HpaII como posición indicadora de la metilación

En la actualidad, se considera la secuenciación por bisulfito como la técnica *gold standard* en los análisis de metilación del ADN. No obstante, el procesado y análisis de las muestras son procesos engorrosos. Además, la lectura de electroferogramas no está exenta de interpretaciones subjetivas y sesgos técnicos (262, 263). Por este motivo, se han desarrollado técnicas alternativas basadas en la determinación de uno o unos pocos sitios CpG. Un gran número de estudios que han mostrado metilación diferencial en islas CpG han utilizado estas aproximaciones. Por ejemplo, para análisis de regiones concretas, la técnica *MSP* (*Methylation specific PCR*) (223) es probablemente el método alternativo más utilizado. En cambio, en estudios a escala genómica, las metodologías que utilizan enzimas de restricción sensibles a la metilación y/o sondas específicas son frecuentemente utilizadas (264, 265). Cada método presenta sus ventajas y desventajas específicas y puede ser más o menos adecuado según el estudio a realizar. Existen diversos trabajos que han realizado un análisis sobre las características de los diferentes métodos y su rendimiento en función de distintos parámetros (p.ej. cantidad de material necesario, resolución, cuantitividad, cobertura genómica, coste computacional, etc.) (127, 244, 264–269). Se recomienda la lectura de estos trabajos en caso de querer profundizar en las comparativas.

Por otro lado, en los últimos años ha aumentado el número de estudios que han analizado muestras humanas mediante *WGBS* (*Whole-Genome Bisulfite Sequencing*) (19, 234, 237, 270–274). Esta situación proporciona un entorno excelente para valorar, tanto desde un punto de vista teórico (244) como práctico (269, 275), las técnicas que reducen la complejidad del estudio mediante una disminución del número de CpGs analizadas. En el presente trabajo se ha determinado la concordancia entre la metilación de las dianas HpaII y las islas CpG que las contienen. Para ello, se han utilizado datos públicos de dos estudios *WGBS* de muestras humanas (234, 237). Por otra parte, se seleccionó la diana HpaII dada su presencia en el 94% de las islas CpG del genoma. No en vano, la diana de restricción de HpaII (CCGG), conjuntamente con la de SmaI (CCCGGG), que incluye la diana HpaII, es la más utilizada en estudios de metilación del ADN (265, 275, 276). Además, no podemos eludir las razones históricas debido a su papel fundamental en el descubrimiento y caracterización inicial de las islas CpG, originalmente conocidas como *HpaII Tiny Fragment* (*HTF*) *islands*.

Nuestros análisis de los datos de Lister *et al.* (234, 237) muestran que la metilación de aquellas dianas HpaII que están dentro de islas CpG, resultan excelentes indicadores de la metilación global de las islas. Estos resultados están en concordancia con otros estudios que utilizan estrategias diferentes (275). Es importante destacar que más allá de la correlación cuantitativa, las dianas HpaII mostraron un alto valor predictivo a nivel cualitativo (la forma más habitual de indicar el estado de metilación del ADN). Por tanto, el valor de metilación de la diana HpaII permite asignar correctamente a la isla CpG, como metilada o desmetilada, con un alto porcentaje de acierto. Además, el uso conjunto de varias dianas HpaII, en aquellas islas CpG que contenían dos o más dianas, proporcionaba predicciones extremadamente exactas del valor de metilación de la isla.

Adicionalmente, nuestros resultados indican que esta alta correlación parece ser independiente de los niveles globales de metilación, siendo éstos bastante diferentes entre las muestras y tipo de tejido. Asimismo, es importante observar que esta correlación se mantiene incluso cuando se utilizan islas CpG definidas mediante otros criterios (86, 93), incluyendo aquellas no presentes en la lista del *UCSC Genome Browser* (**Figura 29**). También cabe destacar que el análisis aquí presentado muestra una proporción mayor en islas metiladas en el conjunto de islas CpG definidas mediante criterios alternativos, que en el grupo de islas del *UCSC Genome Browser* (**Figura 29, histogramas verticales**). Estos resultados son consistentes con el enriquecimiento en *DMRs* (*Differentially Methylated Regions*) observado en las islas CpG no definidas mediante criterios tradicionales (86, 254). Es más, la mayoría de *DMRs* de células de cáncer de colon (277) solapan con las islas CpG definidas mediante modelos de Markov ocultos (86). Esta asociación permite expandir la aplicación de esta aproximación a por ejemplo, el estudio de *DMRs*.

Limitaciones en la aplicación de esta metodología

La metodología aplicada en este trabajo presenta, no obstante, ciertas limitaciones. Su principal desventaja, como toda técnica basada en una reducción de la complejidad, es la incompleta representatividad de todos los elementos a analizar. En nuestro estudio estos elementos son las islas CpG. En el genoma humano, 1.718 islas CpG de 28.226 (6,1%) no contienen ninguna diana HpaII. Sin embargo, aspirar al nivel de cobertura obtenido por Lister *et al.* (234, 237) requiere el uso de aproximaciones extremadamente masivas, tales como *WGBS*, cosa que representa un reto excepcional que limita su

aplicabilidad a tan solo un número reducido de muestras. En este escenario, una aproximación ligeramente incompleta representa una opción más eficiente. Por otro lado, el uso conjunto de varias técnicas basadas en reducción de la complejidad, como sería el uso de varias dianas de restricción, puede permitir una captura mucho mayor de elementos. La utilización de herramientas bioinformáticas, previamente al estudio experimental, permitiría encontrar los conjuntos de enzimas de restricción adecuados.

No obstante, es importante considerar la posible aparición de sesgos en la composición de bases y en el tamaño de los fragmentos, en la aplicación masiva de aproximaciones basadas en HpaII. Sin embargo, otros estudios ya han analizado estos problemas demostrando que tienen un impacto mínimo en los resultados (275).

Por otro lado, cuando se analizaron las distintas características de las islas CpG no se observó ningún sesgo significativo que pudiera afectar a la representatividad de los sitios HpaII. Únicamente se encontró, en las células H1, una mayor frecuencia de sitios discordantes en aquellas islas CpG de mayor tamaño y con un número más alto de dinucleótidos CpG (**Figura 33**). Sin embargo, este sesgo no se asoció con la posición de la diana HpaII dentro de la isla CpG (cerca o lejos de los límites). Se puede especular que gran parte de las discrepancias observadas en células H1, sean debidas a la abundancia relativa de 5-hidroximetilcitosina (5hmC) en las regiones promotoras de genes regulados durante el desarrollo en células madre embrionarias (278). La 5hmC, que no puede ser distinguida de una citosina metilada mediante secuenciación por bisulfito, es considerada un estado transitorio en la desmetilación activa (279) y por tanto, es más probable que confiera un perfil de metilación más heterogéneo a las islas CpG.

Finalmente, cabe destacar que el nivel de metilación de la diana HpaII parece exagerar el estado de metilación de la isla CpG en islas metiladas (**Figura 25**), mientras que en islas CpG desmetiladas no se observó ningún sesgo. Esta sutil sobreestimación de la metilación deja de observarse cuando se utiliza la media de todas las dianas HpaII.

Posibles futuras aplicaciones

Como ya se ha mencionado, la metodología aquí presentada puede ampliarse a diseños que se basen en el uso de varias enzimas de restricción. Sin embargo,

no está limitada únicamente a estrategias basadas en endonucleasas sino que también puede ser aplicada al diseño de otras estrategias de reducción de la complejidad. Por ejemplo, la información de la representatividad del sitio CpG, en relación a su respectivo elemento genómico, puede ayudar en el diseño y selección de sondas específicas para el análisis de la metilación, mediante hibridación con microarrays de ADN transformado con bisulfito (p. ej. la plataforma Infinium de Illumina).

Por otro lado, y aunque en este trabajo se ha limitado al análisis de islas CpG, dado su importante papel funcional en la regulación génica, las dianas HpaII son también frecuentes fuera de las islas CpG. Esto permite que otros elementos genómicos puedan ser analizados mediante estrategias similares, siempre que muestren perfiles de metilación homogéneos. Un ejemplo podrían ser las *CpG island shores*. Actualmente, éstas se definen como regiones de unas 2000 pb alrededor de las islas (277). Se ha observado que en los procesos de diferenciación celular y cáncer, los perfiles de metilación de las *CpG island shores* parecen ser más plásticos que los de las islas CpG (254, 277), convirtiéndose en el objetivo preferido en estudios genómicos.

Adicionalmente, a medida que aparecen nuevos metilomas para diferentes tipos celulares, se puede extender esta estrategia a la evaluación de nuevos marcadores, representantes de otros elementos genómicos más allá de las islas CpG. A su vez, la homogeneidad de la metilación del ADN a lo largo de regiones genómicas puede contribuir a la definición de dominios epigenéticos desconocidos en base a perfiles de metilación característicos. Estos nuevos dominios podrían actuar como elementos funcionales putativos. Las células tumorales representan un objetivo preferente para este tipo de estudios que buscan zonas de regulación alternativas. No obstante, la heterogeneidad celular intrínseca a la mayoría de tumores añade un nivel adicional de dificultad en el análisis e interpretación de la metilación parcial. En tumores, puede coexistir la concordancia en la metilación de los sitios CpG pertenecientes a una misma lectura con las diferencias en la de metilación entre diferentes lecturas, cosa que probablemente se deba a la heterogeneidad celular de la muestra.

En resumen, nuestros resultados proporcionan una validación global de las estrategias basadas en el uso de la enzima de restricción sensible a metilación HpaII. Esta validación puede extenderse a otras aproximaciones de reducción de la complejidad similares. A parte de la alta informatividad y cobertura

ofrecida por estas aproximaciones alternativas, su principal ventaja radica en la drástica reducción de costes, no únicamente en gastos asociados a la generación de datos (trabajo experimental) sino también en el análisis computacional (265, 280–282). La aplicación sistemática de métodos de reducción de la complejidad en combinación con microarrays o secuenciación masiva de nueva generación, puede aportar un fuerte apoyo a la generación de mapas epigenómicos con unos ratios beneficio-coste excelentes.

Interés en las posiciones anómalas

Los resultados obtenidos indican una homogeneidad general en la metilación del genoma humano. No obstante, a pesar de encontrarse globalmente metilado, también muestra patrones de metilación diferenciales entre distintos tipos de células. La disponibilidad de un creciente número de metilomas ha permitido, en los últimos años, realizar estudios comparativos entre distintas líneas celulares y tejidos (277). De esta forma, se han podido identificar un conjunto de zonas con patrones de metilación diferenciales que incluyen regiones específicas de tejido (*T-DMRs*), asociadas a enfermedades como el cáncer (*C-DMRs*) e incluso únicas de líneas celulares reprogramadas (*R-DMRs*) (254). Estas regiones se han encontrado asociadas a cambios observados en la expresión génica entre estos tipos celulares. Cabe destacar que la comparación de la metilación entre muestras a nivel de base nucleotídica resulta necesaria para observar diferencias que pueden quedar enmascaradas por el uso de perfiles medios de metilación (234).

En el estudio presentado se ha pretendido profundizar en el análisis de las diferencias locales en la metilación. Para ello se han analizado posiciones individuales con una metilación discordante respecto a su entorno. En concreto, se han evaluado aquellas que se mantienen desmetiladas cuando su entorno está totalmente metilados. Dada la metilación general del genoma, estos patrones representan anomalías que difícilmente puedan ser azarosas. El número de dinucleótidos CpGs que conforman las secuencias se ha seleccionado, de forma arbitraria, como 11. El tamaño medio de las secuencias obtenidas es de 1 kb. No obstante, la distancia entre las posiciones adyacentes a la discordante es similar a la observada en *T-DMRs*, inferior a 300 pb (277). Dada esta similitud en el tamaño puede entenderse que algunas regiones reportadas como *DMRs* estén conformadas por una única posición CpG discordante.

Al realizar la búsqueda de posiciones anómalas en los perfiles de metilación de H1 e IMR90 se obtuvieron un total de 8.872 y 4.882 secuencias respectivamente. Este valor es un número considerablemente menor al de los resultados obtenidos mediante el experimento de aleatoriedad (aproximadamente 37.000 en H1 y 25.000 en IMR90, **Figura 35**) y extremadamente bajo comparado con las 250.000 dianas HpaII informativas del estudio de correlación. Por otro lado, tan solo una de las posiciones finalmente estudiadas, tras la comparación entre líneas y el descarte mediante el análisis de *SNPs*, intersecaba con una isla CpG. Por tanto, el bajo número de posiciones anómalas y su baja presencia en islas, extiende la utilidad de los estudios basados en técnicas de reducción de la complejidad en estos elementos genómicos.

En estudios comparativos de *DMRs* se ha observado una mayor correlación entre el estado de metilación y la expresión génica, siempre que uno de los puntos a comparar tenía un valor de metilación cercano a 0 (277). Por ese motivo, en nuestra búsqueda de posiciones anómalas diferencialmente metiladas entre las líneas H1 e IMR90, se han analizado únicamente los cambios de 0 a 1 o viceversa. De esta forma, se buscan posiciones que puedan representar un evento cromatínico claro. Por otro lado, el análisis posterior mediante lecturas reales permite obtener un listado certero de posiciones cuyo cambio no es debido a polimorfismos. Durante el análisis se obtuvo un 100% de concordancia en la asignación de *SNPs* con las bases de datos públicas. No obstante, un 63% de posiciones evaluadas como no *SNPs* aparecían también en la base de datos y habrían sido descartadas si únicamente se hubieran utilizado estas fuentes. Este hecho demuestra, en primer lugar, una alta sensibilidad de la aproximación, y en segundo, la necesidad del uso de los datos crudos en los estudios de posiciones individuales anómalas.

Entornos epigenéticos característicos de las posiciones anómalas

Las posiciones seleccionadas, especialmente aquellas que en H1 aparecen metiladas y en IMR90 desmetiladas, muestran claros perfiles de elementos regulatorios observados en anteriores estudios asociados a *DMRs* (4, 283). El enriquecimiento en zonas de hipersensibilidad a DNAsa I, así como los estados cromatínicos asociados, indica que se trata de regiones de cromatina accesibles a la maquinaria transcripcional. Los estados cromatínicos considerados integran un conjunto de marcas epigenéticas para establecer la posible funcionalidad de la región (136, 247). En nuestros análisis de las

posiciones anómalas, los estados que observamos muestran un empobrecimiento en marcas de inactividad, especialmente aquellas que definen la heterocromatina, y un enriquecimiento en estados relacionados con la transcripción. Estos resultados indican que a pesar de encontrarse el resto de la secuencia metilada, las posiciones anómalas se encuentran en regiones de alta actividad transcripcional. Es posible hipotetizar entonces, que la posición anómala desmetilada sea resultado de una actividad regulatoria diferencial en la región.

Por otro lado, la unión de factores de transcripción es una de las características de zonas implicadas en la regulación genética. Los resultados conjuntos del estudio *in silico* de motivos de secuencia y de los datos ENCODE de *ChIP-Seq* muestran una alta prevalencia en factores de transcripción implicados en diferenciación celular, como FOX, GATA, JUN y FOS. Recientemente se ha reportado la implicación de estos factores regulando zonas de hipometilación (283). Es posible que las posiciones anómalas del estudio presentado puedan estar siguiendo una regulación similar. No obstante, cabe mencionar que algunos factores como JUN y FOS se asocian frecuentemente de forma conjunta por lo que es esperable que ambos aparezcan enriquecidos. Adicionalmente, el estudio directo de las secuencias ha mostrado la presencia de lugares de unión para PAX5, un importante factor de transcripción regulador del desarrollo. Además, se observa un enriquecimiento en EP300, un coactivador de la transcripción, que junto a la hipersensibilidad a DNAsa I y la presencia de RNA Pol II son marcas de regiones asociadas a *enhancers* (284).

Este conjunto de características estudiadas: modificaciones de histonas (integradas en estados cromatínicos), hipersensibilidad a DNAsa I, metilación del ADN y ocupación por factores de transcripción definen estructuras cromatínicas asociadas a elementos funcionales no codificantes (143, 144, 285–294). Dada la asociación observada, las posiciones anómalas podrían ser indicadoras de este tipo de regiones regulatorias.

Efecto sobre la expresión génica

El análisis de expresión génica realizado muestra interesantes resultados. Por un lado, los genes MMP1 (*Matrix metalloproteinase-1*) y ALCAM (*Activated leukocyte adhesion molecule*) se han descrito activos en fibroblastos. Especialmente MMP1, también denominada colagenasa de fibroblastos, cuya

función es la degradación del colágeno de la matriz extracelular. En concordancia a esta información, podemos observar en la **Tabla 10** un importante aumento de la expresión entre H1 e IMR90. Interesantemente, este aumento de la expresión es acorde a la desmetilación de la posición anómala. Esto es especialmente relevante en el caso de MMP1, donde la distancia entre la posición anómala y el TSS es de tan solo 266 pb.

Por otro lado, aparecen en la lista genes que se han asociado a diferentes tipos de cáncer. Tanto MITF como ALCAM se han relacionado con melanomas (257–259). MITF tiene efectos protumorigénicos a través de los genes que regula y en el caso de los melanomas juega un papel crucial en su progresión por tratarse de genes relacionados con invasión, migración y metástasis. Por su lado, ALCAM se relaciona con otros numerosos tipos de cáncer debido a su implicación en la adherencia celular (295) e incluso se ha propuesto su uso como marcador en células madres tumorales (296, 297).

Cabe destacar que el cambio más pronunciado se produce en el gen SNRPG, una proteína que forma parte del núcleo estructural de las *snRNPs* (*small nuclear ribonucleoproteins*), un conjunto de proteínas que, junto a otras, forma el espliceosoma, un complejo encargado de la maduración del ARN, una función primordial en la regulación transcripcional.

Finalmente, una de las posiciones anómalas se encuentra en el cuerpo génico de CPSF2 (*cleavage and polyadenilation specificity factor subunit 2*) aunque el *TSS* más cercano pertenece a NDUFB1. A pesar de que no se conoce el papel que juega en el cáncer, CPSF2 se ha descrito recientemente como un indicador de pronóstico en pacientes con carcinoma de tiroides papilar (255, 256). Observamos por tanto como algunas de las posiciones anómalas descritas se sitúan cerca o incluso dentro de genes con una alta importancia clínica.

En resumen, se han encontrado secuencias anómalas de metilación que presentan comportamientos similares a los observados anteriormente en *DMRs*. Éstas fueron identificados basándose únicamente en el perfil de metilación, destacando posiciones con valores discordantes respecto las posiciones de su entorno inmediato. Estas posiciones presentan características asociadas a zonas de alta actividad transcripcional y pueden representar puntos de regulación génica. Además, algunas de ellas se sitúan cerca de genes importantes en la regulación de enfermedades, como el cáncer. Aunque aún son necesarias validaciones experimentales, aproximaciones *in silico* como la

aquí presentada permiten encontrar posiciones de discordancia epigenética que pueden representar dominios funcionales putativos.

6-Conclusiones

- Hemos desarrollado una metodología que permite evaluar la exactitud de las estrategias de análisis de metilación, basadas en el uso de la enzima de restricción sensible a metilación HpaII. Esta metodología puede extenderse a otras aproximaciones de reducción de la complejidad.

- El 97% de las islas CpG analizadas mostraban una excelente correlación de su coeficiente de metilación con sus dianas HpaII.

- La comparación de perfiles de metilación de dos líneas celulares ha permitido identificar 297 posiciones con un estado de metilación divergente respecto al de los dinucleótidos CpG adyacentes.

- Las CpGs con metilación anómala respecto a su entorno se localizan en cromatina accesible y presentan enriquecimiento en marcas de cromatina activa y lugares de unión de factores de transcripción, lo que sugiere una funcionalidad intrínseca.

Bibliografía

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., *et al.* (2001) The Sequence of the Human Genome. *Science*, **291**, 1304–1351.

3. Lander,E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–197.

4. Rivera,C.M. and Ren,B. (2013) Mapping Human Epigenomes. *Cell*, **155**, 39–55.

5. Bernstein,B.E., Meissner,A. and Lander,E.S. (2007) The Mammalian Epigenome. *Cell*, **128**, 669–681.

6. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*, **16**, 6–21.

7. Egger,G., Liang,G., Aparicio,A. and Jones,P.A. (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, **429**, 457–463.

8. Goll,M.G. and Bestor,T.H. (2005) Eukaryotic cytosine methyltransferases. *Annu Rev Biochem*, **74**, 481–514.

9. Margueron,R., Trojer,P. and Reinberg,D. (2005) The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.*, **15**, 163–176.

10. Ptashne,M. (2007) On the use of the word 'epigenetic'. *Curr. Biol.*, **17**, R233–R236.

11. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.

12. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.

13. Zhou,V.W., Goren,A. and Bernstein,B.E. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, **12**, 7–18.

14. Consortium,T.E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

15. Garraway,L.A. and Lander,E.S. (2013) Lessons from the Cancer Genome. *Cell*, **153**, 17–37.

16. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R., *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotech*, **28**, 1045–1048.

17. (Chairperson),T.J.H., Anderson,W., Aretz,A., Barker,A.D., Bell,C., Bernabé,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S., *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

18. Jones,P.A., Archer,T.K., Baylin,S.B., Beck,S., Berger,S., Bernstein,B.E., Carpten,J.D., Clark,S.J., Costello,J.F., Doerge,R.W., *et al.* (2008) Moving AHEAD with an international human epigenome project. *Nature*, **454**, 711–715.

19. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

20. Dawson,M.A. and Kouzarides,T. (2012) Cancer Epigenetics: From Mechanism to Therapy. *Cell*, **150**, 12–27.

21. Berger,S.L., Kouzarides,T., Shiekhattar,R. and Shilatifard,A. (2009) An operational definition of epigenetics. *Genes Dev.*, **23**, 781–783.

22. Language: Disputed definitions (2008) *Nat. News*, **455**, 1023–1028.

23. Cortessis,V.K., Thomas,D.C., Levine,A.J., Breton,C.V., Mack,T.M., Siegmund,K.D., Haile,R.W. and Laird,P.W. (2012) Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Hum. Genet.*, **131**, 1565–1589.

24. Felsenfeld,G. and Groudine,M. (2003) Controlling the double helix. *Nature*, **421**, 448–53.

25. Bird,A. (2007) Perceptions of epigenetics. *Nature*, **447**, 396–398.

26. Kouzarides,T. (2007) Chromatin Modifications and Their Function. *Cell*, **128**, 693–705.

27. Becker,P.B. and Hörz,W. (2002) ATP-dependent nucleosome remodeling. *Annu. Rev. Biochem.*, **71**, 247–273.

28. Narlikar,G.J., Fan,H.-Y. and Kingston,R.E. (2002) Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, **108**, 475–487.

29. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.

30. Feng,S., Jacobsen,S.E. and Reik,W. (2010) Epigenetic reprogramming in plant and animal development. *Science*, **330**, 622–627.

31. Ziller,M.J., Müller,F., Liao,J., Zhang,Y., Gu,H., Bock,C., Boyle,P., Epstein,C.B., Bernstein,B.E., Lengauer,T., *et al.* (2011) Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet.*, **7**, e1002389.

32. Ramsahoye,B.H., Biniszkiewicz,D., Lyko,F., Clark,V., Bird,A.P. and Jaenisch,R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 5237–5242.

33. 5-Methylcytosine (2015) *Wikipedia Free Encycl.*

34. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, **99**, 247–257.

35. Li,E., Bestor,T.H. and Jaenisch,R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

36. Bourc'his,D., Xu,G.L., Lin,C.S., Bollman,B. and Bestor,T.H. (2001) Dnmt3L and the establishment of maternal genomic imprints. *Science*, **294**, 2536–2539.

37. Hata,K., Okano,M., Lei,H. and Li,E. (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Dev. Camb. Engl.*, **129**, 1983–1993.

38. Bestor,T.H. and Ingram,V.M. (1983) Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proc. Natl. Acad. Sci. U. S. A.*, **80**, 5559–5563.

39. Kishikawa,S., Murata,T., Ugai,H., Yamazaki,T. and Yokoyama,K.K. (2003) Control elements of Dnmt1 gene are regulated in cell-cycle dependent manner. *Nucleic Acids Res. Suppl. 2001*.

40. Chuang,L.S., Ian,H.I., Koh,T.W., Ng,H.H., Xu,G. and Li,B.F. (1997) Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. *Science*, **277**, 1996–2000.

41. Sharif,J., Muto,M., Takebayashi,S., Suetake,I., Iwamatsu,A., Endo,T.A., Shinga,J., Mizutani-Koseki,Y., Toyoda,T., Okamura,K., *et al.* (2007) The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, **450**, 908–912.

42. Bostick,M., Kim,J.K., Estève,P.-O., Clark,A., Pradhan,S. and Jacobsen,S.E. (2007) UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, **317**, 1760–1764.

43. Arita,K., Ariyoshi,M., Tochio,H., Nakamura,Y. and Shirakawa,M. (2008) Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature*, **455**, 818–821.

44. Avvakumov,G.V., Walker,J.R., Xue,S., Li,Y., Duan,S., Bronner,C., Arrowsmith,C.H. and Dhe-Paganon,S. (2008) Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature*, **455**, 822–825.

45. Song,J., Rechkoblit,O., Bestor,T.H. and Patel,D.J. (2011) Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science*, **331**, 1036–1040.

46. Du,Z., Song,J., Wang,Y., Zhao,Y., Guda,K., Yang,S., Kao,H.-Y., Xu,Y., Willis,J., Markowitz,S.D., *et al.* (2010) DNMT1 stability is regulated

by proteins coordinating deubiquitination and acetylation-driven ubiquitination. *Sci. Signal.*, **3**, ra80.

47. Estève,P.-O., Chin,H.G., Benner,J., Feehery,G.R., Samaranayake,M., Horwitz,G.A., Jacobsen,S.E. and Pradhan,S. (2009) Regulation of DNMT1 stability through SET7-mediated lysine methylation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 5076–5081.

48. Jones,P.A. and Liang,G. (2009) Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.*, **10**, 805–811.

49. Jeltsch,A. (2006) On the enzymatic properties of DNMT1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme. *Epigenetics*, **1**, 63–66.

50. Jackson-Grusby,L., Beard,C., Possemato,R., Tudor,M., Fambrough,D., Csankovszki,G., Dausman,J., Lee,P., Wilson,C., Lander,E., *et al.* (2001) Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.*, **27**, 31–39.

51. Otani,J. (2009) Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO Rep*, **10**, 1235–1241.

52. Ooi,S.K. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, **448**, 714–717.

53. Conerly,M.L. (2010) Changes in H2A.Z occupancy & DNA methylation during B-cell lymphomagenesis. *Genome Res*, **20**, 1383–1390.

54. Zilberman,D., Coleman-Derr,D., Ballinger,T. and Henikoff,S. (2008) Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, **456**, 125–129.

55. Dong,K.B. (2008) DNA methylation in ES cells requires the lysine methyltransferase G9A but not its catalytic activity. *EMBO J*, **27**, 2691–2701.

56. Epsztejn-Litman,S. (2008) De novo DNA methylation promoted by G9A prevents reprogramming of embryonically silenced genes. *Nat. Struct Mol Biol*, **15**, 1176–1183.

57. Dennis,K., Fan,T., Geiman,T., Yan,Q. and Muegge,K. (2001) LSH, a member of the SNF2 family, is required for genome-wide methylation. *Genes Dev*, **15**, 2940–2944.

58. Zhu,H., Geiman,T.M., Xi,S., Jiang,Q., Schmidtmann,A., Chen,T., Li,E. and Muegge,K. (2006) Lsh is involved in de novo methylation of DNA. *EMBO J.*, **25**, 335–345.

59. Myant,K. and Stancheva,I. (2008) LSH cooperates with DNA methyltransferases to repress transcription. *Mol Cell Biol*, **28**, 215–226.

60. Myant,K. (2011) LSH and G9A/GLP complex are required for developmentally programmed DNA methylation. *Genome Res*, **21**, 83–94.

61. Bhutani,N., Brady,J.J., Damian,M., Sacco,A., Corbel,S.Y. and Blau,H.M. (2010) Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature*, **463**, 1042–1047.

62. Popp,C., Dean,W., Feng,S., Cokus,S.J., Andrews,S., Pellegrini,M., Jacobsen,S.E. and Reik,W. (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, **463**, 1101–1105.

63. Ono,R., Taki,T., Taketani,T., Taniwaki,M., Kobayashi,H. and Hayashi,Y. (2002) LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer Res.*, **62**, 4075–4080.

64. Meyer,C., Kowarz,E., Hofmann,J., Renneville,A., Zuna,J., Trka,J., Ben Abdelali,R., Macintyre,E., De Braekeleer,E., De Braekeleer,M., *et al.* (2009) New insights to the MLL recombinome of acute leukemias. *Leukemia*, **23**, 1490–1499.

65. Huang,Y. and Rao,A. (2014) Connections between TET proteins and aberrant DNA modification in cancer. *Trends Genet.*, **30**, 464–474.

66. He,Y.-F., Li,B.-Z., Li,Z., Liu,P., Wang,Y., Tang,Q., Ding,J., Jia,Y., Chen,Z., Li,L., *et al.* (2011) Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, **333**, 1303–1307.

67. Hashimoto,H., Zhang,X. and Cheng,X. (2013) Selective Excision of 5-Carboxylcytosine by a Thymine DNA Glycosylase Mutant. *J. Mol. Biol.*, **425**, 971–976.

68. Zhang,L., Lu,X., Lu,J., Liang,H., Dai,Q., Xu,G.-L., Luo,C., Jiang,H. and He,C. (2012) Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat. Chem. Biol.*, **8**, 328–330.

69. Maiti,A. and Drohat,A.C. (2011) Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. *J. Biol. Chem.*, **286**, 35334–35338.

70. Kohli,R.M. and Zhang,Y. (2013) TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, **502**, 472–479.

71. Wu,H., D'Alessio,A.C., Ito,S., Xia,K., Wang,Z., Cui,K., Zhao,K., Sun,Y.E. and Zhang,Y. (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, **473**, 389–393.

72. Williams,K. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, **473**, 343–348.

73. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.

74. Selker,E.U., Tountas,N.A., Cross,S.H., Margolin,B.S., Murphy,J.G., Bird,A.P. and Freitag,M. (2003) The methylated component of the Neurospora crassa genome. *Nature*, **422**, 893–897.

75. Bird,A.P., Taggart,M.H. and Smith,B.A. (1979) Methylated and unmethylated DNA compartments in the sea urchin genome. *Cell*, **17**, 889–901.

76. Tweedie,S., Charlton,J., Clark,V. and Bird,A. (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol*, **17**, 1469–75.

77. Montero,L.M., Filipski,J., Gil,P., Capel,J., Martínez-Zapater,J.M. and Salinas,J. (1992) The distribution of 5-methylcytosine in the nuclear genome of plants. *Nucleic Acids Res.*, **20**, 3207–3210.

78. Bird,A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, **8**, 1499–1504.

79. Bird,A., Taggart,M., Frommer,M., Miller,O.J. and Macleod,D. (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, **40**, 91–99.

80. Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.

81. Saxonov,S., Berg,P. and Brutlag,D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 1412–1417.

82. Weber,M., Hellmann,I., Stadler,M.B., Ramos,L., Pääbo,S., Rebhan,M. and Schübeler,D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.

83. Jones,P.A. (1999) The DNA methylation paradox. *Trends Genet. TIG*, **15**, 34–37.

84. Cooper,D.N., Taggart,M.H. and Bird,A.P. (1983) Unmethylated domains in vertebrate DNA. *Nucleic Acids Res.*, **11**, 647–658.

85. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.

86. Wu,H., Caffo,B., Jaffee,H.A., Irizarry,R.A. and Feinberg,A.P. (2010) Redefining CpG islands using hidden Markov models. *Biostatistics*, **11**, 499 –514.

87. Illingworth,R., Kerr,A., Desousa,D., Jørgensen,H., Ellis,P., Stalker,J., Jackson,D., Clee,C., Plumb,R., Rogers,J., *et al.* (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.*, **6**, e22.

88. ENCODE Project Consortium, Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigó,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

89. Illingworth,R.S. and Bird,A.P. (2009) CpG islands – 'A rough guide'. *FEBS Lett.*, **583**, 1713–1720.

90. Kelly,T.K., Miranda,T.B., Liang,G., Berman,B.P., Lin,J.C., Tanay,A. and Jones,P.A. (2010) H2A.Z maintenance during mitosis reveals nucleosome shifting on mitotically silenced genes. *Mol. Cell*, **39**, 901–911.

91. Gal-Yam,E.N., Egger,G., Iniguez,L., Holster,H., Einarsson,S., Zhang,X., Lin,J.C., Liang,G., Jones,P.A. and Tanay,A. (2008) Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 12979–12984.

92. Weber,M., Davies,J.J., Wittig,D., Oakeley,E.J., Haase,M., Lam,W.L. and Schübeler,D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.

93. Illingworth,R.S. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet*, **6**, e1001134.

94. Hashimshony,T., Zhang,J., Keshet,I., Bustin,M. and Cedar,H. (2003) The role of DNA methylation in setting up chromatin structure during development. *Nat. Genet.*, **34**, 187–192.

95. Kass,S.U., Landsberger,N. and Wolffe,A.P. (1997) DNA methylation directs a time-dependent repression of transcription initiation. *Curr. Biol. CB*, **7**, 157–165.

96. Venolia,L. and Gartler,S.M. (1983) Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. *Nature*, **302**, 82–83.

97. Yoder,J.A., Walsh,C.P. and Bestor,T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet. TIG*, **13**, 335–340.

98. Lyon,M.F. (1961) Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature*, **190**, 372–373.

99. Russell,L.B. (1963) Mammalian X-chromosome action: inactivation limited in spread and region of origin. *Science*, **140**, 976–978.

100. Brown,C.J., Hendrich,B.D., Rupert,J.L., Lafrenière,R.G., Xing,Y., Lawrence,J. and Willard,H.F. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527–542.

101. Nesbit,M.N. (1971) X chromosome inactivation mosaicism in the mouse. *Dev. Biol.*, **26**, 252–263.

102. McMahon,A., Fosten,M. and Monk,M. (1983) X-chromosome inactivation mosaicism in the three germ layers and the germ line of the mouse embryo. *J. Embryol. Exp. Morphol.*, **74**, 207–220.

103. Plath,K., Fang,J., Mlynarczyk-Evans,S.K., Cao,R., Worringer,K.A., Wang,H., de la Cruz,C.C., Otte,A.P., Panning,B. and Zhang,Y. (2003) Role of histone H3 lysine 27 methylation in X inactivation. *Science*, **300**, 131–135.

104. Plath,K., Talbot,D., Hamer,K.M., Otte,A.P., Yang,T.P., Jaenisch,R. and Panning,B. (2004) Developmentally regulated alterations in Polycomb repressive complex 1 proteins on the inactive X chromosome. *J. Cell Biol.*, **167**, 1025–1035.

105. Silva,J., Mak,W., Zvetkova,I., Appanah,R., Nesterova,T.B., Webster,Z., Peters,A.H.F.M., Jenuwein,T., Otte,A.P. and Brockdorff,N. (2003) Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev. Cell*, **4**, 481–495.

106. Okamoto,I., Otte,A.P., Allis,C.D., Reinberg,D. and Heard,E. (2004) Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science*, **303**, 644–649.

107. Nora,E.P. and Heard,E. (2010) Chromatin structure and nuclear organization dynamics during X-chromosome inactivation. *Cold Spring Harb. Symp. Quant. Biol.*, **75**, 333–344.

108. Ferguson-Smith,A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.*, **12**, 565–575.

109. Morison,I.M., Ramsay,J.P. and Spencer,H.G. (2005) A census of mammalian imprinting. *Trends Genet. TIG*, **21**, 457–465.

110. Kelsey,G. and Feil,R. (2013) New insights into establishment and maintenance of DNA methylation imprints in mammals. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **368**, 20110336.

111. Abramowitz,L.K. and Bartolomei,M.S. (2012) Genomic imprinting: recognition and marking of imprinted loci. *Curr. Opin. Genet. Dev.*, **22**, 72–78.

112. Smallwood,S.A. and Kelsey,G. (2012) De novo DNA methylation: a germ cell perspective. *Trends Genet*, **28**, 33–42.

113. Kota,S.K. and Feil,R. (2010) Epigenetic transitions in germ cell development and meiosis. *Dev. Cell*, **19**, 675–686.

114. Engel,N., Thorvaldsen,J.L. and Bartolomei,M.S. (2006) CTCF binding sites promote transcription initiation and prevent DNA methylation on the maternal allele at the imprinted H19/Igf2 locus. *Hum. Mol. Genet.*, **15**, 2945–2954.

115. Schoenherr,C.J., Levorse,J.M. and Tilghman,S.M. (2003) CTCF maintains differential methylation at the Igf2/H19 locus. *Nat. Genet.*, **33**, 66–69.

116. Pant,V., Mariano,P., Kanduri,C., Mattsson,A., Lobanenkov,V., Heuchel,R. and Ohlsson,R. (2003) The nucleotides responsible for the direct physical contact between the chromatin insulator protein CTCF and the H19 imprinting control region manifest parent of

origin-specific long-distance insulation and methylation-free domains. *Genes Dev.*, **17**, 586–590.

117. Fedoriw,A.M., Stein,P., Svoboda,P., Schultz,R.M. and Bartolomei,M.S. (2004) Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science*, **303**, 238–240.

118. Sanli,I. and Feil,R. Chromatin mechanisms in the developmental control of imprinted gene expression. *Int. J. Biochem. Cell Biol.*, 10.1016/j.biocel.2015.04.004.

119. Waterston,R.H. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

120. Batzer,M.A. and Deininger,P.L. (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.*, **3**, 370–379.

121. Deininger,P.L. and Batzer,M.A. (1999) Alu repeats and human disease. *Mol. Genet. Metab.*, **67**, 183–193.

122. Jackson,M., Krassowska,A., Gilbert,N., Chevassut,T., Forrester,L., Ansell,J. and Ramsahoye,B. (2004) Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Mol. Cell. Biol.*, **24**, 8862–8871.

123. Trowbridge,J.J., Snow,J.W., Kim,J. and Orkin,S.H. (2009) DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. *Cell Stem Cell*, **5**, 442–449.

124. Broske,A.M. (2009) DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nat. Genet*, **41**, 1207–1215.

125. Hodges,E. (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell*, **44**, 17–28.

126. Deaton,A.M. (2011) Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res*, **21**, 1074–1086.

127. Bock,C. (2012) DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol Cell*, **47**, 633–647.

128. Lock,L.F., Takagi,N. and Martin,G.R. (1987) Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*, **48**, 39–46.

129. Ohm,J.E., McGarvey,K.M., Yu,X., Cheng,L., Schuebel,K.E., Cope,L., Mohammad,H.P., Chen,W., Daniel,V.C., Yu,W., *et al.* (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.*, **39**, 237–242.

130. Schlesinger,Y., Straussman,R., Keshet,I., Farkash,S., Hecht,M., Zimmerman,J., Eden,E., Yakhini,Z., Ben-Shushan,E., Reubinoff,B.E., *et al.* (2007) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.*, **39**, 232–236.

131. Widschwendter,M., Fiegl,H., Egle,D., Mueller-Holzner,E., Spizzo,G., Marth,C., Weisenberger,D.J., Campan,M., Young,J., Jacobs,I., *et al.* (2007) Epigenetic stem cell signature in cancer. *Nat. Genet.*, **39**, 157–158.

132. Challen,G.A. (2012) DNMT3A is essential for hematopoietic stem cell differentiation. *Nat. Genet*, **44**, 23–31.

133. Tian,Z., Tolić,N., Zhao,R., Moore,R.J., Hengel,S.M., Robinson,E.W., Stenoien,D.L., Wu,S., Smith,R.D. and Paša-Tolić,L. (2012) Enhanced top-down characterization of histone post-translational modifications. *Genome Biol.*, **13**, R86.

134. Tan,M., Luo,H., Lee,S., Jin,F., Yang,J.S., Montellier,E., Buchou,T., Cheng,Z., Rousseaux,S., Rajagopal,N., *et al.* (2011) Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification. *Cell*, **146**, 1016–1028.

135. Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

136. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotech*, **28**, 817–825.

137. Bernstein,B.E., Kamal,M., Lindblad-Toh,K., Bekiranov,S., Bailey,D.K., Huebert,D.J., McMahon,S., Karlsson,E.K., Kulbokas,E.J., Gingeras,T.R., *et al.* (2005) Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell*, **120**, 169–181.

138. Kim,T.H., Barrera,L.O., Zheng,M., Qu,C., Singer,M.A., Richmond,T.A., Wu,Y., Green,R.D. and Ren,B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.

139. Pokholok,D.K., Harbison,C.T., Levine,S., Cole,M., Hannett,N.M., Lee,T.I., Bell,G.W., Walker,K., Rolfe,P.A., Herbolsheimer,E., *et al.* (2005) Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell*, **122**, 517–527.

140. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K., *et al.* (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, **125**, 315–326.

141. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, **129**, 823–837.

142. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A., *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

143. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A., *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.*, **107**, 21931–21936.

144. Rada-Iglesias,A., Bajpai,R., Swigut,T., Brugmann,S.A., Flynn,R.A. and Wysocka,J. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283.

145. Heintzman,N.D. and Ren,B. (2009) Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.*, **19**, 541–549.

146. Lee,T.I., Jenner,R.G., Boyer,L.A., Guenther,M.G., Levine,S.S., Kumar,R.M., Chevalier,B., Johnstone,S.E., Cole,M.F., Isono,K., *et al.* (2006) Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. *Cell*, **125**, 301–313.

147. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.-K., Koche,R.P., *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

148. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F., *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

149. Seto,A.G., Kingston,R.E. and Lau,N.C. (2007) The Coming of Age for Piwi Proteins. *Mol. Cell*, **26**, 603–609.

150. Cox,D.N., Chao,A., Baker,J., Chang,L., Qiao,D. and Lin,H. (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev.*, **12**, 3715–3727.

151. Cox,D.N., Chao,A. and Lin,H. (2000) Piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development*, **127**, 503–514.

152. Szakmary,A., Cox,D.N., Wang,Z. and Lin,H. (2005) Regulatory Relationship among piwi, pumilio, and bag-of-marbles in Drosophila Germline Stem Cell Self-Renewal and Differentiation. *Curr. Biol.*, **15**, 171–178.

153. Jia,H., Osak,M., Bogu,G.K., Stanton,L.W., Johnson,R. and Lipovich,L. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, **16**, 1478–1487.

154. Ørom,U.A., Derrien,T., Beringer,M., Gumireddy,K., Gardini,A., Bussotti,G., Lai,F., Zytnicki,M., Notredame,C., Huang,Q., *et al.* (2010)

Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell*, **143**, 46–58.

155. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

156. Lee,J.T. (2012) Epigenetic Regulation by Long Noncoding RNAs. *Science*, **338**, 1435–1439.

157. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K., *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.

158. Szutorisz,H., Canzonetta,C., Georgiou,A., Chow,C.-M., Tora,L. and Dillon,N. (2005) Formation of an Active Tissue-Specific Chromatin Domain Initiated by Epigenetic Marking at the Embryonic Stem Cell Stage. *Mol. Cell. Biol.*, **25**, 1804–1820.

159. Azuara,V., Perry,P., Sauer,S., Spivakov,M., Jørgensen,H.F., John,R.M., Gouti,M., Casanova,M., Warnes,G., Merkenschlager,M., *et al.* (2006) Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.*, **8**, 532–538.

160. Feldman,N., Gerson,A., Fang,J., Li,E., Zhang,Y., Shinkai,Y., Cedar,H. and Bergman,Y. (2006) G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat. Cell Biol.*, **8**, 188–194.

161. Hajkova,P., Erhardt,S., Lane,N., Haaf,T., El-Maarri,O., Reik,W., Walter,J. and Surani,M.A. (2002) Epigenetic reprogramming in mouse primordial germ cells. *Mech. Dev.*, **117**, 15–23.

162. Seki,Y., Hayashi,K., Itoh,K., Mizugaki,M., Saitou,M. and Matsui,Y. (2005) Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Dev. Biol.*, **278**, 440–458.

163. Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.

164. Lee,J., Inoue,K., Ono,R., Ogonuki,N., Kohda,T., Kaneko-Ishino,T., Ogura,A. and Ishino,F. (2002) Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. *Dev. Camb. Engl.*, **129**, 1807–1817.

165. Pen,I., Uller,T., Feldmeyer,B., Harts,A., While,G.M. and Wapstra,E. (2010) Climate-driven population divergence in sex-determining systems. *Nature*, **468**, 436–438.

166. Kim,D.-H. and Sung,S. (2014) Genetic and Epigenetic Mechanisms Underlying Vernalization. *Arab. Book Am. Soc. Plant Biol.*, **12**.

167. Kamakura,M. (2011) Royalactin induces queen differentiation in honeybees. *Nature*, **473**, 478–483.

168. Weaver,I.C.G., Cervoni,N., Champagne,F.A., D'Alessio,A.C., Sharma,S., Seckl,J.R., Dymov,S., Szyf,M. and Meaney,M.J. (2004) Epigenetic programming by maternal behavior. *Nat. Neurosci.*, **7**, 847–854.

169. Uno,E. and Berry,D. (2012) X inactivation and Epigenetics. Walter and Eliza Hall Institute of Medical Research (c) The University of Melbourne

170. Gordon,L., Joo,J.-H.E., Andronikos,R., Ollikainen,M., Wallace,E.M., Umstad,M.P., Permezel,M., Oshlack,A., Morley,R., Carlin,J.B., *et al.* (2011) Expression discordance of monozygotic twins at birth: effect of intrauterine environment and a possible mechanism for fetal programming. *Epigenetics Off. J. DNA Methylation Soc.*, **6**, 579–592.

171. Talens,R.P., Boomsma,D.I., Tobi,E.W., Kremer,D., Jukema,J.W., Willemsen,G., Putter,H., Slagboom,P.E. and Heijmans,B.T. (2010) Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.*, **24**, 3135–3144.

172. Pembrey,M., Saffery,R., Bygren,L.O., Network in Epigenetic Epidemiology and Network in Epigenetic Epidemiology (2014) Human transgenerational responses to early-life experience: potential

impact on development, health and biomedical research. *J. Med. Genet.*, **51**, 563–572.

173. Daxinger,L. and Whitelaw,E. (2012) Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat. Rev. Genet.*, **13**, 153–162.

174. Waterland,R.A. and Jirtle,R.L. (2003) Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol. Cell. Biol.*, **23**, 5293–5300.

175. Morgan,H.D., Sutherland,H.G., Martin,D.I. and Whitelaw,E. (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.*, **23**, 314–318.

176. Wolff,G.L., Kodell,R.L., Moore,S.R. and Cooney,C.A. (1998) Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.*, **12**, 949–957.

177. Kaminen-Ahola,N., Ahola,A., Maga,M., Mallitt,K.-A., Fahey,P., Cox,T.C., Whitelaw,E. and Chong,S. (2010) Maternal ethanol consumption alters the epigenotype and the phenotype of offspring in a mouse model. *PLoS Genet.*, **6**, e1000811.

178. Cooney,C.A., Dave,A.A. and Wolff,G.L. (2002) Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *J. Nutr.*, **132**, 2393S–2400S.

179. Dolinoy,D.C. (2008) The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. *Nutr. Rev.*, **66**, S7–11.

180. Druker,R., Bruxner,T.J., Lehrbach,N.J. and Whitelaw,E. (2004) Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. *Nucleic Acids Res.*, **32**, 5800–5808.

181. Weinhouse,C., Anderson,O.S., Jones,T.R., Kim,J., Liberman,S.A., Nahar,M.S., Rozek,L.S., Jirtle,R.L. and Dolinoy,D.C. (2011) An expression microarray approach for the identification of metastable

epialleles in the mouse genome. *Epigenetics Off. J. DNA Methylation Soc.*, **6**, 1105–1113.

182. Waterland,R.A., Kellermayer,R., Laritsky,E., Rayco-Solon,P., Harris,R.A., Travisano,M., Zhang,W., Torskaya,M.S., Zhang,J., Shen,L., *et al.* (2010) Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.*, **6**, e1001252.

183. Heard,E. and Martienssen,R.A. (2014) Transgenerational Epigenetic Inheritance: Myths and Mechanisms. *Cell*, **157**, 95–109.

184. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Peng,W., Zhang,M.Q., *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

185. Bannister,A.J. and Kouzarides,T. (1996) The CBP co-activator is a histone acetyltransferase. *Nature*, **384**, 641–643.

186. Johnstone,R.W. and Licht,J.D. (2003) Histone deacetylase inhibitors in cancer therapy: is transcription the primary target? *Cancer Cell*, **4**, 13–18.

187. Federico,M. and Bagella,L. (2011) Histone deacetylase inhibitors in the treatment of hematological malignancies and solid tumors. *J. Biomed. Biotechnol.*, **2011**, 475641.

188. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V., Sonkin,D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

189. Schenk,T., Chen,W.C., Göllner,S., Howell,L., Jin,L., Hebestreit,K., Klein,H.-U., Popescu,A.C., Burnett,A., Mills,K., *et al.* (2012) Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. *Nat. Med.*, **18**, 605–611.

190. Mosammaparast,N. and Shi,Y. (2010) Reversal of histone methylation: biochemical and molecular mechanisms of histone demethylases. *Annu. Rev. Biochem.*, **79**, 155–179.

191. Easwaran,H., Johnstone,S.E., Van Neste,L., Ohm,J., Mosbruger,T., Wang,Q., Aryee,M.J., Joyce,P., Ahuja,N., Weisenberger,D., *et al.* (2012) A DNA hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Res.*, **22**, 837–849.

192. Esteller,M. (2008) Epigenetics in cancer. *N Engl J Med*, **358**, 1148–59.

193. Feinberg,A.P. and Tycko,B. (2004) The history of cancer epigenetics. *Nat Rev Cancer*, **4**, 143–53.

194. Baylin,S.B., Fearon,E.R., Vogelstein,B., de Bustros,A., Sharkis,S.J., Burke,P.J., Staal,S.P. and Nelkin,B.D. (1987) Hypermethylation of the 5' region of the calcitonin gene is a property of human lymphoid and acute myeloid malignancies. *Blood*, **70**, 412–7.

195. Esteller,M. (2007) Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet*, **16 Spec No 1**, R50–9.

196. Baylin,S.B. and Jones,P.A. (2011) A decade of exploring the cancer epigenome - biological and translational implications. *Nat. Rev. Cancer*, **11**, 726–734.

197. Feinberg,A.P. and Vogelstein,B. (1983) Hypomethylation of ras oncogenes in primary human cancers. *Biochem. Biophys. Res. Commun.*, **111**, 47–54.

198. Feinberg,A.P. and Vogelstein,B. (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, **301**, 89–92.

199. Feinberg,A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.

200. Ehrlich,M. (2002) DNA methylation in cancer: too much, but also too little. *Oncogene*, **21**, 5400–13.

201. Holliday,R. and Pugh,J.E. (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.

202. Riggs,A.D. (1975) X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, **14**, 9–25.

203. Kuo,K.C., McCune,R.A., Gehrke,C.W., Midgett,R. and Ehrlich,M. (1980) Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. *Nucleic Acids Res.*, **8**, 4763–4776.

204. Annan,R.S., Kresbach,G.M., Giese,R.W. and Vouros,P. (1989) Trace detection of modified DNA bases via moving-belt liquid chromatography-mass spectrometry using electrophoric derivatization and negative chemical ionization. *J. Chromatogr.*, **465**, 285–296.

205. Adouard,V., Dante,R., Niveleau,A., Delain,E., Revet,B. and Ehrlich,M. (1985) The accessibility of 5-methylcytosine to specific antibodies in double-stranded DNA of Xanthomonas phage XP12. *Eur. J. Biochem. FEBS*, **152**, 115–121.

206. Oakeley,E.J., Podestà,A. and Jost,J.P. (1997) Developmental changes in DNA methylation of the two tobacco pollen nuclei during maturation. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 11721–11725.

207. Santos,F., Hendrich,B., Reik,W. and Dean,W. (2002) Dynamic reprogramming of DNA methylation in the early mouse embryo. *Dev. Biol.*, **241**, 172–182.

208. Parle-Mcdermott,A. and Harrison,A. (2011) DNA methylation: a timeline of methods and applications. *Epigenomics Epigenetics*, **2**, 74.

209. Bickle,T.A. and Krüger,D.H. (1993) Biology of DNA restriction. *Microbiol. Rev.*, **57**, 434–450.

210. Cedar,H., Solage,A., Glaser,G. and Razin,A. (1979) Direct detection of methylated cytosine in DNA by use of the restriction enzyme MspI. *Nucleic Acids Res.*, **6**, 2125–2132.

211. Bestor,T.H., Hellewell,S.B. and Ingram,V.M. (1984) Differentiation of two mouse cell lines is associated with hypomethylation of their genomes. *Mol. Cell. Biol.*, **4**, 1800–1806.

212. Wagner,I. and Capesius,I. (1981) Determination of 5-methylcytosine from plant DNA by high-performance liquid chromatography. *Biochim. Biophys. Acta*, **654**, 52–56.

213. Gama-Sosa,M.A., Midgett,R.M., Slagel,V.A., Githens,S., Kuo,K.C., Gehrke,C.W. and Ehrlich,M. (1983) Tissue-specific differences in DNA methylation in various mammals. *Biochim. Biophys. Acta*, **740**, 212–219.

214. Kriaucionis,S. and Heintz,N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, **324**, 929–930.

215. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L., *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.

216. Mullis,K., Faloona,F., Scharf,S., Saiki,R., Horn,G. and Erlich,H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.*, **51 Pt 1**, 263–273.

217. Gonzalgo,M.L., Liang,G., Spruck,C.H., Zingg,J.M., Rideout,W.M. and Jones,P.A. (1997) Identification and characterization of differentially methylated regions of genomic DNA by methylation-sensitive arbitrarily primed PCR. *Cancer Res.*, **57**, 594–599.

218. Liang,G., Gonzalgo,M.L., Salem,C. and Jones,P.A. (2002) Identification of DNA methylation differences during tumorigenesis by methylation-sensitive arbitrarily primed polymerase chain reaction. *Methods San Diego Calif*, **27**, 150–155.

219. Frigola,J., Ribas,M., Risques,R.A. and Peinado,M.A. (2002) Methylome profiling of cancer cells by amplification of inter-methylated sites (AIMS). *Nucleic Acids Res*, **30**, e28.

220. Hayatsu,H., Wataya,Y., Kai,K. and Iida,S. (1970) Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry (Mosc.)*, **9**, 2858–2865.

221. Wang,R.Y., Gehrke,C.W. and Ehrlich,M. (1980) Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res.*, **8**, 4777–4790.

222. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 1827–1831.

223. Herman,J.G., Graff,J.R., Myöhänen,S., Nelkin,B.D. and Baylin,S.B. (1996) Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 9821–9826.

224. Xiong,Z. and Laird,P.W. (1997) COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res.*, **25**, 2532–2534.

225. Wojdacz,T.K. and Dobrovic,A. (2007) Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. *Nucleic Acids Res.*, **35**, e41.

226. Wojdacz,T.K., Dobrovic,A. and Hansen,L.L. (2008) Methylation-sensitive high-resolution melting. *Nat. Protoc.*, **3**, 1903–1908.

227. Jujubix (2011) English: Bisulfite conversion of unmethylated and methylated cytosine DNA residues when treating cytosine and 5-methylcytosine with sodium bisulfite. Often used to determine DNA methylation status.

228. Huang,T.H., Perry,M.R. and Laux,D.E. (1999) Methylation profiling of CpG islands in human breast cancer cells. *Hum. Mol. Genet.*, **8**, 459–470.

229. Weber,M., Davies,J.J., Wittig,D., Oakeley,E.J., Haase,M., Lam,W.L. and Schübeler,D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.

230. Keshet,I., Schlesinger,Y., Farkash,S., Rand,E., Hecht,M., Segal,E., Pikarski,E., Young,R.A., Niveleau,A., Cedar,H., *et al.* (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.*, **38**, 149–153.

231. Beck,S. and Rakyan,V.K. (2008) The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, **24**, 231–237.

232. Reinders,J., Delucinge Vivier,C., Theiler,G., Chollet,D., Descombes,P. and Paszkowski,J. (2008) Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. *Genome Res.*, **18**, 469–476.

233. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

234. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M., *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

235. Maunakea,A.K., Nagarajan,R.P., Bilenky,M., Ballinger,T.J., D'Souza,C., Fouse,S.D., Johnson,B.E., Hong,C., Nielsen,C., Zhao,Y., *et al.* (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, **466**, 253–257.

236. Meissner,A., Gnirke,A., Bell,G.W., Ramsahoye,B., Lander,E.S. and Jaenisch,R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.*, **33**, 5868–5877.

237. Lister,R., Pelizzola,M., Kida,Y.S., Hawkins,R.D., Nery,J.R., Hon,G., Antosiewicz-Bourget,J., O'Malley,R., Castanon,R., Klugman,S., *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.

238. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

239. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078 –2079.

240. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R., *et al.* (2011) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.

241. Illingworth,R.S. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet*, **6**, e1001134.

242. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841 –842.

243. Du,P., Zhang,X., Huang,C.-C., Jafari,N., Kibbe,W., Hou,L. and Lin,S. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

244. Harris,R.A., Wang,T., Coarfa,C., Nagarajan,R.P., Hong,C., Downey,S.L., Johnson,B.E., Fouse,S.D., Delaney,A., Zhao,Y., *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.

245. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M., *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–763.

246. Sherry,S.T., Ward,M.-H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

247. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M., *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

248. Diez-Villanueva,A., Malinverni,R. and Gel,B. (2015) regioneR: Association analysis of genomic regions based on permutation tests. R package version 1.0.3.

249. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–208.

250. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C., Chou,A., Ienasescu,H., *et al.* (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 10.1093/nar/gkt997.

251. Kulakovskiy,I.V., Medvedeva,Y.A., Schaefer,U., Kasianov,A.S., Vorontsov,I.E., Bajic,V.B. and Makeev,V.J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, D195–D202.

252. Vesely,P.W., Staber,P.B., Hoefler,G. and Kenner,L. (2009) Translational regulation mechanisms of AP-1 proteins. *Mutat. Res. Mutat. Res.*, **682**, 7–12.

253. Morello,D., Fitzgerald,M.J., Babinet,C. and Fausto,N. (1990) c-myc, c-fos, and c-jun regulation in the regenerating livers of normal and H-2K/c-myc transgenic mice. *Mol. Cell. Biol.*, **10**, 3185–3193.

254. Doi,A., Park,I.-H., Wen,B., Murakami,P., Aryee,M.J., Irizarry,R., Herb,B., Ladd-Acosta,C., Rho,J., Loewer,S., *et al.* (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.*, **41**, 1350–1353.

255. Nilubol,N., Boufraqech,M., Zhang,L. and Kebebew,E. (2014) Loss of CPSF2 expression is associated with increased thyroid cancer cellular

invasion and cancer stem cell population, and more aggressive disease. *J. Clin. Endocrinol. Metab.*, **99**, E1173–1182.

256. Sung,T.Y., Kim,M., Kim,T.Y., Kim,W.G., Park,Y., Song,D.E., Park,S.-Y., Kwon,H., Choi,Y.M., Jang,E.K., *et al.* (2015) Negative Expression of CPSF2 Predicts a Poorer Clinical Outcome in Patients with Papillary Thyroid Carcinoma. *Thyroid Off. J. Am. Thyroid Assoc.*, **25**, 1020–1025.

257. Garraway,L.A., Widlund,H.R., Rubin,M.A., Getz,G., Berger,A.J., Ramaswamy,S., Beroukhim,R., Milner,D.A., Granter,S.R., Du,J., *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436**, 117–122.

258. Degen,W.G., van Kempen,L.C., Gijzen,E.G., van Groningen,J.J., van Kooyk,Y., Bloemers,H.P. and Swart,G.W. (1998) MEMD, a new cell adhesion molecule in metastasizing human melanoma cell lines, is identical to ALCAM (activated leukocyte cell adhesion molecule). *Am. J. Pathol.*, **152**, 805–813.

259. Jannie,K.M., Stipp,C.S. and Weiner,J.A. (2012) ALCAM Regulates Motility, Invasiveness, and Adherens Junction Formation in Uveal Melanoma Cells. *PLoS ONE*, **7**.

260. Hermann,H., Fabrizio,P., Raker,V.A., Foulaki,K., Hornig,H., Brahms,H. and Lührmann,R. (1995) snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. *EMBO J.*, **14**, 2076–2088.

261. Wang,D.B., Kinoshita,Y., Kinoshita,C., Uo,T., Sopher,B.L., Cudaback,E., Keene,C.D., Bilousova,T., Gylys,K., Case,A., *et al.* (2015) Loss of endophilin-B1 exacerbates Alzheimer's disease pathology. *Brain J. Neurol.*, **138**, 2005–2019.

262. Warnecke,P.M., Stirzaker,C., Melki,J.R., Millar,D.S., Paul,C.L. and Clark,S.J. (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res.*, **25**, 4422–4426.

263. Reins,J., Mossner,M., Richter,L., Kmetsch,A., Thiel,E., Haase,D. and Hofmann,W.-K. (2011) [Letter to the editor]: whole-genome amplification of sodium bisulfite-converted DNA can substantially impact quantitative methylation analysis using pyrosequencing. *BioTechniques*, **50**, 161–164.

264. Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet*, **8**, 286–98.

265. Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*, **11**, 191–203.

266. Pelizzola,M. and Ecker,J.R. (2011) The DNA methylome. *FEBS Lett.*, **585**, 1994–2000.

267. Nair,S.S., Coolen,M.W., Stirzaker,C., Song,J.Z., Statham,A.L., Strbenac,D., Robinson,M.D. and Clark,S.J. (2011) Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics*, **6**, 34–44.

268. Robinson,M.D., Statham,A.L., Speed,T.P. and Clark,S.J. (2010) Protocol matters: which methylome are you actually studying? *Epigenomics*, **2**, 587–598.

269. Rajendram,R., Ferreira,J.C., Grafodatskaya,D., Choufani,S., Chiang,T., Pu,S., Butcher,D.T., Wodak,S.J. and Weksberg,R. (2011) Assessment of methylation level prediction accuracy in methyl-DNA immunoprecipitation and sodium bisulfite based microarray platforms. *Epigenetics*, **6**, 410–415.

270. Laurent,L., Wong,E., Li,G., Huynh,T., Tsirigos,A., Ong,C.T., Low,H.M., Kin Sung,K.W., Rigoutsos,I., Loring,J., *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.

271. Hansen,K.D., Timp,W., Bravo,H.C., Sabunciyan,S., Langmead,B., McDonald,O.G., Wen,B., Wu,H., Liu,Y., Diep,D., *et al.* (2011)

Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

272. Li,Y., Zhu,J., Tian,G., Li,N., Li,Q., Ye,M., Zheng,H., Yu,J., Wu,H., Sun,J., *et al.* (2010) The DNA Methylome of Human Peripheral Blood Mononuclear Cells. *PLoS Biol*, **8**, e1000533.

273. Heyn,H., Li,N., Ferreira,H.J., Moran,S., Pisano,D.G., Gomez,A., Diez,J., Sanchez-Mut,J.V., Setien,F., Carmona,F.J., *et al.* (2012) Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 10522–10527.

274. Heyn,H., Vidal,E., Sayols,S., Sanchez-Mut,J.V., Moran,S., Medina,I., Sandoval,J., Simó-Riudalbas,L., Szczesna,K., Huertas,D., *et al.* (2012) Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient. *Epigenetics*, **7**, 542–550.

275. Suzuki,M., Jing,Q., Lia,D., Pascual,M., McLellan,A. and Greally,J.M. (2010) Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol.*, **11**, R36.

276. Gupta,R., Nagarajan,A. and Wajapeyee,N. (2010) Advances in genome-wide DNA methylation analysis. *BioTechniques*, **49**, iii–xi.

277. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M., *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.

278. Pastor,W.A., Pape,U.J., Huang,Y., Henderson,H.R., Lister,R., Ko,M., McLoughlin,E.M., Brudno,Y., Mahapatra,S., Kapranov,P., *et al.* (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, **473**, 394–397.

279. Williams,K., Christensen,J. and Helin,K. (2012) DNA methylation: TET proteins-guardians of CpG islands? *EMBO Rep.*, **13**, 28–35.

280. Singer,M., Boffelli,D., Dhahbi,J., Schönhuth,A., Schroth,G.P., Martin,D.I.K. and Pachter,L. (2010) MetMap enables genome-scale

Methyltyping for determining methylation states in populations. *PLoS Comput. Biol.*, **6**, e1000888.

281. Krueger,F., Kreck,B., Franke,A. and Andrews,S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.

282. Martin,D.I.K., Singer,M., Dhahbi,J., Mao,G., Zhang,L., Schroth,G.P., Pachter,L. and Boffelli,D. (2011) Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Res.*, **21**, 2049–2057.

283. Schultz,M.D., He,Y., Whitaker,J.W., Hariharan,M., Mukamel,E.A., Leung,D., Rajagopal,N., Nery,J.R., Urich,M.A., Chen,H., *et al.* (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, **523**, 212–216.

284. Kellis,M., Wold,B., Snyder,M.P., Bernstein,B.E., Kundaje,A., Marinov,G.K., Ward,L.D., Birney,E., Crawford,G.E., Dekker,J., *et al.* (2014) Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.*, **111**, 6131–6138.

285. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F., *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.

286. Grosveld,F., van Assendelft,G.B., Greaves,D.R. and Kollias,G. (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell*, **51**, 975–985.

287. Agarwal,S. and Rao,A. (1998) Long-range transcriptional regulation of cytokine gene expression. *Curr. Opin. Immunol.*, **10**, 345–352.

288. Lakshmanan,G., Lieuw,K.H., Grosveld,F. and Engel,J.D. (1998) Partial rescue of GATA-3 by yeast artificial chromosome transgenes. *Dev. Biol.*, **204**, 451–463.

289. Noonan,J.P. and McCallion,A.S. (2010) Genomics of long-range regulatory elements. *Annu. Rev. Genomics Hum. Genet.*, **11**, 1–23.

290. Nardone,J., Lee,D.U., Ansel,K.M. and Rao,A. (2004) Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA. *Nat. Immunol.*, **5**, 768–774.

291. Gross,D.S. and Garrard,W.T. (1988) Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, **57**, 159–197.

292. Li,C.C., Ramirez-Carrozzi,V.R. and Smale,S.T. (2006) Pursuing gene regulation 'logic' via RNA interference and chromatin immunoprecipitation. *Nat. Immunol.*, **7**, 692–697.

293. Weinmann,A.S. and Farnham,P.J. (2002) Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods San Diego Calif*, **26**, 37–47.

294. Johnson,K.D. and Bresnick,E.H. (2002) Dissecting long-range transcriptional mechanisms by chromatin immunoprecipitation. *Methods San Diego Calif*, **26**, 27–36.

295. Burandt,E., Bari Noubar,T., Lebeau,A., Minner,S., Burdelski,C., Jänicke,F., Müller,V., Terracciano,L., Simon,R., Sauter,G., *et al.* (2014) Loss of ALCAM expression is linked to adverse phenotype and poor prognosis in breast cancer: a TMA-based immunohistochemical study on 2,197 breast cancer patients. *Oncol. Rep.*, **32**, 2628–2634.

296. Tachezy,M., Zander,H., Wolters-Eisfeld,G., Müller,J., Wicklein,D., Gebauer,F., Izbicki,J.R. and Bockhorn,M. (2014) Activated Leukocyte Cell Adhesion Molecule (CD166): An 'Inert' Cancer Stem Cell Marker for Non-Small Cell Lung Cancer? *STEM CELLS*, **32**, 1429–1436.

297. Zhang,W.C., Shyh-Chang,N., Yang,H., Rai,A., Umashankar,S., Ma,S., Soh,B.S., Sun,L.L., Tai,B.C., Nga,M.E., *et al.* (2012) Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell*, **148**, 259–272.

Anexo

Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale.

Víctor Barrera and Miguel A. Peinado

11490–11498 Nucleic Acids Research, 2012, Vol. 40, No. 22
doi:10.1093/nar/gks928

11490–11498 *Nucleic Acids Research, 2012, Vol. 40, No. 22* *Published online 12 October 2012*
doi:10.1093/nar/gks928

Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale

Víctor Barrera and Miguel A. Peinado*

Institute of Predictive and Personalized Medicine of Cancer (IMPPC), Badalona, Barcelona, Spain

Received May 28, 2012; Revised August 21, 2012; Accepted September 14, 2012

ABSTRACT

Methylation of a CpG island is a faithful marker of silencing of its associated gene. Different approaches report the methylation status of a CpG island based on the determination of one or a few CpG sites by assuming the homogeneity of methylation along the element. This strategy is frequently applied in both locus-specific and genome-wide studies, but often without a validation of the representativeness of the interrogated CpG site compared with the whole element. We have evaluated the predictive informativeness of the HpaII sites located in CpG islands using data from high-resolution methylome maps, which offer the possibility to assess the methylation homogeneity of each CpG island and to determine the reporter accuracy of single sites as surrogate markers. An excellent correlation was observed between the HpaII and CpG island methylation levels ($r > 0.93$). At the qualitative level, the predictive sensitivity of HpaII was >95% with >92% specificity for methylated CpG islands and >90% sensitivity with >95% specificity for unmethylated CpG islands. This analysis provides a global validation framework for strategies based on the use of the methylation-sensitive HpaII restriction enzyme.

INTRODUCTION

Epigenetic information is encoded as a heritable combination of chemical modifications of both DNA and its packaging histones (1,2). Methylation of the cytosine base within the CpG dinucleotide is the main epigenetic modification of the DNA in mammals (3,4). Most of the human genome is CpG depleted. However, this dinucleotide can be found at close to its expected frequency in small genomic regions (200 bp to a few kb) known as CpG islands (5,6). These areas are usually 'protected' from methylation and are located in the proximal promoter regions of 75% of human genes (3,4,7). Methylated CpG islands are strongly and hereditably repressed (4). Therefore, DNA methylation has been considered as a mark for long-term inactivation (4,8,9). DNA methylation patterns are characteristic of developmental stages and cell differentiation and are also intrinsically associated with multiple pathologies, being cancer a prominent example (3,10–12).

The epigenomic landscape varies markedly across tissue types and between individuals (13,14). Hence, there is not a single reference map, what represents an extraordinary challenge not only for experiment design but also for data management, analysis and interpretation. A considerable effort has been made in the last years to obtain genome-scale maps of DNA methylation and other epigenetic marks in different cell types (15–17). Ambitious initiatives, i.e. the NIH Roadmap Epigenomics Mapping Consortium (www.roadmapepigenomics.org), the Human Epigenome Project (www.epigenome.org) and the Blueprint project (http://www.blueprint-epigenome.eu) are addressed to map DNA methylation, histone modifications and other chromatin features in different cell and tissue types.

A large number of methodologies have been developed for the analysis of DNA methylation at different genomic scales (reviewed in (3,11,18,19)). All data generated until now have been obtained using techniques based on one of these three principles: methylation-sensitive endonucleases, bisulfite conversion or purification of methylated DNA by affinity/specific antibodies (18). Besides direct sequencing of bisulfite converted DNA, which is probably the reference method (20), an extraordinary cornucopia of techniques has found a niche in the Epigenetics labs. This is due to the relatively homogeneous distribution of DNA methylation (or unmethylation) within definite genomic elements. Prominent examples are CpG islands and repeat sequences in which most CpG sites within the element show similar levels of DNA methylation. This uniformity allows the extrapolation of the analysis of a single site or a few sites to the whole CpG island or repeat element. In the foremost studies, this property was instrumental to reveal the global alterations

*To whom correspondence should be addressed. Tel: +34 93 554 3050; Fax: +34 93 465 1472; Email: map@imppc.org

of DNA methylation profiles in cancer cells (21,22). Still nowadays, the analysis of a single CpG site or a few CpG sites as surrogate indicators of the DNA methylation status of the corresponding element is the most prevalent strategy in epigenetic studies at different scales. These approaches are based on either the enzymatic digestion using specific restriction endonucleases or the bisulfite transformation and offer the advantage of high throughput, high sensitivity and relative simplicity of data analysis (18,23).

Most of the studies that use surrogate markers perform some kind of validation and make a global estimation of technique's accuracy. On the other hand, the post-hoc analysis of massive data rarely includes the recognition of bona fide and counterfeit sites which precludes the direct comparison of data generated with different approaches beyond the small subset of elements validated independently. Recent advances in sequencing methods and the development of bioinformatic tools have allowed the generation of single-base resolution maps of human methylomes (24–30). The generation of these high-resolution DNA methylation maps for different cell types, including pathological situations, is likely to represent a milestone in epigenetic studies of similar impact as the sequencing of the human genome. However, an indiscriminate application of such approaches to most DNA methylation studies is nowadays unfeasible.

Assuming that bisulfite sequencing is today's gold standard in DNA methylation analysis and using published results at the genome scale (25,26) as the reference map, we have examined the accuracy of using single CpG sites as surrogate markers for the predefined CpG islands. For pragmatic reasons we report here the analysis of the CpG within the HpaII (CCGG) restriction site, frequently used in genome-scale approaches (11,18). However, other sites may be easily analysed in the same way with our pipeline.

MATERIALS AND METHODS

DNA methylation data acquisition

Data were obtained from two studies, both whole-genome single-base resolution measurements of the methylation by high-throughput bisulfite sequencing. Together they provide data from H1 human embryonic stem cells, IMR90 fetal lung fibroblasts, ADS female adipose stem cells (ADSC) and adipocytes derived from ADSC (ADS-Adi) (25,26). These studies generated 1.16, 1.18, 1.10 and 1.13 billion reads for H1, IMR90, ADSC and ADS-Adi, respectively (ADSC and ADS-Adi reported reads were originally paired-end reads but they were uncoupled and treated as single reads in this analysis). The HpaII-CpG island methylation correlation was also analysed in cell lines: iPSC derived from ADSC (ADS-iPSC), three iPSC lines derived from foreskin fibroblasts (FF-iPSC 6.9, FF-iPSC 19.7, FF-iPSC 19.11), H9 human embryonic stem cells and iPSC derived from IMR90 fibroblasts (IMR90-iPSC) (26). All reads were aligned to the human reference sequence (NCBI build 36/hg18) using the Bowtie program (31). Reads were downloaded from http://neomorph.salk.edu/human_methylome/data.html (H1 and IMR90) and http://neomorph.salk.edu/ips_methylomes/data.html (rest of cell lines), processed to be SAM-like, transformed to the BAM format and indexed using the C++ program SAMTOOLS (32).

The genomic coordinates of the CpG islands (defined according classical criteria (5): GC content of 50% or greater, length >200 bp, and a ratio >0.6 of observed number of CpG dinucleotides to the expected number) and the human genome sequence were downloaded from the UCSC Genome Browser, version hg18 (33). Additional analyses were also performed using CpG islands annotated using experimental (34) and bioinformatic criteria (35). Only reads overlapping fully or partially with these positions were considered (Supplementary Table S1). HpaII positions and CpG islands sequences were obtained by processing the human genome with Perl scripts. Each HpaII was assigned to the corresponding CpG island using the 'intersectBed' function from BEDTOOLS suite (36). A scheme of the data acquisition and processing is shown in Supplementary Figure 1A.

Methylation coefficient calculation

The methylation coefficient was calculated for CpG islands ($ß_C$) and HpaII restriction sites ($ß_H$) using a Python script (available from the authors upon request). In the sequence reads, unmethylated cytosines are visualized as thymines (*T*) due to the bisulfite conversion whereas methylated cytosines remain untransformed (*C*). Hence, the coefficient was defined as the ratio between the number of cytosines and the total number of cytosines and thymines (no. of C / (no. of C + no. of T) (Supplementary Figure 1B). This definition is equivalent to the ß value used in methylation arrays (37) and ranges from 0 (no methylation) to 1 (fully unmethylated). The CpG dinucleotide includes two cytosines, each one on one strand, and reads can cover differently each of them (i.e. if all reads covering a CpG are 5′ to 3′, only information of one cytosine is provided). For that reason, each cytosine within a CpG dinucleotide was processed as an individual genomic position and a methylation coefficient was assigned to it. Our Python script firstly scans each CpG island sequence for CG motif. Once found, it assigns the corresponding nucleotides for that position using the python library pysam and according to reads already indexed (Supplementary Figure S1) and taking into account both Watson and Crick DNA strands. Positions with <5 reads were discarded as they were considered not informative enough. In order to use only positions with trustable methylation information, an additional filter was added: positions with <5 $C+T$ were also discarded. With these filters we obtain the same coverage as reported in a previous analysis (23). Mean and standard deviation (SD) values were obtained for each CpG island taking into account only the valid positions. Only CpG islands with a minimum of informativeness (>25% CpG sites covered by at least 5 reads each) were considered, what resulted in a minimum of 10 693 informative CpG islands for the H1 cell line and

more than 26 000 for the ADS-Adi samples (Supplementary Table S1). The total number of HpaII sites included in the analysis ranged from 32 153 for the H1 cells to more than 250 000 in ADS-Adi (Supplementary Table S1). A second analysis was also performed for all CpG islands covered with at least one informative position to check for the robustness of the reporter informativeness when limited coverage is obtained. When the informativeness restrictions were eliminated, more than 21 000 CpG islands contained at least one HpaII site covered by at least 5 reads (Supplementary Table S1). It is worth to note that the numeric differences between the pairs H1/IMR90 and ADSC/ADS-Adi are caused by the different read coverage between the two studies. Scripts are available upon request.

A number of CpG islands contained more than one HpaII site in the analysed sequence (Supplementary Table S1). In these cases, the mean methylation value of all the HpaII sites was calculated. The reporter value of individual and mean $ß_H$ was calculated as the difference with the respective CpG island methylation value.

A randomized set of data was generated to evaluate possible biases due to the bimodal distribution of methylation. A virtual catalog of CpG islands matching the same size and coverage of the ones included in the study was generated and the valid CpG sites methylation coefficient values were randomly distributed along the CpG islands and used to calculate the corresponding $ß_C$. Hypothetical HpaII sites were randomly chosen matching the actual distribution. The simulation was done with H1 and IMR90 datasets.

Data analysis

The methylation value for the CpG island and HpaII sites together with structural and descriptive information like genomic coordinates or O/E ratio was stored in a mySQL Database to allow rapid retrieval and the easy establishment of relationships (Supplementary Figure S1). Scripts are available upon request. Graphs and derived calculations were generated using the statistical software R. The jitter mode was used in scatter plots with large datasets to improve dot visualization. The complete analysis was performed with the subset of CpG islands with high coverage (Supplementary Table S1) and all the CpG islands represented by at least one sequence read. Most analyses produced identical or very similar results when the filtered (high coverage) or the unfiltered sets were used. For simplicity, only data generated using the filtered set are shown here, except for those cases in which different distributions were observed (Supplementary Figure S2).

RESULTS

Characterization of CpG island DNA methylation internal heterogeneity

The $ß_C$ showed a bimodal distribution with most values near 0 (fully unmethylated) or 1 (fully methylated) with some intermediate values (Supplementary Figure S2), and confirming the enrichment for methylated CpG islands in the H1 cells as compared with the other cell lines as reported (25). Interestingly, this difference disappears when all CpG islands with at least one sequence read were included in the analysis (Supplementary Figure S2). This indicates a methylation dependent bias in the coverage of CpG islands in the H1 sample as methylated CpG islands are better covered than unmethylated ones.

CpG island methylation heterogeneity was represented as the SD of the methylation coefficient among the CpG dinucleotides contained in a CpG island (Figure 1). As a whole, the greatest variability was observed for a small population of CpG islands with intermediate methylation values indicating alternative methylation status of individual CpG positions rather than a homogeneous intermediate methylation level of the CpG sites. Next, we explored the distribution of variability according to the methylation levels. As expected, methylated CpG islands exhibited the lowest levels of internal variability (Supplementary Figure S3). Surprisingly, unmethylated CpG islands exhibited high homogeneity in H1 cells but a broader distribution in the rest of samples, indicating a more relaxed methylation profile. Intermediately methylated CpG islands exhibited higher levels of internal variability, but once again, H1 cells showed less variability (Supplementary Figure S3). Together, these results indicate a high homogeneity in the methylation profiles of methylated CpG islands. Highly unmethylated and, especially, intermediately methylated CpG islands exhibit different levels of heterogeneity, which might suggest that, for a small number of CpG islands, individual CpG sites may not be representative of the global profile.

HpaII site DNA methylation as a proxy of CpG island methylation

To evaluate the predictive value of individual CpG sites contained within a CpG island, the HpaII restriction site was selected (CCGG) as this enzyme is used by multiple locus-specific and genome-scale techniques. A total of 32 153, 77 417, 249 053 and 251 516 HpaII sites located inside the preselected CpG islands (Supplementary Table S1) were informative in H1, IMR90, ADSC and ADS-Adi cells, respectively. The symmetrical DNA methylation of HpaII sites was confirmed by comparing the methylation coefficient of both strands calculated separately (data not shown). Only HpaII sites with five or more informative reads in at least one strand were considered to have high coverage and included in the analysis.

At global scale, an excellent correlation existed between the $ß_H$ and that of its corresponding CpG island (H1, $r = 0.96$, $P < 10^{-15}$; IMR90, $r = 0.93$, $P < 10^{-15}$; ADSC, $r = 0.94$, $P < 10^{-15}$; ADS-Adi, $r = 0.94$, $P < 10^{-15}$) (Figure 2). For the rest of cell lines a high correlation was also observed ($r > 0.94$, Supplementary Figure S4). Further analysis of the data revealed that, in methylated CpG islands, HpaII sites tend to be hypermethylated (differential methylation coefficient <0) as compared with the global methylation coefficient of the respective CpG island in all the cell lines analysed (Figure 3). In unmethylated and intermediately methylated CpG islands no biases were observed.

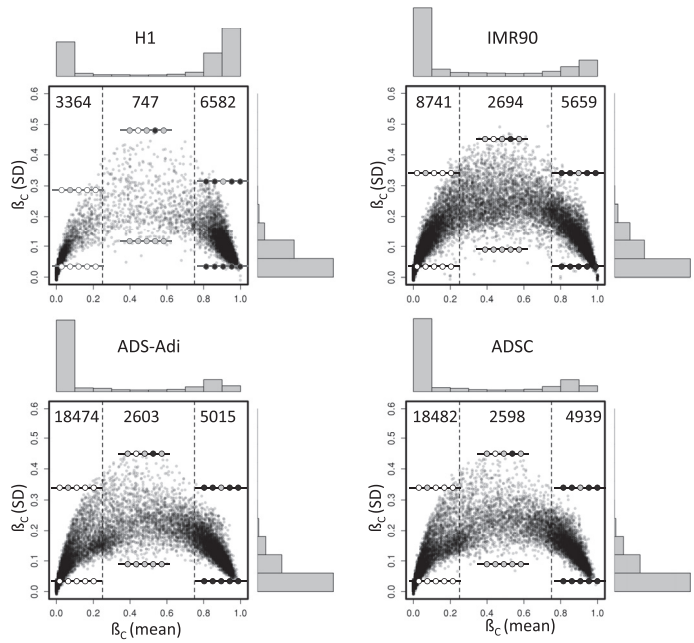


Figure 1. Homogeneity of CpG methylation in CpG islands. The mean of all informative CpG sites located inside each CpG island (CpG island methylation coefficient, $ß_C$) is plotted against the SD for four of the cell lines analysed in this study. Vertical dash lines delimit graph areas containing unmethylated ($ß_C$ mean <0.25) and methylated ($ß_C$ mean >0.75) CpG islands. The numbers of points represented in each area of the graph and the distribution histograms of both axes are shown. Illustrative DNA methylation profiles of CpG islands represented in each area are displayed using lollipop diagrams, in which empty dots represent unmethylated CpG sites, whereas gray-filled dots represent partially methylated sites and black-filled dots fully methylated sites. Only CpG islands with high coverage are displayed.

We have applied this analysis to the CpG islands annotated in the UCSC genome browser, as it represents a referential source of genomic information. Nevertheless, other studies have proposed new methods and criteria to define CpG islands. We have considered two of these studies, one based on the experimental capture of CpG islands (34) and the other on the application of hidden Markov models (35). In both cases, the high correlation was maintained (Supplementary Figure S5). Noteworthy, CpG islands identified experimentally (34) but not annotated in the UCSC showed a similar correlation even the proportion of methylated CpG islands was higher in the newly identified CpG islands than those overlapping with the classical definition (Supplementary Figure S5).

As most studies on DNA methylation report data as binary marks (methylated/unmethylated), a qualitative evaluation of the predictive value of HpaII sites was made. A Receiver Operating Characteristic (ROC) curve analysis was performed to ascertain the optimal cutoff points and accuracy (Supplementary Figure S6). In all cases the area under the curve was above 0.95. The $ß_H$ cutoff point for the methylated CpG islands ($ß_C > 0.75$) ranged from ≥0.40 to ≥0.67 (Supplementary Figure S6). The $ß_H$ cutoff point for the unmethylated CpG islands ($ß_C < 0.25$) ranged from <0.34 to <0.12. Sensitivity and specificity were ≥90% in all cases (Supplementary Figure S6). An additional evaluation was performed in which matching or unmatching scores were set when the absolute difference between the two coefficients ($|ß_C - ß_H|$) was over 0.25. Under this arbitrary criterion, 5.2% (1169 out of 32 153), 8.3% (6455 out of 77 417), 3.13% (7802 out of 249 053) and 3.01% (7575 out of 251 516) of the HpaII sites showed discordant results for H1, IMR90, ADSC and ADS-Adi cells, respectively (Figure 2). When the absolute difference value determining discordance was set to >0.5, the proportion of unmatching data was reduced to 0.9% (300 out of 32 153), 1.3% (1034 out of 77 417), 0.52% (1288 out of 249 053) and 0.50% (1249 out of 251 516), respectively.

Besides the low proportion of discordant sites, a high recurrence was observed, especially in the samples with high coverage (Supplementary Figure S7). For instance, the comparison of 3 samples showed an overlapping of 42% for 2 or more samples and of 16% for the 3 samples (absolute difference >0.25). This represents an extraordinary enrichment, as the total number of informative HpaII sites was above 250 000 and the number of discordant sites per sample is about 3%, which implies that if discordance was randomly distributed, we would

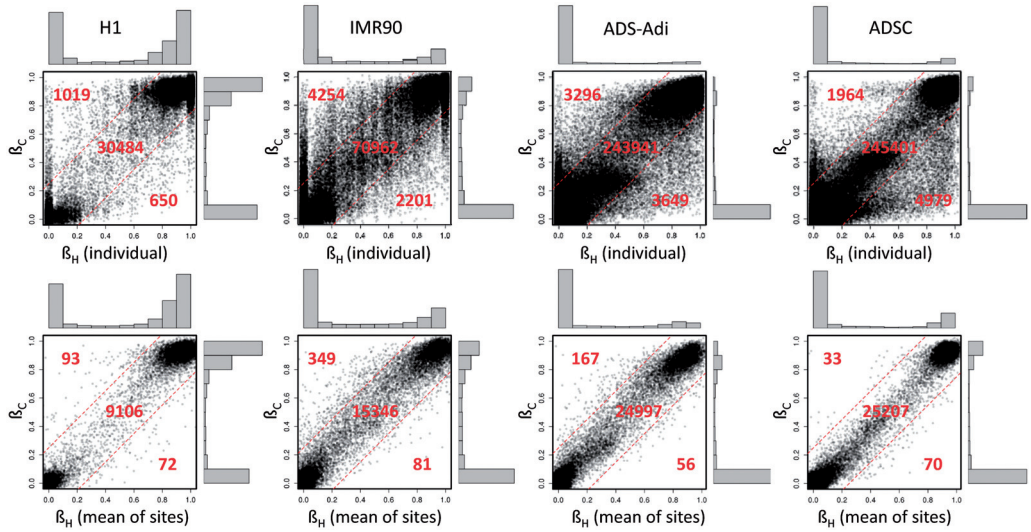


Figure 2. Correlation plots for the methylation coefficient between the HpaII and the corresponding CpG island for H1, IMR90, ADS-Adi and ADSC samples. Dash lines delimit areas with differences >0.25 between the HpaII site and the corresponding CpG island. The numbers of points represented in each area of the graph and the distribution histograms of both axes are shown. Upper panels show correlation for individual HpaII sites with the respective CpG island, lower panels depict the same correlations but comparing the mean of all HpaII sites in any given CpG island. The number of informative CpG islands for each cell line is shown in Supplementary Table S1. Plots representing additional samples and alternative CpG island definitions are shown in Supplementary Figures S4 and S5, respectively.

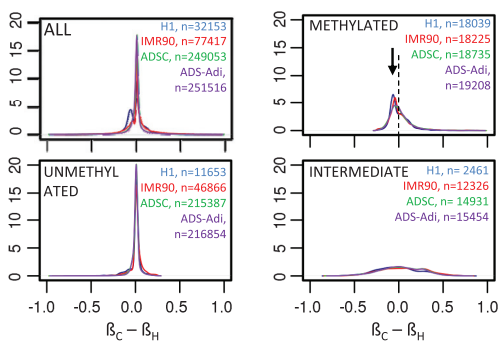


Figure 3. Density plot of the difference between the CpG island methylation coefficient and the respective HpaII methylation coefficient. A bimodal distribution was observed in H1 cells (blue) suggesting that a subpopulation of HpaII sites tend to be hypermethylated as compared with the referenced CpG island. Further exploration of the distribution according to the methylation state of the CpG island (METHYLATED, UNMETHYLATED and INTERMEDIATE) revealed a slight hypermethylation of the HpaII site (as compared with the respective CpG island) in methylated CpG islands (METHYLATED, see arrowhead).

expect about 7 HpaII sites to discordant in the 3 samples, in front of the 2268 observed.

A randomization of CpG sites among all CpG islands (see Materials and Methods) showed no correlation between the hypothetical HpaII sites and the corresponding CpG island ($r = 0.30$ and $r = 0.22$ with data from H1 and IMR90, respectively), demonstrating that the observed correlation is not explained by the bimodal distribution of DNA methylation levels (Suppementary Figure S8).

In those CpG islands with more than one informative HpaII site, the accuracy in the prediction of $ß_C$ was improved by using the average of all the HpaII sites instead of a single one. As expected, mean $ß_H$ exhibited a better correlation with CpG island methylation ($r > 0.98$, Figure 2 lower panels) and the difference between actual $ß_C$ and individual $ß_H$ was dramatically reduced when the mean HpaII methylation was used (Figure 4). At the qualitative level the discordant points were reduced to 1.8, 2.7, 0.95 and 0.89% in H1, IMR90, ADSC and ADS-Adi cells, respectively (discordant points are those with differences in the methylation coefficient above 0.25). These figures were 0.2, 0.1, 0.05 and 0.04% when the difference was set to >0.5.

As a whole, these results indicate that measurement of DNA methylation in HpaII sites (individually or pooled) is a good surrogate of the methylation state of the CpG island, especially when more than one site is used per CpG island.

Features of discordant sites

To get insights into the putative determinants of the atypical methylation in discordant HpaII sites (those points with an absolute difference between the $ß_C$ and the $ß_H > 0.25$) we explored some genomic features (distance from the HpaII to the nearest extreme of the CpG island, CpG island length, total number of CpG

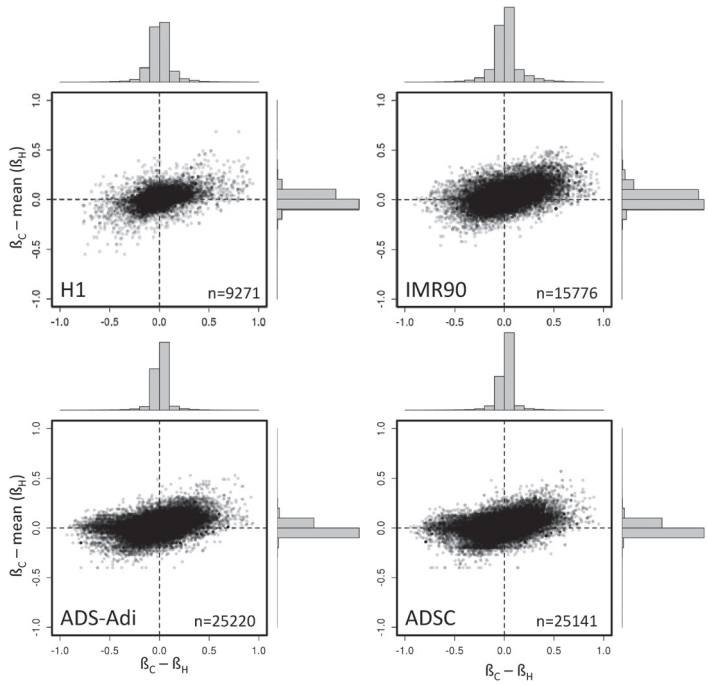


Figure 4. Improvement in the predictive value of HpaII methylation coefficient when using the mean methylation of all HpaII sites within a CpG island. The difference between the CpG island and the HpaII site methylation (*X* axis) exhibits a wider distribution as compared with the use of the mean of all HpaII sites (*Y* axis) contained in the CpG island (see Supplementary Table S1 for the distribution of HpaII in the CpG islands analysed in this study).

positions in the CpG island and Observed versus Expected CG ratio of the CpG island) of these sites separately for hypermethylated and hypomethylated HpaII sites as compared with the respective CpG island (Supplementary Figure S9). No major differences were observed in the distribution of these features, with the exception of a slight increase of discordant sites (both hypo and hypermethylated) in larger CpG islands (Supplementary Figure S9), especially in H1 cells.

Finally, we analysed the possible occurrence of single-nucleotide polymorphisms (SNPs) affecting the CpG dinucleotide in the HpaII site that could explain some discrepancies. SNP data were obtained from the dbSNP build 135 at ftp.ncbi.nih.gov/snp. About 3% of all HpaII sites may contain polymorphisms affecting the detection of methylation. Discordant HpaII sites exhibited a highly significant enrichment of SNPs (up to 9%) as compared with concordant sites (Supplementary Table S2), which suggests that a fraction of the discordances may be due to genetic variation.

DISCUSSION

Bisulfite sequencing is considered the gold standard in DNA methylation analysis but it is a cumbersome procedure. The reading of electropherograms is an intricate task not free of subjective interpretations and technical biases (38,39). Hence, the use of alternative approaches relying on the determination of one or a few CpG sites has been a common detour in a large number of studies reporting differential methylation of CpG islands. Methylation-specific PCR (MSP) is probably the most employed alternative method in targeted studies whereas in genome-wide studies, methylation-sensitive restriction enzymes and specific probes for the methylated/unmethylated sequence are frequently used (11,18). Each method has specific advantages and disadvantages and can be more or less suitable depending on the application. The features of the different methods and their performance in relation to different parameters (i.e. amount of material required, resolution, quantitativeness, genomic coverage, computational cost, etc.) have been addressed in detail elsewhere (11,18,23,40–44).

In the last couple of years, a few studies have performed whole-genome shotgun bisulfite sequencing (WGSBS) of a reduced number of human samples (24–30) providing an excellent framework to inquire the appropriateness of other reduced complexity approaches from a theoretical (23) and a practical (44,45) point of view. Here we have analysed the concordance of HpaII site methylation with that of the inclusive CpG island using data from two WGSBS studies in human samples (25,26). The choice of HpaII was obvious as it is present in 94% of CpG islands in the human genome and together with SmaI (CCCGGG), that includes the

HpaII site (CCGG), is the most widely used methylation-sensitive restriction enzyme for DNA methylation studies (18,45,46). Moreover, we cannot elude historical reasons as this site was instrumental in the discovery and initial characterization of CpG islands originally recognized as HpaII Tiny Fragment (HTF) islands (47,48).

In our analysis of Lister data, the methylation of HpaII sites embedded in CpG islands appears as an excellent reporter of the overall methylation of the CpG island, in agreement with other studies using different strategies (45). Beyond the quantitative correlation, it is important to note that at the qualitative level (the most usual way to report DNA methylation states), HpaII exhibited a very high predictive value that was extremely accurate in CpG islands containing two or more HpaII sites. The high correlation appears to be independent of the global methylation levels, that were quite different among samples, and tissue type. Moreover it is preserved in CpG islands defined by alternative criteria (34,35) and not included in the list of the UCSC genome browser (Supplementary Figure S5). Noteworthy, our analysis also shows that a higher proportion of the CpG islands defined by alternative criteria are methylated as compared with the annotated in UCSC genome browser (Supplementary Figure S5, vertical histograms). This is consistent with the enrichment of differentially methylated regions (DMRs) in these CpG islands (35,49). Moreover, most of DMRs in colon cancer cells (50) overlap with CpG islands defined using hidden Markov models (35), which expands the applicability of this approach.

A limitation in the use of HpaII is that not all CpG islands are represented. In the human genome, 1718 out of 28 226 (6.1%) human CpG islands do not contain a HpaII site. Nevertheless, achieving a similar coverage by WGSBS represents an extraordinary challenge as the one faced in Lister *et al.* (25,26), that limits the application of this kind of approach to reduced sets of samples. Other biases as base composition and size of the fragments should be also considered in the massive application of approaches based on HpaII, but these issues have been already addressed in other studies that have demonstrated that they have a minimal impact on the results (45). When we analysed different features of the CpG islands, no major bias that could affect the representativeness of the HpaII was observed, maybe with the exception of the H1 cells, in which discordant sites tended to be more frequent in longer CpG islands with a high number of CpGs (Supplementary Figure S9). This bias was not associated with the position of the HpaII site within the CpG island (near or far from the edge). We may speculate that many of those discrepancies could be due to the relative abundance of 5-hydroxymethylcytosine (5hmC) in the promoter regions of developmentally regulated genes in embryonic stem cells (51). 5hmC, that cannot be distinguished from methylated cytosine by bisulfite sequencing, is considered to be a transitional state in active demethylation (reviewed in (52)) and hence, it is more likely to confer a heterogeneous methylation profile to the CpG island. Noteworthy, the methylation level of the HpaII site appears to exaggerate the methylation state of the CpG island in methylated CpG islands (Figure 3), whereas in unmethylated CpG islands no bias was observed. This subtle overestimation of methylation is no longer observed when the mean of all HpaII sites is used (data not shown).

The procedure used by us is not limited to strategies based on restriction enzymes but it may be also applied for the design of reduced complexity strategies. In example, information about the representativeness of each CpG site in regard to the respective genomic element may help in the design and selection of specific probes to analyse DNA methylation using hybridization microarrays with bisulfite transformed DNA (i.e. the Infinium platform from Illumina).

In this report we have limited the analysis to CpG islands as they constitute distinct genomic elements with an important and definite functional role in gene regulation. On the other hand, HpaII sites are also frequent outside CpG islands, implying that other genomic elements may be also analysed using similar strategies provided they exhibit homogeneous methylation profiles. One example could be CpG island shores. Currently, CpG island shores are defined as regions within 2000 bp but not inside CpG island (50). In cell differentiation and cancer the methylation profiles of CpG island shores appear to be more plastic than in CpG islands (49,50), becoming a preferential target of genome-wide studies. The availability of new methylomes for different cell types should allow the evaluation of new surrogate markers amenable for other genomic elements beyond CpG islands. In turn, homogeneity of DNA methylation profiles along genomic regions may contribute to define previously unrecognized epigenetic domains as putative functional elements. Cancer cells may represent a preferential target for this type of studies, but the pervasive cell heterogeneity of most tumors involves an additional level of difficulty in the analysis and interpretation of partial methylation. In this case, concordance of CpG methylation in CpG sites within the same read may coexists with methylation heterogeneity between reads, what most probably should be interpreted as an indicator of cell heterogeneity.

In summary, our analysis provides a global validation of strategies based on the use of the methylation-sensitive restriction enzyme HpaII. This validation can be extended to other similar reduced complexity approaches. Besides the high informativeness and coverage offering these alternative approaches, their principal advantage is the drastic reduction in costs not only in expenses associated with data generation (wet lab experiments) but also in computational analysis (18,53–55). The systematic application of reduced complexity methods in combination with microarrays or next-generation sequencing in studies that are not intended to obtain full methylomes will thrust the generation of epigenomic maps with an excellent benefit-cost ratio.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2 and Supplementary Figures 1–9.

ACKNOWLEDGEMENTS

We thank Mireia Jordà, Sergi Lois and Anna Díez for helpful discussions and advice.

FUNDING

PFIS fellowship from Fondo de Investigación Sanitaria [FIS to V.B.]; Spanish Ministry of Science and Innovation [SAF2008/1409, SAF2011/23638 and CSD2006/49]; Generalitat de Catalunya [2009 SGR 1356]. Funding for open access charge: Spanish Ministry of Science and Innovation [SAF2011/23638].

Conflict of interest statement. None declared.

REFERENCES

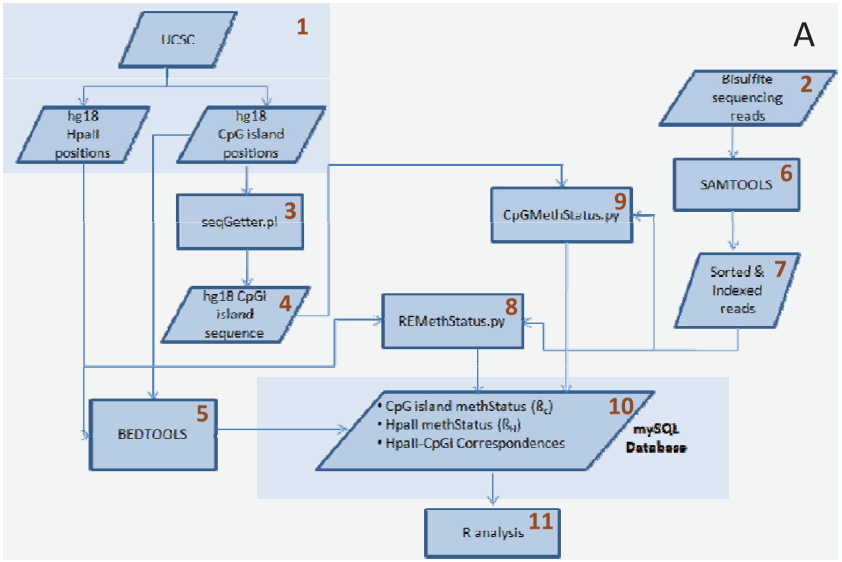
1. Bernstein,B.E., Meissner,A. and Lander,E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
2. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
3. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.
4. Goll,M.G. and Bestor,T.H. (2005) Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, **74**, 481–514.
5. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
6. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
7. Fazzari,M.J. and Greally,J.M. (2004) Epigenomics: beyond CpG islands. *Nat. Rev. Genet.*, **5**, 446–455.
8. Jaenisch,R. and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33**, 245–254.
9. Yoder,J.A., Walsh,C.P. and Bestor,T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*, **13**, 335–340.
10. Feinberg,A.P. and Tycko,B. (2004) The history of cancer epigenetics. *Nat. Rev. Cancer*, **4**, 143–153.
11. Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.*, **8**, 286–298.
12. Jones,P.A. and Baylin,S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
13. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
14. Feinberg,A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.
15. Milosavljevic,A. (2011) Emerging patterns of epigenomic variation. *Trends Genet.*, **27**, 242–250.
16. Beck,S. and Rakyan,V.K. (2008) The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, **24**, 231–237.
17. Weber,M. and Schubeler,D. (2007) Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell Biol.*, **19**, 273–280.
18. Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
19. Jorda,M. and Peinado,M.A. (2010) Methods for DNA methylation analysis and applications in colon cancer. *Mutat. Res.*, **693**, 84–93.
20. Clark,S.J., Statham,A., Stirzaker,C., Molloy,P.L. and Frommer,M. (2006) DNA methylation: bisulphite modification and analysis. *Nat. Protoc.*, **1**, 2353–2364.
21. Gama-Sosa,M.A., Wang,R.Y., Kuo,K.C., Gehrke,C.W. and Ehrlich,M. (1983) The 5-methylcytosine content of highly repeated sequences in human DNA. *Nucleic Acids Res.*, **11**, 3087–3095.
22. Feinberg,A.P. and Vogelstein,B. (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, **301**, 89–92.
23. Harris,R.A., Wang,T., Coarfa,C., Nagarajan,R.P., Hong,C., Downey,S.L., Johnson,B.E., Fouse,S.D., Delaney,A., Zhao,Y. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.
24. Laurent,L., Wong,E., Li,G., Huynh,T., Tsirigos,A., Ong,C.T., Low,H.M., Kin Sung,K.W., Rigoutsos,I., Loring,J. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.
25. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.-M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
26. Lister,R., Pelizzola,M., Kida,Y.S., Hawkins,R.D., Nery,J.R., Hon,G., Antosiewicz-Bourget,J., O'Malley,R., Castanon,R., Klugman,S. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.
27. Hansen,K.D., Timp,W., Bravo,H.C., Sabunciyan,S., Langmead,B., McDonald,O.G., Wen,B., Wu,H., Liu,Y., Diep,D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.
28. Li,Y., Zhu,J., Tian,G., Li,N., Li,Q., Ye,M., Zheng,H., Yu,J., Wu,H., Sun,J. *et al.* (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.*, **8**, e1000533.
29. Heyn,H., Vidal,E., Sayols,S., Sanchez-Mut,J.V., Moran,S., Medina,I., Sandoval,J., Simo-Riudalbas,L., Szczesna,K., Huertas,D. *et al.* (2012) Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient. *Epigenetics*, **7**, 542–550.
30. Heyn,H., Li,N., Ferreira,H.J., Moran,S., Pisano,D.G., Gomez,A., Diez,J., Sanchez-Mut,J.V., Setien,F., Carmona,F.J. *et al.* (2012) Distinct DNA methylomes of newborns and centenarians. *Proc. Natl Acad. Sci. USA.*, **109**, 10522–10527.
31. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
32. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
33. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
34. Illingworth,R.S., Gruenewald-Schneider,U., Webb,S., Kerr,A.R., James,K.D., Turner,D.J., Smith,C., Harrison,D.J., Andrews,R. and Bird,A.P. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.*, **6**, e1001134.
35. Wu,H., Caffo,B., Jaffee,H.A., Irizarry,R.A. and Feinberg,A.P. (2010) Redefining CpG islands using hidden Markov models. *Biostatistics*, **11**, 499–514.
36. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
37. Du,P., Zhang,X., Huang,C.C., Jafari,N., Kibbe,W.A., Hou,L. and Lin,S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.
38. Warnecke,P.M., Stirzaker,C., Melki,J.R., Millar,D.S., Paul,C.L. and Clark,S.J. (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res.*, **25**, 4422–4426.
39. Reins,J., Mossner,M., Richter,L., Kmetsch,A., Thiel,E., Haase,D. and Hofmann,W.K. (2011) [Letter to the editor]: whole-genome amplification of sodium bisulfite-converted DNA can substantially impact quantitative methylation analysis using pyrosequencing. *Biotechniques*, **50**, 161–164.
40. Pelizzola,M. and Ecker,J.R. (2011) The DNA methylome. *FEBS Lett.*, **585**, 1994–2000.

41. Nair,S.S., Coolen,M.W., Stirzaker,C., Song,J.Z., Statham,A.L., Strbenac,D., Robinson,M.W. and Clark,S.J. (2011) Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics*, **6**, 34–44.
42. Robinson,M.D., Statham,A.L., Speed,T.P. and Clark,S.J. (2010) Protocol matters: which methylome are you actually studying? *Epigenomics*, **2**, 587–598.
43. Bock,C., Tomazou,E.M., Brinkman,A.B., Müller,F., Simmer,F., Gu,H., Jäger,N., Gnirke,A., Stunnenberg,H.G. and Meissner,A. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.
44. Rajendram,R., Ferreira,J.C., Grafodatskaya,D., Choufani,S., Chiang,T., Pu,S., Butcher,D.T., Wodak,S.J. and Weksberg,R. (2011) Assessment of methylation level prediction accuracy in methyl-DNA immunoprecipitation and sodium bisulfite based microarray platforms. *Epigenetics*, **6**, 410–415.
45. Suzuki,M., Jing,Q., Lia,D., Pascual,M., McLellan,A. and Greally,J.M. (2010) Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol.*, **11**, R36.
46. Gupta,R., Nagarajan,A. and Wajapeyee,N. (2010) Advances in genome-wide DNA methylation analysis. *Biotechniques*, **49**, iii–xi.
47. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
48. Bird,A. (2009) On the track of DNA methylation: an interview with Adrian Bird by Jane Gitschier. *PLoS Genet.*, **5**, e1000667.
49. Doi,A., Park,I.H., Wen,B., Murakami,P., Aryee,M.J., Irizarry,R., Herb,B., Ladd-Acosta,C., Rho,J., Loewer,S. *et al.* (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.*, **41**, 1350–1353.
50. Pastor,W.A., Pape,U.J., Huang,Y., Henderson,H.R., Lister,R., Ko,M., McLoughlin,E.M., Brudno,Y., Mahapatra,S., Kapranov,P. *et al.* (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, **473**, 394–397.
51. Williams,K., Christensen,J. and Helin,K. (2012) DNA methylation: TET proteins-guardians of CpG islands? *EMBO Rep.*, **13**, 28–35.
52. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.
53. Singer,M., Boffelli,D., Dhahbi,J., Schonhuth,A., Schroth,G.P., Martin,D.I. and Pachter,L. (2010) MetMap enables genome-scale Methyltyping for determining methylation states in populations. *PLoS Comput. Biol.*, **6**, e1000888.
54. Krueger,F., Kreck,B., Franke,A. and Andrews,S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.
55. Martin,D.I., Singer,M., Dhahbi,J., Mao,G., Zhang,L., Schroth,G.P., Pachter,L. and Boffelli,D. (2011) Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Res.*, **21**, 2049–2057.

Supplementary Figures and Tables

Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale

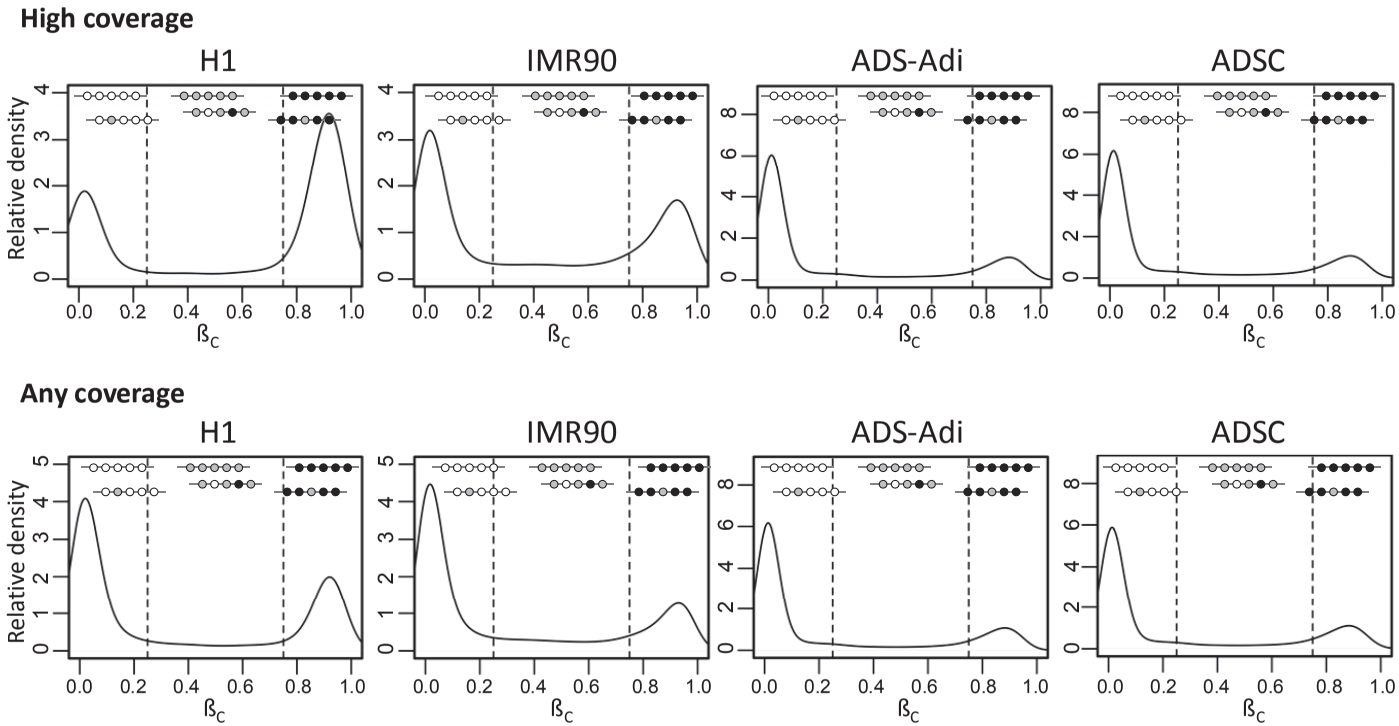
Víctor Barrera, Miguel A. Peinado
Institute of Predictive and Personalized Medicine of Cancer (IMPPC),
Badalona, Barcelona, Spain



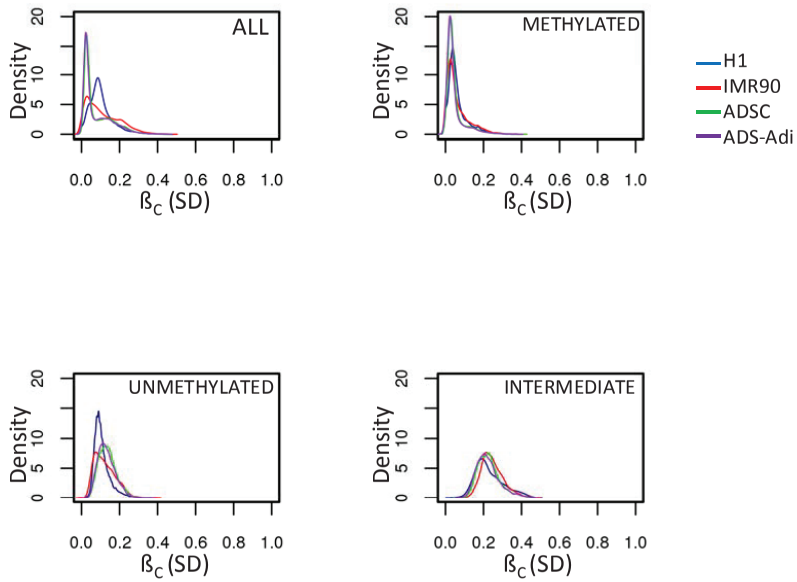
B

```
Genome browser        acta[cg]at[cg]atagat[cg]acctagg[cg]gtgga t[cg]ttca
Bisulfite converted   atta[tg]at[tg]atagattgatttaggt[tg]ggtgga t[tg]ttta

READs                  tta[cg]at[cg]atag        aggt[cg]ggtgga t[tg]t
                      atta[cg]at[cg]ata           t[tg]ggtgga t[tg]ttt
                            t[cg]atagattgat    gt[cg]ggtgga t[tg]tt
                         a[tg]at[cg]atagat        ggtgga t[tg]ttta
                                  agattgatttagg
                            [cg]at[cg]atagat  atttaggt[cg]ggtgg
                      atta[tg]at[cg]ataga          t[cg]ggtgga t[tg]ttta
COUNT     Total            5   6       2         5       5
          C                3   6       0         4       0
          T                2   0       2         1       5
Methylation coefficient  0.6  1       0        0.8      0

CpG island Methylation coefficient (ß_C): 0.48
HpaII methylation coefficient (ß_H): 0.8
Reporter value: ß_C - ß_H = -0.32
Absolute difference: |ß_C - ß_H| = 0.32
```

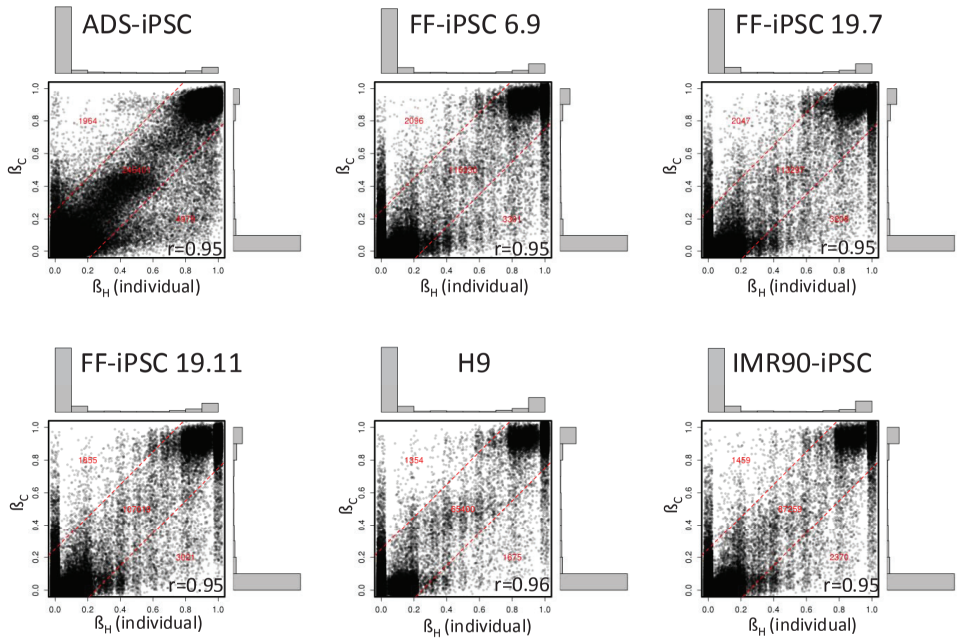
Suppl. Fig. 1. **(A)** Scheme of the data acquisition and processing. **1**. HpaII and CpG island coordinates obtained from UCSC databases (hg18 assembly, http://genome.ucsc.edu). **2**. Bisulfite reads downloaded from the original studies (ref. 25 and 26) web site (see Methods). **3**. Perl script for CpG island sequence obtention . **4**. CpG island sequence file. **5**. Use of intersectBED function from BEDTOOLS Suite to assign each HpaII site to its corresponding CpG island. **6**. Use of SAMTOOLS utilities to process the bisulfite reads from a sam file to a sorted and indexed bam file. **7**. bam file with reads information. **8**. Python scripts for HpaII methylation data calculation. Assigns methylation values ($ß_H$) to each HpaII site taking as input HpaII coordinates (1) and processed bisulfite reads (7). **9**. Python scripts for CpG island methylation data calculation. Reports mean methylation coefficient ($ß_C$) and standard deviation for each CpG island. It takes as input, CpG island sequence (4) for CpG dinucleotide search and processed bisulfite reads (7). **10**. Data storage in a MySQL Database. **11**. Statistical analysis.
(B) Simulated example to illustrate calculation of methylation coefficients. CpGs are underlined and the HpaII site is in a box. Calculations were performed for the two DNA strands (only one is shown here) and integrated into a single measure for each CpG site. The example contains 5 informative CpG sites and the values for the different parameters used in this study are calculated and illustrates a discordant site HpaII is hypermethylated (0.8) as compared with the CpG island (0.48).



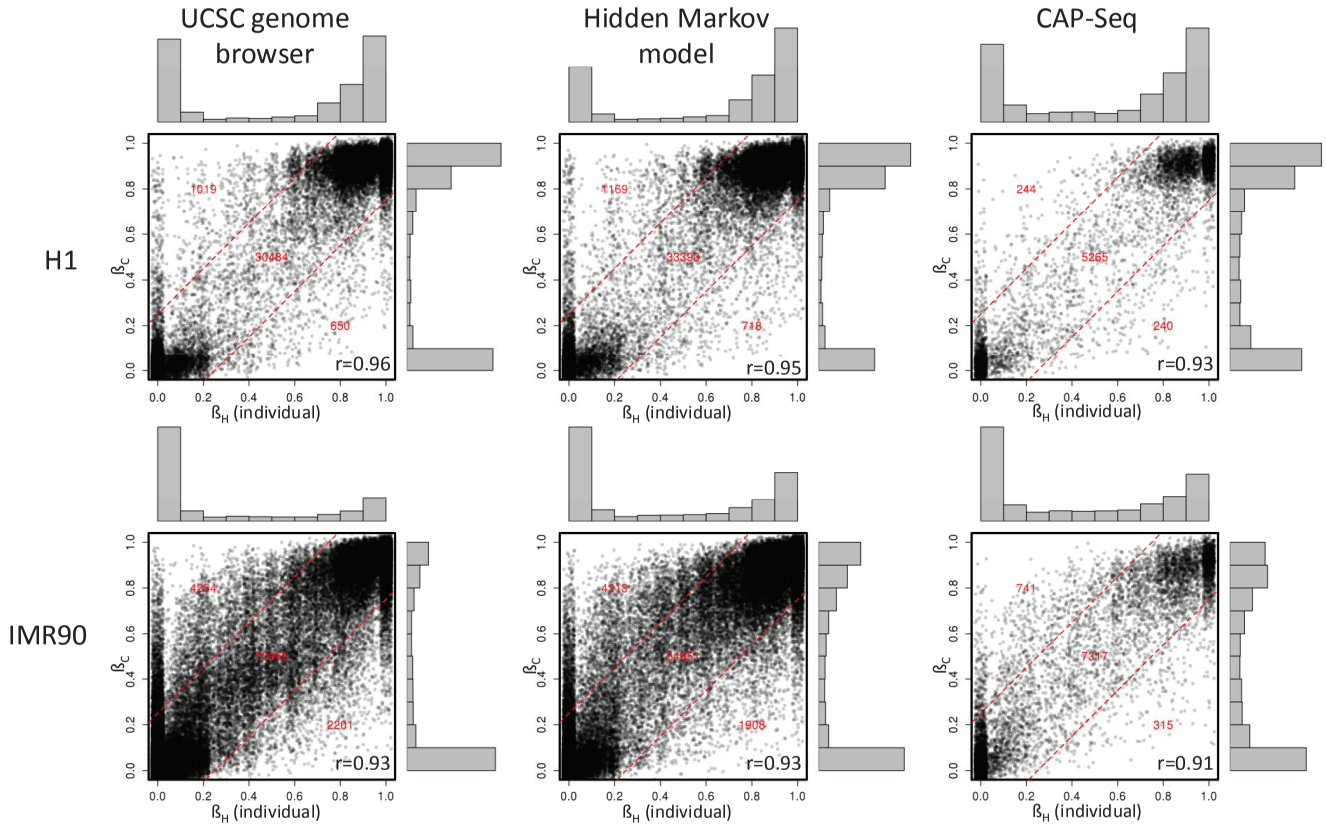
Suppl. Fig. 2: Distribution of the methylation coefficient of CpG islands with high coverage (upper panels) and all CpG islands irrespective of the coverage (lower panels) in the four cell lines. Vertical dash lines delimit graph areas containing unmethylated ($ß_C$ mean <0.25) and methylated ($ß_C$ mean >0.75) CpG islands. Illustrative DNA methylation profiles of CpG islands represented in each region displayed using lollipop diagrams, in which empty dots represent unmethylated CpG sites, while grey filled dots represent partially methylated sites and black filled dots fully methylated sites.



Suppl. Fig. 3: Density plots for the Standard Deviation (SD) of the CpG island methylation coefficient in CpG islands according to the mean coefficient. METHYLATED, $ß_C > 0.75$; UNMETHYLATED, $ß_C < 0.25$; INTERMEDIATE, $ß_C$: 0.25 to 0.75.



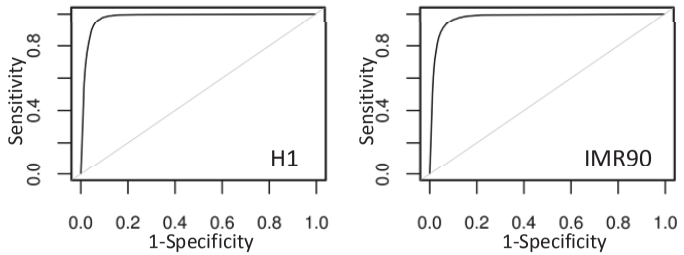
Suppl. Figure 4. Correlation plots for the methylation coefficient between the HpaII ($ß_H$) and the corresponding CpG island ($ß_C$) for six cell lines. Dash lines delimit areas with differences >0.25 between the HpaII site and the corresponding CpG island, the number of dots in each area is indicated. Histogram on top and to the right of each graph indicate the distribution of points along the axes. Only CpG islands with high coverage were considered.



Suppl. Figure 5. Correlation plots for the methylation coefficient between the HpaII ($ß_H$) and the corresponding CpG island ($ß_C$) for H1 and IMR90 cell lines in CpG islands as defined by classical criteria (UCSC browser), hidden Markov models (reference 35) and detected by CAP-Seq (reference 34). In the last case, only those CpG islands not included in the UCSC list are displayed. Dash lines delimit areas with differences >0.25 between the HpaII site and the corresponding CpG island, the number of dots in each area is indicated. The numbers of points represented in each area of the graph and the distribution histograms of both axes are shown. Only CpG islands with high coverage were included in this analysis.

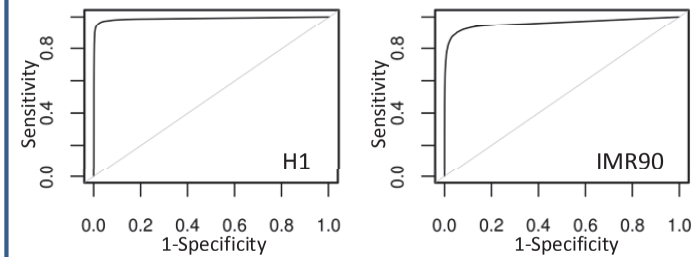
Barrera V 2012

Methylated CpG islands ($ß_c > 0.75$)



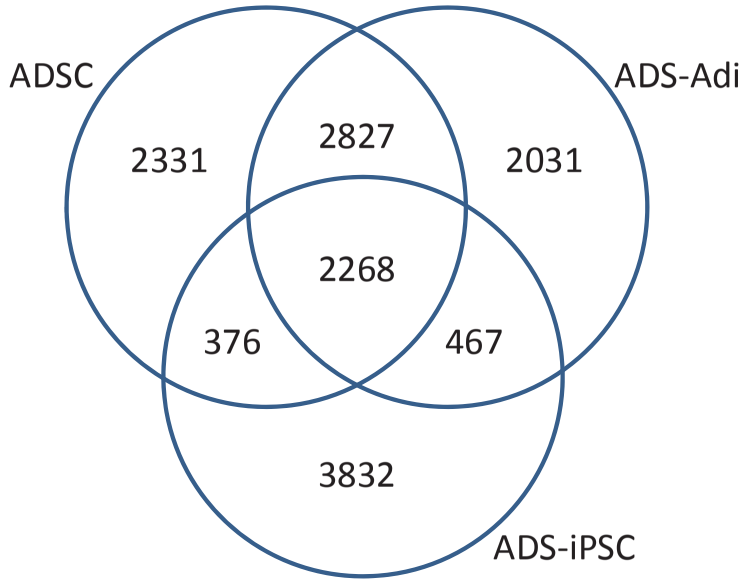
$ß_c > 0.75$ (METHYLATED)	H1	IMR90	ADS-Adi	ADSC
CUTOFF ($ß_H$)	≥0.67	≥0.56	≥0.43	≥0.40
SENSITIVITY (%)	96.1	95.0	98.1	98.3
SPECIFICITY (%)	92.7	91.9	95.9	95.6
AREA UNDER THE CURVE	0.976	0.974	0.990	0.990

Unmethylated CpG islands ($ß_c < 0.25$)

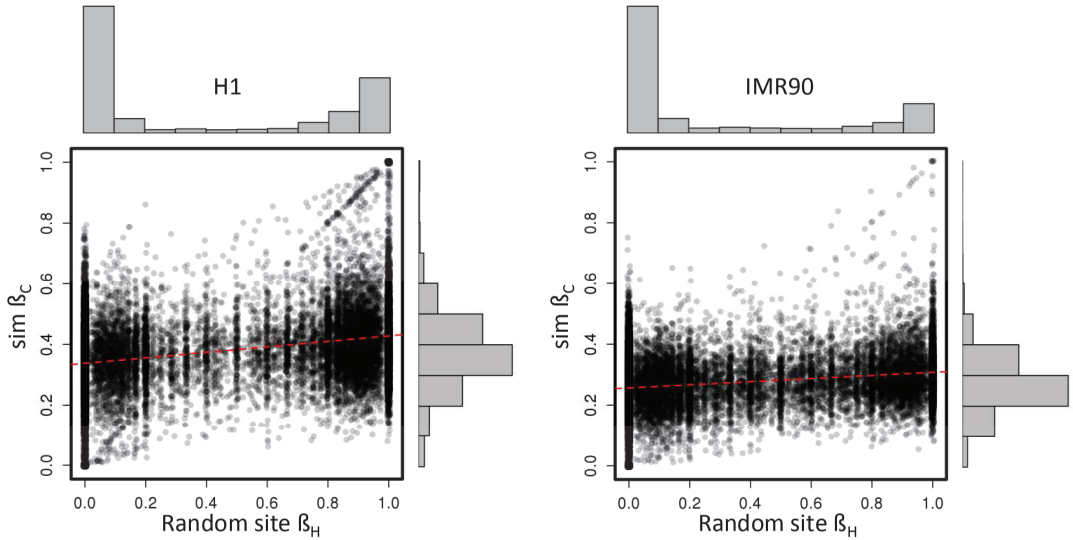


$ß_c < 0.25$ (UNMETHYLATED)	H1	IMR90	ADS-Adi	ADSC
CUTOFF ($ß_H$)	<0.34	<0.17	<0.12	<0.13
SENSITIVITY (%)	96.1	89.9	92.4	91.9
SPECIFICITY (%)	97.7	94.7	96.0	95.9
AREA UNDER THE CURVE	0.989	0.959	0.968	0.966

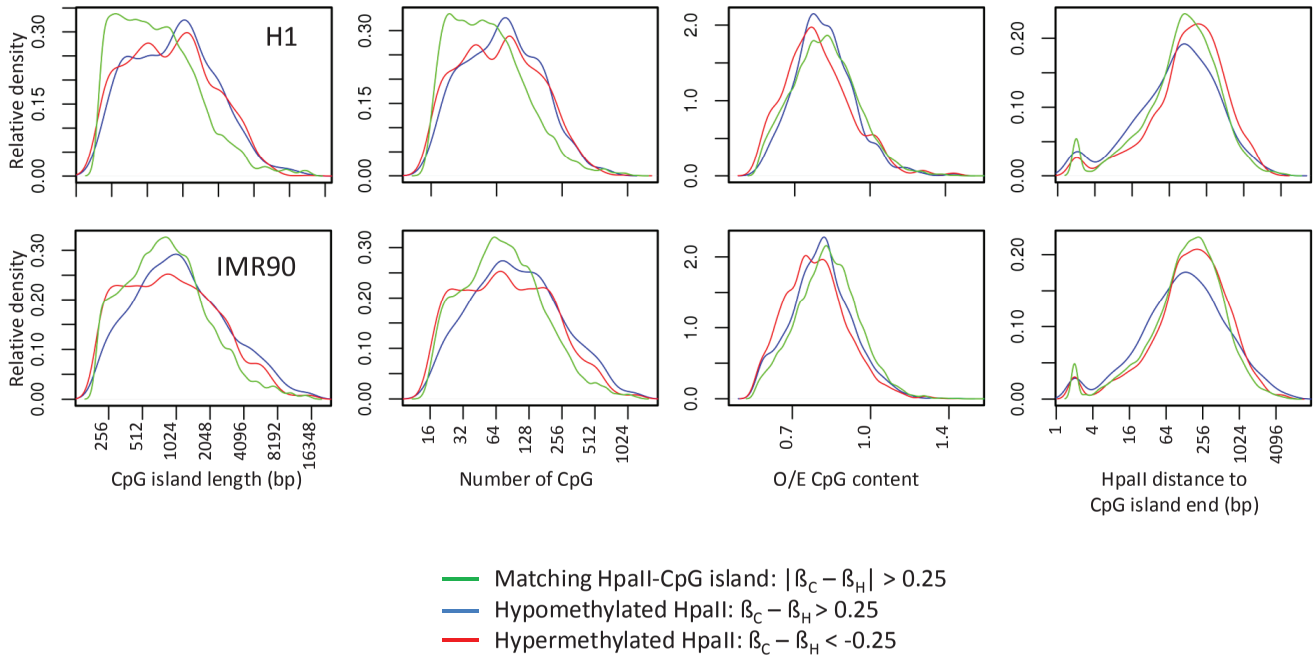
Suppl. Fig. 6. Receiver Operating Characteristic (ROC) curve analysis of the predictive value of HpaII methylation ($ß_H$) on the respective CpG island methylation coefficient ($ß_C$). To perform the analysis, a CpG island was considered methylated when the mean methylation coefficient of all informative CpG sites $ß_C$ >0.75. On the other side, it was considered unmethylated when $ß_C$ <0.25.



Suppl. fig. 7. Recurrence of discordant HpaII sites among three cell lines. Venn diagrams show the number of sites with discordant methylation levels (absolute difference >0.25) between the HpaII site and the mean coefficient of the respective CpG island. The number of informative HpaII sites was about 250,000 for all the samples.



Suppl. fig. 8: Correlation plots for the methylation coefficient between a simulated HpaII restriction site (random site ß$_H$) and a simulated CpG island (sim ß$_C$) using data from the two cell lines H1 and IMR90. Line for the linear model has been added.



Suppl. Fig. 9. Genomic features of CpG islands classified according to the concordance with the HpaII site.

SUPPLEMENTARY TABLE 1. Number of reads, CpG islands and informative HpaII sites considered in the study.

Cells	H1		IMR90		ADS-Adi		ADSC	
Total no of reads [1]	1,154,658,045		1,183,875,099		1,131,768,326		1,0985,572,398	
No of reads considered [2]	2,490,724		3,378,352		8,755,104		8,162,058	
	High Coverage[3]	All [4]	High Coverage[3]	All [4]	High Coverage[3]	All [4]	High Coverage[3]	All [4]
No of informative CpG islands	10,693	25,770	17,094	26,617	26,092	26,742	26,019	26,719
No of informative HpaII sites	32,153	68,622	77,417	113,343	251,516	252,521	249,053	250,110
CpG islands with informative HpaII sites[5]	9,271	21,002	15,776	24,038	25,220	25,567	25,141	25,521
CpG islands with 1 informative HpaII site	2,154	5,838	2,320	3,849	1,624	1,784	1,612	1,774
CpG islands with 2 informative HpaII sites	2,168	4,925	2,773	4,263	2,253	2,327	2,262	2,366
CpG islands with >2 informative HpaII sites	4,949	10,239	10,683	15,926	21,343	21,456	21,267	21,381

[1]From Lister *et al.* (references 25 and 26)
[2]Overlapping with CpG islands
[3]CpG islands with a >25% informative CpG sites. A CpG site is considered informative when it is covered by at least five reads
[4]All CpG islands with at least 1 informative CpG position (≥5 informative reads) considered
[5] Number of CpG islands with a HpaII site: 26,508

SUPPLEMENTARY TABLE 2. Frequency of SNPs in HpaII sites classified according to the difference of methylation with the respective CpG island

Cells	H1	IMR90
Concordant[1] (SNP/total)	1025/30484	1810/70962
	3.4%	2.5%
Discordant[2] (SNP/total)	159/1669	310/6415
	9.5%	4.8%
Fisher test	$p<10^{-15}$	$p<10^{-15}$

[1]Concordant: Absolute difference between CpG island methylation coefficient and HpaII methylation coefficient <0.25.
[2]Discordant: Absolute difference between CpG island methylation coefficient and HpaII methylation coefficient >0.25. SNPs data were obtained from the dbSNP build 135 at ftp.ncbi.nih.gov/snp.